Sample complexity of population recovery

Yury Polyanskiy   YP@MIT.EDU
Department of EECS, MIT, Cambridge, MA, USA

Ananda Theertha Suresh   THEERTHA@GOOGLE.COM
Google Research, New York, NY, USA

Yihong Wu   YIHONG.WU@YALE.EDU
Department of Statistics and Data Science, Yale University, New Haven, CT, USA

Abstract

The problem of population recovery refers to estimating a distribution based on incomplete or corrupted samples. Consider a random poll of sample size n conducted on a population of individuals, where each pollee is asked to answer d binary questions. We consider one of the two polling impediments:

- in lossy population recovery, a pollee may skip each question with probability ϵ;

- in noisy population recovery, a pollee may lie on each question with probability ϵ.

Given n lossy or noisy samples, the goal is to estimate the probabilities of all 2^d binary vectors simultaneously within accuracy δ with high probability.

This paper settles the sample complexity of population recovery. For lossy model, the optimal sample complexity is $\tilde{\Theta}(\delta^{-2\max\{\frac{\epsilon}{1-\epsilon},1\}})$, improving the state of the art by Moitra and Saks in several ways: a lower bound is established, the upper bound is improved and the result depends at most on the logarithm of the dimension. Surprisingly, the sample complexity undergoes a phase transition from parametric to nonparametric rate when ϵ exceeds $1/2$. For noisy population recovery, the sharp sample complexity turns out to be more sensitive to dimension and scales as $\exp(\Theta(d^{1/3}\log^{2/3}(1/\delta)))$ except for the trivial cases of $\epsilon = 0, 1/2$ or 1.

For both models, our estimators simply compute the empirical mean of a certain function, which is found by pre-solving a linear program (LP). Curiously, the dual LP can be understood as Le Cam's method for lower-bounding the minimax risk, thus establishing the statistical optimality of the proposed estimators. The value of the LP is determined by complex-analytic methods.

1. Introduction

1.1. Formulation

The problem of *population recovery* refers to estimating an unknown distribution based on incomplete or corrupted samples. Initially proposed by Dvir et al. (2012); Wigderson and Yehudayoff (2012) in the context of learning DNFs with partial observations and further investigated in Batman et al. (2013); Moitra and Saks (2013); Lovett and Zhang (2015); De et al. (2016b), this problem can also be viewed as a special instance of learning mixtures of discrete distributions in the framework of Kearns et al. (1994).

The setting of population recovery is the following: Let $\mathcal{P}_d$ be the set of all probability distributions over the hypercube $\{0,1\}^d$. Let $P \in \mathcal{P}_d$ be an unknown probability distribution and $X \stackrel{\text{def}}{=} (X_1, X_2, \ldots X_d) \sim P$. Instead of observing X, we observe its noisy version according to one of the two observation models:

- **Lossy population recovery**: For each i, Y_i is obtained by passing X_i independently through the binary erasure channel with erasure probability ϵ, where $Y_i = X_i$ with probability $1 - \epsilon$, and $Y_i =?$, with probability ϵ.

- **Noisy population recovery**: For each i, Y_i is obtained by passing X_i independently through the binary symmetric channel with error probability ϵ, where $Y_i = X_i$ with probability $1 - \epsilon$, and $Y_i = 1 - X_i$, with probability ϵ.

Given independent noisy or lossy samples, the goal of population recovery is to estimate the underlying distribution. Specifically, let $X^{(1)}, \ldots, X^{(n)}$ be independently drawn from P and we observe their noisy or lossy versions, denoted by $Y^{(1)}, \ldots, Y^{(n)}$, and aim to estimate the probabilities of all strings within δ simultaneously, i.e., $\|P - \widehat{P}\|_\infty \leq \delta$ with high probability.[1] In the absence of erasures or errors, the problem is simply that of distribution estimation and the empirical distribution is near-optimal. If the samples are lossy or noisy, the distribution of each sample is a mixture of exponentially (in d) many product distributions with mixing weights given by the input distribution P. This problem is hence a special case of learning mixtures of discrete product distributions introduced in Kearns et al. (1994) and more recently in Feldman et al. (2008); Li et al. (2015).

One of the key observations from Dvir et al. (2012) is that both the sample and algorithmic complexity of estimating P_x for all $x \in \{0,1\}^d$ is largely determined by those of estimating P_x for a single x, which, without loss of generality, can be assumed to be the zero string. This problem is referred to as *individual recovery*.[2] Note that one can convert any estimator $\widehat{P}_0$ to $\widehat{P}_x$ by XOR-ing with x the samples and applying the estimator $\widehat{P}_0$. However, naively applying union bound over all strings inflates both the time complexity and the error probability by a factor 2^d, which is unacceptable. The clever workaround in Dvir et al. (2012) is to leverage the special structure of Hamming space by recursively solving the problem on lower-dimensional subspaces. This argument is further explained in Appendix A along with an improved analysis. Specifically, given any estimator $\widehat{P}_0$ using n samples and time complexity t such that $\mathbb{E}|\widehat{P}_0 - P_0| \leq \delta$, it can be converted to a distribution estimator $\widehat{P}$ such that $\|\widehat{P} - P\|_\infty \leq \delta$ with probability at least $1 - \tau$ with $n \cdot \log \frac{d}{\delta\tau}$ samples and time complexity $t \cdot \frac{d}{\delta} \log \frac{d}{\delta\tau}$. Therefore the problem of population recovery is equivalent, both statistically and algorithmically, to the problem of individual recovery.

To understand the statistical fundamental limit of this problem, we consider the *minimax risk*, defined as:

$$R^*(n, d) = \inf_{\widehat{P}_0} \sup_P \mathbb{E}_P[(\widehat{P}_0 - P_0)^2]. \tag{1}$$

---

1. Equivalently, up to constant factors, we need to output a set of strings $S \subset \{0,1\}^d$ and $\widehat{P}_x$ for each $x \in S$, such that for all $x \in S$, $|\widehat{P}_x - P_x| \leq \delta$ and for all $x \notin S$, $P_x \leq \delta$. The point is that even when dimension d is large the list S can be kept to a finite size of order $\frac{1}{\delta}$.

2. This can be viewed as the combinatorial counterpart of estimating the density at a point, a well-studied problem in nonparametric statistics, where given n iid samples drawn from a density f, the goal is to estimate $f(0)$ cf. Tsybakov (2009).

To be consistent with the existing literature, the main results in this paper are phrased in terms of *sample complexity*:

$$n^*(\delta, d) = \min\{n : R^*(n, d) \le \delta^2\}, \tag{2}$$

with subscript L or N denoting the lossy or noisy observation model. Up to constant factors, $n^*(\delta, d)$ is also the minimal sample size such that P_0 can be estimated within an additive error of δ with probability, say, $1/3$. The focus on this paper is to obtain sharp bounds on both $n_\mathsf{L}^*(d, \delta)$ and $n_\mathsf{N}^*(d, \delta)$ and computationally efficient estimators with provable optimality. We make no assumption on the support size of the underlying distribution.

1.2. Prior work

To review the existing results, we start with lossy population recovery. Recall that ϵ is the erasure probability. A polynomial-time estimator is given in Dvir et al. (2012) that succeeds for $\epsilon \le 0.635$, which was subsequently improved to $\epsilon \le 1/\sqrt{2}$ by Batman et al. (2013). The state of the art is Moitra and Saks (2013) who proposed a polynomial time algorithm that works for all $\epsilon < 1$, with sample complexity:

$$n_\mathsf{L}^*(d, \delta) \lesssim \left(\frac{d}{\delta}\right)^{\frac{2}{1-\epsilon} \log \frac{2}{1-\epsilon}}. \tag{3}$$

It is worth noting that for the lossy model, most of the estimators for P_0 are of the following form:

$$\widehat{P}_0 = \frac{1}{n} \sum_{i=1}^{n} g(w_i) \tag{4}$$

where $w_i = w(Y^{(i)})$ is the number of ones in the ith sample $Y^{(i)}$. Such an estimator is referred to as a *linear estimator* since (4) can be equivalently written as a linear combination $\widehat{P}_0 = \frac{1}{n} \sum_{j=0}^{d} g(j) N_j$, where N_j is the number of samples of Hamming weight j. Both Moitra and Saks (2013) and De et al. (2016b) focus on efficient algorithms to construct the coefficient g.

The problem of noisy population recovery was first studied by Wigderson and Yehudayoff (2012). In contrast to the lossy model which imposes no assumption on the input, all provable results for the noisy model are obtained under a "sparsity" assumption, namely, the distribution P has bounded support size k. An algorithm for noisy recovery is proposed in Wigderson and Yehudayoff (2012) with sample complexity $k^{\log k}$. This was further improved by Lovett and Zhang (2015) to $k^{\log \log k}$. Recently De et al. (2016b) proposed a recovery algorithm with sample complexity that is polynomial in $(k/\delta)^{1/(1-2\epsilon)^4}$ and d, where ϵ is the probability to flip each bit.

Despite these exciting advances, several fundamental questions remain unanswered:

- Is it sufficient to restrict to the family of linear estimators of the form (4)?

- What is the optimal sample complexity of population recovery in terms of ϵ, d, δ?

- Estimators such as (4) only depend on the number of ones in the sample. For lossy population recovery, in fact the number of ones and zeros in each sample together

3

constitute a sufficient statistic (see Remark 5) and a natural question arises: is the number of zeros in the samples almost uninformative for estimating P_0?

- For noisy population recovery, is the assumption of bounded support size necessary for achieving polynomial (in d) sample complexity?

As summarized in the next two subsections, the main results in this paper settle all of these questions. Specifically, we find the optimal sample complexity and show that linear estimators (4) suffice to achieve it. For lossy population recovery, the number of zeros indeed can be ignored. For the noisy model, without the bounded support assumption the sample complexity scales superpolynomially in the dimension.

1.3. New results

The main results of this paper provide sharp characterizations of the sample complexity for both lossy and noisy population recovery, as well as computationally efficient estimators, shown optimal by minimax lower bounds. We start from lossy recovery. The next result determines the optimal sample complexity up to a polylog factor, which turns out to be *dimension-free*.

Theorem 1 (Lossy population recovery) *Assume $\delta < 1/e$. There exist universal constants c_1, c_2 such that the following hold. For the low-erasure regime of $\epsilon \leq 1/2$,*

$$\frac{c_1}{\delta^2} \leq \sup_{d \in \mathbb{N}} n_{\mathsf{L}}^*(\delta, d) \leq \frac{1}{\delta^2}. \tag{5}$$

For the high-erasure regime of $\epsilon > 1/2$,

$$c_2 \Big(e\delta \log \frac{1}{\delta} \Big)^{-\frac{2\epsilon}{1-\epsilon}} \leq \sup_{d \in \mathbb{N}} n_{\mathsf{L}}^*(\delta, d) \leq \delta^{-\frac{2\epsilon}{1-\epsilon}}. \tag{6}$$

Furthermore, for (5) and (6), the supremum can be replaced by any $d \geq 1$ and $d \gtrsim \delta^{-\frac{2\epsilon}{1-\epsilon}} \mathsf{polylog}(\frac{1}{\delta})$, respectively.

Remark 2 (Elbow effects) *An important result in the statistics literature on nonparametric and high-dimensional functional estimation, the* elbow effect *refers to the phenomenon that there exists a critical regularity parameter (e.g., smoothness) below which the rate of estimation is parametric $O(1/n)$ and above which, it becomes nonparametric (slower than $1/n$) in the number of samples n. Taking a nonparametric view of the lossy population recovery problem by considering the input as binary sequence of infinite length, what we proved is the following characterization of the minimax estimation error:*

$$\inf_{\widehat{P_0}} \sup_P \mathbb{E}_P[(\widehat{P_0} - P_0)^2] = \begin{cases} \Theta(n^{-1}) & 0 \leq \epsilon \leq \frac{1}{2} \\ n^{-\frac{1-\epsilon}{\epsilon}} \mathsf{polylog}(n) & \frac{1}{2} < \epsilon < 1 \end{cases} \tag{7}$$

which exhibits the following elbow phenomenon:

4

- *In the* low-erasure *regime of $\epsilon \leq 1/2$, it is possible to achieve the parametric rate of $1/n$ in any dimension. Clearly this is the best one can hope for even when $d = 1$, in which case one wants to estimate the bias of a coin based on n samples but unfortunately ϵ-fraction of the samples are lost.*

- *In the* high-erasure *regime of $\epsilon > 1/2$, the infinite dimensionality of the problem kicks in and the optimal rate of convergence becomes nonparametric and strictly slower than $1/n$.*

Since elbow effects are known for estimating non-linear (e.g., quadratic) functionals (see, e.g., Ibragimov et al. (1987); Cai and Low (2005); Fan et al. (2015)), it is somewhat surprising that it is also manifested in lossy population recovery where the goal is to estimate a simple linear functional, namely, P_0.

The lossy population recovery problem should also be contrasted with classical statistical problems in the presence of *missing data*, *such as low-rank matrix completion and principle component analysis. In the latter case, it has been shown in Lounici (2014) that if an ϵ faction of the coordinates of each sample are randomly erased, this effectively degrades the sample size from n to $n\epsilon^2$. For lossy population recovery, Theorem 1 shows there is a phase transition in the effect of missing observations: when more than half of the data are erased ($\epsilon > 1/2$), the rate of convergence are penalized and the effective sample size drops from n to $n^{(1-\epsilon)/\epsilon}$.*

Next we turn to the noisy population recovery. Perhaps surprisingly, the sample complexity is no longer dimension-free and in fact grows with the dimension d super-polynomially. The next result determines it up to a constant in the exponent.

Theorem 3 (Noisy population recovery)   *Let $\mu(\epsilon) \triangleq \frac{\epsilon(1-\epsilon)}{(1-2\epsilon)^2}$. For all $d \geq 1$,*

$$\min\left\{\exp(c_1(1-2\epsilon)^2 d), \exp\left(c_1\left(d\mu(\epsilon)\log^2\frac{1}{\delta}\right)^{1/3}\right)\right\} \leq n_N^*(\delta, d) \leq \exp\left(c_2\left(d\mu(\epsilon)\log^2\frac{1}{\delta}\right)^{1/3}\right).$$

where c_1, c_2 are universal constants and the upper and lower bounds hold for all $\delta < 1$ and all $\delta < 1/3$, respectively.

Theorem 3 shows that to estimate P_0 within a constant accuracy δ, the optimal sample size scales as

$$n_N^*(\delta, d) = \exp\left(\Theta\left(\left(\frac{\epsilon(1-\epsilon)}{(1-2\epsilon)^2}d\log^2\frac{1}{\delta}\right)^{1/3}\right)\right).$$

which is superpolynomial in the dimension. This shows that the assumption made by Kearns et al. (1994); Wigderson and Yehudayoff (2012) and subsequent work that the input distribution has bounded support size is in fact crucial for achieving polynomial sample complexity. Indeed both algorithms in Lovett and Zhang (2015); De et al. (2016b) are based on Fourier analysis of Boolean functions and exploit the sparsity of the distribution and they are not of the linear form (4).

Finally, we mention that a subset of our results was discovered independently by De et al. (2017) using similar techniques.

1.4. Technical contributions

To describe our approach, we start from the constructive part. For both lossy and noisy population recovery, we also focus on linear estimator (4). Choosing the coefficient vector g to minimize the worst-case mean squared error $\mathbb{E}(P_0 - \widehat{P}_0)^2$ leads to the following linear programming (LP):

$$\min_{g \in \mathbb{R}^{d+1}} \|\Phi^\top g - e_0\|_\infty + \frac{1}{\sqrt{n}}\|g\|_\infty, \tag{8}$$

where $e_0 = (1, 0, \ldots, 0)^\top$ and Φ is a column stochastic matrix (probability transition kernel) that describes the conditional distribution of the output Hamming weight given the input weight. Here the first and second term in (8) correspond to the bias and standard deviation respectively. In fact, for lossy recovery the (unique) unbiased estimator is a linear one corresponding to $g = (\Phi^\top)^{-1} e_0$. When the erasure probability $\epsilon \leq \frac{1}{2}$, this vector has bounded entries and hence the variance is $O(\frac{1}{n})$. This has already been noticed in (Dvir et al., 2012, Sec. 6.2). However, when $\epsilon > \frac{1}{2}$, the variance of the unbiased estimator is exponentially large and the LP (8) aims to achieve the best bias-variance tradeoff.

Surprisingly, we show that the value of the above LP also produces a lower bound that applies to *any* estimator. This is done by relating the dual program of (8) to Le Cam's two-point method for proving minimax lower bound Le Cam (1986). Section 2 formalizes this argument and introduces a general framework of characterizing the minimax risk of estimating linear functionals of discrete distributions by means of linear programming.

The bulk of the paper is devoted to evaluating the LP (8) for the lossy (Section 3) and noisy model (Section 4). The common theme is to recast the dual LP in the function space on the complex domain, consider its H^∞-relaxation, and use tools from complex analysis to bound the value. Similar proof technique was previously employed in Moitra and Saks (2013) to upper-bound the value of the dual LP in order to establish the sample complexity upper bound in (3). Here we tighten the analysis to obtain the optimal exponent and dimension-free result. Furthermore, we show that the dual LP not only provides an upper bound on the estimation error, it also gives minimax lower bound via the general result in Section 2 together with a refinement based on Hellinger distance. To show the impossibility results we need to bound the value of the dual LP from below, which we do by demonstrating an explicit solution that plays the role of the least favorable prior in the minimax lower bound.

Initially, we were rather surprised to have non-trivial complex-analytic methods such Hadamard three-lines theorem emerge in this purely statistical problem. Realizing to have rediscovered portions of Moitra and Saks (2013), we were further surprised when a new estimator for the trace reconstruction problem[3] was recently proposed by Nazarov and Peres (2016) and De et al. (2016a), whose design crucially relied on complex-analytic methods. In fact, one of the key steps of Nazarov and Peres (2016) relies on the results from Borwein and Erdélyi (1997), which we also use in Section 4 to obtain sharp bounds for noisy population recovery.

---

3. This problem, in short, aims to reconstruct a binary string based on its independent observations through a memoryless deletion channel.

Finally, in Section 5 we propose an alternative linear estimator, which requires a slightly worse sample complexity

$$\delta^{-2\max\{\frac{3\epsilon-1}{1-\epsilon},1\}},$$

having the exponent off by a factor of at most 2 in the worst case of $\epsilon \to 1$. The advantage of this estimator is that it is explicit and does not require pre-solving a d-dimensional LP. The estimator is obtained by applying the *smoothing* technique introduced in Orlitsky et al. (2016) that averages a sequence of randomly truncated unbiased estimators to achieve good bias-variance tradeoff.

1.5. Notations

Let $\mathrm{Bin}(n,p)$ denote the binomial distribution with parameters n and p. For a pair of probability distributions P and Q, let $P \otimes Q$ and $P * Q$ denote their product distribution and convolution, respectively. Let $\mathrm{TV}(P,Q) = \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q|$, $H^2(P,Q) = \int (\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q})^2$ and $H(P,Q) = \sqrt{H^2(P,Q)}$ denote the total variation, squared and non-squared Hellinger distance, respectively. Throughout the paper, for any $\epsilon \in (0,1)$, let $\bar{\epsilon} \triangleq 1 - \epsilon$. For a family of parametric distribution $\{P_\theta : \theta \in \Theta\}$ and a prior π on Θ, with a slight abuse of notation we write $\mathbb{E}_{\theta \sim \pi}[P_\theta]$ to denote the mixture distribution $\int \pi(\mathrm{d}\theta)P_\theta$. For a holomorphic function $f(z)$ on the unit disk $D \subset \mathbb{C}$ we let $[z^k]f(z)$ denote the k-th coefficient of its Maclaurin series.

2. Estimating linear functionals and duality of Le Cam's method

In this section, we relate the general problem of estimating linear functionals of distributions to the solution of a *minimal total variation* linear program. The latter corresponds to the optimization of a pair of distributions in Le Cam's two-point method and thus produces minimax lower bounds. Surprisingly, we also show that dual linear program produces an estimator, thereby explaining the general tightness of the two-point method.

Theorem 4 *Let Θ and $\mathcal{X}$ be finite sets and $\{P_\theta : \theta \in \Theta\}$ be a collection of distributions on $\mathcal{X}$. Let π be a distribution (prior) on Θ and let $X_1, \ldots, X_n$ be iid samples from the mixture distribution[4] $\sum_{\theta \in \Theta} \pi(\theta)P_\theta$. To estimate a linear functional of π*

$$F(\pi) \triangleq \langle \pi, h \rangle = \mathbb{E}_{\theta \sim \pi} h(\theta),$$

define the minimax quadratic risk

$$R_n \triangleq \inf_{\widehat{F}} \sup_\pi \mathbb{E}_\pi[(\widehat{F}(X_1, \ldots, X_n) - F(\pi))^2]. \tag{9}$$

Without loss of generality[5] we assume that there exists θ_+ and θ_- such that $h(\theta_-) \leq 0 \leq h(\theta_+)$. Then we have

$$\frac{1}{64}\delta\left(\frac{1}{n}\right)^2 \leq R_n \leq \delta\left(\frac{1}{\sqrt{n}}\right)^2, \tag{10}$$

---

4. In other words, for each $i = 1, \ldots, n$, θ_i is drawn iid from π and $X_i \sim P_{\theta_i}$ independently.

5. This is because R_n is unchanged if h is replaced by $h - c\mathbf{1}$ for any constant c. Note, however, that the value of (11) does change.

where $\delta(t)$ is given by the following linear program

$$\delta(t) \triangleq \max_{\Delta \in \mathbb{R}^\Theta} \left\{ \langle \Delta, h \rangle : \|\Delta\|_1 \leq 1, \|\Phi\Delta\|_1 \leq t \right\}, \tag{11}$$

where $\Phi = (\Phi_{x,\theta})_{x \in \mathcal{X}, \theta \in \Theta}$ is an $|\mathcal{X}| \times |\Theta|$ column-stochastic matrix with $\Phi_{x,\theta} \triangleq P_\theta(x)$.

Remark 5 *Particularizing the above framework to the problem of population recovery, the subsequent two sections will be devoted to computing the LP value $\delta(t)$ for the lossy and noisy model, respectively. Below we specify the setting as a concrete example. Since the probability of the zero vector is* permutation-invariance, *it is sufficient[6] to consider permutation-invariant distributions and the relevant parameter is the Hamming weight of the input string, denoted by θ, which is distributed according a distribution π supported on the parameter space $\Theta = \{0, 1, \ldots, d\}$. The goal, in turn, is to estimate $\langle \pi, e_0 \rangle = \pi(0)$. Each sample is sufficiently summarized into their* types:

- *For lossy population recovery: $\mathcal{X} = \{0, \ldots, d\}^2$ and $X = (U, V)$, where U and V denote the number of ones and zeros, respectively, in the output. More precisely, we have*

$$X|\theta \sim \mathrm{Bin}(\theta, \bar{\epsilon}) \otimes \mathrm{Bin}(d - \theta, \bar{\epsilon}). \tag{12}$$

- *For noisy population recovery: $\mathcal{X} = \{0, \ldots, d\}$ where X is the number of ones in output, and so the model is given by*

$$X|\theta \sim \mathrm{Bin}(\theta, \bar{\epsilon}) * \mathrm{Bin}(d - \theta, \epsilon) \tag{13}$$

Remark 6 *As will be evident from the proof, the square-root gap in the LP characterization (10) is due the usage of the bound $\mathrm{TV}(P^{\otimes n}, Q^{\otimes n}) \leq n\mathrm{TV}(P, Q)$, which is often loose. To close this gap, we resort to the Hellinger distance and for lossy population recovery we show in Section 3 that*

$$R_n = \delta \left(\frac{1}{\sqrt{n}} \right)^2 \mathsf{polylog}(n).$$

For noisy population recovery, it turns out that $\delta(t)$ is exponentially small in $\mathsf{polylog}(\frac{1}{t})$ and hence (10) is sufficiently tight (see Section 4).

Proof [Proof of Theorem 4] To prove the right-hand inequality in (10) we consider the following estimator

$$\widehat{F}(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} g(X_i) \tag{14}$$

for some $g \in \mathbb{R}^\mathcal{X}$. Then

$$\mathbb{E}_\pi[\widehat{F}(X_1, \ldots, X_n)] = \sum_{\theta \in \Theta} \pi(\theta) \sum_{x \in \mathcal{X}} g(x) P_\theta(x) = \pi^\top \Phi^\top g = \langle \pi, \Phi^\top g \rangle$$

---

6. Indeed, for the worst-case formulation (1), the least favorable distribution P is *permutation-invariant*. Furthermore, if the input string X has a permutation-invariant distribution, so is the distribution of the lossy or noisy observation Y, for which a sufficient statistic is its *type*.

and the bias is

$$|\mathbb{E}_\pi[\widehat{F}] - F(\pi)| \leq \|\pi\|_1 \|\Phi^\top g - h\|_\infty = \|\Phi^\top g - h\|_\infty.$$

For variance,

$$\mathsf{var}_\pi(\widehat{F}) = \frac{1}{n}\mathsf{var}_\pi(g(X)) \leq \frac{1}{n}\|g\|_\infty^2.$$

This shows

$$\sqrt{R_n} \leq \sup_\pi \left(\mathbb{E}_\pi[|\widehat{F} - F(\pi)|^2]\right)^{\frac{1}{2}} \leq \|\Phi^\top g - h\|_\infty + \frac{1}{\sqrt{n}}\|g\|_\infty. \tag{15}$$

We now optimize the right-hand side of (15) over all g as follows:

$$\min_{g \in \mathbb{R}^{\mathcal{X}}} \|\Phi^\top g - h\|_\infty + \frac{1}{\sqrt{n}}\|g\|_\infty \tag{16}$$

$$= \min_g \max_{\|y\|_1 \leq 1, \|z\|_1 \leq 1} \langle y, \Phi^\top g - h \rangle + \frac{1}{\sqrt{n}}\langle z, g \rangle \tag{17}$$

$$= \min_g \max_{\|y\|_1 \leq 1, \|z\|_1 \leq 1} \left\langle \Phi y + \frac{1}{\sqrt{n}}z, g \right\rangle - \langle y, h \rangle \tag{18}$$

$$= \max_{\|y\|_1 \leq 1, \|z\|_1 \leq 1} \min_g \left\langle \Phi y + \frac{1}{\sqrt{n}}z, g \right\rangle - \langle y, h \rangle \tag{19}$$

$$= \max_{\|y\|_1 \leq 1, \|\Phi y\|_1 \leq 1/\sqrt{n}} \langle y, h \rangle = \delta\left(\frac{1}{\sqrt{n}}\right), \tag{20}$$

where in (16) we write min since the optimization can clearly be restricted to a suitably large ℓ_∞-box, (17) is by the dual representation of norms, (18) is by linearity, (19) is by von Neumann's minimax theorem for bilinear functions, and (20) follows since the inner minimization in (19) is $-\infty$ unless $\Phi y + \frac{1}{\sqrt{n}}z = 0$. In fact, the equality of (16) and (20) also follows from the strong duality of finite-dimensional LP.

To prove the left-hand bound in (10), we first introduce an auxiliary linear program:

$$\tilde{\delta}(t) \triangleq \max_\Delta \{\langle \Delta, h \rangle : \|\Delta\|_1 \leq 1, \langle \Delta, \mathbf{1} \rangle = 0, \|\Phi\Delta\|_1 \leq t\}, \tag{21}$$

where Φ and Δ are as in (11).

We recall the basics of the Le Cam's two-point method Le Cam (1986). If there exist two priors on Θ, under which the distributions of the samples are not perfectly distinguishable, i.e., of a small total variation, then the separation in the functional values constitutes a lower bound that holds for all estimators. Specifically, for estimating linear functionals under the quadratic risk, we have Le Cam (1986)

$$R_n \geq \frac{1}{8} \max_{\pi,\pi'} \langle \pi - \pi', h \rangle^2 (1 - \mathrm{TV}((\Phi\pi)^{\otimes n}, (\Phi\pi')^{\otimes n}), \tag{22}$$

where π, π' are probability vectors on Θ. Since $\mathrm{TV}((\Phi\pi)^{\otimes n}, (\Phi\pi')^{\otimes n}) \leq n\mathrm{TV}(\Phi\pi, \Phi\pi')$, we have

$$R_n \geq \frac{1}{16} \left(\max_{\pi,\pi'} \left\{\langle \pi - \pi', h \rangle : \mathrm{TV}(\Phi\pi, \Phi\pi') \leq \frac{1}{2n}\right\}\right)^2. \tag{23}$$

9

Now, we observe that

$$\max_{\pi,\pi'}\left\{\langle \pi - \pi', h\rangle : \mathrm{TV}(\Phi\pi, \Phi\pi') \leq \frac{1}{2n}\right\} = 2\tilde{\delta}\left(\frac{1}{2n}\right). \tag{24}$$

Indeed, the optimizer Δ in (21) can clearly be chosen so that $\|\Delta\|_1 = 1$. Next, decompose $\Delta = \Delta_+ - \Delta_-$, where $\Delta_\pm(x) \triangleq \max(\pm\Delta(x), 0)$. From $\langle \Delta, 1\rangle = 0$ we conclude that $\pi = 2\Delta_+$ and $\pi' = 2\Delta_-$ are valid probability distributions on Θ and, furthermore,

$$\mathrm{TV}(\Phi\pi, \Phi\pi') = \|\Phi\Delta\|_1, \quad \langle \pi - \pi', h\rangle = 2\langle \Delta, h\rangle.$$

For the reverse direction, simply take $\Delta = \frac{1}{2}(\pi - \pi')$.

Overall, from (23)-(24) we get

$$R_n \geq \frac{1}{4}\tilde{\delta}\left(\frac{1}{2n}\right)^2.$$

To complete the proof the lower-bound in (10) we invoke property 2 (to convert to $\delta(t)$) and property 1 (with $\lambda = 1/2$) from the following lemma (proved in Appendix B). ∎

Lemma 7 (Properties of $\delta(t)$ and $\tilde{\delta}(t)$)

1. *For any $\lambda \in [0, 1]$ we have*

$$\delta(\lambda t) \geq \lambda\delta(t), \qquad \tilde{\delta}(\lambda t) \geq \lambda\tilde{\delta}(t).$$

2. *Assuming that there exist $\theta_\pm$ such that $h(\theta_+) \geq 0 \geq h(\theta_-)$ we have*

$$\tilde{\delta}(t) \leq \delta(t) \leq 2\tilde{\delta}(t). \tag{25}$$

3. *For any non-constant h, there exists $C = C(h) > 0$ such that[7]*

$$\tilde{\delta}(t) \geq C(h)t.$$

Remark 8 *Although Theorem 4 is in terms of the mean-square error, it is easy to obtain high-probability bound using standard concentration inequalities. Consider the estimator (14) with g being the solution to the LP (11) with $t = \frac{1}{\sqrt{n}}$. Since the bias satisfies $|\mathbb{E}[\widehat{F} - F]| \leq \delta(\frac{1}{\sqrt{n}})$ and the standard deviation is at most $\frac{1}{\sqrt{n}}\|g\|_\infty \leq \delta(\frac{1}{\sqrt{n}})$, Hoeffding inequality implies the Gaussian concentration $\mathbb{P}[|\widehat{F} - F| \geq t\delta(1/\sqrt{n})] \leq \exp(-ct^2)$ for all sufficiently large t and some absolute constant c.*

---

7. This property is not used for establishing Theorem 4. We prove it because it implies that $1/n$ (parametric rate) is always a lower bound in (10).

3. Lossy population recovery

Capitalizing on the general framework introduced in Section 2, we prove the sample complexity bounds for lossy population recovery announced in Theorem 1. The outline is the following:

1. In Section 3.1 we obtain sharp bounds on the value of the LP (11) in the infinite-dimensional case, with the restriction that the estimator is allowed to depend only on the number of ones in the erased samples (Proposition 9). In particular, the upper bound on the LP value leads to the sample complexity upper bound in Theorem 1 for any dimension d.

2. The lower bound is proved in Section 3.2 in two steps: (i) By resorting to the Hellinger distance, in Lemma 12 we remove the square-root gap in the general Theorem 4. (ii) Recall from Remark 5 that it is sufficient to consider estimators that are functions of the number of zeros and ones. To complete the proof, we show that the number of zeros provides negligible information for estimating the probability of the zero vector.

3.1. Solving the linear programming by H^∞-relaxation

In this subsection we determine the value of the LP (11) for lossy population recovery where the estimator is restricted to be functions on the output Hamming weight. In view of (12), to apply Theorem 4, here Φ is the transition matrix from the input to the output Hamming weight:

$$\Phi_{ij} = P(w(Y) = i | w(X) = j) = \begin{cases} \binom{j}{i}(1-\epsilon)^i \epsilon^{j-i} & i \leq j \\ 0 & i > j \end{cases} \tag{26}$$

and the linear functional corresponds to $h = e_0$. In other words, (11) reduces to

$$\delta(t) = \sup_\Delta \left\{ \Delta_0 : \|\Delta\|_1 \leq 1, \left\| \sum_{j \geq 0} \Delta_j \mathrm{Bin}(j, 1-\epsilon) \right\|_1 \leq t \right\}. \tag{27}$$

For notational convenience, let us define the following "min-TV" LP that is equivalent to (27):

$$t(\delta) \triangleq \inf_\Delta \left\{ \left\| \sum_{j \geq 0} \Delta_j \mathrm{Bin}(j, 1-\epsilon) \right\|_1 : \Delta_0 \geq \delta, \|\Delta\|_1 \leq 1 \right\}. \tag{28}$$

Clearly we have $\delta(t(\delta)) = \delta$.

Proposition 9 *If $\epsilon \leq \frac{1}{2}$, then*

$$\delta \leq t(\delta) \leq 2(1-\epsilon)\delta. \tag{29}$$

If $\epsilon > \frac{1}{2}$, then

$$\delta^{\frac{\epsilon}{1-\epsilon}} \leq t(\delta) \leq \left(e\delta \log \frac{1}{\delta} \right)^{\frac{\epsilon}{1-\epsilon}} \tag{30}$$

where the right inequality holds provided that $\delta < 1/e$.

Remark 10 *Thanks to Theorem 4, the lower bound on $t(\delta)$ in Proposition 9 immediately translates into the following upper bound on the MSE of estimating P_0 in any dimension:*

$$\sup_{d \geq 1} R^*(n,d) \leq \delta(1/\sqrt{n})^2 \leq n^{-\min\left\{\frac{1-\epsilon}{\epsilon}, 1\right\}},$$

without any hidden constants. The upper bound on $t(\delta)$ require additional work to yield matching minimax lower bounds; this is done in Section 3.2.

Proof [Proof of Proposition 9] Let D be the open unit disk and $\bar{D}$ the closed unit disk on the complex plane. For analytic functions on D we introduce two norms:

$$\|f\|_{H^\infty(D)} \triangleq \sup_{z \in D} |f(z)|, \qquad \|f\|_A \triangleq \sum_{n \geq 0} |a_n|,$$

where a_n are the Taylor coefficients of $f(z) = \sum_n a_n z^n$. Functions with bounded A-norm form a space known as the Wiener algebra (under multiplication). It is clear that every A-function is also continuous on $\bar{D}$. Furthermore, by the maximal modulus principle, any function continuous on $\bar{D}$ and analytic on D satisfies:

$$\|f\|_{H^\infty(D)} = \sup_{z \in \partial D} |f(z)|, \tag{31}$$

and therefore:

$$\|f\|_{H^\infty} \leq \|f\|_A. \tag{32}$$

In general, there is no estimate in the opposite direction;[8] nevertheless, we can estimate the A-norm using the H^∞-norm over a larger domain: Suppose $f \in H^\infty(rD)$ for $r > 1$, then

$$\|f\|_A \leq \frac{r}{r-1} \|f\|_{H^\infty(rD)}. \tag{33}$$

Indeed, from Cauchy's integral formula we estimate coefficients a_n as

$$|a_n| \leq r^{-n} \|f\|_{H^\infty(rD)} \tag{34}$$

and then sum over $n \geq 0$ to get (33).

Now, to every sequence $\{\Delta_j : j \geq 0\}$ with finite ℓ_1-norm, we associate a function in A as

$$f(z) \triangleq \sum_{j \geq 0} \Delta_j z^j,$$

which, in case when Δ is a probability distribution, corresponds its generating function. Note that the channel Φ maps the distribution of the input Hamming weight $w(X)$ into that of the output Hamming weight $w(Y)$ via a linear transformation. Equivalently, we

---

8. That is, the space A is a strict subset of $H^\infty(D)$. This is easiest to see by noticing that all functions in A are continuous on the unit circle, while $f(z) = \frac{1}{(1-z)^i}$ is in H^∞ but discontinuous on the unit circle. Another example of a function in H^∞ but not in A is $f(z) = \exp(\frac{1+z}{1-z})$, but it takes effort to show it is not in A. Also note that the linear functional $f \mapsto f(1)$ defined on polynomials is bounded in both A-norm and $H^\infty(D)$ norm. However, in A it admits a unique extension to all of A, while in $H^\infty(D)$ there are different incompatible extensions (existence is from Hahn-Banach).

can describe the action Φ on the function f in terms how input generating functions are mapped to that of the output as follows:

$$\mathbb{E}[z^{w(X)}] \overset{\Phi}{\mapsto} \mathbb{E}[z^{w(Y)}] = \mathbb{E}[z^{\mathrm{Bin}(w(X),\bar{\epsilon})}] = \mathbb{E}[(\epsilon z + \bar{\epsilon})^X]$$

that is, $f(z) \mapsto f(\epsilon z + \bar{\epsilon})$. To this end, define the following operator, also denoted by Φ, as

$$(\Phi f)(z) \triangleq f(\epsilon + \bar{\epsilon}z), \tag{35}$$

which is known as a composition operator on a unit disk.[9] Therefore, the channel maps Δ linearly into another ℓ_1-sequence $\Delta' = \sum_j \Delta_j \mathrm{Bin}(j, \bar{\epsilon})$, which are exactly the coefficients of Φf. Indeed, $\Delta'_j = \sum_{k \geq j} \Delta_k \binom{k}{j} \bar{\epsilon}^j \epsilon^{k-j}$ and

$$\sum_{j \geq 0} \Delta'_j z^j = \sum_{k \geq 0} \Delta_k \sum_{j=0}^{k} \binom{k}{j} (\bar{\epsilon}z)^j \epsilon^{k-j} = \sum_{k \geq 0} \Delta_k (\bar{\epsilon}z + \epsilon)^k = (\Phi f)(z).$$

Then, with the above identification, the LP (28) can be recast as

$$t(\delta) = \inf\{\|\Phi f\|_A : \|f\|_A \leq 1, f(0) \geq \delta\}. \tag{36}$$

From (35) an important observation is that Φf restricts f to a *horodisk*

$$D_\tau \triangleq \bar{\tau} + \tau D = \{z \in \mathbb{C} : |z - \bar{\tau}| \leq \tau\} \tag{37}$$

which shrinks as τ decreases from 1 to 0 (see Fig. 1). Thus, $\|\Phi f\|_{H^\infty(D)} = \sup_{z \in D_\epsilon} |f(z)|$.
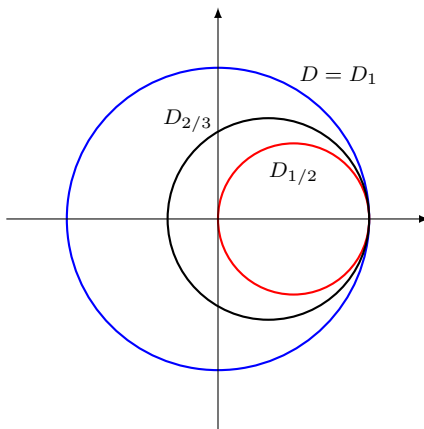

Figure 1: Horodisks.

In view of (32), it is clear that $t(\delta)$ is lower bounded by

$$t_1(\delta) \triangleq \inf\{\|\Phi f\|_{H^\infty(D)} : \|f\|_{H^\infty(D)} \leq 1, f(0) \geq \delta\}, \tag{38}$$

9. Despite the simple definition, characterizing properties of such operators is a rather difficult task cf. Cowen (1988).

Furthermore, if $\epsilon \leq 1/2$ then $0 \in \bar{D}_{\bar{\epsilon}}$ and thus $t_1(\delta) \geq \delta$. This proves the lower bound in (29). To show the upper bound, Take $f(z) = \delta(1 - z)$, which is feasible since $f(0) = \delta$ and $\|f\|_A = 2\delta < 1$. Clearly, $(\Phi f)(z) = \delta\bar{\epsilon}(1 - z)$ which gives $t(\delta) \leq 2\bar{\epsilon}\delta$.

In the remainder of the proof we focus on the non-trivial case of $\epsilon > 1/2$. Note that the Möbius transform $z \mapsto \frac{z-1}{z+1}$ maps the right half-plane onto the D so that: a) the imaginary axis gets mapped to the unit circle ∂D; b) the line $1 + i\mathbb{R}$ gets mapped to the horocircle $\partial D_{1/2}$ that passes through 0; c) the line $\frac{\epsilon}{\bar{\epsilon}} + i\mathbb{R}$ gets mapped to the horocircle $\partial D_{\bar{\epsilon}}$. Then by Hadamard's three-lines theorem (see, e.g., (Simon, 2011, Theorem 12.3)) we have that for any function $f \in H^\infty(D)$:

$$\sup_{z \in D_{1/2}} |f(z)| \leq \left(\sup_{z \in D} |f(z)|\right)^{\frac{1-2\bar{\epsilon}}{\epsilon}} \left(\sup_{z \in D_{\bar{\epsilon}}} |f(z)|\right)^{\frac{\bar{\epsilon}}{\epsilon}}. \tag{39}$$

Since any feasible solution f to (38) has $f(0) \geq \delta$ and $\|f\|_{H^\infty(D)} \leq 1$, we conclude that

$$t(\delta) \geq t_1(\delta) \geq \delta^{\frac{\epsilon}{\bar{\epsilon}}},$$

proving the lower bound in (30).

To show the upper bound, we demonstrate an explicit feasible solution for (36). Choose $\alpha < 1$ such that

$$\beta \triangleq \frac{\delta}{1 - \alpha} \leq 1. \tag{40}$$

The main idea is to choose the function so that the comparison inequality (39) is tight. To this end, recall that Hadamard three-lines theorem holds with equality for exponential function. This motivates us to consider the following mother function

$$g(z) \triangleq \beta^{\frac{1+z}{1-z}}.$$

Note that for any $\eta \in \mathbb{R}$, for the horodisk D_η defined in (37), $z \mapsto \frac{1+z}{1-z}$ maps the horocircle ∂D_η back to the straight line $\frac{\bar{\eta}}{\eta} + i\mathbb{R}$. Since $0 \leq \beta \leq 1$,

$$\|g\|_{H^\infty(D_\eta)} = \beta^{\inf_{z \in \partial D_\eta} \operatorname{Re}(\frac{1+z}{1-z})} = \beta^{\frac{\bar{\eta}}{\eta}}. \tag{41}$$

In particular, since $D = D_1$, by setting $\eta = 1$ we get $\|g\|_{H^\infty(D)} = 1$. Next, for $\alpha \in (0, 1)$ define the scaled function

$$f_\alpha(z) \triangleq (1 - \alpha)g(\alpha z), \tag{42}$$

which is a feasible solution to (36). Indeed, $f_\alpha(0) = (1 - \alpha)g(0) = (1 - \alpha)\beta = \delta$, by definition. Furthermore, invoking the estimate (33) with $r = 1/\alpha$, we have

$$\|f_\alpha\|_A \leq \frac{1}{1 - \alpha}\|f_\alpha\|_{H^\infty(D/\alpha)} = \|g\|_{H^\infty(D)} = 1. \tag{43}$$

Setting $\eta = 1 - \alpha\epsilon$ and $r = \frac{\eta}{\alpha\bar{\epsilon}} > 1$ since $\alpha < 1$, we have again from the reverse estimate (33)

$$\|\Phi f_\alpha\|_A \leq \frac{1}{1 - \frac{1}{r}}\|\Phi f_\alpha\|_{H^\infty(rD)} = \frac{1 - \alpha}{1 - \frac{1}{r}}\|g\|_{H^\infty(\alpha\epsilon + \alpha\bar{\epsilon}rD)} = (1 - \alpha\epsilon)\|g\|_{H^\infty(D_\eta)}. \tag{44}$$

14

By (41) and let $\alpha = 1 - b$, we have

$$\|\Phi f_\alpha\|_A \le (1 - \alpha\epsilon)\beta^{\frac{\bar{\eta}}{\eta}} = (1 - \alpha\epsilon)\left(\frac{\delta}{1-\alpha}\right)^{\frac{\alpha\epsilon}{1-\alpha\epsilon}} \le \exp\left\{\frac{\epsilon}{1-\epsilon}(1-b)\log\frac{\delta}{b}\right\} \qquad (45)$$

Choosing $b = \frac{1}{\log\frac{1}{\delta}}$ fulfills (40) and yields the upper bound in (30). ∎

Remark 11 *We compare the proof techniques of the lower bound part of Proposition 9 with that of Moitra and Saks (2013). The sample complexity bound (3) is also obtained by bounding the value of the dual LP from above via the H^∞-norm relaxation. The suboptimality of the exponent $\frac{1}{1-\epsilon}\log\frac{2}{1-\epsilon}$ in (3) seems to stem from the application of the Hadamard three-circle theorem, which is applicable to three concentric circles centered at the origin. In comparison, the sharp result in Proposition 9 is obtained by comparing the value of any feasible f on three horocircles (see Fig. 1), which are images of three horizontal lines under the Möbius transform and Hadamard three-lines theorem is readily applicable yielding the optimal exponent $\frac{\epsilon}{1-\epsilon}$. The upper bound part of Proposition 9 is new.*

3.2. Tight statistical lower bounds

Based on the general theory in Theorem 4, to prove the minimax lower bound announced in Theorem 1, there are two things to fix:

(a) why numbers of 0's provides negligible information for estimating the probability of the zero vector;

(b) how to fix the square-root gap in the general lower bound in Theorem 4.

We start with the second task. Recall the relation between the total variation and the Hellinger distance Tsybakov (2009)

$$\frac{1}{2}H^2 \le \mathrm{TV} \le H\sqrt{1 - H^2/4}. \qquad (46)$$

and the tensorization property of the Hellinger distance

$$H^2(P^{\otimes n}, Q^{\otimes n}) = 2 - 2(1 - H^2(P, Q)/2)^n.$$

Applying both to the original Le Cam's method (22), after simple algebra we get the following minimax lower bound (still for estimators based on the number of ones only):

$$\sqrt{R_n} \ge \frac{1}{4}\max_{\pi,\pi'}\left\{\pi(0) - \pi'(0) : H^2(\Phi\pi, \Phi\pi') \le \frac{1}{2n}\right\}. \qquad (47)$$

for any constant C, where Φ is given in (26), i.e. $\Phi\pi = \mathbb{E}_{\theta\sim\pi}[\mathrm{Bin}(\theta, \bar{\epsilon})]$.

It remains to show that for specific models, e.g., lossy population recovery, we have the following locally quadratic-like behavior: given the optimal Δ to (28), one can find feasible π and π' for (47) so that $\pi - \pi' \approx \Delta$ and $H^2(\Phi\pi, \Phi\pi') \lesssim \mathrm{TV}(\Phi\pi, \Phi\pi')^2$, such that the lower bound in (46) is essentially tight. This is done in the next lemma. To construct the pair

π and π', the main idea is to perturb a fixed distribution μ by $\pm\Delta$. In this case it turns out the center μ can be chosen to be a geometric distribution. Furthermore, recall that the near-optimal solution used in Proposition 9 deals with infinite sequence Δ in (28), which, in the context of population recovery, corresponds to input strings of infinite length.[10] It turns out that it suffices to consider $d = \Omega(\log^2 \frac{1}{\delta})$.

Lemma 12 *Fix $\epsilon > \frac{1}{2}$. For every $\delta < \frac{1}{2e}$, there exists a pair of probability distributions π and π' on $\mathbb{Z}_+$ such that $|\pi(0) - \pi'(0)| \geq \delta$ and*

$$H^2(\Phi\pi, \Phi\pi') \leq C \left(e\delta \log \frac{1}{\delta} \right)^{\frac{2\epsilon}{1-\epsilon}}, \tag{48}$$

where $C = 4$. Furthermore, if $C = 36$, both distributions can be picked to be supported on $\{0, \ldots, d\}$ provided that $d \geq \frac{2\epsilon}{1-\epsilon} \log^2 \frac{1}{\delta}$.

Proof Let $g, \beta, \alpha, b, \eta, r$ be as in the proof of Proposition 9. In particular, $\alpha = 1 - \frac{1}{\log \frac{1}{\delta}}$, $\eta = 1 - \alpha\epsilon$, $\beta = \delta \log \frac{1}{\delta}$, and $r = \frac{1-\alpha\epsilon}{\alpha\bar{\epsilon}}$. Set

$$f(z) = \bar{\alpha}g(\alpha z) - \bar{\alpha}g(\alpha).$$

Let $\Delta_k \triangleq [z^k] f(z)$. Then $\Delta_0 = \bar{\alpha}g(0) - \bar{\alpha}g(\alpha) = \delta - (\delta \log \frac{1}{\delta})^{2\log \frac{1}{\delta}-1}$ and $\Delta_k = \bar{\alpha}\alpha^k[z^k]f(z)$. Since $\delta \leq 1/(2e)$ by assumption, we have

$$\delta > \Delta_0 \geq \frac{\delta}{2}.$$

Furthermore, recall from the proof of Proposition 9 that $\|g\|_{H^\infty(D)} = 1$ and thus

$$|\Delta_k| \leq \bar{\alpha}\alpha^k, \qquad k \geq 1. \tag{49}$$

Consider the following geometric distribution μ on $\mathbb{Z}_+$:

$$\mu(k) \triangleq \bar{\alpha}\alpha^k.$$

Define now π and π' via

$$\pi(k) \triangleq \mu(k) + \Delta_k, \quad \pi'(k) \triangleq \mu(k) - \Delta_k.$$

Note that $\mu(0) = \bar{\alpha} = \frac{1}{\log \frac{1}{\delta}} \geq \delta > \Delta_0$, which implies $\pi(0) \geq 0$ and $\pi'(0) \geq 0$. Furthermore, from (49) we get that $\pi(k), \pi'(k) \geq 0$ for all $k \in \mathbb{Z}_+$. Since $f(1) = \sum_k \Delta_k = 0$ we conclude that π and π' are indeed probability distributions satisfying

$$\pi(0) - \pi'(0) \geq \delta.$$

---

10. For the simple case of $\epsilon \leq 1/2$, the construction in Proposition 9 is a degree-1 polynomial. This amounts to considering a single bit and using a pair of Bernoulli distributions to establish the optimality of the parametric rate $1/n$, which is standard.

Next, notice that since $(\sqrt{1+r} - \sqrt{1-r})^2 \le 2r^2$ for all $r \in [0,1]$ we get

$$H^2(\Phi\pi, \Phi\pi') = \sum_{k \ge 0} \left(\sqrt{\Phi\mu(k) + \Phi\Delta(k)} - \sqrt{\Phi\mu(k) - \Phi\Delta(k)} \right)^2 \le 2 \sum_{k \ge 0} \frac{\Phi\Delta(k)^2}{\Phi\mu(k)} . \qquad (50)$$

Elementary calculation (e.g. from (35) and $\sum_k z^k \mu(k) = \frac{1-\alpha}{1-\alpha z}$) shows that

$$\Phi\mu(k) = \left(1 - \frac{1}{r} \right) r^{-k} . \qquad (51)$$

We also know from the proof of Proposition 9 and $g(\alpha) = \beta^{\frac{1+\alpha}{1-\alpha}} \le \beta^{\frac{\bar{\eta}}{\eta}}$ that

$$\|\Phi f\|_{H^\infty(rD)} \le \bar{\alpha}\beta^{\frac{\bar{\eta}}{\eta}} + \bar{\alpha}g(\alpha) \le 2\bar{\alpha}\beta^{\frac{\bar{\eta}}{\eta}} .$$

Therefore, from (34) we get

$$|\Phi\Delta(k)| = |[z^k]f(\bar{\epsilon}z + \epsilon)| \le 2\bar{\alpha}\beta^{\frac{\bar{\eta}}{\eta}}r^{-k} . \qquad (52)$$

Plugging (51) and (52) into (50) we obtain as in (45)

$$H^2(\Phi\pi, \Phi\pi') \le 4\bar{\alpha}^2\beta^{2\bar{\eta}/\eta}(1 - 1/r)^{-2} = 4(1 - \alpha\epsilon)^2\beta^{2\bar{\eta}/\eta} \le 4\left(e\delta \log \frac{1}{\delta} \right)^{\frac{2\epsilon}{1-\epsilon}} . \qquad (53)$$

To prove the second part, we replace π, π' by their conditional version, denote by $\tilde{\pi}(k) = \frac{\pi(k)}{\pi([0,d])}$ for $k \le d$ and similarly for $\tilde{\pi}'$. Then

$$\pi((d,\infty)) + \pi'((d,\infty)) \le 2\mu((d,\infty)) = 2 \sum_{k>d} \bar{\alpha}\alpha^k = 2\alpha^{d+1} \le 2\exp\left(-\frac{d}{\log\frac{1}{\delta}} \right) \le 2\delta^{\frac{2\epsilon}{1-\epsilon}}, \qquad (54)$$

where the last inequality holds provided that $d \ge \frac{2\epsilon}{1-\epsilon} \log^2 \frac{1}{\delta}$. For Hellinger,

$$H(\Phi\pi, \Phi\pi') \ge H(\Phi\tilde{\pi}, \Phi\tilde{\pi}') - H(\Phi\pi, \Phi\tilde{\pi}) - H(\Phi\pi', \Phi\tilde{\pi}') \qquad (55)$$

$$\ge H(\Phi\tilde{\pi}, \Phi\tilde{\pi}') - H(\pi, \tilde{\pi}) - H(\pi', \tilde{\pi}') \qquad (56)$$

$$= H(\Phi\tilde{\pi}, \Phi\tilde{\pi}') - \sqrt{2 - 2\sqrt{\pi([0,d])}} - \sqrt{2 - 2\sqrt{\pi'([0,d])}} \qquad (57)$$

$$\ge H(\Phi\tilde{\pi}, \Phi\tilde{\pi}') - 4\delta^{\frac{\epsilon}{1-\epsilon}} , \qquad (58)$$

where (55) is from the triangle inequality for Hellinger distance, (56) is from the data processing inequality for the latter, (57) is explicit computation and finally in (58) we used (54) and the fact that $\sqrt{2 - 2\sqrt{1-x}} \le \sqrt{2x}$ for all $0 \le x \le 1$. In view of (53), this completes the proof of (48) with $C = 36$. ∎

Finally, we put everything together.

Proof [Proof of Theorem 1] The upper-bound on the sample complexity follows from Theorem 4 and Proposition 9.

17

For the lower bound, we need to show that the number of zeros carries almost no information. This is done by dimension expansion. Indeed, set $d' \gg d$ and extend π and π' to distributions on $\{0, \ldots, d'\}$ by zero-padding. The intuition is that if the input vector contains at most d ones, then the number of zeros of the output is distributed approximately as $\mathrm{Bin}(d', \bar{\epsilon})$, almost independent of the input. We make this idea precise next.

Fix $\delta = 4\delta' > 0$, set $d = \frac{2\epsilon}{1-\epsilon} \log^2 \frac{1}{\delta}$ and consider probability distributions π and π' constructed in Lemma 12 with $H^2(\Phi\pi, \Phi\pi') \leq h_1$, where

$$h_1 = C \left(e\delta \log \frac{1}{\delta} \right)^{\frac{2\epsilon}{1-\epsilon}}. \tag{59}$$

Take $d' = \frac{16\bar{\epsilon}d^2}{\epsilon h_1}$ and note that according to Lemma 21 in Appendix B we have

$$H^2(\mathrm{Bin}(d' - s, \bar{\epsilon}), \mathrm{Bin}(d', \bar{\epsilon})) \leq \frac{4\bar{\epsilon}d^2}{\epsilon d'} = \frac{h_1}{4} \qquad \forall s \in \{0, \ldots, d\}. \tag{60}$$

It suffices to show

$$H^2(\mathbb{E}_{\theta\sim\pi}[\mathrm{Bin}(\theta, \bar{\epsilon}) \otimes \mathrm{Bin}(d' - \theta, \bar{\epsilon})], \mathbb{E}_{\theta\sim\pi'}[\mathrm{Bin}(\theta, \bar{\epsilon}) \otimes \mathrm{Bin}(d' - \theta, \bar{\epsilon})]) \leq 4h_1, \tag{61}$$

Indeed, assuming (61) we can conclude from (47) that

$$n_{\mathsf{L}}^*(\delta', d') \geq \frac{1}{8h_1},$$

which in view of (59) is precisely the statement of the theorem. To show (61) consider the following chain:

$$H(\mathbb{E}_{\theta\sim\pi}[\mathrm{Bin}(\theta, \bar{\epsilon}) \otimes \mathrm{Bin}(d' - \theta, \bar{\epsilon})], \mathbb{E}_{\theta\sim\pi'}[\mathrm{Bin}(\theta, \bar{\epsilon}) \otimes \mathrm{Bin}(d' - \theta, \bar{\epsilon})])$$

$$\leq H(\mathbb{E}_{\theta\sim\pi}[\mathrm{Bin}(\theta, \bar{\epsilon})], \mathbb{E}_{\theta\sim\pi'}[\mathrm{Bin}(\theta, \bar{\epsilon})]) + \sqrt{h_1} \tag{62}$$

$$= H(\Phi\pi, \Phi\pi') + \sqrt{h_1} \leq 2\sqrt{h_1} \tag{63}$$

where (62) is an application of the following Lemma with $\tau = \sqrt{h_1}/2$. ∎

Lemma 13 *Suppose there exists Q^* such that $H(Q^*, Q_\theta) \leq \tau$ for any $\theta \in \Theta$. Then for any distributions π and π' on Θ,*

$$H(\mathbb{E}_{\theta\sim\pi}[P_\theta \otimes Q_\theta], \mathbb{E}_{\theta\sim\pi'}[P_\theta \otimes Q_\theta]) \leq H(\mathbb{E}_{\theta\sim\pi} P_\theta, \mathbb{E}_{\theta\sim\pi'} P_\theta) + 2\tau.$$

Proof The triangle inequality yields

$$H(\mathbb{E}_{\theta\sim\pi}[P_\theta \otimes Q_\theta], \mathbb{E}_{\theta\sim\pi'}[P_\theta \otimes Q_\theta]) \leq H(\mathbb{E}_{\theta\sim\pi}[P_\theta \otimes Q^*], \mathbb{E}_{\theta\sim\pi'}[P_\theta \otimes Q^*])$$
$$+ H(\mathbb{E}_{\theta\sim\pi}[P_\theta \otimes Q_\theta], \mathbb{E}_{\theta\sim\pi}[P_\theta \otimes Q^*])$$
$$+ H(\mathbb{E}_{\theta\sim\pi'}[P_\theta \otimes Q_\theta], \mathbb{E}_{\theta\sim\pi'}[P_\theta \otimes Q^*]).$$

Here

$$H(\mathbb{E}_{\theta\sim\pi}[P_\theta \otimes Q^*], \mathbb{E}_{\theta\sim\pi'}[P_\theta \otimes Q^*]) = H(\mathbb{E}_{\theta\sim\pi}[P_\theta] \otimes Q^*, \mathbb{E}_{\theta\sim\pi'}[P_\theta] \otimes Q^*) = H(\mathbb{E}_{\theta\sim\pi}[P_\theta], \mathbb{E}_{\theta\sim\pi'}[P_\theta])$$

and, by convexity, $H(\mathbb{E}_{\theta\sim\pi}[P_\theta \otimes Q_\theta], \mathbb{E}_{\theta\sim\pi'}[P_\theta \otimes Q^*]) \leq \mathbb{E}_{\theta\sim\pi} H(Q_\theta, Q^*) \leq \tau$. ∎

4. Noisy population recovery

In this section we bound the value of the LP (11) for noisy population recovery. As previously mentioned in Remark 6, since in this case the sample complexity turns out to grow super-polynomially, the general result in Theorem 4 suffices to produce the sample complexity bound announced in Theorem 3 and there is no need to consider refined Hellinger-based minimax lower bound as developed in Section 3.2 for lossy population recovery.

We begin by specializing the LP in (11) to the noisy population recovery setting with error probability ϵ, where the transition matrix Φ is given by (13) throughout this section. Recall from (13) that conditioned on the input Hamming weight $w(X) = j$, the output Hamming weight $w(Y)$ is distributed as the convolution $\mathrm{Bin}(j, \bar{\epsilon}) * \mathrm{Bin}(d - j, \epsilon)$. Thus, similarly to (28), we consider the equivalent formulation:

$$t(\delta, d) \triangleq \min_{\Delta \in \mathbb{R}^{d+1}} \left\{ \left\| \sum_{j=0}^{d} \Delta_j \mathrm{Bin}(j, \bar{\epsilon}) * \mathrm{Bin}(d - j, \epsilon) \right\|_1 : \Delta_0 \geq \delta, \|\Delta\|_1 \leq 1 \right\}. \tag{64}$$

Since we can always negate the observed bits, in the sequel we shall assume, without loss of generality, that

$$\epsilon < 1/2.$$

To recast (64) as an optimization problem in terms of functions (polynomials), we note that the channel Φ maps the generating function of the input weight to that of the output as follows:

$$\mathbb{E}[z^{w(X)}] \overset{\Phi}{\mapsto} \mathbb{E}[z^{w(Y)}] = \mathbb{E}[z^{\mathrm{Bin}(w(X), \bar{\epsilon}) + \mathrm{Bin}(d - w(X), \epsilon)}] = \mathbb{E}[(\bar{\epsilon}z + \epsilon)^{w(X)}(\epsilon z + \bar{\epsilon})^{d - w(X)}]$$

that is,

$$f(z) \mapsto (\Phi f)(z) \triangleq f\left(\frac{\bar{\epsilon}z + \epsilon}{\epsilon z + \bar{\epsilon}} \right) (\epsilon z + \bar{\epsilon})^d. \tag{65}$$

Thus, each degree-d polynomial is mapped to another via

$$\sum_{i=0}^{d} a_i z^i \mapsto \sum_{i=0}^{d} a_i (\bar{\epsilon}z + \epsilon)^i (\epsilon z + \bar{\epsilon})^{d-i}.$$

Therefore, using the A-norm introduced in Section 3.1, we have:

$$t(\delta, d) = \min \left\{ \left\| f\left(\frac{\bar{\epsilon}z + \epsilon}{\epsilon z + \bar{\epsilon}} \right) (\epsilon z + \bar{\epsilon})^d \right\|_A : f(0) \geq \delta, \|f\|_A \leq 1, \deg f \leq d \right\}. \tag{66}$$

As usual, by choosing $f \equiv \delta$, we have the trivial bound

$$t(\delta, d) \leq \delta. \tag{67}$$

The next proposition provides a sharp characterization:

Proposition 14

Assume that $\epsilon < \frac{1}{2}$ and $d \geq 1$. Define

$$\mu(\epsilon) \triangleq \frac{\epsilon(1-\epsilon)}{(1-2\epsilon)^2}. \tag{68}$$

There exist absolute constant c, c' such that the following hold: For all $\delta < 1$,

$$t(\delta, d) \geq \exp\left\{-c\left(d\mu(\epsilon)\log^2\frac{e}{\delta}\right)^{1/3}\right\}. \tag{69}$$

For all $\delta < 1/3$,

$$t(\delta, d) \leq \max\left\{\exp(-(1-2\epsilon)^2 d), \exp\left(-c'\left(d\mu(\epsilon)\log^2\frac{1}{\delta}\right)^{1/3}\right)\right\}. \tag{70}$$

Again, thanks to Theorem 4, (69) and (70) translate into the sample complexity lower and upper bound in Theorem 3, respectively.

The rest of this section is devoted to proving Proposition 14 using H^∞-relaxations (Section 4.1), with upper and lower bound shown in Section 4.2 and Section 4.3, respectively.

4.1. Two H^∞-relaxations

Compared to the analysis of lossy population recovery, it turns out that the upper estimate (33) on A-norm using H^∞-norm over a larger disk, which we relied on in Section 3.1 to deal with the composition operator (35), is not useful for the more complicated operator (65) and more intricate analysis is thus required. Specifically, we need to consider two H^∞-relaxations of (66), the latter of which retains the A-norm constraint:

$$t_1(\delta, d) = \min\left\{\left\|f\left(\frac{\bar{\epsilon}z + \epsilon}{\epsilon z + \bar{\epsilon}}\right)(\epsilon z + \bar{\epsilon})^d\right\|_{H^\infty(D)} : f(0) \geq \delta, \|f\|_{H^\infty(D)} \leq 1, \deg f \leq d\right\} \tag{71}$$

$$t_2(\delta, d) = \min\left\{\left\|f\left(\frac{\bar{\epsilon}z + \epsilon}{\epsilon z + \bar{\epsilon}}\right)(\epsilon z + \bar{\epsilon})^d\right\|_{H^\infty(D)} : f(0) \geq \delta, \|f\|_A \leq 1, \deg f \leq d\right\}. \tag{72}$$

Since $\|p\|_{H^\infty(D)} \leq \|p\|_A \leq d\|p\|_{H^\infty(D)}$ for any degree-d polynomials p, we have, in turn,

$$t_1(\delta, d) \leq t(\delta, d) \leq dt_2(\delta, d). \tag{73}$$

Next we simplify (71) and (72) via a change of variable. note that as long as $\epsilon \neq \frac{1}{2}$, the Möbius transform $z \mapsto \frac{\bar{\epsilon}z + \epsilon}{\epsilon z + \bar{\epsilon}}$ maps the unit circle to itself. Therefore

$$\left\|f\left(\frac{\bar{\epsilon}z + \epsilon}{\epsilon z + \bar{\epsilon}}\right)(\epsilon z + \bar{\epsilon})^d\right\|_{H^\infty(D)} \overset{(31)}{=} \sup_{z \in \partial D}\left|f\left(\frac{\bar{\epsilon}z + \epsilon}{\epsilon z + \bar{\epsilon}}\right)(\epsilon z + \bar{\epsilon})^d\right|$$

$$= \sup_{w \in \partial D}|f(w)h(w)| \overset{(31)}{=} \|fh\|_{H^\infty(D)},$$

where

$$h(w) \triangleq \left(\frac{1 - 2\epsilon}{\bar{\epsilon} - \epsilon w}\right)^d$$

20

is analytic on $\bar{D}$ since $\epsilon < 1/2$ by assumption. Hence (71) and (72) are equivalent to

$$t_1(\delta, d) = \min\left\{\|fh\|_{H^\infty(D)} : f(0) = \delta, \|f\|_{H^\infty(D)} \leq 1, \deg f \leq d\right\}, \tag{74}$$

$$t_2(\delta, d) = \min\left\{\|fh\|_{H^\infty(D)} : f(0) = \delta, \|f\|_A \leq 1, \deg f \leq d\right\}. \tag{75}$$

4.2. Lower bound on t_1

Rewriting $h(z)$ as $h(z) = \frac{1}{(1-c(z-1))^d}$, where $c \triangleq \frac{\epsilon}{1-2\epsilon}$, we have for all $\theta \in [-\pi, \pi]$,

$$|h(e^{i\theta})| = (1 + 2c(1+c)(1 - \cos\theta))^{-d/2} \geq (1 + c(1+c)\theta^2)^{-d/2} \geq e^{-\mu d\theta^2/2}$$

where $\mu = \mu(\epsilon)$ is defined in (68). Next, from (Borwein and Erdélyi, 1997, Corollary 3.2) we conclude that for any feasible f for (74), we have

$$\sup_{\theta \in (-a/2, a/2)} |f(e^{i\theta})| \geq (\delta/e)^{c_1/a},$$

where c_1 is an absolute constant. Thus,

$$\|fh\|_{H^\infty(D)} \geq \sup_{\theta \in (-a/2, a/2)} |f(e^{i\theta})||h(e^{i\theta})| \geq e^{-\mu da^2/2 - \frac{c_1}{a}\log\frac{e}{\delta}}.$$

Setting $a = \left(\frac{2c_1 \log\frac{e}{\delta}}{\mu d}\right)^{\frac{1}{3}}$ we get the following non-asymptotic bound:

$$t_1(\delta, d) \geq \exp\left\{-c_2\left(\mu(\epsilon)d\log^2\frac{e}{\delta}\right)^{1/3}\right\},$$

where c_2 is an absolute constant.

4.3. Upper bound on t_2

In view of (73), we will show that for all $\epsilon < 1/2$, $\delta \in (0, 1/3)$ and all $d \in \mathbb{N}$ we have

$$t(\delta, d) \leq \max\left\{\exp(-(1 - 2\epsilon)^2 d), \left(\frac{d^4\mu}{\log^4\frac{1}{\delta}}\right)^{1/3}\exp\left\{-c'\left(d\mu\log^2\frac{1}{\delta}\right)^{1/3}\right\}\right\}. \tag{76}$$

for some absolute constant c'.

First consider $d \leq \frac{1}{(1-2\epsilon)^2}\log\frac{1}{\delta}$. By the trivial bound (67), we have

$$t(\delta, d) \leq \exp(-(1 - 2\epsilon)^2 d). \tag{77}$$

In the sequel we shall assume that

$$d \geq \frac{1}{(1 - 2\epsilon)^2}\log\frac{1}{\delta}. \tag{78}$$

To construct a near optimal solution for (75), we modify the feasible solution previously used in the proof of Proposition 9. Let

$$g(z) = (1 - z)^2\delta^{\frac{1+z}{1-z}}, \tag{79}$$

21

and $f(z) = g(\alpha z)$ with $\alpha \in (0, 1)$ to be chosen. Let $\tilde{g}$ be the degree-d truncation of the Taylor expansion of $g(z) = \sum_n a_n z^n$, and $\tilde{f}(z) = \tilde{g}(\alpha z)$. Note that $\tilde{f}(0) = \tilde{g}(0) = g(0) = \delta$. Next, instead of invoking (33) which estimates the A-norm by the H^∞-norm over a bigger disk and turns out to be too loose here, the next lemma (proved in Appendix B) uses the H^∞-norm of the derivative:

Lemma 15 *For any $\delta > 0$, $\|g\|_A \le A(\delta) \triangleq 4(\log \frac{1}{\delta} + 3)$.*

Therefore

$$\|\tilde{f}\|_A = \sum_{n=0}^{d} \alpha^n |a_n| \le \sum_{n \ge 0} |a_n| = \|g\|_A \le A(\delta).$$

By Cauchy's inequality (34), the Taylor series coefficients of f satisfy $|a_n| \le \|g\|_{H^\infty(D)} = 1$ and hence

$$\|\tilde{f} - f\|_A \le \frac{\alpha^{d_0+1}}{1 - \alpha}. \tag{80}$$

Next we bound the objective function. Rewriting $h(z)$ as $h(z) = \frac{1}{(1 - c(z-1))^d}$, where $c \triangleq \frac{\epsilon}{1 - 2\epsilon}$, we recall that

$$\|h\|_{H^\infty(D)} = 1.$$

Thus, we have

$$\begin{aligned}
\|\tilde{f}h\|_{H^\infty(D)} &\le \|fh\|_{H^\infty(D)} + \|(f - \tilde{f})h\|_{H^\infty(D)} \\
&\overset{(80)}{\le} \|g(\alpha z)h(z)\|_{H^\infty(D)} + \frac{\alpha^{d+1}}{1 - \alpha} \\
&\le 4\|\delta^{\frac{1+\alpha z}{1-\alpha z}} h(z)\|_{H^\infty(D)} + \frac{\alpha^{d+1}}{1 - \alpha},
\end{aligned} \tag{81}$$

where in the last step we used $\|(1 - \alpha z)^2\|_{H^\infty(D)} \le 4$. The first term is bounded by the following lemma (proved in Appendix B):

Lemma 16 *If*

$$d \ge 2 \log \frac{1}{\delta} \max\left\{ \frac{1}{\mu(1 - \alpha)^3}, \frac{1}{1 - \alpha} \right\} \tag{82}$$

where $\mu = \mu(\epsilon)$ is defined in (68), then

$$\|\delta^{\frac{1+\alpha z}{1-\alpha z}} h(z)\|_{H^\infty(D)} = \delta^{\frac{1+\alpha}{1-\alpha}}. \tag{83}$$

Finally, set

$$\alpha = 1 - \left(\frac{2}{d\mu} \log \frac{1}{\delta} \right)^{1/3}, \tag{84}$$

which, in view of the assumption (78), fulfills the condition (82) in the Lemma 16. Combining (81) and (83) yields $\|\tilde{f}h\|_{H^\infty(D)} \le \delta^{\frac{1+\alpha}{1-\alpha}} + \frac{\alpha^{d+1}}{1-\alpha}$. Since $\alpha^d \le \exp(-(1-\alpha)d)$, the assumption (78) together with (84) implies that $d \ge \log \frac{1}{\delta}$ and further $\alpha^d \le \delta^{\frac{1}{1-\alpha}}$. Hence

$$\|\tilde{f}h\|_{H^\infty(D)} \le \frac{1}{1-\alpha} \delta^{\frac{1}{1-\alpha}} = \left(\frac{d\mu}{2 \log \frac{1}{\delta}} \right)^{1/3} \exp\left\{ -\left(\frac{d\mu}{2} \log^2 \frac{1}{\delta} \right)^{1/3} \right\}. \tag{85}$$

In summary, for any $\delta \in (0,1)$, we have constructed $\tilde{f}$ such that $\tilde{f}(0) = \delta$, $\|\tilde{f}\|_A \leq A(\delta) = 4(\log \frac{1}{\delta} + 3)$ and $\|\tilde{f}h\|_{H^\infty(D)}$ is bounded by (85). Rescaling by $A(\delta)$, we conclude there exists a universal constant c_2, such that for any $\delta' = \frac{\delta}{A(\delta)} \in (0, \frac{1}{3})$,

$$t_2(\delta', d) \leq c_2 \left(\frac{d\mu}{\log^4 \frac{1}{\delta'}} \right)^{1/3} \exp\left\{ -c_2 \left(d\mu \log^2 \frac{1}{\delta'} \right)^{1/3} \right\},$$

which, in view of (73), yields (76).

5. Smoothed estimators for lossy population recovery

In this section, we construct an explicit estimator for lossy population recovery that is optimal up to a factor of 2 in the exponent. We start from the unbiased estimator for P_0 and then modify it via the *smoothing technique* proposed in Orlitsky et al. (2016).

Recall that for the linear estimator g in (4), its bias is bounded by $\|\Phi^\top g - e_0\|_\infty$ and the standard deviation is at most $\frac{1}{\sqrt{n}}\|g\|_\infty$, where the matrix Φ is given by (26) which is an upper triangular matrix with non-zero diagonals. As mentioned in Section 1, the unique unbiased estimator is a linear estimator with coefficients $g^{\mathrm{u}} = (\Phi^\top)^{-1}e_0$. Direct calculation shows that $g_j^{\mathrm{u}} = (-\frac{\epsilon}{\bar{\epsilon}})^j$ for $j = 0, 1, \ldots, d$.[11]

Note that $\|g^{\mathrm{u}}\|_\infty = \max(1, (\epsilon/\bar{\epsilon})^{d+1})$. Hence, for $\epsilon \leq 1/2$, $\|g^{\mathrm{u}}\|_\infty = 1$ and the unbiased estimator has a sample complexity at most $O(\delta^{-2})$. For $\epsilon > 1/2$, the coefficients increase exponentially in d which results in high variance. To alleviate this issue, we modify the estimator via the smoothing technique proposed in Orlitsky et al. (2016). The main idea of smoothing is to introduce an independent random integer L, truncate the unbiased estimator after the Lth term, and average the truncated estimators according to an appropriately chosen distribution that aims to balance the bias and variance. Equivalently, this amounts to multiplying the coefficients of the unbiased estimator with a tail probability that modulates the exponential growth. To this end, define the smoothed estimator $\widehat{P}_0^{\mathrm{u}}$ as a linear estimator (4) with the coefficient g given by

$$g_i^{\mathrm{s}} \triangleq g_i^{\mathrm{u}} \cdot \mathbb{P}(L \geq i).$$

The following theorem gives the sample complexity guarantee for Poisson smoothing.

Theorem 17 *Let $\epsilon > 1/2$. Let L be Poisson distributed with mean $\lambda = \frac{1-\epsilon}{3\epsilon-1} \log n$. Then*

$$\sup_{P \in \mathcal{P}_d} \mathbb{E}_P[(\widehat{P}_0^{\mathrm{u}} - P_0)^2] \leq 4n^{-\frac{1-\epsilon}{3\epsilon-1}}.$$

Therefore, the sample complexity of the smoothed estimator is at most $O\left(\delta^{-\frac{2(3\epsilon-1)}{1-\epsilon}}\right)$.

Proof We first bound the variance by the moment generating function of L. Observe that

$$\|g^{\mathrm{s}}\|_\infty = \max_i |g_i^{\mathrm{u}}| \cdot \mathbb{P}(L \geq i) = \max_i (\epsilon/\bar{\epsilon})^i \cdot \mathbb{P}(L \geq i) \leq \mathbb{E}_L[(\epsilon/\bar{\epsilon})^L] = e^{\lambda \frac{2\epsilon-1}{1-\epsilon}}, \qquad (86)$$

---

11. This can also be obtained from consider the inverse operator of (35), which is again a composition operator $(\Phi^{-1}f)(z) = f((z - \epsilon)/\bar{\epsilon})$.

where the inequality follows from the assumption that $\epsilon/\bar\epsilon > 1$. To bound the bias term, note that $(\Phi^\top g^{\mathrm s})_0 = 1$ and hence

$$\|\Phi^\top g^{\mathrm s} - e_0\|_\infty \le \max_{1 \le j \le d} |(\Phi^\top g^{\mathrm s})_j|. \tag{87}$$

For any $j > 0$,

$$(\Phi^\top g^{\mathrm s})_j = \sum_{i=0}^d \Phi_{ij} g_i^{\mathrm s} = \sum_{i=0}^j \binom{j}{i} \epsilon^j (-1)^i \mathbb{P}(L \ge i) \triangleq \epsilon^j f(j). \tag{88}$$

Observe that

$$
\begin{aligned}
f(j) &= \sum_{i=0}^j \binom{j}{i} (-1)^i \mathbb{P}(L \ge i) \\
&= \sum_{i=0}^j \left(\binom{j-1}{i-1} + \binom{j-1}{i} \right) (-1)^i \mathbb{P}(L \ge i) \\
&= \sum_{i=0}^j \binom{j-1}{i-1} (-1)^i \mathbb{P}(L \ge i) + \sum_{i=0}^j \binom{j-1}{i} (-1)^i \mathbb{P}(L \ge i) \\
&= \sum_{i=0}^{j-1} \binom{j-1}{i} (-1)^{i+1} \mathbb{P}(L \ge i+1) + \sum_{i=0}^{j-1} \binom{j-1}{i} (-1)^i \mathbb{P}(L \ge i) \\
&= \sum_{i=0}^{j-1} \binom{j-1}{i} (-1)^i \mathbb{P}(L = i).
\end{aligned}
$$

If $L \sim \mathrm{Poi}(\lambda)$, then

$$f(j) = \sum_{i=0}^{j-1} \binom{j-1}{i} (-1)^i \mathbb{P}(L = i) = \sum_{i=0}^{j-1} \binom{j-1}{i} (-1)^i e^{-\lambda} \frac{\lambda^i}{i!} = e^{-\lambda} L_{j-1}(\lambda),$$

where $L_k(\lambda) = \sum_{i=0}^k \frac{(-\lambda)^i}{i!} \binom{k}{i}$ is the Laguerre polynomial of degree k. Since $|L_k(\lambda)| \le e^{\lambda/2}$ for all $k \ge 0$ cf. (Abramowitz and Stegun, 1964, 22.14.12), we have $|f(j)| \le e^{-\lambda/2}$ for all $j \ge 1$. Hence by (87) and (88),

$$\|\Phi^\top g^{\mathrm s} - e_0\|_\infty \le e^{-\lambda/2}.$$

Combining the above equation with (86) yields

$$\|\Phi^\top g^{\mathrm s} - e_0\|_\infty + \frac{\|g^{\mathrm s}\|_\infty}{\sqrt n} \le e^{-\lambda/2} + \frac{e^{\lambda \frac{2\epsilon - 1}{1 - \epsilon}}}{\sqrt n}.$$

The theorem follows by setting $\lambda = \frac{1-\epsilon}{3\epsilon - 1} \log n$. ∎

Appendix A. Converting individual recovery to population recovery

In this appendix we describe the algorithm in Dvir et al. (2012) that relates individual recovery to population recovery and converts any estimator for P_0 to a distribution estimator. The main ingredient is induction on the length d and the observation that for any x, one can convert any estimator $\widehat{P}_0$ for P_0 to one for P_x:

$$\widetilde{P}_x = \widehat{P}_0(Y_1 \oplus x, \dots, Y_n \oplus x),$$

which inherits the performance guarantee of $\widehat{P}_0$ under both lossy and noisy model.

Let xa denote the sequence obtained by concatenating sequences x and a. Consider the following algorithm:

- Input δ.

- Initialize: $\mathcal{S}_1 = \{0, 1\}$.

- For each i from 2 to d:

    - $\mathcal{S}_i \leftarrow \cup_{x \in \mathcal{S}_{i-1}}\{x0, x1\}$.
    - For each x in $\mathcal{S}_i$, compute the estimate $\widetilde{P}_x = \widehat{P}_0(Y_1 \oplus x, \dots, Y_n \oplus x)$.
    - Let $\mathcal{S}_i \leftarrow \mathcal{S}_i \setminus \{x : \widetilde{P}_x \leq 2\delta\}$.

- Assign $\widehat{P}_x = \widetilde{P}_x$ for all $x \in \mathcal{S}_d$ and $\widehat{P}_x = 0$ otherwise.

The following lemma, applicable to both lossy and noisy recovery, improves the original guarantee in Dvir et al. (2012) which depends on the support size of the distribution.

Lemma 18 *If there is an algorithm $\widehat{P}_0$ that estimates P_0 to an accuracy of δ using n samples with probability at least $1 - \alpha$, then there is an algorithm that with probability $\geq 1 - 9d\alpha/(4\delta)$ satisfies,*

$$\max_{x \in \{0,1\}^d} |P_x - \widehat{P}_x| \leq 4\delta.$$

The run time for the algorithm is $O(d/\delta \cdot (t + nd))$, where t is the time it takes to compute $\widehat{P}_0$.

Remark 19 *Lemma 18 shows that the sample complexity of population recovery (estimating all probabilities) is within a logarithmic factor of that of individual recovery (estimating $\widehat{P}_0$), namely, $n^*(d, \delta)$. To see this, consider an estimator $\widehat{P}_0$ that achieves $\mathbb{E}(\widehat{P}_0 - P_0)^2 \leq \delta^2$ so that $|\widehat{P}_0 - P_0| \leq 2\delta$ with probability at least $1/4$. By the usual split-sample-then-median method,[12] $O(n^*(\delta, d) \cdot \log \frac{d}{\delta\tau})$ samples suffices to boost the probability of estimating P_0 within δ to $1 - \frac{\delta\tau}{d}$, which, in view of Lemma 18, suffices to estimate all probabilities within δ with probability $1 - \tau$.*

---

12. That is, divide all the samples into $\log \frac{1}{\tau}$ batches, apply the same estimator to each batch and take the batchwise median.

Proof We first show that with high probability all sequences with probability $\geq 4\delta$ remain in $\mathcal{S}_d$. Observe that if a sequence x has probability $\geq 4\delta$, then all of its prefixes also has probability $\geq 4\delta$. By definition, $\mathbb{P}(|P_x - \widehat{P}_x| \geq 2\delta) \leq \alpha$. Hence, the probability that x is discarded is at most $\alpha(d-1)$ by the union bound and the probability that any such sequence is absent in $\mathcal{S}_d$ is at most $(d-1)\alpha/(4\delta)$.

We now show by induction that $|\mathcal{S}_d| \leq 1/\delta$ with probability $\geq 1 - 2(d-1)\alpha/\delta$. Suppose $|\mathcal{S}_{i-1}| \leq 1/\delta$ with probability $\geq 1 - 2(i-2)\alpha/\delta$, which holds for $i-1 = 1$. Then,

$$\mathbb{P}(\exists y \in \mathcal{S}_i : P_y \leq \delta) \leq \sum_{y \in \cup_{x \in \mathcal{S}_{i-1}}\{x0,x1\}} \mathbb{P}(\widehat{P}_y > 2\delta, P_y \leq \delta)$$

$$\overset{(a)}{\leq} \sum_{y \in \cup_{x \in \mathcal{S}_{i-1}}\{x0,x1\}} \alpha$$

$$= 2\alpha|\mathcal{S}_{i-1}|$$

where (a) follows from the assumption that $\mathbb{P}(|\widehat{P}_y - P_y| \leq \delta) \geq 1 - \alpha$ for all y. Hence, by the inductive hypothesis and the union bound, the probability that $|\mathcal{S}_i|$ exceeds $1/\delta$ is at most $2(i-1)\alpha/\delta$.

Combing the above two results, we get that with probability $\leq 9(d-1)\alpha/(4\delta)$, $\mathcal{S}_d$ contains all symbols with probability $\geq 4\delta$ and $|\mathcal{S}_d| \leq 1/\delta$. Conditioned on this event, with probability at least $1 - \alpha/\delta$, $|P_x - \widehat{P}_x| \leq \delta$, for all $x \in \mathcal{S}_d$. Hence,

$$\max_{x \in \{0,1\}^d} |P_x - \widehat{P}_x| = \max\left(\max_{x \in S_d} |P_x - \widehat{P}_x|, \max_{x \notin S_d} P_x\right) \leq \max\left(\max_{x \in S_d} |P_x - \widehat{P}_x|, 4\delta\right) \leq 4\delta.$$

By the union bound the total error probability is $9(d-1)\alpha/(4\delta) + \alpha/\delta$. ■

Appendix B. Proofs of auxiliary results

Proof [Proof of Lemma 7]

1. Notice that maximizer Δ for $\delta(t)$ yields a feasible solution $\lambda\Delta$ in the program for $\delta(\lambda t)$.

2. The first inequality of (25) is obvious. Now consider Δ is the maximizer for $\delta(t)$. Let $\langle \Delta, \mathbf{1} \rangle = \epsilon$. Since Φ is column-stochastic, we have $\Phi^\top \mathbf{1} = \mathbf{1}$ and thus

$$\langle \Delta, \mathbf{1} \rangle = \langle \Phi\Delta, \mathbf{1} \rangle \leq \|\Phi\Delta\|_1 \|\mathbf{1}\|_\infty \leq t$$

and thus $|\epsilon| \leq t$. Then, let $\pi_\pm$ be distributions on Θ with $\langle \pi_-, h \rangle \leq 0 \leq \langle \pi_+, h \rangle$ (their existence follows from that of $\theta_\pm$). Let $\pi_0 = \pi_-$ if $\epsilon > 0$ and $\pi_0 = \pi_+$ otherwise. Define

$$\tilde{\Delta} \triangleq \frac{1}{2}\Delta - \frac{\epsilon}{2}\pi_0.$$

From $\|\pi_-\|_1 = 1$, $\|\Phi r\|_1 \leq \|r\|_1$ and triangle inequality we have:

$$\|\tilde{\Delta}\| \leq \frac{1}{2} + \frac{|\epsilon|}{2} \leq 1$$

$$\|\Phi\tilde{\Delta}\| \leq \frac{1}{2}\|\Phi\Delta\|_1 + \frac{|\epsilon|}{2} \leq t$$

$$\langle \tilde{\Delta}, \mathbf{1} \rangle = \frac{1}{2}\langle \Delta, \mathbf{1} \rangle - \frac{\epsilon}{2} = 0$$

$$\langle \tilde{\Delta}, h \rangle = \frac{1}{2}\langle \Delta, h \rangle - \frac{\epsilon}{2}\langle \pi_0, h \rangle \geq \frac{1}{2}\langle \Delta, h \rangle \ .$$

Thus, we get that $\tilde{\delta}(t) \geq \frac{1}{2}\delta(t)$.

3. Let z be such that $\langle z, \mathbf{1} \rangle = 0$ and $\|z\|_1 \leq 1$ and set $\Delta = tz$. Then $\|\Phi\Delta\|_1 \leq \|\Phi\|_{\ell_1 \to \ell_1}\|\Delta\|_1 \leq t$. Choose z to maximize $\langle z, h \rangle$ gives the desired result with $C(h) = \frac{h_{\max} - h_{\min}}{2}$, where $h_{\max} > h_{\min}$ are the maximal and minimal values of $h(\theta)$.

∎

Proof [Proof of Lemma 15] Recall the following upper bound on the A-norm in terms of the H^∞-norm of the derivative, which is a simple consequence of Parseval's identity and the Cauchy-Schwartz inequality (see, e.g., Newman (1975)):

$$\|g\|_A \leq \|g\|_{H^\infty(D)} + 2\|g'\|_{H^\infty(D)}. \tag{89}$$

Recall that $\|\delta^{\frac{1+z}{1-z}}\|_{H^\infty(D)} = 1$. We have $\|g\|_{H^\infty(D)} \leq 4$. Furthermore, $g'(z) = 2\delta^{\frac{1+z}{1-z}}(\log\delta + z - 1)$. We have $\|g'\|_{H^\infty(D)} \leq 2(\log\frac{1}{\delta} + 2)$ and hence the lemma. ∎

Remark 20 (Alternative proof of Lemma 15) *In fact with more refined analysis it is possible to choose*

$$g(z) = (1 - z)\delta^{\frac{1+z}{1-z}}$$

as opposed to (79). Of course, in this case the bound (89) based on derivatives fails. Instead, we can express the Taylor expansion coefficients of g in terms of Laguerre polynomials. Recall the generalized Laguerre polynomial $L_n^{(\nu)}(x)$ and its generating function (Gradshteyn and Ryzhik, 2007, 8.975):

$$\sum_{n\geq0} z^n L_n^{(\nu)}(x) = \frac{1}{(1-z)^{1+\nu}}\exp\left(\frac{zx}{z-1}\right), \quad |z| < 1.$$

Since $g(z) = (1 - z)\delta\delta^{\frac{2}{1-z}}$, its Taylor expansion $g(z) = \sum_{n\geq0} a_n z^n$ is given by $a_n = \delta L_n^{(-2)}(x)$ with $x = 2\log\frac{1}{\delta}$ and hence $\|g\|_A = \sum_{n\geq0}|L_n^{(-2)}(x)|$. Recall the classical bound of Szegö (cf. (Szegö, 1975, (7.6.10))) on Laguerre polynomials: for $\nu \leq -\frac{1}{2}$, as $n \to \infty$, $|L_n^{(\nu)}(x)| = O(n^{\nu/2-1/4})$ uniformly in any compact interval of $[0,\infty)$. Therefore $a_n =$

$O(n^{-5/4})$ *and hence* $\|g\|_A$ *is finite. Furthermore, recall Fejér's sharp asymptotics of La-guerre polynomials (Szegö, 1975, Theorem 8.22.1): for fixed* $\nu \in \mathbb{R}$ *as* $n \to \infty$, $L_n^{(\nu)}(x) = \sqrt{\pi} e^{x/2} x^{-\nu/2-1/4} n^{\nu/2-1/4} \cos(2\sqrt{nx} - \pi(1/4 + \nu/2)) + O(n^{\nu/2-3/4})$, *where* $x > 0$ *and the re-mainder is uniform in any compact interval of* $(0, \infty)$. *This suggests that* $\|g\|_A = O((\log \frac{1}{\delta})^{5/4})$.

Proof [Proof of Lemma 16] Note that $\|\delta^{\frac{1+\alpha z}{1-\alpha z}} h(z)\|_{H^\infty(D)} = \max_{|z|=1} \left| \delta^{\frac{1+\alpha z}{1-\alpha z}} \frac{1}{(1-c(z-1))^d} \right|$, where

$$\left| \frac{\delta^{\frac{1+\alpha e^{i\theta}}{1-\alpha e^{i\theta}}}}{(1-c(e^{i\theta}-1))^d} \right| = \delta^{\mathrm{Re}(\frac{1+\alpha e^{i\theta}}{1-\alpha e^{i\theta}})} |1 - c(e^{i\theta} - 1)|^{-d}$$

$$= \delta^{\frac{1-\alpha^2}{1-2\alpha\cos\theta+\alpha^2}} (1 + 2c(1+c)(1-\cos\theta))^{-d/2} = e^{-F(1-\cos\theta)}$$

where $F(t) \triangleq \frac{1-\alpha^2}{(1-\alpha)^2+2\alpha t} \log \frac{1}{\delta} + \frac{d}{2} \log(1 + 2t\mu)$ and $\mu = \mu(\epsilon)$ as defined in (68) since $c = \frac{\epsilon}{1-2\epsilon}$ and $c(1+c) = \mu$. The goal is to identify conditions so that the H^∞-norm is achieved at $\theta = 0$ ($z = 1$). Let $s = 2t\mu, A = \frac{d}{2(1-\alpha^2)\log\frac{1}{\delta}}, B = (1-\alpha)^2, C = \alpha/\mu$. Then $F(t) = (1-\alpha^2)\log\frac{1}{\delta}(A\log(1+s) + \frac{1}{B+Cs}) \triangleq \delta(s)$. Straightforward calculation shows that $F'(t) \geq 0$ for $t \in [0, 2]$ provided that

$$A \geq \frac{C}{B^2}, \quad A \geq \frac{1}{2B},$$

that is, $d \geq \log \frac{1}{\delta} \max\{\frac{\alpha(1-\alpha^2)}{\mu(1-\alpha)^4}, \frac{1-\alpha^2}{(1-\alpha)^2}\}$, which is ensured by (82). Hence $\|\delta^{\frac{1+\alpha z}{1-\alpha z}} h(z)\|_{H^\infty(D)} = e^{-F(0)}$ as claimed. ∎

Finally, the next result is used in Section 3.2:

Lemma 21 *For any* $d' \geq d \geq 1$ *and* $0 \leq p < 1$, $H^2(\mathrm{Bin}(d', p), \mathrm{Bin}(d' - d, p)) \leq \frac{4pd^2}{(1-p)d'}$.

Proof [Proof of Lemma 21] It suffices to show that for any $n \geq 1$, $H^2(\mathrm{Bin}(n, p), \mathrm{Bin}(n - 1, p)) \leq \frac{p}{(1-p)n}$. Indeed, since $H^2(P, Q) \leq \chi^2(P\|Q) = \int \frac{(dP)^2}{dQ} - 1$, we have

$$\chi^2(\mathrm{Bin}(n-1, p)\|\mathrm{Bin}(n, p)) = \mathbb{E}_{X \sim \mathrm{Bin}(n,p)}\left[\left(\frac{n-X}{n(1-p)}\right)^2 \mathbf{1}_{\{X < n\}}\right] \leq \frac{p}{n(1-p)}.$$

Then by the triangle inequality of the Hellinger distance, we have $H(\mathrm{Bin}(d', p), \mathrm{Bin}(d' - d, p)) \leq \sqrt{\frac{p}{1-p}} \sum_{n=d'-d+1}^{d'} \frac{1}{\sqrt{n}} \leq \sqrt{\frac{p}{1-p}} \int_{d'-d}^{d'} \frac{1}{\sqrt{x}} dx \leq \sqrt{\frac{p}{1-p}} \frac{2d}{\sqrt{d'}}$. ∎

Acknowledgement

References

M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Wiley-Interscience, New York, NY, 1964.

Lucia Batman, Russell Impagliazzo, Cody Murray, and Ramamohan Paturi. Finding heavy hitters from lossy or noisy data. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 347–362. Springer, 2013.

Peter Borwein and Tamás Erdélyi. Littlewood-type problems on subarcs of the unit circle. *Indiana University Mathematics Journal*, 46(4):1323, 1997.

T Tony Cai and Mark G Low. Nonquadratic estimators of a quadratic functional. *The Annals of Statistics*, pages 2930–2956, 2005.

Carl C Cowen. Linear fractional composition operators on H^2. *Integral equations and operator theory*, 11(2):151–160, 1988.

Anindya De, Ryan O'Donnell, and Rocco Servedio. Optimal mean-based algorithms for trace reconstruction. *arXiv preprint arXiv:1612.03148*, 2016a.

Anindya De, Michael Saks, and Sijian Tang. Noisy population recovery in polynomial time. *arXiv preprint arXiv:1602.07616*, 2016b.

Anindya De, Ryan O'Donnell, and Rocco Servedio. Sharp bounds for population recovery. *arXiv preprint arXiv:1703.01474*, 2017.

Zeev Dvir, Anup Rao, Avi Wigderson, and Amir Yehudayoff. Restriction access. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 19–33. ACM, 2012.

Jianqing Fan, Philippe Rigollet, and Weichen Wang. Estimation of functionals of sparse covariance matrices. *The Annals of Statistics*, 43(6):2706, 2015.

Jon Feldman, Ryan O'Donnell, and Rocco A Servedio. Learning mixtures of product distributions over discrete domains. *SIAM Journal on Computing*, 37(5):1536–1564, 2008.

I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals Series and Products.* Academic Press, New York, NY, seventh edition, 2007.

I.A. Ibragimov, A.S. Nemirovskii, and R.Z. Khas'minskii. Some problems on nonparametric estimation in gaussian white noise. *Theory of Probability & Its Applications*, 31(3):391–406, 1987.

Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 273–282. ACM, 1994.

Lucien Le Cam. *Asymptotic methods in statistical decision theory.* Springer-Verlag, New York, NY, 1986.

Jian Li, Yuval Rabani, Leonard J Schulman, and Chaitanya Swamy. Learning arbitrary statistical mixtures of discrete distributions. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 743–752. ACM, 2015.

Karim Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

Shachar Lovett and Jiapeng Zhang. Improved noisy population recovery, and reverse Bonami-Beckner inequality for sparse functions. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 137–142. ACM, 2015.

Ankur Moitra and Michael Saks. A polynomial time algorithm for lossy population recovery. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 110–116. IEEE, 2013.

F. Nazarov and Y. Peres. Trace reconstruction with $\exp(O(n^{1/3}))$ samples. *arXiv preprint arXiv:1612.03599*, 2016.

DJ Newman. A simple proof of Wieners $1/f$ theorem. *Proceedings of the American Mathematical Society*, 48(1):264–265, 1975.

Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences (PNAS)*, 113(47):13283–13288, 2016.

Barry Simon. *Convexity: An analytic viewpoint*. Cambridge University Press, 2011.

G. Szegö. *Orthogonal polynomials*. American Mathematical Society, Providence, RI, 4th edition, 1975.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Verlag, New York, NY, 2009.

Avi Wigderson and Amir Yehudayoff. Population recovery and partial identification. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 390–399. IEEE, 2012.