Mach Learn (2016) 102:29–56
DOI 10.1007/s10994-015-5489-9

Population recovery and partial identification

Avi Wigderson[1] · **Amir Yehudayoff**[2]

Received: 3 April 2014 / Accepted: 5 March 2015 / Published online: 1 April 2015

Abstract We study several problems in which an *unknown* distribution over an *unknown* population of vectors needs to be recovered from partial or noisy samples, each of which nearly completely erases or obliterates the original vector. For example, consider a distribution p over a population $V \subseteq \{0,1\}^n$. A noisy sample v' is obtained by choosing v according to p and flipping each coordinate of v with probability say 0.49 independently. The problem is to recover V, p as efficiently as possible from noisy samples. Such problems naturally arise in a variety of contexts in learning, clustering, statistics, computational biology, data mining and database privacy, where loss and error may be introduced by nature, inaccurate measurements, or on purpose. We give fairly efficient algorithms to recover the data under fairly general assumptions. Underlying our algorithms is a new structure we call a *partial identification (PID) graph* for an arbitrary finite set of vectors over any alphabet. This graph captures the extent to which certain subsets of coordinates in each vector distinguish it from other vectors. PID graphs yield strategies for dimension reductions and re-assembly of statistical information, and so may be useful in other applications as well. The quality of our algorithms (sequential and parallel runtime, as well as numerical stability) critically depends on three parameters of PID graphs: width, depth and cost. The combinatorial heart of this work is showing that *every* set of vectors possesses a PID graph in which all three parameters are small (we prove some limitations on their trade-offs as well). We further give an efficient algorithm to find such near-optimal PID graphs for any set of vectors. Our efficient PID graphs imply general algorithms for these recovery problems, even when loss or noise are just below the information-theoretic limit. In the learning/clustering context this gives a new algorithm for learning mixtures of binomial distributions (with known marginals)

Editor: Tamas Horvath.

✉ Amir Yehudayoff
amir.yehudayoff@gmail.com

Avi Wigderson
avi@math.ias.edu

[1] Institute for Advanced Study, Princeton, NJ, USA

[2] Technion–IIT, Haifa, Israel

whose running time depends only quasi-polynomially on the number of clusters. We discuss implications to privacy and coding as well.

Keywords Noise recovery · Identification · Partial information

1 Introduction

Recovery of information from lossy or noisy data is common to many scientific and commercial endeavours. In this work we introduce a new structure we call a partial identification (PID) graph that helps in solving such recovery problems.

The use of IDs, features of individuals that *uniquely* identify them in a given population, is ubiquitous. Fingerprints and retinal scans, or simply names and addresses, are useful for identifying people. In other populations, various appropriate subsets of features are used. In some populations, however, some IDs may be too long, and thus too expensive to maintain, e.g., when storage is limited. In these cases, *partial* IDs can be used: instead of a long list of features per individual, we can maintain a shorter list that is more efficient.[1] This efficiency has a cost, as it may introduce ambiguity and "imposters." PID graphs represent the imposter-structure of a given choice of PIDs. A detailed discussion and formal definitions appear in Sect. 1.3.

We provide fairly efficient algorithms for recovery and clustering of information from lossy/noisy data. These algorithms use PID graphs to identify useful local sub-systems of a given system, and combine local pieces of information to a global one. The complexity of our algorithms crucially depends on the PID graphs they use. A substantial part of our algorithms, therefore, is a procedure that builds efficient PID graphs.

A roadmap As the introduction is long, here is a roadmap to it. In Sect. 1.1 we give intuition and formally define population recovery problems, samplers and our main results. In Sect. 1.2 we explain why a simpler "distribution recovery" problem is sufficient to solve. Sects. 1.3 and 1.4 respectively explain how parameters of PID graphs control our algorithms, and how to efficiently achieve good parameters. Section 1.6 extensively discusses our work in various contexts which considered similar problems and methods.

1.1 Recovery problems

Let us start with two scenarios that motivate the kind of problems we consider in this paper. We then formulate them more precisely.

Recovery from lossy samples Imagine that you are a paleontologist, who wishes to determine the population of dinosaurs that roamed the Earth before the hypothetical meteorite lead to their extinction. Typical observations of dinosaurs consist of finding a few teeth of one here, a tailbone of another there, perhaps with some more luck a skull and several vertebrae of a third, and rarely a near complete skeleton of a fourth. Each observation belongs to one of several species. Using these fragments, you are supposed to figure out the population of dinosaurs, namely, a complete description of (say) the bone skeleton of each species, *and* the fraction that each species occupied in the entire dinosaur population. Even assuming that our probability of finding the remains of a particular species, e.g. Brontosaurus, is the same as its fraction in the population, the problem is clear: while complete features identify the species, fragments may be common to several. Even with knowledge of the complete description of

[1] It will be clear that in the settings we consider, hashing and related techniques are irrelevant, and one has to use actual features of individuals.

each type (which we a-priori do not have), it is not clear how to determine the distribution from such partial descriptions. A modern-day version of this problem is analyzing the partial Netflix matrix, where species are user types, features are movies, and each user ranked only relatively few movies.

Recovery from noisy samples Imagine you are a social scientist who wants to discover behavioral traits of people and their correlations. You devise a questionnaire where each yes/no question corresponds to a trait. You then obtain a random sample of subjects to fill the questionnaire. Many of the traits, however, are private, so most people are unwilling to answer such questions truthfully, fearing embarrassment or social consequences if these are discovered.[2] To put your subjects at ease you propose that instead of filling the questionnaire truthfully, they fill it randomly: as follows: to each question, they flip the true answer with probability 0.49, independently of all others. This surveying method (typically applied when there is one question of interest) is known as "randomized response" and was initiated by Warner (1965). Is privacy achieved? Can one recover the original truthful information about the original sequences of traits and their distribution in the population? Observe that typical samples look very different than the sequences generating them. Indeed, can one synthesize sensitive databases for privacy by publishing very noisy versions of their records?

These examples illustrate some of the many settings in which lossy or noisy samples of concrete data is available, and from which one would like to efficiently recover a full description of the original data. Losses and noise may be an artifact of malicious or random actions, and the entity attempting recovery may be our friends or foes. We formalize below the general setting and show that, surprisingly, accurate recovery is possible, as long as samples bear minimal correlation with *some* data item. This is achieved even when *both* the dimension (number of "attributes") and the population size (number of "clusters") are large!

1.1.1 Definitions and results

Fix an alphabet Σ, integer parameters n, k and an accuracy parameter $\alpha > 0$. Assume, as in the examples above, there is a population of vectors $V \subseteq \Sigma^n$ of size (at most) k, and a probability distribution π on the population V. In the following, we shall not know π or even V, but rather only get lossy or noisy samples of them. Our goal will be to fully recover both V and π.

The *population recovery problem* (PRP) is to (approximately) reconstruct V and π, up to the accuracy parameter α. Namely, given independent lossy or noisy samples, find a set V' containing every element $v \in V$ for which $\pi(v) > \alpha$, and a distribution π' on V' such that for every $v \in V'$ we have $|\pi(v) - \pi'(v)| < \alpha$.

We start by focusing on two processes for generating samples (which we will later generalize): lossy sampling and noisy sampling. This type of recovery problem was first formulated in Dvir et al. (2012), for the model of lossy sampling. The sampling procedures we consider are parametrized by $\mu, \nu \in [0, 1]$. When μ and ν are 0, the samples provide zero information on the underlying distribution, and when μ and ν are 1, the samples provide full information (i.e., there is no noise or loss).

Lossy samples Let $0 \leq \mu \leq 1$. A *lossy* sample v' with parameter μ is generated randomly as follows. First, an element $v \in V$ is picked at random with probability $\pi(v)$. Then, independently at random, every coordinate v_i of v is replaced by the symbol "?" (assumed not to be

[2] A possibly worse scenario is when looking for information concerning medical conditions, where leakage of the truth may cause one to lose medical insurance.

in Σ) with probability $1-\mu$, and is left untouched with probability μ. The result is v'. For example, a lossy sample with parameter 0 is a vector of question marks, and a lossy sample with parameter 1 gives full access to V.

Noisy samples Let $0 \leq \nu \leq 1$, and for simplicity assume $\Sigma = \{0,1\}$. A *noisy* sample v' with parameter μ is generated as follows. First, an element $v \in V$ is picked at random with probability $\pi(v)$. Then, independently at random, every bit v_i of v is flipped with probability $(1-\nu)/2$, and left untouched with probability $(1+\nu)/2$. The result is v'. For example, a noisy sample with parameter 0 is a uniform random vector from the discrete cube (regardless of V, π).

The nature of the recovery problem, namely, that we only have access to samples, requires that we allow algorithms to err with some small probability $\delta > 0$. The main question we study is "for which noise/loss parameters is there an efficient reconstruction algorithm, hopefully that runs in time polynomial in all other parameters $|\Sigma|, n, k, 1/\alpha, \log(1/\delta)$?"

It is not hard to see that, in the noisy model, if $\nu = 0$ it is information theoretically impossible to solve the recovery problem. If $\nu = o(1)$ then it is impossible to solve the recovery problem in time polynomial in k, where $o(1)$ means as $k \to \infty$. For example, consider the case of distinguishing all even-weight strings from all odd-weight strings of length $n = \log k$. The statistical distance between the two distributions on samples is at most $(2\nu)^n$, since it is at most $2^n \left| \sum_{j=0}^{n} \binom{n}{j} (-1)^j ((1+\nu)/2)^j ((1-\nu)/2)^{n-j} \right|$. Similarly, for $\mu = 0$ the lossy recovery problem is information theoretically impossible, and for $\mu = o(1)$ it is impossible to solve it in time polynomial in k, where $o(1)$ means as $k \to \infty$.

In all other cases, namely when $\mu, \nu > 0$ are constants, we design algorithms for these two problems, which run in time polynomial in $|\Sigma|, (nk)^{\log k}, 1/\alpha$ and $\log(1/\delta)$. The exponent of the running-time polynomial depends logarithmically on $1/\mu, 1/\nu$, respectively. The question of reducing the dependence on k to polynomial is the main open problem. We now state the main results, starting with the lossy case and then describing the noisy case.

Theorem 1.1 *Let $0 < \mu < 1$. There is an algorithm that, given an integer k, $\alpha > 0$, $0 < \delta < 1$ and μ, runs in time $\mathsf{poly}_{\mu}(|\Sigma|, (nk)^{\log k}, 1/\alpha, \log(1/\delta))$ and solves the recovery problem for an n-dimensional database of size k over Σ, given lossy samples with parameter μ, with error α and probability at least $1-\delta$.*

Independence of erasures of different coordinates is not necessary for the algorithm stated in the theorem to perform well. Indeed, the algorithm works as long as every set of size at most $\log k$ coordinates has a non-negligible probability (say β) of having no question marks. The running time in this case may be slowed down by a factor of $\mathsf{poly}(1/\beta)$. For example, as long as the question marks are $(\log k)$-wise independent, we have $1/\beta = (1/\mu)^{\log k} = \mathsf{poly}(k)$.

In Dvir et al. (2012) a completely different algorithm was given for the lossy case, which is fully polynomial in all parameters including k, but works only as long as $\mu > 0.365$. A recent and beautiful analysis of Moitra and Saks (2013) that is based on the behaviour of complex harmonic functions showed that the algorithm of Dvir et al. (2012) in fact runs in polynomial time for every $\mu > 0$.

Our main result for recovery from noisy samples is similar to the lossy one.

Theorem 1.2 *Let $0 < \nu < 1$. There is an algorithm that, given an integer k, $\alpha > 0$ and $0 < \delta < 1$ and ν, runs in time $\mathsf{poly}_{\nu}((kn)^{\log k}, 1/\alpha, \log(1/\delta))$ and solves the recovery problem for an n-dimensional database of size k over $\Sigma = \{0,1\}$ with noisy samples with parameter ν, error α and probability at least $1-\delta$.*

For $n > \log k$, there is an algorithm that runs in time $\mathsf{poly}_{\nu}(n, (1/\alpha)^{\log k}, (1/\nu)^{\log k}, \log(1/\delta), k^{\log^2 k})$ and performs the same task even without *knowing the noise parameter ν.*

Theorem 1.2 implies, in particular, that for $\log k < n$ the distribution p is unique in that for every other distribution $\overline{p}$ that is far from p and for every noise parameter $\overline{\nu}$, the distribution on samples that $\overline{p}, \overline{\nu}$ yield is far from the one produced by p, ν (for $\log k = n$, there is no such uniqueness).

The study of recovering information from noisy samples appears already in the work of Kearns et al. (1994). There it is called "learning mixtures of hamming balls" and two algorithms are presented. One with a similar running time to our algorithm, but that only guarantees small divergence, that is, the population V is not reconstructed exactly but only up to hamming distance roughly $\sqrt{n}$. A second that basically solves PRP using noisy samples, but runs in time exponential in k.

In a followup work, Lovett and Zhang (2015) provided a more efficient learning algorithm for PRP that runs in time $\mathsf{poly}(n, k^{\log\log(k)})$. Their algorithm uses Fourier analytic properties of sparse boolean functions, and is therefore quite different than ours. Their strategy however is specific to noisy samples, whereas our holds for more general samplers.

With an appropriate definition of noise, one can generalize this theorem to every alphabet Σ. Moreover, it can also be generalized to the case that a different (but known) noise probability is applied in every coordinate. As in the lossy case, we do not need the full independence of the noise but only $(\log k)$-wise independence. We do, however, require knowledge of the noise parameter ν (or at least some polynomially good estimate of it).

More general samplers A third sampler one can consider is a *hybrid sampler* combining the lossy and noisy ones. In the hybrid model, a sample is generated by first erasing all but a μ fraction of the coordinates, and then flipping each one of the remaining coordinates with probability $(1-\nu)/2$. A fourth sampler is a noisy sampler where every coordinate i is flipped with probability $(1-\nu_i)/2$ that depends on i. Even in these more complex situations we can achieve the same bounds on recovery (in general knowledge of the noise model is required for our algorithm to perform well). In fact, when proving the above theorems, we consider even a more general type of samplers that we define in Sect. 3.1 below.

Our algorithms use the following strategy. Due to the partial, distorted and inconsistent information provided by samples, we focus on small *windows* (subsets of coordinates). The strategy consists of two phases. In the first phase, we collect statistics for many judiciously chosen windows. This phase may be viewed as a collection of dimension reductions of the original problem. Since windows are small (of logarithmic size), we can afford running time that is exponential in the window size, and so we can get pretty accurate information for each of the chosen windows. However, as many vectors may agree on their projection on a small window, we need to "decouple" the statistics thus obtained. This is done in the second phase, where special care is needed to control accumulation of errors. The implementation of this strategy is guided by a PID graph.

To build the PID graph though, we need to know the underlying population (which is actually part of what we want to learn in PRP). In the next section we discuss the simpler problem of learning when we already know the population. It already requires some of the key ideas. The full solution of PRP is discussed in Sect. 4, and interestingly it implicitly uses the PID graph construction (although V is unknown).

1.2 Recovery when vectors are known

The *distribution recovery problem* (DRP) is a sub-problem of PRP, in which the set V of vectors in the population is actually given to the algorithm, and the only problem is estimating the unknown probability distribution π on V.

An important reduction provided in Dvir et al. (2012) shows that PRP can be efficiently reduced to DRP. If from V we can deduce π, we can find V as well. In other words, we may assume V is known even though it is not. This somewhat counter-intuitive statement has a very simple proof and moreover is quite general. It works for all *uniform* samplers of V, in particular, the ones we discussed above. We note that it is very similar in spirit to the prefix constructions in other learning settings, such as Goldreich and Levin (1989), Kushilevitz and Mansour (1993). For completeness (and since it is not stated in this generality in Dvir et al. 2012), we provide the formal definition of a uniform sampler and a proof sketch in the appendix.

Although the reduction from Dvir et al. (2012) is quite general it is not sufficient for us, due to the way errors accumulate in our algorithm.[3] Later on we explain how to remove the assumption that we know V. We also mention that, in a follow-up work, Batman et al. (2013) show that recovering a super-set of V, but not π, can be done in polynomial time from lossy samples and quasi-polynomial time from noisy ones.

Assume for now that we know V and only need to estimate π. The approach taken here, which is very different from that of Dvir et al. (2012), is to first efficiently construct a "succinct" representation for vectors in V. To motivate it, let us first observe that DRP is much easier if the number n of "features" of an individual was small, say $n \leq O(\log k)$ where $|V| = k$. This allows us to use algorithms that run in exponential time in n.

The following proposition states that exponential running time suffices for solving the problem (and, in fact, for a more general family of samplers; see Sect. 3.1 for details).

Proposition 1.3 *DRP can be solved in both sample models in time exponential in n.*

How can we make use of the proposition when n is large (as it typically is)? For this we introduce a generic way to reduce the dimension: Use PID graphs.

1.3 Dimension reductions and PID graphs

At this point V is available to us and we can preprocess it. Our approach is to pick, for every individual $v \in V$, a small representative set of features $S_v \subseteq [n]$, of size $|S_v| \leq \log k$. Then, solve DRP on the window S_v, ignoring all other coordinates, in time $\mathsf{poly}(k)$ using the lemma above. For a vector v and a set S, denote by $v[S]$ the restriction of v to the entries in S. If $v[S_v]$, the pattern of v in coordinates S_v, *uniquely* identifies v in V (namely it is an ID), then we directly get an estimate of $\pi(v)$. Unfortunately, such short IDs do not exist for all vectors in a general population V. For some vectors v the pattern on S_v is shared by other *imposters* $u \in V$. Thus the set S_v is only a *partial ID* (PID) for v, and the statistics obtained from the DRP solver on S_v counts the total mass of π of all imposters u of v. Recovering π from this information is clearly a linear system. This linear system is not always invertible, so we first need to ensure that it is. But, even when it is invertible, recall that we only have estimates, and inverting the system can exponentially blow up the accuracy parameter.

To study the cost of this computation, it is natural to capture "imposter relations" in a graph. A *PID graph* $G = (V, E, \{S_v : v \in V\})$ is a labelled directed graph defined as follows. The vertices of G are the elements of $V \subseteq \Sigma^n$. Every vertex $v \in V$ is labeled by a PID $S_v \subseteq [n]$. Directed edges go from every vertex v to every imposter u of v, namely, $u \neq v$ such that $u[S_v] = v[S_v]$.

We also require that a PID graph is acyclic (this is essential for it to be useful in applications). For every V, there are choices of PIDs that yield an acyclic graph (for example,

[3] We thank an anonymous referee for pointing this out.

$S_v = [n]$ for all v), and there are choices that yield a graph with directed cycles (like $S_v = \emptyset$ for all v). In this work, we focus on choices that yield acyclic graphs, and achieve efficient constructions under this requirement.

The set of *imposters* $I(v)$ of v contains v itself and all vertices u such that (v, u) is an edge in G. Let $p(v) = \sum_{u \in I(v)} \pi(u)$. Since G is acyclic, the linear system expressing p in terms of π is *triangular*. So, knowing p, we can compute π by backwards substitution. But, as we shall only have an approximation p' of p, we need to ensure that the initial error of p' is small enough so that accumulated error in π' will be small as well.

The accumulation of errors in this computation is affected by $\mathsf{depth}(G)$, the length of the longest directed path in it. This motivates using a graph G of smallest possible depth. The depth also dominates the running time of any parallel algorithm for this task. This is a secondary reason for minimizing the depth.

Depth is just a crude parameter controlling error accumulation. A more precise one is $\mathsf{cost}(G)$, defined recursively for acyclic G according to its topological order by $\mathsf{cost}(v) = 1$ on every sink v in G, and

$$\mathsf{cost}(v) = 1 + \sum_{u \in I(v): u \neq v} \mathsf{cost}(u).$$

Note that $\mathsf{cost}(v)$ is simply the number of directed paths (of all lengths) emanating at v. Define

$$\mathsf{cost}(G) = \max\{\mathsf{cost}(v) : v \in V\}.$$

Trivially,

$$\mathsf{cost}(G) \leq |V|^{\mathsf{depth}(G)}.$$

As we already explained, the runtime of solving each dimension-reduced DRP problem on S_v is exponential in $|S_v|$. This aspect is captured by the *width* of G,

$$\mathsf{width}(G) = \max\{|S_v| : v \in V\}.$$

We conclude by showing how the runtime of the full DRP algorithm depends on these three parameters.

Theorem 1.4 *Let G be a PID graph for a population V of size $|V| = k$.*

- *Given G, the distribution recovery problem with accuracy $\alpha > 0$ for V can be solved, with probability at least $1 - \delta$, using lossy samples with parameter μ in time T for $T = \mathsf{poly}_{\mu}(nk|\Sigma| \log(1/\delta)/\alpha) \cdot \mathsf{cost}(G) \cdot (1/\mu)^{\mathsf{width}(G)}$. Moreover, it can be solved in parallel using T processors in time $\mathsf{depth}(G) \cdot \mathsf{poly} \log T$.*
- *Given G, the distribution recovery problem with accuracy $\alpha > 0$ for V can be solved, with probability at least $1 - \delta$, using noisy samples with parameter ν in time T for $T = \mathsf{poly}_{\nu}(nk \log(1/\delta)/\alpha) \cdot \mathsf{cost}(G) \cdot (1/\nu)^{\mathsf{width}(G)}$. Moreover, it can be solved in parallel using T processors in time $\mathsf{depth}(G) \cdot \mathsf{poly} \log T$.*

1.4 Efficient PID graphs

We are now faced with constructing, for any population V, a PID graph with the smallest possible width and depth (and cost). Minimizing only one of these parameters is easy; considering the following examples will help getting some intuition.

To minimize depth, pick $S_v = [n]$ for all v. Since these PIDs are actually IDs, there are no imposters and the depth is zero. The width, however, is n, which would lead to an exponential

time algorithm. Note that if we insist on IDs, the width may have to be n, e.g. when V consists of all unit vectors e_i as well as the all zeros vector $\bar{0}$. While each e_i has an ID of size one, the only ID for the vector $\bar{0}$ is $[n]$.

On the other extreme, let us see how we can easily make the width at most logarithmic. In this case, however, the depth is possibly k, which again leads to an exponential running time (due to error-accumulation). First observe[4] that in any V, there must be at least one member v with an ID S_v of size at most $\log |V|$. We can now remove v from the population, and consider the set $\tilde{V} = V \backslash \{v\}$. We can now find a vector $\tilde{v}$ which has a short ID in $\tilde{V}$. Even though $S_{\tilde{v}}$ is an ID of $\tilde{v}$ in $\tilde{V}$, the vector v may still be an imposter of $\tilde{v}$ in V. Continuing in this fashion results in a set of PIDs, each of which has size at most $\log k$, but may yield a graph of depth k. Indeed, if our set of vectors V is simply all $n+1$ vectors of the form $1^i 0^{n-i}$, then all PIDs thus obtained have size at most 1, but their graph is a path of length $n+1$. This graph happens to have small cost, but similar examples in which the cost is exponential can be given as well.

We prove that width, depth (and thus cost) can simultaneously be made small. Moreover, we can find such PIDs efficiently.

Theorem 1.5 *There is an algorithm that, given V of size k, runs in time* $\mathsf{poly}(n,k)$*, and finds a PID S_v for every v in V so that the corresponding PID graph G is acyclic with* $\mathsf{width}(G) \leq \log k$ *and* $\mathsf{depth}(G) \leq \log k$*. Specifically,* $\mathsf{cost}(G) \leq k^{\log k}$.

Observe that finding the smallest possible ID for a given v in V is exactly the hyper-graph vertex-cover problem which is NP-hard to compute.

It would be preferable to obtain short PIDs whose graph has a polynomial upper bound on the cost. Such a bound would automatically yield a polynomial running time for our algorithms. Unfortunately, such a bound cannot be obtained.

Proposition 1.6 *For every k there is a database V of size k over $\Sigma = \{0, 1\}$ so that* every *PID graph for V of width at most $O(\log k)$ has super-polynomial cost, at least $k^{\Omega(\log \log k)}$.*

This example shows that in order to solve DRP in polynomial time new ideas are required. We note, however, that for most databases V, the cost of a PID graph is much smaller. So, for an "average input" our algorithm runs faster.

1.5 Recovery when vectors are unknown

The discussion above focused on the case that the population V is known. An important part of our main goal is, however, to recover V itself. This turns out to be fairly simple using the ideas presented so far. The full algorithm appears in Sect. 4.

Here is a rough description of our recovery algorithm. Although V is not known, we do know that it has a PID graph of width w and depth d, both are at most $\log k$. The PID graph for V can therefore be thought of having d levels. The top most level corresponds to vectors v in V that have a unique ID, and the lower levels only have PIDs. Based on this knowledge, as we explain in more detail below, we shall perform an iterative process. Starting at the top, each iteration shall reveal one more level of the graph G, based on the knowledge obtained in previous iterations. Since we have an a priori bound on the number of levels, we know when to stop. We shall also be able to bound the error terms sufficiently well.

[4] Indeed, to find such v, S_v take the left-most (non-constant) coordinate in $[n]$, and keep only vectors in V that have the least popular value in this coordinate. Now do it again, and again, until you are left with only one vector v. This process chooses at most $\log |V|$ coordinates, which serve as an ID for v.

Learning the top most level: There are $\binom{n}{w} \leq n^{\log k}$ subsets of $[n]$ that are potential unique IDs for the vectors in the top most level. Given such a potential unique ID S, we can check in time roughly $\mathsf{poly}(n, 2^w) = \mathsf{poly}(n, k)$ whether S is a unique ID for some vector v_S in V or not (similarly to Proposition 1.3). Once we find such an S and v_S, we can also recover $v_S[S]$, and we can recover the extension of v_S to a coordinate $i \notin S$ by solving DRP in $S \cup \{i\}$. This again takes time $\mathsf{poly}(n, k)$.

The lower levels We have thus recovered the top most level $V_0 \subseteq V$, and have actually obtained a pretty accurate estimate of π on V_0. To recover the second level, we just consider the top most level in $V \setminus V_0$, by reweighing the samples from V using our estimate of π on V_0. We thus estimate π on the second level V_1, and so forth. Since the depth of this procedure is small (at most $\log k$), the overall numerical error is under control.

1.6 Discussion and related work

In this section we discuss our results and our methods in comparison to similar ones in the literature. There are numerous works dealing with a wide variety of recovery from lossy and noisy information, and we cannot hope to survey them all. These arise in various fields, and sometimes with different motivations, assumptions and terminology, and we try to relate them to ours. The partition to topics below is somewhat arbitrary and certainly overlapping.

Learning and clustering The population recovery problem from lossy samples was introduced in Dvir et al. (2012). It arose naturally from the analysis of PAC-learning algorithms in a general non-black-box learning framework they introduced, called *restriction access*. Here we introduced noisy (and more general) sample-models. The algorithms we describe suggest that the restriction access model can be generalized to allow noise as well as loss. We note that our recovery algorithms are very different than the ones in Dvir et al. (2012).

Another natural way to view the noisy model is as follows. The unknown vectors v in our database are k cluster *centers*, and the noise generates a mixture of *binomial* distributions around these centers, with unknown mixture coefficients determined by π. This problem was considered when the binomial noise is *unknown*, and Blum et al. (2009), Feldman et al. (2008) give algorithms to learn such mixtures. A different closely related problem is the well-studied learning mixtures of Gaussians problem (see Kalai et al. 2012 and references therein). The most crucial difference between these works and ours is that we assume to know the binomial distributions around the cluster, and previous works do not. This is, of course, less general, but allows us to achieve a better dependence on k, the number of centers. For example, previous algorithms are polynomial only for constant k, but ours is polynomial even for k polynomial in $2^{\sqrt{\log n}}$.

Another difference is that in the Gaussian noise model the distribution is continuous and the vectors naturally reside in high dimensional Euclidean space, whereas the binomial noise is discrete and vectors reside in high dimensional combinatorial cube. The Gaussian distribution is spherically symmetric and Euclidean space allows projections in all directions, whereas in the Boolean case symmetry is restricted and allows only projection on coordinates.

Teaching Theoretical frameworks for teaching have also been considered. A learner and a teacher wish to interact so that after the teacher gives a few examples of classification to the student, the student is able to perform classification by herself. A specific type of interaction is using compression schemes (see e.g. Floyd and Warmuth 1995; Littlestone and Warmuth 1986). Compression schemes are very similar to PID graphs. A *compression scheme* corresponds to the function $v \mapsto S_v$, except that in a compression scheme this map should be so that $v[S_v \cup S_{v'}] \neq v'[S_v \cup S_{v'}]$ for all $v \neq v'$ in V. The *size* of a compression

scheme $\max\{|S_v|\}$ corresponds to the width of the corresponding PID graph. Every PID graph yields a compression scheme. There are always compression schemes of size at most $\log |V|$ and the smallest possible size of compression schemes is still not fully understood. Our construction gives compression schemes with the additional property of small depth, which may be useful for teaching as well.

Codes Yet another, quite different view of our result is as a surprising robustness of repetition codes. Imagine the following way to encode a *multi-set* of k binary vectors of length n: simply make many (say m) copies of each. We show that this simple repetition code tolerates the following vicious kind of attack. First, the km vectors are arbitrarily permuted. Next, in each vector, a random 99 % of the coordinates are erased (replaced by question marks). Finally, the surviving bits are each flipped with probability 49 %. It is not clear that this leaves enough information to recover the original data. We show that for rate $m = \mathsf{poly}(n, k^{\log k})$, recovery is not only information theoretically possible but can be performed in polynomial time. Of course, determining the minimal rate for which this is possible is an intriguing question (relevant to the next item as well).

Privacy Resolving the conflict between individual privacy in large databases and their usability for statistical, medical and social research is an important research area. Again, we can do no justice to the volumes of work on the subject, and we resort to generalities. One general principle that is used is to obliterate individual records and perturb the data by applying to it random noise of some form, while allowing access to some statistics or aggregate information of reasonable quality. In most settings under consideration, a family of queries to the database is allowed, and then noise can be applied to the data itself, or to the result of the query, or to both (see examples in Liew et al. 1999; Lefons et al. 1983; Beck 1980; Traub et al. 1984; Agrawal and Srikant 2000; Dinur and Nissim 2003; Dwork and Smith 2008; Dwork et al. 2006 and their many references). An important part of these works includes defining notions of individual privacy, finding mechanisms which ensure privacy under general families of queries and/or proving that no such mechanisms exist for other families of queries. Most impossibility results are about noise applied to query answers, as e.g. in Dinur and Nissim (2003) who show how (almost all) of the database can be recovered if the noise is not sufficiently high to be practically useless. Our work has a similar flavor, and shows that under some (quite general) conditions, information can be recovered from highly noisy data. Thus it may be viewed as another general impossibility result for privacy of *databases*. Yet it seems that the same result may be viewed positively when applied to the privacy of *questionnaires*; it suggests that individuals who are asked to answer sensitive questions can be allowed to honestly answer each, then randomly flip each answer with probability (say) 0.49 (thus providing complete deniability). Our algorithms can process such noisy collection of questionnaires and completely recover the prevailing patterns of honest answers.

Dimension reduction High dimension of data is often considered a curse for efficient algorithms. Often the best known (or even best possible) running time is exponential in the dimension. One common remedy used in numerous algorithms for such data is *dimension reduction*. The most famous one is the Johnson–Lindenstrauss reduction (Johnson and Lindenstrauss 1984), showing that a random linear projection of k points in Euclidean space to a subspace of dimension $O(\log k)$ nearly preserves all their pairwise Euclidean distances. Similar reductions in many settings and problems allow much faster processing of the data in the lower-dimensional space, and then "lifting" the solution to the original space, often losing something in its quality. Typically, reductions are random, and are done once (though there are exceptions to both: sometimes reductions move to a subspace determined by some

top singular eigenvectors, and e.g. in the aforementioned Gaussian mixtures learning, many 1-dimensional projections are used). Also in our algorithms we use many different dimension reductions, projecting the data onto very different small subsets of the coordinates. But for the problems we deal with random projections are useless in general, and judicious choice is needed to guarantee all properties of our PID graphs.

Identification Unique identifiers, IDs for short, are prevalent for people, animals, cars, computers etc. Often these IDs are artificial *additions* to the individuals, like driver licenses and barcodes. But sometimes such additions are impossible or inconvenient. Another possibility is to let IDs consist of a genuine subset of the features of the individual. One example is fingerprints and retinal scans used for people. A second one is data records in databases, which typically have their IDs being a single key feature of a record. When space is not an issue, or when the data is structured in advance to have short IDs, the use of unique IDs is easy and fully justified. But in situations where the set of individuals is unstructured, short IDs for every individual, distinguishing it from all others may not always be possible. In this paper we introduce the idea of using *partial IDs*, or PIDs, which are subsets of features per individual that may not distinguish it completely from all others, but in which some control is available about its set of imposters. In this work we had specific requirements (arising from analyzing our algorithms) from our set of PIDs, and we could satisfy them all with very short PIDs (even when no short IDs exist). In different settings one may need different properties, and we just suggest the possibility that PIDs may be of use elsewhere. A particular setting of large collections of unstructured (and noisy/lossy) set of sequences one may consider are those arising in computational biology, e.g. from some sequencing or phylogeny of large collections of large molecules like DNA and proteins. Is there any use of PIDs for the algorithmic problems such data gives rise to? A related work is Matsen et al. (2008) where certain small-dimensional witnesses are used to study phylogenetic mixtures.

Reconstruction Reconstruction problems appear in various settings abound, and we just mention a few. The famous graph reconstruction conjecture of Kelly and Ulam (see e.g. Nash-Williams 1978), now open for half a century, states that an unlabeled n-vertex graph is always uniquely defined by its (unordered multi-set of) n induced subgraphs obtained by removing each of the vertices. Another intriguing one is Mossell's subsequence problem (Mossell 2008) arising from computational biology (Holenstein et al. 2008), asking if an n-bit sequence is uniquely identified (with, say, a noticeable gap of $1/n$) by the probability measure on it subsequences, obtained by independently removing each symbol with probability $1/2$ (and removing the gaps!). Both problems are information theoretic about reconstruction of data from lossy models *different* than the ones considered here, which seem difficult even if we allow inefficient algorithms! A different reconstruction problem, in which efficiency is important, is Blum's junta learning problem (Blum 1994), which can be cast as a recovery from a noisy model different than ours.

1.7 Organization

The technical part of paper essentially has two parts. The first part, Sect. 2, contains the full discussion of PID graphs: their construction with good parameters, and a tradeoff showing near optimality of the parameters we achieve. The second part, Sect. 3, contains our recovery algorithms and their analysis. In Sect. 6, the final part of the paper, we explain the reduction from PRP to DRP.

2 PID graphs

In this section we study PID graphs. We first describe our main result, an efficient algorithm for constructing PID graphs of logarithmic width and depth for any set of vectors. As noted, these bounds only imply a quasi-polynomial upper bound on the cost. We next describe a set of vectors for which any logarithmic width PID graph has super-polynomial cost.

Our construction of an efficient PID graph is iterative: Start with a PID graph of logarithmic width in a similar fashion to as described in Sect. 1.4. Then, iteratively improve it. Each iteration consists of some local improvement based on the procedure "extend" below. Finally, when no more extensions are possible, the overall structure is (perhaps surprisingly) as desired.

We start with setting some useful notation. Fix a set $V \subseteq \Sigma^n$ of size $|V| = k$. We construct PIDs S_v for the vectors $v \in V$ by an iterative procedure, and so we redefine the set of imposters $I(v, S)$ with respect to any subset $S \subseteq [n]$ to explicitly depend on S. As before it is the set of all $u \in V$ (including v itself) which have the same pattern on S as v does, namely

$$I(v, S) = \{u \in V : u[S] = v[S]\}.$$

By convention, $v[\emptyset] = u[\emptyset]$ for all v, u.

2.1 Extension of a PID

A key procedure in the algorithm is a "greedy extension" of a PID: For $v \in V$ and $S \subseteq [n]$, we seek to successively add coordinates to S which shrink the number of imposters by a factor of 2, as long as we can. Formally, this procedure $\mathsf{Extend}(v, S)$ produces a superset of S and is defined recursively as follows (in extend, the set V is a fixed set of vector that the procedures knows).

***Extend* (*for a fixed V*)**

Input: A vector $v \in V$ and a set $S \subseteq [n]$.
Recursion base: Let J be the set of i in $[n] \backslash S$ so that

$$|I(v, S \cup \{i\})| \leq |I(v, S)|/2.$$

If $J = \emptyset$, output

$$\mathsf{Extend}(v, S) = S.$$

Recursive step: Otherwise, let[5] $i = \min J$, and compute

$$\mathsf{Extend}(v, S) = \mathsf{Extend}(v, S \cup \{i\}).$$

Call S *maximal* for v if $\mathsf{Extend}(v, S) = S$. A simple but crucial claim summarizes the properties of this procedure.

Claim 2.1 (Properties of extension) *For every $v \in V$ and $S \subseteq [n]$, if $T = \mathsf{Extend}(v, S)$ then*

- $S \subseteq T$,
- $|I(v, T)| \leq |I(v, S)| \cdot 2^{|S|-|T|}$,
- *T is maximal for v, and*

[5] Any choice of i in J will do. Minimum is chosen for concreteness.

- *If $u \neq v$ and $u \in I(v, T)$, then T is* not *maximal for u.*

Proof The first property follows as we just add coordinates. The second as we halve the size with each addition of i. The third by the halting definition. The fourth follows as T is maximal for v, and since any imposter $u \neq v$ has some coordinate $i \notin T$ for which $u_i \neq v_i$: $|I(u, T \cup \{i\})| \leq |I(v, T)| - |I(v, T \cup \{i\})| < |I(v, T)|/2$. □

Let us give some intuition for the algorithm constructing the PID graph before formally describing it. Our algorithm will perform many such Extend sub-routines. We initialize it by setting $S_v = \mathsf{Extend}[v, \emptyset]$ for all v. Then we repeatedly and judiciously choose a vector u and a subset S (not necessarily its PID), and assign the new S_u to be $\mathsf{Extend}(u, S)$. We insist that whenever we activate $\mathsf{Extend}(v, S)$, we have $|I(v, S)| \leq k2^{-|S|}$. This guarantees that each S_v during the running of the algorithm, and especially the final ones, has size at most $\log k$ and so the graph produced has small width.

To take care of the depth we ensure that eventually, for every v, *every* imposter $u \neq v$ has a PID S_u that is strictly larger than S_v. Thus along every directed path PID-size increases, and as it has maximum value $\log k$, no path can be longer than $\log k$. This condition also ensures that no cycles exist in the graph. The main question is how to ensure this "monotonicity" property (which certainly does not hold initially). The answer is simple: fix it whenever it is violated. Specifically, as long as some imposter $u \neq v$ in $I(v, S_v)$ with $|S_u| \leq |S_v|$ exists, replace S_u with $\mathsf{Extend}(u, S_v)$. The properties of extension in the claim above guarantee a new S_u which is longer than the current S_v (and also longer than the previous S_u).

These fixings continue as long as necessary, so all we need to argue is termination and efficiency. The overall PID-size strictly increases with every such step. Since their total size cannot exceed $k \log k$ by the width bound, we get a $k \log k$ upper bound on the number of invocations of Extend. The bound on the running time of the algorithm easily follows.[6]

We now formally define the algorithm and analyze it.

PID construction algorithm

Input: A set $V \subseteq \Sigma^n$.
Initialize: For every $v \in V$ set $S_v = \mathsf{Extend}(v, \emptyset)$.
Iterate: While there exists v and $u \neq v$ with $u \in I(v, S_v)$ and $|S_u| \leq |S_v|$ set

$$S_u = \mathsf{Extend}(u, S_v).$$

Output: The set of final PIDs S_v and their graph G.

Theorem 2.2 *The algorithm above terminates in at most $k \log k$ iterations, and produces a PID graph G with $\mathsf{width}(G) \leq \log k$ and $\mathsf{depth}(G) \leq \log k$ (and so also $\mathsf{cost}(G) \leq k^{\log k}$).*

The theorem follows from the following claim, that summarizes the invariants maintained by the algorithm, and conclusions from them. The invariants follow from the properties of Extend in Claim 2.1 above, and can be easily proved by induction on the iterations of the algorithm.

Claim 2.3 (Properties of the PID algorithm)

- *If $u \neq v$ in V were chosen in some iteration, then $|\mathsf{Extend}(u, S_v)| > |S_v| \geq |S_u|$.*
- *The total length of PIDs, $\sum_{v \in V} |S_v|$, strictly increases at every iteration.*

[6] An obvious implementation gives a running time of $(n + k)^2 \mathsf{poly} \log k$. It would be interesting to find a near-linear-time implementation.

- *For every $v \in V$ and every S_v that is obtained while the algorithm runs,* $1 \leq |I(v, S_v)| \leq 2^{-|S_v|} \cdot |V|$. *Specifically, the size of S_v never exceeds* $\log k$.
- *The total length of PIDs never exceeds* $k \log k$.
- *The algorithm halts after at most* $k \log k$ *iterations.*
- *Let G be the PID graph the algorithm computed. Then, along every path the size of the corresponding PIDs strictly increases.*

2.2 A costly set of vectors

We now describe, for every constant C, an example of a database $|V|$ of size k so that every PID graph for V of width at most $C \log k$ has cost at least $k^{\Omega(\log\log(k)/\log(2C))}$. This set of vectors is very simple, it consists of all n-bit vectors of (appropriately) bounded Hamming weight.

Proof of Proposition 1.6 First, fix some parameters: Let $C \geq 1$, m be a large integer, and $n = \lceil Dm \rceil$ for $D = 20C^2$. The set V consists of all vector in $\{0, 1\}^n$ of (Hamming) weight at most m. The size of V is therefore

$$|V| = k = \sum_{j \leq m} \binom{n}{j} \leq (2eD)^m \leq D^{2m}.$$

Let G be a PID graph for V of width at most $C \log k \leq (2C \log D)m \leq n/2 - m$.

For any choice of PIDs, for every $\ell < m$, and for every vector v of weight ℓ, the number of imposters at level $\ell + 1$ of v is at least $n - \ell - C \log k \geq n/2$ (at worst S_v is disjoint from the 1's of v). The number of paths starting at the all zero vector is, therefore, at least $(n/2)^m$. So, since $m \geq \Omega(\log(k)/\log(2C))$,

$$\mathsf{cost}(G) \geq (n/2)^m \geq 2^{\Omega(m \log n)} \geq k^{\Omega(\log\log(k)/\log(2C))}.$$

□

3 Distribution recovery algorithms

In this section we describe two distribution recovery algorithms, for lossy and for noisy samples (and also remark on recovery from the hybrid model combining loss and noise). That is, we prove Theorem 1.4. In both cases, there is some known database V and we wish to estimate a probability distribution π on it, using random imperfect samples. We assume that the algorithms get a PID graph G of V, and specifically a PID S_v for all v in V.

The two algorithms for the two cases share a two-phase structure: an estimation phase and an aggregation phase. In the estimation phase, the PID graph is used to extract a list of quantities that are useful to estimate, and these estimations are obtained. In the aggregation phase, the estimations obtained are numerically combined into an estimation of π.

We now discuss the two phases in more detail. Both use the PID graph G, and specifically will estimate π by first estimating an auxillary vector p, which for every v in V is defined by

$$p(v) = \sum_{u \in I(v)} \pi(u),$$

where $I(v)$ is the set of imposters of u in G. We shall see how all parameters of the PID graph, width, depth and cost affect the performance of these two phases.

In the *estimation phase*, the algorithm obtains an (L_∞) estimate p' of p. As we shall see, these estimates amount to solving k distribution recovery problems in small dimension (at most $\mathsf{width}(G)$), by projecting the data to the coordinates of each of the PIDs S_v. As $\mathsf{width}(G) \leq \log k$, these recovery algorithms can afford to use time exponential in the dimension! Computing p' from lossy samples in this regime is simple, and uses only the well-known Hoeffding bound. Computing p' from noisy samples, even in this regime, is more elaborate (and, in fact, a-priori may be impossible). In both cases, a simple parallel implementation of the algorithm is possible. In both cases, the quality of the approximation p' of p needed (which determines running time and number of samples) follows from the discussion of the aggregation phase below.

In the aggregation phase, the estimate p' of p, is used to compute an estimate π' of π. Clearly, the vector p is obtained from π by a linear transformation: there exists a $k \times k$ real matrix $M = M(G)$ so that

$$p = M\pi.$$

Since G is acyclic, the matrix M is triangular with 1's on the diagonal: π can be computed from p by back substitution using induction on the structure of G,

$$\pi(v) = p(v) - \sum_{u \in I(v): u \neq v} \pi(u). \tag{3.1}$$

This process is easily parallelized, and can be seen to take exactly $\mathsf{depth}(G)$ rounds. The main issue is, of course, that we do not have p available but rather an approximation p' of it. A priori such a matrix M may have exponentially small condition number (in k), and thus may require exponentially good approximation p' of p, which will, in turn, cause the first phase to require exponentially many samples. We show, however, that the error aggregation depends only on $\mathsf{cost}(G)$. Lemma 3.1 below states that using p' instead of p in (3.1) above yields an estimate π' satisfying

$$\left\|\pi - \pi'\right\|_\infty \leq \mathsf{cost}(G) \left\|p - p'\right\|_\infty. \tag{3.2}$$

Specifically, we only need to estimate p up to an additive factor $\alpha/\mathsf{cost}(G)$ in order to get an α approximation of π. When G has cost $k^{\log k}$, as we achieve, such estimates can be obtained in quasi-polynomial running time (instead of exponential).

Summarizing, the high-level description of the algorithms is:

Recovery algorithm

Input: A database $V \subseteq \Sigma^n$ and a PID graph G for V.
Estimation: Obtain a $(\alpha/\mathsf{cost}(G))$-estimate p' of p.
Aggregation: Compute $\pi' = M^{-1}p'$, using (3.1) according to the topological order on G.
Output: The estimate π' of a distribution π on V.

We note that $\mathsf{cost}(G)$ can be easily computed, and in many cases it can be much smaller than the worst-case super-polynomial one.

We now formally discuss the two phases. In the estimation phase (which is the only phase that depends on the underlying sampler), we consider each sampler model separately. The aggregation phase is the same for both.

3.1 The estimation phase

We now show how both algorithms compute an estimate p' of p: for every v in V, it will hold that

$$\Pr\left[|p(v) - p'(v)| \geq \varepsilon\right] \leq \delta,$$

with

$$\varepsilon = \alpha/\mathsf{cost}(G)$$

(a simple union bound implies that this estimate jointly holds for all v, with high probability).

In Sect. 1.2 of the introduction we intuitively explained how each of the lossy and noisy cases are handled when n is small. In this section we unify the two, and actually define and analyze a much more general sampler that extends both. This generality also helps to understand the essence of this phase of the algorithm.

A general sampler We start by describing a general sampling procedure over Σ_2^n, given a distribution π on Σ_1^n. As we shall explain, the lossy and noisy cases are specific instances of it. The sampler is determined by n stochastic (i.e., with non negative entries and columns that sum up to 1) *transition* matrices $T_1, \ldots, T_n$, each in $\mathbb{R}^{\Sigma_2 \times \Sigma_1}$. These n matrices and π define a distribution on samples v' in Σ_2^n. First, choose v according to π. Then, for each i in $[n]$, if $v_i = \sigma_1$ in Σ_1, then v'_i is random: it is σ_2 in Σ_2 with probability $(T_i)_{\sigma_2,\sigma_1}$.

Let us consider the lossy and noisy cases. In both cases, all transition matrices $T_1, \ldots, T_n$ are the same. In the lossy case, $\Sigma_1 = \{0, 1\}$ and $\Sigma_2 = \{0, 1, ?\}$, and the transition matrix T_{loss} is

$$(T_{loss})_{\sigma_2,\sigma_1} = \begin{cases} \mu & \sigma_2 = \sigma_1, \\ 1-\mu & \sigma_2 =?, \\ 0 & \text{otherwise.} \end{cases}$$

In the noisy case, $\Sigma_1 = \Sigma_2 = \{0, 1\}$ and the transition matrix in every coordinate is

$$T_{noise} = \begin{bmatrix} T_{0,0} & T_{0,1} \\ T_{1,0} & T_{1,1} \end{bmatrix} = \begin{bmatrix} (1+\nu)/2 & (1-\nu)/2 \\ (1-\nu)/2 & (1+\nu)/2 \end{bmatrix}. \tag{3.3}$$

A simple generalization yields, e.g., a noise model where each coordinate i in $[n]$ is perturbed with a different probability $(1 - \nu_i)/2$. As we shall see, our algorithm works in this case as well.

Minimal singular value Clearly, not all samplers allow recovery of π. The crucial property of $T_1, \ldots, T_n$ that allows recovery is the following. For each transition matrix T, let $\mathsf{ms}(T)$ be the minimal singular value of T, that is, the minimal square-root of an eigenvalue of the symmetric matrix $T^t T$, with T^t the transposition of T. For example, in the lossy case, $T_{loss}^t T_{loss} = \mu^2 I + (1-\mu)^2 J$ with J the all-one matrix, so

$$\mathsf{ms}(T_{loss}) = \mu.$$

In the noisy case, T_{noise} is symmetric, and

$$\mathsf{ms}(T_{noise}) = \min\{1, \nu\} = \nu.$$

Define

$$\mathsf{ms}(T_1, \ldots, T_n) = \min\{\mathsf{ms}(T_i) : i \in [n]\}.$$

Recovery is possible only when $\mathsf{ms}(T_1, \ldots, T_n) > 0$. Indeed, let p be say the uniform distribution on Σ_1, and assume that the right kernel of T is not $\{0\}$. There is therefore a vector $z \neq 0$ so that $Tz = 0$ so that $\|z\|_\infty < 1/|\Sigma_1|$. Hence,

$$0 = \sum_{\sigma_2} (Tz)(\sigma_2) = \sum_{\sigma_2} \sum_{\sigma_1} T_{\sigma_2,\sigma_1} z(\sigma_1) = \sum_{\sigma_1} z(\sigma_1) \sum_{\sigma_2} T_{\sigma_2,\sigma_1} = \sum_{\sigma_1} z(\sigma_1).$$

Thus $p' = p + z$ is a distribution on Σ_1 that satisfies $Tp = Tp'$. This means that samples generated by p are distributed exactly like samples from p', which means that p, p' can not be distinguished. From now on we thus assume $\mathsf{ms}(T_1, \ldots, T_n) > 0$. For both the lossy model and the noisy model, the corresponding ms is a constant bounded away from zero.

The estimation algorithm and its analysis We now move to describe our algorithm for estimating p using a general sampler defined by $T_1, \ldots, T_n$. Start by obtaining a set V' of size

$$t = \mathsf{poly}\left(\mathsf{cost}(G), 1/\alpha, \log(1/\delta), (|\Sigma_1||\Sigma_2|/\mathsf{ms}(T))^{\mathsf{width}(G)}\right)$$

of random samples $V' = \{v'_1, \ldots, v'_t\}$. The exact value of t will be determined later on. For each $v \in V$, use the PID S_v to perform dimension reduction on this sample to estimate $p(v)$. These computations can all be done in parallel. Fix v in V, and denote $S = S_v$. We describe how to estimate $p(v)$.

The transition matrix of vectors on the coordinates S is the tensor product $T = T^{(S)} = \otimes_{i \in S} T_i$. That is, for v_1 in Σ_1^S and v_2 in Σ_2^S, the number T_{v_1,v_2} is the probability of seeing v_2 given that v_1 was chosen (using π). Observe

$$\mathsf{ms}(T) \geq \mathsf{ms}(T_1, \ldots, T_n)^s \geq \mathsf{ms}(T_1, \ldots, T_n)^{\mathsf{width}(G)}.$$

For every v_2 in Σ_2^S, let $q(v_2)$ be the probability that v_2 is the output of the sampler (when restricted to entries in S). Since $\mathsf{ms}(T_1, \ldots, T_n) > 0$, there is a unique r so that

$$q = Tr.$$

What is r? The number $r(v_1)$, for every v_1 in Σ_1^S, is the probability that a vector v that is chosen according to π is so that $v[S] = v_1$. Specifically,

$$p(v) = r(v[S]).$$

In other words, we just wish to estimate one entry in r.

The vector q can be easily approximated: for each v_2 in Σ_2^S, set

$$q'(v_2) = \frac{|\{v' \in V' : v'[S] = v_2\}|}{|V'|},$$

the fraction among the given samples of the pattern v_2. The vector q' is a good estimate of its expectation $q = \mathbb{E}q'$. Let r' be the unique vector so that

$$q' = Tr'$$

(it can be efficiently computed using q', T, in time polynomial in t). Set

$$p'(v) = r'(v[S]).$$

We are almost ready to conclude. Using Hoeffding's bound, for t as claimed,

$$\Pr\left[\left\|q - q'\right\|_2 \geq \varepsilon \cdot \mathsf{ms}(T)\right] \leq \delta.$$

Since $q - q' = T(r - r')$, if $\|q - q'\|_2 \le \varepsilon \cdot \mathsf{ms}(T)$, then $\|r - r'\|_2 \le \varepsilon$. So, overall

$$\Pr\left[|p(v) - p'(v)| \ge \varepsilon\right] \le \Pr\left[\|r - r'\|_2 \ge \varepsilon\right] \le \delta,$$

as required.

The following summarizes this part of the algorithm.

Estimation phase

Input: A vector v with a PID $S = S_v \subseteq [n]$, a transition matrix $T = T^{(S)}$ with $\mathsf{ms}(T) > 0$, and samples generated by π and T.
Statistics: Obtain an $(\varepsilon \cdot \mathsf{ms}(T))$-estimate q' of q.
Linear algebra: Compute r', the unique vector so that $q' = Tr'$.
Output: The estimate $p'(v) = r'(v[S])$.

Comments A (somewhat) more efficient procedure that is based on Hamming distances, in the spirit of Dvir et al. (2012), is also applicable in this case.

The same algorithm works also when the sampling procedure is only $\mathsf{width}(G)$-wise independent, since all calculations are based on local data (concerning sets of entries of size at most $\mathsf{width}(G)$).

3.2 The aggregation phase: error bounds

We now explain how to bound the accumulation of error. Let M be the matrix defined by the PID graph G for V as above. That is, $p = M\pi$ means

$$p(v) = \sum_{u \in I(v)} \pi(u).$$

We prove the following estimate on the increase of error M^{-1} can cause.

Lemma 3.1 *For any vector w,*

$$\left\|M^{-1}w\right\|_\infty \le \mathsf{cost}(G)\,\|w\|_\infty .$$

The lemma clearly implies (3.2).

Proof Denote $w' = M^{-1}w$. We in fact prove that for every $v \in V$,

$$\left|w'_v\right| \le \mathsf{cost}(v)\,\|w\|_\infty .$$

This follows by induction on the structure of G. When v is a sink in G, we have $w'_v = w_v$, so the claim holds. Otherwise, since $w = M^{-1}w'$, we have $w_v = \sum_{u \in I(v)} w'_u$. So, by induction,

$$\left|w'_v\right| = \left|w_v - \sum_{u \in I(v): u \ne v} w'_u\right| \le \|w\|_\infty \left(1 + \sum_{u \in I(v): u \ne v} \mathsf{cost}(u)\right) = \mathsf{cost}(v)\,\|w\|_\infty .$$

□

3.3 Unknown noise parameter

So far, we have always assumed knowledge of the noise and erasure parameters. For lossy samples, this assumption can be easily removed by estimating the probability in which question marks appear. In the noisy case, it is less obvious how to remove this assumption. We now explain how to accomplish that.

The key observation is that any noise operation makes a sparse distribution (i.e. with support-size at most k) highly non-sparse (i.e. full support). This distinction is, of course, only possible due to assumption $n > \log k$. Our algorithm can be thought of as "de-noising" a distribution. So, going over all noise options, and de-noising according to all of them, the only sparse distribution should be obtained for the true noise parameter. This is indeed what happens.

Consider a population $\overline{V} \subset \{0,1\}^n$ of size $|\overline{V}| = k$, and a distribution $\overline{p}$ on $\overline{V}$. Let

$$m = \lfloor \log k \rfloor + 1$$

and let $V \subset \{0,1\}^m$ be the projection of $\overline{V}$ to the first m coordinates, and let p be the distribution on $\{0,1\}^m$ that is the projection of $\overline{p}$ to the first m coordinates.

Our goal is to estimate the noise parameter ν given random noisy samples. That is, given an accuracy parameter $\beta > 0$, compute, w.h.p., a number ν' so that $|\nu' - \nu| \leq \beta$. The running time we shall achieve is $(C/(\nu\beta))^m$, for $C > 0$ a large constant. As we explain at the end of this subsection, to use ν' in the recovery algorithm instead of ν, we need β to be smaller than $\alpha \cdot \mathsf{poly}_\nu(k^{-\log k})$, establishing the bound in the second part of Theorem 1.2.

We describe an algorithm that uses a lower bound on ν. Namely, our algorithm will get as input a number ν_0 so that $\nu \geq \nu_0$. This assumption can be easily removed (without increasing the running time by much) by trying values for ν_0 in $\{1, 1/2, 1/4, \ldots\}$ until finding one that works.

For $\nu \geq 0$, let $T(\nu)$ be the noise operator with parameter ν on $\{0,1\}^m$, as defined in Sect. 3.1 (the tensor power of the matrix defined in (3.3)). That is, we get access to samples distributed according to q, for

$$q = T(\nu)p.$$

Finding noise

Estimation: Set

$$\varepsilon = (c\nu_0\beta)^m,$$

for a fixed small constant $c > 0$ so that the following holds. Obtain (in time polynomial in $1/\varepsilon$, with high probability, by collecting samples, as in Sect. 3.1) an estimate q' of q so that

$$q' = q + e$$

satisfies

$$\|e\|_\infty \leq \varepsilon^2.$$

Enumerating options: Consider the following optional noise parameters: for every $0 \leq \ell \leq O(\log(1/\nu_0)/\varepsilon^2)$, set

$$\nu_\ell = \nu_0(1+\varepsilon^2)^\ell \leq 1$$

and

$$p_\ell = T(1/\nu_\ell)q'.$$

Let Λ be the set of ℓ so that p_ℓ is close to a probability distribution, that is, there is a distribution r_ℓ so that $\|p_\ell - r_\ell\|_\infty \leq \varepsilon$.

Making a decision: Define the *sparsity* of p_ℓ as the number of $v \in \{0,1\}^m$ so that $p_\ell(v) > \varepsilon$. Choose ℓ' in Λ so that $p_{\ell'}$ has minimal sparsity among all ℓ in Λ, and output

$$\nu' = \nu_{\ell'}.$$

(If Λ is empty, abort.)

It remains to prove correctness: w.h.p.[7], the running time is bounded as claimed, and when the algorithm terminates, the set Λ is not empty, and

$$\nu/\nu' \in [1-\beta, 1+\beta].$$

We shall use the following simple properties of the noise operator.

Lemma 3.2 *1. The noise operator yields a group representation: $T(\nu)T(\nu') = T(\nu\nu')$ for every $\nu, \nu' > 0$, and $T(1)$ is identity.*
2. Let $0 \le \gamma \le 1$. The minimal entry in the matrix $T(1-\gamma)$ is $(\gamma/2)^m$. The maximal entry in the matrix $T(1-\gamma)$ is 1.
3. Let $\nu > 1$. The maximal entry in the matrix $T(\nu)$ is ν^m.

We first show that a good estimate for ν yields a sparse distribution.

Claim 3.3 *Let $\tilde{\nu}$ be the smallest noise in $\{\nu_\ell\}$ that is at least ν. Let $\tilde{\ell}$ be so that $\tilde{\nu} = \nu_{\tilde{\ell}}$ and let $\tilde{p} = p_{\tilde{\ell}}$. Then, w.h.p., $\|\tilde{p} - p\|_\infty \le \varepsilon$. Specifically, $\tilde{\ell}$ is in Λ and the sparsity of $\tilde{p}$ is at most k.*

Proof Observe

$$\tilde{p} = p + (T(\nu/\tilde{\nu}) - I)p + T(1/\tilde{\nu})e, \tag{3.4}$$

where I is the identity matrix. By Lemma 3.2,

$$\|T(1/\tilde{\nu})e\|_\infty \le (1/\nu_0)^m \varepsilon^2 \le \varepsilon/2.$$

By choice of $\tilde{\nu}$,

$$\|T(\nu/\tilde{\nu}) - I\|_\infty \le \left|\frac{1-(1-\varepsilon^2)}{2}\right| \le \varepsilon/2.$$

□

We now show that noise parameters that are far from the true ν yield non-sparse distributions.

Claim 3.4 *Let $\tilde{\nu}, \tilde{\ell}, \tilde{p}$ be as above. Let $\hat{\ell}$ in Λ be so that $\hat{\nu} = \nu_{\hat{\ell}}$ admits*

$$\tilde{\nu}/\hat{\nu} \notin [1-\beta/2, 1+\beta/2].$$

Let $\hat{p} = p_{\hat{\ell}}$. Then, w.h.p., the sparsity of $\hat{p}$ is larger than k.

Proof By definition $T(\tilde{\nu})\tilde{p} = T(\hat{\nu})\hat{p}$. There are two cases to consider. The first is that $\tilde{\nu} > \hat{\nu}$. In this case, all entries of $T(\hat{\nu}/\tilde{\nu})$ are at least $(\beta/4)^m$ and at most 1. Since $\hat{\ell} \in \Lambda$, all entries of $\tilde{p} = T(\hat{\nu}/\tilde{\nu})\hat{p}$ are at least $(\beta/4)^m - 2^m\varepsilon \ge \varepsilon$. But this is false due to Claim 3.3. So this case cannot happen. The second case is $\hat{\nu} > \tilde{\nu}$. Since $\tilde{\ell} \in \Lambda$, similarly to the first case, the sparsity of $\hat{p}$ is full. □

[7] That is, when the assumed bound on the norm of e holds. The same holds for "w.h.p." below.

So, since ℓ' was chosen to have minimal sparsity,

$$\tilde{\nu}/\nu' \in [1-\beta/2, 1+\beta/2].$$

Overall,

$$\nu/\nu' = (\nu/\tilde{\nu}) \cdot (\tilde{\nu}/\nu') \in [1-\beta, 1+\beta].$$

We now explain why it suffices for the population recovery algorithm (Sect. 4 below) to make sure that β is smaller than $\alpha \cdot \mathsf{poly}_{\nu}(k^{-\log k})$. The basic procedure in the algorithm is described in Sect. 3.1 above, and can be summarized as follows. There is a distribution π on V that we wish to learn, and our access to π is say via noisy samples. The algorithm only attempts to learn the projection π_S of π to some set $S \subset [n]$ of coordinates of size $|S| = m \leq 1 + \log k$. Let T be the real noise operator $T = (T(\nu))^{\otimes m}$ which we do not know, and let T' be the known noise operator $T' = (T(\nu'))^{\otimes m}$. The algorithm computes a good approximation q' of $q = T\pi_S$. It then approximates π_S by $r' = (T(1/\nu'))^{\otimes m} q'$ instead of $r = (T(1/\nu))^{\otimes m} q'$. This estimate is pretty accurate (using Lemma 3.2):

$$\begin{aligned}\|r - r'\|_\infty &= \|\big((T(\nu/\nu'))^{\otimes m} - I\big)(T(1/\nu))^{\otimes m})q'\|_\infty \\ &\leq ((1+\beta)^m - 1)(1/\nu)^m \leq \alpha \cdot \mathsf{poly}(k^{-\log k}).\end{aligned}$$

This implies that when using ν' (instead of ν) in the population recovery problem, the resulting calculations may be at most $\alpha \cdot \mathsf{poly}(k^{-\log k})$ far from the value when computing with ν, which suffices for the algorithm to perform well.

4 A population recovery algorithm

Here we prove Theorem 1.2, that is, we describe an algorithm that solves the PRP problem (without knowing V).

For a population $V \subset \{0,1\}^n$ of size $|V| = k$ and for an integer $t > 0$, define $SID_t(V)$ as the set of vectors that have short IDs, that is, the set of v in V so that there is $S_v \subset [n]$ of size $|S_v| \leq t$ so that for all $v' \neq v$ in V we have $v'[S] \neq v[S]$. As noted above, the set $SID_{\log k}(V)$ is never empty. We also use the following notation: Given a family $P_0, P_1, \ldots, P_d$ of subsets of V, denote $P_{\leq d} = \bigcup_{i=0}^{d} P_i$ and $P_{<d} = \bigcup_{i=0}^{d-1} P_i$.

We define an optimal partition P of V. It is the partition of V to $d+1$ sets $P_0, P_1, \ldots, P_d$ defined inductively by $P_0 = SID_{\log k}(V)$ and for $i > 0$, if $P_{<i} = V$ then $d = i-1$ and if $P_{<i} \neq V$ then

$$P_i = SID_{\log k}(V \backslash P_{<i}).$$

The integer d is called the *depth* of the partition. For v in V, the *depth* of v in P is the integer $d_v \in \{0, 1, \ldots, d\}$ so that $v \in P_{d_v}$.

Theorem 1.5 implies that P is shallow.

Lemma 4.1 *The depth of the optimal partition P is at most* $\log k$.

Proof Let G' be the PID graph constructed in Theorem 1.5 with sets S'_v for v in V. The observation is that for every v in V, the depth of v in P is at most the depth of v in G'. This can be proved by induction on the depth in G'. □

4.1 The algorithm

The algorithm does not know the population V and so does not know P as well. It, however, trusts Lemma 4.1 and stops after $\log k$ iterations. Here is a schematic outline of algorithm, which is meant to provide a rough outline of the structure of the algorithm. Some details are therefore missing from it, and they are discussed at length in the remainder of this section.

Population recovery (schematic description)

Input: A list of noisy samples and parameters[8] k, α, δ and ν.

Initialize: Set $V_0 = \emptyset$.

Iterate: For $d = 0, 1, 2, \ldots, \log k$, do the following.

Go over all $S \subset [n]$ of size $|S| = \log k$ and all $u \in \{0,1\}^{\log k}$.

(*) If there is a unique $v \in V \backslash V_{<d}$ so that $v[S] = u$ then estimate $\pi(v)$ and add v to V_d.

If at some point $|V_{\leq d}| > k$, abort.

The key computational step is (*). It may seem that it requires knowing V, but as we shall see we can accomplish (*) without knowing V.

We now describe the algorithm in more detail. The algorithm provides the following features. Its numerical error is small, i.e., the accuracy parameter is at most $\alpha \cdot k^{O(\log k)}$. The probability of error is at most $(nk)^{O(\log k)}\delta$ and the running time is at most polynomial in $(nk)^{\log k}\log(1/\delta)/\alpha$. Dividing α, δ by the appropriate factor yields the promised guarantees.

In the following, S is a subset of $[n]$ of size $|S| = \log k$, u is a vector in $\{0,1\}^{\log k}$, i is an element of $[n]\backslash S$ and $b \in \{0,1\}$. Denote

$$V_{S,u} = \{v \in V : v[S] = u\}$$

and

$$V_{S,u,i,b} = \{v \in V : v[S] = u, v_i = b\}.$$

The case $\mathbf{d = 0}$. Here we explain how to find $SIDI_{\log k}(V)$. The intuition is that if $v \in SIDI_{\log k}(V)$ then there is a set S_v of size at most $\log k$ so that for every $i \notin S_v$, there is a unique extension in V of $v[S_v]$ in coordinate i, that is, $V_{S_v,v[S_v],i,v_i} = \{v\}$ and $V_{S_v,v[S_v],i,1-v_i} = \emptyset$.

Go over all S, u and perform the following procedure, which returns the unique vector v in V so that $v[S] = u$ if such exists:

***Unique recovery for* S, u**

Compute: Obtain, with probability at least $1 - \delta$, an $(\alpha/4)$-estimate $p'_{S,u}$ of $\pi(V_{S,u})$ using samples, as described in Sect. 3.1.

If $p'_{S,u} \leq \alpha$, then abort for these S, u.

Otherwise, obtain, with probability at least $1 - \delta$, an $(\alpha/(4k))$-estimate $p'_{S,u,i,b}$ for $\pi(V_{S,u,i,b})$ for every $i \notin S$ and $b \in \{0,1\}$.

If both $p'_{S,u,i,1}$ and $p'_{S,u,i,0}$ are at least $\alpha/(4k)$ then abort for these S, u.

Otherwise, for every $i \notin S$ there is a unique $b_i \in \{0,1\}$ so that $p'_{S,u,i,b_i} \geq \alpha/2$.

Update population: Let v be the vector defined as $v[S] = u$ and $v_i = b_i$ for all $i \notin S$.

Add v to V_0, set $S_v = S$ and $\pi'(v) = p'_{S,u}$.

The case $\mathbf{d > 0}$. Here we aim at finding $SIDI_{\log k}(V \backslash V_{<d})$, using the approximate probabilities we already obtained for $V_{<d}$. Since the approximations in step d depend on steps

[8] The assumption that ν is given as input can be removed: Instead of ν we may compute an estimate ν' of ν, and use ν' in the computations that follow without harming the accuracy too much, as described in Sect. 3.3.

$0, \ldots, d-1$, the accuracy of the computation deteriorates exponentially with d. By a union bound on all previous errors, the accuracy guaranteed is roughly at most k^d times the individual accuracy α. Since the depth is only logarithmic, the overall accuracy is as claimed.

Go over all S, u and perform the following procedure:

***Unique recovery for* S, u *with corrections using* $V_{<d}$**

Compute: Obtain an α-estimate $q_{S,u}$ of $\pi(V_{S,u})$ using noisy samples, with probability at least $1-\delta$.
Compute

$$p'_{S,u} = q_{S,u} - \sum_{v' \in V_{S,u} \cap V_{<d}} \pi'(v').$$

(UR.i.) If $p'_{S,u} \leq (4k)^{2d}\alpha$, then abort for these S, u.
Otherwise, for every $i \notin S$ and $b \in \{0,1\}$, obtain, with probability at least $1-\delta$, an α-estimate $q_{S,u,i,b}$ for $\pi(V_{S,u,i,b})$.
Compute

$$p'_{S,u,i,b} = q_{S,u,i,b} - \sum_{v' \in V_{S,u,i,b} \cap V_{<d}} \pi'(v').$$

(UR.ii.) If both $p'_{S,u,i,1}$ and $p'_{S,u,i,0}$ are at least $(4k)^{2d-1}\alpha$ then abort for these S, u.
Otherwise, for every $i \notin S$ there is a unique $b_i \in \{0,1\}$ so that $p'_{S,u,i,b_i} \geq (4k)^{2d-1}\alpha$.

Update population: Let v be the vector defined as $v[S] = u$ and $v_i = b_i$ for all $i \notin S$.
Add v to V_d, set $S_v = S$ and $\pi'(v) = p'_{S,u}$.

4.2 Analysis

Below, we write w.h.p. which stands for "with high probability" instead of writing "with probability at least $1-(kn)^{O(\log k)}\delta$." The event that holds w.h.p. is that all approximations the algorithm made are actually correct. The number of approximations made is at most $2n(1+\log k)n^{\log k}$, and each fails with probability at most δ.

Lemma 4.2 *W.h.p., the following properties hold for every* $d \in \{0, 1, \ldots, \log k\}$*:*

1. *If v is in the set V_d found by the algorithm then v is in the original set V.*
2. *For every v in V_d,*

$$|\pi'(v) - \pi(v)| \leq (4k)^{2d}\alpha.$$

3. *For every v in V_d and for every $v' \neq v$ in $I(v, S_v)\backslash V_{<d}$ where I is the imposter-set defined in Sect. 2,*

$$\pi(v') \leq (4k)^{2d}\alpha/(2k).$$

4. *If v is in V and $\pi(v) \geq 2(4k)^{2d_v}\alpha$ then v is in $V_{\leq d_v}$ where d_v is the depth of v in the optimal partition P.*

This lemma implies Theorem 1.2 since we found a set $V' = \bigcup_{d \leq \log k} V_d$ so that by item 4 if $v \in V$ is so that $\pi(v)$ is large then $v \in V'$ and further by item 2 we have a good estimate $\pi'(v)$ of $\pi(v)$.

Proof The proof is by induction on d. We omit the proof of the induction base since its proof is basically the same as that of the induction step.

1. Assume that v is in V_d and that it was generated from S, u. Thus, $p'_{S,u} > (4k)^{2d}\alpha$, and for every $i \notin S$, we have $p'_{S,u,i,1-v_i} < (4k)^{2d-1}\alpha$. A useful observation is that, w.h.p.,

$$|p'_{S,u} - \pi(V_{S,u}\backslash V_{<d})| \le |q_{S,u} - \pi(V_{S,u})| + \sum_{v' \in V_{S,u} \cap V_{<d}} \left|\pi'(v') - \pi(v')\right|$$
$$\le \alpha + k(4k)^{2d-2}\alpha \le (4k)^{2d-1}\alpha. \tag{4.1}$$

Therefore, w.h.p.,

$$\pi(V_{S,u}\backslash V_{<d}) \ge (4k)^{2d}\alpha - (4k)^{2d-1}\alpha > (4k)^{2d}\alpha/2.$$

So there is v' in $V_{S,u}\backslash V_{<d}$ so that

$$\pi(v') > (4k)^{2d}\alpha/(2k).$$

Similarly to (4.1), for $i \notin S$ and $b \in \{0, 1\}$,

$$|p'_{S,u,i,b} - \pi(V_{S,u,i,b}\backslash V_{<d})| \le \alpha + k(4k)^{2d-2}\alpha$$

which implies

$$\pi(V_{S,u,i,1-v_i}\backslash V_{<d}) < (4k)^{2d-1}\alpha + \alpha + k(4k)^{2d-2}\alpha < (4k)^{2d}\alpha/(2k).$$

For every $i \notin S$ we have

$$\pi(V_{S,u,i,v'_i}\backslash V_{<d}) \ge \pi(v').$$

So $v[S] = v'[S]$ and for all $i \notin S$ we have $1 - v_i \ne v'_i$ or $v_i = v'_i$. Thus $v = v'$ is in V.

2. By the induction hypothesis (items 1. and 3.) and (4.1),

$$|\pi'(v) - \pi(v)| \le |p'_{S,u} - \pi(V_{S,u}\backslash V_{<d})| + |\pi(V_{S,u}\backslash V_{<d}) - \pi(v)|$$
$$\le (4k)^{2d-1}\alpha + k(4k)^{2d}\alpha/(2k) \le (4k)^{2d}\alpha.$$

3. Let v in V_d, let $S = S_v$, and let $u = v[S]$. Thus, $I(v, S_v) = V_{S,u}$. By choice,

$$p'_{S,u,i,v_i} \ge (4k)^{2d-1}\alpha.$$

Assume towards a contradiction that there is $v' \ne v$ in $V_{S,u}\backslash V_{<d}$ so that

$$\pi(v') > (4k)^{2d}\alpha/(2k).$$

There is $i \notin S$ so that $v'_i = 1 - v_i$. Thus,

$$\pi(V_{S,u,i,1-v_i}\backslash V_{<d}) \ge \pi(v') > (4k)^{2d}\alpha/(2k).$$

So, w.h.p., as in (4.1),

$$p'_{S,u,i,1-v_i} > (4k)^{2d}\alpha/(2k) - \alpha - k(4k)^{2d-2}\alpha > (4k)^{2d}\alpha/(4k).$$

This means that the algorithm actually aborted (inspection UR.ii. failed), which is a contradiction.

4. Assume v is in V, has depth $d = d_v$ in the optimal P and $\pi(v) \ge 2(4k)^{2d}\alpha$. If v is in $V_{<d}$ then we are done. There are therefore S, u so that $v[S] = u$ and every $v' \ne v$ in $I(v, S)$ has depth $d_{v'} \le d - 1$ in P. Partition $I(v, S)\backslash\{v\}$ to two sets: I_+ the set of v' so that $\pi(v') \ge 2(4k)^{2d_{v'}}\alpha$ and $I_- = I(v, S)\backslash(\{v\} \cup I_+)$. By induction, every v' in I_+ is in $V_{<d}$. Since

$$\pi(V_{S,u}\backslash V_{<d}) \ge \pi(v) \ge 2(4k)^{2d}\alpha,$$

using (4.1),

$$p'_{S,u} \geq (4k)^{2d}\alpha.$$

This means that S, u passed the first inspection (UR.i.). For every $i \notin S$, the set $V_{S,u,i,1-v_i} \setminus V_{<d}$ consists of element of I_- only, which means

$$\pi(V_{S,u,i,1-v_i} \setminus V_{<d}) \leq k2(4k)^{2d-2}\alpha.$$

So, w.h.p.,

$$\begin{aligned} p'_{S,u,i,1-v_i} &= q_{S,u,i,1-v_i} - \sum_{v' \in V_{S,u,i,1-v_i} \cap V_{<d}} \pi'(v') \\ &\leq \left| q_{S,u,i,1-v_i} - \pi(V_{S,u,i,1-v_i}) \right| \\ &\quad + \left| \pi(V_{S,u,i,1-v_i} \cap V_{<d}) - \sum_{v' \in V_{S,u,i,1-v_i} \cap V_{<d}} \pi'(v') \right| + \pi(V_{S,u,i,1-v_i} \setminus V_{<d}) \\ &\leq \alpha + k(4k)^{2d-2}\alpha + 2k(4k)^{2d-2}\alpha < (4k)^{2d-1}\alpha. \end{aligned}$$

This means that S, u passed the second inspection (UR.ii.) and that v is indeed in $V_{\leq d}$.

□

5 Future work

This work introduced algorithms for certain recovery problems of databases from highly noisy samples. Our algorithms run in quasi-polynomial time (in the worst case: the running time depends on the underlying database and in many cases is far more efficient). We do not know if polynomial time algorithms for these problems exist, and for which noise parameters. The first question towards understanding the complexity of this recovery problem is the information theoretic one: Does a polynomial number of samples suffice to accurately estimate the underlying distribution in the worst case, even if one does not care about computational complexity. If positive, this answer may then lead to a polynomial time algorithm for every database.

Key ingredients in our solution are PIDs and PID graphs. Specifically, we give an efficient construction of a PID graph of width and depth $\log k$ for any database of size k. The cost of the constructed PID graph is thus at most $k^{\log k}$, which basically implies the quasi-polynomial time upper bound on the running time of the recovery algorithms. We also prove that the cost of a PID graph of width $\log k$ cannot, in general, be smaller than $k^{\log\log k}$. Understanding the smallest cost of a logarithmic-width PID graph remains an open question. The super-polynomial lower bound also suggests that a faster algorithm will necessarily need to go beyond PIDs.

The type of recovery problems we consider here are, in a sense, the simplest non-trivial ones. There are several natural and more general recovery problems. We mention two out of many possible. One is a noisy recovery problem where every coordinate has a different *unknown* noise parameter. A more challenging second variant is the following. Think of V as *centers* around which noise occurs. Assume that around every center there is a different noise distribution. Can (or under what conditions) the distribution on centers be accurately estimated using (known or unknown) different-noise-per-center samples? Such a result will be closer in spirit to what is known for Gaussian noise.

Acknowledgments We thank Zeev Dvir for helpful discussions. We thank Sanjeev Arora, Avrim Blum, Russell Impagliazzo, Dick Karp, Yishay Mansour and Elchanan Mossel for helpful comments on an earlier version of this work. We thank an anonymous referee for finding an error in a previous version of text (concerning the usage of the PRP to DRP reduction). The research of the first author is partially supported by NSF Grants CCF-0832797 and DMS-0835373. The second author is a Horev fellow, and is supported by the Taub Foundation. His research is supported by the Israeli Science Foundation (503/11) and the Binational Science Foundation (2010089).

Conflict of interest The authors declare that they have no conflict of interest.

6 Appendix: A reduction between PRP and DRP

We now explain how to reduce PRP to DRP under quite general conditions. Recall that the difference between the two problems is that in PRP we do not know the population V whereas in DRP we do know it. So DRP should be easier to solve, but as the following reduction shows in some cases the two problems are equivalent. The reason we include this part in the text is that this is a general reduction that could be helpful elsewhere (but unfortunately is not sufficient for us).

We start by formally defining *uniform samplers*. A uniform sampler is clearly more general than the two above (which acts independently in each coordinate). It includes many others, for example ones in which every vector generates a different distribution, as used in some clustering and learning mixtures of distributions, e.g. Kalai et al. (2012).

Definition A *sampler* f is a length preserving[9] map from Σ^* to probability distributions on Σ^*. A sample $v' \in \Sigma^n$ is generated by a sampler f and a distribution π on Σ^n by sampling v according to π and sampling v' according to $f(v)$. A sampler f is *uniform* if for every π on Σ^n and every $t \leq n$, the following diagram is commutative: denoting by P_t the projection onto coordinates in $[t]$,

$$\begin{array}{ccc} \pi & \xrightarrow{P_t} & P_t(\pi) \\ \downarrow f & & \downarrow f \\ v' & \xrightarrow{P_t} & v'' \end{array}$$

That is, the following two processes produce the same distribution on samples v'' in Σ^t: (i) Choose v according to π, and choose v'' according to $f(P_t(v))$. (ii) Choose v according to π, choose v' according to $f(v)$, and set $v'' = P_t(v')$.

Here is an example for the sampler in the case of noisy samples with parameter ν: For every $v \in \{0,1\}^n$, the distribution $f(v)$ is that of $v \oplus w \in \{0,1\}^n$, where $w_1, \ldots, w_n$ in $\{0,1\}$ are i.i.d. so that $\mathbb{E}w_1 = (1-\nu)/2$.

Sketch of reduction The algorithm for solving PRP reconstructs V going column-by-column by "extending and removing":

Let $V'_1 = \Sigma$. By projecting the unknown V to the first coordinate we get a PRP problem with $V'_1 = \Sigma^1$. The underlying distribution π_1 on Σ^1 is so that $\pi_1(\sigma)$ is the cumulative weight of π on vectors whose first entry is σ. The uniformity of the sampler implies that given a sample v' to PRP on V, by projecting to the first coordinate, we can generate a legal query to PRP on V_1. Using oracle access, compute $\pi'_1 = A_{DRP}(V'_1)$. If the DRP algorithm worked,

[9] For every n, the sampler f maps $v \in \Sigma^n$ to a distribution $f(v)$ on Σ^n.

π_1' is a good approximation of π_1. Let V_1 be the subset of V_1' after removing its non-useful part: V_1 is the set of all symbols σ in V_1' so that $\pi_1'(\sigma) \geq \alpha/2$.

Continue inductively: for $t > 1$, let V_t' be the extension of V_{t-1} by Σ, $V_t' = V_{t-1} \times \Sigma$. Using oracle access and due to uniformity of sampler (as above), compute $\pi_t' = A_{DRP}(V_t')$, using only the first t entries of the random samples to PRP on V. Let V_t be the set of all vectors v in V_t' so that $\pi_t'(v) \geq \alpha/2$.

Finally, output V_n and π_n'. □

References

Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. *ACM SIGMOD Record*, *29*(2), 439–450.

Batman, L., Impagliazzo, R., Murray, C., & Paturi, R. (2013). Finding heavy hitters from partial or noisy data. In *APPROX-RANDOM, 2013* (pp. 347–362).

Beck, L. (1980). A security mechanism for statistical data bases. *ACM Transactions of Databases*, *5*(3), 316–338.

Blum, A. (1994). Relevant examples and relevant features: Thoughts from computational learning theory. In *AAAI fall symposium on 'relevance'*.

Blum, A., Coja-Oghlan, A., Frieze, A., & Zhou, S. (2009). Separating populations with wide data: A spectral analysis. *Electronic Journal of Statistics*, *3*, 76–113.

Dinur, I., Nissim, K. (2003). Revealing information while preserving privacy. In *PODS* (pp. 202–210).

Dvir, Z., Rao, A., Wigderson, A., Yehudayoff, A. (2012). Restriction access. In *ITCS* (pp. 19–33).

Dwork, C., McSherry, F., Nissim, K., Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *TCC* (pp. 265–284).

Dwork, C., Smith, A. (2008). Differential privacy for statistics: What we know and what we want to learn. *In NCHS/CDC data confidentiality workshop*.

Feldman, J., O'Donnell, R., & Servedio, R. (2008). Learning mixtures of product distributions over discrete domains. *SIAM Journal on Computing*, *37*(5), 1536–1564.

Floyd, S., & Warmuth, M. K. (1995). Sample compression, learnability, and the Vapnik–Chervonenkis dimension. *Machine Learning*, *21*(3), 269–304.

Goldreich, O., Levin, L. (1989). A generic hardcore predicate for any one-way function. In *STOC* (pp. 25–30).

Holenstein, T., Mitzenmacher, M., Panigrahy, R., Wieder, U. (2008). Trace reconstruction with constant deletion probability and related results. In *SODA* (pp. 389–398).

Johnson, W., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, *26*, 189–206.

Kalai, A., Moitra, A., & Valiant, G. (2012). Disentangling Gaussians. *Communications of the ACM*, *55*(2), 113–120.

Kearns, M., Mansour, Y., Ron, D., Rubinfel, R., Schapire, R. E., Sellie, L. (1994). On the learnability of discrete distributions. In *STOC* (pp. 273–282).

Kushilevitz, E., & Mansour, Y. (1993). Learning decision trees using the fourier spectrum. *SIAM Journal of Computing*, *22*(6), 1331–1348.

Lefons, E., Silvestri, A., Tangorra, F. (1983). An analytic approach to statistical databases. In *International conference on very large data bases* (pp. 260–274).

Liew, C. K., Choi, U. J., & Liew, C. J. (1999). A data distortion by probability distribution. *Communications of the ACM*, *42*(10), 89–94.

Littlestone, N., Warmuth, M.K. (1986). Relating data compression and learnability. Unpublished, obtainable at http://www.cse.ucsc.edu/manfred/pubs/T1.

Lovett, S., Zhang, J. (2015). Improved noisy population recovery, and reverse Bonami–Beckner inequality for sparse functions. In *STOC*.

Matsen, F. A., Mossel, E., & Steel, M. (2008). Mixed-up trees: The structure of phylogenetic mixtures. *Bulletin of Mathematical Biology*, *70*(4), 1115–1139.

Moitra, A., Saks, M. (2013). A polynomial time algorithm for lossy population recovery. In *CoRR* arXiv:1302.1515

Mossell, E. (2008). The subsequence problem. Unpublished.

Nash-Williams, C. (1978). The reconstruction problem. In *Selected topics in graph theory I*. Academic Press: New York.

Traub, J., Yemini, Y., & Wozniakowski, H. (1984). The statistical security of a statistical database. *ACM Transactions on Database Systems*, *9*(4), 672–679.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*, 63–69.