Optimal disclosure risk assessment

Federico Camerlenghi[1]*, Stefano Favaro[2], Zacharie Naulet[3], Francesca Panero[4]

[1]Università di Milano - Bicocca
[2]Univerisità di Torino and Collegio Carlo Alberto
[3]Université Paris-Saclay
[4]University of Oxford

March 14, 2020

Abstract

Protection against disclosure is a legal and ethical obligation for agencies releasing microdata files for public use. Consider a microdata sample of size n from a finite population of size $\bar{n} = n + \lambda n$, with $\lambda > 0$, such that each sample record contains two disjoint types of information: identifying categorical information and sensitive information. Any decision about releasing data is supported by the estimation of measures of disclosure risk, which are defined as discrete functionals of the number of sample records with a unique combination of values of identifying variables. The most common measure is arguably the number τ_1 of sample unique records that are population uniques. In this paper, we first study nonparametric estimation of τ_1 under the Poisson abundance model for sample records. We introduce a class of linear estimators of τ_1 that are simple, computationally efficient and scalable to massive datasets, and we give uniform theoretical guarantees for them. In particular, we show that they provably estimate τ_1 all of the way up to the sampling fraction $(\lambda + 1)^{-1} \propto (\log n)^{-1}$, with vanishing normalized mean-square error (NMSE) for large n. We then establish a lower bound for the minimax NMSE for the estimation of τ_1, which allows us to show that: i) $(\lambda+1)^{-1} \propto (\log n)^{-1}$ is the smallest possible sampling fraction for consistently estimating τ_1; ii) estimators' NMSE is near optimal, in the sense of matching the minimax lower bound, for large n. This is the main result of our paper, and it provides a rigorous answer to an open question about the feasibility of nonparametric estimation of τ_1 under the Poisson abundance model and for a sampling fraction $(\lambda + 1)^{-1} < 1/2$.

1 Introduction

Protection against disclosure is a legal and ethical obligation for agencies releasing microdata files for public use. Any decision about release requires a careful assessment of the risk of disclosure, which is supported by the estimation of measures of disclosure risk (Willenborg and de Waal (2001)). Let consider a microdata sample $\boldsymbol{X}(n) = (X_1, \ldots, X_n)$ from a finite population of size $\bar{n} > n$ and, without loss of generality, assume that each X_i is a record containing two disjoint types of information for the i-th individual: identifying information and sensitive information. Identifying information consists of a set of categorical variables which might be matchable to known units of the population. A risk of disclosure results from the possibility that an intruder might succeed in identifying a microdata unit through such a matching, and hence be able to

[*] Corresponding author federico.camerlenghi@unimib.it

disclose sensitive information on this unit. To quantify the risk of disclosure, sample records $\boldsymbol{X}(n)$ are typically cross-classified according to identifying variables. That is, $\boldsymbol{X}(n)$ is partitioned in $K_n \leq n$ cells, with $Y_j(\boldsymbol{X}, n)$ being the number of X_i's belonging to cell j, for $j = 1, \ldots, K_n$, such that $\sum_{1 \leq j \leq K_n} Y_j(\boldsymbol{X}, n) = n$; we refer to the number of occurrences $Y_j(\boldsymbol{X}, n)$ as the sample frequency of cell j. Then, a risk of disclosure arises from cells in which both sample frequencies and population frequencies are small. Of special interest are cells with frequency 1 (singletons or uniques) since, assuming no errors in the matching process or data sources, for these cells the match is guaranteed to be correct. This has motivated inferences on measures of disclosure risk that are suitable functionals of the number of uniques, the most common being the number τ_1 of sample uniques which are also population uniques. We refer to Skinner et al. (1994) for a comprehensive account on measures of disclosure risk.

In this paper, we first study nonparametric estimation of the discrete functional τ_1 under the Poisson abundance model for sample records. The Poisson abundance model is arguably the most natural, and weak, assumption to infer τ_1 (Bethlehem et al. (1990) and Skinner and Shlomo (2008)). If $\bar{n} = n + \lambda n$, with $\lambda > 0$, the model assumes that: i) the population records $(X_1, \ldots, X_{n+\lambda n})$ can be ideally extended to a sequence $\boldsymbol{X} = (X_i)_{i \geq 1}$, of which $\boldsymbol{X}(n)$ is an observable subsample; ii) the X_i's are independent and identically distributed as an unknown distribution $(p_j)_{j \geq 1}$, where p_j is the probability of the j-th cell in which $\boldsymbol{X}$ may be cross-classified; iii) the sample size is a Poisson random variable N with mean n, in symbols $N \sim \text{Poiss}(n)$. Then, sample records $\boldsymbol{X}(N) = (X_1, \ldots, X_N)$ result in K_N cells with $Y_j(\boldsymbol{X}, N)$ being the sample frequency of cell j, for $j = 1, \ldots, K_N$, such $Y_j(\boldsymbol{X}, N) \sim \text{Poiss}(np_j)$, $Y_{j_1}(\boldsymbol{X}, N)$ is independent of $Y_{j_2}(\boldsymbol{X}, N)$ for any $j_1 \neq j_2$, and $\sum_{1 \leq j \leq K_N} Y_j(\boldsymbol{X}, N) = N$. Skinner and Elliot (2002) first raised the problem of nonparametric estimation of τ_1 under the Poisson abundance model, leaving that as an open problem. In particular, they discussed about the feasibility of nonparametric estimation of τ_1, arguing that it is an intrinsically difficult problem. The problem shares the well-known difficulties of the classical problem of estimating the number of unseen species (Good and Toulmin (1956) and Efron and Thisted (1976)). Indeed nonparametric estimators of τ_1 may be "unreasonable" since they are subject to serious upward bias and high variance for small sampling fractions of the population, i.e. $(\lambda + 1)^{-1} < 1/2$ or, in other words, for n smaller than a half of the population $\bar{n}$.

Under the Poisson abundance model for sample records $\boldsymbol{X}(n)$ from the population $(X_1, \ldots, X_{n+\lambda n})$, we introduce a class of nonparametric linear estimators of τ_1 that are simple, computationally efficient and scalable to massive datasets. We show that our estimators admit an interpretation as (smoothed) nonparametric empirical Bayes estimators in the sense of Robbins (1956), and we prove theoretical guarantees for them that hold uniformly for any distribution $(p_j)_{j \geq 1}$. In particular, we show that our estimators provably estimate τ_1 all of the way up to the sampling fraction $(\lambda + 1)^{-1} \propto (\log n)^{-1}$ of the population, with vanishing normalized mean-square error (NMSE) as n becomes large. Then, by relying on recent techniques developed in Wu and Yang (2019) in the context of nonparametric estimation of the support size of discrete distributions, we establish a lower bound for the minimax NMSE for the estimation of τ_1. This result allows us to show that $(\lambda + 1)^{-1} \propto (\log n)^{-1}$ is the smallest possible sampling fraction of the population for consistently estimating τ_1, and that the estimators' NMSE is near optimal, in the sense of matching the minimax lower bound, for a large sample size n. This is the main result of the present paper, and it provides a rigorous answer to the question raised by Skinner and Elliot (2002) about the feasibility of nonparametric estimation of τ_1 under the Poisson abundance model and for a sampling fraction $(\lambda + 1)^{-1} < 1/2$. Indeed our result shows that nonparametric estimation of τ_1 has uniformly provable guarantees, in terms of vanishing NMSE for large n, if and only if $(\lambda + 1)^{-1} \propto (\log n)^{-1}$.

Starting from the seminal work of Bethlehem et al. (1990), in the last three decades a full range of parametric and semiparametric approaches, both frequentist and Bayesian, has been proposed for making inference on τ_1. See, e.g., Skinner et al. (1994), Samuels (1998) Reiter (2005), Rinott and Shlomo (2006), Skinner and Shlomo (2008), Manrique-Vallier and Reiter (2012), Manrique-Vallier and Reiter (2014), Carota et al. (2015) and Carota et al. (2018). A common thread of these works has been the enrichment of the classical Poisson abundance model with stronger modeling assumptions: while early approaches were focused on parametric Bayesian modeling of the random partition induced by the cross classification of sample records, recent approaches focused on semiparametric modeling of the associations among identifying variables, typically by means of complex Bayesian hierarchical latent class models. All approaches in the literature are shown to empirically estimate τ_1, even for relatively small sampling fractions, but without any provable guarantees. The approach we propose in the present paper may be viewed as the natural nonparametric counterpart of the parametric empirical Bayes approach, in the sense of Efron and Morris (1973), introduced in Bethlehem et al. (1990) and further developed in Skinner et al. (1994) and Rinott and Shlomo (2006). Besides being the first nonparametric approach to the estimation of τ_1 under the Poisson abundance model, our approach stands out for being the first to give theoretical guarantees on the performance of the proposed class of estimators.

The paper is structured as follows. In Section 2 we introduce a class of nonparametric estimators for τ_1, and we show that they provably estimate τ_1 all of the way up to the sampling fraction $(\lambda+1)^{-1} \propto (\log n)^{-1}$, with vanishing NMSE for large sample size n. In Section 3 we show that $(\lambda+1)^{-1} \propto (\log n)^{-1}$ is the smallest possible sampling fraction of the population which guarantees a vanishing NMSE, and that estimators' NMSE is near optimal for large n. Section 4 contains a discussion of our results, their interplay with other discrete functional estimation problems, and remaining open challenges. Proofs are deferred to the Appendix, whereas technical results and numerical illustrations are available as online supplementary material.

2 A nonparametric estimator of τ_1

We consider an infinite sequence of observations $\boldsymbol{X}$, and we assume that $\boldsymbol{X}(N) = (X_1, \ldots, X_N)$ is the microdata sample of random size N under the Poisson abundance model. We suppose that $\boldsymbol{X}(N)$ is a subsample of $(X_1, \ldots, X_{M+N})$, where $M \sim \text{Poiss}(\lambda n)$, with $\lambda > 0$ and independent of N. In the present framework $(X_{N+1}, \ldots, X_{N+M})$ may be seen as the unobservable population. When sample records are cross-classified according to identifying variables, the sample $(X_1, \ldots, X_N)$ results partitioned in $K_N \leq N$ cells with corresponding sample frequencies $(Y_1(\boldsymbol{X}, N), \ldots, Y_{K_N}(\boldsymbol{X}, N))$ such that $\sum_{1 \leq j \leq K_N} Y_j(\boldsymbol{X}, N) = N$. Hereafter we denote by $Z_i(\boldsymbol{X}, N)$ the number of cells with frequency i, and by $Z_{\bar{i}}(\boldsymbol{X}, N)$ the number of cells with frequency greater or equal than i, for any index $i \geq 1$. We are interested in estimating the number τ_1 of sample uniques which are also population uniques, namely the following discrete functional

$$\tau_1(\mathbf{X}, N, M) = \sum_{j \geq 1} \mathbb{1}_{\{Y_j(\boldsymbol{X}, N)=1\}} \mathbb{1}_{\{Y_j(\boldsymbol{X}, N+M)=1\}},$$

where $\mathbb{1}$ denotes the indicator function. We recall that the frequency counts, defined as $Y_j(\boldsymbol{X}, N) = \sum_{1 \leq i \leq N} \mathbb{1}_{\{X_i = j\}}$, are distributed according to a Poisson distribution with parameter np_j, where p_j is the unknown probability associated to the j-th cell, that is $p_j \in [0, 1]$ for $j \geq 1$ such that $\sum_{j \geq 1} p_j = 1$. We will denote by $\boldsymbol{Y}(\boldsymbol{X}, N) := (Y_1(\boldsymbol{X}, N), \ldots)$ the whole sequence of the cell's frequency counts. We remark that, under the Poisson abundance model, the $Y_j(\boldsymbol{X}, N)$'s are independent random variables variables and, in addition, $Y_j(\boldsymbol{X}, N+M) - Y_j(\boldsymbol{X}, N)$ is independent of $Y_j(\boldsymbol{X}, N)$, for any $j \geq 1$: these properties follow from standard statistical arguments.

When the sample size n is fixed, the independence property of the $Y_j(\boldsymbol{X}, n)$'s falls down and approximation arguments are required to handle such a situation.

To fix the notation, in the sequel we will write $f \lesssim g$, for two generic functions f and g, if and only if (iff) there exists a universal constant $C > 0$ such that $f(x) \leq Cg(x)$; we will further write $f \asymp g$ whenever both $f \lesssim g$ and $g \lesssim f$ are satisfied. Let us denote by $\mathscr{P}$ the set of all possible distributions over the set of natural numbers $\mathbb{N}$, i.e. $\mathscr{P} := \{P = \sum_{j\geq 1} p_j \delta_j : p_j \in [0,1], \text{ with} \sum_{j\geq 1} p_j = 1\}$, where δ_j denotes the Dirac measure centered at $j \in \mathbb{N}$. An estimator of $\tau_1(\boldsymbol{X}, N, M)$ is understood to be a measurable function $\hat{\rho}_1(\boldsymbol{X}(N), N)$ depending on the available sample $\boldsymbol{X}(N)$ and the actual size of the observed sample N. We will evaluate the performance of a generic estimator $\hat{\rho}_1(\boldsymbol{X}(N), N)$ of $\tau_1(\boldsymbol{X}, N, M)$, by its worst–case NMSE, defined as

$$\mathscr{E}_{\lambda,n}(\hat{\rho}_1(\boldsymbol{X}(N), N)) := \sup_{P\in\mathscr{P}} \frac{\mathbb{E}[(\hat{\rho}_1(\boldsymbol{X}(N), N) - \tau_1(\boldsymbol{X}, N, M))^2]}{n^2}, \tag{1}$$

where $\mathbb{E}[(\hat{\rho}_1(\boldsymbol{X}(N), N) - \tau_1(\boldsymbol{X}, N, M))^2]$ is the mean squared error (MSE) of $\hat{\rho}_1$ under the model (P, n, λ), also denoted by $\text{MSE}[\hat{\rho}_1(\boldsymbol{X}(N), N)]$. Since $\text{MSE}[\hat{\rho}_1(\boldsymbol{X}(N), N)]$ does not vanish as $n \to +\infty$, it is common to evaluate the performance of an estimator for $\tau_1(\boldsymbol{X}, N, M)$ in terms of the NMSE (see, e.g., Orlitsky et al. (2016)). The NMSE is indeed the MSE of $\hat{\rho}_1(\boldsymbol{X}(N), N)$ normalized by the maximum value of $\tau_1(\boldsymbol{X}, N, M)$ (which is exactly n, given $N = n$), and hence the performance of $\hat{\rho}_1(\boldsymbol{X}(N), N)$ is evaluated in terms of the rate of convergence to 0 of the NMSE as $n \to +\infty$.

A nonparametric estimator for $\tau_1(\boldsymbol{X}, N, M)$ may be simply deduced by comparing expectations. Indeed, under the Poisson abundance model, it easy to see that

$$\mathbb{E}[\tau_1(\boldsymbol{X}, N, M)] = \sum_{i\geq 0} (-1)^i \lambda^i (i+1) \mathbb{E}[Z_{i+1}(\boldsymbol{X}, N)]. \tag{2}$$

See Appendix A.1 for details on the derivation of identity (2). In particular, according to identity (2) we can define the following estimator of $\tau_1(\boldsymbol{X}, N, M)$:

$$\hat{\tau}_1(\boldsymbol{X}(N), N) = \sum_{i\geq 0} (-1)^i (i+1) \lambda^i Z_{i+1}(\boldsymbol{X}, N). \tag{3}$$

By construction $\hat{\tau}_1(\boldsymbol{X}(N), N)$ is an unbiased estimator of $\mathbb{E}[\tau_1(\boldsymbol{X}, N, M)]$, that is $\mathbb{E}[\hat{\tau}_1(\boldsymbol{X}(N), N)] = \mathbb{E}[\tau_1(\boldsymbol{X}, N, M)] = \sum_{j\geq 1} np_j e^{-(\lambda+1)np_j}$. The estimator $\hat{\tau}_1(\boldsymbol{X}(N), N)$ admits a natural interpretation as a nonparametric empirical Bayes estimator in the sense of Robbins (1956). More precisely, $\hat{\tau}_1(\boldsymbol{X}(N), N)$ is the posterior expectation of $\mathbb{E}[\tau_1(\boldsymbol{X}, N, M)]$ with respect to an unknown prior distribution on the p_i's that is estimated from the $Y_j(\mathbf{X}, N)$. See Appendix A.2 for details. This observation makes the estimator (3) the natural nonparametric counterpart of the parametric empirical Bayes estimator, in the sense of Efron and Morris (1973), introduced in Bethlehem et al. (1990).

Theorem 1 *For any positive reals x and y let $\lfloor x \rfloor$ denote the integer part of x and let $x \vee y$ denote the maximum between x and y. If $\lambda < 1$, for any $P \in \mathscr{P}$ and for any $n > 0$*

$$\begin{aligned} &\text{Var}[\tau_1(\boldsymbol{X}, N, M) - \hat{\tau}_1(\boldsymbol{X}(N), N)] \\ &\qquad\qquad \leq \Psi^2(\lambda) \mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] - \frac{\mathbb{E}[Z_1(\boldsymbol{X}, N+M)]}{\lambda+1}, \end{aligned} \tag{4}$$

where in (4) we defined $\Psi(\lambda) = (j^*+1)\lambda^{j^*}$ such that $j^* = \lfloor (2\lambda - 1)/(1-\lambda) \rfloor \vee 0$.

The proof of Theorem 1 is deferred to Appendix A.3. According to Theorem 1, for $\lambda < 1$ one has that $\mathrm{Var}[\tau_1(\boldsymbol{X}, N, M) - \hat{\tau}_1(\boldsymbol{X}(N), N)] \lesssim n$ upon noticing that $\mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] \leq \mathbb{E}[N] = n$. That is, in expectation, $\hat{\tau}_1(\boldsymbol{X}(N), N)$ approximate $\tau_1(\boldsymbol{X}, N, M)$ to within n. We formalize these observations in the next corollary.

Corollary 1 *If $\lambda < 1$ is fixed then $\mathscr{E}_{\lambda,n}(\hat{\tau}_1(\boldsymbol{X}(N), N)) \leq W(\lambda)/n$, for any $n \geq 1$ and for some constant $W(\lambda)$ depending only on λ.*

Corollary 1 legitimates $\hat{\tau}_1(\boldsymbol{X}(N), N)$ as an estimator of $\tau_1(\boldsymbol{X}, N, M)$ under the assumption $\lambda < 1$. Unfortunately, this assumption is unrealistic in the context of disclosure risk assessment, where the size λn of the unobserved population is typically much bigger than the size n of the observed sample. The variance bound in Theorem 1 reveals that the assumption $\lambda < 1$ is necessary to obtain a finite estimate of the variance. This variance issue of $\hat{\tau}_1(\boldsymbol{X}(N), N)$ is determined by the geometrically increasing magnitude of the coefficients $(i+1)(-\lambda)^i$. Indeed, as $\lambda \geq 1$, the estimator $\hat{\tau}_1(\boldsymbol{X}(N), N)$ grows superlinearly as $(i+1)(-\lambda)^i$ for the largest i such that $Z_{i+1}(\boldsymbol{X}, N) > 0$, thus eventually far exceeding $\tau_1(\boldsymbol{X}, N, M)$ that grows at most linearly. Then $\hat{\tau}_1(\boldsymbol{X}(N), N)$ is useless for $\lambda \geq 1$, thus requiring an adjustment via suitable smoothing techniques. To fix this issue we follow ideas developed by Good and Toulmin (1956), Efron and Thisted (1976) and Orlitsky et al. (2016) in the context of the nonparametric estimation of the number of unseen species. We propose a smoothed version of $\hat{\tau}_1(\boldsymbol{X}(N), N)$ by truncating the series (3) at an independent random location L, and then averaging over the distribution of L, i.e.,

$$\begin{aligned}\hat{\tau}_1^L(\boldsymbol{X}(N), N) &= \mathbb{E}_L\left[\sum_{i=1}^{L}(-1)^i(i+1)\lambda^i Z_{i+1}(\boldsymbol{X}, N)\right] \\ &= \sum_{i\geq 0}(-1)^i(i+1)\lambda^i \mathbb{P}(L \geq i) Z_{i+1}(\boldsymbol{X}, N).\end{aligned} \tag{5}$$

For any $\lambda \geq 1$, as the the index i in (5) increases, the tail probability $\mathbb{P}[L \geq j]$ compensate for the exponential growth of $(i+1)(-\lambda)^i$, thereby stabilizing the variance. In the next theorem we show that for $\lambda \geq 1$ the estimator $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$ is biased for $\mathbb{E}[\tau_1(\boldsymbol{X}, N, M)]$, and we provide a bound for the MSE of $\hat{\tau}_1(\boldsymbol{X}(N), N)$.

Theorem 2 *Let $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$ be the estimator of $\tau_1(\mathbf{X}, N, M)$ defined in (5). If $\lambda \geq 1$ then*

$$\begin{aligned}&\mathbb{E}[\hat{\tau}_1^L(\boldsymbol{X}(N), N)] \\ &\quad = \mathbb{E}[\tau_1(\boldsymbol{X}, N, M)] + \sum_{j\geq 1} e^{-p_j n(\lambda+1)} p_j n \int_0^{\lambda n p_j} e^s \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right] \mathrm{d}s\end{aligned} \tag{6}$$

and

$$\begin{aligned}&\mathrm{MSE}[\hat{\tau}_1^L(\boldsymbol{X}(N), N)] \\ &\quad \leq (\mathbb{E}_L[(L+1)\lambda^L])^2 \mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] - \frac{\mathbb{E}[Z_1(\boldsymbol{X}, N+M)]}{\lambda+1} \\ &\qquad + \left(\sum_{j\geq 1} e^{-p_j n(\lambda+1)} p_j n \int_0^{\lambda n p_j} e^s \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right] \mathrm{d}s\right)^2.\end{aligned} \tag{7}$$

The proof of Theorem 2 is in Appendix A.4. Choosing different smoothing distributions for L yields different estimators for $\tau_1(\boldsymbol{X}, N, M)$. Following Orlitsky et al. (2016), we consider two distributions for L: i) a Poisson distribution with parameter $\beta > 0$; ii) a Binomial distribution with parameter $(x_0, 2/(\lambda+2))$. To choose the parameter β of the Poisson distribution and the parameter x_0 of the Binomial distribution, one should look for $\tilde{\beta}$ and $\tilde{x}_0$ which minimizes the MSE bound (7). Once the values of $\tilde{\beta}$ and $\tilde{x}_0$ are determined explicitly, we are able to obtain a "limit of predictability" for $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$. That is, for some $\delta > 0$ we are able to specify the maximum value of the sampling fraction λ for which $\mathscr{E}_{\lambda,n}(\hat{\tau}_1^L(\boldsymbol{X}(N), N)) < \delta$. This gives a provable (performance) guarantee for the estimation of $\tau_1(\boldsymbol{X}, N, M)$ in terms of λ.

Proposition 1 *Let L be a Poisson random variable with parameter β. Then*

$$\mathrm{MSE}[\hat{\tau}_1^L(\boldsymbol{X}(N), N)] \leq e^{-2\beta} n^2 + n e^{2\beta(2\lambda-1)}. \tag{8}$$

The right-hand side of (8) in minimized by setting $\tilde{\beta} = \log(n/(2\lambda-1))/(4\lambda)$, for any $\lambda \geq 1$. Moreover, if L is a Poisson random variable with parameter $\tilde{\beta}$ then

$$\mathscr{E}_{n,\lambda}(\hat{\tau}_1^L(\boldsymbol{X}(N), N)) \leq \frac{A(\lambda)}{n^{1/(2\lambda)}}, \tag{9}$$

and for any $\delta \in (0,1)$

$$\lim_{n\to+\infty} \frac{\max\left\{\lambda : \mathscr{E}_{n,\lambda}(\hat{\tau}_1^L(\boldsymbol{X}(N), N)) \leq \delta\right\}}{\log(n)} \geq \frac{1}{2\log(A/\delta)} \tag{10}$$

where $A(\lambda)$, is continuous in $[1, +\infty)$ with $\lim_{\lambda\to+\infty} A(\lambda) = 1$ and $A = \max_{\lambda\geq 1} A(\lambda) < +\infty$.

See Appendix A.5 for the proof of Proposition 1. A result similar to Proposition 1 holds true when the random variable L is assumed to be distributed according to a Binomial distribution. This result is stated in the next proposition, and its proof is omitted since it is along lines similar to the proof of Proposition 1.

Proposition 2 *For any positive reals x and y let $\lfloor x \rfloor$ denote the integer part of x. Let L be a Binomial random variable with parameter $(x_0, 2/(\lambda+2))$. Then*

$$\mathrm{MSE}[\hat{\tau}_1^L(\boldsymbol{X}(N), N)] \leq n \left(\frac{\lambda}{\lambda+2}\right)^{2x_0} \left[3^{10x_0/3} + n\left(\frac{\lambda}{2(\lambda+1)}\right)^2\right] \tag{11}$$

and the choice $\tilde{x}_0 = \lfloor (3/10)\log_3(n\lambda^2/((\lambda+1)(\lambda^2(3^{10/3}-1) - 4\lambda - 4)))\rfloor$ minimizes the right-hand side of (11), for any $\lambda \geq 1$. Moreover, if L is a Binomial random variable with parameter $(\tilde{x}_0, 2/(\lambda+2))$ then

$$\mathscr{E}_{n,\lambda}(\hat{\tau}_1^L(\boldsymbol{X}(N), N)) \leq \frac{C(\lambda)}{n^{3\log_3(1+2/\lambda)/5}}, \tag{12}$$

and for any $\delta \in (0,1)$

$$\lim_{n\to+\infty} \frac{\max\left\{\lambda : \mathscr{E}_{n,\lambda}(\hat{\tau}_1^L) \leq \delta\right\}}{\log(n)} \geq \frac{6}{5\log(3)\log(C/\delta)} \tag{13}$$

where $C(\lambda)$ is continuous in $[1, +\infty)$ with $\lim_{\lambda\to+\infty} C(\lambda) = 1$ and $C = \max_{\lambda\geq 1} C(\lambda)$.

3 Optimality of the proposed estimators

In Section 2 we have introduced two different estimators of $\tau_1(\boldsymbol{X}, N, M)$, and we have provided guarantees of their performance, as $n \to +\infty$, in terms of the NMSE. We have already remarked that the case $\lambda \geq 1$ is the most interesting one for estimating the disclosure risk $\tau_1(\boldsymbol{X}, N, M)$. Indeed in the context of disclosure risk assessment the fraction of the unobserved sample λ is usually much larger than 1. Throughout the section we assume that $\lambda \geq 1$ and we prove that the proposed estimator $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$ is essentially optimal. More precisely we determine a lower bound for the best worst–case NMSE, defined as

$$\mathscr{E}(\lambda, n) := \inf_{\hat{\rho}_1} \mathscr{E}_{\lambda,n}(\hat{\rho}_1(\boldsymbol{X}(N), N)) \tag{14}$$

where the infimum in the previous definition runs over all possible estimators $\hat{\rho}_1$ of $\tau_1(\boldsymbol{X}, N, M)$. We will then see that the determined lower bound essentially matches with the upper bound (9). In the sequel we refer to $\mathscr{E}(\lambda, n)$ as the (normalized) minimax risk. The theorem provides us with a lower bound for $\mathscr{E}(\lambda, n)$.

Theorem 3 *Assume that* $\liminf_{n\to+\infty}(1+\lambda) > e^2$. *Then, there exists a universal constant* $K > 0$ *such that, for any* n *sufficiently large, we have that*

$$\mathscr{E}(\lambda, n) \geq K \cdot \begin{cases} 1 & \text{if } \lambda + 1 > \log(n) \\ \frac{1+\lambda}{\log(n)} \left(\frac{\sqrt{\log(n)}}{n(1+\lambda)} \right)^{e^2/(1+\lambda)} & \text{if } \lambda + 1 \leq \log(n) \end{cases} \tag{15}$$

According to Theorem 3, it is clear that the lower bound on the (normalized) minimax risk goes to zero if $\lambda + 1 = o(\log(n))$ and the rate is provided by the following Corollary.

Corollary 2 *Assume that* $1 + \lambda > e^2$. *Then there exist universal constants* $c > 0$ *and* $c' > 0$ *such that, for any* n *sufficiently large, we have that*

$$\mathscr{E}(\lambda, n) \geq c \frac{1}{n^{c'/\lambda}}. \tag{16}$$

Corollary 2 is a consequence of Theorem 3, indeed, when $\lambda + 1 > \log(n)$ the two lower bounds in (15) and (16) are constants, whereas if $\lambda + 1 \leq \log(n)$ it is easy to observe that the leading term in (15), as $n \to +\infty$, is of order $1/n^{c'/\lambda}$ as in (16) for some $c' > 0$. One may easily see that every constant $c' > e^2$ works in (16). Corollary 2 provides us with a lower bound for the NMSE of any estimator of the disclosure risk $\tau_1(\boldsymbol{X}, N, M)$. The lower bound (16) has an important implication: without imposing any parametric assumption on the model, one can estimate $\tau_1(\boldsymbol{X}, N, M)$ with vanishing NMSE all the way up to $\lambda \propto \log n$. It is then impossible to determine an estimator having provable guarantees, in terms of vanishing NMSE, when $\lambda = \lambda(n)$ goes to $+\infty$ much faster than $\log(n)$, as a function of n. By the "limit of predictability" (10) determined for the estimator $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$, we conclude that the proposed estimator is optimal, because its "limit of predictability" matches (asymptotically) with its maximum possible value $\lambda \propto \log(n)$.

3.1 Guideline for the proof of Theorem 3

We present the main ingredients for the proof of Theorem 3. Hereafter we will write $\mathbb{E}_P^{n,\lambda}$ (resp. $\mathbb{P}_P^{n,\lambda}$) in order to make explicit the dependence of the expected value (resp. the probability measure) w.r.t. P, the parameter n of the Poisson random variable N and λ. The proof of

Theorem 3 relies on the method of the two fuzzy hypotheses (Tsybakov (2009)), which allows to reduce the proof of Theorem 3 to the problem of finding the best polynomial approximation to some functions. A similar approach has been recently considered by Wu and Yang (2016, 2019) in the context of nonparametric estimation of the support size of discrete distributions. Some steps of the proof of Theorem 3 are similar to that of Wu and Yang (2016), and therefore they are omitted here, in favor of highlighting only the key differences. For the sake of completeness, the whole proof is offered in the online supplementary material.

Lemma 1 and Lemma 2 below are used in the proof of Theorem 3, and they constitutes the essential difference between the proof of Theorem 3 and the proof of the minimax lower bound in the work of Wu and Yang (2016). Lemma 1 and Lemma 2 are proved in Appendix B.1 and Appendix B.2, respectively.

Lemma 1 *The following identity holds true*

$$\mathscr{E}(\lambda, n) = \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}[(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}(\boldsymbol{Y}(\boldsymbol{X}, N)))^2],$$

where the infinimum in the previous equation is understood to be taken with respect to all measurable maps $\hat{\rho} : \mathbb{N}^{\mathbb{N}} \to \mathbb{R}$.

Remark that the definition of the minimax risk in (14) allows for estimators depending on the whole sample $\boldsymbol{X}(N)$, while $\tau_1(\boldsymbol{X}, N, M)$ depends only on the frequencies $\boldsymbol{Y}(\boldsymbol{X}, N+M)$ and $\boldsymbol{Y}(\boldsymbol{X}, N)$. Thus, in view of Lemma 1, there should be no gain of information in using estimators depending on $\boldsymbol{X}(N)$ over estimators depending only on the frequencies $\boldsymbol{Y}(\boldsymbol{X}, N)$. Investigation of the proof of Lemma 1 shows that for all estimators $\hat{\tau}_1$, the estimator $\hat{\rho}$ obtained by symmetrizing $\hat{\tau}_1$ and taking the expectation conditional on $\boldsymbol{Y}(\boldsymbol{X}, N)$ has always risk smaller or equal than $\hat{\tau}_1$. This may be viewed as a form of *Rao-Blackwellisation* of $\hat{\tau}_1$, where $\boldsymbol{Y}(\boldsymbol{X}, N)$ acts as a sufficient statistics for τ_1, in the sense that $\hat{\rho}$ never depends on the distribution of $\boldsymbol{X}$[1].

Besides being of self-interest for the reasons previously invoked, Lemma 1 crucially makes the proof of Theorem 3 easier by remarking that $(\boldsymbol{X}, k) \mapsto \boldsymbol{Y}(\boldsymbol{X}, k)$ is nicely distributed under the Poisson model. The Lemma 1 constitutes the starting point of the proof of Theorem 3. The rest of the proof consists on applying the reduction scheme of Wu and Yang (2019) Wu and Yang (2016) to the expression in Lemma 1. The major difference with the aforementioned paper is that we have to find the best, uniform on some interval, polynomial approximation of the map $x \mapsto \exp(-2Bx)$ for arbitrary $B > 0$ instead of the map $x \mapsto \log(x)$ considered in Wu and Yang (2016).

To be more precise, for $a, b \in \mathbb{R}$, we let $\mathsf{C}[a, b]$ denote the space of continuous functions on $[a, b]$, and for any $L \in \mathbb{Z}_+$ we let $\mathsf{P}_L[a, b] \subset \mathsf{C}[a, b]$ denote the space of polynomials of degree no more than L on $[a, b]$. For any $f \in \mathsf{C}[a, b]$, the best polynomial (of degree at most L) approximation to f is defined as

$$E_L(f, [a, b]) := \inf\{\sup\{|f(x) - q(x)| : \ x \in [a, b]\} : \ q \in \mathsf{P}_L[a, b]\}. \tag{17}$$

Then, our main result on the best, uniform on some interval, polynomial approximation of the of the map $x \mapsto \exp(-2Bx)$, is stated in the following lemma, proved in Appendix B.2. The rate of approximation is given in term of the function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\varphi(x) := 1 - \sqrt{1 + x^2} + x \operatorname{arcsinh}(x). \tag{18}$$

[1] We want to emphasize that τ_1 is not a parameter of the model, and thus the notion of a sufficient statistics is here ambiguous.

Lemma 2 *Let $\xi > 1$ and $g_\xi : [\xi^{-1}, 1] \to \mathbb{R}_+$ be such that $g_\xi(x) := \exp\{-2B_\xi x\}$ with $B_\xi = (\xi/2)(1 + O(\xi^{-1}))$ as $\xi \to \infty$. Then, for every $\zeta > 0$, there exist constants $K, \xi_0 > 0$ such that for all $\xi > \xi_0$ and all $0 < L \leq \zeta\xi$,*

$$E_L(g_\xi, [\xi^{-1}, 1]) \geq K \cdot \begin{cases} 1 & \text{if } 0 < L \leq \sqrt{\xi/2}, \\ \frac{\sqrt{\xi}}{L} \exp\big\{ - \frac{\xi}{2}\varphi\big(\frac{2L}{\xi}\big)\big\} & \text{if } \sqrt{\xi/2} < L < \zeta\xi. \end{cases}$$

It is worth discussing how the previous result can be of interest beyond its use in this paper. Approximation theory usually focuses on the regime $L/\xi \to \infty$, where the error of approximation is known to be super-exponential in L. This regime is omitted here since it is a classical result and we only need the regime $L/\xi \to \gamma$ for some constant $\gamma \geq 0$ in the proof of Theorem 3. Approximation in the latter regime is much more difficult, as emphasized by Lemma 2, and was not studied before to the best of our knowledge.

The proof of Lemma 2 uses the well-known duality between best polynomial approximation and best trigonometric polynomial approximation. Using the orthogonality of trigonometric polynomials, we are able to reduce the problem into finding a good lower bound on $\max_{K\in\mathbb{N}} Ke^{-C} I_{L+4K}(C)$, where I_k are the modified Bessel function of the first kind (see (Olver et al., 2010, pg. 248)), and $C \approx \xi/2$. Then, the most delicate and final step consist on establishing the double asymptotic of $I_k(C)$ as $k \to \infty$ and $C \to \infty$, with the constraint that $\sqrt{C} \leq k \lesssim C$.

Finally, we note that the lower bound in Lemma 2 is essentially sharp, *i.e.*, up to determining the value of the constant K. The matching upper-bound is derived in the supplementary material by analyzing the rate of convergence of Chebychev polynomials approximation of increasing orders[2].

4 Discussion

Skinner and Elliot (2002) first raised the problem of nonparametric estimation of τ_1 under the Poisson abundance model for sample records, and they left that as an open problem in the field of disclosure risk assessment. In this paper we first considered the problem of Skinner and Elliot (2002), and we presented a rigorous solution to it. In particular, we introduced a class of nonparametric estimators of τ_1, and we gave uniform theoretical guarantees for them. Firstly, we showed that our estimators provably estimate τ_1 all of the way up to the sampling fraction $(\lambda + 1)^{-1} \propto (\log n)^{-1}$, with vanishing NMSE as n becomes large. Secondly, and most importantly, we proved that: i) $(\lambda + 1)^{-1} \propto (\log n)^{-1}$ is the smallest possible sampling fraction of the population for consistently estimating τ_1; ii) estimators' NMSE is near optimal, in the sense of matching the minimax lower bound, for large n. Besides being the first study on nonparametric inference for τ_1 under the Poisson abundance model, our work is the first to provide theoretical guarantees on the estimation of τ_1. Indeed, despite the large number of contributions to the estimation of τ_1, all of them proposed parametric and semiparametric approaches that empirically estimate τ_1, but without provable guarantees. In particular, to be best of our knowledge, none of the contributions considers a rigorous study on the interplay between the estimation of τ_1 and λ.

The problem of estimating τ_1 belongs to a broad class of discrete functional estimation problems, commonly known as species sampling problems. Consider a population of individuals $(X_i)_{i\geq 1}$ belonging to different "species" $(S_j)_{j\geq 1}$ with unknown proportions $(p_j)_{j\geq 1}$. Given an initial observable samples of size n from the population, species sampling problems refer to

[2] The upper-bound is given for completeness, but it is not needed for the purpose of establishing the minimax lower bound.

the estimation of features of the population or features of λn additional unobservable samples. Recent noteworthy works on species sampling problems are concerned with the estimation of the following discrete functionals: support size (e.g., Valiant and Valiant (2013) and Wu and Yang (2019)); entropy (e.g., Jiao et al. (2015) and Wu and Yang (2016)); missing mass (e.g., Ohannessian and Dahleh (2012), Mossel and Ohannessian (2019) and Ben-Hamou et al. (2017)); number of unseen species (e.g. Efron and Thisted (1976) and Orlitsky et al. (2016)). Interest in these quantities first appeared in ecology, and it has grown in the recent years driven by challenging applications in biosciences, physical sciences, machine learning, engineering, theoretical computer science, information theory, etc. Our study on τ_1 contributes to these recent literature, by studying a new discrete functional of interest in the context of disclosure risk assessment.

While τ_1 is known to be the most common measure of disclosure risk (Bethlehem et al. (1990) and Skinner et al. (1994)), one might consider alternative measures by broadening the definition of "uniqueness". For instance, Fienberg and Makov (1998) considered a measure of disclosure risk defined in terms of the number of cells with frequency less or equal than 2. In general, one may consider

$$\tau_{r_N,r_M}(\boldsymbol{X},N,M)=\sum_{j\geq 1}\mathbb{1}_{\{Y_j(\boldsymbol{X},N)\leq r_N\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)\leq r_M\}},$$

namely the number of cells with sample frequency less or equal than r_N which have population frequency less or equal than r_M. A nonparametric estimator of τ_{r_N,r_M} and an upper bound for the corresponding NMSE can be derived along lines similar to those applied in this paper for τ_1. Regarding a lower bound on the NMSE, however, things get more challenging. Technically, the main difference would be in the approximation theory involved. Instead of finding the best (uniform) polynomial approximation to $x\mapsto\exp\{-Bx\}$ on some interval, we would have to find the best polynomial approximation to $x\mapsto q(x)\exp\{-Bx\}$ where q is some polynomial. As we are concerned with lower bounds, this turns out to be a much more challenging problem. The interest in τ_{r_N,r_M} is not only motivated in context of disclosure risk assessment, but also in the broad area of biosciences. Indeed, the discrete functional τ_{0,r_M} corresponds to the number of unseen rare species in additional unobservable samples, which is a natural refinement of the number of unseen species considered in Orlitsky et al. (2016). Work on these problems is ongoing.

A Nonparametric estimators of the disclosure risk: proofs

For the sake of simplifying notations, throughout this section we write τ_1 instead of $\tau_1(\boldsymbol{X},N,M)$, $\hat{\tau}_1$ instead of $\hat{\tau}_1(\boldsymbol{X}(N),N)$, and $\hat{\tau}_1^L$ instead of $\hat{\tau}_1^L(\boldsymbol{X}(N),N)$.

A.1 Details for the determination of the estimator (2)

First observe that, according to the definition of τ_1, we can write the following identities

$$\mathbb{E}[Z_i(\boldsymbol{X},N)]=\sum_{j\geq 1}\mathbb{P}(Y_j(\mathbf{X},N)=i)=\sum_{j\geq 1}e^{-np_j}\frac{(np_j)^i}{i!}. \tag{A.1}$$

Then $\mathbb{E}[\tau_1]=\sum_{j\geq 1}\mathbb{P}(Y_j(\mathbf{X},N)=1)\mathbb{P}(Y_j(\mathbf{X},N+M)-Y_j(\mathbf{X},N)=0)=\sum_{j\geq 1}np_je^{-np_j}e^{-\lambda np_j}$, and by a direct application of Taylor series expansion of the exponential function $e^{-\lambda np_j}$, for any $j \geq 1$, we can write the following expression

$$\mathbb{E}[\tau_1] = \sum_{i\geq 0} \frac{(-1)^i \lambda^i}{i!} \sum_{j\geq 1} (np_j)^{i+1} e^{-np_j} = \sum_{i\geq 0} (-1)^i \lambda^i (i+1) \mathbb{E}[Z_{i+1}(\boldsymbol{X}, N)],$$

where the last equality follows from a direct application of the identity displayed in (A.1).

A.2 Empirical Bayes approach to determine (3)

The estimator $\hat{\tau}_1$ admits a natural interpretation as a nonparametric empirical Bayes estimator in the sense of Robbins (1956), i.e., it is the posterior expectation of $\mathbb{E}[\tau_1]$ with respect to an empirical nonparametric prior distribution on the unknown p_j's. Specifically, note that $\mathbb{E}[\tau_1] = \sum_{j=1}^{+\infty} e^{-(\lambda+1)np_j} np_j$, and assume that the p_j's are independent and distributed according to the empirical cumulative distribution function $G(p)$ of $p_{i_1}, \dots, p_{i_k}$, corresponding to the k distinct cells arising from the cross classification of the initial sample, namely $G(p) := k^{-1} \sum_{1\leq t\leq k} \mathbb{1}_{\{p_{i_t} \leq p\}}$. Consider a cell j containing x individuals out of the initial sample of size N, where $x \geq 0$, then from Equation (9) of Robbins (1956)

$$\varphi_n(x) := \frac{\int e^{-(\lambda+1)np} npe^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)}{\int e^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)} \tag{A.2}$$

is the Bayes estimator of the quantity $e^{-(\lambda+1)np_j} np_j$ appearing $\mathbb{E}[\tau_1]$, for a cell j which contains x individuals out of the initial sample of size N. Now, rewrite $\varphi_n(x)$ as

$$\begin{aligned} \varphi_n(x) &= \frac{\int e^{-(\lambda+1)np} npe^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)}{\int e^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)} \\ &= \frac{\sum_{i\geq 0} \frac{(-(\lambda+1))^i}{i!x!} (x+i+1)! \int \frac{(np)^{x+i+1}}{(x+i+1)!} e^{-np} G(\mathrm{d}p)}{\int e^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)} \\ &= \frac{\sum_{i\geq 0} \frac{(-(\lambda+1))^i}{i!x!} (x+i+1)! \mathbb{E}[Z_{x+i+1}(\boldsymbol{X}, N)]}{\mathbb{E}[Z_x(\boldsymbol{X}, N)]}. \end{aligned}$$

Then the nonparametric Bayes estimator of $\mathbb{E}[\tau_1]$ is obtained summing up over all the possible cross classification of the observed cells, where we replace $\mathbb{E}[Z_x(\boldsymbol{X}, N)]$ by their empirical counterparts $Z_x(\boldsymbol{X}, N)$. Specifically, we can write the following

$$\begin{aligned} \hat{\tau}_1 &= \sum_{x\geq 0} Z_x(\boldsymbol{X}, N) \frac{\sum_{i\geq 0} \frac{(-(\lambda+1))^i}{i!x!} (x+i+1)! Z_{x+i+1}(\boldsymbol{X}, N)}{Z_x(\boldsymbol{X}, N)} \\ &= \sum_{i\geq 0} (i+1) Z_{i+1}(\boldsymbol{X}, N) \sum_{x=0}^{i} \frac{i!}{(i-x)!x!} (-(\lambda+1))^{i-x} \\ &= \sum_{i\geq 0} (-1)^i \lambda^i (i+1) Z_{i+1}(\boldsymbol{X}, N), \end{aligned}$$

which coincides with the estimator (3) obtained by means of the identity displayed in (3).

A.3 Proof of Theorem 1

Because of the independence of the random variables $\{Y_j(\boldsymbol{X}, N)\}_{j\geq 1}$, we may write the variance $\mathrm{Var}(\tau_1 - \hat{\tau}_1)$ as follows

$$
\begin{aligned}
&\mathrm{Var}(\tau_1 - \hat{\tau}_1)\\
&= \sum_{j\geq 1} \mathrm{Var}\left(\sum_{i\geq 0}(-1)^i(i+1)\lambda^i \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right)\\
&\quad = \sum_{j\geq 1} \mathbb{E}\left[\sum_{i\geq 0}(-1)^i(i+1)\lambda^i \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right]^2\\
&\quad = \sum_{j\geq 1} \mathbb{E}\left[\sum_{i\geq 1} a_i \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} + \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\left(a_0 - \mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right)\right]^2,
\end{aligned}
$$

where we have defined $a_i := (-1)^i(i+1)\lambda^i$. Now, observe that the events $\{(Y_j(\boldsymbol{X}, N) = i)\}_{i\geq 1}$ are all disjoint, hence the variance $\mathrm{Var}(\tau_1 - \hat{\tau}_1)$ may be rewritten as

$$
\begin{aligned}
&\sum_{j\geq 1} \mathbb{E}\left[\sum_{i\geq 1} a_i^2 \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} + \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\left(a_0 - \mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right)^2\right]\\
&\qquad = \sum_{j\geq 1} \mathbb{E}\left[\sum_{i\geq 0} a_i^2 \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right]
\end{aligned}
$$

observing that $a_0 = 1$. Thus, simple calculations show that we can bound $\mathrm{Var}(\tau_1 - \hat{\tau}_1)$ as

$$
\begin{aligned}
\mathrm{Var}(\tau_1 - \hat{\tau}_1) &\leq \max_{j\geq 0} |a_j|^2 \mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] - \sum_{j\geq 1} e^{-n(\lambda+1)p_j} np_j\\
&= \max_{i\geq 0} |a_i|^2 \mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] - \frac{1}{\lambda+1}\mathbb{E}[Z_1(\boldsymbol{X}, N+M)]. \qquad \text{(A.3)}
\end{aligned}
$$

It remains to show that the a_i's have a maximum for $\lambda < 1$, which is attained when $i = i^* := \lfloor (2\lambda - 1)/(1-\lambda)\rfloor \vee 0$. Hence the thesis follows by (A.3), since $\max_{i\geq 0} |a_i| = \Psi(\lambda)$.

A.4 Proof of Theorem 2

First we focus on the determination of the bound (6), concerning the bias. Remember the definition of both $\hat{\tau}_1^L$ and τ_1 to write

$$
\mathbb{E}[\hat{\tau}_1^L - \tau_1] = -\mathbb{E}\left[\sum_{i\geq 0}(-1)^i(i+1)\lambda^i \mathbb{P}(L \leq i-1) Z_{i+1}(\boldsymbol{X}, N)\right]
$$

where we have observed that non–smoothed estimator $\hat{\tau}_1$ is unbiased. It is now easy to see that

$$
\mathbb{E}[\hat{\tau}_1^L - \tau_1] = -\mathbb{E}\left[\sum_{i\geq 0}(-1)^i(i+1)\lambda^i \mathbb{P}(L \leq i-1) Z_{i+1}(\boldsymbol{X}, N)\right]
$$

$$
\begin{aligned}
&= -\mathbb{E}\left[\sum_{i\geq 1}(-1)^i(i+1)\lambda^i\mathbb{P}(L\leq i-1)\sum_{j\geq 1}\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}}\right]\\
&= -\sum_{i\geq 1}\sum_{j\geq 1}(-1)^i(i+1)\lambda^i\mathbb{P}(L\leq i-1)\mathbb{P}(Y_j(\boldsymbol{X},N)=i+1)\\
&= -\sum_{i\geq 1}\sum_{j\geq 1}(-1)^i(i+1)\lambda^i\mathbb{P}(L\leq i-1)e^{-np_j}\frac{(np_j)^{i+1}}{(i+1)!}\\
&= -\sum_{j\geq 1}e^{-np_j}np_j\sum_{i\geq 1}(-1)^i\frac{(\lambda np_j)^i}{i!}\mathbb{P}(L\leq i-1). \qquad \text{(A.4)}
\end{aligned}
$$

Now we focus on the evaluation of the sum with respect to i. If we set $y := \lambda np_j$ then

$$
\sum_{i\geq 1}\frac{(-y)^i}{i!}\mathbb{P}(L\leq i-1) = \sum_{i=1}^{+\infty}\frac{(-y)^i}{i!}\sum_{k=0}^{i-1}\mathbb{P}(L=k) = \sum_{k=0}^{+\infty}\mathbb{P}(L=k)\sum_{i=k+1}^{+\infty}\frac{(-y)^i}{i!}
$$

and remembering the definition of the incomplete gamma function we obtain that

$$
\begin{aligned}
\sum_{i\geq 1}(-1)^i\frac{y^i}{i!}\mathbb{P}(L\leq i-1) &= \sum_{k=0}^{+\infty}\mathbb{P}(L=k)\frac{e^{-y}}{k!}\int_0^{-y}\tau^k e^{-\tau}\mathrm{d}\tau\\
&= -\sum_{k=0}^{+\infty}\mathbb{P}(L=k)\frac{e^{-y}}{k!}\int_0^{y}(-s)^k e^{s}\mathrm{d}s\\
&= -e^{-y}\int_0^y e^s\mathbb{E}_L\left[\frac{(-s)^L}{L!}\right]\mathrm{d}s.
\end{aligned}
$$

Putting the previous expression in (A.4) and observing that $y = \lambda np_j$, (6) immediately follows. Now, in order to bound the variance of the difference between τ_1 and its estimator $\hat{\tau}_1^L$, recall that $\{Y_j(\boldsymbol{X},N)\}_{j\geq 1}$ are independent. Then,

$$
\begin{aligned}
\mathrm{Var}(\hat{\tau}_1^L-\tau_1) &= \mathrm{Var}\Bigg(\sum_{i\geq 0}(-1)^i(i+1)\lambda^i Z_{i+1}(\boldsymbol{X},N)\mathbb{P}(L\geq i)\\
&\qquad\qquad - \sum_{j=1}^{+\infty}\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\Bigg)\\
&= \sum_{j=1}^{+\infty}\mathrm{Var}\Big(\sum_{i=0}^{+\infty}(-1)^i(i+1)\lambda^i\mathbb{P}(L\geq i)\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}}\\
&\qquad\qquad - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\Big)\\
&= \sum_{j=1}^{+\infty}\mathrm{Var}\left(\sum_{i=0}^{+\infty}a_i\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right),
\end{aligned}
$$

having defined $a_i := (-1)^i(i+1)\lambda^i\mathbb{P}(L\geq i)$ for any $i\geq 0$. Therefore, we can write

$$
\mathrm{Var}(\hat{\tau}_1^L-\tau_1)
$$

$$
\begin{aligned}
&\leq \sum_{j=1}^{+\infty} \mathbb{E}\left[\left(\sum_{i=0}^{+\infty} a_i \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right)^2\right] \\
&= \sum_{j=1}^{+\infty} \mathbb{E}\left[\sum_{i=1}^{+\infty} a_i^2 \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} + \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}(a_0 - \mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}})^2\right]
\end{aligned}
$$

where we have used the incompatibility of the events $\{(Y_j(\boldsymbol{X}, N) = i)\}$ for different values of j. We can proceed with the upper bound for the variance as follows

$$
\begin{aligned}
&\mathrm{Var}(\hat{\tau}_1^L - \tau_1) \\
&\quad = \sum_{j=1}^{+\infty} \mathbb{E}\left[\sum_{i=0}^{+\infty} a_i^2 \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right] \\
&\quad \leq \max_{i\geq 0} |a_i|^2 \mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] - \sum_{j=1}^{+\infty} \mathbb{E}\left[\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right] \\
&\quad = \max_{i\geq 0} |a_i|^2 \mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] - \sum_{j=1}^{+\infty} e^{-\lambda n p_j} e^{-n p_j} n p_j \\
&\quad = \max_{i\geq 0} |a_i|^2 \mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] - \frac{1}{\lambda+1}\mathbb{E}[Z_1(\boldsymbol{X}, N+M)]. \qquad \text{(A.5)}
\end{aligned}
$$

Now, let observe that we can estimate the maximum value of the $|a_i|$'s as follows

$$
\begin{aligned}
\max_{i\geq 0} |a_i| &= \max_{i\geq 0}(i+1)\lambda^i \mathbb{P}(L \geq i) = \max_{i\geq 0}(i+1)\lambda^i \sum_{k=i}^{+\infty} \mathbb{P}(L = k) \\
&\leq \max_{i\geq 0} \sum_{k=i}^{+\infty}(i+1)\lambda^i \mathbb{P}(L = k) \leq \sum_{k=0}^{+\infty}(k+1)\lambda^k \mathbb{P}(L = k) \\
&= \mathbb{E}_L[(L+1)\lambda^L].
\end{aligned}
$$

Replacing $\max_{i\geq 0} |a_i|$ with $\mathbb{E}_L[(L+1)\lambda^L]$ in (A.5), the upper bound of $\mathrm{Var}(\hat{\tau}_1^L - \tau_1)$ becomes

$$
\mathrm{Var}(\hat{\tau}_1^L - \tau_1) \leq (\mathbb{E}_L[(L+1)\lambda^L])^2 \mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] - \frac{\mathbb{E}[Z_1(\boldsymbol{X}, N+M)]}{\lambda+1}.
$$

The proof is completed by putting together the previous upper bound for the variance and the one for the bias (6), from which the bound on the MSE (7) easily follows.

A.5 Proof of Proposition 1

To prove (8) we use Theorem 2, bounding the two terms appearing in (7) separately. In order toobtain an estimate of first term on the right-hand side of (7), we note that for any $y > 0$ the following holds

$$
\begin{aligned}
-e^{-y}\int_0^y e^s \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right]\mathrm{d}s &= -e^{-y}\int_0^y e^s \sum_{k=0}^{+\infty} e^{-\beta}\frac{\beta^k}{k!}\frac{(-s)^k}{k!}\mathrm{d}s \\
&= -e^{-y-\beta}\int_0^y e^s \sum_{k=0}^{+\infty} \frac{(\beta s)^k(-1)^k}{\Gamma(k+1)k!}\mathrm{d}s
\end{aligned}
$$

Recall that the Bessel polynomial (see Olver et al. (2010)) is defined as $J_0(z) := \sum_{k=0}^{+\infty} \frac{(-1)^k z^{2k}}{2^{2k}\Gamma(k+1)k!}$, and that $|J_0(z)| \le 1$. Therefore, we obtain the following inequality

$$\left| -e^{-y} \int_0^y e^s \mathbb{E}_L \left[\frac{(-s)^L}{L!} \right] \mathrm{d}s \right| \le e^{-(y+\beta)} \int_0^y e^s |J_0(2\sqrt{s\beta})| \mathrm{d}s \le e^{-\beta}(1 - e^{-y}).,$$

which may be applied to bound the first term on the right-hand side of (7), with $y = \lambda n p_j$. Precisely,

$$\begin{aligned} &\left| \sum_{j\ge1} e^{-p_j n(\lambda+1)} p_j n \int_0^{\lambda n p_j} e^s \mathbb{E}_L \left[\frac{(-s)^L}{L!} \right] \mathrm{d}s \right| \\ &\qquad \le \sum_{j\ge1} e^{-np_j} np_j e^{-\beta}(1 - e^{-\lambda n p_j}) \le e^{-\beta} \sum_{j=1}^{+\infty} e^{-np_j} np_j \\ &\qquad = e^{-\beta}\mathbb{E}[Z_1(\boldsymbol{X}, N)] \le e^{-\beta}\mathbb{E}[N] = e^{-\beta} n. \end{aligned} \tag{A.6}$$

In order to upper bound the other term on the right-hand side of (7), we observe that

$$\begin{aligned} \mathbb{E}_L[(L+1)\lambda^L] = \sum_{k=0}^{+\infty} e^{-\beta} \frac{\beta^k}{k!} \lambda^k (k+1) &= e^{-\beta} \left(\sum_{k=1}^{+\infty} \frac{(\beta\lambda)^k}{(k-1)!} + \sum_{k=0}^{+\infty} \frac{(\beta\lambda)^k}{k!} \right) \\ &= e^{-\beta}(e^{\beta\lambda} + \beta\lambda e^{\beta\lambda}) = e^{\beta(\lambda-1)}(1+\beta\lambda), \end{aligned}$$

hence we get

$$\begin{aligned} (\mathbb{E}_L[(L+1)\lambda^L])^2 \mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] - \frac{1}{\lambda+1}\mathbb{E}[Z_1(\boldsymbol{X}, N+M)] \\ \le n e^{2\beta(\lambda-1)}(1+\beta\lambda)^2. \end{aligned} \tag{A.7}$$

Using (A.6) and (A.7), one can now estimate the MSE (7) in the Poisson case and (8) follows. Because of (8) the NMSE can be bounded from above by

$$\mathscr{E}_{n,\lambda}(\hat{\tau}_1^L) \le e^{-2\beta} + \frac{e^{2\beta(\lambda-1)}(1+\beta\lambda)^2}{n}$$

using the exponential inequality $1 + x \le e^x$ we get

$$\mathscr{E}_{n,\lambda}(\hat{\tau}_1^L) \le e^{-2\beta} + \frac{e^{2\beta(2\lambda-1)}}{n}. \tag{A.8}$$

It is easy to show that the right-hand side of (A.8) is minimized when β equals $\frac{1}{4\lambda}\log\left(\frac{n}{2\lambda-1}\right)$. Therefore, it is easy to observe that the inequality (A.8) becomes

$$\mathscr{E}_{n,\lambda}(\hat{\tau}_1^L) \le \frac{1}{n^{1/(2\lambda)}} \cdot \frac{2\lambda}{(2\lambda-1)^{1-1/(2\lambda)}} \tag{A.9}$$

hence the second bound (9) follows provided that $A(\lambda) := \frac{2\lambda}{(2\lambda-1)^{1-1/(2\lambda)}}$. Now we can prove the "limit of predictability" in the Poisson case, indeed thanks to (9) we have

$$\mathscr{E}_{n,\lambda}(\hat{\tau}_1^L) \le \frac{A}{n^{1/(2\lambda)}},$$

besides observe that

$$\frac{A}{n^{1/(2\lambda)}} \leq \delta$$

is satisfied if and only if $\lambda \leq \frac{\log(n)}{2\log(A/\delta)} =: \lambda^*$. As a consequence the maximum value of λ for which the inequality $\mathscr{E}_{n,\lambda}(\hat{\tau}_1^L) \leq \delta$ is satisfied, is bigger or equal than λ^*, i.e.,

$$\max\left\{\lambda : \ \mathscr{E}_{n,\lambda}(\hat{\tau}_1^L) \leq \delta\right\} \geq \frac{\log(n)}{2\log(A/\delta)}.$$

Then the thesis follows by taking the limit of the previous inequality as $n \to +\infty$.

B Proofs related to the lower bound

B.1 Proof of Lemma 1

First, it is obvious that

$$\mathscr{E}(\lambda, n) \leq \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}[(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}(\boldsymbol{Y}(\boldsymbol{X}, N)))^2].$$

We now prove that the previous is indeed an inequality by deriving a lower bound that matches. Let $n > 0$ be fixed. By definition, for every $\varepsilon > 0$ there exists an estimator $\hat{\rho}_1$ such that

$$\begin{aligned}
\mathscr{E}(\lambda, n) &\geq \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}[(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}_1(\boldsymbol{X}(N), N))^2] - \varepsilon \\
&= \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}[\mathbb{E}_P^{n,\lambda}[(\tau_1(\boldsymbol{X}, N, M) \\
&\qquad\qquad - \hat{\rho}_1(\boldsymbol{X}(N), N))^2 \mid \boldsymbol{Y}(\boldsymbol{X}, N), \boldsymbol{Y}(\boldsymbol{X}, N+M)]] - \varepsilon \\
&\geq \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}[(\tau_1(\boldsymbol{X}, N, M) - \mathbb{E}_P^{n,\lambda}[\hat{\rho}_1(\boldsymbol{X}(N), N) \mid \boldsymbol{Y}(\boldsymbol{X}, N)])^2] - \varepsilon
\end{aligned} \tag{B.1}$$

where the last line follows by Jensen's inequality and by observing that

$$\begin{gathered}
\mathbb{E}_P^{n,\lambda}[\tau_1(\boldsymbol{X}, N, M) \mid \boldsymbol{Y}(\boldsymbol{X}, N), \boldsymbol{Y}(\boldsymbol{X}, N+M)] = \tau_1(\boldsymbol{X}, N, M), \quad \text{and,} \\
\mathbb{E}_P^{n,\lambda}[\hat{\rho}_1(\boldsymbol{X}(N), N) \mid \boldsymbol{Y}(\boldsymbol{X}, N), \boldsymbol{Y}(\boldsymbol{X}, N+M)] = \mathbb{E}_P^{n,\lambda}[\hat{\rho}_1(\boldsymbol{X}(N), N) \mid \boldsymbol{Y}(\boldsymbol{X}, N)].
\end{gathered}$$

To see that the last equation is true, remark that $\boldsymbol{Y}(\boldsymbol{X}, N+M) - \boldsymbol{Y}(\boldsymbol{X}, N)$ is independent of $\boldsymbol{Y}(\boldsymbol{X}, N)$ and depends only on $(X_{N+1}, \ldots, X_{N+M})$. Now we claim that $\hat{\rho}_1$ can be chosen such that for any $k \in \mathbb{Z}_+$ and any permutation $\sigma_k(\boldsymbol{X}(k))$ of the data, it holds $\hat{\rho}_1(\boldsymbol{X}(k), k) = \hat{\rho}_1(\sigma_k(\boldsymbol{X}(k)), k)$. We delay the proof of the claim to later. Now assume the claim is true. Given k and $\boldsymbol{Y}(\boldsymbol{X}, k)$, we can construct the functional

$$G(\boldsymbol{Y}(\boldsymbol{X}, k), k) := (\underbrace{1, \ldots, 1}_{\times Y_1(\boldsymbol{X}, k)}, \underbrace{2, \ldots, 2}_{\times Y_2(\boldsymbol{X}, k)}, \ldots).$$

Since $\hat{\rho}_1$ is invariant under permutations of the data, we have for any $P \in \mathscr{P}$,

$$\begin{aligned}
&\mathbb{E}_P^{n,\lambda}[\hat{\rho}_1(\boldsymbol{X}(N), N) \mid \boldsymbol{Y}(\boldsymbol{X}, N)] \\
&\qquad = \mathbb{E}_P^{n,\lambda}\big[\mathbb{E}_P^{n,\lambda}[\hat{\rho}_1(\boldsymbol{X}(N), N) \mid \boldsymbol{Y}(\boldsymbol{X}, N), N] \mid \boldsymbol{Y}(\boldsymbol{X}, N)\big] \\
&\qquad = \mathbb{E}_P^{n,\lambda}\big[\mathbb{E}_P^{n,\lambda}[\hat{\rho}_1(G(\boldsymbol{Y}(\boldsymbol{X}, N), N), N) \mid \boldsymbol{Y}(\boldsymbol{X}, N), N] \mid \boldsymbol{Y}(\boldsymbol{X}, N)\big] \\
&\qquad = \mathbb{E}_P^{n,\lambda}[\hat{\rho}_1(G(\boldsymbol{Y}(\boldsymbol{X}, N), N), N) \mid \boldsymbol{Y}(\boldsymbol{X}, N)] \\
&\qquad = \hat{\rho}_1(G(\boldsymbol{Y}(\boldsymbol{X}, N), N), N).
\end{aligned}$$

The last line follows because $N = \sum_{j\geq 1} Y_j(\boldsymbol{X}, N)$, and hence N is completely determined by $\boldsymbol{Y}(\boldsymbol{X}, N)$. Therefore, we have proved that the conditional expected value of $\hat{\rho}_1(\boldsymbol{X}(N), N)$, given $\boldsymbol{Y}(\boldsymbol{X}, N)$ does not depend on P. Thus, (B.1) implies,

$$\begin{aligned}\mathscr{E}(\lambda, n) &\geq \sup_{P\in\mathscr{P}} n^{-2}\mathbb{E}_P^{n,\lambda}[(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}_1(G(\boldsymbol{Y}(\boldsymbol{X}, N), N), N))^2] - \varepsilon \\ &\geq \inf_{\hat{\rho}} \sup_{P\in\mathscr{P}} n^{-2}\mathbb{E}_P^{n,\lambda}[(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}(\boldsymbol{Y}(\boldsymbol{X}, N)))^2] - \varepsilon.\end{aligned}$$

Since the previous is true for all $\varepsilon > 0$, the conclusion follows.

We now prove the claim we have used in the previous argument, i.e. that *$\hat{\rho}_1$ can be chosen such for any $k \in \mathbb{Z}_+$ and any permutation $\sigma_k(\boldsymbol{X}(k))$ of the data, it holds $\hat{\rho}_1(\boldsymbol{X}(k), k) = \hat{\rho}_1(\sigma_k(\boldsymbol{X}(k)), k)$.* When $k = 0$, then the claim is trivial, hence we assume without loss of generality that $k \geq 1$. We will prove that for any estimator $\hat{\rho}_1$, there is a symmetric estimator $\hat{t}_1$ with a risk no more than the risk of $\hat{\rho}_1$. Let $\hat{\rho}_1$ be arbitrary. Construct $\hat{t}_1$ such that for any $k \in \mathbb{N}$

$$\hat{t}_1(\boldsymbol{X}(k), k) := \frac{1}{|\{\sigma_k\}|} \sum_{\{\sigma_k\}} \hat{\rho}_1(\sigma_k(\boldsymbol{X}(k)), k).$$

Clearly $\hat{t}_1$ has the desired invariance property under permutations. Moreover, by Jensen's inequality,

$$\begin{aligned}&\mathbb{E}_P^{n,\lambda}[(\tau_1(\boldsymbol{X}, N, M) - \hat{t}_1(\boldsymbol{X}(N), N))^2] \\ &\qquad = \mathbb{E}_P^{n,\lambda}\Big[\mathbb{E}_P^{n,\lambda}\Big[\Big(\frac{1}{|\{\sigma_N\}|} \sum_{\{\sigma_N\}} (\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}_1(\sigma_N(\boldsymbol{X}(N)), N)\Big)^2 \mid N\Big]\Big] \\ &\qquad \leq \mathbb{E}_P^{n,\lambda}\Big[\mathbb{E}_P^{n,\lambda}\Big[\frac{1}{|\{\sigma_N\}|} \sum_{\{\sigma_N\}} (\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}_1(\sigma_N(\boldsymbol{X}(N)), N))^2 \mid N\Big]\Big]\end{aligned}$$

Now remark that for all $(k, k') \in \mathbb{Z}_+^2$ the map $\boldsymbol{X} \mapsto \tau_1(\boldsymbol{X}, k, k')$ is invariant under any permutations of the k first entries of $\boldsymbol{X}$. Moreover, $\boldsymbol{X}$ is an i.i.d. vector, then the last display implies that

$$\begin{aligned}&\mathbb{E}_P^{n,\lambda}[(\tau_1(\boldsymbol{X}, N, M) - \hat{t}_1(\boldsymbol{X}(N), N))^2] \\ &\qquad \leq \mathbb{E}_P^{n,\lambda}\Big[\mathbb{E}_P^{n,\lambda}\Big[\frac{1}{|\{\sigma_N\}|} \sum_{\{\sigma_N\}} (\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}_1(\boldsymbol{X}(N), N))^2 \mid N\Big]\Big] \\ &\qquad = \mathbb{E}_P^{n,\lambda}[(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}_1(\boldsymbol{X}(N), N))^2].\end{aligned}$$

The conclusion follows by taking the supremum over $P \in \mathscr{P}$ both sides of the last display.

B.2 Proof of Lemma 2

In the whole proof, we drop the subscripts ξ whenever it is convenient.

Let $\sigma : [-1, 1] \to [\xi^{-1}, 1]$ be such that $\sigma(x) := (1 - \xi^{-1})(x + 1)/2 + \xi^{-1}$. Notice that σ is bijective. By translating and rescaling, we claim that $E_L(g, [\xi^{-1}, 1]) = E_L(g \circ \sigma, [-1, 1])$. To see that this is true, remark that for all $p \in \mathsf{P}_L[-1, 1]$ we have $\|g \circ \sigma - p\|_\infty = \|g - p \circ \sigma^{-1}\|_\infty \geq E_L(g, [\xi^{-1}, 1])$. This shows that $E_L(g \circ \sigma, [-1, 1]) \geq E_L(g, [\xi^{-1}, 1])$. The same steps using σ^{-1} show that $E_L(g \circ \sigma, [-1, 1]) \leq E_L(g, [\xi^{-1}, 1])$. Hence $E_L(g, [\xi^{-1}, 1]) = E_L(g \circ \sigma, [-1, 1])$.

For the sake of simplicity, we let $C := B(1-\xi^{-1})$ and $\gamma_C : [-1,1] \to \mathbb{R}_+$ is defined by $\gamma_C(x) = \exp\{-C(x+1)\}$. From the discussion in the previous paragraph, we have indeed reduced the problem to finding $E_L(\gamma_C, [-1,1])$. This is because

$$\begin{aligned} E_L(g, [\xi^{-1}, 1]) = E_L(g \circ \sigma, [-1,1]) &= \exp\{-2B\xi^{-1}\} E_L(\gamma_C, [-1,1]) \\ &= e(1+o(1)) E_L(\gamma_C, [-1,1]). \end{aligned}$$

To find a lower bound on $E_L(\gamma_C, [-1,1])$, we will exploit the well-known relationship between uniform approximation on the interval by polynomials and uniform approximation of periodic even functions by trigonometric polynomials. We write $\mathsf{CE}[-1,1]$ the space of continuous and even functions on $[-1,1]$, and for any $L \in \mathbb{Z}_+$ we let $\mathsf{TP}_L[-1,1]$ denote the set of even trigonometric polynomials of degree at most L, *i.e.* $\mathsf{TP}_L[-1,1]$ is

$$\Big\{ T \in \mathsf{CE}[-1,1] : \; T(x) = \textstyle\sum_{k=0}^{L} a_k \cos(\pi k x), \; a_k \in \mathbb{R}, \; x \in [-1,1] \Big\}.$$

We furthermore define the periodization operator $P : \mathsf{C}[-1,1] \to \mathsf{CE}[-1,1]$ such that $Pf(\theta) = f(\cos(\pi\theta))$ for all $f \in \mathsf{C}[-1,1]$ and all $\theta \in [-1,1]$. Then, it is well-known (see for instance the Theorem 14.8.1 in Davidson and Donsig (2009(@)) that

$$E_L(\gamma_C, [-1,1]) = \inf\{\|P\gamma_C - T\|_\infty : \; T \in \mathsf{TP}_L[-1,1]\}. \tag{B.2}$$

We will now bound the right-hand side of (B.2) by a technique inspired from Newman and Rivlin (1976), which works as well for our setting. For any $K \in \mathbb{N}$, we define the trigonometric polynomial $T_K : [-1,1] \to \mathbb{C}$ such that

$$T_K(\theta) := e^{i\pi(L+1)\theta} \Big\{ \sum_{k=0}^{K-1} e^{i2\pi k\theta} \Big\}^2.$$

Then, by orthogonality of the trigonometric polynomials, we have that

$$\int_{-1}^{-1} |T_K(\theta)| \,\mathrm{d}\theta = \sum_{j=0}^{K-1} \sum_{k=0}^{K-1} \int_{-1}^{1} e^{i2\pi(j-k)\theta} \,\mathrm{d}\theta = 2K. \tag{B.3}$$

By definition, for every $\varepsilon > 0$ we can find a $Q \in \mathsf{TP}_L[-1,1]$ such that $\|P\gamma_C - Q\|_\infty \leq E_L(\gamma_C, [-1,1]) + \varepsilon$. Choose such Q, and remark that (B.3) implies,

$$\begin{aligned} \Big| \int_{-1}^{1} (P\gamma_C(\theta) - Q(\theta)) T_K(\theta) \,\mathrm{d}\theta \Big| &\leq \|P\gamma_C - Q\|_\infty \int_{-1}^{1} |T_K(\theta)| \,\mathrm{d}\theta \\ &\leq 2K\{E_L(\gamma_C, [-1,1]) + \varepsilon\}. \end{aligned}$$

On the other hand remark that Q is a trigonometric polynomial of degree at most L, while T_K is a trigonometric polynomial of degree strictly greater than L. Therefore Q is orthogonal to T_K. Moreover, the last display is true for all $\varepsilon > 0$ and for all $K \in \mathbb{N}$, thus it must be the case that

$$E_L(\gamma_C, [-1,1]) \geq \max_{K \in \mathbb{N}} \frac{1}{2K} \Big| \int_{-1}^{1} P\gamma_C(\theta) T_K(\theta) \,\mathrm{d}\theta \Big|. \tag{B.4}$$

Interestingly, we can compute the previous integral. Namely,

$$\int_{-1}^{1} P\gamma_C(\theta) T_K(\theta) \,\mathrm{d}\theta = \sum_{j=0}^{K-1} \sum_{k=0}^{K-1} \int_{-1}^{1} \gamma_C(\cos(\pi\theta)) e^{i\pi\theta(L+1+2j+2k)} \,\mathrm{d}\theta$$

$$= 2(-1)^{L+1}\sum_{j=0}^{K-1}\sum_{k=0}^{K-1} e^{-C} I_{L+1+2j+2k}(C),$$

where $I_\nu(z) := \frac{1}{\pi}\int_0^\pi e^{z\cos(t)}\cos(\nu t)\mathrm{d}t$ is the modified Bessel function (see (Olver et al., 2010, pg. 248)); in particular (Olver et al., 2010, formula 10.32.3). More precisely, from the above considerations and the fact that the modified Bessel functions are non–negative, we deduce that

$$\Big|\int_{-1}^{1} P\gamma_C(\theta)T_K(\theta)\,\mathrm{d}\theta\Big| = 2\sum_{j=0}^{K-1}\sum_{k=0}^{K-1} e^{-C} I_{L+1+2j+2k}(C).$$

Soni (1965) proved that $I_{k+1}(z) \le I_k(z)$ for all $k \in \mathbb{N}$ and all $z > 0$. Hence, we obtain from the last display and (B.4) the bound

$$E_L(\gamma_C, [-1,1]) \ge \max_{K\in\mathbb{N}} K e^{-C} I_{L+4K}(C). \tag{B.5}$$

In the next lemma, We obtain a bound on the modified Bessel function $z \mapsto I_k(z)$ which remains tighter than the classical bound derived in Luke (1972) when $z \ge k$. The proof of the lemma is to be found in Section B.3.

Lemma B.1 *Assume $k \in \mathbb{N}$ and assume that $C > 8\sqrt{1+(k/C)^2}$. Then,*

$$e^{-C} I_k(C) > \frac{\exp\{-C\varphi(k/C)\}}{2e^4(1+(k/C)^2)^{1/4}\sqrt{C}}.$$

For $\alpha, \beta \in \mathbb{R}$ to be chosen accordingly, we define $K_* := \alpha\sqrt{C}$ if $L < \sqrt{C}$, or $K_* := \beta C/L$ if $L \ge \sqrt{C}$. In view of (B.5), it is clear that $E_L(\gamma_C,[-1,1]) \ge K_* e^{-C} I_{L+4K_*}(C)$. Consider now the case where $L < \sqrt{C}$, then

$$0 \le \frac{L+4K_*}{C} = \frac{L+\alpha\sqrt{C}}{C} < \frac{\alpha+1}{\sqrt{C}}.$$

Thus, $C\varphi((L+4K_*)/C) = O(1)$ as $C \to \infty$, $(L+4K_*)/C \to 0$ as $C \to \infty$, and $C > 8\sqrt{1+(L+4K_*)^2/C^2}$ when C gets large enough. We then obtain from Lemma B.1 that in this case,

$$E_L(\gamma_C,[-1,1]) > \frac{\alpha\sqrt{C}(1+o(1))\exp\{-C\varphi((L+4K_*)/C)\}}{2e^2\sqrt{C}} \gtrsim 1,$$

at least for C large enough. We now consider the case $L \ge \sqrt{C}$. In this case, we have,

$$0 \le \frac{L+4K_*}{C} = \frac{L+\beta C/L}{C} = \frac{L}{C} + \frac{\beta}{L} \le \frac{L}{C} + \frac{\beta}{\sqrt{C}}.$$

Because by assumption there is a constant $\zeta > 0$ such that $L \le \zeta C$, then $(L+4K_*)/C \le \zeta + o(1)$ as $C \to \infty$, and thus we have $C > 8\sqrt{1+(L+4K_*)^2/C^2}$ when C is large enough. Then, we can apply Lemma B.1 to find that as $C \to \infty$,

$$\begin{aligned} E_L(\gamma_C,[-1,1]) &> \frac{(\beta C/L)(1+o(1))\exp\{-C\varphi((L+4K_*)/C)\}}{4e^2\sqrt{C}(1+(L/C)^2)^{1/4}} \\ &\gtrsim \sqrt{\frac{C}{L^2}}\exp\Big\{-C\varphi\Big(\frac{L}{C}+\frac{\beta}{L}\Big)\Big\}, \end{aligned}$$

at least for C large enough. Further, it can be seen that $|\varphi'(x)| \leq |x|$ (see for instance Section S6.3 in the supplementary material). Then, by Taylor expansion,

$$\varphi\Big(\frac{L}{C} + \frac{\beta}{L}\Big) \leq \varphi\Big(\frac{L}{C}\Big) + \Big(\frac{L}{C} + \frac{\beta}{L}\Big)\frac{\beta}{L},$$

and thus, $C\varphi(L/C + \beta/L) \leq C\varphi(L/C) + \beta(1 + C\beta/L^2) \leq C\varphi(L/C) + \beta(1+\beta)$. It follows,

$$E_L(\gamma_C, [-1,1]) \gtrsim \sqrt{\frac{C}{L^2}} \exp\Big\{ - C\varphi\Big(\frac{L}{C}\Big)\Big\}.$$

With similar arguments, $C\varphi(L/C) = (\xi/2)\varphi(2L/\xi) + O(1)$ as $\xi \to \infty$.

B.3 Proof of Lemma B.1

The proof relies on the well known series representation of the modified Bessel function (see (Olver et al., 2010, formula 10.25.2)), namely we have whenever $k \in \mathbb{N}$,

$$I_k(z) = \sum_{p=0}^{\infty} \frac{1}{p!(p+k)!}\Big(\frac{z}{2}\Big)^{2p+k}. \tag{B.6}$$

Conveniently, all the terms in the summation are non-negative, which we will exploit to get our lower bound. By Stirling's formula, when $k \geq 1$, for any $p \geq 0$

$$(p+k)! \leq e\sqrt{(p+k)}\exp\{-(p+k) + (p+k)\log(p+k)\},$$

and for any $p \geq 1$, we have $p! \leq e\sqrt{p}\exp\{-p + p\log p\}$. For convenience, let define the functions $\phi_{z,k} : \mathbb{R}_+^* \to \mathbb{R}_+$, such that for any $x, z \in \mathbb{R}_+^*$ and any $k \in \mathbb{N}$,

$$\phi_{z,k}(x) := -z + 2x + k - x\log x - (x+k)\log(x+k) + (2x+k)\log(z/2).$$

Hence, because each term in the series expansion of (B.6) is non-negative, we get the estimate,

$$e^{-z}I_k(z) \geq e^{-z}\sum_{p\geq 1}\frac{1}{p!(p+k)!}\Big(\frac{z}{2}\Big)^{2p+k} \geq \frac{1}{e^2}\sum_{p\geq 1}\frac{\exp\{\phi_{z,k}(p)\}}{\sqrt{p(p+k)}}. \tag{B.7}$$

Notice that,

$$\phi'_{z,k}(x) = -\log(x) - \log(x+k) + 2\log(z/2), \qquad \phi''_{z,k}(x) = -\frac{1}{x} - \frac{1}{x+k}.$$

Thus, $\phi_{z,k}$ admits a unique non-negative extremum at x_0 solution to $x_0(x_0+k) = z^2/4$, that is,

$$x_0 = \frac{-k + \sqrt{k^2+z^2}}{2}, \text{ and, } \quad \phi''_{z,k}(x_0) = -\frac{4}{z}\sqrt{1+(k/z)^2} < 0.$$

Henceforth x_0 is indeed the unique maximum of the function $\phi_{z,k}$ on $\mathbb{R}_+$. We let p_0 smallest integer larger than x_0. Then $p_0 \geq 1$ and we have, by Taylor expansion that for any $p \geq p_0$ there is a $\bar{p} \in (x_0, p)$

$$\begin{aligned}\phi_{z,k}(p) &= \phi_{z,k}(x_0) + \phi'_{z,k}(x_0)(p-x_0) + \frac{1}{2}\phi''_{z,k}(\bar{p})(p-x_0)^2 \\ &= \phi_{z,k}(x_0) + \frac{1}{2}\phi''_{z,k}(\bar{p})(p-x_0)^2.\end{aligned}$$

Remark that, because $\bar{p} \geq x_0$,

$$\phi''_{z,k}(\bar{p}) = -\frac{1}{\bar{p}} - \frac{1}{\bar{p}+k} \geq -\frac{1}{x_0} - \frac{1}{x_0+k} = -\frac{4}{z}\sqrt{1+(k/z)^2}.$$

Then, for any $p \geq p_0$,

$$\phi_{z,k}(p) \geq \phi_{z,k}(x_0) + \frac{1}{2}\phi''_{z,k}(x_0)(p-x_0)^2 = \phi_{z,k}(x_0) - \frac{2\sqrt{1+(k/z)^2}}{b}(p-x_0)^2.$$

Therefore,

$$e^{-z}I_k(z) \geq \frac{\exp\{\phi_{z,k}(x_0)\}}{e^2} \sum_{p \geq p_0} \frac{\exp\{\phi''_{z,k}(x_0)(p-x_0)^2/2\}}{\sqrt{p(p+k)}}.$$

Let p_1 be the largest integer such that $-\phi''_{z,k}(x_0)(p_1-x_0)^2 \leq 2$. Remark that whenever $z > 2(1+(k/z)^2)^{1/2}$, we have $p_1 \geq x_0+1$, which is always the case in the conditions of the lemma. Because the summand is the previous is monotonically decreasing for $p \geq p_0$, we get the bound,

$$e^{-z}I_k(z) \geq \frac{\exp\{\phi_{z,k}(x_0)\}}{e^4} \frac{(p_1-p_0)}{\sqrt{p_1(p_1+k)}} \geq \frac{\exp\{\phi_{z,k}(x_0)\}}{e^4} \frac{(p_1-x_0)-1}{\sqrt{p_1(p_1+k)}}.$$

But, by the definition of p_1, we have that $p_1+1-x_0 > \sqrt{2/(-\phi''_{z,k}(x_0))}$. Therefore, whenever $z > 8(1+(k/z)^2)^{1/2}$, by the definition of $\phi''_{z,k}(x_0)$,

$$\begin{aligned} e^{-z}I_k(z) &\geq \frac{\exp\{\phi_{z,k}(x_0)\}}{e^4\sqrt{-\phi''_{z,k}(x_0)p_1(p_1+k)}}\left\{\sqrt{2} - 2\sqrt{-\phi''_{z,k}(x_0)}\right\} \\ &\geq \frac{\sqrt{2}\exp\{\phi_{z,k}(x_0)\}}{2e^4\sqrt{-\phi''_{z,k}(x_0)p_1(p_1+k)}}. \end{aligned}$$

Also,

$$\begin{aligned} p_1(p_1+k) &= x_0(x_0+k) + (p_1^2 - x_0^2) + (p_1-x_0)k \\ &= x_0(x_0+k) + (p_1-x_0)(p_1+x_0+k) \\ &= x_0(x_0+k) + (p_1-x_0)^2 + (p_1-x_0)(2x_0+k). \end{aligned}$$

But we have that $x_0(x_0+k) = z^2/4$, $(p_1-x_0)^2 \leq -2/\phi''_{z,k}(x_0)$, and $2x_0+k = z\sqrt{1+(k/z)^2}$. Thus,

$$\begin{aligned} p_1(p_1+k) &\leq \frac{z^2}{4} + \frac{2}{-\phi''_{z,k}(x_0)} + \sqrt{\frac{2(1+(k/z)^2)}{-\phi''_{z,k}(x_0)}}z \\ &= \frac{z^2}{4} + \frac{z}{2\sqrt{1+(k/z)^2}} + \frac{z^{3/2}}{\sqrt{2}}[1+(k/z)^2]^{1/4} \\ &= \frac{z^2}{4}\left\{1 + \frac{z^{-1/2}[1+(k/z)^2]^{1/4}}{\sqrt{2}} + \frac{z^{-1}}{2\sqrt{1+(k/z)^2}}\right\}. \end{aligned}$$

Therefore, whenever $z > 8(1+(k/z)^2)^{1/2}$,

$$p_1(p_1+k) \leq \frac{z^2}{4}\left\{1 + \frac{1}{4} + \frac{1}{16}\right\} \leq \frac{21}{64}z^2 < \frac{z^2}{2}.$$

Hence,

$$e^{-z}I_k(z) > \frac{\exp\{\phi_{z,k}(x_0)\}}{e^4\sqrt{-\phi''_{z,k}(x_0)z}} = \frac{\exp\{\phi_{z,k}(x_0)\}}{2e^4(1+(k/z)^2)^{1/4}\sqrt{z}}.$$

After some algebra, we find that

$$\begin{aligned}\phi_{z,k}(x_0) &= -z + z\sqrt{1+(k/z)^2}\\ &\quad - (z/2)\{-(k/z)+\sqrt{1+(k/z)^2}\}\log\{-(k/z)+\sqrt{1+(k/z)^2}\}\\ &\quad - (z/2)\{(k/z)+\sqrt{1+(k/z)^2}\}\log\{(k/z)+\sqrt{1+(k/z)^2}\}\\ &= -z + z\sqrt{1+(k/z)^2} - z\cdot(k/z)\operatorname{arcsinh}(k/z) = -z\varphi(k/z).\end{aligned}$$

Acknowledgements

The authors are grateful to an Associate Editor and two anonymous Referees for all their comments, corrections, and suggestions which improved remarkably the paper. Federico Camerlenghi and Stefano Favaro received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257. Federico Camerlenghi and Stefano Favaro gratefully acknowledge the financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022.

References

Ben-Hamou, A., Boucheron, S. and Ohannessian, M.I. (2017). Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli* **23**, 249–287.

Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990). Disclosure control of microdata. *J. Amer. Statist. Assoc.* **85**, 38–45.

Carota, C., Filippone, M., Leombruni, R. and Polettini, S. (2015). Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models. *Ann. Appl. Statist.* **9**, 525–546.

Carota, C., Filippone, M. and Polettini, S. (2018). Assessing Bayesian nonparametric log-linear models: an application to disclosure risk estimation. *Preprint: arXiv:1801.05244*

Davidson, K. and Donsig, A. (2009). Real analysis and applications: theory in practice. *Springer Science and Business Media.*

Efron, B. and Morris, C (1973). Stein's estimation rule and its competitors - an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68**, 117–130.

Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63**, 435–447.

Fienberg, S.E. and Makov, U.E. (1998). Condentiality, uniqueness, and disclosure limitation for categorical data *J. Off. Stat.* **14**, 385–397.

Good, I.J. and Toulmin, G.H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43**, 45–63.

Jiao, J., Venkat, K., Han, Y. and Weissman, T. (2015). Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Th.* **61**, 2835–2885.

Luke, Y.L. (1972). Inequalities for generalized hypergeometric functions. *J. Approximation Theory*, **5**, 41–65.

Manrique-Vallier, D. and Reiter, J.P. (2012). Estimating identification disclosure risk using mixed membership models. *J. Amer. Statist. Assoc.* **107** 1385–1394.

Manrique-Vallier, D. and Reiter, J.P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *J. Comput. Graph. Statist.* **23** 1061–1079.

Mossel, E. and Ohannessian, M.I. (2019) On the impossibility of learning the missing mass. *Entropy* **21**, 28.

Newman, D.J. and Rivlin, T.J. (1976). Approximation of monomials by lower degree polynomials. *Aequationes Math.* **14**, 451–455.

Ohannessian, M.I. and Dahleh, M.A. (2012). Rare probability estimation under regularly varying heavy tails. *J. Mach. Learn. Res.* **23**, 1–24.

Olver, F.W.J., Lozier, D.W., Boisvert, R.F. and Clark, C.W. (2010). *NIST handbook of mathematical functions*, Cambridge University Press.

Orlitsky, A., Suresh, A.T. and Wu, Y. (2017). Optimal prediction of the number of unseen species. *Proc. Natl. Acad. Sci. USA* **113**, 13283–13288.

Reiter, J.P. (2005). Estimating risks of identification disclosure in microdata. *J. Amer. Statist. Assoc.* **100**, 1103–1112.

Rinott, Y. and Shlomo, N. (2006). A generalized negative binomial smoothing model for sample disclosure risk estimation. In *Privacy in Statistical Databases. Lecture Notes in Computer Science*, Springer, Berlin.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. 3rd Berkeley Symp.*,**1**, 157–163.

Samuels, S.M. (1998). A Bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *J. Off. Statist.* **14**, 373–383.

Skinner, C.J. and Elliot, M.J. (2002). A measure of disclosure risk for microdata. *J. Roy. Statist. Soc. B* **64**, 855–867.

Skinner, C., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata. *J. Off. Stat.* **10**, 31–51.

Skinner, and Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models. *J. Amer. Statist. Assoc.* **103**, 989–1001.

Soni, R.P. (1965). On an inequality for modified Bessel functions. *J. Math. and Phys.* **44**, 1–4.

Tsybakov, A. B. (2009) *Introduction to nonparametric estimation.* Springer Science and Business Media.

Valiant, P. and Valiant, G. (2013). Estimating the unseen: iomproved estimators for entropy and other properties. *Adv. Neur. Info. Proc. Sys.* **27**, 2157–2165.

Willenborg, L. and de Waal, T. (2001) *Elements of statistical disclosure control.* Springer, New York.

Wu, Y. and Yang, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inform. Theory* **62**, 3702–3720.

Wu, Y. and Yang, P. (2019). Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *Ann. Statist.*, **47**, 857–883.

Supplementary material for “Optimal disclosure risk assessment”

Federico Camerlenghi, Stefano Favaro, Zacharie Naulet, Francesca Panero

S1 Organization of the document

This document is the companion paper to the article *Optimal Disclosure Risk Assessment*, by the same authors. It complements the result of the main paper in the following way:

- In Section S2, we give the complete proof of the minimax lower bound given in Theorem 3 of the main document, with all details.
- In Section S3, we present an illustration on synthetic data of the estimators introduced in Section 2. We compare our estimator with various estimators from the existing literature.
- In Section S4, we demonstrate that the lower bound $E_L(g_\xi, [\xi^{-1}, 1])$ derived in Lemma 2 of the main document is sharp (up to constants). The proof is constructive and exhibits that Chebychev polynomials achieve the bound.
- Finally, Sections S5 and S6 contain the proofs of the auxiliary results, respectively for the minimax lower bound and the tightness of the lower bound on $E_L(g_\xi, [\xi^{-1}, 1])$.

S2 Complete proof of the minimax lower bound

This section is devoted to the complete proof of the minimax lower bound stated in the main document, that is Theorem 3. Unless specified otherwise, the notations and conventions are the same as in the main document. We recall that the minimax risk is defined as

$$\mathscr{E}(\lambda, n) := \inf_{\hat{\rho}_1} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}[(\hat{\rho}_1(\boldsymbol{X}(N), N) - \tau_1(\boldsymbol{X}, N, M))^2], \tag{S1}$$

where the infinimum is taken over all estimators $\hat{\rho}_1$. To obtain a lower bound on the last display, we adapt the reduction scheme of Wu and Yang (2019, 2016) which is based on the method of the *two fuzzy hypotheses* Tsybakov (2009). More precisely, the proof consists on the following steps.

Step 1 The very first step is to use Lemma 1 in the main document. We recall that Lemma 1 shows that the infinimum in equation (S1) can be restricted over estimators depending only on $(\boldsymbol{X}(N), N)$ through $\boldsymbol{Y}(\boldsymbol{X}, N)$. The details for this step are in the main document and omitted here. We recall the result

$$\mathscr{E}(\lambda, n) = \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}[(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}(\boldsymbol{Y}(\boldsymbol{X}, N)))^2]. \tag{S2}$$

Step 2 The rhs of equation (S2) does not look like a classical minimax bound because $\tau_1(\boldsymbol{X}, N, M)$ is a random variable and not a function of $P \in \mathscr{P}$ (though its distribution is). In order to reduce the problem to a classical minimax problem, we show that τ_1 is sufficiently concentrated around its expectation so that $\tau_1(\boldsymbol{X}, N, M)$ can be traded (asymptotically as $n \to \infty$) for $\bar{\tau}_1(P, n, \lambda) \coloneqq \mathbb{E}_P^{n,\lambda}[\tau_1(\boldsymbol{X}, N, M)]$ under the model P. This is made formal in the next proposition, proved in Section S5.1.

Proposition S1. *Let $\boldsymbol{Y}_N$ denote the random variable $(\boldsymbol{X}, N) \mapsto \boldsymbol{Y}(\boldsymbol{X}, N)$. Then for any $\lambda, n > 0$ the following is true,*

$$\mathscr{E}(\lambda, n) \geq \frac{1}{2} \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}[(\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] - n^{-1}. \tag{S3}$$

Remark that we dropped-out the superscript λ in $\mathbb{E}_P^{n,\lambda}$ in Proposition S1 as the argument in the expectation is independent of M, and thus its distribution depends on λ only through $\bar{\tau}_1(P, n, \lambda)$.

Step 3 The reduction scheme of Wu and Yang (2019, 2016) involves the construction of (fuzzy) hypotheses that are not probability distributions, but only quasi probability distributions. Namely, to use their reduction scheme, we need to show that trading $\mathscr{P}$ for a suitable set of quasi probability distributions $\mathscr{P}'$ in equation (S3) does not affect the bound too much.

For $S \in \mathbb{N}$, $\xi, \delta > 0$ to be chosen accordingly at the end of the day, we define $\mathscr{P}'$ as

$$\mathscr{P}' \coloneqq \left\{ \textstyle\sum_{k=1}^S p_k \delta_k : \; p_k \in [0, \xi S^{-1}], \; |\sum_{k=1}^S p_k - 1| \leq \delta \right\}. \tag{S4}$$

Here and after, under $\mathbb{P}_P^{n,\lambda}$ with $n > 0$ and $P \in \mathscr{P}'$, the random variable $\boldsymbol{Y}_N$ is understood as a vector of independent Poisson random variables with intensities $(np_1, \dots, np_S, 0, \dots)$, with $\sum_{j=1}^S p_j$ not necessarily equal to one, and $(P, n, \lambda) \mapsto \bar{\tau}_1(P, n, \lambda)$ is extended trivially from $\mathscr{P}$ to $\mathscr{P}'$ by letting $\bar{\tau}_1(P, n, \lambda) \coloneqq n \sum_{j=1}^S p_j e^{-n(1+\lambda)p_j}$, $P \in \mathscr{P}'$. Then we have the following proposition, proved in Section S5.2.

Proposition S2. *Let define $n' \coloneqq (1 + \delta)n$ and let S, ξ, δ as defined previously. Then, $\mathscr{E}(\lambda, n)$ is bounded from below by*

$$\frac{1}{4n^2} \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{E}_P^{n',\lambda}[(\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] - \frac{1}{n} - \Big(1 + \frac{n\xi(1+\lambda)}{S(1-\delta)}\Big)^2 \delta^2.$$

This implies that for any $\varepsilon > 0$, $\mathscr{E}(\lambda, n)$ is bounded from below by

$$\frac{\varepsilon^2}{4} \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{P}_P^{n',\lambda}\big(|\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N)| > n\varepsilon\big) - \frac{1}{n} - \Big(1 + \frac{n\xi(1+\lambda)}{S(1-\delta)}\Big)^2 \delta^2.$$

Step 4 The next step involves applying the *method of the two fuzzy hypotheses* Tsybakov (2009) to the result of Proposition S2. The next lemma is an adaptation of (Tsybakov, 2009, Section 2.7.4) to our setting. Its proof is to be found in Section S5.3.

Lemma S1 (Method of the two fuzzy hypotheses)**.** *Let $\mathcal{M}(\mathbb{N})$ denote the space of all measures on $\mathbb{N}$, endowed with canonical σ-algebra. Let $Q_1 = \sum_{j=1}^S q_{1,j}\delta_j$ and $Q_2 = \sum_{j=1}^S q_{2,j}\delta_j$ be independent random variables taking values in $\mathcal{M}(\mathbb{N})$. Also let $\mathscr{P}'$ and ε as defined previously. Assume that for some $0 < \alpha, \beta, \gamma < 1$ with $2\alpha + 2\beta + \gamma \leq 1$ and with n' defined as above the following hold:*

1. *$\mathbb{P}(Q_1 \notin \mathscr{P}') \leq \alpha$ and $\mathbb{P}(Q_2 \notin \mathscr{P}') \leq \alpha$;*
2. *$\mathbb{P}(|\bar{\tau}_1(Q_j, n, \lambda) - \mathbb{E}[\bar{\tau}_1(Q_j, n, \lambda)]| > n\varepsilon/2) \leq \beta$ for $j = 1, 2$;*
3. *$\mathbb{E}[\bar{\tau}_1(Q_1, n, \lambda)] \geq \mathbb{E}[\bar{\tau}_1(Q_2, n, \lambda)] + n\varepsilon$;*
4. *$\mathsf{TV}(\mathbb{E}[\otimes_{j=1}^S \mathrm{Poiss}(n' q_{1,j})], \mathbb{E}[\otimes_{j=1}^S \mathrm{Poiss}(n' q_{2,j})]) \leq \gamma$. Here $\mathsf{TV}(P, Q)$ is used to denote the total-variation distance between probability measures P and Q.*

Then,

$$\inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{P}_P^{n',\lambda}(|\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\mathbf{Y}_N)| > n\varepsilon) \geq \frac{1}{2}\Big(1 - 2\alpha - 2\beta - \gamma\Big).$$

Step 5 The next step consists on constructing the hypotheses that will be used in conjunction with Lemma S1 and Proposition S2 to establish the minimax lower bound. The construction relies on ideas from Wu and Yang (2019, 2016).

For some $L \in \mathbb{N}$ to be determined later, but satisfying $L \leq K_1 \xi$ for some constant $K_1 > 0$, we let U and V be two random variables taking values in $[0, \xi S^{-1}]$ such that $\mathbb{E}[U] = \mathbb{E}[V] = S^{-1}$ and when n is large enough,

$$\mathbb{E}[U^k] = \mathbb{E}[V^k] \quad \forall k \in \{0, \dots, L+1\},$$
$$\mathbb{E}[U e^{-n(1+\lambda)U}] \geq \mathbb{E}[V e^{-n(1+\lambda)V}] + S^{-1}\varepsilon.$$

The existence of such random variables is guaranteed by Lemma S2 below, proven in Section S5.4, for the appropriate choice of S, ξ, L and ε.

Lemma S2. *Let $L \in \mathbb{N}$ and $\xi > 0$ such that $L \leq K_1 \xi$ for some $K_1 > 0$. Let $S = \lceil n(1+\lambda) \rceil$. Then there exists $K_2 > 0$ (depending only on K_1) and two random variables U and V taking values in $[0, \xi S^{-1}]$ such that,*

$$\mathbb{E}[U^k] = \mathbb{E}[V^k] \quad \forall k \in \{0, \dots, L+1\},$$
$$\mathbb{E}[U] = \mathbb{E}[V] = S^{-1}, \quad \mathrm{Var}(U) \leq \xi S^{-2}, \quad \mathrm{Var}(V) \leq \xi S^{-2},$$
$$\mathbb{E}[U e^{-n(1+\lambda)U}] \geq \mathbb{E}[V e^{-n(1+\lambda)V}] + S^{-1} K_2 \min\big\{1, \sqrt{\xi/L^2}\exp(-L^2/\xi)\big\}.$$

Then we let $(U_1, \dots, U_S)$, respectively $(V_1, \dots, V_S)$, be an independent vector of i.i.d. copies of U, respectively V, and we let

$$Q_1 = \sum_{j=1}^S U_j \delta_j, \quad \text{and,} \quad Q_2 = \sum_{j=1}^S V_j \delta_j.$$

The next proposition establishes conditions under which Q_1 and Q_2 as defined above meet the criteria of Lemma S1. The first two items are consequences of Bernstein's and Hoeffding's inequalities (respectively), item 3 is straightforward, and the last item is an immediate corollary of (Wu and Yang, 2019, Lemma 6). The proof is given in Section S5.5.

Proposition S3. *The following items are true.*

1. *Assume that $\mathrm{Var}(U) \leq \xi S^{-2}$, $\mathrm{Var}(V) \leq \xi S^{-2}$, and $S\delta^2 \geq 2\xi(1 + \delta/3)\log(2/\alpha)$. Then $\mathbb{P}(Q_1 \notin \mathscr{P}') \leq \alpha$ and $\mathbb{P}(Q_2 \notin \mathscr{P}') \leq \alpha$.*
2. *Assume that $S\varepsilon^2 \geq 2\xi \log(2/\beta)$. Then $\mathbb{P}(|\bar{\tau}_1(Q_1, n, \lambda) - \mathbb{E}[\bar{\tau}_1(Q_1, n, \lambda)]| > n\varepsilon/2) \leq \beta$. The same is also true for Q_2.*
3. *$\mathbb{E}[\bar{\tau}_1(Q_1, n, \lambda)] \geq \mathbb{E}[\bar{\tau}_1(Q_2, n, \lambda)] + n\varepsilon$.*
4. *Assume that $2\log(2)LS \geq n\xi(1+\delta)$ and $\gamma(2S)^{L+2}(L+2)! \geq 4S(n\xi(1+\delta))^{L+2}$. Then $\mathsf{TV}(\mathbb{E}[\otimes_{j=1}^S \mathrm{Poiss}(n' U_j)], \mathbb{E}[\otimes_{j=1}^S \mathrm{Poiss}(n' V_j)]) \leq \gamma$.*

Step 6 The proof of Theorem 3 follows from combining Propositions S2, S3, and Lemma S1, by choosing the constants α, β, γ and variables $\varepsilon, S, \xi, \delta, L$ accordingly. We now make explicit the choice for these constants and variables.

In the following for any $x > 0$ the notations $\lceil x \rceil$ stands for the smallest integer greater or equal than x. Then, for constants $c_0, c_1 > 0$ to be determined we choose

$$S = \lceil n(1+\lambda) \rceil, \tag{S5}$$

$$\delta = c_0 \varepsilon / \xi, \tag{S6}$$

$$\xi = (2c_1/e) \min\{(1+\lambda)\log n,\, \log^2 n\}. \tag{S7}$$

For another constant $c_2 > 0$ to be determined, we further define $A(\lambda, n) > 0$ to be the solution to

$$A(\lambda, n) \log A(\lambda, n) = c_1^{-1} + c_1^{-1} \frac{\log(1+\lambda) - (1/2)\log\log(n) + \log(c_2)}{\log(n)}.$$

Then we pick (remark that this ensures that $L \le K_1 \xi$ for some $K_1 > 0$, as requested previously),

$$L = \begin{cases} \lceil 2c_1 \log(n) \rceil & \text{if } 1+\lambda > \log(n), \\ \lceil c_1 A(\lambda, n) \log(n) \rceil & \text{if } 1+\lambda \le \log(n), \end{cases} \tag{S8}$$

and for $c_3 > 0$ to be determined,

$$\varepsilon = c_3 \cdot \begin{cases} 1 & \text{if } 1+\lambda > \log(n), \\ \frac{1}{\sqrt{\log(n)}} \cdot \sqrt{\frac{2(1+\lambda)}{ec_1 A(\lambda,n)^2}} \cdot n^{-\frac{ec_1 A(\lambda,n)^2}{2(1+\lambda)}} & \text{if } 1+\lambda \le \log(n). \end{cases} \tag{S9}$$

With this choice, we obtain the next proposition, proved in Section S5.6.

Proposition S4. *Let $\alpha = \beta = \gamma = 1/10$, and let $S, \xi, \delta, L, \varepsilon$ as in Equations S5, S6, S7, S8 and S9. Then,*

1. $(1 + \frac{n\xi(1+\lambda)}{S(1-\delta)})^2 \delta^2 \le c_0^2 \varepsilon^2 (1 + o(1))$ *as $n \to \infty$;*

2. *If* $\liminf_n \left\{ \frac{1+\lambda}{ec_1 A(\lambda,n)^2} \right\} > 1$ *then there exists $n_0 > 0$ such that for all $n \ge n_0$ it holds $S\delta^2 \ge 2\xi(1 + \delta/3)\log(2/\alpha)$;*

3. *If* $\liminf_n \left\{ \frac{1+\lambda}{ec_1 A(\lambda,n)^2} \right\} > 1$ *then there exists $n_0 > 0$ such that for all $n \ge n_0$ it holds $S\varepsilon^2 \ge 2\log(2/\beta)$;*

4. *For any $K_2 > 0$ the constant $c_3 > 0$ can be chosen such that $\varepsilon \le K_2 \min\{1,\, \sqrt{\xi/L^2}\exp(-L^2/\xi)\}$; In conjunction with Lemma S2 this guarantees the existence of U and V used in Step 5.*

5. *If $c_2 > 0$ is large enough, then there exists $n_0 > 0$ such that for all $n \ge n_0$ we have $2\log(2)LS \ge n\xi(1+\delta)$ and $\gamma(2S)^{L+2}(L+2)! \ge 4S(n\xi(1+\delta))^{L+2}$.*

Therefore, as a consequence of Propositions S2, S3 and Lemma S1, when c_0, c_1, c_2, c_3 are appropriately chosen, if $1 + \lambda > ec_1 A(\lambda, n)^2$, and if n gets large enough,

$$\mathscr{E}(\lambda, n) \ge \Big(\frac{1}{16} - c_0^2 + o(1)\Big)\varepsilon^2.$$

Step 7 In view of Equation S9, the choice of c_1 shall be made cautiously. Indeed, the next proposition shows that $c_1 = 1/e$ is the optimal choice. The result of the next proposition also allows to get the final expression for the lower bound in $\mathscr{E}(\lambda, n)$, thus finishing the proof of Theorem 3. The proof of Proposition S5 is to be found in Section S5.7.

Proposition S5. *Let $c_1 = 1/e$. Then whenever $1 + \lambda \leq \log(n)$ we have $A(\lambda, n) = e + o(1)$ as $n \to \infty$. Furthermore when $1 + \lambda \leq \log(n)$, as $n \to \infty$,*

$$c_1 A(\lambda, n)^2 \log(n) \leq e \log(n) + e \log \frac{c_2(1+\lambda)}{\sqrt{\log(n)}} + o(1).$$

S3 Numerical illustrations

We present an illustration on synthetic data of the estimators introduced in Section 2. We also consider other estimators of τ_1 that have been proposed in the literature of disclosure risk assessment: i) two parametric empirical Bayes estimators of τ_1 proposed by Bethlehem et al. (1990) and Skinner et al. (1994); ii) a naive nonparametric estimator of τ_1; iii) a Bayesian nonparametric estimator of τ_1 proposed by Samuels (1998). A common feature of these estimators, as well as our class of nonparametric estimators, is that they rely on the Poisson abundance model for modeling the random partition induced by the cross-classified sample records. More recent approaches, not considered here, focus on modeling associations among identifying variables by log-linear models, local smoothing polynomials and hierarchical latent models. E.g., Manrique-Vallier and Reiter (2012), Manrique-Vallier and Reiter (2014), Carota et al. (2015) and Carota et al. (2018). In particular, the Bayesian hierarchical semiparametric models of Carota et al. (2015) and Carota et al. (2018) show a remarkable better performance than models for random partitions, at the cost of an increasing computational effort for the need of Markov chain Monte Carlo methods for posterior approximation.

The approach of Bethlehem et al. (1990) is a parametric empirical Bayes approach in the sense of Efron and Morris (1973). It relies on the following modeling assumption for the cells' frequencies of the population: $Y_j(\boldsymbol{X}, \bar{n}) \sim \text{Poiss}(\bar{n} p_j)$, where $\bar{n}$ is the size of the entire population. Bethlehem et al. (1990) also assumed a Gamma prior distribution over the probabilities associated to each cell, namely $p_j \sim \text{Gam}(\alpha, \beta)$. One should specify the p_j's under the condition $\sum_{j=1}^{K_{\bar{n}}} p_j = 1$, however, for the sake of simplicity, Bethlehem et al. (1990) assumed that $\sum_{j=1}^{K_{\bar{n}}} \mathbb{E}[p_j] = 1$, which is tantamount to saying that $\alpha = 1/(K_{\bar{n}}\beta)$. Under these modeling assumptions, Bethlehem et al. (1990) proposed an estimator of the expected value of total number $T_1(\boldsymbol{X}, \bar{n})$ of population uniques, i.e.,

$$T_1(\boldsymbol{X}, \bar{n}) := \sum_{j=1}^{K_{\bar{n}}} \mathbb{1}_{\{Y_j(\boldsymbol{X}, \bar{n})=1\}}. \tag{S10}$$

Under the above Poisson-Gamma model, $\mathbb{E}[T_1(\boldsymbol{X}, \bar{n})] = \bar{n}(1+\bar{n}\beta)^{-(1+\alpha)}$, which depends on the parameters α and β, with the condition $\alpha = 1/(K\beta)$. Parameters can be easily estimated via maximum likelihood, as we have done in the subsequent numerical experiments. If $K_{\bar{n}}$ is not available, Bethlehem et al. (1990) suggested to estimate $K_{\bar{n}}$ assuming a uniform distribution over the cells, hence

$$\hat{K}_{\bar{n}} = \frac{\bar{n} K_n}{\sum_{j=1}^{K_n} \mathbb{1}_{\{Y_j(\boldsymbol{X}, n)=1\}}},$$

where n is the size of the observed sample and K_n stands for the number of distinct cells dictated by the sample of size n. If $\hat{\alpha}$ and $\hat{\beta}$ denote the maximum likelihood estimators of α and

β, respectively, then an estimator of $T_1(\boldsymbol{X}, \bar{n})$ is $\hat{T}_1 = \bar{n}(1+\bar{n}\hat{\beta})^{-(1+\hat{\alpha})}$. Bethlehem et al. (1990) then suggested a corresponding estimator of τ_1 as the sample portion of $\hat{T}_1$. More precisely, they proposed

$$\hat{\tau}_1^B = \frac{n}{\bar{n}}\hat{T}_1 = n(1+\bar{n}\hat{\beta})^{-(1+\hat{\alpha})}. \tag{S11}$$

as an estimator of τ_1. Skinner et al. (1994) improved the estimator (S11). In particular, still under the Poisson-Gamma model, they considered directly the problem of estimating τ_1. In particular, they proposed the following estimator

$$\hat{\tau}_1^S := K_n \left(\frac{1+\bar{n}\hat{\beta}}{1+n\hat{\beta}}\right)^{-(1+\hat{\alpha})}, \tag{S12}$$

where the prior parameters α and β can be estimated via maximum likelihood. The estimators proposed in Section 2, due to their nonparametric empirical Bayes interpretation in the sense of Robbins (1956), may be considered as the natural nonparametric counterparts of the empirical Bayes estimator (S12).

Besides parametric estimators of τ_1, we also consider two nonparametric estimators. A naive nonparametric estimator of τ_1 relies on the intuition that a natural estimator of τ_1 is the sampling fraction, with respect to the population, of the number of sample uniques. This estimator was first discussed in Bethlehem et al. (1990) and Skinner et al. (1994), and it is defined as follows

$$\hat{\tau}_1^{\mathscr{N}} := Z_1(\boldsymbol{X}, n)\frac{n}{\bar{n}}. \tag{S13}$$

Samuels (1998) exploits Bayesian nonparametric ideas, and in particular a Dirichlet process prior (Ferguson (1973)) on the p_j's to derive a smoothed version of the naive estimator (S13). In particular, Samuels (1998) suggested the following estimator

$$\hat{\tau}_1^{\mathscr{D}} := Z_1(\boldsymbol{X}, n)\frac{n+\vartheta-1}{\bar{n}+\vartheta-1}, \tag{S14}$$

where ϑ is the concentration parameter of the Dirichlet process prior. It is well-known (see, e.g. Ferguson (1973)) that the maximum likelihood estimator of ϑ can be obtained by solving, with respect to ϑ, the equation $K_n = \sum_{1\leq j\leq n-1}\vartheta/(\vartheta+j)$.

We study the behavior of the Normalized Mean Squared Error (NMSE), with respect to the sampling fraction $(1+\lambda)^{-1}$, for the collection of estimators of τ_1 introduced before. In order to do that, we generate a collection of synthetic tables with C cells, where $C = 3\cdot 10^6$ in all our experiments. The population size is fixed to $\bar{n} = 10^6$, and we evaluate the NMSE for different values of the sample size $n = \bar{n}(\lambda+1)^{-1}$. The true probabilities $(p_j)_{j\geq 1}$ of cells are generated according to different types of distributions: the Zipf distribution, i.e., $p_j \propto j^{-s}$ for some $s > 0$, the uniform distribution over the total number of cells and the uniform Dirichlet distribution. Each Figure corresponds to a different choice of the distribution over the cells' probabilities: the Zipf distribution with respective parameter $s = 0.6, 0.8, 1$ (Figures S1–S3), the uniform distribution (Figure S4), the uniform Dirichlet distribution with respective parameter $\beta = 0.5, 1$ (Figures S5–S6). Each figure shows how the NMSE varies as a function of the sampling fraction $(1+\lambda)^{-1}$ for different estimators: i) the nonparametric estimator with Binomial smoothing $\hat{\tau}_1^{L_b}$, see Proposition 2; ii) the nonparametric estimator with Poisson smoothing $\hat{\tau}_1^{L_p}$, see Proposition 1; iii) the naive nonparametric estimator $\hat{\tau}_1^{\mathscr{N}}$; iv) the Bayesian nonparametric estimator $\hat{\tau}_1^{\mathscr{D}}$; v) the parametric empirical Bayes estimator $\hat{\tau}_1^B$; vi) the parametric empirical Bayes estimator $\hat{\tau}_1^S$. All experiments are averaged over 100 iterations and the empirical bands represent one standard deviation from the mean of the corresponding estimates.

The sampling fractions considered in our simulation study are above the limiting threshold $(\log n)^{-1}$. Within this range of sampling fractions, we do not observe a clear behavior for the performance of the estimators. It is apparent that in most of the simulated scenarios our estimator outperforms as the sampling fraction $(1+\lambda)^{-1}$ increases from the limiting threshold $(\log n)^{-1}$. From Figure S5, the Bayesian nonparametric estimator $\hat{\tau}_1^{\mathscr{D}}$ provides the smallest NMSE; this behaviour is not surprising since data are drawn from a Dirichlet distribution. In Figures S1–S3, better performances are achieved by the estimators $\hat{\tau}_1^{L_b}$ and $\hat{\tau}_1^{L_p}$. We further observe that the choice of the smoothing distribution L for $\hat{\tau}_1^{L}$, i.e. the Binomial smoothing or the Poisson smoothing, is crucial with respect to the performance of the corresponding estimators. In all the simulated scenarios the Binomial smoothing displays a better performance than the Poisson smoothing. Finally in Table S1, we report the estimates of τ_1 (with empirical confidence bands) when $(\lambda+1)^{-1}=1/5$ for all the choices of the cells' probabilities, from the left to right: the Zipf distribution with parameter $s=0.6, 0.8, 1$, the uniform distribution, the uniform Dirichlet distribution with parameter $\beta=0.5, 1$. All experiments are averaged over 100 iterations and the empirical intervals represent one standard deviation from the mean of the corresponding estimates. From Table S1, we can deduce similar considerations as before.

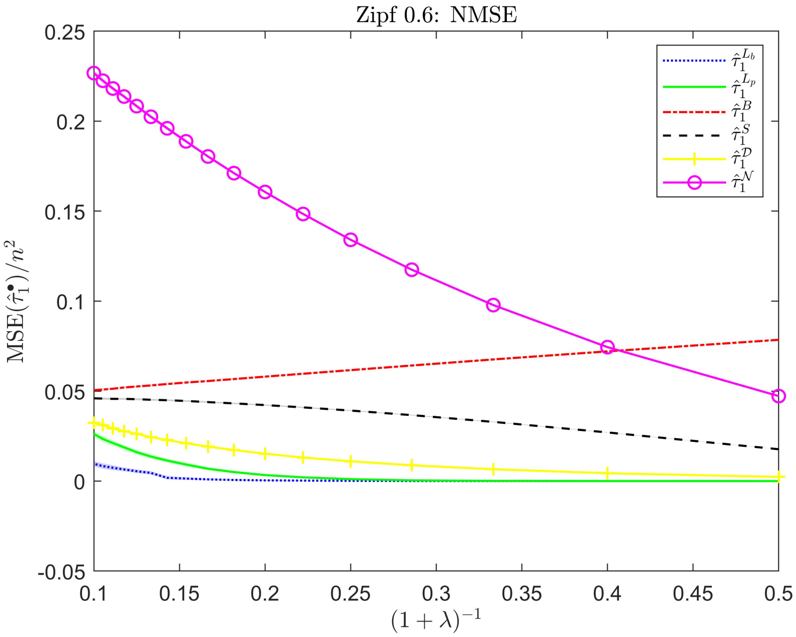

Figure S1: The normalized mean squared error as a function of the sampling fraction $(1+\lambda)^{-1}$ when the distribution of the cell's probabilities is a Zipf with parameter $s=0.6$. Each curve corresponds to a different estimator of τ_1: i) the nonparametric estimator with Binomial smoothing $\hat{\tau}_1^{L_b}$; ii) the nonparametric estimator with Poisson smoothing $\hat{\tau}_1^{L_p}$; iii) the naive nonparametric estimator $\hat{\tau}_1^{\mathscr{N}}$; iv) the Bayesian nonparametric estimator $\hat{\tau}_1^{\mathscr{D}}$; v) the parametric empirical Bayes estimator $\hat{\tau}_1^{B}$; vi) the parametric empirical Bayes estimator $\hat{\tau}_1^{S}$. The shaded bands corresponds to one standard deviation.

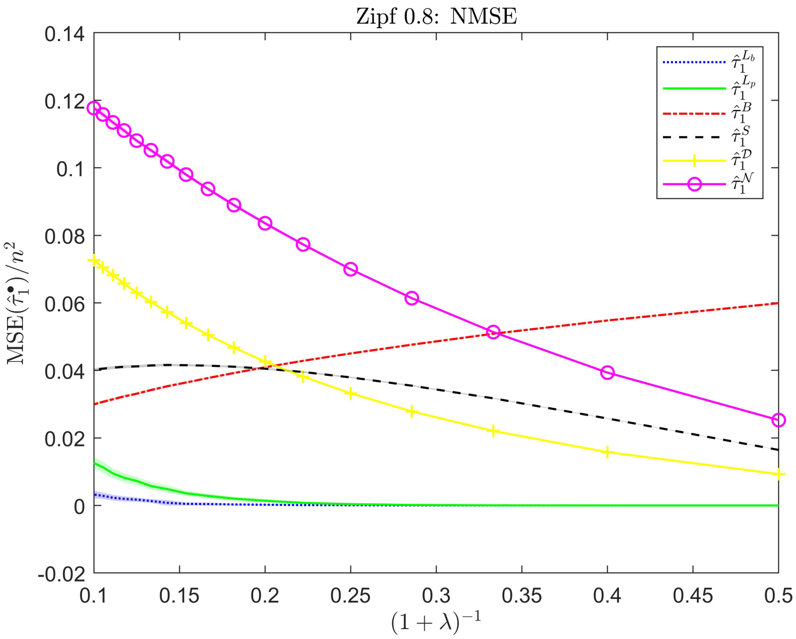

Figure S2: The normalized mean squared error as a function of the sampling fraction $(1+\lambda)^{-1}$ when the distribution of the cell's probabilities is a Zipf with parameter $s = 0.8$. Each curve corresponds to a different estimator of τ_1: i) the nonparametric estimator with Binomial smoothing $\hat{\tau}_1^{L_b}$; ii) the nonparametric estimator with Poisson smoothing $\hat{\tau}_1^{L_p}$; iii) the naive nonparametric estimator $\hat{\tau}_1^{\mathscr{N}}$; iv) the Bayesian nonparametric estimator $\hat{\tau}_1^{\mathscr{D}}$; v) the parametric empirical Bayes estimator $\hat{\tau}_1^{B}$; vi) the parametric empirical Bayes estimator $\hat{\tau}_1^{S}$. The shaded bands corresponds to one standard deviation.

S4 Tightness of the approximation lower bound

We show that a suitable Chebychev polynomial approximation of the exponential function achieves (up to a multiplicative constant) the lower bound of Lemma 2 in the main document.

In view of Section B.2 in the main document, letting $\gamma_C : [-1,1] \to \mathbb{R}$ such that $\gamma_C(x) \coloneqq e^{-C(x+1)}$, it is enough to find a sequence of polynomial $(q_L)_{L\geq 1}$ such that q_L has degree at most L and for a constant $K > 0$,

$$L \leq \sqrt{C} \Longrightarrow \sup_{x\in[-1,1]} |\gamma_C(x) - q_L(x)| \leq K, \tag{S15}$$

and,

$$\sqrt{C} \leq L \leq \zeta C \Longrightarrow \sup_{x\in[-1,1]} |\gamma_C(x) - q_L(x)| \leq K\frac{L}{\sqrt{C}}e^{-C\varphi(L/C)}, \tag{S16}$$

at least when C is large enough, and with φ defined in Equation (18) in the main document. If $L \leq \sqrt{C}$, then we pick $q_L(x) = 0$ identically, so that the equation (S15) is trivially satisfied with any $K \geq 1$, because $|\gamma_C(x)| \leq 1$. Thus it suffices to establish (S16). For any $k \geq 0$, we let $T_k : [-1,1] \to \mathbb{R}$ the k-th order Chebychev polynomial, defined uniquely through the equality

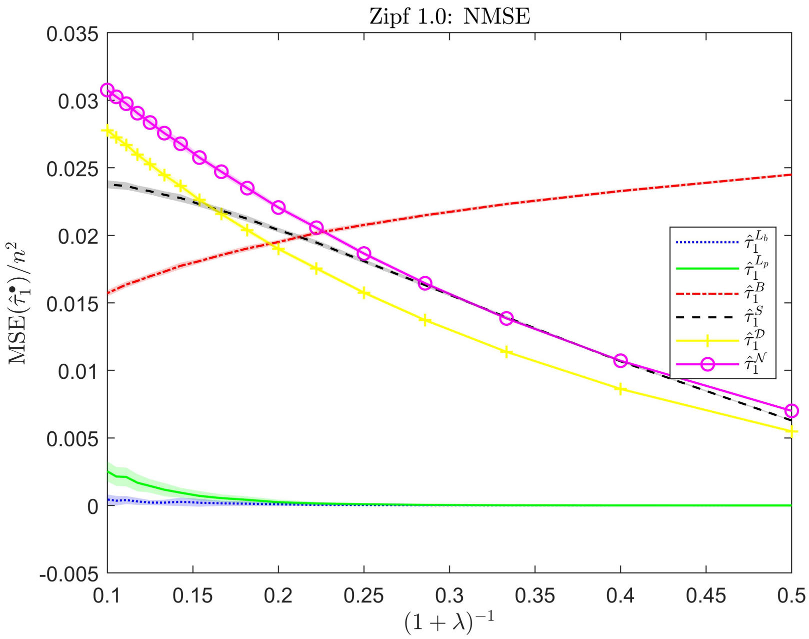

Figure S3: The normalized mean squared error as a function of the sampling fraction $(1+\lambda)^{-1}$ when the distribution of the cell's probabilities is a Zipf with parameter $s = 1.0$. Each curve corresponds to a different estimator of τ_1: i) the nonparametric estimator with Binomial smoothing $\hat{\tau}_1^{L_b}$; ii) the nonparametric estimator with Poisson smoothing $\hat{\tau}_1^{L_p}$; iii) the naive nonparametric estimator $\hat{\tau}_1^{\mathscr{N}}$; iv) the Bayesian nonparametric estimator $\hat{\tau}_1^{\mathscr{D}}$; v) the parametric empirical Bayes estimator $\hat{\tau}_1^{B}$; vi) the parametric empirical Bayes estimator $\hat{\tau}_1^{S}$. The shaded bands corresponds to one standard deviation.

$T_k(\cos(\theta)) = \cos(k\theta)$, for all $\theta \in [-\pi, \pi]$. Then, we choose,

$$q_L(x) := \sum_{k=0}^{L} a_k(C) \cdot T_k(x), \quad a_k(C) := \int_{-1}^{1} \frac{e^{-C(x+1)} T_k(x)}{\sqrt{1-x^2}} \mathrm{d}x. \tag{S17}$$

We collect in the next Lemma several facts about the polynomial q_L and its coefficients $a_k(C)$ which will be used to derive the rate of approximation of q_L to γ_C, in the uniform norm.

Lemma S3. *The following items are true.*

1. *$a_k(C) = \pi(-1)^k e^{-C} I_k(C)$ for all $k \geq 0$, where I_k is the modified Bessel function of the first kind (see (Olver et al., 2010, pg. 248)).*

2. *The series $q_\infty := \sum_{k=0}^{\infty} a_k(C) T_k$ converges uniformly in $[-1, 1]$, and $q_\infty(x) = \gamma_C(x)$ for all $x \in [-1, 1]$.*

3. *For all $D > 0$ there exists $B_0 > 0$ such that for all $B \geq B_0$ and for all $k \geq \max\{BC, 2\}$, we have the bound $|a_k(C)| \leq e^{-Dk}$.*

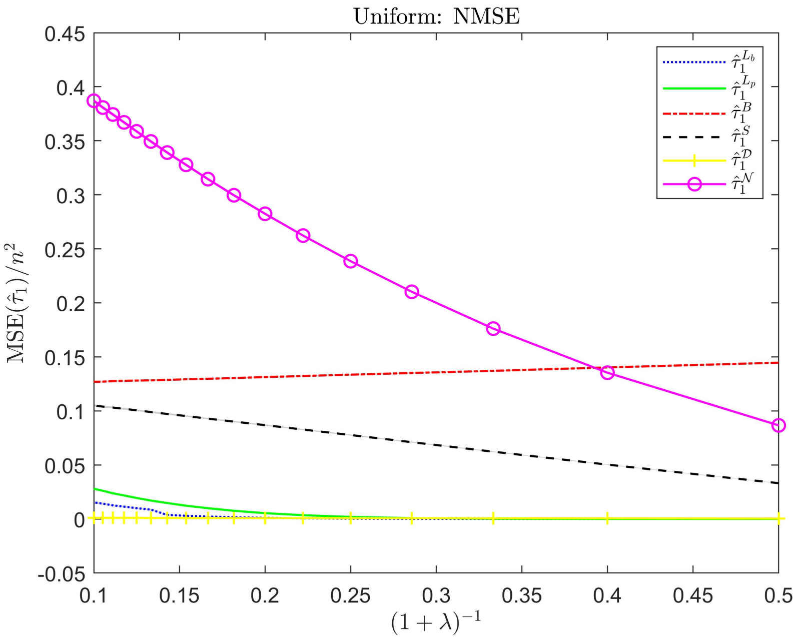

Figure S4: The normalized mean squared error as a function of the sampling fraction $(1+\lambda)^{-1}$ when the cell's probabilities are uniform distributed. Each curve corresponds to a different estimator of τ_1: i) the nonparametric estimator with Binomial smoothing $\hat{\tau}_1^{L_b}$; ii) the nonparametric estimator with Poisson smoothing $\hat{\tau}_1^{L_p}$; iii) the naive nonparametric estimator $\hat{\tau}_1^{\mathscr{N}}$; iv) the Bayesian nonparametric estimator $\hat{\tau}_1^{\mathscr{D}}$; v) the parametric empirical Bayes estimator $\hat{\tau}_1^{B}$; vi) the parametric empirical Bayes estimator $\hat{\tau}_1^{S}$. The shaded bands corresponds to one standard deviation.

4. *For all $B > 0$ there exists $C_0 > 0$ such that for all $C > C_0$, for all $\sqrt{C} \leq L < k \leq BC$, we have the bound*

$$|a_k(C)| \leq \sqrt{2\pi} \cdot \frac{\exp\{-C\varphi(k/C)\}}{\sqrt{C}}.$$

Using the results of the previous lemma, we obtain the following corollary on the error of the best uniform polynomial approximation to γ_C on $[-1,1]$, written $E_L(\gamma_C, [-1,1])$.

Corollary S1. *For all $\zeta > 0$ there exists $C_0 > 0$ such that for all $C > C_0$ and for all $\sqrt{C} \leq L \leq \zeta C$*

$$E_L(\gamma_C, [-1,1]) \leq \sqrt{4\pi(1+\zeta^2)} \cdot \frac{\sqrt{C}}{L} e^{-C\varphi(L/C)}.$$

Furthermore, the polynomial q_L defined in (S17) *achieves the previous upper bound; and in view of [...] in the main document, this bound is the best possible, up to a multiplicative constant.*

S5 Remaining proofs for the minimax lower bound

This section gather all the proofs of the propositions and lemma stated in Section S2.

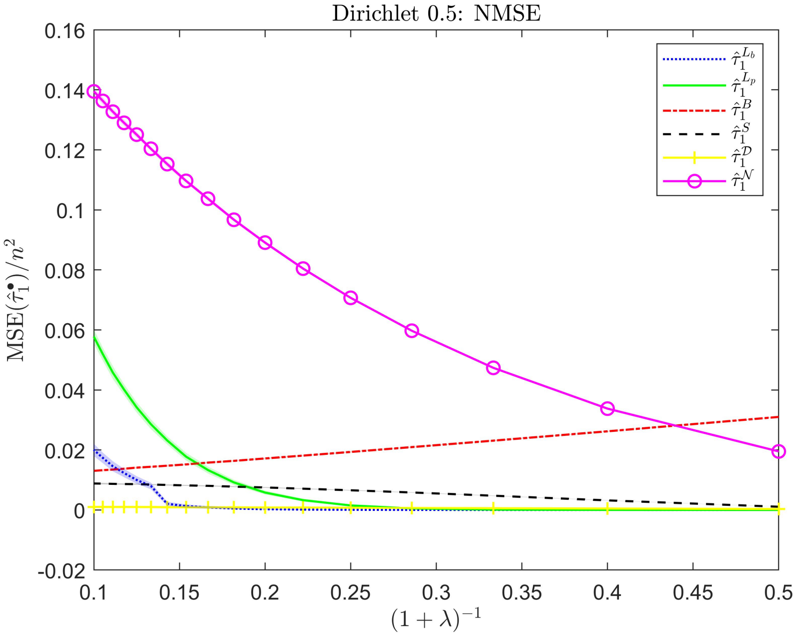

Figure S5: The normalized mean squared error as a function of the sampling fraction $(1+\lambda)^{-1}$ when the distribution of the cell's probabilities is a uniform Dirichlet distribution with respective parameter $\beta = 0.5$. Each curve corresponds to a different estimator of τ_1: i) the nonparametric estimator with Binomial smoothing $\hat\tau_1^{L_b}$; ii) the nonparametric estimator with Poisson smoothing $\hat\tau_1^{L_p}$; iii) the naive nonparametric estimator $\hat\tau_1^{\mathscr{N}}$; iv) the Bayesian nonparametric estimator $\hat\tau_1^{\mathscr{D}}$; v) the parametric empirical Bayes estimator $\hat\tau_1^{B}$; vi) the parametric empirical Bayes estimator $\hat\tau_1^{S}$. The shaded bands corresponds to one standard deviation.

S5.1 Proof of Proposition S1

Using Jensen's inequality we deduce that

$$\begin{aligned}\mathscr{E}(\lambda,n) &= \inf_{\hat\rho}\sup_{P\in\mathscr{P}} n^{-2}\mathbb{E}_P^{n,\lambda}[\mathbb{E}_P^{n,\lambda}[(\tau_1(\boldsymbol{X},N,M)-\hat\rho(\boldsymbol{Y}(\boldsymbol{X},N)))^2 \mid \boldsymbol{Y}(\boldsymbol{X},N)]]\\ &\geq \inf_{\hat\rho}\sup_{P\in\mathscr{P}} n^{-2}\mathbb{E}_P^{n,\lambda}[(\mathbb{E}_P^{n,\lambda}[\tau_1(\boldsymbol{X},N,M)\mid \boldsymbol{Y}(\boldsymbol{X},N)]-\hat\rho(\boldsymbol{Y}(\boldsymbol{X},N)))^2].\end{aligned}$$

Note that there is no explicit dependency on $\boldsymbol{X}$ and M anymore in the last display, but only on the random variable $(\boldsymbol{X},N)\mapsto \boldsymbol{Y}(\boldsymbol{X},N)$ which, under P, is distributed as an infinite vector of independent Poisson random variables with parameters $(np_1,np_2,\dots)$. Besides observe also that $N=\sum_{j\geq1}Y_j(\boldsymbol{X},N)$. Let define

$$\begin{aligned}\tilde\tau_1(\boldsymbol{Y}_N,P,n,\lambda) &:= \mathbb{E}_P^{n,\lambda}[\tau_1(\boldsymbol{X},N,M)\mid \boldsymbol{Y}(\boldsymbol{X},N)]\\ &= \sum_{j\geq1}\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{E}_P^{n,\lambda}[\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)-Y_j(\boldsymbol{X},N)=0\}}\mid \boldsymbol{Y}(\boldsymbol{X},N)].\end{aligned}$$

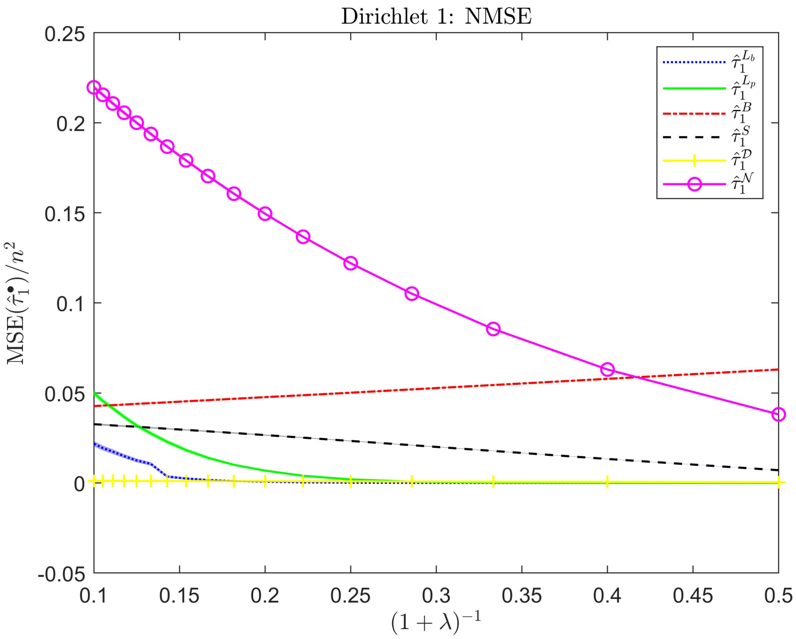

Figure S6: The normalized mean squared error as a function of the sampling fraction $(1+\lambda)^{-1}$ when the distribution of the cell's probabilities is a uniform Dirichlet distribution with respective parameter $\beta = 1.0$. Each curve corresponds to a different estimator of τ_1: i) the nonparametric estimator with Binomial smoothing $\hat\tau_1^{L_b}$; ii) the nonparametric estimator with Poisson smoothing $\hat\tau_1^{L_p}$; iii) the naive nonparametric estimator $\hat\tau_1^{\mathscr{N}}$; iv) the Bayesian nonparametric estimator $\hat\tau_1^{\mathscr{D}}$; v) the parametric empirical Bayes estimator $\hat\tau_1^{B}$; vi) the parametric empirical Bayes estimator $\hat\tau_1^{S}$. The shaded bands corresponds to one standard deviation.

Remark that $(Y_j(\boldsymbol{X}, N+M) - Y_j(\boldsymbol{X}, N) : j \in \mathbb{N})$ is independent of $\boldsymbol{Y}(\boldsymbol{X}, N)$ and is a collection of independent Poisson random variables with intensities $(\lambda n p_j : j \in \mathbb{N})$. Henceforth, we get

$$\tilde\tau_1(\boldsymbol{Y}_N, P, n, \lambda) = \sum_{j\geq 1} e^{-\lambda n p_j} \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}, \tag{S18}$$

and besides, since we abusively let $\boldsymbol{Y}_N$ denote the random variable $(\boldsymbol{X}, N) \mapsto \boldsymbol{Y}(\boldsymbol{X}, N)$,

$$\mathscr{E}(\lambda, n) \geq \inf_{\hat\rho} \sup_{P\in\mathscr{P}} n^{-2}\mathbb{E}_P^{n,\lambda}[(\tilde\tau_1(\boldsymbol{Y}_N, P, n, \lambda) - \hat\rho(\boldsymbol{Y}_N))^2]. \tag{S19}$$

We now trade $\tilde\tau_1(\boldsymbol{Y}_N, P, n, \lambda)$ for its expectation whowh we define as $\bar\tau_1(P, n, \lambda) \coloneqq \mathbb{E}_P^{n,\lambda}[\tau_1(\boldsymbol{X}, N, M)]$. Recall that under P the vector $\boldsymbol{Y}_N$ is distributed as independent Poisson with parameters $(np_1, np_2, \dots)$. Hence,

$$\bar\tau_1(P, n, \lambda) = \sum_{j\geq 1} e^{-\lambda n p_j} \mathbb{E}_P^{n,\lambda}[\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}] = n \sum_{j\geq 1} p_j e^{-(1+\lambda) n p_j}.$$

Similarly, for any $P \in \mathscr{P}$,

	Zipf 0.6	Zipf 0.8	Zipf 1
True τ_1	112780	82254	42397
$\hat{\tau}_1^{L_b}$	$116533 \in (115361, 117704)$	$84478 \in (83041, 85916)$	$43370 \in (41980, 44760)$
$\hat{\tau}_1^{L_p}$	$124242 \in (123380, 125104)$	$89443 \in (88195, 90690)$	$45307 \in (44188, 46427)$
$\hat{\tau}_1^{\mathscr{N}}$	$32623 \in (32580, 32666)$	$24436 \in (24386, 24485)$	$12593 \in (12555, 12630)$
$\hat{\tau}_1^{\mathscr{D}}$	$88030 \in (87699, 88362)$	$40983 \in (40833, 41133)$	$14740 \in (14688, 14792)$
$\hat{\tau}_1^{B}$	$64587 \in (64525, 64650)$	$41815 \in (41714, 41915)$	$14362 \in (14312, 14412)$
$\hat{\tau}_1^{S}$	$71651 \in (71543, 71759)$	$42022 \in (41900, 42145)$	$13738 \in (13690, 13787)$

	Uniform	Dirichlet 0.5	Dirichlet 1
True τ_1	143375	92849	112468
$\hat{\tau}_1^{L_b}$	$149823 \in (149127, 150520)$	$95806 \in (94658, 96955)$	$117449 \in (116465, 118433)$
$\hat{\tau}_1^{L_p}$	$157967 \in (157408, 158526)$	$108040 \in (107174, 108907)$	$128879 \in (128150, 129607)$
$\hat{\tau}_1^{\mathscr{N}}$	$37424 \in (37392, 37457)$	$33133 \in (33086, 33181)$	$35147 \in (35110, 35184)$
$\hat{\tau}_1^{\mathscr{D}}$	$149121 \in (148619, 149623)$	$98586 \in (98178, 98993)$	$118696 \in (118285, 119106)$
$\hat{\tau}_1^{B}$	$71141 \in (71110, 71172)$	$66620 \in (66568, 66672)$	$68820 \in (68782, 68858)$
$\hat{\tau}_1^{S}$	$84631 \in (84565, 84697)$	$75504 \in (75404, 75604)$	$79853 \in (79776, 79930)$

Table S1: Estimation of τ_1 for several simulated scenarios, when the size of the population is $\bar{n} = 10^6$ and $(\lambda+1)^{-1} = 1/5$. Each column corresponds to a different choice of the distribution over the cells' probabilities. The first line displays the true value of τ_1, while the other rows contain the estimates and the empirical bands based on one standard deviation. All the experiments are averaged over 100 iterations.

$$\mathbb{E}_P^{n,\lambda}[(\tilde{\tau}_1(\boldsymbol{Y}_N, P, n, \lambda) - \bar{\tau}_1(P, n, \lambda))^2] \\ = \sum_{j\geq 1} np_j e^{-(1+2\lambda)np_j}\left\{1 - np_j e^{-np_j}\right\} \leq n. \quad \text{(S20)}$$

Thus from (S19) and Young's inequality, we find that

$$\mathscr{E}(\lambda, n) \geq \frac{1}{2n^2} \inf_{\hat{\rho}} \sup_{P\in\mathscr{P}} \mathbb{E}_P^{n,\lambda}[(\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] \\ - \frac{1}{n^2}\mathbb{E}_P^{n,\lambda}[(\tilde{\tau}_1(\boldsymbol{Y}_N, P, n, \lambda) - \bar{\tau}_1(P, n, \lambda))^2].$$

That is using (S20),

$$\mathscr{E}(\lambda, n) \geq \frac{1}{2} \inf_{\hat{\rho}} \sup_{P\in\mathscr{P}} n^{-2}\mathbb{E}_P^{n,\lambda}[(\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] - n^{-1}.$$

S5.2 Proof of Proposition S2

For any $P \in \mathscr{P}'$ we let $\tilde{P}(\cdot) := P(\cdot)/P(\mathbb{N})$, so that $\tilde{P} \in \mathscr{P}$ is a probability measure. We write $\tilde{p}_j := p_j/P(\mathbb{N})$, $j \in \{1, \ldots, S\}$. Furthermore we let $m(P) := n\sum_{j=1}^{S} p_j$. Then since $\boldsymbol{Y}_N$ is a vector of independent Poisson random variables, is clear that for any $P \in \mathscr{P}'$

$$\mathbb{E}_{\tilde{P}}^{n,\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] = \mathbb{E}_{P}^{m(P),\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2]. \tag{S21}$$

We now choose $\hat{\tau}$ to be an estimator satisfying for some $\zeta > 0$

$$\begin{aligned}\sup_{P\in\mathscr{P}'} \mathbb{E}_{P}^{m(P),\lambda}&[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\tau}(\boldsymbol{Y}_N))^2] \\ &\leq \inf_{\hat{\rho}} \sup_{P\in\mathscr{P}'} \mathbb{E}_{P}^{m(P),\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] + \zeta.\end{aligned}$$

This is always possible for any $\zeta > 0$. Furthermore remark that $m(P) \leq (1+\delta)n = n'$, so that $m(P)/n' \leq 1$ always when $P \in \mathscr{P}'$. Let $P \in \mathscr{P}'$ be fixed, and let $\boldsymbol{W} = (W_1, W_2, \dots)$ such that conditional on $\boldsymbol{Y}_N$, the random variables W_j are independent binomial random variables with parameters $(Y_j, m(P)/n')$. Then define $\tilde{\tau}(\boldsymbol{Y}_N) \coloneqq \mathbb{E}[\hat{\tau}(\boldsymbol{W})] \mid \boldsymbol{Y}_N]$. By Jensen's inequality,

$$\begin{aligned}\mathbb{E}_{P}^{n',\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \tilde{\tau}(\boldsymbol{Y}_N))^2] &= \mathbb{E}_{P}^{n',\lambda}[(\mathbb{E}[\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\tau}(\boldsymbol{W}) \mid \boldsymbol{Y}_N])^2] \\ &\leq \mathbb{E}_{P}^{n',\lambda}[\mathbb{E}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\tau}(\boldsymbol{W}))^2 \mid \boldsymbol{Y}_N]] \\ &= \mathbb{E}_{P}^{m(P),\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\tau}(\boldsymbol{Y}_N))^2] \\ &\leq \inf_{\hat{\rho}} \sup_{P\in\mathscr{P}'} \mathbb{E}_{P}^{m(P),\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] + \zeta.\end{aligned}$$

Taking the supremum over $P \in \mathscr{P}'$ on the lhs of the last display, and using that the infinimum over $\hat{\rho}$ will be always smaller than the value at $\tilde{\tau}$, we find using (S21) that

$$\begin{aligned}\inf_{\hat{\rho}} \sup_{P\in\mathscr{P}} &\mathbb{E}_{P}^{n,\lambda}[(\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] \\ &= \inf_{\hat{\rho}} \sup_{P\in\mathscr{P}'} \mathbb{E}_{\tilde{P}}^{n,\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] \\ &= \inf_{\hat{\rho}} \sup_{P\in\mathscr{P}'} \mathbb{E}_{P}^{m(P),\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] \\ &\geq \inf_{\hat{\rho}} \sup_{P\in\mathscr{P}'} \mathbb{E}_{P}^{n',\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] - \zeta.\end{aligned}$$

Since the previous is true for all $\zeta > 0$, we indeed have proven

$$\begin{aligned}\inf_{\hat{\rho}} \sup_{P\in\mathscr{P}} &\mathbb{E}_{P}^{n,\lambda}[(\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] \\ &\geq \inf_{\hat{\rho}} \sup_{P\in\mathscr{P}'} \mathbb{E}_{P}^{n',\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2].\end{aligned} \tag{S22}$$

To finish the proof of the proposition, we will now show that $\bar{\tau}_1(\tilde{P}, n)$ in (S22) can be traded for $\bar{\tau}_1(P, n, \lambda)$ at small cost. Remark that by Young's inequality, for any $P \in \mathscr{P}'$ and any $\hat{\rho}$,

$$\begin{aligned}\mathbb{E}_{P}^{n',\lambda}&[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] \\ &\geq \frac{1}{2}\mathbb{E}_{P}^{n',\lambda}[(\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] - (\bar{\tau}_1(P, n, \lambda) - \bar{\tau}_1(\tilde{P}, n, \lambda))^2,\end{aligned} \tag{S23}$$

with

$$\bar{\tau}_1(P, n, \lambda) - \bar{\tau}_1(\tilde{P}, n, \lambda)$$

$$= n\sum_{j=1}^{S}(\tilde{p}_j - p_j)e^{-(1+\lambda)np_j} - n\sum_{j=1}^{S}\tilde{p}_j e^{-(1+\lambda)np_j}\Big\{1 - e^{n(1+\lambda)(p_j-\tilde{p}_j)}\Big\}.$$

Hence,

$$|\bar{\tau}_1(P,n,\lambda) - \bar{\tau}_1(\tilde{P},n,\lambda)| \\ \leq n\sum_{j=1}^{S}|\tilde{p}_j - p_j| + n\sum_{j=1}^{S}\tilde{p}_j e^{-(1+\lambda)np_j}|1 - e^{n(1+\lambda)(p_j-\tilde{p}_j)}|. \quad \text{(S24)}$$

The first term of the rhs of the last display is easily seen to be bounded by $n\delta$ since $|p_j - \tilde{p}_j| = \tilde{p}_j|\sum_{k=1}^{S} p_k - 1| \leq \delta\tilde{p}_j$ for all $j = 1, \ldots, S$. For the second term, we use that $0 \leq 1 - e^{-x} \leq x$ for all $x \geq 0$. Hence, if $p_j \leq \tilde{p}_j$ we have,

$$\begin{aligned}\tilde{p}_j e^{-(1+\lambda)np_j}|1 - e^{n(1+\lambda)(p_j-\tilde{p}_j)}| &= \tilde{p}_j e^{-(1+\lambda)np_j}(1 - e^{-n(1+\lambda)(\tilde{p}_j-p_j)}) \\ &\leq n(1+\lambda)|\tilde{p}_j - p_j| \cdot \tilde{p}_j \\ &\leq n\delta(1+\lambda) \cdot \tilde{p}_j^2,\end{aligned}$$

while if $p_j > \tilde{p}_j$

$$\begin{aligned}\tilde{p}_j e^{-(1+\lambda)np_j}|1 - e^{n(1+\lambda)(p_j-\tilde{p}_j)}| &= \tilde{p}_j e^{-(1+\lambda)n\tilde{p}_j}(1 - e^{-n(1+\lambda)(p_j-\tilde{p}_j)}) \\ &\leq n(1+\lambda)|\tilde{p}_j - p_j| \cdot \tilde{p}_j \\ &\leq n\delta(1+\lambda) \cdot \tilde{p}_j^2.\end{aligned}$$

Therefore in any cases the second term of the rhs of Equation (S24) is bounded above by $n^2\delta(1+\lambda)\sum_{j=1}^{S}\tilde{p}_j^2$, and thus

$$|\bar{\tau}_1(P,n,\lambda) - \bar{\tau}_1(\tilde{P},n,\lambda)| \leq n\delta + n^2\delta(1+\lambda)\sum_{j=1}^{S}\tilde{p}_j^2 \leq \Big(1 + \frac{n\xi(1+\lambda)}{S(1-\delta)}\Big)n\delta.$$

This estimate combined with (S22) and (S23) completes the proof for the first inequality of the proposition. The second inequality simply follows from the first by an application of Markov's inequality.

S5.3 Proof of Lemma S1

The proof is a trivial adaptation of the classical Le Cam method with two fuzzy hypotheses, as also described in Tsybakov (2009).

Let $\hat{\rho}$ be fixed but arbitrary and let define for convenience the events $A_n(P;\hat{\rho}) \coloneqq \{\boldsymbol{Y}_N : |\bar{\tau}_1(P,n,\lambda) - \hat{\rho}(\boldsymbol{Y}_N)| > n\varepsilon\}$. Since the average is always less or equal than the supremum over $\mathscr{P}'$, we establish that

$$\begin{aligned}&\sup_{P\in\mathscr{P}'}\mathbb{P}_P^{n',\lambda}(A_n(P;\hat{\rho})) \\ &\qquad\geq \frac{1}{2}\mathbb{E}\big[\mathbb{P}_{Q_1}^{n',\lambda}(A_n(Q_1;\hat{\rho}))\mathbf{1}_{\mathscr{P}'}(Q_1)\big] + \frac{1}{2}\mathbb{E}\big[\mathbb{P}_{Q_2}^{n',\lambda}(A_n(Q_2;\hat{\rho}))\mathbf{1}_{\mathscr{P}'}(Q_2)\big] \\ &\qquad\geq \frac{1}{2}\mathbb{E}\big[\mathbb{P}_{Q_1}^{n',\lambda}(A_n(Q_1;\hat{\rho}))\big] + \frac{1}{2}\mathbb{E}\big[\mathbb{P}_{Q_2}^{n',\lambda}(A_n(Q_2;\hat{\rho}))\big] - \alpha,\end{aligned}$$

where for the last line we have used the item 1 of the Lemma.

Now let define the events $B_n(Q_j;\hat{\rho}) := \{\boldsymbol{Y}_N \,:\, |\mathbb{E}[\bar{\tau}_1(Q_j,n,\lambda)] - \hat{\rho}(\boldsymbol{Y}_N)| > n\varepsilon/2\}$, for $j=1,2$. Under item 2 of the Lemma, it is rapidly obtained from the last display that

$$\begin{aligned}
&\sup_{P\in\mathscr{P}'} \mathbb{P}_P^{n',\lambda}(A_n(P;\hat{\rho}))\\
&\qquad \geq \frac{1}{2}\mathbb{E}\big[\mathbb{P}_{Q_1}^{n',\lambda}(B_n(Q_1;\hat{\rho}))\big] + \frac{1}{2}\mathbb{E}\big[\mathbb{P}_{Q_2}^{n',\lambda}(B_n(Q_2;\hat{\rho}))\big] - \alpha - \beta\\
&\qquad = \frac{1}{2}\mathbb{E}\big[1 - \mathbb{P}_{Q_1}^{n',\lambda}(B_n(Q_1;\hat{\rho})^c) + \mathbb{P}_{Q_2}^{n',\lambda}(B_n(Q_2;\hat{\rho}))\big] - \alpha - \beta.
\end{aligned}$$

But under item 3 of the lemma, we have that $B_n(Q_1;\hat{\rho})^c \subseteq B_n(Q_2;\hat{\rho})$. Moreover under Q_j, $j=1,2$, $\boldsymbol{Y}_N$ is a vector of independent Poisson random variables with parameters $(n'q_{j,1},\dots,n'q_{j,S},0,\dots)$ and thus by the classical Le Cam's trick the last equation is bounded by

$$\begin{aligned}
&\sup_{P\in\mathscr{P}'} \mathbb{P}_P^{n',\lambda}(A_n(P;\hat{\rho}))\\
&\qquad\qquad \geq \frac{1}{2}\Big(1 - \mathsf{TV}\big(\mathbb{E}[\otimes_{j=1}^S \mathrm{Poiss}(n'q_{1,j})],\, \mathbb{E}[\otimes_{j=1}^S \mathrm{Poiss}(n'q_{2,j})]\big)\Big) - \alpha - \beta\\
&\qquad\qquad \geq \frac{1}{2}\Big(1 - \gamma - 2\alpha - 2\beta\Big),
\end{aligned}$$

where the last line follows from the item 4 of the Lemma. Since the rhs of the last display is independent of $\hat{\rho}$, the conclusion of the Lemma follows.

S5.4 Proof of Lemma S2

The proof of Lemma S2 follows the guidelines used in the papers Wu and Yang (2019, 2016), relating the problem of the existence of the random variables to the problem of finding the best polynomial approximation to some function.

For $a,b\in\mathbb{R}$, we let $\mathsf{C}[a,b]$ denote the space of continuous functions on $[a,b]$, and for any $L\in\mathbb{Z}_+$ we let $\mathsf{P}_L[a,b]\subset\mathsf{C}[a,b]$ denote the space of polynomials of degree no more than L on $[a,b]$. For any $f\in\mathsf{C}[a,b]$, the best polynomial (of degree at most L) approximation to f is defined as

$$E_L(f,[a,b]) := \inf\{\sup\{|f(x)-q(x)| \,:\, x\in[a,b]\} \,:\, q\in\mathsf{P}_L[a,b]\}.$$

For the sake of simplicity, we define $B := n(1+\lambda)\xi/(2S)$. We also define $g:[\xi^{-1},1]\to\mathbb{R}_+$ such that $g(x) := \exp\{-2Bx\}$. It is a classical result that for any $L\in\mathbb{N}$ we can find random variables X and Y taking values in $[\xi^{-1},1]$ and such that

$$\begin{aligned}
\mathbb{E}[X^k] &= \mathbb{E}[Y^k], \qquad k=0,\dots,L,\\
\mathbb{E}[g(X)] &= \mathbb{E}[g(Y)] + E_L(g,[\xi^{-1},1]).
\end{aligned}$$

The proof of the existence of such random variables can be found for instance in Wu and Yang (2016, 2019) for a constructive argument, or for instance in Lepski et al. (1999) using the Hahn-Banach theorem and a duality argument.

We now assume that we have at our disposal the random variables X and Y of the previous paragraph, and we write P_X and P_Y their distributions. The construction of the random variables U and V is done using the trick introduced in Wu and Yang (2016, Lemma 4). Namely, we let U and V having respective distributions on $[0,\xi S^{-1}]$

$$P_U(\mathrm{d}x) := \big(1-\mathbb{E}[(\xi X)^{-1}]\big)\delta_0 + (Sx)^{-1}P_{\xi X/S}(\mathrm{d}x),$$

$$P_V(\mathrm{d}x) := \big(1 - \mathbb{E}[(\xi Y)^{-1}]\big)\delta_0 + (Sx)^{-1} P_{\xi Y/S}(\mathrm{d}x).$$

Because $X, Y \geq \xi^{-1}$ almost-surely, then $\mathbb{E}[(\xi X)^{-1}] \leq 1$ and $\mathbb{E}[(\xi Y)^{-1})] \leq 1$. Indeed from Wu and Yang (2016, Lemma 4), P_U and P_V are proper probability distributions on $[0, \xi S^{-1}]$ satisfying

$$\begin{aligned}
&\mathbb{E}[U] = \mathbb{E}[V] = 1/S, \qquad \mathbb{E}[U^k] = \mathbb{E}[V^k], \qquad k = 0, \dots, L+1, \\
&\mathbb{E}[U \exp\{-n(1+\lambda)U\}] = \mathbb{E}[V \exp\{-n(1+\lambda)V\}] + S^{-1} E_L(g, [\xi^{-1}, 1]).
\end{aligned}$$

Furthermore, it is clear that,

$$\mathbb{E}[U^2] = \frac{1}{S} \int x \, P_{\xi x/S}(\mathrm{d}x) = \frac{\xi \mathbb{E}[X]}{S^2} \leq \frac{\xi}{S^2}.$$

Hence $\mathrm{Var}(U) \leq \xi/S^2$. It is obvious that we also have $\mathrm{Var}(V) \leq \xi/S^2$. Thus, the proof of the theorem is finished by obtaining a lower bound on the best polynomial approximation $E_L(g, [\xi^{-1}, 1])$. This is a consequence of the Lemma 2 in the main paper since $L \leq K_1 \xi$, $B = (\xi/2)(1 + O(\xi^{-1}))$, and also because

$$\frac{\xi}{2} \varphi\Big(\frac{2L}{\xi}\Big) \leq \frac{\xi}{2} \cdot \frac{1}{2} \Big(\frac{2L}{\xi}\Big)^2 = \frac{L^2}{\xi},$$

by using the facts about φ derived in Section S6.3.

S5.5 Proof of Proposition S3

Here we prove separately all the items stated in Proposition S3.

Proof of item 1. The proof is an immediate consequence of Bernstein's inequality using that $\mathrm{Var}(U) \leq \xi S^{-2}$ and $0 \leq U \leq \xi S^{-1}$. Similarly for V. □

Proof of item 2. The proof for Q_1 and Q_2 are identical, thus we only prove the result for Q_1. By definition, we have that

$$\bar{\tau}_1(Q_1, n, \lambda) = n \sum_{j=1}^{S} U_j e^{-n(1+\lambda)U_j}.$$

Whence, $\bar{\tau}_1(Q_1, n, \lambda)$ is a sum of i.i.d. random variables taking values in $[0, n\xi S^{-1}]$. By Hoedffding's inequality,

$$\mathbb{P}\big(|\bar{\tau}_1(Q_1, n, \lambda) - \mathbb{E}[\bar{\tau}_1(Q_1, n, \lambda)]| > n\varepsilon/2\big) \leq 2 \exp\Big\{ -\frac{S\varepsilon^2}{2\xi} \Big\}.$$

The conclusion follows from simple algebraic manipulations. □

Proof of item 3. This is immediate by remarking that $\mathbb{E}[\bar{\tau}_1(Q_1, n, \lambda)] = nS\mathbb{E}[Ue^{-n(1+\lambda)U}]$ and $\mathbb{E}[\bar{\tau}_1(Q_2, n, \lambda)] = nS\mathbb{E}[Ve^{-n(1+\lambda)V}]$. □

Proof of item 4. Since $(U_1, \dots, U_S)$ and $(V_1, \dots, V_S)$ are independent and i.i.d vectors, we obtain immediately that

$$\mathsf{TV}(\mathbb{E}[\otimes_{j=1}^S \mathrm{Poiss}(n'U_j)],\, \mathbb{E}[\otimes_{j=1}^S \mathrm{Poiss}(n'V_j)])$$

$$= S\mathsf{TV}(\mathbb{E}[\mathrm{Poiss}(n'U)], \mathbb{E}[\mathrm{Poiss}(n'V)]). \quad \text{(S25)}$$

Since $0 \leq U, V \leq \xi S^{-1}$ almost-surely, we obtain from (Wu and Yang, 2019, Lemma 6),

$$\mathsf{TV}\Big(\mathbb{E}[\mathrm{Poiss}(n'U)], \mathbb{E}[\mathrm{Poiss}(n'V)]\Big) \leq \frac{1}{(L+2)!}\Big(\frac{n'\xi}{2S}\Big)^{L+2}\Big(2 + 2^{n'\xi/(2S)-L} + 2^{n'\xi/(2\log(2)S)-L}\Big).$$

Recall that $n' = n(1+\delta)$, thus under the conditions of the proposition we have $n'\xi/(2S) \leq n'\xi/(2\log(2)S) \leq L$, and hence from the last display we obtain that

$$\mathsf{TV}\Big(\mathbb{E}[\mathrm{Poiss}(n'U)], \mathbb{E}[\mathrm{Poiss}(n'V)]\Big) \leq \frac{4}{(L+2)!}\Big(\frac{n\xi(1+\delta)}{2S}\Big)^{L+2}. \quad \text{(S26)}$$

Then the conclusion follows by combining Equations (S25) and (S26). □

S5.6 Proof of Proposition S4

Here we prove separately all the items stated in Proposition S4.

Proof of item 1. From the definitions of ξ, S and δ, we immediately see that $(1 + \frac{n\xi(1+\lambda)}{S(1-\delta)})^2\delta^2 \leq (1 + \frac{\xi}{1-\delta})^2 c_0^2\varepsilon^2/\xi^2 = c_0^2\varepsilon^2(1+o(1))$, because $\xi \to \infty$ and $\varepsilon = O(1)$ (the latter fact is easier to see a posteriori). □

Proof of items 2 and 3. The case $1 + \lambda > \log(n)$ is straightforward, thus we focus only on $1+\lambda \leq \log(n)$. For the sake of simplicity, we define $r := \sqrt{\log(n)/(1+\lambda)}$ and $y := \sqrt{ec_1}A(\lambda, n)$, so that $\varepsilon = \sqrt{2}c_3(ry)^{-1}\exp(-r^2y^2/2)$. Then from the definitions of S, δ and ξ,

$$\begin{aligned} S\delta^2 &\geq c_0^2 n(1+\lambda)\frac{\varepsilon^2}{\xi^3}\cdot\xi \\ &\gtrsim n\cdot\max\Big\{\frac{1}{(1+\lambda)^2\log^3(n)}, \frac{1+\lambda}{\log^6(n)}\Big\}\varepsilon^2\cdot\xi \\ &\gtrsim \xi\cdot n\cdot\max\Big\{\frac{1}{(1+\lambda)^2\log^3(n)}, \frac{1+\lambda}{\log^6(n)}\Big\}\frac{1}{(ry)^2}e^{-r^2y^2}. \end{aligned}$$

But under the assumption of the Proposition, have $\liminf_n \big\{\frac{\log(n)}{r^2y^2}\big\} > 1$, which entails that for n large enough $S\delta^2 \geq 2\xi(1+\delta/3)\log(20)$. The proof of item 3 is similar. □

Proof of item 4. This is an immediate consequence of the definitions of ε, L and ξ. □

Proof of item 5, Case $1+\lambda \leq \log(n)$*.* Note that in this case we have $\xi = (2c_1/e)(1+\lambda)\log(n)$ and $L = \lceil c_1 A(\lambda, n)\log(n)\rceil$. For n large enough such that $0 < \delta \leq e\log(2) - 1$ (this always happens, see for instance the remark in the proof of item 1), we have

$$\begin{aligned} 2\log(2)LS &\geq 2\log(2)c_1A(\lambda,n)\log(n)\cdot n(1+\lambda) \\ &= n\xi\cdot e\log(2)A(\lambda,n) \\ &\geq n\xi\cdot e\log(2) \\ &\geq n\xi(1+\delta), \end{aligned}$$

where the third line follows because $\lambda \geq 0$ and from the definition of $A(\lambda, n)$ by remarking that $a \log a \geq 0 \Rightarrow a \geq 1$.

Further, using that $(L+2)! \geq L^2 L!$, and because $L \leq K_1 \xi$ implies that $(1+\delta)^{L+2} \lesssim (1+\delta)^L \leq e^{L\delta} \leq e^{K_1 c_1 \varepsilon} \lesssim 1$, we have

$$\begin{aligned}\frac{4S}{(L+2)!}\Big(\frac{n\xi(1+\delta)}{2S}\Big)^{L+2} &\lesssim \frac{S}{L^2}\Big(\frac{n\xi}{2S}\Big)^2 \frac{1}{L!}\Big(\frac{n\xi}{2S}\Big)^L \\ &= \frac{S}{L^2}\Big(\frac{c_1 \log(n)}{e}\Big)^2 \frac{1}{L!}\Big(\frac{c_1 \log(n)}{e}\Big)^L \\ &\lesssim \frac{S \log^2(n)}{L^{5/2}}\Big(\frac{c_1 \log(n)}{L}\Big)^L,\end{aligned}$$

where the last line follows from Stirling's formula. Using the definitions of $S \lesssim n(1+\lambda)$, L and $A(\lambda, n)$, we deduce that

$$\begin{aligned}\frac{4S}{(L+2)!}\Big(\frac{n\xi(1+\delta)}{2S}\Big)^{L+2} &\lesssim \frac{n(1+\lambda)A(\lambda,n)^{-c_1 A(\lambda,n)\log(n)}}{A(\lambda,n)^{5/2}\log^{1/2}(n)} \\ &= \frac{1}{c_2 A(\lambda,n)^{5/2}} \leq \frac{1}{c_2}.\end{aligned}$$

Therefore by choosing $c_2 > 0$ large enough we obtain that $\gamma(2S)^{L+2}(L+2)! \geq 4S(n\xi(1+\delta))^{L+2}$. □

Proof of item 5, Case $1+\lambda > \log(n)$. Note that in this case we have $\xi = (2c_1/e)\log^2(n)$ and $L = \lceil 2c_1 \log(n) \rceil$. For n large enough such that $0 < \delta \leq 2e\log(2) - 1$ (this always happens, see for instance the remark in the proof of item 1), we have

$$\begin{aligned}2\log(2)LS &\geq 4c_1 \log(2) n(1+\lambda)\log(n) \\ &\geq 4c_1 \log(n) n \log^2(n) \\ &= n\xi \cdot 2e\log(2) \\ &\geq n\xi(1+\delta).\end{aligned}$$

Proceeding along similar lines as for the case $1+\lambda \leq \log(n)$, it is easily found that as $n \to \infty$ we have

$$\frac{4S}{(L+2)!}\Big(\frac{n\xi(1+\delta)}{2S}\Big)^{L+2} \to 0,$$

and hence certainly that $\gamma(2S)^{L+2}(L+2)! \geq 4S(n\xi(1+\delta))^{L+2}$ when n gets large enough. □

S5.7 Proof of Proposition S5

We define the function $\varphi : \mathbb{R}_+ \to \mathbb{R}$ such that $\varphi(x) = x\log(x)$. When $1+\lambda \leq \log(n)$, it is clear that $A(\lambda, n)$ converges to the solution of $\varphi(x) = c_1^{-1} = e$, hence $A(\lambda, n) \to e$, which proves the first claim.

For the second claim, let define,

$$\Delta_n := e\frac{\log(1+\lambda) - (1/2)\log\log(n) + \log(c_2)}{\log(n)}.$$

For n large enough such that $\Delta_n > -1$, it is clear than $A(\lambda, n) \geq 0$. Furthermore, by a Taylor expansion of φ near $x = e$, we find that there is a $\bar{x}$ in the line segment between $A(\lambda, n)$ and e

such that

$$\begin{aligned}\varphi(A(\lambda,n)) &= \varphi(e)+\varphi'(e)(A(\lambda,n)-e)+\frac{\varphi''(\bar{x})}{2}(A(\lambda,n)-e)^2\\ &\geq \varphi(e)+\varphi'(e)(A(\lambda,n)-e),\end{aligned}$$

because $\varphi''(x) = 1/x > 0$ whenever $x > 0$. Since $\varphi(A(\lambda,n)) - \varphi(e) = \Delta_n$, $\varphi(e) = e$, and $\varphi'(e) = 2$, we deduce that for those n large,

$$0 \leq A(\lambda,n) \leq e + \Delta_n/2.$$

Therefore,

$$\begin{aligned}e^{-1}A(\lambda,n)^2\log(n) &\leq e\log(n)+\Delta_n\log(n)+\frac{\Delta_n^2\log(n)}{4e}\\ &= e\log(n)+e\log\frac{c_2(1+\lambda)}{\sqrt{\log(n)}}+o(1).\end{aligned}$$

This concludes the proof.

S6 Proofs related to the upper-bound on the best polynomial approximation

In this section, we give the proofs of the results stated in Section S4, regarding the construction of a polynomial of degree no more than L achieving the approximation error of the Lemma 2 in the main document.

S6.1 Proof of Lemma S3

Below we prove the items stated in Lemma S3. The proofs mainly consist on driving the formula for $a_k(C)$ and getting sharp estimates on $|a_k(C)|$ for various regimes governed by the ratio k/C.

Proof of item (1). By doing the change of variable $x \mapsto \cos(\theta)$ in the definition of $a_k(C)$, and using that $T_k(\cos\theta) = \cos(k\theta)$ we obtain

$$\begin{aligned}a_k(C) &= e^{-C}\int_{-1}^{1}\frac{e^{-Cx}T_k(x)}{\sqrt{1-x^2}}\mathrm{d}x\\ &= e^{-C}\int_0^{\pi}e^{-C\cos\theta}\cos(k\theta)\mathrm{d}\theta\\ &= \pi(-1)^k e^{-C}I_k(C),\end{aligned}$$

where we used (Olver et al., 2010, formula 10.32.3). □

Proof of item (2). The uniform convergence of the series is an immediate consequence of the fact that $|T_n(x)| \leq 1$ for all $x \in [-1,1]$ and the upper bound estimate on $|a_k(C)|$ obtained just after in item (3). □

Proof of item (3). To prove the item, we use the classical bound on the modified Bessel function obtained by Luke (1972). Indeed, for any $k \geq BC$, we have

$$0 \leq \pi e^{-C}I_k(C) \leq \frac{\pi}{k!}\Big(\frac{C}{2}\Big)^k$$

$$
\begin{aligned}
&\leq \frac{\pi}{\sqrt{2\pi k}}\Big(\frac{eC}{2k}\Big)^k \\
&\leq \sqrt{\frac{\pi}{2k}}\exp\Big\{-k\log\Big(\frac{2B}{e}\Big)\Big\},
\end{aligned}
$$

where the first line comes from Luke (1972), and the second line by Stirling's approximation. For $k \geq 2$ we have $\pi/(2k) \leq 1$. Thus, it is enough to take $B_0 = e^{(1+D)}/2$, which concludes the proof. □

Proof of item (4). We follow a similar path as in the Section B.3 of the main document. Indeed, we can remark by Stirling's formula that for any $p, k \geq 0$ we have

$$
(p+k)! \geq \sqrt{2\pi}(p+k)^{p+k+1/2}e^{-(p+k)}, \qquad p! \geq \sqrt{2\pi}p^{p+1/2}e^{-p}.
$$

Then, by defining $\phi_{z,k}(x)$ as in Section B.3 of the main document, we obtain the upper bound,

$$
\begin{aligned}
e^{-C}I_k(C) &\leq \frac{e^{-C}}{k!}\Big(\frac{C}{2}\Big)^k + \sum_{p\geq 1}\frac{1}{p!(p+k)!}\Big(\frac{C}{2}\Big)^{2p+k} \\
&\leq \frac{e^{-C}}{k!}\Big(\frac{C}{2}\Big)^k + \frac{1}{2\pi}\sum_{p\geq 1}\frac{\exp\{\phi_{C,k}(p)\}}{\sqrt{p(p+k)}}. \qquad \text{(S27)}
\end{aligned}
$$

We consider the first term of the rhs of the previous display. By Stirling's formula, we have

$$
\begin{aligned}
\frac{e^{-C}}{k!}\Big(\frac{C}{2}\Big)^k &\leq \frac{e^{-C}}{\sqrt{2\pi k}}\Big(\frac{eC}{2k}\Big)^k \\
&= \frac{1}{\sqrt{2\pi k}}\exp\Big\{-C + k\log\Big(\frac{eC}{2k}\Big)\Big\} \\
&= \frac{1}{\sqrt{2\pi k}}\exp\{\phi_{C,k}(0)\},
\end{aligned}
$$

where $\phi_{C,k}(0)$ is defined by extending $\phi_{C,k}$ at zero by continuity. We remark that,

$$
\begin{aligned}
&\phi_{C,k}(0) - \phi_{C,k}(x_0) \\
&\quad = -2x_0 + x_0\log x_0 - k\log k + (x_0+k)\log(x_0+k) - 2x_0\log\frac{z}{2} \\
&\quad = -2x_0 - k\log k + k\log(x_0+k) + x_0\Big(\log x_0 + \log(x_0+k) - 2\log\frac{z}{2}\Big) \\
&\quad = -2x_0 + k\log\Big(1+\frac{x_0}{k}\Big) - x_0\phi'_{C,k}(x_0) \\
&\quad = -2x_0 + k\log\Big(1+\frac{x_0}{k}\Big) \\
&\quad \leq -x_0.
\end{aligned}
$$

It follows,

$$
\frac{e^{-C}}{k!}\Big(\frac{C}{2}\Big)^k \leq \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{C}}\cdot\sqrt{\frac{C}{2\pi k}}e^{-x_0} = \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{C}}\cdot o(1)
$$

as $C \to \infty$, by remarking that $C/k \lesssim \sqrt{C}$ and that $k \leq BC$, hence $C/k \geq B^{-1}$ and $x_0 \geq B'k > B'\sqrt{C} \to \infty$ for a universal constant $B' > 0$.

We now consider the second term in the rhs of (S27). We let p_0 be the integer defined in Section B.3 of the main document, that is $x_0 < p_0 \leq x_0 + 1$ is integer and $\phi'_{z,k}(x_0) = 0$. Recall that $x_0 \geq B'k > B'\sqrt{C} \to \infty$ for a universal constant $B' > 0$. Let $G_1 > 0$ be a constant to be chosen accordingly later, and let $A_1 \in \mathbb{N}$ be the only integer such that

$$x_0 - G_1\sqrt{x_0 \log(x_0)} - 1 < A_1 \leq x_0 - G_1\sqrt{x_0 \log(x_0)}.$$

By the previous discussion, we have $1 < A_1 < x_0$ at least for L large enough. Similarly, we let $G_2 > 0$ a constant to be chosen accordingly, and we let $A_2 \in \mathbb{N}$ be the only integer such that

$$x_0 + G_2(1 + \sqrt{x_0}) \log(x_0) \leq A_2 < x_0 + G_2(1 + \sqrt{x_0}) \log(x_0) + 1.$$

Obviously $A_2 > x_0$. Then we decompose the sum in the rhs of (S27) as

$$\underbrace{\sum_{p=1}^{A_1} \frac{\exp\{\phi_{C,k}(p)\}}{\sqrt{p(p+k)}}}_{S_1} + \underbrace{\sum_{p=A_1+1}^{A_2} \frac{\exp\{\phi_{C,k}(p)\}}{\sqrt{p(p+k)}}}_{S_2} + \underbrace{\sum_{p>A_2} \frac{\exp\{\phi_{C,k}(p)\}}{\sqrt{p(p+k)}}}_{S_3}.$$

The conclusion of the proof follows by gathering the bounds for S_1, S_2, and S_3, which are derived in the paragraphs below, and by using that $\phi_{C,k}(x_0) = -C\varphi(k/C)$.

Bound on S_1 Let $p \in [1, A_1]$. We remark by a Taylor expansion that $\phi_{C,k}(p) = \phi_{C,k}(x_0) + \frac{1}{2}\phi''_{C,k}(\bar{p})(p - x_0)^2$ for some $\bar{p} \in (1, x_0)$. As for Section B.3 of the main document, we see that $\phi''_{C,k}(\bar{p}) \leq -1/x_0$. Therefore, remarking that $(p - x_0)^2 \geq G_1^2 x_0 \log(x_0)$ for any $1 \leq p \leq A_1$ (at least for L large enough),

$$\begin{aligned} S_1 &\leq \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{k}} \sum_{p=1}^{A_1} \exp\Big(-\frac{(p - x_0)^2}{2x_0}\Big) \\ &\leq \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{C}} \cdot A_1\sqrt{\frac{C}{k}} x_0^{-G_1^2/2} \\ &= \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{C}} \cdot o(1), \end{aligned}$$

where the last line follows by choosing G_1 large enough, because $A_1 \leq x_0$, $C/k \lesssim \sqrt{C}$, and $x_0 \geq B'\sqrt{C} \to \infty$.

Bound on S_2 Let $A_1 < p \leq A_2$. Then, $|p - x_0| = O\big(\sqrt{x_0}\log(x_0)\big)$ as $C \to \infty$. Further, it is easily seen that, as $C \to \infty$,

$$\begin{aligned} \sup_{x \in [A_1, A_2]} |\phi'''_{C,k}(x)| &= \sup_{x \in [A_1, A_2]} \Big(\frac{1}{x^2} + \frac{1}{(x+k)^2}\Big) \\ &\leq 2 \sup_{x \in [A_1, A_2]} \frac{1}{x^2} \\ &= \frac{2(1 + o(1))}{x_0^2}. \end{aligned}$$

Therefore, by Taylor expansion, and as $C \to \infty$,

$$\phi_{C,k}(p) = \phi_{C,k}(x_0) + \frac{1}{2}\phi''(x_0)(p - x_0)^2 + O\Big(\frac{x_0^{3/2}\log^3(x_0)}{x_0^2}\Big)$$

$$= \phi_{C,k}(x_0) + \frac{1}{2}\phi''(x_0)(p - x_0)^2 + o(1).$$

It follows,

$$\begin{aligned}
S_2 &\leq (1 + o(1)) \cdot \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{A_1(A_1 + k)}} \sum_{p=A_1+1}^{A_2} \exp\left(\frac{1}{2}\phi''(x_0)(p - x_0)^2\right) \\
&\leq (1 + o(1)) \cdot \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{x_0(x_0 + k)}} \sum_{p=-\infty}^{\infty} \exp\left(\frac{1}{2}\phi''(x_0)p^2\right) \\
&\leq (1 + o(1)) \cdot \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{x_0(x_0 + k)}} \left\{1 + 2\sum_{p=1}^{\infty} \exp\left(\frac{1}{2}\phi''(x_0)p^2\right)\right\} \\
&\leq (1 + o(1)) \cdot \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{x_0(x_0 + k)}} \left\{1 + 2\int_0^{\infty} \exp\left(\frac{1}{2}\phi''(x_0)t^2\right) \mathrm{d}t\right\} \\
&\leq (1 + o(1)) \cdot \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{x_0(x_0 + k)}} \left\{1 + \sqrt{\frac{2\pi}{-\phi''_{C,k}(x_0)}}\right\}.
\end{aligned}$$

It is proven in the Section B.3 of the main document that $x_0(x_0 + k) = C^2/4$ and $-\phi''_{C,k}(x_0) = (4/C)\sqrt{1 + (k/C)^2}$. It follows, as $C \to \infty$,

$$\begin{aligned}
S_2 &\leq \sqrt{2\pi}(1 + o(1)) \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{C} \cdot \left(1 + (k/C)^2\right)^{1/4}} \\
&\leq \sqrt{2\pi}(1 + o(1)) \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{C}}.
\end{aligned}$$

Bound on S_3 Let $p > A_2$. Remark that for L large enough we also have $p > x_0 + \sqrt{x_0}$. Then, by performing two Taylor expansions, we find that there is $\bar{x} \in (x_0, x_0 + \sqrt{x_0})$ and $\bar{p} \in (x_0 + \sqrt{x_0}, p)$ such that

$$\begin{aligned}
\phi_{C,k}(p) &= \phi_{C,k}(x_0 + \sqrt{x_0}) + \phi'_{C,k}(\bar{p})(p - x_0 - \sqrt{x_0}) \\
&= \phi_{C,k}(x_0) + \frac{1}{2}\phi''_{C,k}(\bar{x})x_0 + \phi'_{C,k}(\bar{p})(p - x_0 - \sqrt{x_0}) \\
&\leq \phi_{C,k}(x_0) + \phi'_{C,k}(\bar{p})(p - x_0 - \sqrt{x_0}),
\end{aligned}$$

where the last line follows because $\phi''_{C,k}(\bar{x}) < 0$. By the results of Section B.3 of the main document, we can also see that

$$\begin{aligned}
\phi'_{C,k}(\bar{p}) &= \phi_{C,k}(x_0) - \log\frac{\bar{p}}{x_0} - \log\frac{\bar{p} + k}{x_0 + k} \\
&= -\log\frac{\bar{p}}{x_0} - \log\frac{\bar{p} + k}{x_0 + k} \\
&\leq -\log\frac{\bar{p}}{x_0} \\
&\leq -\log\left(1 + \frac{1}{\sqrt{x_0}}\right) \\
&\leq -\frac{1}{1 + \sqrt{x_0}}.
\end{aligned}$$

Hence because $p > x_0 + \sqrt{x_0}$,

$$\phi_{C,k}(p) \leq \phi_{C,k}(x_0) - \frac{p - x_0 - \sqrt{x_0}}{1 + \sqrt{x_0}}.$$

It follows,

$$\begin{aligned} S_3 &\leq \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{x_0(x_0+k)}} \sum_{p > A_2} \exp\Big(- \frac{p - x_0 - \sqrt{x_0}}{1 + \sqrt{x_0}} \Big) \\ &\leq \frac{e \cdot \exp\{\phi_{C,k}(x_0)\}}{\sqrt{x_0(x_0+k)}} \cdot x_0^{-G_2} \cdot \sum_{p \geq 0} \exp\Big(- \frac{p}{1 + \sqrt{x_0}} \Big) \\ &= \frac{e \cdot \exp\{\phi_{C,k}(x_0)\}}{\sqrt{x_0(x_0+k)}} \cdot x_0^{-G_2} \cdot (1 + \sqrt{x_0}). \end{aligned}$$

It is shown in Section B.3 of the main document that $x_0(x_0+k) = C^2/4$. Therefore, for $C \to \infty$ and G_2 sufficiently large,

$$S_3 = \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{C}} \cdot o(1). \qquad \square$$

S6.2 Proof of Corollary S1

By item (2) of Lemma S3, we obtain immediately that

$$E_L(\gamma_C, [-1,1]) \leq \sup_{x \in [-1,1]} |q_L(x) - \gamma_C(x)| \leq \sum_{k > L} |a_k(C)|. \tag{S28}$$

We let L' be the largest integer smaller than BC, for $B > 0$ large enough. Then, by the item (3) of Lemma S3, for any $D > 0$ we can choose B_0 such that for all $B > B_0$,

$$\sum_{k > L'} |a_k(C)| \leq \sum_{k > L'} e^{-Dk} \leq \frac{e^{-DBC}}{e^D - 1}.$$

By taking B, D sufficiently large, the contribution of the previous display in the rhs of (S28) is negligible. It remains to bound the sum from $L+1$ to L' (note that for B large enough, we have $L' > L$). By the item item (4) of Lemma S3, we obtain that

$$\begin{aligned} \sum_{k=L+1}^{L'} |a_k(C)| &\leq \sqrt{2\pi} \sum_{k=L+1}^{L'} \frac{\exp\{-C\varphi(k/C)\}}{\sqrt{C}} \\ &\leq \sqrt{2\pi} \int_L^\infty \frac{\exp\{-C\varphi(x/C)\}}{\sqrt{C}} \mathrm{d}x \\ &= \sqrt{2\pi C} \int_{L/C}^\infty \exp\{-C\varphi(x)\} \mathrm{d}x, \end{aligned}$$

where the second line follows because φ is monotone increasing on (L, ∞), because $\varphi' > 0$ (see for instance Section S6.3). Interestingly, the function φ' is also monotone increasing $(L/C, \infty)$, because $\varphi'' > 0$ (see again Section S6.3). Hence, $u \geq L/C \Leftrightarrow \varphi'(u) \geq \varphi'(L/C)$, and by Markov's inequality,

$$\sum_{k=L+1}^{L'} |a_k(C)| \leq \sqrt{\frac{2\pi}{C}} \int_{L/C}^\infty \frac{C\varphi'(u) \exp\{-C\varphi(u)\}}{\varphi'(L/C)} \mathrm{d}u$$

$$= \sqrt{2\pi} \cdot \frac{\exp\{-C\varphi(L/C)\}}{\varphi'(L/C) \cdot \sqrt{C}}.$$

Now we remark that by a Taylor expansion we have $u \in (0, L/C)$, that is $u \in (0, \zeta)$, such that $\varphi'(L/C) = \varphi'(0) + \varphi''(u) \cdot L/C = \varphi''(u) \cdot L/C$. In view of Section S6.3, we deduce that

$$\varphi'(L/C) \geq \frac{1}{\sqrt{1+\zeta^2}} \cdot \frac{L}{C},$$

and thus,

$$\sum_{k=L+1}^{L'} |a_k(C)| \leq \sqrt{2\pi(1+\zeta^2)} \cdot \frac{\sqrt{C}}{L} e^{-C\varphi(L/C)}.$$

S6.3 Some results about the function φ

In this section, we collect some facts about the function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ defined in (18) of the main document. It is convenient to rewrite φ as

$$\varphi(x) := 1 - \sqrt{1+x^2} + \frac{1}{2}(-x + \sqrt{1+x^2}) \log(-x + \sqrt{1+x^2}) \\ + \frac{1}{2}(x + \sqrt{1+x^2}) \log(x + \sqrt{1+x^2}).$$

Then,

$$\varphi'(x) = -\frac{(-x + \sqrt{1+x^2}) \log(-x + \sqrt{1+x^2})}{2\sqrt{1+x^2}} + \frac{(x + \sqrt{1+x^2}) \log(x + \sqrt{1+x^2})}{2\sqrt{1+x^2}},$$
$$\varphi''(x) = \frac{1}{(1+x^2)^{1/2}}, \qquad \varphi'''(x) = -\frac{x}{(1+x^2)^{3/2}}.$$

By a Taylor expansion of φ near 0, we find that there is a $y \in (0, x)$ such that

$$\varphi(x) = \varphi(0) + \varphi'(0)x + \frac{1}{2}\varphi''(0)x^2 + \frac{1}{6}\varphi'''(y)x^3 \leq \frac{x^2}{2},$$

because $\varphi(0) = \varphi'(0) = 0$ and $\varphi'''(y) \leq 0$ for all $y \geq 0$ by the computations above. Similarly, there is $y \in (0, x)$ such that,

$$|\varphi'(x)| \leq |\varphi'(0)| + |\varphi''(y)||x| \leq |x|.$$

References

Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990). Disclosure control of microdata. *J. Amer. Statist. Assoc.* **85**, 38–45.

Carota, C., Filippone, M., Leombruni, R. and Polettini, S. (2015). Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models. *Ann. Appl. Statist.* **9**, 525–546.

Carota, C., Filippone, M. and Polettini, S. (2018). Assessing Bayesian nonparametric log-linear models: an application to disclosure risk estimation. *Preprint: arXiv:1801.05244*

Efron, B. and Morris, C (1973). Stein's estimation rule and its competitors - an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68**, 117–130.

Ferguson, T.S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *Ann. Statist.* **1**, 209–230.

Lepski, O., Nemirovski, A. and Spokoiny, V. (1999). On estimation of the L_r norm of a regression function. *Probab. Theory and Related Fields* **113**, 221–253.

Luke, Y.L. (1972). Inequalities for generalized hypergeometric functions. *J. Approximation Theory*, **5**, 41–65.

Manrique-Vallier, D. and Reiter, J.P. (2012). Estimating identification disclosure risk using mixed membership models. *J. Amer. Statist. Assoc.* **107** 1385–1394.

Manrique-Vallier, D. and Reiter, J.P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *J. Comput. Graph. Statist.* **23** 1061–1079.

Olver, F.W.J., Lozier, D.W., Boisvert, R.F. and Clark, C.W. (2010). *NIST handbook of mathematical functions*, Cambridge University Press.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. 3rd Berkeley Symp.*,**1**, 157–163.

Samuels, S.M. (1998). A Bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *J. Off. Statist.* **14**, 373–383.

Skinner, C., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata. *J. Off. Stat.* **10**, 31–51.

Tsybakov, A. B. (2009) *Introduction to nonparametric estimation.* Springer Science & Business Media.

Wu, Y. and Yang, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inform. Theory* **62**, 3702–3720.

Wu, Y. and Yang, P. (2015). Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *Ann. Statist.*, to appear.