Linear Regression With Shuffled Data: Statistical and Computational Limits of Permutation Recovery

Ashwin Pananjady, Martin J. Wainwright, and Thomas A. Courtade

***Abstract*— Consider a noisy linear observation model with an unknown permutation, based on observing $y = \Pi^* A x^* + w$, where $x^* \in \mathbb{R}^d$ is an unknown vector, Π^* is an unknown $n \times n$ permutation matrix, and $w \in \mathbb{R}^n$ is additive Gaussian noise. We analyze the problem of permutation recovery in a random design setting in which the entries of matrix A are drawn independently from a standard Gaussian distribution and establish sharp conditions on the signal-to-noise ratio, sample size n, and dimension d under which Π^* is exactly and approximately recoverable. On the computational front, we show that the maximum likelihood estimate of Π^* is NP-hard to compute for general d, while also providing a polynomial time algorithm when $d = 1$.**

***Index Terms*— Correspondence estimation, permutation recovery, unlabeled sensing, information-theoretic bounds, random projections.**

I. Introduction

RECOVERY of a vector based on noisy linear measurements is the classical problem of linear regression, and is arguably the most basic form of statistical estimator. A variant, the "errors-in-variables" model [3], allows for errors in the measurement matrix; classical examples include additive or multiplicative noise [4]. In this paper, we study a form of errors-in-variables in which the measurement matrix is perturbed by an unknown permutation of its rows.

More concretely, we study an observation model of the form

$$y = \Pi^* A x^* + w, \tag{1}$$

where $x^* \in \mathbb{R}^d$ is an unknown vector, $A \in \mathbb{R}^{n \times d}$ is a measurement (or design) matrix, Π^* is an unknown $n \times n$ permutation matrix, and $w \in \mathbb{R}^n$ is observation noise. We refer to the setting where $w = 0$ as the *noiseless case*. As with linear regression, there are two settings of interest, corresponding to whether the design matrix is **(i)** deterministic (the fixed design case), or **(ii)** random (the random design case).

There are also two complementary problems of interest: recovery of the unknown Π^*, and recovery of the unknown x^*. In this paper, we focus on the former problem; the latter problem is also known as unlabelled sensing [5].

The observation model (1) is frequently encountered in scenarios where there is uncertainty in the order in which measurements are taken. An illustrative example is that of sampling in the presence of jitter [6], in which the uncertainty about the instants at which measurements are taken results in an unknown permutation of the measurements. A similar synchronization issue occurs in timing and molecular channels [7]. Here, identical molecular tokens are received at the receptor at different times, and their signatures are indistinguishable. The vectors of transmitted and received times correspond to the signal and the observations, respectively, where the latter is some permuted version of the former with additive noise. A recent paper [8] also points out an application of this observation model in flow cytometry.

Another such scenario arises in multi-target tracking problems [9]. For example, in the robotics problem of simultaneous localization and mapping [10], the environment in which measurements are made is unknown, and part of the problem is to estimate relative permutations between measurements. Archaeological measurements [11] also suffer from an inherent lack of ordering, which makes it difficult to estimate the chronology. Another compelling example of such an observation model is in data anonymization, in which the order, or "labels", of measurements are intentionally deleted to preserve privacy. The inverse problem of data de-anonymization [12] is to estimate these labels from the observations.

Also, in large sensor networks, it is often the case that the number of bits of information that each sensor records and transmits to the server is exceeded by the number of bits it transmits in order to identify itself to the server [13]. In applications where sensor measurements are linear, model (1) corresponds to the case where each sensor only sends its measurement but not its identity. The server is then tasked with recovering sensor identites, or equivalently, with determining the unknown permutation.

Manuscript received October 15, 2016; revised October 12, 2017; accepted November 5, 2017. Date of publication November 21, 2017; date of current version April 19, 2018. This work was supported in part by NSF under Grant CCF-1528132 and Grant CCF-0939370 (Center for Science of Information), in part by the Office of Naval Research MURI under Grant DOD-002888, in part by the Air Force Office of Scientific Research under Grant AFOSR-FA9550-14-1-001, in part by the Office of Naval Research under Grant ONR-N00014, and in part by the National Science Foundation under Grant CIF-31712-23800. This work was presented in part at the 2016 Allerton Conference on Communication, Control and Computing [1] and is available on Arxiv [2].

A. Pananjady and T. A. Courtade are with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: ashwinpm@berkeley.edu).

M. J. Wainwright is with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA, and also with the Department of Statistics, University of California at Berkeley, Berkeley, CA 94720 USA.

Communicated by G. Moustakides, Associate Editor for Sequential Methods.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2017.2776217

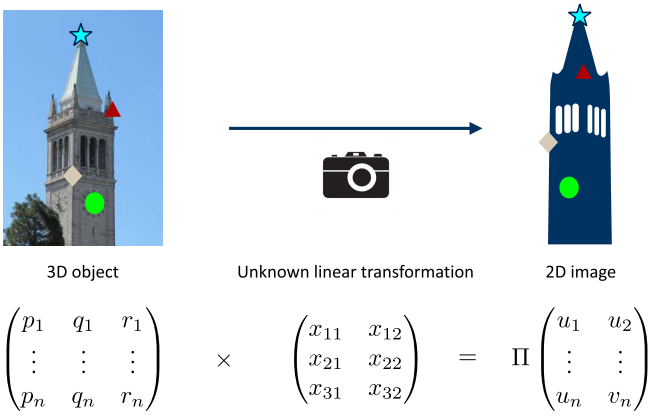

Fig. 1. Example of pose and correspondence estimation. The camera introduces an unknown linear transformation corresponding to the pose. The unknown permututation represents the correspondence between points, which is shown in the picture via coloured shapes, and needs to be estimated.

The pose and correspondence estimation problem in image processing [14], [15] is also related to the observation model (1). The capture of a 3D object by a 2D image can be modelled by an unknown linear transformation called the "pose", and an unknown permutation representing the "correspondence" between points in the two spaces. One of the central goals in image processing is to identify this correspondence information, which in this case is equivalent to permutation estimation in the linear model. An illustration of the problem is provided in Figure 1. Image stitching from multiple camera angles [16] also involves the resolution of unknown correspondence information between point clouds of images.

The discrete analog of the model (1) in which the vectors x^* and y, and the matrix A are all constrained to belong to some finite alphabet/field corresponds to the permutation channel studied by Schulman and Zuckerman [17], with A representing the (linear) encoding matrix. However, techniques for the discrete problem do not carry over to the continuous problem (1).

Another line of work that is related in spirit to the observation model (1) is the genome assembly problem from shotgun reads [18], in which an underlying vector $x^* \in \{A, T, G, C\}^d$ must be assembled from an unknown permutation of its continuous sub-vector measurements, called "reads". Two aspects, however, render it a particularization of our observation model, besides the obvious fact that x^* in the genome assembly problem is constrained to a finite alphabet: (i) in genome assembly, the matrix A is fixed and consists of shifted identity matrices that select sub-vectors of x^*, and (ii) the permutation matrix of genome assembly is in fact a block permutation matrix that permutes sub-vectors instead of coordinates as in equation (1).

It is worth noting that both the permutation recovery and vector recovery problems have an operational interpretation in applications. Permutation recovery is equivalent to "correspondence estimation" in vision tasks [14], [15], and vector recovery is equivalent to "pose estimation". In sensor network examples, permutation recovery corresponds to sensor identification [13], while vector recovery corresponds to signal estimation. Clearly, accurate permutation estimation allows for recovery of the regression vector, while the reverse may not be true. From a theoretical standpoint, such a distinction is similar to the difference between the problems of subset selection [19] and sparse vector recovery [20] in high dimensional linear regression, where studying the model selection and parameter estimation problems together helped advance our understanding of the statistical model in its entirety.

A. Related Work

Previous work related to the observation model (1) can be broadly classified into two categories – those that focus on x^* recovery, and those focussed on recovering the underlying permutation. We discuss the most relevant results below.

1) Latent Vector Estimation: The observation model (1) appears in the context of compressed sensing with an unknown sensor permutation [21]. The authors consider the matrix-based observation model $Y = \Pi^* A X^* + W$, where X^* is a matrix whose columns are composed of multiple, unknown, sparse vectors. Their contributions include a branch and bound algorithm to recover the underlying X^*, which they show to perform well empirically for small instances under the setting in which the entries of the matrix A are drawn i.i.d. from a Gaussian distribution.

In the context of pose and correspondence estimation, the paper [15] considers the noiseless observation model (1), and shows that if the permutation matrix maps a sufficiently large number of positions to themselves, then x^* can be recovered reliably.

In the context of molecular channels, the model (1) has been analyzed for the case when x^* is some random vector, $A = I$, and w represents non-negative noise that models delays introduced between emitter and receptor. Rose *et al.* [7] provide lower bounds on the capacity of such channels. In particular, their results yield closed-form lower bounds for some special noise distributions, e.g., exponentially random noise.

A more recent paper [5] that is most closely related to our model considers the question of when the equation (1) has a unique solution x^*, i.e., the identifiability of the noiseless model. The authors show that if the entries of A are sampled i.i.d. from any continuous distribution with $n \geq 2d$, then equation (1) has a unique solution x^* with probability 1. They also provide a converse showing that if $n < 2d$, any matrix A whose entries are sampled i.i.d. from a continuous distribution does not (with probability 1) have a unique solution x^* to equation (1). While the paper shows uniqueness, the question of designing an efficient algorithm to recover a solution, unique or not, is left open. The paper also analyzes the stability of the noiseless solution, and establishes that x^* can be recovered exactly when the SNR goes to infinity.

Since a version of this paper was made available online [2], a few recent papers have also considered variants of the observation model (1). Elhami *et al.* [22] show that there is a careful choice of the measurement matrix A such that it is possible to recover the vector x^* in time $\mathcal{O}(dn^{d+1})$ in the noiseless case. Hsu *et al.* [23] show that the vector x^* can be recovered efficiently in the noiseless setting (1) when the

design matrix A is i.i.d. Gaussian. They also demonstrate that in the noisy setting, it is not possible to recover the vector x^* reliably unless the signal-to-noise ratio is sufficiently high. See Section 4 of their paper [23] for a detailed comparison of their results with our own. Our own follow-up work [24] establishes the minimax rate of prediction for the more general multivariate setting, and proposes an efficient algorithm for that setting with guaranteed recovery provided some technical conditions are satisfied. Haghighatshoar and Caire [25] consider a variant of the observation model (1) in which the permutation matrix is replaced by a row selection matrix, and provide an alternating minimization algorithm with theoretical guarantees.

We also briefly compare the model (1) with the problem of vector recovery in unions of subspaces, studied widely in the compressive sensing literature [26], [27]. In the compressive sensing setup, the vector x^* lies in the union of finitely many subspaces, and must be recovered from linear measurements with a random matrix, without a permutation. In our model, on the other hand, the vector x^* is unrestricted, and the observation y lies in the union of $n!$ subspaces – one for each permutation. While the two models share a superficial connection, results do not carry over from one to the other in any obvious way. In fact, our model is fundamentally different from traditional compressive sensing, since the unknown permutation acts on the *row space* of the design matrix A. In contrast, restricting x^* to a union of subspaces (or more specifically, restricting its sparsity) influences the column space of A.

2) Latent Permutation Estimation: While our paper seems to be the first to consider permutation recovery in the linear regression model (1), there are many related problems for which permutation recovery has been studied. We mention only those that are most closely related to our work.

The problem of feature matching in machine learning [28] bears a superficial resemblance to our observation model. There, observations take the form $Y = X^* + W$ and $Y' = \Pi^* X^* + W'$, with all of (X^*, Y, Y', W, W') representing matrices of appropriate dimensions, and the goal is to recover Π^* from the tuple (Y, Y'). The paper [28] establishes minimax rates on the separation between the rows of X^* (as a function of problem parameters n, d, σ) that allow for exact permutation recovery.

The problem of statistical seriation [29] involves an observation model of the form $Y = \Pi^* X^* + W$, with the matrix X^* obeying some shape constraint. In particular, if the columns of X^* are unimodal (or, as a special case, monotone), then Flammarion *et al.* [29] establish minimax rates for the problem in the prediction error metric $\|\widehat{\Pi}\widehat{X} - \Pi^* X^*\|_F^2$ by analyzing the least squares estimator. The seriation problem without noise was also considered by Fogel *et al.* [30] in the context of designing convex relaxations to permutation problems.

Estimating network structure from co-occurrence data [31], [32] also involves the estimation of multiple underlying permutations from unordered observations. In particular, Rabbat *et al.* [31] consider the problem of estimating the structure of an unknown directed graph from multiple, unordered paths formed by simple random walks on the graph. They propose an EM-type algorithm with importance sampling to tackle the problem. The fundamental limits of such a problem were later considered by Gripon and Rabbat [32].

Permutation estimation has also been considered in other observation models involving matrices with structure, particularly in the context of ranking [33]–[35], or even more generally, in the context of *identity management* [36]. While we mention both of these problems because are related in spirit to permutation recovery, the problem setups do not bear too much resemblance to our linear model (1).

Algorithmic approaches to solving for Π^* in equation (1) are related to the multi-dimensional assignment problem. In particular, while finding the correct permutation mapping between two vectors minimizing some loss function between them corresponds to the 1-dimensional assignment problem, here we are faced with an assignment problem between subspaces. While we do not elaborate on the vast literature that exists on solving variants on assignment problems, we note that broadly speaking, assignment problems in higher dimensions are much harder than the 1-D assignment problem. A survey on the quadratic assignment problem [37] and references therein provide examples and methods that are currently used to solve these problems.

B. Contributions

Our primary contribution addresses permutation recovery in the noisy version of observation model (1), with a random design matrix A. In particular, when the entries of A are drawn i.i.d. from a standard Gaussian matrix, we show sharp conditions on the SNR under which exact permutation recovery is possible. We also derive necessary conditions for approximate permutation recovery to within a prescribed Hamming distortion.

We also briefly address the computational aspect of the permutation recovery problem. We show that the information theoretically optimal estimator we propose for exact permutation recovery is NP-hard to compute in the worst case. For the special case of $d = 1$, however, we show that it can be computed in polynomial time. Our results are corroborated by numerical simulations.

C. Organization

The remainder of this paper is organized as follows. In the next section, we set up notation and formally state the problem. In Section III, we state our main results and discuss some of their implications. We provide proofs of the main results in Section IV, deferring the more technical lemmas to the appendices.

II. Background and Problem Setting

In this section, we set up notation, state the formal problem, and provide concrete examples of the noiseless version of our observation model by considering some fixed design matrices.

A. Notation

Since most of our analysis involves metrics involving permutations, we introduce all the relevant notation in this section.

Permutations are denoted by π and permutation matrices by Π. We use $\pi(i)$ to denote the image of an element i under the permutation π. With a minor abuse of notation, we let $\mathcal{P}_n$ denote both the set of permutations on n objects as well as the corresponding set of permutation matrices. We sometimes use the compact notation y_π (or y_Π) to denote the vector y with entries permuted according to the permutation π (or Π).

We let $\mathsf{d_H}(\pi, \pi')$ denote the Hamming distance between two permutations. More formally, we have $\mathsf{d_H}(\pi, \pi') := \#\{i \mid \pi(i) \neq \pi'(i)\}$. Additionally, we let $\mathsf{d_H}(\Pi, \Pi')$ denote the Hamming distance between two permutation matrices, which is to be interpreted as the Hamming distance between the corresponding permutations.

The notation v_i denotes the ith entry of a vector v. We denote the ith standard basis vector in $\mathbb{R}^d$ by e_i. We use the notation $a_i^\top$ to refer to the ith row of A. We also use the standard shorthand notation $[n] := \{1, 2, \ldots, n\}$.

We also make use of standard asymptotic $\mathcal{O}$ notation. Specifically, for two real sequences f_n and g_n, $f_n = \mathcal{O}(g_n)$ means that $f_n \leq C g_n$ for a universal constant $C > 0$, and $f_n = \Omega(g_n)$ denotes that the relation $g_n = \mathcal{O}(f_n)$ holds. Lastly, all logarithms denoted by log are to the base e, and we use c_1, c_2, etc. to denote absolute constants that are independent of other problem parameters.

B. Formal Problem Setting and Permutation Recovery

As mentioned in the introduction, we focus exclusively on the noisy observation model in the random design setting. In other words, we obtain an n-vector of observations y from the model (1) with $n \geq d$ to ensure identifiability, and with the following assumptions:

Signal Model: The vector $x^* \in \mathbb{R}^d$ is fixed, but unknown. We note that this is different from the *adversarial* signal model of Unnikrishnan *et al.* [5], and we provide clarifying examples in Section II-C.

Measurement Matrix: The measurement matrix $A \in \mathbb{R}^{n \times d}$ is a random matrix of i.i.d. standard Gaussian variables chosen without knowledge of x^*. Our assumption on i.i.d. standard Gaussian designs easily extends to accommodate the more general case when rows of A are drawn i.i.d. from the distribution $\mathcal{N}(0, \Sigma)$. In particular, writing $A = W\sqrt{\Sigma}$, where W in an $n \times d$ standard Gaussian matrix and $\sqrt{\Sigma}$ denotes the symmetric square root of the (non-singular) covariance matrix Σ, our observation model takes the form

$$y = \Pi^* W \sqrt{\Sigma} x^* + w,$$

and the unknown vector is now $\sqrt{\Sigma}x^*$ in the model (1).

Noise Variables: The vector $w \sim \mathcal{N}(0, \sigma^2 I_n)$ represents uncorrelated noise variables, each of (possibly unknown) variance σ^2. As will be made clear in the analysis, our assumption that the noise is Gaussian also readily extends to accommodate i.i.d. σ-sub-Gaussian noise. Additionally, the permutation noise represented by the unknown permutation matrix Π^* is arbitrary.

The main recovery criterion addressed in this paper is that of exact permutation recovery, which is formally described below. We also address the problem of approximate permutation recovery.

Exact Permutation Recovery: The problem of exact permutation recovery is to recover Π^*, and the risk of an estimator is evaluated on the 0-1 loss. More formally, given an estimator of Π^* denoted by $\widehat{\Pi} : (y, A) \to \mathcal{P}_n$, we evaluate its risk by

$$\Pr\{\widehat{\Pi} \neq \Pi^*\} = \mathbb{E}\left[\mathbf{1}\{\widehat{\Pi} \neq \Pi^*\}\right], \tag{2}$$

where the probability in the LHS is taken over the randomness in y induced by both A and w.

Approximate Permutation Recovery: It is reasonable to think that recovering Π^* up to some distortion is sufficient for many applications. Such a relaxation of exact permutation recovery allows the estimator to output a $\widehat{\Pi}$ such that $\mathsf{d_H}(\widehat{\Pi}, \Pi^*) \leq D$, for some distortion D to be specified. The risk of such an estimator is again evaluated on the 0-1 loss of this error metric, given by $\Pr\{\mathsf{d_H}(\widehat{\Pi}, \Pi^*) \geq D\}$, with the probability again taken over both A and w. While our results are derived mainly in the context of exact permutation recovery, they can be suitably modified to also yield results for approximate permutation recovery.

We now provide some examples in which the noiseless version of the observation model (1) is identifiable.

C. Illustrative Examples of the Noiseless Model

In this section, we present two examples to illustrate the problem of permutation recovery and highlight the difference between our signal model and that of Unnikrishnan *et al.* [5].

Example 1: Consider the noiseless case of the observation model (1). Let v_i, v_i' $(i = 1, 2, \ldots, d)$ represent i.i.d. continuous random variables, and form the design matrix A by choosing

$$a_{2i-1}^\top := v_i e_i^\top \text{ and } a_{2i}^\top = v_i' e_i^\top, \quad i = 1, 2, \ldots, d.$$

Note that $n = 2d$. Now consider our fixed but unknown signal model for x^*. Since the permutation is arbitrary, our observations can be thought of as the unordered set $\{v_i x_i^*, v_i' x_i^* \mid i \in [d]\}$. With probability 1, the ratios $r_i := v_i / v_i'$ are distinct for each i, and also $v_i x_i^* \neq v_j x_j^*$ with probability 1, by assumption of a fixed x^*. Therefore, there is a one to one correspondence between the ratios r_i and x_i^*. All ratios are computable in time $\mathcal{O}(n^2)$, and x^* can be exactly recovered. Using this information, we can also exactly recover Π^*.

Example 2: A particular case of this example was already observed by Unnikrishnan *et al.* [5], but we include it to illustrate the difference between our signal model and the adversarial signal model. Form the fixed design matrix A by including 2^{i-1} copies of the vector e_i among its rows. We therefore[1] have $n = \sum_{i=1}^{d} 2^{i-1} = 2^d - 1$.

Our observations therefore consist of 2^{i-1} repetitions of x_i^* for each $i \in [d]$. The value of x_i^* can therefore be recovered by simply counting the number of times it is repeated, with our choice of the number of repetitions also accounting for cases when $x_i^* = x_j^*$ for some $i \neq j$. Notice that we can now recover *any* vector x^*, even those chosen adversarially with knowledge of the A matrix. Therefore, such a design

[1] Unnikrishnan *et al.* [5] proposed that e_i be repeated i times, but it is easy to see that this does not ensure recovery of an adversarially chosen x^*.

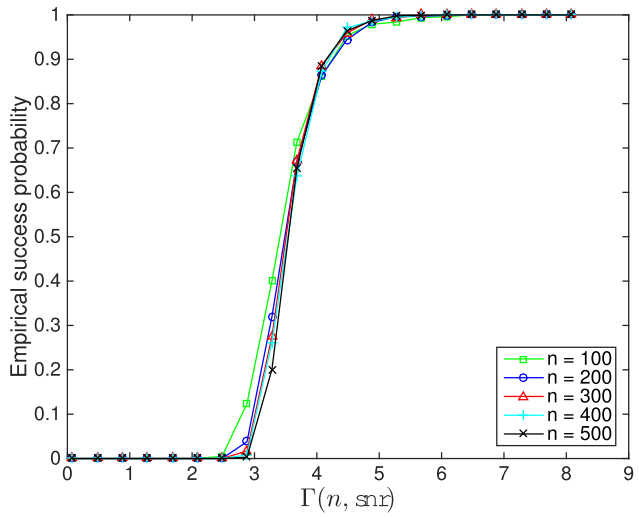

Fig. 2. Empirical frequency of the event $\{\widehat{\Pi}_{\mathsf{ML}} = \Pi^*\}$ over 1000 independent trials with $d = 1$, plotted against $\Gamma(n, \mathsf{snr})$ for different values of n. The probability of successful permutation recovery undergoes a phase transition as $\Gamma(n, \mathsf{snr})$ varies from 3 to 5. This is consistent with the prediction of Theorems 1 and 2.

matrix allows for an *adversarial* signal model, in the flavor of compressive sensing [20].

Having provided examples of the noiseless observation model, we now return to the noisy setting of Section II-B, and state our main results.

III. Main Results

In this section, we state our main theorems and discuss their consequences. Proofs of the theorems can be found in Section IV.

A. Statistical Limits of Exact Permutation Recovery

Our main theorems in this section provide necessary and sufficient conditions under which the probability of error in exactly recovering the true permutation goes to zero.

In brief, provided that d is sufficiently small, we establish a threshold phenomenon that characterizes how the signal-to-noise ratio $\mathsf{snr} := \frac{\|x^*\|_2^2}{\sigma^2}$ must scale relative to n in order to ensure identifiability. More specifically, defining the ratio

$$\Gamma(n, \mathsf{snr}) := \frac{\log(1 + \mathsf{snr})}{\log n},$$

we show that the maximum likelihood estimator recovers the true permutation with high probability provided $\Gamma(n, \mathsf{snr}) \gg c$, where c denotes an absolute constant. Conversely, if $\Gamma(n, \mathsf{snr}) \ll c$, then exact permutation recovery is impossible. For illustration, we have plotted the behaviour of the maximum likelihood estimator for the case when $d = 1$ in Figure 2. Evidently, there is a sharp phase transition between error and exact recovery as the ratio $\Gamma(n, \mathsf{snr})$ varies from 3 to 5.

Let us now turn to more precise statements of our results. We first define the maximum likelihood estimator (MLE) as

$$(\widehat{\Pi}_{\mathsf{ML}}, \widehat{x}_{\mathsf{ML}}) = \arg \min_{\substack{\Pi \in \mathcal{P}_n \\ x \in \mathbb{R}^d}} \|y - \Pi A x\|_2^2. \tag{3}$$

The following theorem provides an upper bound on the probability of error of $\widehat{\Pi}_{\mathsf{ML}}$, with (c_1, c_2) denoting absolute constants.

Theorem 1: For any $d < n$ and $\epsilon < \sqrt{n}$, if

$$\log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) \geq \left(c_1 \frac{n}{n-d} + \epsilon\right) \log n, \tag{4}$$

then $\Pr\{\widehat{\Pi}_{\mathsf{ML}} \neq \Pi^*\} \leq c_2 n^{-2\epsilon}$.

Theorem 1 provides conditions on the signal-to-noise ratio $\mathsf{snr} = \frac{\|x^*\|_2^2}{\sigma^2}$ that are sufficient for permutation recovery in the non-asymptotic, noisy regime. In contrast, the results of Unnikrishnan *et al.* [5] are stated in the limit $\mathsf{snr} \to \infty$, without an explicit characterization of the scaling behavior.

We also note that Theorem 1 holds for all values of $d < n$, whereas the results of Unnikrishnan *et al.* [5] require $n \geq 2d$ for identifiability of x^* in the noiseless case. Although the recovery of Π^* and x^* are not directly comparable, it is worth pointing out that the discrepancy also arises due to the difference between our fixed and unknown signal model, and the adversarial signal model assumed in the paper [5].

We now turn to the following converse result, which complements Theorem 1.

Theorem 2: For any $\delta \in (0, 2)$, if

$$2 + \log\left(1 + \frac{\|x^*\|_2^2}{\sigma^2}\right) \leq (2 - \delta) \log n, \tag{5}$$

then $\Pr\{\widehat{\Pi} \neq \Pi^*\} \geq 1 - c_3 e^{-c_4 n \delta}$ *for any estimator* $\widehat{\Pi}$.

Theorem 2 serves as a "strong converse" for our problem, since it guarantees that if condition (5) is satisfied, then the probability of error of any estimator goes to 1 as n goes to infinity. Indeed, it is proved using the strong converse argument for the Gaussian channel [38], which yields a converse result for any *fixed* design matrix A (see (15)). It is also worth noting that the converse result of Theorem 2 holds uniformly over d.

Taken together, Theorems 1 and 2 provide a crisp characterization of the problem when $d \leq pn$ for some fixed $p < 1$. In particular, setting ϵ and δ in Theorems 1 and 2 to be small constants and letting n grow, we recover the threshold behavior of identifiability in terms of $\Gamma(n, \mathsf{snr})$ that was discussed above and illustrated in Figure 2. In the next section, we find that a similar phenomenon occurs even with approximate permutation recovery.

When d can be arbitrarily close to n, the characterization obtained using these bounds is no longer sharp. In this regime, we conjecture that Theorem 1 provides the correct characterization of the limits of the problem, and that Theorem 2 can be sharpened.

B. Limits of Approximate Permutation Recovery

The techniques we used to prove results for exact permutation recovery can be suitably modified to obtain results for approximate permutation recovery to within a Hamming distortion D. In particular, we show the following converse result for approximate recovery.

Theorem 3: For any $2 < D \leq n-1$, *if*

$$\log\left(1+\frac{\|x^*\|_2^2}{\sigma^2}\right) \leq \frac{n-D+1}{n}\log\left(\frac{n-D+1}{2e}\right), \quad (6)$$

then $\Pr\{\mathsf{d_H}(\widehat{\Pi}, \Pi^*) \geq D\} \geq 1/2$ *for any estimator* $\widehat{\Pi}$.

Note that for any $D \leq pn$ with $p \in (0,1)$, Theorems 1 and 3 provide a set of sufficient and necessary conditions for approximate permutation recovery that match up to constant factors. In particular, the necessary condition resembles that for exact permutation recovery, and the same SNR threshold behaviour is seen even here.

Remark 1: The converse results given by Theorem 2 and Theorem 3 hold even when the estimator has exact knowledge of x^*.

C. Computational Aspects

In the previous sections, we considered the MLE given by equation (3) and analyzed its statistical properties. However, since equation (3) involves a combinatorial minimization over $n!$ permutations, it is unclear if $\widehat{\Pi}_{\mathsf{ML}}$ can be computed efficiently. The following theorem addresses this question.

Theorem 4: For $d = 1$, *the MLE* $\widehat{\Pi}_{\mathsf{ML}}$ *can be computed in time* $\mathcal{O}(n\log n)$ *for any choice of the measurement matrix* A. *In contrast, if* $d = \Omega(n)$, *then* $\widehat{\Pi}_{\mathsf{ML}}$ *is NP-hard to compute.*

The algorithm used to prove the first part of the theorem involves a simple sorting operation, which introduces the $\mathcal{O}(n\log n)$ complexity. We emphasize that the algorithm assumes no prior knowledge about the distribution of the data; for every given A and y, it returns the optimal solution to problem (3).

The second part of the theorem asserts that the algorithmic simplicity enjoyed by the $d = 1$ case does not extend to general d. The proof proceeds by a reduction from the NP-complete partition problem. We stress here that the NP-hardness claim holds over worst case input instances. In particular, it does not preclude the possibility that there exists a polynomial time algorithm that solves problem (3) with high probability when A is chosen randomly as in our original setting. In fact, since this paper was posted online, Hsu *et al.* [23] have proposed an efficient algorithm for the noiseless version of the problem with a random design matrix A, by leveraging the connection to the partition (subset-sum) problem. It is also worth noting that the same paper [23] provides a $(1+\epsilon)$-approximation algorithm for the MLE objective (3) running in time $\mathcal{O}\left(\left(\frac{n}{\epsilon}\right)^{\mathcal{O}(d)}\right)$.

IV. Proofs of Main Results

In this section, we prove our main results. Technical details are deferred to the appendices. Throughout the proofs, we assume that n is larger than some universal constant. The case where n is smaller can be handled by changing the constants in our proofs appropriately. We also use the notation c, c' to denote absolute constants that can change from line to line. Technical lemmas used in our proofs are deferred to the appendix.

We begin with the proof of Theorem 1. At a high level, it involves bounding the probability that any fixed permutation is preferred to Π^* by the estimator. The analysis requires precise control on the lower tails of χ^2-random variables, and tight bounds on the norms of random projections, for which we use results derived in the context of dimensionality reduction by Dasgupta and Gupta [39].

In order to simplify the exposition, we first consider the case when $d = 1$ in Section IV-A, and later make the necessary modifications for the general case in Section IV-B. In order to understand the technical subtleties, we recommend that the reader fully understand the $d = 1$ case along with the technical lemmas before moving on to the proof of the general case.

A. Proof of Theorem 1: $d = 1$ case

Recall the definition of the maximum likelihood estimator

$$(\widehat{\Pi}_{\mathsf{ML}}, \widehat{x}_{\mathsf{ML}}) = \arg\min_{\Pi\in\mathcal{P}_n}\min_{x\in\mathbb{R}^d}\|y-\Pi Ax\|_2^2.$$

For a fixed permutation matrix Π, assuming that A has full column rank,[2] the minimizing argument x is simply $(\Pi A)^\dagger y$, where $X^\dagger = (X^\top X)^{-1}X^\top$ represents the pseudoinverse of a matrix X. By computing the minimum over $x \in \mathbb{R}^d$ in the above equation, we find that the maximum likelihood estimate of the permutation is given by

$$\widehat{\Pi}_{\mathsf{ML}} = \arg\min_{\Pi\in\mathcal{P}_n}\|P_\Pi^\perp y\|_2^2, \quad (7)$$

where $P_\Pi^\perp = I - \Pi A(A^\top A)^{-1}(\Pi A)^\top$ denotes the projection onto the orthogonal complement of the column space of ΠA.

For a fixed $\Pi \in \mathcal{P}_n$, define the random variable

$$\Delta(\Pi, \Pi^*) := \|P_\Pi^\perp y\|_2^2 - \|P_{\Pi^*}^\perp y\|_2^2. \quad (8)$$

For any permutation Π, the estimator (7) prefers the permutation Π to Π^* if $\Delta(\Pi, \Pi^*) \leq 0$. The overall error event occurs when $\Delta(\Pi, \Pi^*) \leq 0$ for some Π, meaning that

$$\{\widehat{\Pi}_{\mathsf{ML}} \neq \Pi^*\} = \bigcup_{\Pi\in\mathcal{P}_n\setminus\Pi^*}\{\Delta(\Pi, \Pi^*) \leq 0\}. \quad (9)$$

Equation (9) holds for any value of d. We shortly specialize to the $d = 1$ case. Our strategy for proving Theorem 1 boils down to bounding the probability of each error event in the RHS of equation (9) using the following key lemma, proved in Section V-A.1. Technically speaking, the proof of this lemma contains the meat of the proof of Theorem 1, and the interested reader is encouraged to understand these details before embarking on the proof of the general case. Recall the definition of $\mathsf{d_H}(\Pi, \Pi')$, the Hamming distance between two permutation matrices.

Lemma 1: For $d = 1$ *and any two permutation matrices* Π *and* Π^*, *and provided* $\frac{\|x^*\|_2^2}{\sigma^2} > 1$, *we have*

$$\Pr\{\Delta(\Pi, \Pi^*) \leq 0\} \leq c'\exp\left(-c\,\mathsf{d_H}(\Pi, \Pi^*)\log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right)\right).$$

We are now ready to prove Theorem 1.

[2] An $n \times d$ i.i.d. Gaussian random matrix has full column rank with probability 1 as long as $d \leq n$

Proof of Theorem 1 for $d = 1$: Fix $\epsilon > 0$ and assume that the following consequence of condition (4) holds:

$$c \log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) \geq (1+\epsilon)\log n, \tag{10}$$

where c is the same as in Lemma 1. Now, observe that

$$\begin{aligned}
&\Pr\{\widehat{\Pi}_{\mathsf{ML}} \neq \Pi^*\} \\
&\leq \sum_{\Pi \in \mathcal{P}_n \setminus \Pi^*} \Pr\{\Delta(\Pi, \Pi^*) \leq 0\} \\
&\overset{\text{(i)}}{\leq} \sum_{\Pi \in \mathcal{P}_n \setminus \Pi^*} c' \exp\left(-c\, \mathsf{d}_{\mathsf{H}}(\Pi, \Pi^*) \log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right)\right) \\
&\leq c' \sum_{2 \leq k \leq n} n^k \exp\left(-c\, k \log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right)\right) \\
&\overset{\text{(ii)}}{\leq} c' \sum_{2 \leq k \leq n} n^{-\epsilon k} \\
&\leq c' \frac{1}{n^\epsilon (n^\epsilon - 1)},
\end{aligned}$$

where step (i) follows since $\#\{\Pi : \mathsf{d}_{\mathsf{H}}(\Pi, \Pi^*) = k\} \leq n^k$, and step (ii) follows from condition (10). Relabelling the constants in condition (10) proves the theorem. □

In the next section, we prove Theorem 1 for the general case.

B. Proof of Theorem 1: Case $d \in \{2, 3, \ldots, n-1\}$

In order to be consistent, we follow the same proof structure as for the $d = 1$ case. Recall the definition of $\Delta(\Pi, \Pi^*)$ from equation (8). We begin with an equivalent of the key lemma to bound the probability of the event $\{\Delta(\Pi, \Pi^*) \leq 0\}$. As in the $d = 1$ case, this constitutes the technical core of the result.

Lemma 2: For any $1 < d < n$, any two permutation matrices Π and Π^* at Hamming distance h, and provided $\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) n^{-\frac{2n}{n-d}} > \frac{5}{4}$, we have

$$\begin{aligned}
&\Pr\{\Delta(\Pi, \Pi^*) \leq 0\} \\
&\quad \leq c' \max\left[\exp\left(-n \log \frac{n}{2}\right), \right. \\
&\qquad \left. \exp\left(ch\left(\log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) - \frac{2n}{n-d}\log n\right)\right)\right].
\end{aligned} \tag{11}$$

We prove Lemma 2 in Section V-B.1. Taking it as given, we are ready to prove Theorem 1 for the general case.

Proof of Theorem 1, General Case: As before, we use the union bound to prove the theorem. We begin by fixing some $\epsilon \in (0, \sqrt{n})$ and assuming that the following consequence of condition (4) holds:

$$c \log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) \geq \left(1 + \epsilon + c\frac{2n}{n-d}\right)\log n. \tag{12}$$

Now define $b(k) := \sum_{\Pi : \mathsf{d}_{\mathsf{H}}(\Pi, \Pi^*) = k} \Pr\{\Delta(\Pi, \Pi^*) \leq 0\}$. Applying Lemma 2 then yields

$$\begin{aligned}
&\frac{(n-k)!}{n!} b(k) \\
&\quad \leq c' \max\left\{\exp\left(-n \log \frac{n}{2}\right), \right. \\
&\qquad \left. \exp\left(-ck\left(\log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) - \frac{2n}{n-d}\log n\right)\right)\right\}.
\end{aligned} \tag{13}$$

We upper bound $b(k)$ by splitting the analysis into two cases.

a) Case 1: If the first term attains the maximum in the RHS of inequality (13), then for all $2 \leq k \leq n$, we have

$$\begin{aligned}
b(k) &\leq c' n! \exp(-n \log n + n \log 2) \\
&\overset{\text{(i)}}{\leq} c' e \sqrt{n} \exp(-n \log n + n \log 2 + -n + n \log n) \\
&\overset{\text{(ii)}}{\leq} \frac{c'}{n^{2\epsilon+1}},
\end{aligned}$$

where inequality (i) follows from the well-known upper bound $n! \leq e\sqrt{n}\left(\frac{n}{e}\right)^n$, and inequality (ii) holds since $\epsilon \in (0, \sqrt{n})$.

b) Case 2: Alternatively, if the maximum is attained by the second term in the RHS of inequality (13), then we have

$$\begin{aligned}
b(k) &\leq n^k c' \exp\left(-ck\left(\log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) - \frac{2n}{n-d}\log n\right)\right) \\
&\overset{\text{(iii)}}{\leq} c' n^{-\epsilon k},
\end{aligned}$$

where step (iii) follows from condition (12).

Combining the two cases, we have

$$b(k) \leq \max\{c' n^{-\epsilon h}, c n^{-2\epsilon - 1}\} \leq \left(c' n^{-\epsilon h} + c n^{-2\epsilon-1}\right).$$

The last step is to use the union bound to obtain

$$\begin{aligned}
\Pr\{\widehat{\Pi}_{\mathsf{ML}} \neq \Pi^*\} &\leq \sum_{2 \leq k \leq n} b(k) \\
&\leq \sum_{2 \leq k \leq n} \left(c' n^{-\epsilon h} + c n^{-2\epsilon-1}\right) \\
&\overset{\text{(iv)}}{\leq} c n^{-2\epsilon},
\end{aligned} \tag{14}$$

where step (iv) follows by a calculation similar to the one carried out for the $d = 1$ case. Relabelling the constants in condition (12) completes the proof. □

C. Proof of Theorem 2

We begin by assuming that the design matrix A is fixed, and that the estimator has knowledge of x^* a-priori. Note that the latter cannot make the estimation task any easier. In proving this lower bound, we can also assume that the entries of Ax^* are distinct, since otherwise, perfect permutation recovery is impossible.

Given this setup, we now cast the problem as one of coding over a Gaussian channel. Toward this end, consider the codebook

$$\mathcal{C} = \{\Pi A x^* \mid \Pi \in \mathcal{P}_n\}.$$

We may view ΠAx^* as the codeword corresponding to the permutation Π, where each permutation is associated to one of $n!$ equally likely messages. Note that each codeword has power $\|Ax^*\|_2^2$.

The codeword is then sent over a Gaussian channel with noise power equal to $\sum_{i=1}^n \sigma^2 = n\sigma^2$. The decoding problem is to ascertain from the noisy observations which message was sent, or in other words, to identify the correct permutation.

We now use the non-asymptotic strong converse for the Gaussian channel [40]. In particular, using Lemma 11 (see Appendix B-C) with $R = \frac{\log n!}{n}$ then yields that for any $\delta' > 0$, if

$$\frac{\log n!}{n} > \frac{1+\delta'}{2}\log\left(1+\frac{\|Ax^*\|_2^2}{n\sigma^2}\right),$$

then for any estimator $\widehat{\Pi}$, we have $\Pr\{\widehat{\Pi} \neq \Pi\} \geq 1 - 2\cdot 2^{-n\delta'}$. For the choice $\delta' = \delta/(2-\delta)$, we have that if

$$(2-\delta)\log\left(\frac{n}{e}\right) > \log\left(1+\frac{\|Ax^*\|_2^2}{n\sigma^2}\right), \tag{15}$$

then $\Pr\{\widehat{\Pi} \neq \Pi\} \geq 1 - 2\cdot 2^{-n\delta/2}$. Note that the only randomness assumed so far was in the noise w and the random choice of Π.

We now specialize the result for the case when A is Gaussian. Toward that end, define the event

$$\mathcal{E}(\delta) = \left\{1+\delta \geq \frac{\|Ax^*\|_2^2}{n\|x^*\|_2^2}\right\}.$$

Conditioned on the event $\mathcal{E}(\delta)$, it can be verified that condition (5) implies condition (15). We also have

$$\begin{aligned}\Pr\{\mathcal{E}(\delta)\} &= 1 - \Pr\left\{\frac{\|Ax^*\|_2^2}{n\|x^*\|_2^2} > 1+\delta\right\}\\ &\overset{\text{(i)}}{\geq} 1 - c'e^{-cn\delta},\end{aligned}$$

where step (i) follows by using the sub-exponential tail bound (see Lemma 9 in Appendix B-B), since $\frac{\|Ax^*\|_2^2}{\|x^*\|_2^2} \sim \chi_n^2$.

Putting together the pieces, we have that provided condition (5) holds,

$$\begin{aligned}\Pr\{\widehat{\Pi} \neq \Pi^*\} &\geq \Pr\{\widehat{\Pi} \neq \Pi^* | \mathcal{E}(\delta)\}\Pr\{\mathcal{E}(\delta)\}\\ &= (1 - 2\cdot 2^{-n\delta/2})(1 - c'e^{-cn\delta})\\ &\geq 1 - c'e^{-cn\delta}.\end{aligned}$$

□

D. Proof of Theorem 3

We now prove Theorem 3 for approximate permutation recovery. For any estimator $\widehat{\Pi}$, we denote by the indicator random variable $E(\widehat{\Pi}, D)$ whether or not the $\widehat{\Pi}$ has acceptable distortion, i.e., $E(\widehat{\Pi}, D) = \mathbb{I}[\mathsf{d_H}(\widehat{\Pi}, \Pi^*) \geq D]$, with $E = 1$ representing the error event. For Π^* picked uniformly at random in $\mathcal{P}_n$, Lemma 6 stated and proved in Section V-C lower bounds the probability of error as:

$$\Pr\{E(\widehat{\Pi}, D) = 1\} \geq 1 - \frac{I(\Pi^*; y, A) + \log 2}{\log n! - \log \frac{n!}{(n-D+1)!}}.$$

Applying the chain rule for mutual information yields

$$\begin{aligned}I(\Pi^*; y, A) &= I(\Pi^*; y|A) + I(\Pi^*, A)\\ &\overset{\text{(i)}}{=} I(\Pi^*; y|A)\\ &= \mathbb{E}_A\left[I(\Pi^*; y|A=\alpha)\right],\end{aligned} \tag{16}$$

where step (i) follows since Π^* is chosen independently of A. We now evaluate the mutual information term $I(\Pi^*; y|A = \alpha)$, which we denote by $I_\alpha(\Pi^*; y)$. Letting $H_\alpha(y) := H(y|A=\alpha)$ denote the conditional entropy of y given a fixed realization of A, we have

$$\begin{aligned}I_\alpha(\Pi^*; y) &= H_\alpha(y) - H_\alpha(y|\Pi^*)\\ &\overset{\text{(ii)}}{\leq} \tfrac{1}{2}\log\det \operatorname{cov} yy^\top - \tfrac{n}{2}\log\sigma^2,\end{aligned}$$

where the covariance is evaluated with $A = \alpha$, and in step (ii), we have used two key facts:

(a) Gaussians maximize entropy for a fixed covariance, which bounds the first term, and

(b) For a fixed realization of Π^*, the vector y is composed of n uncorrelated Gaussians. This leads to an explicit evaluation of the second term.

Now taking expectations over A and noting that $\operatorname{cov} yy^\top \preceq \mathbb{E}_w\left[yy^\top\right]$, we have from the concavity of the log determinant function and Jensen's inequality that

$$\begin{aligned}I(\Pi^*; y|A) &= \mathbb{E}_A\left[I_\alpha(\Pi^*; y)\right]\\ &\leq \frac{1}{2}\log\det\mathbb{E}\left[yy^\top\right] - \frac{n}{2}\log\sigma^2,\end{aligned} \tag{17}$$

where the expectation in the last line is now taken over randomness in both A and w.

Furthermore, using the AM-GM inequality for PSD matrices $\det X \leq \left(\frac{1}{n}\operatorname{trace} X\right)^n$, and by noting that the diagonal entries of the matrix $\mathbb{E}\left[yy^\top\right]$ are all equal to $\|x^*\|_2^2 + \sigma^2$, we obtain

$$I(\Pi^*; y, A) \leq \frac{n}{2}\log\left(1+\frac{\|x^*\|_2^2}{\sigma^2}\right).$$

Combining the pieces, we now have that $\Pr\{\widehat{\Pi} \neq \Pi^*\} \geq 1/2$ if

$$n\log\left(1+\frac{\|x^*\|_2^2}{\sigma^2}\right) \leq (n-D+1)\log\left(\frac{n-D+1}{2e}\right), \tag{18}$$

which completes the proof. □

E. Proofs of Computational Aspects

In this section, we prove Theorem 4 by providing an efficient algorithm for the $d = 1$ case and showing NP-hardness for the $d > 1$ case.

1) Proof of Theorem 4: d = 1 Case: In order to prove the theorem, we need to show an algorithm that performs the optimization (7) efficiently. Accordingly, note that for the case when $d = 1$, equation (7) can be rewritten as

$$\begin{aligned}\widehat{\Pi}_{\mathsf{ML}} &= \arg\max_{\Pi} \|a_\Pi^\top y\|^2\\ &= \arg\max_{\Pi}\max\left\{a_\Pi^\top y, -a_\Pi^\top y\right\}\\ &= \arg\min_{\Pi}\max\left\{\|a_\Pi - y\|_2^2, \|a_\Pi + y\|_2^2\right\},\end{aligned} \tag{19}$$

where the last step follows since $2a_{\Pi}^{\top}y = \|a\|^2+\|y\|^2-\|a_{\Pi}-y\|_2^2$, and since the first two terms do not involve optimizing over Π.

Once the optimization problem has been written in the form (19), it is easy to see that it can be solved in polynomial time. In particular, using the fact that for fixed vectors p and q, $\|p_{\Pi}-q\|$ is minimized for Π that sorts p according to the order of q, we see that Algorithm 1 computes $\widehat{\Pi}_{\mathsf{ML}}$ exactly. This is a classical result [41] known as the rearrangement inequality (see, e.g., [2, Example 2]).

Algorithm 1 Exact Algorithm for Implementing Equation (7) for the Case When $d = 1$

Input: design matrix (vector) a, observation vector y

$\Pi_1 \leftarrow$ permutation that sorts a according to y

$\Pi_2 \leftarrow$ permutation that sorts $-a$ according to y

$\widehat{\Pi}_{\mathsf{ML}} \leftarrow \arg\max\{|a_{\Pi_1}^{\top}y|, |a_{\Pi_2}^{\top}y|\}$

Output: $\widehat{\Pi}_{\mathsf{ML}}$

The procedure defined by Algorithm 1 is clearly the correct thing to do in the noiseless case: in this case, x^* is a scalar value that scales the entries of a, and so the correct permutation can be identified by a simple sorting operation. Two such operations suffice, one to account for when x^* is positive and one more for when it is negative. Since each sort operation takes $\mathcal{O}(n\log n)$ steps, Algorithm 1 can be executed in nearly linear time. □

2) Proof of Theorem 4: NP-Hardness: In this section, we show that given a vector matrix pair $(y, A) \in \mathbb{R}^n \times \mathbb{R}^{n\times d}$, it is NP-hard to determine whether the equation $y = \Pi Ax$ has a solution for a permutation matrix $\Pi \in \mathcal{P}_n$ and vector $x \in \mathbb{R}^d$. Clearly, this is sufficient to show that the problem (3) is NP-hard to solve in the case when A and y are arbitrary.

Our proof involves a reduction from the PARTITION problem, the decision version of which is defined as the following.

Definition 1 (PARTITION): Given d integers $b_1, b_2, \ldots, b_d$, does there exist a subset $S \subset [d]$ such that

$$\sum_{i\in S} b_i = \sum_{i\in[d]\setminus S} b_i?$$

It is well known [43] that PARTITION is NP-complete. Also note that asking whether or not equation (1) has a solution (Π, x) is equivalent to determining whether or not there exists a permutation π and a vector x such that $y_{\pi} = Ax$ has a solution. We are now ready to prove the theorem.

Given an instance $b_1, \cdots b_d$ of PARTITION, define a vector $y \in \mathbb{Z}^{2d+1}$ with entries

$$y_i := \begin{cases} b_i, & \text{if } i \in [d] \\ 0, & \text{otherwise.} \end{cases}$$

Also define the $2d+1 \times 2d$ matrix

$$A := \begin{bmatrix} I_{2d} \\ 1_d^{\top} \; -1_d^{\top} \end{bmatrix}.$$

Clearly, the pair (y, A) can be constructed from $b_1, \cdots b_d$ in time polynomial in $n = 2d+1$. We now claim that $y_{\pi} = Ax$ has a solution (Π, x) if and only if there exists a subset $S \subset [d]$ such that $\sum_{i\in S} b_i = \sum_{i\in[d]\setminus S} b_i$.

By converting to row echelon form, we see that $y_{\pi} = Ax$ if and only if

$$\sum_{i|\pi(i)\le d} y_{\pi(i)} = \sum_{i|\pi(i)>d} y_{\pi(i)}, \tag{20}$$

and equation (20) holds, by construction, if and only if for $S = \{i \mid \pi(i) \le d\} \cap [d]$, we have

$$\sum_{i\in S} b_i = \sum_{i\in[d]\setminus S} b_i.$$

This completes the proof. □

V. Discussion

We analyzed the problem of exact permutation recovery in the linear regression model, and provided necessary and sufficient conditions that are tight in most regimes of n and d. We also provided a converse for the problem of approximate permutation recovery to within some Hamming distortion. It is still an open problem to characterize the fundamental limits of exact and approximate permutation recovery for all regimes of n, d and the allowable distortion D. In the context of exact permutation recovery, we believe that the limit suggested by Theorem 1 is tight for all regimes of n and d, but showing this will likely require a different technique. In particular, as pointed out in Remark 1, all of our lower bounds assume that the estimator is provided with x^* as side information; it is an interesting question as to whether stronger lower bounds can be obtained without this side information.

On the computational front, many open questions remain. The primary question concerns the design of computationally efficient estimators that succeed in similar SNR regimes. We have already shown that the maximum likelihood estimator, while being statistically optimal for moderate d, is computationally hard to evaluate in the worst case. Showing a corresponding hardness result for random A with noise is also an open problem. Finally, while this paper mainly addresses the problem of permutation recovery, the complementary problem of recovering x^* is also interesting, and we plan to investigate its fundamental limits in future work.

Appendix A
Proofs of Technical Lemmas

In this section, we provide statements and proofs of the technical lemmas used in the proofs of our main theorems.

A. Supporting Proofs for Theorem 1: $d = 1$ Case

We provide a proof of Lemma 1 in this section; see Section IV-A for the proof of Theorem 1 (case $d = 1$) given Lemma 1. We begin by restating the lemma for convenience.

1) Proof of Lemma 1: Before the proof, we establish notation. For each $\delta > 0$, define the events

$$\mathcal{F}_1(\delta) = \left\{ |\|P_{\Pi^*}^{\perp}y\|_2^2 - \|P_{\Pi}^{\perp}w\|_2^2| \ge \delta \right\}, \text{ and} \tag{21a}$$

$$\mathcal{F}_2(\delta) = \left\{ \|P_{\Pi}^{\perp}y\|_2^2 - \|P_{\Pi}^{\perp}w\|_2^2 \le 2\delta \right\}. \tag{21b}$$

Evidently,

$$\{\Delta(\Pi, \Pi^*) \leq 0\} \subseteq \mathcal{F}_1(\delta) \cup \mathcal{F}_2(\delta). \tag{22}$$

Indeed, if neither $\mathcal{F}_1(\delta)$ nor $\mathcal{F}_2(\delta)$ occurs

$$\begin{aligned}\Delta(\Pi, \Pi^*) &= \left(\|P_\Pi^\perp y\|_2^2 - \|P_\Pi^\perp w\|_2^2\right) - \left(\|P_{\Pi^*}^\perp y\|_2^2 - \|P_\Pi^\perp w\|_2^2\right)\\ &> 2\delta - \delta\\ &= \delta.\end{aligned}$$

Thus, to prove Lemma 1, we shall bound the probability of the two events $\mathcal{F}_1(\delta)$ and $\mathcal{F}_2(\delta)$ individually, and then invoke the union bound. Note that inequality (22) holds for all values of $\delta > 0$; it is convenient to choose $\delta^* := \frac{1}{3}\|P_\Pi^\perp \Pi^* A x^*\|_2^2$. With this choice, the following lemma bounds the probabilities of the individual events over randomness in w conditioned on a given A. Its proof is postponed to the end of the section.

Lemma 3: For any $\delta > 0$ and with $\delta^* = \frac{1}{3}\|P_\Pi^\perp \Pi^* A x^*\|_2^2$, we have

$$\Pr_w\{\mathcal{F}_1(\delta)\} \leq c' \exp\left(-c\frac{\delta}{\sigma^2}\right), \text{ and} \tag{23a}$$

$$\Pr_w\{\mathcal{F}_2(\delta^*)\} \leq c' \exp\left(-c\frac{\delta^*}{\sigma^2}\right). \tag{23b}$$

The next lemma, proved in Section V-A.3, is needed in order to incorporate the randomness in A into the required tail bound. It is convenient to introduce the shorthand $T_\Pi := \|P_\Pi^\perp \Pi^* A x^*\|_2^2$.

Lemma 4: For $d = 1$ and any two permutation matrices Π and Π^* at Hamming distance h, we have

$$\Pr_A\{T_\Pi \leq t\|x^*\|_2^2\} \leq 6\exp\left(-\frac{h}{10}\left[\log\frac{h}{t} + \frac{t}{h} - 1\right]\right) \tag{24}$$

for all $t \in [0, h]$.

We now have all the ingredients to prove Lemma 1.

Proof of Lemma 1: Applying Lemma 3 and using the union bound yields

$$\begin{aligned}\Pr_w\{\Delta(\Pi, \Pi^*) \leq 0\} &\leq \Pr_w\{\mathcal{F}_1(\delta^*)\} + \Pr_w\{\mathcal{F}_2(\delta^*)\}\\ &\leq c' \exp\left(-c\frac{T_\Pi}{\sigma^2}\right).\end{aligned} \tag{25}$$

Combining bound (25) with Lemma 4 yields

$$\begin{aligned}&\Pr\{\Delta(\Pi, \Pi^*) \leq 0\}\\ &\quad\leq c' \exp\left(-c\frac{t\|x^*\|_2^2}{\sigma^2}\right)\Pr_A\{T_\Pi \geq t\|x^*\|_2^2\}\\ &\qquad + \Pr_A\{T_\Pi \leq t\|x^*\|_2^2\}\\ &\quad\leq c' \exp\left(-c\frac{t\|x^*\|_2^2}{\sigma^2}\right)\\ &\qquad + 6\exp\left(-\frac{h}{10}\left[\log\frac{h}{t} + \frac{t}{h} - 1\right]\right),\end{aligned} \tag{26}$$

where the last inequality holds provided that $t \in [0, h]$, and the probability in the LHS is now taken over randomness in both w *and* A.

Using the shorthand $\mathsf{snr} := \frac{\|x^*\|_2^2}{\sigma^2}$, setting $t = h\frac{\log \mathsf{snr}}{\mathsf{snr}}$, and noting that $t \in [0, h]$ since $\mathsf{snr} > 1$, we have

$$\begin{aligned}&\Pr\{\Delta(\Pi, \Pi^*) \leq 0\}\\ &\quad\leq c' \exp\left(-ch\log \mathsf{snr}\right)\\ &\qquad + 6\exp\left(-\frac{h}{10}\left[\log\left(\frac{\mathsf{snr}}{\log \mathsf{snr}}\right) + \frac{\log \mathsf{snr}}{\mathsf{snr}} - 1\right]\right).\end{aligned}$$

It is easily verified that for all $\mathsf{snr} > 1$, we have

$$\log\left(\frac{\mathsf{snr}}{\log \mathsf{snr}}\right) + \frac{\log \mathsf{snr}}{\mathsf{snr}} - 1 > \frac{\log \mathsf{snr}}{4}. \tag{27}$$

Hence, after substituting for snr, we have

$$\Pr\{\Delta(\Pi, \Pi^*) \leq 0\} \leq c' \exp\left(-ch\log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right)\right). \tag{28}$$

□

2) Proof of Lemma 3: We prove each claim of the lemma separately.

a) Proof of claim (23a)*:* To start, note that by definition of the linear model, we have $\|P_{\Pi^*}^\perp y\|_2^2 = \|P_{\Pi^*}^\perp w\|_2^2$. Letting Z_ℓ denote a χ^2 random variable with ℓ degrees of freedom, we claim that

$$\|P_{\Pi^*}^\perp w\|_2^2 - \|P_\Pi^\perp w\|_2^2 = Z_k - \tilde{Z}_k,$$

where $k := \min(d, \mathsf{d_H}(\Pi, \Pi^*))$.

For the rest of the proof, we adopt the shorthand $\Pi \setminus \Pi' := \mathsf{range}(\Pi A) \setminus \mathsf{range}(\Pi' A)$, and $\Pi \cap \Pi' := \mathsf{range}(\Pi A) \cap \mathsf{range}(\Pi' A)$. Now, by the Pythagorean theorem, we have

$$\|P_{\Pi^*}^\perp w\|_2^2 - \|P_\Pi^\perp w\|_2^2 = \|P_\Pi w\|_2^2 - \|P_{\Pi^*} w\|_2^2.$$

Splitting it up further, we can then write

$$\|P_\Pi w\|_2^2 = \|P_{\Pi\cap\Pi^*} w\|_2^2 + \|(P_\Pi - P_{\Pi\cap\Pi^*})w\|_2^2,$$

where we have used the fact that $P_{\Pi\cap\Pi^*}P_\Pi = P_{\Pi\cap\Pi^*} = P_{\Pi\cap\Pi^*}P_{\Pi^*}$.

Similarly for the second term, we have $\|P_{\Pi^*} w\|_2^2 = \|P_{\Pi\cap\Pi^*} w\|_2^2 + \|(P_{\Pi^*} - P_{\Pi\cap\Pi^*})w\|_2^2$, and hence,

$$\begin{aligned}\|P_\Pi w\|_2^2 - \|P_{\Pi^*} w\|_2^2 = \|(P_\Pi - P_{\Pi\cap\Pi^*})w\|_2^2 - \|(P_{\Pi^*}\\ - P_{\Pi\cap\Pi^*})w\|_2^2.\end{aligned}$$

Now each of the two projection matrices above has rank[3] $\mathsf{dim}(\Pi \setminus \Pi^*) = k$, which completes the proof of the claim. To prove the lemma, note that for any $\delta > 0$, we can write

$$\Pr\{\mathcal{F}_1(\delta)\} \leq \Pr\{|Z_k - k| \geq \delta/2\} + \Pr\{|\tilde{Z}_k - k| \geq \delta/2\}.$$

Using the sub-exponential tail-bound on χ^2 random variables (see Lemma 9 in Appendix B-B) completes the proof. □

[3]With probability 1

b) Proof of claim (23b)*:* We begin by writing

$$\Pr_w\{\mathcal{F}_2(\delta)\} = \Pr_w\left\{\underbrace{\|P_\Pi^\perp \Pi^* A x^*\|_2^2 + 2\langle P_\Pi^\perp \Pi^* A x^*, P_\Pi^\perp w\rangle}_{R(A,w)} \le 2\delta\right\}.$$

We see that conditioned on A, the random variable $R(A, w)$ is distributed as $\mathcal{N}(T_\Pi, 4\sigma^2 T_\Pi)$, where we have used the shorthand $T_\Pi := \|P_\Pi^\perp \Pi^* A x^*\|_2^2$.

So applying standard Gaussian tail bounds (see, for example, Boucheron *et al.* [44]), we have

$$\Pr_w\{\mathcal{F}_2(\delta)\} \le \exp\left(-\frac{(T_\Pi - 2\delta)^2}{8\sigma^2 T_\Pi}\right).$$

Setting $\delta = \delta^* := \frac{1}{3}T_\Pi$ completes the proof. □

3) Proof of Lemma 4: In the case $d = 1$, the matrix A is composed of a single vector $a \in \mathbb{R}^n$. Recalling the random variable $T_\Pi = \|P_\Pi^\perp \Pi^* A x^*\|_2^2$, we have

$$\begin{aligned} T_\Pi &= (x^*)^2\left(\|a\|_2^2 - \frac{1}{\|a\|_2^2}\langle a_\Pi, a\rangle^2\right) \\ &\overset{\text{(i)}}{\ge} (x^*)^2\left(\|a\|_2^2 - |\langle a, a_\Pi\rangle|\right) \\ &= \frac{(x^*)^2}{2}\min\left(\|a - a_\Pi\|_2^2, \|a + a_\Pi\|_2^2\right), \end{aligned}$$

where step (i) follows from the Cauchy Schwarz inequality. Applying the union bound then yields

$$\Pr\{T_\Pi \le t(x^*)^2\} \le \Pr\{\|a - a_\Pi\|_2^2 \le 2t\} + \Pr\{\|a + a_\Pi\|_2^2 \le 2t\}.$$

Let Z_ℓ and $\tilde{Z}_\ell$ denote (not necessarily independent) χ^2 random variables with ℓ degrees of freedom. We split the analysis into two cases.

a) Case $h \ge 3$: Lemma 7 from Appendix B-A guarantees that

$$\frac{\|a - a_\Pi\|_2^2}{2} \overset{d}{=} Z_{h_1} + Z_{h_2} + Z_{h_3}, \text{ and} \tag{29a}$$

$$\frac{\|a + a_\Pi\|_2^2}{2} \overset{d}{=} \widetilde{Z}_{h_1} + \widetilde{Z}_{h_2} + \widetilde{Z}_{h_3} + \widetilde{Z}_{n-h}, \tag{29b}$$

where $\overset{d}{=}$ denotes equality in distribution and $h_1, h_2, h_3 \ge \frac{h}{5}$ with $h_1 + h_2 + h_3 = h$. An application of the union bound then yields

$$\Pr\{\|a - a_\Pi\|_2^2 \le 2t\} \le \sum_{i=1}^3 \Pr\left\{Z_{h_i} \le t\frac{h_i}{h}\right\}.$$

Similarly, provided that $h \ge 3$, we have

$$\begin{aligned} \Pr\{\|a + a_\Pi\|_2^2 \le 2t\} &\le \Pr\{\widetilde{Z}_{h_1} + \widetilde{Z}_{h_2} + \widetilde{Z}_{h_3} + \widetilde{Z}_{n-h} \le t\} \\ &\overset{\text{(ii)}}{\le} \Pr\{\widetilde{Z}_{h_1} + \widetilde{Z}_{h_2} + \widetilde{Z}_{h_3} \le t\} \\ &\overset{\text{(iii)}}{\le} \sum_{i=1}^3 \Pr\left\{\widetilde{Z}_{h_i} \le t\frac{h_i}{h}\right\}, \end{aligned}$$

where inequality (ii) follows from the non-negativity of Z_{n-h}, and the monotonicity of the CDF; and inequality (iii) from the union bound. Finally, bounds on the lower tails of χ^2 random variables (see Lemma 8 in Appendix B-B) yield

$$\begin{aligned} \Pr\left\{Z_{h_i} \le t\frac{h_i}{h}\right\} &= \Pr\left\{\widetilde{Z}_{h_i} \le t\frac{h_i}{h}\right\} \\ &\overset{\text{(iv)}}{\le} \left(\frac{t}{h}\exp\left(1 - \frac{t}{h}\right)\right)^{h_i/2} \\ &\overset{\text{(v)}}{\le} \left(\frac{t}{h}\exp\left(1 - \frac{t}{h}\right)\right)^{h/10}. \end{aligned}$$

Here, inequality (iv) is valid provided $\frac{th_i}{h} \le h_i$, or equivalently, if $t \le h$, whereas inequality (v) follows since $h_i \ge h/5$ and the function $xe^{1-x} \in [0, 1]$ for all $x \in [0, 1]$. Combining the pieces proves Lemma 4 for $h \ge 3$.

b) Case $h = 2$: In this case, we have

$$\frac{\|a - a_\Pi\|_2^2}{2} \overset{d}{=} 2Z_1, \quad \text{and} \quad \frac{\|a + a_\Pi\|_2^2}{2} \overset{d}{=} 2\widetilde{Z}_1 + \widetilde{Z}_{n-2}.$$

Proceeding as before by applying the union bound and Lemma 8, we have that for $t \le 2$, the random variable T_Π obeys the tail bound

$$\begin{aligned} \Pr\{T_\Pi \le t(x^*)^2\} &\le 2\left(\frac{t}{2}\exp\left(1 - \frac{t}{2}\right)\right)^{1/2} \\ &\le 6\left(\frac{t}{h}\exp\left(1 - \frac{t}{h}\right)\right)^{h/10}, \text{ for } h = 2. \end{aligned}$$

□

B. Supporting Proofs for Theorem 1: $d > 1$ Case

We now provide a proof of Lemma 2. See Section IV-B for the proof of Theorem 1 (case $d > 1$) given Lemma 2. We begin by restating the lemma for convenience.

1) Proof of Lemma 2: The first part of the proof is exactly the same as that of Lemma 1. In particular, Lemma 3 applies without modification to yield a bound identical to the inequality (25), given by

$$\Pr_w\{\Delta(\Pi, \Pi^*) \le 0\} \le c' \exp\left(-c\frac{T_\Pi}{\sigma^2}\right), \tag{30}$$

where $T_\Pi = \|P_\Pi^\perp \Pi^* A x^*\|_2^2$, as before.

The major difference from the $d = 1$ case is in the random variable T_Π. Accordingly, we state the following parallel lemma to Lemma 4.

Lemma 5: For $1 < d < n$, any two permutation matrices Π and Π^* at Hamming distance h, and $t \le hn^{-\frac{2n}{n-d}}$, we have

$$\Pr_A\{T_\Pi \le t\|x^*\|_2^2\} \le 2\max\{T_1, T_2\}, \tag{31}$$

where

$$T_1 = \exp\left(-n\log\frac{n}{2}\right), \text{ and}$$

$$T_2 = 6\exp\left(-\frac{h}{10}\left[\log\left(\frac{h}{tn^{\frac{2n}{n-d}}}\right) + \frac{tn^{\frac{2n}{n-d}}}{h} - 1\right]\right).$$

The proof of Lemma 5 appears in Section V-B.2. We are now ready to prove Lemma 2.

Proof of Lemma 2: We prove Lemma 2 from Lemma 5 and equation (30) by an argument similar to the one before.

In particular, in a similar vein to the steps leading up to equation (26), we have

$$\Pr\{\Delta(\Pi,\Pi^*)\le 0\} \le c'\exp\left(-c\frac{t\|x^*\|_2^2}{\sigma^2}\right)+\Pr_A\{T_\Pi\le t\|x^*\|_2^2\}. \quad (32)$$

We now use the shorthand $\mathsf{snr} := \frac{\|x^*\|_2^2}{\sigma^2}$ and let $t^* = h\frac{\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)}{\mathsf{snr}}$. Noting that $\mathsf{snr}\cdot n^{-\frac{2n}{n-d}} > 5/4$ yields $t^*\le hn^{-\frac{2n}{n-d}}$, we set $t=t^*$ in inequality (32) to obtain

$$\Pr\{\Delta(\Pi,\Pi^*)\le 0\} \le c'\exp\left(-ch\log sn^{-\frac{2n}{n-d}}\right)+\Pr_A\{T_\Pi\le t^*\|x^*\|_2^2\}. \quad (33)$$

Since $\Pr_A\{T_\Pi\le t^*\|x^*\|_2^2\}$ can be bounded by a maximum of two terms (31), we now split the analysis into two cases depending on which term attains the maximum.

a) Case 1: First, suppose that the second term attains the maximum in inequality (31), i.e., $\Pr_A\{T_\Pi \le t^*\|x^*\|_2^2\} \le 12\exp\left(-\frac{h}{10}\left[\log\left(\frac{h}{t^*n^{\frac{2n}{n-d}}}\right)+\frac{t^*n^{\frac{2n}{n-d}}}{h}-1\right]\right)$. Substituting for t^*, we have

$$\Pr_A\{T_\Pi\le t^*\|x^*\|_2^2\} \le 12\exp\left(-\frac{h}{10}\log\left(\frac{\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}}{\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)}\right)\right) \cdot\exp\left(-\frac{h}{10}\left[\frac{\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)}{\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}}-1\right]\right).$$

We have $\mathsf{snr}\cdot n^{-\frac{2n}{n-d}} > \frac{5}{4}$, a condition which leads to the following pair of easily verifiable inequalities:

$$\log\left(\frac{\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}}{\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)}\right)+\frac{\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)}{\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}}-1 \ge \frac{\log \mathsf{snr}\cdot n^{-\frac{2n}{n-d}}}{4}, \text{ and} \quad (34a)$$

$$\log\left(\frac{\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}}{\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)}\right)+\frac{\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)}{\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}}-1 \le 5\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right). \quad (34b)$$

Using inequality (34a), we have

$$\Pr_A\{T_\Pi\le t^*\|x^*\|_2^2\}\le 12\exp\left(-ch\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)\right). \quad (35)$$

Inequality (34b) will be useful in the second case to follow. Now using inequalities (35) and (33) together yields

$$\Pr\{\Delta(\Pi,\Pi^*)\le 0\}\le c'\exp\left(-ch\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)\right). \quad (36)$$

It remains to handle the second case.

b) Case 2: Suppose now that $\Pr_A\{T_\Pi\le t^*\|x^*\|_2^2\}\le 2\exp\left(-n\log\frac{n}{2}\right)$, i.e., that the first term in RHS of inequality (31) attains the maximum when $t=t^*$. In this case, we have

$$\begin{aligned}\exp\left(-n\log\frac{n}{2}\right) &\ge 6\exp\left(-\frac{h}{10}\left[\log\left(\frac{h}{t^*n^{\frac{2n}{n-d}}}\right)+\frac{t^*n^{\frac{2n}{n-d}}}{h}-1\right]\right)\\ &\overset{\text{(i)}}{\ge} c'\exp\left(-ch\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)\right),\end{aligned}$$

where step (i) follows from the right inequality (34b). Now substituting into inequality (33), we have

$$\begin{aligned}\Pr\{\Delta(\Pi,\Pi^*)\le 0\} &\le c'\exp\left(-ch\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)\right)+2\exp\left(-n\log\frac{n}{2}\right)\\ &\le c'\exp\left(-n\log\frac{n}{2}\right). \end{aligned}\quad (37)$$

Combining equations (36) and (37) completes the proof of Lemma 2. □

2) Proof of Lemma 5: We begin by reducing the problem to the case $x^*=e_1\|x^*\|_2$, where e_1 represents the first standard basis vector in $\mathbb{R}^d$. In particular, if $Wx^*=e_1\|x^*\|_2$ for a $d\times d$ unitary matrix W and writing $A=\widetilde{A}W$, we have by rotation invariance of the Gaussian distribution that the the entries of $\widetilde{A}$ are distributed as i.i.d. standard Gaussians. It can be verified that $T_\Pi = \|I-\Pi\widetilde{A}(\widetilde{A}^\top\widetilde{A})^{-1}(\Pi\widetilde{A})^\top\Pi^*\widetilde{A}e_1\|_2^2\|x^*\|_2^2$. Since $\widetilde{A}\overset{d}{=}A$, the reduction is complete.

In order to keep the notation uncluttered, we denote the first column of A by a. We also denote the span of the first column of ΠA by S_1 and that of the last $d-1$ columns of ΠA by S_{-1}. Denote their respective orthogonal complements by $S_1^\perp$ and $S_{-1}^\perp$. We then have

$$\begin{aligned}T_\Pi &= \|x^*\|_2^2\|P_\Pi^\perp a\|_2^2\\ &= \|x^*\|_2^2\|P_{S_{-1}^\perp\cap S_1^\perp}a\|_2^2\\ &= \|x^*\|_2^2\|P_{S_{-1}^\perp\cap S_1^\perp}P_{S_1^\perp}a\|_2^2.\end{aligned}$$

We now condition on a. Consequently, the subspace $S_1^\perp$ is a *fixed* $(n-1)$-dimensional subspace. Additionally, $S_{-1}^\perp\cap S_1^\perp$ is the intersection of a uniformly random $(n-(d-1))$-dimensional subspace with a fixed $(n-1)$-dimensional subspace, and is therefore a uniformly random $(n-d)$-dimensional subspace within $S_1^\perp$. Writing $u=\frac{P_{S_1^\perp}a}{\|P_{S_1^\perp}a\|_2}$, we have

$$T_\Pi \overset{d}{=} \|x^*\|_2^2\|P_{S_{-1}^\perp\cap S_1^\perp}u\|_2^2\|P_{S_1^\perp}a\|_2^2.$$

Now since $u\in S_1^\perp$, note that $\|P_{S_{-1}^\perp\cap S_1^\perp}u\|_2^2$ is the squared length of a projection of an $(n-1)$-dimensional unit vector onto a uniformly chosen $(n-d)$-dimensional subspace. In other words, denoting a uniformly random projection from m dimensions to k dimensions by P_k^m and noting that u is a unit vector, we have

$$\|P_{S_{-1}^\perp\cap S_1^\perp}u\|_2^2\overset{d}{=}\|P_{n-d}^{n-1}v_1\|_2^2\overset{\text{(i)}}{=}1-\|P_{d-1}^{n-1}v_1\|_2^2,$$

where v_1 represents a fixed standard basis vector in $n-1$ dimensions. The quantities P_{n-d}^{n-1} and P_{d-1}^{n-1} are projections onto orthogonal subspaces, and step (i) is a consequence of the Pythagorean theorem.

Now removing the conditioning on a, we see that the term for $d > 1$ can be lower bounded by the corresponding T_Π for $d = 1$, but scaled by a random factor – the norm of a random projection. Using $T_\Pi^1 := \|P_{S_1}^\perp a\|_2^2 \|x^*\|_2^2$ to denote T_Π when $d = 1$, we have

$$T_\Pi = (1 - X_{d-1}) T_\Pi^1, \tag{38}$$

where we have introduced the shorthand $X_{d-1} = \|P_{d-1}^{n-1} v_1\|_2^2$.

We first handle the random projection term in equation (38) using Lemma 10 in Appendix B-B. In particular, substituting $\beta = (1-z)\frac{n-1}{d-1}$ in inequality (47) yields

$$\begin{aligned}
\Pr\{1 - X_{d-1} \le z\} &\le \left(\frac{n-1}{d-1}\right)^{(d-1)/2} \left(\frac{z(n-1)}{n-d}\right)^{(n-d)/2} \\
&\overset{\text{(i)}}{\le} \sqrt{\binom{n-1}{d-1}} \sqrt{\binom{n-1}{n-d}} z^{\frac{n-d}{2}} \\
&= \binom{n-1}{d-1} z^{\frac{n-d}{2}} \\
&\overset{\text{(ii)}}{\le} 2^{n-1} z^{\frac{n-d}{2}},
\end{aligned}$$

where in steps (i) and (ii), we have used the standard inequality $2^n \ge \binom{n}{r} \ge \left(\frac{n}{r}\right)^r$. Now setting $z = n^{\frac{-2n}{n-d}}$, which ensures that $(1-z)\frac{n-1}{d-1} > 1$ for all $d < n$ and large enough n, we have

$$\Pr\{1 - X_{d-1} \le n^{\frac{-2n}{n-d}}\} \le \exp\left(-n \log \tfrac{n}{2}\right). \tag{39}$$

Applying the union bound then yields

$$\begin{aligned}
&\Pr\{T_\Pi \le t\|x^*\|_2^2\} \\
&\quad \le \Pr\{1 - X_{d-1} \le n^{\frac{-2n}{n-d}}\} + \Pr\{T_\Pi^1 \le t n^{\frac{2n}{n-d}} \|x^*\|_2^2\}. \quad (40)
\end{aligned}$$

We have already computed an upper bound on $\Pr\{T_\Pi^1 \le t n^{\frac{2n}{n-d}} \|x^*\|_2^2\}$ in Lemma 4. Applying it yields that provided $t \le h n^{-\frac{2n}{n-d}}$, we have

$$\begin{aligned}
&\Pr\{T_\Pi^1 \le t n^{\frac{2n}{n-d}} \|x^*\|_2^2\} \\
&\qquad \le 6\left(\frac{t n^{\frac{2n}{n-d}}}{h} \exp\left(1 - \frac{t n^{\frac{2n}{n-d}}}{h}\right)\right)^{h/10}. \quad (41)
\end{aligned}$$

Combining equations (41) and (39) with the union bound (40) and performing some algebraic manipulation then completes the proof of Lemma 5. □

C. Supporting Proofs for Theorem 3

The following technical lemma was used in the proof of Theorem 3, and we recall the setting for convenience. For any estimator $\widehat{\Pi}$, we denote by the indicator random variable $E(\widehat{\Pi}, D)$ whether or not the $\widehat{\Pi}$ has acceptable distortion, i.e., $E(\widehat{\Pi}, D) = \mathbb{I}[\mathsf{d_H}(\widehat{\Pi}, \Pi^*) \ge D]$, with $E = 1$ representing the error event. Assume Π^* is picked uniformly at random in $\mathcal{P}_n$.

Lemma 6: The probability of error is lower bounded as

$$\Pr\{E(\widehat{\Pi}, D) = 1\} \ge 1 - \frac{I(\Pi^*; y, A) + \log 2}{\log n! - \log \frac{n!}{(n-D+1)!}}. \tag{42}$$

Proof of Lemma 6: We use the shorthand $E := E(\widehat{\Pi}, D)$ in this proof to simplify notation. Proceeding by the usual proof of Fano's inequality, we begin by expanding $H(E, \Pi^*|y, A = a, \widehat{\Pi})$ in two ways:

$$\begin{aligned}
&H(E, \Pi^*|y, A, \widehat{\Pi}) \\
&\quad = H(\Pi^*|y, A, \widehat{\Pi}) + H(E|\Pi^*, y, A, \widehat{\Pi}) \quad (43a) \\
&\quad = H(E|y, A, \widehat{\Pi}) + H(\Pi^*|E, y, A, \widehat{\Pi}). \quad (43b)
\end{aligned}$$

Since $\Pi^* \to (y, A) \to \widehat{\Pi}$ forms a Markov chain, we have $H(\Pi^*|y, A, \widehat{\Pi}) = H(\Pi^*|y, A)$. Non-negativity of entropy yields $H(E|\Pi^*, y, A, \widehat{\Pi}) \ge 0$. Since conditioning cannot increase entropy, we have $H(E|y, A, \widehat{\Pi}) \le H(E) \le \log 2$, and $H(\Pi^*|E, y, A, \widehat{\Pi}) \le H(\Pi^*|E, \widehat{\Pi})$. Combining all of this with the pair of equations (43) yields

$$\begin{aligned}
&H(\Pi^*|y, A) \\
&\quad \le H(\Pi^*|E, \widehat{\Pi}) + \log 2 \\
&\quad = \Pr\{E = 1\} H(\Pi^*|E = 1, \widehat{\Pi}) \\
&\qquad + (1 - \Pr\{E = 1\})\, H(\Pi^*|E = 0, \widehat{\Pi}) + \log 2. \quad (44)
\end{aligned}$$

We now use the fact that uniform distributions maximize entropy to bound the two terms as $H(\Pi^*|E = 1, \widehat{\Pi}) \le H(\Pi^*) = \log n!$, and $H(\Pi^*|E = 0, \widehat{\Pi}) \le \log \frac{n!}{(n-D+1)!}$, where the last inequality follows since $E = 0$ reveals that Π^* is within a Hamming ball of radius $D-1$ around $\widehat{\Pi}$, and the cardinality of that Hamming ball is $\frac{n!}{(n-D+1)!}$.

Substituting back into inequality (44) yields

$$\begin{aligned}
&\Pr\{E = 1\}\left(\log n! - \log \frac{n!}{(n-D+1)!}\right) + H(\Pi^*) \\
&\qquad \ge H(\Pi^*|y, A) - \log 2 - \log \frac{n!}{(n-D+1)!} + \log n!,
\end{aligned}$$

where we have added the term $H(\Pi^*) = \log n!$ to both sides. Simplifying then yields inequality (42). □

Appendix B
Auxiliary Results

In this section, we prove a preliminary lemma about permutations that is useful in many of our proofs. We also derive tight bounds on the lower tails of χ^2-random variables and state an existing result on tail bounds for random projections.

D. Independent Sets of Permutations

In this section, we prove a combinatorial lemma about permutations. Given a Gaussian random vector $Z \in \mathbb{R}^n$, we use this lemma to characterize the distribution of $Z \pm \Pi Z$ as a function of the permutation Π. In order to state the lemma, we need to set up some additional notation. For a permutation π on k objects, let G_π denote the corresponding undirected incidence graph, i.e., $V(G_\pi) = [k]$, and $(i, j) \in E(G_\pi)$ iff $j = \pi(i)$ or $i = \pi(j)$.

Lemma 7: Let π be a permutation on $k \ge 3$ objects such that $\mathsf{d_H}(\pi, I) = k$. Then the vertices of G_π can be partitioned into three sets V_1, V_2, V_3 such that each is an independent set, and $|V_1|, |V_2|, |V_3| \ge \lfloor \frac{k}{3} \rfloor \ge \frac{k}{5}$.

Proof: Note that for any permutation π, the corresponding graph G_π is composed of cycles, and the vertices in each

cycle together form an independent set. Consider one such cycle. We can go through the vertices in the order induced by the cycle, and alternate placing them in each of the 3 partitions. Clearly, this produces independent sets, and furthermore, having 3 partitions ensures that the last vertex in the cycle has some partition with which it does not share edges. If the cycle length $C \equiv 0 \pmod 3$, then each partition gets $C/3$ vertices, otherwise the smallest partition has $\lfloor C/3 \rfloor$ vertices. The partitions generated from the different cycles can then be combined (with relabelling, if required) to ensure that the largest partition has cardinality at most 1 more than that of the smallest partition. ∎

E. Tail Bounds on χ^2 Random Variables and Random Projections

In our analysis, we require tight control on lower tails of χ^2 random variables. The following lemma provides one such bound.

Lemma 8: Let Z_ℓ denote a χ^2 random variable with ℓ degrees of freedom. Then for all $p \in [0, \ell]$, we have

$$\Pr\{Z_\ell \leq p\} \leq \left(\frac{p}{\ell}\exp\left(1-\frac{p}{\ell}\right)\right)^{\ell/2} = \exp\left(-\frac{\ell}{2}\left[\log\frac{\ell}{p}+\frac{p}{\ell}-1\right]\right). \quad (45)$$

Proof: The lemma is a simple consequence of the Chernoff bound. In particular, we have for all $\lambda > 0$ that

$$\begin{aligned}\Pr\{Z_\ell \leq p\} &= \Pr\{\exp(-\lambda Z_\ell) \geq \exp(-\lambda p)\} \\ &\leq \exp(\lambda p)\mathbb{E}\left[\exp(-\lambda Z_\ell)\right] \\ &= \exp(\lambda p)(1+2\lambda)^{-\frac{\ell}{2}}. \end{aligned} \quad (46)$$

where in the last step, we have used $\mathbb{E}\left[\exp(-\lambda Z_\ell)\right] = (1+2\lambda)^{-\frac{\ell}{2}}$, which is valid for all $\lambda > -1/2$. Minimizing the last expression over $\lambda > 0$ then yields the choice $\lambda^* = \frac{1}{2}\left(\frac{\ell}{p}-1\right)$, which is greater than 0 for all $0 \leq p \leq \ell$. Substituting this choice back into equation (46) proves the lemma. ∎

We also state the following lemma for general sub-exponential random variables (see, e.g., Boucheron *et al.* [44]). We use it in the context of χ^2 random variables.

Lemma 9: Let X be a sub-exponential random variable. Then for all $t > 0$, we have

$$\Pr\{|X - \mathbb{E}[X]| \geq t\} \leq c' e^{-ct}.$$

Lastly, we require tail bounds on the norms of random projections, a problem that has been studied extensively in the literature on dimensionality reduction. The following lemma, a consequence of the Chernoff bound, is taken from Dasgupta and Gupta [39, Lemma 2.2b].

Lemma 10 ([39]): Let x be a fixed n-dimensional vector, and let P_d^n be a projection matrix from n-dimensional space to a uniformly randomly chosen d-dimensional subspace, where $d \leq n$. Then we have for every $\beta > 1$ that

$$\Pr\{\|P_d^n x\|_2^2 \geq \frac{\beta d}{n}\|x\|_2^2\} \leq \beta^{d/2}\left(1+\frac{(1-\beta)d}{n-d}\right)^{(n-d)/2}. \quad (47)$$

F. Strong Converse for Gaussian Channel Capacity

The following result due to Shannon [38] provides a strong converse for the Gaussian channel. The non-asymptotic version as stated here was also derived by Yoshihara [40].

Lemma 11 ([40]): Consider a vector Gaussian channel on n coordinates with message power P and noise power σ^2, whose capacity is given by $\overline{R} = \log\left(1+\frac{P}{\sigma^2}\right)$. For any codebook $\mathcal{C}$ with $|\mathcal{C}| = 2^{nR}$, if for some $\epsilon > 0$ we have

$$R > (1+\epsilon)\overline{R},$$

then the probability of error $p_e \geq 1 - 2\cdot 2^{-n\epsilon}$ for n large enough.

REFERENCES

[1] A. Pananjady, M. J. Wainwright, and T. A. Courtade, "Linear regression with an unknown permutation: Statistical and computational limits," in *Proc. IEEE 54th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2016, pp. 417–424.

[2] A. Pananjady, M. J. Wainwright, and T. A. Courtade. (Aug. 2016). "Linear regression with an unknown permutation: Statistical and computational limits." [Online]. Available: https://arxiv.org/abs/1608.02902

[3] R. J. A. Little and D. B. Rubin, *Statistical Analysis With Missing Data*. Hoboken, NJ, USA: Wiley, 2014.

[4] P.-L. Loh and M. J. Wainwright, "Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2012, pp. 2601–2605.

[5] J. Unnikrishnan, S. Haghighatshoar, and M. Vetterli. (2015). "Unlabeled sensing with random linear measurements." [Online]. Available: https://arxiv.org/abs/1512.00115

[6] A. Balakrishnan, "On the problem of time jitter in sampling," *IRE Trans. Inf. Theory*, vol. 8, no. 3, pp. 226–236, Apr. 1962.

[7] C. Rose, I. S. Mian, and R. Song. (2012). "Timing channels with multiple identical quanta." [Online]. Available: https://arxiv.org/abs/1208.1070

[8] A. Abid, A. Poon, and J. Zou. (May. 2017). "Linear regression with shuffled labels." [Online]. Available: https://arxiv.org/abs/1705.01342

[9] A. B. Poore and S. Gadaleta, "Some assignment problems arising from multiple target tracking," *Math. Comput. Modeling*, vol. 43, nos. 9–10, pp. 1074–1091, 2006.

[10] S. Thrun and J. J. Leonard, "Simultaneous localization and mapping," in *Springer Handbook of Robotics*. Berlin, Germany: Springer, 2008, pp. 871–889.

[11] W. S. Robinson, "A method for chronologically ordering archaeological deposits," *Amer. Antiquity*, vol. 16, no. 4, pp. 293–301, 1951.

[12] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Security Privacy (SP)*, May 2008, pp. 111–125.

[13] L. Keller, M. J. Siavoshani, C. Fragouli, K. Argyraki, and S. Diggavi, "Identity aware sensor networks," in *Proc. IEEE INFOCOM*, Apr. 2009, pp. 2177–2185.

[14] P. David, D. Dementhon, R. Duraiswami, and H. Samet, "SoftPOSIT: Simultaneous pose and correspondence determination," *Int. J. Comput. Vis.*, vol. 59, no. 3, pp. 259–284, 2004.

[15] M. Marques, M. Stošić, and J. Costeira, "Subspace matching: Unique solution to point matching with geometric constraints," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 1288–1294.

[16] S. Mann, "Compositing multiple pictures of the same scene," in *Proc. 46th IST Annu. Conf.*, 1993, pp. 50–52.

[17] L. J. Schulman and D. Zuckerman, "Asymptotically good codes correcting insertions, deletions, and transpositions," *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 2552–2557, Nov. 1999.

[18] X. Huang and A. Madan, "CAP$_3$: A DNA sequence assembly program," *Genome Res.*, vol. 9, no. 9, pp. 868–877, 1999.

[19] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, Dec. 2009.

[20] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[21] V. Emiya, A. Bonnefoy, L. Daudet, and R. Gribonval, "Compressed sensing with unknown sensor permutation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 1040–1044.
[22] G. Elhami, A. Scholefield, B. B. Haro, and M. Vetterli, "Unlabeled sensing: Reconstruction algorithm and theoretical guarantees," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4566–4570.
[23] D. Hsu, K. Shi, and X. Sun. (2017). "Linear regression without correspondence." [Online]. Available: https://arxiv.org/abs/1705.07048
[24] A. Pananjady, M. J. Wainwright, and T. A. Courtade, "Denoising linear models with permuted data," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 446–450.
[25] S. Haghighatshoar and G. Caire. (2017). "Signal recovery from unlabeled samples." [Online]. Available: https://arxiv.org/abs/1701.08701
[26] Y. M. Lu and M. N. Do, "A theory for sampling signals from a union of subspaces," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2334–2345, Jun. 2008.
[27] T. Blumensath, "Sampling and reconstructing signals from a union of linear subspaces," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4660–4671, Jul. 2011.
[28] O. Collier and A. S. Dalalyan, "Minimax rates in permutation estimation for feature matching," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 162–192, 2016.
[29] N. Flammarion, C. Mao, and P. Rigollet. (2016). "Optimal rates of statistical seriation." [Online]. Available: https://arxiv.org/abs/1607.02435
[30] F. Fogel, R. Jenatton, F. Bach, and A. D'Aspremont, "Convex relaxations for permutation problems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1016–1024.
[31] M. G. Rabbat, M. A. T. Figueiredo, and R. D. Nowak, "Network inference from co-occurrences," *IEEE Trans. Inf. Theory*, vol. 54, no. 9, pp. 4053–4068, Sep. 2008.
[32] V. Gripon and M. Rabbat, "Reconstructing a graph from path traces," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2013, pp. 2488–2492.
[33] N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainwright, "Stochastically transitive models for pairwise comparisons: Statistical and computational issues," *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 934–959, Feb. 2017.
[34] S. Chatterjee, "Matrix estimation by universal singular value thresholding," *Ann. Stat.*, vol. 43, no. 1, pp. 177–214, 2015.
[35] A. Pananjady, C. Mao, V. Muthukumar, M. J. Wainwright, and T. A. Courtade, "Worst-case vs average-case design for estimation from fixed pairwise comparisons," *CoRR*, Jul. 2017. [Online]. Available: http://arxiv.org/abs/1707.06217
[36] J. Huang, C. Guestrin, and L. Guibas, "Fourier theoretic probabilistic inference over permutations," *J. Mach. Learn. Res.*, vol. 10, pp. 997–1070, 2009.
[37] E. M. Loiola, N. M. M. de Abreu, P. O. Boaventura-Netto, P. Hahn, and T. Querido, "A survey for the quadratic assignment problem," *Eur. J. Oper. Res.*, vol. 176, no. 2, pp. 657–690, 2007.
[38] C. E. Shannon, "Probability of error for optimal codes in a Gaussian channel," *Bell Syst. Tech. J.*, vol. 38, no. 3, pp. 611–656, 1959.
[39] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of Johnson and Lindenstrauss," *Random Struct. Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.
[40] K.-I. Yoshihara, "Simple proofs for the strong converse theorems in some channels," *Kodai Math. Seminar Rep.*, vol. 16, no. 4, pp. 213–222, 1964.
[41] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*. Cambridge, MA, USA: Cambridge Univ. Press, 1952.
[42] A. Vince, "A rearrangement inequality and the permutahedron," *Amer. Math. Monthly*, vol. 97, no. 4, pp. 319–323, 1990. [Online]. Available: http://dx.doi.org/10.2307/2324517
[43] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*. North Chelmsford, MA, USA: Courier Corporation, 1998.
[44] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, U.K.: Oxford Univ. Press, 2013.

Ashwin Pananjady is a PhD candidate in the Department of Electrical Engineering and Computer Sciences (EECS) at the University of California, Berkeley. He received the B.Tech. degree (with honors) in electrical engineering from IIT Madras in 2014. His research interests include statistics, optimization, and information theory. He is a recipient of the Governor's Gold Medal from IIT Madras, and an Outstanding Graduate Student Instructor Award from UC Berkeley.

Martin J. Wainwright is currently Chancellor's Professor at the University of California at Berkeley, with a joint appointment between the Department of Statistics and the Department of Electrical Engineering and Computer Sciences (EECS). He received a Bachelor's degree in Mathematics from University of Waterloo, Canada, and Ph.D. degree in EECS from Massachusetts Institute of Technology (MIT). His research interests include high-dimensional statistics, information theory, statistical machine learning, and optimization theory. Among other awards, he has received Best Paper Awards from IEEE Signal Processing Society (2008), IEEE Information Theory Society (2010); a Medallion Lectureship (2013) from the Institute of Mathematical Statistics (IMS); a Section Lectureship at the International Congress of Mathematicians (2014); the COPSS Presidents' Award (2014) from the Joint Statistical Societies; and the IMS David Blackwell Lectureship (2017).

Thomas A. Courtade received the B.Sc. degree (summa cum laude) in electrical engineering from Michigan Technological University, Houghton, MI, USA, in 2007, and the M.S. and Ph.D. degrees from the University of California, Los Angeles (UCLA), CA, USA, in 2008 and 2012, respectively. He is an Assistant Professor with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA. Prior to joining UC Berkeley in 2014, he was a Postdoctoral Fellow supported by the NSF Center for Science of Information. Prof. Courtade's honors include a Distinguished Ph.D. Dissertation Award and an Excellence in Teaching Award from the UCLA Department of Electrical Engineering, and a Jack Keil Wolf Student Paper Award for the 2012 International Symposium on Information Theory. He was the recipient of a Hellman Fellowship in 2016.