Monotone Matrix Estimation via Robust Deconvolution

Devavrat Shah DEVAVRAT@MIT.EDU
Dogyoon Song DGSONG@MIT.EDU
LIDS, SDSC, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139

Abstract

The goal of deconvolution is in estimating the distribution of a random variable based on its noisy observations. The goal of matrix estimation is in estimating the entries of a large matrix from observed entries, which are noisy versions of entries in a small fraction of the entire matrix. We study the rate of convergence for estimation of matrices with a certain monotonicity property. It turns out to be equivalent to solving a robust version of the deconvolution problem. As the main result of this paper, we provide a simple, intuitive algorithm for matrix estimation which extends the works by Fan (1991) and Delaigle et al. (2008). We show that our computationally efficient method achieves near optimal minimax rate for the matrix estimation as well as robust deconvolution. This rate is within a constant factor to the rate achieved by the kernel deconvolution estimator in the classical setup.

Keywords: Deconvolution, Matrix estimation, Density estimation, Latent variable model, Minimax rate

1. Introduction

Deconvolution is a statistical inverse problem to estimate the unknown density f_X of a random variable X based on observations of random variable Z whose density takes the form $f_Z = T(f_X)$ for some transformation T. For example, let the observed random variable be $Z = X + N$, with N being independent, identically distributed noise; the density $f_Z = f_X * f_N$ with f_N being noise density and $*$ representing convolution. In this case, estimating f_X is effectively the process of deconvolution.

In a large body of such problems, including density deconvolution and errors-in-variables regression, the transformation T is commonly assumed to be known. In the simplest scenario, we have n independent observations of Z from which its density is estimated, thereby leading to estimation of $f_X = T^{-1}(f_Z)$ since T is known. Fan (1991) discussed how well the unknown density and its cumulative distribution function (CDF) can be estimated by nonparametric kernel methods with certain smoothness conditions imposed on the density f_X. In this celebrated work, they not only address how to estimate the density and compute the rate of convergence, but they also discuss how difficult the deconvolution problem is and how the difficulty depends on the noise characteristic. The work provides insights on the optimal rates of convergence and the best estimators in terms of the rates of convergence.

However, the noise density f_N and hence the transformation T may not be known a priori in many real-world applications. To overcome the challenge, it is often assumed that additional samples from replicated or validation data are available to estimate f_N. For example, samples of replicated contaminated data in the form of repeated measurements as in Delaigle et al. (2008), or sometimes direct samples from the error distribution are assumed available. Another line of works

have suggested to estimate the scale of the error distribution, but they require a particular parametric model for the noise density and even restrictive smoothness assumptions on the signal distribution in some cases.

In this paper, we consider a generalization of the deconvolution problem stated above that arises naturally in the context of matrix estimation. The problem of matrix estimation is as follows. We are given a partial observation of a data matrix $Z = [Z_{ij}] \in \mathbb{R}^{m \times n}$ which is generated as per the so-called *latent variable model*. Specifically, each row $i \in [m] = \{1, \ldots, m\}$ and column $j \in [n]$ are associated with latent parameters $\theta^{(i)}_{row}, \theta^{(j)}_{col} \in [0, 1]$ respectively. There is also a latent function $g : [0, 1] \times [0, 1] \to \mathbb{R}$. The random variables Z_{ij} are independent across i, j and are generated as $Z_{ij} = g(\theta^{(i)}_{row}, \theta^{(j)}_{col}) + N_{ij}$ where N_{ij} are independent, identically distributed noise random variables. The distribution of noise random variables is unknown. We observe each of Z_{ij} with probability $p \in (0, 1]$, independently. The goal is to recover the "mean" matrix $A = [A_{ij}]$ where $A_{ij} = \mathbb{E}[Z_{ij}] = g(\theta^{(i)}_{row}, \theta^{(j)}_{col})$. Ideally, we wish to retrieve a good estimate of A with as small p as possible.

Now consider row $i \in [m]$ of matrix A. Recovering it requires knowing $g(\theta^{(i)}_{row}, \cdot)$ where $\cdot \in \{\theta^{(j)}_{col},\ j \in [n]\}$. Now learning $g(\theta^{(i)}_{row}, \cdot),\ \cdot \in [0, 1]$ boils down to learning distribution of random variable $X^i = f(\theta^{(i)}_{row}, U)$, where U is uniform on $[0, 1]$. That is, matrix estimation problem is about learning m distributions, $X^i,\ i \in [m]$ simultaneously from their noisy samples. This is like the setup of Delaigle et al. (2008), but harder. Because, in the setup of Delaigle et al. (2008), we had *repeated* measurements while we have only a *single* measurement here. To articulate this, consider $m = 1$: it is impossible to learn the distribution corresponding to $X^1 = g(\theta^{(1)}_{row}, U)$ when the additive noise is unknown because of the lack of *repeated* measurements as required in Delaigle et al. (2008). For m large enough, as we shall show, even though above difficulty remains, we can utilize "commonality" between columns to create a "noisy version" of repeated measurements by looking across a row. And this requires a robust version of the method introduced in Delaigle et al. (2008) which is an important contribution of this work. Using this improved "collective deconvolution" method, we show that for the class of matrix estimation problem considered here, our efficient algorithm provides a minimax rate that is nearly optimal.

To enable "commonality" as mentioned above, we utilize the monotonicity property of the matrix. Precisely, we assume there exists a permutation of columns which leads to rearranging entries in all the rows in a monotone nondecreasing manner simultaneously. This assumption has similarity to the strong stochastic transitivity in rank aggregation (see Shah et al. (2016)) context and degree monotonicity in graphon estimation context; see Bickel and Chen (2009) and Chan and Airoldi (2014) for example. We note that our model is asymmetric unlike graphon which is symmetric. Due to limitation of space, further reviews on related works can be found in the Appendix.

1.1. Our Contributions

As the main contribution of this work, as noted earlier, we present a robust extension of the works by Fan (1991) and Delaigle et al. (2008) with the near optimal rate of convergence in terms of mean squared error. Ours is a neighborhood-based matrix completion method that operates with a very sparse data set. Technically, the refined use of concentration inequalities and chaining in the proofs can be interesting in its own right.

The key technical contribution is the noise density estimation algorithm described in Section 4 (see noise density estimation procedure and Algorithm 2 in Appendix E for more details) and

its analysis. It is aimed at imitating the setup of repeated measurements by detecting the columns having column features close to each other.

Our estimation algorithm (Algorithm 1) first estimates the column features for every column by taking average values, and then estimate the noise density using the estimated column features. The regularity assumption on the latent function with respect to the column features is used in this noise density estimation step. Thereafter, the latent function, or the inverse of the signal CDF, can be restored exploiting the estimated noise density. We also analyze the consistency and the rate of convergence of the proposed algorithm, which is summarized as Theorem 9 (see Theorem 3 for a simplified version). A full description of the algorithm and its analysis can be found in Appendix E.

The algorithmic upper bounds and the information-theoretic lower bounds for the rates of convergence under three different noise scenarios are summarized in Table 1.

Table 1: Mean Squared Error of function estimation depending on the noise models.

Noise Model	Algorithmic upper bound	Info-theoretic lower bound
Noiseless	$O\left(\frac{1}{(n-1)p}\right)$ Theorem 1	$\Omega\left(\frac{1-p}{(n-1)p}\right)$ Theorem 4
Supersmooth known distribution	$O\left(\left(\log np\right)^{-\frac{2}{\beta}}\right)$ Theorem 2	$\Omega\left((1-p)\left(\log(n-1)p\right)^{-\frac{3}{\beta}}\right)$ Theorem 5
Supersmooth unknown distribution	$O\left((\log np)^{-\frac{2}{\beta}}\right)$ Theorem 3	same as above

1.2. Organization

The paper is organized as follows. In Section 2, we state the problem of interest and our model assumptions. In Section 3, we present our main theoretical results, exhibiting the rates of convergence of our algorithm and its near optimality. We describe our proposed algorithm with a generic recipe and some details for the noisy scenario with unknown noise distribution in Section 4. We provide a sketch of the proof in Section 5, including core lemmas for analysis, however, the full details of the analysis and proof are deferred until Appendix E.

The lower bounds on MSE are stated in Theorem 4 and 5. The proof of these two theorems can be found in Appendix B. For comparison with easier noise scenarios, we discuss the algorithm and analysis adapted to noiseless setup (Appendix C) and to noisy setup when the noise distribution is known a priori (Appendix D).

2. Setup

2.1. Problem Statement

We wish to estimate matrix $A \in \mathbb{R}^{m\times n}$ from its partial, and possibly noisy observations $Z \in \mathbb{R}^{m\times n}$. Let $\mathcal{O} \subset [m]\times[n]$ denote the set of indices for which Z_{ij} is observed; Z_{ij} are such that $\mathbb{E}\left[Z(i,j)\right] = A(i,j)$. In this paper, we assume the additive noise model

$$Z(i,j) = A(i,j) + N(i,j), \quad \forall (i,j) \in \mathcal{O},$$

where $N(i,j)$ are independent and identically distributed random variable with zero mean: $\mathbb{E}[N(i,j)] = 0$. For $(i,j) \in [m] \times [n] \setminus \mathcal{O}$, $Z(i,j)$ is not observed, denoted as $Z(i,j) = \star$. We shall assume that each entry $(i,j) \in [m] \times [n]$ belongs to $\mathcal{O}$ with probability $p \in (0,1]$ independently.

We assume a nonparametric model for the matrix A: each row $i \in [m]$ and column $j \in [n]$ is associated with latent features $\theta^{(i)}_{row}, \theta^{(j)}_{col} \in [0,1] \subset \mathbb{R}$, and the (i,j)-th entry of matrix A takes the form

$$A(i,j) = g\left(\theta^{(i)}_{row}, \theta^{(j)}_{col}\right) \tag{1}$$

for some latent measurable function $g : [0,1]^2 \to \mathbb{R}$. However, this representation is not unique, because we can apply an invertible transform to the domain (latent feature space) and take the push-forward of the latent function with respect to the transform, so that $A(i,j)$ remains the same under the new representation. Therefore, estimation of the latent function g is an ill-posed problem, and we would rather focus on prediction of the values $A(i,j)$ for $(i,j) \in [m] \times [n]$.

Problem 1 *Given a data matrix $Z \in \mathbb{R}^{m \times n}$, can we recover the true parameter matrix $A \in \mathbb{R}^{m \times n}$ under the aforementioned setup in an algorithmically efficient manner?*

2.2. Performance Metric

Given an estimator $\varphi : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$, which returns the estimate $\hat{A} = \varphi(Z)$ of matrix A using Z, we use the mean-squared error (MSE) to evaluate the performance:

$$MSE(\varphi) = \mathbb{E}\left[\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\hat{A}(i,j) - A(i,j)\right)^2\right]. \tag{2}$$

We call the estimator φ to be consistent if MSE vanishes as the problem size (m,n) increases, i.e.

$$\lim_{m,n \to \infty} MSE(\varphi) = 0.$$

With these notations, the refined problem of interest is as follows.

Problem 2 *If consistent recovery in Problem 1 is possible for p large enough, how fast does the MSE converge to 0 as a function of p, m and n?*

2.3. Operating Model Assumptions

In addition to the assumptions for the additive noise model presented in Section 2.1, we assume some additional properties for the latent function g (see Eq. (1)) and the noise distribution.

2.3.1. ASSUMPTIONS ON THE LATENT FUNCTION

In addition to measurability, cetain types of smoothness conditions are usually imposed on the latent function, such as Lipschitz- or Hölder continuity. In this paper, we will focus on the class of functions $g : [0,1]^2 \to \mathbb{R}$, which are bounded, monotone increasing (Eq. (3)) and (l, L) bi-Lipschitz (Eq. (4)) with respect to the second argument. That is to say,

$$y_1 \le y_2 \implies g(x,y_1) \le g(x,y_2), \quad \forall x \in [0,1], \qquad \text{and} \tag{3}$$

$$\exists l, L > 0 \quad \text{s.t.} \quad 0 < l \le \frac{g(x,y_2) - g(x,y_1)}{y_2 - y_1} \le L < \infty, \quad \forall x, \forall y_1 \ne y_2. \tag{4}$$

However, we impose no further restrictions on g with regard to the first argument.

A bi-Lipschitz mapping is injective, and is a bijection onto its image. Therefore, for each $x \in [0,1]$, we can define the inverse of $g(x,\cdot) : [0,1] \to [g(x,0), g(x,1)]$, as $g_x^{-1} : [g(x,0), g(x,1)] \to [0,1]$. It is easy to check that g^{-1} is also monotone increasing and $(\frac{1}{L}, \frac{1}{l})$ bi-Lipschitz.

2.3.2. ASSUMPTIONS ON THE NOISE

We assume noise is symmetric with mean zero, and sub-Gaussian with parameter σ, i.e., $\mathbb{E}\left[e^{tX}\right] \leq e^{\frac{t^2\sigma^2}{2}}$, $\forall t \in \mathbb{R}$. In addition, we assume the noise is supersmooth (see Appendix L.2.1, cf. Fan (1991); Delaigle et al. (2008) for more detail), i.e., there exist $B > 1$, and $\beta, \gamma > 0$ such that

$$B^{-1} \exp\left(-\gamma|t|^\beta\right) \leq \phi_N(t) \leq B \exp\left(-\gamma|t|^\beta\right), \qquad \forall t \in \mathbb{R}, \tag{5}$$

where $\phi_N(t)$ is the characteristic function of the noise distribution. For example, Gaussian noise is a typical example of super-smooth noise with parameter $\beta = 2$. As the name suggests, supersmooth noise is smoother than the class of 'ordinary-smooth' noise (cf. Fan (1991) for definition), which has polynomially decaying tail in the Fourier domain.

2.4. Recapping the Model

For a succinct representation of the model introduced so far, we introduce three matrices of the same size, $A, N, M \in \mathbb{R}^{m \times n}$. Specifically, A is the matrix which we would like to estimate. N is a random matrix of size (m, n), whose entries are drawn i.i.d. as per a noise distribution. M is a random binary masking matrix with each entry being 1 with probability p and 0 with probability $1-p$, independently. The observation matrix Z is such that $Z(i,j) = A(i,j) + N(i,j)$ if $M(i,j) = 1$, and $Z(i,j) = \star$ if $M(i,j) = 0$ regardless of the value of $A(i,j) + N(i,j)$.

3. Main Results

We present main results of our work by answering Problems 1 and 2 respectively. We provide simple estimation algorithms that require robust deconvolution method. The convergence rate for MSE under these algorithms are contrasted with lower bound results which primarily follow from the classical literature in function approximation and deconvolution.

3.1. Algorithmic Upper Bounds on MSE

We build up towards our main result by considering increasing order of difficulty in terms of assumption on noise model: (1) *Noiseless*: $N(i,j) = 0$ for all $(i,j) \in [m] \times [n]$; (2) *Known noise*: the noise distribution is known; and (3) *Unknown noise*: the noise distribution is unknown and has to be also estimated. Again, the main result is the scenario (3) with unknown noise, however, the other two cases help in building solution up and are presented for completeness. The following three main theorems explicitly state upper bounds on the MSE rate for each noise scenario, which turn out to be (near-) optimal in comparison with Theorems 4 and 5. We present the theorems in the language of matrix estimation, however, the algorithm proposed in Section 4 essentially recovers the underlying latent function, namely, graphon.

Theorem 1 (Informal; noiseless) *In the noiseless scenario, there is a polynomial time algorithm* $\breve{\varphi} : Z \mapsto \hat{A}$*, which consistently estimates* A *from a data matrix* Z *with* $MSE\left(\breve{\varphi}\right) = O\left(\frac{1}{(n-1)p}\right)$.

Theorem 2 (Informal; known noise) *In the known noise scenario, there is a polynomial time algorithm* $\tilde{\varphi} : Z \mapsto \hat{A}$*, which consistently estimates* A *from a data matrix* Z *with* $MSE\left(\tilde{\varphi}\right) = O\left((\log np)^{-\frac{2}{\beta}}\right)$.

Theorem 3 (Informal; unknown noise) *In the unknown noise scenario, there is a polynomial time algorithm* $\hat{\varphi} : Z \mapsto \hat{A}$*, which consistently estimates* A *from a data matrix* Z *with* $MSE\left(\hat{\varphi}\right) = O\left((\log np)^{-\frac{2}{\beta}}\right)$.

The full statements and the proofs of these theorems can be found in Appendices C, D, and E, respectively with corresponding adaptations of the estimation algorithm and their analysis. In a nutshell, the proposed algorithm consists of a two separate procedures of estimating the column features (quantiles) of all columns and then estimating the latent function (the inverse of signal CDF) for all rows. We show our proposed algorithm achieves the (near-) optimal rate of MSE in all three noise scenarios.

We remark that the MSE converges to 0 as $m, n \to \infty$ as long as $p = \omega\left(n^{-1}\right)$, regardless of the noise assumption. Even when there is nontrivial noise, our algorithm attains a vanishing MSE upper bound as long as $p = \omega(\max\{m^{-1}, n^{-1}\})$. This provides a positive answer to Problem 1.

However, answering to Problem 2, we require a technical condition for the aspect ratio between m and n when there is nontrivial noise. It is necessary to have $(\log np)^{\frac{2}{\beta}} \ll mp \ll n$ to achieve the MSE upper bound as described in the Theorems 2 and 3. This condition stems from our analysis; our proposed algorithm does not require it. The condition ensures that the error in function estimation dominates the error in column feature estimation in the noisy scenarios. Note that this condition is easily satisfied in most setups, and that there is no such restriction in the noiseless scenario.

3.2. Information-theoretic Lower Bounds on MSE

In order to argue the lower bound on the MSE rate for any estimation procedure, we show there exists a pair of latent functions, which are not possible to distinguish beyond certain resolution (the lower bound) by any algorithm from given data. Specifically, we show that for any given data $\theta^{(i)}_{row}$, $\theta^{(j)}_{col}$, $Z(i,j)$, there exist two functions g and $g^\dagger$ which would generate identical data at the sampling points, yet are significantly different. Suppose that there is an oracle algorithm φ^* which has access not only to $Z(i,j)$ but also to $\theta^{(i)}_{row}$, $\theta^{(j)}_{col}$. However, since $g\left(\theta^{(i)}_{row}, \theta^{(j)}_{col}\right) = g^\dagger\left(\theta^{(i)}_{row}, \theta^{(j)}_{col}\right)$ for all (i,j) such that $M(i,j) = 1$, even an oracle cannot tell if the data is generated as per either g or $g^\dagger$ based on the given data. No algorithm can outperform the oracle, and therefore, the MSE cannot be smaller than the squared L^2 distance between g and $g^\dagger$. The details of the argument are provided in Appendix B.

Theorem 4 (Informal; noiseless) *In the noiseless scenario, for any estimation algorithm* φ*, there exists a hard instance for which* $MSE(\varphi) = \Omega\left(\frac{1-p}{(n-1)p}\right)$.

Theorem 5 (Informal; additive noise) *In the additive noise scenario, for any estimation algorithm* φ*, there exists a hard instance for which* $MSE(\varphi) = \Omega\left((1-p)\big(\log(n-1)p\big)^{-3/\beta}\right)$.

4. Algorithm

4.1. Generic Recipe

We shall use a "generic" recipe for estimation in all three scenarios considered in this work: noiseless, noisy with known noise distribution, and noisy with unknown noise distribution. The generic algorithm is adapted for each setup to deal with the effect of the noise. Due to limitation of space, we shall provide details in Section 4.2 for the scenario when the noise distribution is unknown, which is the most challenging case. However, details for all the scenarios as well as accompanying analysis for each setup can be found in the Appendix C (noiseless); D (known noise) and E (unknown noise). The generic algorithm for each of the three scenarios is as follows:

Algorithm 1: Generic recipe of the algorithm

1. Estimate the latent feature (=quantile) $\theta^{(j)}_{col}$ of column j. Let it be denoted by $\hat{q}(j)\ \ j \in [n]$.
2. Estimate $F^{(i)} = g^{-1}_{x=\theta^{(i)}_{row}}$ on row i, which is the inverse of the latent function $g\left(\theta^{(i)}_{row}, \cdot\right)$ restricted on the first coordinate. Let it be denoted by $\hat{F}^{(i)},\ i \in [m]$.
3. Plug in the estimates: $\hat{A}(i,j) = \hat{g}^{(i)}(\hat{q}(j)),\ i \in [m],\ j \in [n]$, where $\hat{g}^{(i)} = \left(\hat{F}^{(i)}\right)^{-1}$.

We note that, by assumption, for any given $x \in [0,1]$ the latent function $g(x,\cdot) : [0,1] \to \mathbb{R}$ along the second dimension is continuous and monotone increasing in our model, and hence invertible. The inverse of g (for a fixed x), namely, $g^{-1}(x,\cdot) : \mathbb{R} \to [0,1]$, can be viewed as a cumulative distribution function for a certain distribution on $\mathbb{R}$. In short, for each row $i \in [m]$, we can consider the latent function restricted to $x = \theta^{(i)}_{row}$, i.e., $g(x,\cdot)$, as the inverse of the cumulative distribution function of signal along row i. The estimation of $F^{(i)}$ changes depending upon whether it is noiseless or noisy with known / unknown noise distribution.

4.2. Details

We describe details of the steps outlined in the generic algorithm above for the most challenging scenario with unknown noise distribution. We will be brief here due to space limitation. However, further details can be found in the Appendix E.

4.2.1. SOME NOTATIONS

For $i \in [m],\ j \in [n]$, let

$$\mathcal{B}_i = \{j' \in [n] : M(i,j') = 1\} \text{ and } \mathcal{B}^j = \{i' \in [m] : M(i',j) = 1\}. \tag{6}$$

Define Heaviside step function $H : \mathbb{R} \to \left\{0, \frac{1}{2}, 1\right\}$ using the indicator function $\mathbb{I}$ as $H(x) = \frac{1}{2}\big(\mathbb{I}\{x > 0\} + \mathbb{I}\{x \geq 0\}\big)$. That is, $\sum_{j_2=1}^{n} H\big(Z(i,j_1) - Z(i,j_2)\big)$ is the number of entries $Z(i,j)$ in row i whose value smaller than $Z(i,j_1)$ while $Z(i,j_1)$ (and indices with value equal to it) being counted with weight $\frac{1}{2}$.

4.2.2. STEP 1: ESTIMATING $\theta_{col}^{(j)}$ BY $\hat{q}_{marg}(j),\ j \in [n]$.

Given $Z \in \mathbb{R}^{m \times n}$, define Z_{marg} as the column average of observed data. That is, $Z_{\text{marg}}(j) = \frac{\sum_{i=1}^{m} M(i,j)Z(i,j)}{\sum_{i=1}^{m} M(i,j)}$ if $\mathcal{B}^n \neq \emptyset$. If $\mathcal{B}^j = \emptyset$, we let $Z_{\text{marg}}(j) = \frac{1}{2}$ by default. Then, for $j \in [n]$ let

$$\hat{q}_{\text{marg}}(j) = \frac{1}{n} \sum_{j'=1}^{n} H\left(Z_{\text{marg}}(j) - Z_{\text{marg}}(j')\right). \tag{7}$$

4.2.3. STEP 2: ESTIMATING $F^{(i)} = g^{-1}_{x=\theta_{row}^{(i)}}$ BY $\hat{F}^{(i)},\ i \in [m]$.

Each entry in the row i can be viewed as a sum of two independent random variables: the first random variable is $g(\theta_{row}^{(i)}, \theta_{col}^{(j)})$ with the randomness induced due to that in the column parameter $\theta_{col}^{(j)}$ that are sampled uniformly from $[0, 1]$; the second random variable is the additive noise. Therefore, the empirical CDF of the observations gives good estimation of distribution of the summation of these two random variables. However, the interest is in recovering the distribution of the first random variable. If we do know the distribution of the second random variable, we can deconvolute the effect of noise by deconvolution kernel estimator.

Putting it other way, we wish to recover distribution of random variable X, but we observe samples of $Z = X + N$ instead of X. And we do not know the distribution of N. Due to independence, we know that $\phi_Z(t) = \phi_X(t)\phi_N(t)$ for all $t \in \mathbb{R}$, where ϕ_Z, ϕ_X, ϕ_N denote the characteristic function of random variable Z, X and N respectively. To overcome the challenge of unknown noise distribution, we estimate the noise characteristic function first and then estimate the CDF using kernel deconvolution, but with an additional ridge parameter to avoid division by zero.

Indeed, this is known as deconvolution kernel density estimator in literature. We shall adopt prior results Carroll and Hall (1988); Fan (1991); Delaigle et al. (2008) to our setting. In particular, in the prior setting, to estimate noise distribution, it is assumed that for a given *fixed* instance of X, we have multiple noisy observations, e.g. $X + N_1, \ldots, X + N_k$ with k large enough. In our setting, it is effectively *one* sample per instance of X. So it is not straightforward to estimate noise distribution. We overcome this challenge as follows (further details can be found in Appendix L).

Noise Density Estimation. We shall explain how to produce estimation $\hat{\phi}_N(t)$ for noise distribution using pairs of observations from rows $i \in [m]$. To begin with, suppose that we can repeatedly observe the same instance X_i of target random variable up to independent additive noise, i.e., $Z_{ij} = X_i + N_{ij}$ with N_{ij} independent. Although we don't know the value of X_i, we can see that the difference in the observed data entries is equal to the difference between two independent noise instances: $Z_{i1} - Z_{i2} = (X_i + N_{i1}) - (X_i + N_{i2}) = N_{i1} - N_{i2}$. Assuming symmetry in the noise distribution, $N_{i1} - N_{i2} \equiv N_{i1} + N_{i2}$. Therefore, $\phi_{N_{i1}-N_{i2}}(t) = \phi_N(t)^2$. From symmetry of N, we know that $\phi_N(t)$ is real-valued. Moreover, it is positive because N is supersmooth. Therefore, we can estimate $\phi_N(t)$ by taking square root of the (the absolute value of) estimate $\hat{\phi}_{N_1-N_2}(t)$ as

$$\hat{\phi}_N(t) = \hat{\phi}_{N_1-N_2}(t)^{\frac{1}{2}} = \left| \frac{1}{n} \sum_{i=1}^{n} \cos\left[t\left(N_{i1} - N_{i2}\right)\right] \right|^{\frac{1}{2}}.$$

However, the repeated measurement assumption is not feasible because we have *at most* one measurement for a given index (i, j). Despite this challenge, we can still hope to obtain *almost* repeated

samples from observations in a given row, if we choose columns $j_1, j_2 \in [n]$ that have *very* similar features $\theta_{col}^{(j_1)} \approx \theta_{col}^{(j_2)}$ so that

$$Z(i,j_1) - Z(i,j_2) = \underbrace{[A(i,j_1) - A(i,j_2)]}_{\approx 0,\ \because \theta_{col}^{(j_1)} \approx \theta_{col}^{(j_2)}} + [N(i,j_1) - N(i,j_2)] \approx N(i,j_1) - N(i,j_2).$$

This intuition leads to the following procedure. For each $i \in [m]$, we produce different estimates $\hat{\phi}_N$, namely, $\hat{\phi}_{N,i}$ using data only from the rows $i' \in [m] \setminus i$.

1. Let $\mathcal{T} := \big\{(i, j_1, j_2) \in [m] \times [n]^2 : M(i,j_1) = M(i,j_2) = 1 \text{ and } \hat{q}_{marg}(j_1) \approx \hat{q}_{marg}(j_2)\big\}$.
2. For $i \in [m]$, define $\mathcal{T}_i$ as $\mathcal{T}_i := \Big\{(i', j_1, j_2) \in \mathcal{T} : i' \neq i\Big\}$.
3. For $i \in [m]$, estimate $\hat{\phi}_{N,i}(t) = \left|\frac{1}{|\mathcal{T}_i|} \sum_{(i,j_1,j_2)\in\mathcal{T}_i} \cos\Big[t\left(Z(i,j_1) - Z(i,j_2)\right)\Big]\right|^{1/2}$.

Intuitively, $\mathcal{T}$ is the set of index triples to imitate the repeated measurements: Algorithm 2 in Appendix E for its construction. For each row i, we estimate the noise characteristic function $\hat{\phi}_{N,i}$ by using $\mathcal{T}_i$, which is a subset of $\mathcal{T}$ tailored to exclude the data from row i.

Estimating $\hat{F}^{(i)}$. Recall $\mathcal{B}_i = \{j \in [n] : M(i,j) = 1\}$ (see Eq. (6)). We define the kernel smoothed CDF estimator with unknown noise density as follows. Given constants D_1, D_2 such that $D_1 \leq \inf_{x,y\in[0,1]} g(x,y)$ and $D_2 \geq \sup_{x,y\in[0,1]} g(x,y)$,

$$\hat{F}^{(i)}(z) = \begin{cases} \int_{D_1}^{z} \hat{f}^{(i)}(w)dw, & \text{if } z < D_2, \\ 1, & \text{if } z \geq D_2, \end{cases} \tag{8}$$

where

$$\hat{f}^{(i)}(z) = \frac{1}{h|\mathcal{B}_i|} \sum_{j\in\mathcal{B}_i} \hat{L}\left(\frac{z - Z(i,j)}{h}\right) \qquad \text{and} \qquad \hat{L}(z) = \frac{1}{2\pi} \int e^{-\mathbf{i}tz} \frac{\phi_K(t)}{\hat{\phi}_{N,i}\left(\frac{t}{h}\right) + \rho} dt.$$

The kernel bandwidth parameter $h = (4\gamma)^{\frac{1}{\beta}} \left(\log|\mathcal{B}_i|\right)^{-\frac{1}{\beta}}$ where β and γ are smoothness parameters for the noise (see Eq. (5)). We choose the ridge parameter $\rho = |\mathcal{B}_i|^{-7/24}$. We choose a kernel K satisfying the following conditions: (i) K is symmetric, i.e., $K(x) = K(-x)$ for all $x \in \mathbb{R}$; and (ii) ϕ_K is supported within $[-1, 1]$. More details can be found in Remark 19 in Appendix D.

4.2.4. STEP 3: ESTIMATING $A(i,j)$ BY $\hat{A}(i,j)$, $i \in [m], j \in [n]$.

For each $i \in [m]$, let $\hat{g}^{(i)} = \left(\hat{F}^{(i)}\right)^{-1}$ denote the quantile function (right pseudo-inverse) associated with $\hat{F}^{(i)}$. Plugging Eq. (7) into it leads to the estimate of matrix entry:

$$\hat{A}(i,j) = \hat{g}^{(i)}\left(\hat{q}_{marg}(j)\right). \tag{9}$$

5. Sketch of the Proof of Theorem 3

Here we provide a sketch of the proof of main Theorem 3. Details can be found in Appendix E. The key to establishing this result is arguing that each of the three steps of the algorithm detailed in Section 4.2 succeeds. This is what we do next.

5.1. Step 1 Works

Recall D_1, D_2 are some constants such that $D_1 \leq \inf_{x,y\in[0,1]} g(x,y)$ and $D_2 \geq \sup_{x,y\in[0,1]} g(x,y)$. We define two other constants $C_1 \equiv \frac{l^2}{2(D_2-D_1)^2}$ and $C_2 \equiv \frac{l^2}{8\sigma^2}$, which depend on model parameters l, σ. We define a threshold for quantile estimation $t_q^* \equiv \frac{4\sqrt{\pi}}{\sqrt{mp}}\left(\frac{\sqrt{e^{C_1}}+\sqrt{2}}{\sqrt{C_1}} + \frac{\sqrt{e^{C_2}}+\sqrt{2}}{\sqrt{C_2}}\right)$. Next we establish that the quantile estimates concentrate around the true column features.

Lemma 6 *For any* $t \geq 2t_q^* = \Theta\left(\frac{1}{\sqrt{mp}}\right)$,

$$\mathbb{P}\left(\left|\hat{q}_{marg}(j) - \theta_{col}^{(j)}\right| > t\right) \leq \exp\left(-\frac{n}{6}\left(t - t_q^*\right)\right) + \exp\left(-\frac{nt^2}{2}\right) + \exp\left(-\frac{mp}{8}\right).$$

Proof [Sketch] Consider an ideal estimator $\hat{q}_*(j) = \frac{1}{n}\sum_{j'=1}^{n} H\left(\theta_{col}^{(j)} - \theta_{col}^{(j')}\right)$, which has access to the hidden column features. Now

$$\left|\hat{q}_{\text{marg}}(j) - \theta_{col}^{(j)}\right| \leq \left|\hat{q}_{\text{marg}}(j) - \hat{q}_*(j)\right| + \left|\hat{q}_*(j) - \theta_{col}^{(j)}\right|.$$

Due to uniform distribution on $[0,1]$ for column parameters, one expects

$$\left|\hat{q}_*(j) - \theta_{col}^{(j)}\right| \approx \Theta(\frac{1}{\sqrt{n}}).$$

Therefore, to obtain error bound of $t \geq t_q^*$ (we assume $mp \ll n$), it boils down to controlling $\left|\hat{q}_{\text{marg}}(j) - \hat{q}_*(j)\right|$. We obtain a probabilistic tail upper bound for $|\hat{q}_{marg}(j) - \hat{q}_*(j)|$ by rewriting it as the sum of indicator which is dominated by a certain binomial random variable. This leads to the desired claimed bound. Please see Appendix G for details. ■

5.2. Step 2 Works

We define thresholds t_0^* and T_0^* relevant for CDF estimation in AppendixE.2.2 (cf. Eqs. (57), (58)). In effect, they are such that $t_0^* = O\left(\left(\log np\right)^{-1/\beta}\right)$ and $T_0^* = t_0^* + C\frac{\left(\log 2np\right)^{1/\beta}}{(np)^{5/24}}$ for some constant C. We define an event which we shall show to hold with high-probability as follows: for $i \in [m]$,

$$E_{(i)} \equiv \left\{\frac{np}{2} \leq |\mathcal{B}_i| \leq 2np\right\}.$$

Finally, we take a note of "remainder term" $\tilde{\Psi}_{m,n,p}$, defined precisely in Eq. (59) which turns out to be $o(1)$ with scaling of m, n, p. Now we state the main result about Step 2 of the algorithm working.

Lemma 7 *For any* $i \in [m]$, *and for any* $t \geq T_0^*$,

$$\mathbb{P}\left(\sup_{z\in[D_1,D_2]}\left|\tilde{F}^{(i)}(z) - F^{(i)}(z)\right| > t \,\middle|\, E_{(i)}\right) \leq (2np)^{\frac{1}{6}}\exp\left(\frac{-\left(\frac{np}{2}\right)^{5/12}(t-t_0^*)^2}{8C_4^2\left(\log(2np)\right)^{\frac{2}{\beta}}}\right) + \tilde{\Psi}_{m,n,p}.$$

The constant $C_4 = \frac{BK_{max}(D_2-D_1)}{\pi(4\gamma)^{\frac{1}{\beta}}}$, *where* $B \geq 1$ *is a noise model parameter (see Eq. (5)) and* $K_{max} = \sup_t |\phi_K(t)|$.

Proof [Sketch] Details can be found in Appendix I, here we provide a very succinct summary. The main idea of the proof is to decompose the desired probability into three pieces using triangle inequality and the union bound: (1) the variance of $\hat{F}^{(i)}(z)$ (Eq. (99)); (2) the bias of the CDF estimator in the known noise setup (Eq. (100)); and (3) the discrepancy in the estimator between the known noise and the unknown noise scenarios, i.e., $\mathbb{E}\left[\hat{F}^{(i)}(z)\right] - \mathbb{E}\left[\tilde{F}^{(i)}(z)\right]$ (letting $\tilde{F}$ denote the estimator with known ϕ_N) (Eq. (101)). Since $[D_1, D_2]$ is a compact set, we can obtain the desired bound by chaining technique whenever the supremum over $[D_1, D_2]$ is considered.

Controlling (1) is accomplished by applying McDiarmid's inequality and the result is stated as Lemma 44. There is an upper bound for (2), which is stated in Lemma 29 and its proof is based on the upper bound result by Fan (1991).

The most challenging aspect of this Lemma (and of this paper) is to establish that (3) is well-behaved. This requires us to identify a set of events that hold with high-enough-probability, and conditioned on those events, the desired bound holds. The set of events are listed in Appendix I.6. A sequence of Lemmas in the first three subsections of I precisely argue what the above statement claims. All in all, when the dust settles, we obtain the desired claim of this Lemma. ■

5.3. Step 3 Works

Using Lemmas 6 and 7, we establish a probabilistic tail bound on $|\hat{A}(i,j) - A(i,j)|$ and then integrate it to obtain a bound on Mean-Squared-Error (MSE).

5.3.1. PROBABILISTIC TAIL BOUND

For given choice of parameters $t > 0$ and L, m, n, p, t_q^*, T_0^*, we define two conditions:

$$E_1 = \left\{ t \le 4Lt_q^* \right\} \qquad \text{and} \qquad E_2 = \left\{ t \le 2LT_0^* \right\}. \tag{10}$$

Theorem 8 *For each $(i,j) \in [m] \times [n]$, for any $t \ge 0$,*

$$\begin{aligned} \mathbb{P}\left(\left| \hat{A}(i,j) - A(i,j) \right| > t \right) \le \exp\left(-\frac{n(t - 2Lt_q^*)}{12L} \right) \mathbb{I}\{E_1^c\} \\ + (2np)^{\frac{1}{6}} \exp\left(\frac{-\left(\frac{np}{2}\right)^{5/12}}{8C_4^2 \left(\log(2np)\right)^{\frac{2}{\beta}}} (t - t_0^*)^2 \right) \mathbb{I}\{E_2^c\} \\ + \exp\left(-\frac{nt^2}{8L^2} \right) + \mathbb{I}\{E_1\} + \mathbb{I}\{E_2\} + \Psi_{m,n,p}, \end{aligned} \tag{11}$$

where t_0^*, T_0^* and $\Psi_{m,n,p}$ are some functions of m, n, p, which do not depend on t.

In above, $\Psi_{m,n,p}$ is defined in (63). It can be seen that the terms in the last line of (11) decay to 0 at an exponential rate as $\min(m,n)p \to \infty$, independent of t.

Proof Let $\theta^* \equiv F^{(i)}\left(\hat{A}(i,j)\right) = F^{(i)}\left(\hat{g}^{(i)}\left(\hat{q}_{marg}(j)\right)\right)$. Now $|\theta^* - \hat{q}_{marg}(j)| \le \left\| \hat{F}^{(i)} - F^{(i)} \right\|_\infty$ because $\hat{F}^{(i)}$ is continuous. Since $\hat{A}(i,j) = \hat{g}^{(i)}\left(\hat{q}_{marg}(j)\right) = g\left(\theta_{row}^{(i)}, \theta^*\right)$, and g is (l,L)-biLipschitz,

$$\left| \hat{A}(u,i) - A(i,j) \right| = \left| g\left(\theta_{row}^{(i)}, \theta_{col}^{(j)}\right) - g\left(\theta_{row}^{(i)}, \theta^*\right) \right| \le L\left(\left| \theta_{col}^{(j)} - \hat{q}_{marg}(j) \right| + \left\| \hat{F}^{(i)} - F^{(i)} \right\|_\infty \right).$$

If $\left|\theta_{col}^{(j)} - \hat{q}_{marg}(j)\right| \leq \frac{t}{2L}$, $\left\|\hat{F}^{(i)} - F^{(i)}\right\|_\infty \leq \frac{t}{2L}$ then $\left|\hat{A}(u,i) - A(i,j)\right| \leq t$. Therefore

$$\mathbb{P}\left(\left|\hat{A}(i,j) - A(i,j)\right| > t\right)$$
$$\leq \mathbb{P}\left(\left|\hat{q}_{marg}(j) - \theta_{col}^{(j)}\right| > \frac{t}{2L}\right) + \mathbb{P}\left(\sup_{z\in\mathbb{R}}\left|\hat{F}^{(i)}(z) - F^{(i)}(z)\right| > \frac{t}{2L}\middle| E_{(i)}\right) + \mathbb{P}\left(E_{(i)}^c\right)$$

by applying the union bound. Now, we can conclude the proof by applying Lemmas 6 and 7. ■

5.3.2. MEAN SQUARED ERROR

Let $\hat{\varphi}$ denote the estimator which maps Z to $\hat{A}$. The mean squared error of estimator $\hat{\varphi}$ is given as

$$MSE\left(\hat{\varphi}\right) = \int_0^\infty 2u\mathbb{P}\left(\left|\hat{A}(i,j) - A(i,j)\right| > u\right) du. \tag{12}$$

Define

$$c(n,p) \equiv \frac{\left(\frac{np}{2}\right)^{5/12}}{8C_4^2\left(\log(2np)\right)^{\frac{2}{\beta}}}.$$

Theorem 9 (The Full Version of Main theorem 3) *The mean squared error of the deconvolution kernel estimator $\hat{\varphi}$ is bounded above as follows:*

$$MSE\left(\hat{\varphi}\right) \leq 4L^2T_0^{*2} + (4Lt_q^*)^2 + 4Lt_q^*\sqrt{\frac{3L\pi}{n}}$$
$$+ 4L^2(2np)^{\frac{1}{6}}\left[\frac{1}{c(n,p)} + t_0^*\sqrt{\frac{\pi}{c(n,p)}}\right] + \frac{8L^2}{n} + \frac{288L^2}{n^2} + \Psi_{m,n,p}\Big(D_2 - D_1\Big)^2.$$

We remark that $4L^2T_0^{*2}$ is the dominant term, which scales as $O\left((\log np)^{-\frac{2}{\beta}}\right)$ (see Eq. (56) for definition of T_0^*). As a result, the upper bound diminishes to 0 at the rate of $(\log np)^{-\frac{2}{\beta}}$ as $mp, np \to \infty$.

6. Discussion

We end this paper with two remarks. First, there is an exponential gap in the mean squared error between the noiseless setup and noisy setup where measurements are corrupted by super-smooth additive noise. The gap is natural because recovery from noisy measurements should be more difficult, but it is surprising to observe an exponential gap. We note that the exponential degradation stems from the super-smooth assumption on the noise, and we strongly believe it is possible to obtain a similar result with only a polynomial gap when the noise is ordinary smooth (i.e., the noise characteristic function has a polynomially decaying tail).

Second, it is noteworthy that we do not have column features available at our hand, unlike the setup in those existing literature. However, we are still able to evaluate our estimated function at unknown points to reconstruct the matrix and the asymptotically optimal rate is achieved. This was possible because we are not estimating a single function, but collectively estimating a set of functions and a kind of collaboration is happening between the functions. If the column features (or extrinsic covariates) need not be estimated but are available from other sources, our task truly reduces to learning row-wise distributions and we obtain the same bounds.

References

Edo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.

E.M. Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981 – 2014, 2008.

P.J. Bickel and A. Chen. A nonparametric view of network models and newman-girvan and other modularities. In *Proc. Natl. Acad. Sci.*, pages 21068 – 21073, 2009.

P.J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *Ann. Statist.*, 39:2280 – 2301, 2011.

Paul P Biemer, Robert M Groves, Lars E Lyberg, Nancy A Mathiowetz, and Seymour Sudman. *Measurement errors in surveys*, volume 173. John Wiley & Sons, 2011.

J Martin Bland and DouglasG Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476):307–310, 1986.

Ralph A. Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4), 1952.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Raymond J. Carroll and Peter Hall. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186, 1988.

Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. CRC press, 2006.

Stanley Chan and Edoardo Airoldi. A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning*, pages 208–216, 2014.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

Sourav Chatterjee, Persi Diaconis, and Allan Sly. Random graphs with a given degree sequence. *The Annals of Applied Probability*, pages 1400–1435, 2011.

Alexander P. Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 1979.

Aurore Delaigle, Peter Hall, and Alexander Meister. On deconvolution with repeated measurements. *The Annals of Statistics*, pages 665–685, 2008.

Luc Devroye. Consistent deconvolution in density estimation. *Canadian Journal of Statistics*, 17 (2):235–239, 1989.

Peter J Diggle and Peter Hall. A fourier approach to nonparametric deconvolution of a density estimate. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 523–531, 1993.

Graham Dunn. *Design and analysis of reliability studies: The statistical evaluation of measurement errors*. Edward Arnold Publishers, 1989.

Graham Dunn Dunn. *Statistical evaluation of measurement errors: Design and analysis of reliability studies*. John Wiley & Sons, 2009.

Michael Eliasziw, S Lorraine Young, M Gail Woodbury, and Karen Fryday-Field. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. *Physical therapy*, 74(8):777–788, 1994.

Jianqing Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pages 1257–1272, 1991.

Jianqing Fan. Adaptively local one-dimensional subproblems with application to a deconvolution problem. *The Annals of Statistics*, pages 600–610, 1993.

Ravi Sastry Ganti, Laura Balzano, and Rebecca Willett. Matrix completion under monotonic single index models. In *Advances in Neural Information Processing Systems*, pages 1864–1872, 2015.

Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

John L Jaech. *Statistical analysis of measurement errors*, volume 2. John Wiley & Sons Incorporated, 1985.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the 45th annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

RH Keshavan, A Montanari, and S Oh. Matrix completion from a few entries. *IEEE Trans. Inf. Theory*, 56(6), 2009.

Ashish Khetan and Sewoong Oh. Reliable crowdsourcing under the generalized dawid-skene model. *arXiv preprint arXiv:1602.03481)*, 2016.

Olga Klopp, Alexandre B Tsybakov, Nicolas Verzelen, et al. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316–354, 2017.

S. N. Kudryavtsev. Recovering a function with its derivatives from function values at a given number of points. *Russian Academy of Sciences Izvestiya Mathematics*, 45(3)::505?528, 1991.

C.E. Lee, Y. Li, D. Shah, and Song D. Blind regression: Nonparametric regression for latent variable models via collaborative filtering. In *Advances in Neural Information Processing Systems*, pages 2155–2163, 2016.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley and Sons, 1959.

John Mendelsohn and John Rice. Deconvolution of microfluorometric histograms with b splines. *Journal of the American Statistical Association*, 77(380):748–753, 1982.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

Michael H Neumann and O Hössjer. On the effect of estimating the error density in nonparametric deconvolution. *Journal of Nonparametric Statistics*, 7(4):307–330, 1997.

Angelika Rohde, Alexandre B Tsybakov, et al. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

Karl Rohe, Sourav Chatterjee, Bin Yu, et al. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Nihar B. Shah, Sivaraman Balakrishnan, Adityanand Guntuboyina, and Martin J. Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. In *International Conference on Machine Learning*, pages 11–20, 2016.

Nathan Srebro, Noga Alon, and Tommi S Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In *Advances In Neural Information Processing Systems*, pages 1321–1328, 2004.

Leonard A Stefanski. Rates of convergence of some estimators in a class of deconvolution problems. *Statistics & Probability Letters*, 9(3):229–235, 1990.

Leonard A. Stefanski and Raymond J. Carroll. Deconvolving kernel density estimators. *Statistics*, 21(2):169–184, 1990.

Louis L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4), 1927.

Matt P Wand and M Chris Jones. *Kernel smoothing*. Crc Press, 1994.

Patrick J Wolfe and Sofia C Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.

Jiaming Xu. Rates of convergence of spectral methods for graphon estimation. *arXiv preprint arXiv:1709.03183*, 2017.

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Estimating network edge probabilities by neighborhood smoothing. *arXiv preprint arXiv:1509.08588*, 2015.

Dengyong Zhou, Qiang Liu, John C. Platt, Christopher Meek, and Nihar B. Shah. Regularized minimax conditional entropy for crowdsourcing. *arXiv preprint arXiv:1503.07240)*, 2015.

Appendix A. Related Works

Early works on the problem of density estimation under the assumption of known measurement error distribution focused on addressing how to estimate the unknown density and compute the rates of convergence of the methods for specific error distributions. These early works include Carroll and Hall (1988), Devroye (1989), Fan (1993), Mendelsohn and Rice (1982), Stefanski and Carroll (1990), Stefanski (1990). Among those vast amount of literature, Fan (1991) discusses how difficulty of the deconvolution problem depend on the dispersion of the noise by introducing the notion of supersmooth and ordinary smooth noise, thereby providing insights on the nonparametric deconvolution.

Subsequently, the problem of density estimation with unknown error density, which is also estimated from samples of the error itself, has been considered; see Diggle and Hall (1993) and Neumann and Hössjer (1997), for example. In particular, the setup where there are replicated measurements for each inherently different samples–with errors being independent and the intrinsic signal of the observations being the same among repeated measurements–drew much attention. For example, Jaech (1985) described an experimental setup where the uranium concentration is repeatedly measured for several fuel pellets; Biemer et al. (2011) discusses repeated observations in a social science context, e.g., in surveys. There are also a plenty of works under the setup on medical and clinical research, for example, Bland and Altman (1986) on lung function, Dunn (1989) on a brain-related study, Eliasziw et al. (1994) on physiotherapy for the knee, etc. Further medical examples can be found in Carroll et al. (2006) and Dunn (2009).

Delaigle et al. (2008) argues that even in such a setting of unknown error density with repeated measurements, a modified kernel deconvolution estimator using the estimated error density and a ridge parameter to avoid division-by-zero achieves the same first order property as the original kernel deconvolution estimator considered in Carroll and Hall (1988), Fan (1991).

Our problem of interest is closely related to, but not limited to the problem of matrix completion. It is because we are not only recovering the matrix as a stack of numbers, but the underlying latent functions and column features.

There have been a huge amount of intellectual advances in the matrix completion, especially in spectral approaches such as matrix factorization. This method is based on the observation that all matrices admit a unique singular value decomposition, and its goal is to recover the target matrix by estimating row and column singular vectors from the partial noisy observation. Since Srebro et al. (2004) suggested to use low-rank matrix approximation in this context, many statistically efficient estimators based on optimization have been suggested. They prove that $rn \log n$ samples out of n^2 entries suffice to impute the missing entries by matrix factorization, where r is rank of the matrix to recover; see Candès and Recht (2009), Candès and Tao (2010), Rohde et al. (2011), Keshavan et al. (2009), Negahban and Wainwright (2012), Jain et al. (2013), for example.

However, many of these approaches require that the matrix is of low rank ($r \ll n$) to achieve a sensible sample complexity. As Ganti et al. (2015) pointed out, a simple nonlinear entrywise transformation can produce a matrix of high rank, although there are only a few free model parameters.

Latent variable model is a more general model and it subsumes the low rank model as a special case where the latent features are r dimensional vectors and the latent function is given as their inner product (or a bilinear function). Chatterjee (2015) proposed the universal singular value thresholding (USVT) estimator inspired by low-rank matrix approximation and he argued that the USVT estimator provides an accurate estimate for any Lipschitz function under latent variable model.

However, with his analysis based on step function approximation (stochastic block model approximation), to obtain a consistent estimate for an $n \times n$ matrix, $\Omega\left(n^{2-\frac{2}{r+2}}\right)$ observations out of n^2 are required, where r stands for the dimension of the latent spaces where the row and column latent variables are drawn from. The rate of USVT is further investigated in a more recent work by Xu (2017).

In contrast, Lee et al. (2016) suggested a similarity-based estimator for collaborative filtering and they proved that their estimator requires $\Omega\left(n^{\frac{3}{2}+\delta}\right)$ for any small $\delta > 0$ out of n^2 for consistency of the estimator, as long as $r = o(\log n)$. They reported that the bottleneck in sample complexity was the overlap requirement between pairs of rows, which necessitates $np^2 \gg 1$, which is a commonly observed phenomenon in neighbor-based approaches.

When interpreted as matrix completion method, the algorithm suggested in this paper can avoid this restrictive overlap requirement by using distribution signatures, such as moments of distribution (in fact, the characteristic function is used). For that purpose we additionally assumed monotonicity of the latent function with respect to the column feature. We will discuss later that this monotonicity assumption is required only 1) when our goal is to estimate the matrix entries, or 2) when the noise density has to be estimated. The assumption is not necessary if we are to estimate only the distributions, or equivalently, the latent function.

This flavor of monotonicity assumption is quite common in crowdsourcing and ranking literature. For example, the Dawid-Skene model suggested in Dawid and Skene (1979) and its generalization (see Zhou et al. (2015), Khetan and Oh (2016)) assumes each worker i and task j are respectively assigned latent features p_i and q_j in the interval $[0, 1]$. Roughly, p_i denotes the competence of worker i and q_j denotes the difficulty of task j. Actually our assumption is weaker than this, because we assume monotonicity only for the column features while this line of works assumes monotonicity in both directions.

Similarly, in the literature of rank aggregation from pairwise comparison, the Bradley-Terry-Luce model (Bradley and Terry (1952), Luce (1959)) and the Thurstone model (Thurstone (1927)) are in the mainstreams. Some generalization of it such as nonparametric Bradley-Terry model by Chatterjee (2015) and Strong Stochastic Transitivity Shah et al. (2016) are suggested, but they still share the monotonicity at the core.

Another related field of research is that of graphon estimation. A graphon is a measurable function $W : [0, 1]^2 \to [0, 1]$, which was originally introduced as a limit object of the connectivity pattern in graph instances, but it is now also widely used as a generative model in the study of large networks.

Suggested as a nonparametric framework for the analysis of networks, estimating graphon has gained huge interest in the scene of modern statistics. The framework relates to stochastic blockmodels Airoldi et al. (2008), Rohe et al. (2011) and degree-based models Bickel and Chen (2009), Chatterjee et al. (2011), Bickel et al. (2011). Theory and algorithm for the consistency and the rate of convergence for the graphon estimation have been pursued via numerous approaches including Wolfe and Olhede (2013), Airoldi et al. (2013), Zhang et al. (2015). Recently, Gao et al. (2015) and Klopp et al. (2017) discussed the optimal minimax rate of convergence, but unfortunately their algorithms are not computationally tractable.

Appendix B. Lower bounds: Proofs of Theorems 4 and 5

B.1. Lower Bounding MSE By L^2 Function Distance

Here, we establish that the MSE of any estimator can be lower bounded by L^2 distance between "estimated" latent function and actual latent function. This will be useful steps towards establishing the desired bounds in Theorems 4 and 5 since for each context, we will identify hard instance of latent functions that will be difficult to estimate in terms of L^2 distances. To that end, we shall assume that our estimator ϕ has access to an oracle that provide information about latent parameter associated with each column. Clearly, the lower bound on MSE for such a powerful estimator will be lower bound on MSE for any valid estimator.

Recall that the L^2 norm of a function g defined on $[0,1]$ is defined as

$$\|g\|_{L^2[0,1]} = \left(\int_0^1 |g(x)|^2 \, dx\right)^{1/2}. \tag{13}$$

We use a subscript to explicitly indicate the function is estimated from a certain number of sample observations, i.e., we let $\hat{g}_\nu$ denote an estimated function for g from ν sample points. Also recall the definition of MSE from Eq. (2): for estimator $\varphi : Z \mapsto \hat{A}^\varphi$,

$$MSE(\varphi) = \mathbb{E}\left[\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\hat{A}^\varphi(i,j) - A(i,j)\right)^2\right].$$

Lemma 10 *For any algorithm $\varphi : Z \mapsto \hat{A}^\varphi$,*

$$MSE\left(\varphi\right) \geq (1-p)\mathbb{E}_{\theta_{row}^{(1)},\theta_{col}^{(-1)},\nu}\left[\left\|\hat{g}_\nu^\varphi(\theta_{row}^{(1)},\cdot) - g(\theta_{row}^{(1)},\cdot)\right\|_{L^2[0,1]}^2\right],$$

where $\nu \sim Binomial(n-1,p)$ and $\theta_{col}^{(-1)}$ denotes $\{\theta_{col}^{(j)} : j \in [n], j \neq 1\}$.

Proof Given an algorithm $\varphi : Z \mapsto \hat{A}^\varphi$, we first reduce the expression of MSE as follows:

$$\begin{aligned}
MSE(\varphi) &= \mathbb{E}\left[\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\hat{A}^\varphi(i,j) - A(i,j)\right)^2\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\left(\hat{A}^\varphi(i,j) - A(i,j)\right)^2\right] \\
&= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left(\hat{A}^\varphi(1,j) - A(1,j)\right)^2\right] && \because \text{rows are exchangeable} \\
&= \mathbb{E}\left[\left(\hat{A}^\varphi(1,1) - A(1,1)\right)^2\right] && \because \text{columns are exchangeable} \\
&= \mathbb{E}\left[\left(\hat{g}^\varphi(\theta_{row}^{(1)},\theta_{col}^{(1)}) - g(\theta_{row}^{(1)},\theta_{col}^{(1)})\right)^2\right].
\end{aligned} \tag{14}$$

Note that $\hat{g}^\varphi$ is a function estimated based on the data $\{Z(i,j) : (i,j) \in \mathcal{O}\}$.

Now suppose that the algorithm φ is equipped with an oracle, i.e., it can access to the true value of the latent features for $(i,j) \in \mathcal{O}$. For an oracle algorithm having access to $\theta^{(j)}_{col}$, there is no information utilized to estimate $\hat{A}^{\varphi}(1,1)$ from the observations $\{Z(i,j) : (i,j) \in \mathcal{O}, i \neq 1\}$. Note that no regularity is assumed over the first coordinate of latent functions g in our model other than the monotonicity assumption. The restriction of function g at two different row features, $g(\theta^{(1)}_{row}, \cdot)$ and $g(\theta^{(2)}_{row}, \cdot)$, can be very different and the only information transferrable from one row to another is the pairwise order between column features. Since φ already has access to the true column features, the mutual information between $\hat{A}^{\varphi}(1,1)$ and $\{Z(i,j) : (i,j) \in \mathcal{O}, i \neq 1\}$ is zero.

Recall that M is the masking matrix, i.e., $M(i,j) = 1$ if and only if $(i,j) \in \mathcal{O}$; otherwise, $M(i,j) = 0$. We let $M(1,\cdot)$ denote the first row of the matrix M, and let $\nu = \sum_{j=2}^{n} M(1,j)$ denote the number of observed entries in row 1, excluding $(1,1)$. Note that ν is a random variable distributed as per binomial distribution $Binomial(n-1,p)$. We use $\theta^{(-1)}_{col}$ as a shorthand notation to denote $\{\theta^{(j)}_{col} : j \in [n], j \neq 1\}$. Assuming φ perfectly restores g at $\left\{\left(\theta^{(i)}_{row}, \theta^{(j)}_{col}\right) : (i,j) \in \mathcal{O}\right\}$, it follows that

$$
\begin{aligned}
Eq.(14) &= \mathbb{E}_{\theta^{(1)}_{row}}\left[\mathbb{E}_{\theta^{(1)}_{col},\theta^{(-1)}_{col},M}\left[\left(\hat{g}^{\varphi}(\theta^{(1)}_{row},\theta^{(1)}_{col}) - g(\theta^{(1)}_{row},\theta^{(1)}_{col})\right)^2 \middle| \theta^{(1)}_{row}\right]\right] \\
&= \mathbb{E}_{\theta^{(1)}_{row}}\left[\mathbb{E}_{\theta^{(1)}_{col},\theta^{(-1)}_{col},M(1,\cdot)}\left[\left(\hat{g}^{\varphi}(\theta^{(1)}_{row},\theta^{(1)}_{col}) - g(\theta^{(1)}_{row},\theta^{(1)}_{col})\right)^2 \middle| \theta^{(1)}_{row}\right]\right] \qquad \because \text{oracle} \\
&\geq \mathbb{E}_{\theta^{(1)}_{row}}\left[\mathbb{E}_{\theta^{(1)}_{col},\theta^{(-1)}_{col},M(1,\cdot)}\left[\left(\hat{g}^{\varphi}(\theta^{(1)}_{row},\theta^{(1)}_{col}) - g(\theta^{(1)}_{row},\theta^{(1)}_{col})\right)^2 \mathbb{I}\{M(1,1) \neq 1\} \middle| \theta^{(1)}_{row}\right]\right] \\
&\geq (1-p)\mathbb{E}_{\theta^{(1)}_{row},\theta^{(-1)}_{col},\nu}\left[\left\|\hat{g}^{\varphi}_{\nu}(\theta^{(1)}_{row},\cdot) - g(\theta^{(1)}_{row},\cdot)\right\|^2_{L^2[0,1]}\right].
\end{aligned}
$$

■

In the subsequent sections, we investigate lower bounds on $\left\|\hat{g}^{\varphi}_{\nu}(\theta^{(1)}_{row},\cdot) - g(\theta^{(1)}_{row},\cdot)\right\|^2_{L^2[0,1]}$ to establish Theorems 4 and 5. Without loss of generality, we may assume our matrix is a 1 by n matrix due to the oracle argument. To further establish lower bound, we shall suppose that given $\{\theta^{(1)}_{row}\}$, $\{\theta^{(j)}_{row}\}_{j\in[n]}$, $\{Z(1,j) : j \in [n] \text{ and } M(1,j) = 1\}$, the algorithm φ can perfectly estimate the function $g(\theta^{(1)}_{row}, x)$ for $x \in \{\theta^{(j)}_{col} : j \in [n], M(1,j) = 1\}$. Then we show that there exists an adversarial function $g^{\dagger}$ such that

1. $g\left(\theta^{(1)}_{row}, \theta^{(j)}_{col}\right) = g^{\dagger}\left(\theta^{(1)}_{row}, \theta^{(j)}_{col}\right)$ for all $j \in [n]$ such that $M(1,j) = 1$, and
2. $\left\|g(\theta^{(1)}_{row},\cdot) - g^{\dagger}(\theta^{(1)}_{row},\cdot)\right\|_{L^2[0,1]}$ is sufficiently large.

Then, there is no way for φ to distinguish $g^{\dagger}$ from g based on the data, φ would return the same output g even if the latent function g were replaced with $g^{\dagger}$. Therefore, $\left\|g(\theta^{(1)}_{row},\cdot) - g^{\dagger}(\theta^{(1)}_{row},\cdot)\right\|_{L^2[0,1]}$ establishes a lower bound on $MSE(\varphi)$. More detailed argument for the noiseless case (Appendix B.2) and noisy case (Appendix B.3) will follow.

B.2. Proof of Theorem 4

In this section, we show that for any slice of true latent function $g_1 := g(\theta^{(1)}_{row}, \cdot) : [0,1] \to \mathbb{R}$ and for any set of sampling points $y_1, \dots, y_\nu \in [0,1]$, there exists an adversarial function $g_1^\dagger : [0,1] \to \mathbb{R}$ such that $g_1(y) = g_1^\dagger(y)$ for all $y \in \{y_1, \dots, y_\nu\}$, yet $\|g_1 - g_1^\dagger\|_2^2 \geq \frac{c}{\nu}$ for some universal constant c, independent of g_1 and ν. This claim follows from a classical result in function approximation theory.

We define some notations before introducing the function approximation lemma. Recall that the L^1 norm of a function $g : [0,1] \to \mathbb{R}$ is defined as $\|g\|_{L^1[0,1]} := \int_0^1 |g(x)|dx$ (see Eq. (13) for comparison with L^2 norm). We let $L^1[0,1] := \left\{g : [0,1] \to \mathbb{R} : \|g\|_{L^1[0,1]} < \infty\right\}$ denote the space of functions with finite L^1 norm, i.e., integrable functions. We also recall that $C^\infty[0,1]$ is the space of functions defined on $[0,1]$, which are infinitely differentiable. Lastly, we call a function g to be δ-Lipschitz if $\|g(y_1) - g(y_2)\| \leq \delta\|y_1 - y_2\|$ for any two points y_1, y_2 in the domain of g.

Theorem 11 (Kudryavtsev (1991), Lemma 4.4, simplified) *There exists a universal constant c such that for every $\nu \in \mathbb{N}$, and for any $y_1, \dots, y_\nu \in [0,1]$, there exists a δ-Lipschitz function $h \in L^1[0,1] \cap C^\infty[0,1]$ for which*

1. *$h(y_i) = 0$, for all $i = 1, \dots, \nu$, and*
2. *$\|h\|_{L^2[0,1]} \geq c\frac{\delta}{\sqrt{\nu}}$.*

We use this theorem to prove Theorem 4.

Theorem 12 (Full version of Theorem 4) *In the noiseless scenario, for any estimation algorithm φ, there exists a hard instance for which*

$$MSE(\varphi) \geq (1-p)\frac{c^2\delta^2}{(n-1)p}.$$

Proof [Proof of Theorem 12] Choose a positive real number $\delta < \frac{L-l}{2}$. Consider a bounded function $g : [0,1]^2 \to \mathbb{R}$, which is $(l+\delta, L-\delta)$ bi-Lipschitz with respect to the second argument. We suppose that given any data $\{\theta^{(1)}_{row}\}$, $\{\theta^{(j)}_{row}\}_{j\in[n]}$, $\{g(\theta^{(1)}_{row}, \theta^{(j)}_{col}) : j \in [n] \text{ and } M(1,j) = 1\}$, algorithm φ can perfectly restore the function $g(\theta^{(1)}_{row}, \cdot)$ from data.

Let $\nu := \sum_{j=2}^n M(1,j)$ denote the number of samples observed. By Theorem 11, there exists a δ-Lipschitz function $h \in L^1[0,1] \cap C^\infty[0,1]$ which satisfies

1. $h(\theta^{(j)}_{col}) = 0$, for all $j \in [n]$ such that $M(1,j) = 1$, and
2. $\|h\|_{L^2[0,1]} \geq c\frac{\delta}{\sqrt{\nu}}$.

Now we consider an adversarial function $g^\dagger : [0,1]^2 \to \mathbb{R}$ (for given data), which is defined as

$$g^\dagger(x,y) = g(x,y) + h(y)\mathbb{I}\left\{x = \theta^{(1)}_{row}\right\}.$$

First of all, we remark that $g^\dagger$ is a valid latent function which satisfies every criterion in our model (see Section 2), because h is continuous and δ-Lipschitz. If the latent function g were replaced with $g^\dagger$ by an adversary, the algorithm φ could not recognize that from given data because

$h(\theta_{col}^{(j)}) = 0$, for all $j \in [n]$ such that $M(1,j) = 1$. Therefore, φ would still return $\hat{g}^{\varphi} = g$ instead of yielding $\hat{g}^{\varphi} = g^{\dagger}$ even though the true latent function is now $g^{\dagger}$.

This leads to the following lower bound, regardless of $\theta_{row}^{(1)} \in [0,1]$:

$$\left\| \hat{g}_{\nu}^{\varphi}(\theta_{row}^{(1)}, \cdot) - g^{\dagger}(\theta_{row}^{(1)}, \cdot) \right\|_{L^2[0,1]}^2 = \left\| g(\theta_{row}^{(1)}, \cdot) - g^{\dagger}(\theta_{row}^{(1)}, \cdot) \right\|_{L^2[0,1]}^2 = \|h\|_{L^2[0,1]}^2 \geq \frac{c^2\delta^2}{\nu}.$$

Inserting this back to Lemma 10, we can conclude the following MSE lower bound even if φ is an algorithm which can perfectly estimate g from a finite number of samples. Recall that ν denotes the number of observations used to estimate $\hat{g}^{\varphi}$ and it is a random variable distributed as per $Binomial(n-1, p)$.

$$\begin{aligned}
MSE\left(\varphi\right) &\geq (1-p)\mathbb{E}_{\theta_{row}^{(1)}, \theta_{col}^{(-1)}, \nu}\left[\left\| \hat{g}_{\nu}^{\varphi}(\theta_{row}^{(1)}, \cdot) - g^{\dagger}(\theta_{row}^{(1)}, \cdot) \right\|_{L^2[0,1]}^2 \right] \\
&\geq (1-p)\mathbb{E}_{\theta_{col}^{(-1)}, \nu}\left[\min_{\theta_{row}^{(1)} \in [0,1]} \left\| \hat{g}_{\nu}^{\varphi}(\theta_{row}^{(1)}, \cdot) - g^{\dagger}(\theta_{row}^{(1)}, \cdot) \right\|_{L^2[0,1]}^2 \right] \\
&\geq (1-p)\mathbb{E}_{\theta_{col}^{(-1)}, \nu}\left[\frac{c^2\delta^2}{\nu} \right] \\
&\geq (1-p)\mathbb{E}_{\nu}\left[\frac{c^2\delta^2}{\nu} \right] \\
&\geq (1-p)\frac{c^2\delta^2}{\mathbb{E}_{\nu}\left[\nu\right]} && \because \text{Jensen's inequality} \\
&= (1-p)\frac{c^2\delta^2}{(n-1)p}.
\end{aligned}$$

The lower bound essentially quantifies the uncertainty between two functions g and $g^{\dagger}$ which could have generated the same data to feed algorithm φ. We have shown a lower bound for an oracle algorithm, which has access to the latent features $\theta_{row}^{(1)}$ and $\{\theta_{col}^{(j)}\}_{j \in [n]: M(1,j)=1}$ and can perfectly restore a certain latent function. Since no algorithm can outperform an oracle, this lower bound holds for any algorithm, i.e., for any algorithm φ, there exists a hard instance to estimate. ■

B.3. Proof of Theorem 5

When the measurements are convoluted by a supersmooth additive noise (see Eq. (5) for definition), it gets exponentially harder to estimate the underlying function. We adopt the lower bound result from Fan (1991) to prove our MSE lower bound which supports this claim.

For that purpose, we first remark that we can interpret a slice of latent function, $g(\theta_{row}^{(1)}, \cdot)$, as the (pseudo-) inverse of a cumulative distribution function $F^{(1)}$. That is to say, if $g(\theta_{row}^{(1)}, y) = z$ for $y \in [0,1]$, we can rewrite it as $F^{(1)}(z) = y$ with the support of the distribution $F^{(1)}$ being the same with the range of $g(\theta_{row}^{(1)}, \cdot)$. Since the latent function g is bi-Lipschitz, the distribution $F^{(1)}$ is absolutely continuous, and it which admits a probability density $f^{(1)}$.

Fan (1991) defined the following class of density parametrized by three parameters m, B, and $0 \leq \alpha < 1$.

$$\mathcal{C}_{m,\alpha,B} = \left\{ f(x) : \left| f^{(m)}(x) - f^{(m)}\left(x + \delta\right) \right| \leq B\delta^{\alpha} \right\},$$

Since our density $f^{(i)}$ is the derivative of $F^{(i)}$, it satisfies $f^{(i)}(z) \leq \frac{1}{l}$ by the inverse function theorem. Therefore, for any valid latent function $g : [0,1]^2 \to \mathbb{R}$, $f^{(1)} = \frac{d}{dz}F^{(1)} = \frac{d}{dz}g^{-1}\left(\theta^{(1)}_{row}, \cdot\right)$ belongs to Fan's class $\mathcal{C}_{0,0,\frac{1}{l}}$.

The following hardness result is excerpted from Fan (1991). We let ν denote the number of measurements corrupted by additive noise.

Theorem 13 (Fan (1991), Theorem 4, simplified) *For any x_0, no estimator $\hat{T}_N$ can estimate $T(f) = f^{(\lambda)}(x_0)$ with the constraint $f \in \mathcal{C}_{m,\alpha,B}$ faster than $O\left((\log \nu)^{-(m+\alpha-\lambda)/\beta}\right)$, i.e., there is a universal constant $c > 0$ such that*

$$\sup_{f \in \mathcal{C}_{m,\alpha,B}} \mathbb{E}\left[\left(\hat{T}_\nu - T(f)\right)^2\right] > c\,(\log \nu)^{-2(m+\alpha-\lambda)/\beta}. \tag{15}$$

Since the cumulative distribution function can be considered as the anti-derivative of the density, or the derivative of "order -1" as discussed in Fan (1991) Theorem 6 and Section 4, we have for any $x \in \mathbb{R}$,

$$\sup_{f \in \mathcal{C}_{0,0,\frac{1}{l}}} \mathbb{E}\left[(\hat{F}_\nu(x) - F(x))^2\right] > c\,(\log \nu)^{-2/\beta}, \tag{16}$$

by inserting $m = 0, \alpha = 0, B = \frac{1}{l}$, and $\lambda = -1$ to Eq. (15).

Now we are ready to use this result to prove Theorem 5.

Theorem 14 (Full version of Theorem 5) *In the additive noise scenario, for any estimation algorithm φ, there exists a hard instance for which*

$$MSE(\varphi) \geq \frac{(1-p)l^2c^{3/2}}{6\sqrt{2}}\big(\log(n-1)p\big)^{-3/\beta}.$$

Proof [Proof of Theorem 14] We let $\hat{F}$ and F denote the pseudo-inverse of $\hat{g}\left(\theta^{(1)}_{row}, \cdot\right)$ and $g\left(\theta^{(1)}_{row}, \cdot\right)$, respectively. Since we assumed the latent function $g\left(\theta^{(1)}_{row}, \cdot\right)$ is (l, L) bi-Lipschitz for any $\theta^{(1)}_{row} \in [0,1]$, its inverse function $F = g^{-1}\left(\theta^{(1)}_{row}, \cdot\right)$ is continuous, monotone increasing over $[0,1]$ and $\left(\frac{1}{L}, \frac{1}{l}\right)$ bi-Lipschitz. Therefore, we can treat F as an absolutely continuous distribution function and its derivative f belongs to Fan's class $\mathcal{C}_{m,\alpha,B}$ with $m = 0, \alpha = 0$, and $B = \frac{1}{l}$.

Suppose that given any data $\{\theta^{(1)}_{row}\}$, $\{\theta^{(j)}_{row}\}_{j\in[n]}$, $\{g(\theta^{(1)}_{row}, \theta^{(j)}_{col}) : j \in [n] \text{ and } M(1,j) = 1\}$, algorithm φ returns an estimate of the latent function $\hat{g}^\varphi_\nu(\theta^{(1)}_{row}, \cdot)$. Here, $\nu := \sum_{j=2}^n M(1,j)$ in the subscript denotes the number of samples used for estimation of $\hat{g}^\varphi_\nu$.

Let $\hat{F}^\varphi_\nu := (\hat{g}^\varphi_\nu)^{-1}(\theta^{(1)}_{row}, \cdot)$. We may assume $\hat{g}^\varphi_\nu$ is nondecreasing, because g is monotone increasing from the model assumption. In fact, $g\left(\theta^{(1)}_{row}, \cdot\right)$ is assumed to be not only monotone increasing, but (l, L) bi-Lipschitz. Therefore, F is $\left(\frac{1}{L}, \frac{1}{l}\right)$ bi-Lipschitz.

Let $z^* := \arg\max_{z\in\mathbb{R}} \mathbb{E}\left[(\hat{F}^\varphi_\nu(z) - F(z))^2\right]$. Then let $y^* = F(z^*)$ and $\hat{y}^*_\nu := \hat{F}^\varphi_\nu(z^*)$ denote the image of z^* under F and $\hat{F}^\varphi_\nu$, respectively. Note that $\hat{F}^\varphi_\nu$ is a random function, and hence, $\hat{y}^*_\nu$ is a random variable. Subsequently, we define $\hat{z}^*_\nu := F^{-1}(\hat{y}^*_\nu)$.

Without loss of generality, we may assume $y^* \leq \hat{y}^*_\nu$ and it follows that $\hat{z}^*_\nu \geq z^*$. Then for $y \in [y^*, \hat{y}^*_\nu]$,

$$\begin{aligned} g\left(\theta^{(1)}_{row}, y\right) - \hat{g}^\varphi_\nu\left(\theta^{(1)}_{row}, y\right) &\geq g\left(\theta^{(1)}_{row}, y\right) - \hat{g}^\varphi_\nu\left(\theta^{(1)}_{row}, \hat{y}^*_\nu\right) \\ &= g\left(\theta^{(1)}_{row}, y\right) - g\left(\theta^{(1)}_{row}, y^*\right) \\ &\geq l(y - y^*). \end{aligned}$$

From the definition of L^2 distance, it follows that

$$\begin{aligned} \left\|\hat{g}^\varphi_\nu\left(\theta^{(1)}_{row}, \cdot\right) - g\left(\theta^{(1)}_{row}, \cdot\right)\right\|^2_{L^2[0,1]} &= \int_0^1 \left|\hat{g}^\varphi_\nu\left(\theta^{(1)}_{row}, y\right) - g\left(\theta^{(1)}_{row}, y\right)\right|^2 dy \\ &\geq \int_{y^*}^{\hat{y}^*_\nu} \left|\hat{g}^\varphi_\nu\left(\theta^{(1)}_{row}, y\right) - g\left(\theta^{(1)}_{row}, y\right)\right|^2 dy \\ &\geq \int_{y^*}^{\hat{y}^*_\nu} l^2 \left|y - y^*\right|^2 dy \\ &= \frac{l^2}{3}\left|\hat{y}^*_\nu - y^*\right|^3 \\ &= \frac{l^2}{3}\left|\hat{F}^\varphi_\nu(z^*) - F(z^*)\right|^3. \end{aligned} \tag{17}$$

Recall from Lemma 10 that for any algorithm $\varphi : Z \mapsto \hat{A}^\varphi$,

$$MSE\left(\varphi\right) \geq (1-p)\mathbb{E}_{\theta^{(1)}_{row}, \theta^{(-1)}_{col}, \nu}\left[\left\|\hat{g}^\varphi_\nu(\theta^{(1)}_{row}, \cdot) - g(\theta^{(1)}_{row}, \cdot)\right\|^2_{L^2[0,1]}\right],$$

where $\nu \sim Binomial(n-1, p)$ and $\theta^{(-1)}_{col}$ denotes $\{\theta^{(j)}_{col} : j \in [n], j \neq 1\}$. If we restrict our latent function to take the form $g\left(\theta^{(i)}_{row}, \theta^{(j)}_{col}\right) = g_2(\theta^{(j)}_{col})$ for some $g_2 : [0,1] \to \mathbb{R}$, then we can remove the expectation with respect to $\theta^{(1)}_{row}$. From Eq. (17), it follows that

$$\begin{aligned} MSE\left(\varphi\right) &= (1-p)\mathbb{E}_{\theta^{(-1)}_{col}, \nu}\left[\frac{l^2}{3}\left|\hat{F}^\varphi_\nu(z^*) - F(z^*)\right|^3\right] \\ &\geq \frac{(1-p)l^2}{3}\mathbb{E}_\nu\left[\mathbb{E}_{\theta^{(-1)}_{col}}\left[\left|\hat{F}^\varphi_\nu(z^*) - F(z^*)\right|^2\right]^{3/2}\right] \qquad \because \text{Jensen's inequality} \end{aligned}$$

By Theorem 13–more precisely, by Eq. (16)–for any ν, there exists a latent function g_2 (and corresponding $f \in \mathcal{C}_{0,0,\frac{1}{l}}$ such that for any oracle algorithm φ,

$$\mathbb{E}_{\theta^{(-1)}_{col}}\left[\left|\hat{F}^\varphi_\nu(z^*) - F(z^*)\right|^2\right] \geq \frac{c}{2}\left(\log \nu\right)^{-2/\beta}.$$

All in all, there exists a hard instance of latent function g such that

$$\begin{aligned} MSE\left(\varphi\right) &\geq \frac{(1-p)l^2}{3}\mathbb{E}_\nu\left[\left(\frac{c}{2}\left(\log\nu\right)^{-2/\beta}\right)^{3/2}\right] \\ &= \frac{(1-p)l^2c^{3/2}}{6\sqrt{2}}\mathbb{E}_\nu\left[\left(\log\nu\right)^{-3/\beta}\right] \\ &\geq \frac{(1-p)l^2c^{3/2}}{6\sqrt{2}}\left(\log\mathbb{E}_\nu\left[\nu\right]\right)^{-3/\beta} && \because \text{Jensen's inequality} \\ &= \frac{(1-p)l^2c^{3/2}}{6\sqrt{2}}\left(\log(n-1)p\right)^{-3/\beta}. \end{aligned}$$

We can apply Jensen's inequality because $\left(\log x\right)^{-3/\beta}$ is convex when $x > 1$, for any $\beta > 0$. ■

Appendix C. Proof of Theorem 1: Noiseless Scenario

In this section, we prove Theorem 1 establishing an upper bound on MSE achievable in the noiseless setup. This is done by evaluating MSE for a specific algorithm. We start by describing the algorithm followed by evaluating its performance in terms of MSE.

C.1. Algorithm Description

We shall use a "generic" recipe for estimation in all three scenarios considered in this work: noiseless, noisy with known noise distribution and noisy with unknown noise distribution. The only change in each case would be how we handle the noise.

C.1.1. GENERIC DESCRIPTION

1. Estimate the latent feature (or quantile) $\theta^{(j)}_{col}$ of column $j \in [n]$. Let it be $\hat{q}(j)$.
2. Estimate $F^{(i)} = g^{-1}_{x=\theta^{(i)}_{row}}$ on row i, which is the inverse of the latent function $g\left(\theta^{(i)}_{row}, \cdot\right)$ restricted on the first coordinate. Let it be $\hat{F}^{(i)},\ i \in [m]$.
3. Estimate $\hat{g}^{(i)} = \left(\hat{F}^{(i)}\right)^{-1},\ i \in [m]$.
4. Plug in estimate: $\hat{A}(i,j) = \hat{g}^{(i)}(\hat{q}(j)),\ i \in [m],\ j \in [n]$, where $\hat{g}^{(i)} = \left(\hat{F}^{(i)}\right)^{-1}$.

By assumption, the function $g(x,\cdot)$ is invertible for every $x \in [0,1]$ since $g(x,\cdot) : [0,1] \to \mathbb{R}$ is continuous and monotonically increasing. Let the inverse (given fixed x) be denoted as $g^{-1}(x,\cdot) : \mathbb{R} \to [0,1]$. That is, $g^{-1}(x,\cdot)$ can be viewed as a cumulative distribution function for distribution on $\mathbb{R}$. In short, for each row $i \in [m]$, we can consider the hidden latent function restricted to $x = \theta^{(i)}_{row}$, $g(x,\cdot)$, as the inverse of the cumulative distribution function along row i (see Appendix J, Definitions 46 and 47 for details).

The first two steps of the algorithm will vary across scenarios to account for noise.

C.1.2. DETAILED DESCRIPTION: NOISELESS SETUP

Notations. For $i \in [m]$, we let $\mathcal{B}_i$ denote the set of column indices for which $Z(i,j)$ is observed (similarly, $\mathcal{B}^j$ denotes the set of row indices for $j \in [n]$, respectively), that is

$$\mathcal{B}_i = \{j' \in [n] : M(i,j') = 1\} \text{ and } \mathcal{B}^j = \{i' \in [m] : M(i',j) = 1\}. \tag{18}$$

Define indicator function

$$\mathbb{I}\{\text{condition}\} = \begin{cases} 1, & \text{if condition is true,} \\ 0, & \text{if condition is false.} \end{cases} \tag{19}$$

Define Heaviside step function $H : \mathbb{R} \to \{0, \frac{1}{2}, 1\}$ as

$$H(x) = \frac{1}{2}\big(\mathbb{I}\{x > 0\} + \mathbb{I}\{x \geq 0\}\big) = \begin{cases} 1, & \text{if } x > 0, \\ \frac{1}{2}, & \text{if } x = 0, \\ 0, & \text{if } x < 0. \end{cases} \tag{20}$$

That is, $\sum_{j_2=1}^{n} H\big(Z(i,j_1) - Z(i,j_2)\big)$ is the number of entries $Z(i,j)$ in row i whose value smaller than $Z(i,j_1)$ while $Z(i,j_1)$ itself is counted with weight $\frac{1}{2}$.

Now the details of the steps of the algorithm.

1. $\hat{q}(j)$: Estimate of $\theta_{col}(j)$, $j \in [n]$. Given $Z \in \mathbb{R}^{m \times n}$ and $j \in [n]$, for $i \in \mathcal{B}^j$ define

$$\hat{q}_i(j) = \frac{\sum_{j'=1}^{n} M(i,j') H\big(Z(i,j) - Z(i,j')\big)}{\sum_{j'=1}^{n} M(i,j')}. \tag{21}$$

Subsequently, define estimation of $\theta_{col}(j)$ as

$$\hat{q}(j) = \begin{cases} \frac{1}{2}, & \text{if } \mathcal{B}^j = \emptyset, \text{ else} \\ \hat{q}_{i^*(j)}(j), & \text{where } i^*(j) \text{ is randomly chosen from } \mathcal{B}^j. \end{cases} \tag{22}$$

2. $\breve{F}^{(i)}$: Estimate of $F^{(i)} = g^{-1}_{x=\theta^{(i)}_{row}}$, $i \in [m]$. For $z \in \mathbb{R}$, define

$$\breve{F}^{(i)}(z) = \frac{\sum_{j=1}^{n} M(i,j) \mathbb{I}\{Z(i,j) \leq z\}}{\sum_{j=1}^{n} M(i,j)}. \tag{23}$$

3. and 4. $\breve{A}(i,j)$: Estimate of $A(i,j)$, $i \in [n], j \in [m]$. For each $i \in [m]$, let $\breve{g}^{(i)} = \left(\breve{F}^{(i)}\right)^{-1}$ denote the quantile function (right pseudo-inverse) associated with $\breve{F}^{(i)}$. Plugging in Eq. (22) into it leads to the estimate of matrix entry:

$$\breve{A}(i,j) = \breve{g}^{(i)}\left(\hat{q}(j)\right), \quad \forall (i,j) \in [m] \times [n]. \tag{24}$$

By definition, $\breve{F}^{(i)}$ is simply the empirical cumulative distribution function. Hence, by Glivenko-Cantelli theorem, it follows that it is a consistent estimator for $F^{(i)}$. Using the Dvoretzky-Kiefer-Wolfowitz inequality (see Appendix J, Lemma 49), we obtain concentration of $\breve{F}^{(i)}$ around $F^{(i)}$. This is summarized in Lemma 16.

C.2. Algorithm Analysis

We start by establishing two key results needed for establishing proof of Theorem 1. To that end, note that $\hat{q}_i(j)$ is the average of $\sum_{j'=1}^{n} M(i,j')$ independent random variables as per our model. Therefore, by Chernoff bound, for each i, it concentrates around its expectation, which is the true parameter $\theta_{col}^{(j)}$ of interest. This explain the choice of (21)-(22). This is summarized in Lemma 15.

By definition, $\breve{F}^{(i)}$ is simply the empirical cumulative distribution function. Hence, by Glivenko-Cantelli theorem, it follows that it is a consistent estimator for $F^{(i)}$. Using the Dvoretzky-Kiefer-Wolfowitz inequality (see Appendix J, Lemma 49), we obtain concentration of $\breve{F}^{(i)}$ around $F^{(i)}$. This is summarized in Lemma 16.

Finally, we obtain the error bound for estimation $\breve{A}(i,j)$ in Lemma 17. This will further lead to proof of Theorem 1.

We will use the following definition in what follows.

$$D_{max} \equiv \sup_{x,y\in[0,1]} g(x,y) \quad \text{and} \quad D_{min} \equiv \inf_{x,y\in[0,1]} g(x,y),$$

$$L \equiv \sup_{x,y_1\neq y_2\in[0,1]} \frac{g(x,y_2)-g(x,y_1)}{y_2-y_1} \text{ and } l \equiv \inf_{x,y_1\neq y_2\in[0,1]} \frac{g(x,y_2)-g(x,y_1)}{y_2-y_1}.$$

C.2.1. CONCENTRATION OF $\hat{q}(j)$ AROUND $\theta_{col}(j)$

We state the following.

Lemma 15 *When there is no noise ($N = 0$) in the model, for any $j \in [n]$, the quantile estimator $\hat{q}(j)$ (see Eq. (22)) concentrates to $\theta_{col}^{(j)}$ with high probability:*

$$\mathbb{P}\left(\left|\hat{q}(j) - \theta_{col}^{(j)}\right| \geq t\right) \leq 2\exp\left(-2\left|\mathcal{B}_{i^*}\right| t^2\right),$$

where i^* denote the row index chosen in Eq. (22).

Note that, when $\mathcal{B}^j = \emptyset$ and $\hat{q}(j)$ is chosen to be $\frac{1}{2}$, we shall use i^* as any index leading to $\mathcal{B}_{i^*} \subset \mathcal{B}^j$ being $\emptyset$ and hence $\mathbb{P}\left(\left|\hat{q}(j) - \theta_{col}^{(j)}\right| \geq t\right) \leq 2$, which is always true! The proof of the above Lemma can be found in Section F.

C.2.2. CONCENTRATION OF $\breve{F}^{(i)}$ AROUND $F^{(i)}$.

We state the following.

Lemma 16 (Concentration of noiseless CDF estimation) *When there is no noise in the model, the empirical cumulative distribution function (ECDF) $\breve{F}^{(i)}$ (Eq. (23)) uniformly concentrates to the true CDF $F^{(i)} = g^{-1}_{x=\theta_{row}^{(i)}}$, that is for each $i \in [m]$,*

$$\mathbb{P}\left(\sup_{z\in\mathbb{R}}\left|\breve{F}^{(i)}(z) - F^{(i)}(z)\right| > t\right) \leq 2\exp\left(-2|\mathcal{B}_i|t^2\right).$$

Proof The proof is a direct application of Dvoretzky-Kiefer-Wolfowitz inequality (see Lemma 49). ■

C.3. Completing Proof of Theorem 1

We complete the proof of Theorem 1 using Lemmas 15 and 16. To that end, we first state exponential tail bound on error in estimation, $|\breve{A}(i,j) - A(i,j)|$ in Lemma 17 and then using it, obtain bound on Mean-Square-Error (MSE) to conclude the proof in Theorem 18.

C.3.1. TAIL BOUND ON $|\breve{A}(i,j) - A(i,j)|$.

Theorem 17 (Probabilistic bound: noiseless) *For each $(i,j) \in [m] \times [n]$ and $t \geq 0$,*

$$\mathbb{P}\left(\left|\breve{A}(i,j) - A(i,j)\right| > t\right) \leq 2\exp\left(-mp\right) + 4\exp\left(-(n-1)p\left(1 - \exp\left(-\frac{2t^2}{9L^2}\right)\right)\right). \quad (25)$$

Proof Let $\breve{g}^{(i)} = \left(\breve{F}^{(i)}\right)^{-1}$ denote the quantile function (right pseudo-inverse) associated with $\breve{F}^{(i)}$. Note that $A(u,i) = g\left(\theta^{(i)}_{row}, \theta^{(j)}_{col}\right)$ and $\breve{A}(i,j) = \breve{g}^{(i)}\left(\hat{q}(j)\right)$. Let $\theta^* := F^{(i)}\left(\breve{A}(i,j)\right) = F^{(i)}\left(\breve{g}^{(i)}\left(\hat{q}(j)\right)\right)$. We can observe that $|\theta^* - \hat{q}(j)| \leq 2\left\|\breve{F}^{(i)} - F^{(i)}\right\|_\infty$.

By definition of uniform norm, at the point of continuity, we have that $|\theta^* - \hat{q}(j)| \leq \left\|\breve{F}^{(i)} - F^{(i)}\right\|_\infty$. Else if $\breve{g}^{(i)}\left(\hat{q}(j)\right)$ is a jump discontinuity of $\breve{F}^{(i)}$, then it follows that for any $\delta > 0$, $\breve{F}^{(i)}\left(\breve{g}^{(i)}\left(\hat{q}(j)\right) - \delta\right) \leq \hat{q}(j) \leq \breve{F}^{(i)}\left(\breve{g}^{(i)}\left(\hat{q}(j)\right)\right)$. Since $F^{(i)}$ is assumed to be continuous, $\left\|\breve{F}^{(i)} - F^{(i)}\right\|_\infty \geq \frac{1}{2}\sup_y \lim_{\delta\to 0+} \breve{F}^{(i)}(y) - \breve{F}^{(i)}(y-\delta)$. Therefore, $|\theta^* - \hat{q}(j)| \leq 2\left\|\breve{F}^{(i)} - F^{(i)}\right\|_\infty$.

Since $\breve{A}(i,j) = \breve{g}^{(i)}\left(\hat{q}(j)\right) = g\left(\theta^{(i)}_{row}, \theta^*\right)$, and g is (l,L)-biLipschitz,

$$\begin{aligned}
\left|A(i,j) - \breve{A}(i,j)\right| &= \left|g\left(\theta^{(i)}_{row}, \theta^{(j)}_{col}\right) - g\left(\theta^{(i)}_{row}, \theta^*\right)\right| \\
&\leq L\left|\theta^{(j)}_{col} - \theta^*\right| \\
&\leq L\left(\left|\theta^{(j)}_{col} - \hat{q}(j)\right| + |\hat{q}(j) - \theta^*|\right) \\
&\leq L\left(\left|\theta^{(j)}_{col} - \hat{q}(j)\right| + 2\left\|\breve{F}^{(i)} - F^{(i)}\right\|_\infty\right).
\end{aligned}$$

If $\left|\theta^{(j)}_{col} - \hat{q}(j)\right| \leq \frac{t}{3L}$ and $\left\|\breve{F}^{(i)} - F^{(i)}\right\|_\infty \leq \frac{t}{3L}$, then $\left|A(u,i) - \breve{A}(i,j)\right| \leq t$. Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\left|\breve{A}(i,j) - A(i,j)\right| > t\right) \\
&\quad\leq \mathbb{P}\left(\left|\hat{q}(j) - \theta^{(j)}_{col}\right| > \frac{t}{3L}\right) + \mathbb{P}\left(\sup_{z\in\mathbb{R}}\left|\breve{F}^{(i)}(z) - F^{(i)}(z)\right| > \frac{t}{3L}\right) \\
&\quad\leq 2\exp\left(-\frac{2\left|\mathcal{B}_{i^*}\right|t^2}{9L^2}\right) + 2\exp\left(-\frac{2\left|\mathcal{B}_i\right|t^2}{9L^2}\right),
\end{aligned}$$

where the last inequality follows from Lemma 15 and Lemma 16. Recall that i^* denote the row index chosen in the algorithm (see Eq. (22))

Note that $|\mathcal{B}_i|$ is the sum of n independent Bernoulli random variables with parameter p under our Bernoulli model. Therefore, it takes integral value in $\{0, 1, \ldots, n\}$ following Binomial(n,p) distribution.

$|\mathcal{B}_{i^*}|$ follows a slightly different distribution. By algorithm description (see Eq. (22)), $|\mathcal{B}_{i^*}| = 0$ if and only if $\mathcal{B}^j = \emptyset$, whose probability is $(1-p)^m$. For $i \in \mathcal{B}^j$, it is already conditioned that $M(i,j) = 1$. Therefore,

$$\mathbb{P}\left(|\mathcal{B}_{i^*}| = k\right) = \begin{cases} (1-p)^m, & \text{if } k = 0, \\ [1-(1-p)^m]\binom{n-1}{k-1} p^{k-1}(1-p)^{n-k}, & \text{if } k \geq 1. \end{cases}$$

As a last step, we will marginalize out $|\mathcal{B}_i|$ and $|\mathcal{B}_{i^*}|$.

$$\begin{aligned}
&\mathbb{P}\left(\left|\breve{A}(i,j) - A(i,j)\right| > t\right) \\
&\quad = \sum_{k_1,k_2} \left[\mathbb{P}\left(\left|\breve{A}(i,j) - A(i,j)\right| > t \,\middle|\, |\mathcal{B}_i| = k_1, |\mathcal{B}_{i^*}| = k_2\right)\right. \\
&\qquad \left. \times \mathbb{P}\left(|\mathcal{B}_i| = k_1, |\mathcal{B}_{i^*}| = k_2\right)\right] \\
&\quad \leq \sum_{k_1} 2\exp\left(-\frac{2k_1t^2}{9L^2}\right)\mathbb{P}\left(|\mathcal{B}_i| = k_1\right) \qquad (26) \\
&\qquad + \sum_{k_2} 2\exp\left(-\frac{2k_2t^2}{9L^2}\right)\mathbb{P}\left(|\mathcal{B}_{i^*}| = k_2\right). \qquad (27)
\end{aligned}$$

We can further simplify the last two terms as follows:

$$\begin{aligned}
Eq.(26) &= \sum_{k_1} 2\exp\left(-\frac{2k_1t^2}{9L^2}\right)\binom{n}{k_1}p^{k_1}(1-p)^{n-k_1} \\
&= 2\sum_{k_1}\binom{n}{k_1}\left[p\exp\left(-\frac{2t^2}{9L^2}\right)\right]^{k_1}(1-p)^{n-k_1} \\
&= 2\left[1-p\left(1-\exp\left(-\frac{2t^2}{9L^2}\right)\right)\right]^n \qquad \because \text{binomial theorem} \\
&= 2\left[1-\frac{np}{n}\left(1-\exp\left(-\frac{2t^2}{9L^2}\right)\right)\right]^n \\
&\leq 2\exp\left(-np\left(1-\exp\left(-\frac{2t^2}{9L^2}\right)\right)\right).
\end{aligned}$$

The inequality in the last line holds because $\left(1+\frac{a}{n}\right)^n \leq e^a$ for any $a \in \mathbb{R}$ and any $n \in \mathbb{N}$.

In a similar manner,

$$\begin{aligned}
Eq.(27) &= 2(1-p)^m + 2\left[1-(1-p)^m\right] \\
&\qquad \times \sum_{k_2=1}^{n} \exp\left(-\frac{2k_2t^2}{9L^2}\right)\binom{n-1}{k_2-1}p^{k_2-1}(1-p)^{n-k_2} \\
&\leq 2(1-p)^m + 2\left[1-(1-p)^m\right] \\
&\qquad \times \exp\left(-\frac{2t^2}{9L^2}\right)\exp\left(-(n-1)p\left(1-\exp\left(-\frac{2t^2}{9L^2}\right)\right)\right) \\
&\leq 2\exp\left(-mp\right) + 2\exp\left(-(n-1)p\left(1-\exp\left(-\frac{2t^2}{9L^2}\right)\right)\right).
\end{aligned}$$

Putting everything together

$$\mathbb{P}\left(\left|\breve{A}(i,j) - A(i,j)\right| > t\right) \leq 2\exp\left(-mp\right) + 4\exp\left(-(n-1)p\left(1-\exp\left(-\frac{2t^2}{9L^2}\right)\right)\right).$$

■

C.3.2. MEAN SQUARED ERROR

Let $\breve{\varphi}$ denote the estimator which maps Z to $\breve{A}$. Then

$$\begin{aligned}
MSE\left(\breve{\varphi}\right) &= \mathbb{E}\left[\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\breve{A}(i,j) - A(i,j)\right)^2\right] \\
&= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}\left[\left(\breve{A}(i,j) - A(i,j)\right)^2\right] && \because \text{linear} \\
&= \mathbb{E}\left[\left(\breve{A}(1,1) - A(1,1)\right)^2\right] && \because \text{exchangeable} \\
&= \int_0^\infty \mathbb{P}\left(\left(\breve{A}(1,1) - A(1,1)\right)^2 > t\right) dt && \because \text{positive} \\
&= \int_0^\infty \mathbb{P}\left(\left|\breve{A}(1,1) - A(1,1)\right| > \sqrt{t}\right) dt && \because u = \sqrt{t} \\
&= \int_0^\infty 2u\mathbb{P}\left(\left|\breve{A}(1,1) - A(1,1)\right| > u\right) du. && (28)
\end{aligned}$$

Now it remains to integrate the tail bounds obtained in the previous section to conclude our first main theorem. In general, we can derive the following formulae from integration by substitution

$$\int_0^\infty ue^{-au^2}ds = \int_0^\infty \frac{1}{2a}e^{-z}dz = -\frac{1}{2a}e^{-z}\Big|_0^\infty = \frac{1}{2a}, \tag{29}$$

$$\int_0^\infty ue^{-au}du = \int_0^\infty \frac{z}{a^2}e^{-z}dz = \frac{\Gamma(2)}{a^2} = \frac{1}{a^2}. \tag{30}$$

These formulae will be frequently used, because many of our error bound have such forms. Also, from the model assumption and the construction of the estimators, the estimation error is bounded

$$\left|\breve{A}(i,j) - A(i,j)\right| \leq D_{max} - D_{min} \equiv D,$$

where $D \equiv D_{max} - D_{min}$, a constant independent of m, n.

Theorem 18 (The Full Version of Main theorem 1; Noiseless MSE) *The mean squared error of the noiseless estimator $\breve{\varphi}$ is bounded above as follows:*

$$MSE\left(\breve{\varphi}\right) \leq \frac{18L^2\exp\left(\frac{2}{9L^2}\right)}{(n-1)p} + D^2\left[2\exp(-mp) + 4\exp\left(-(n-1)p\left(1-e^{-\frac{2}{9L^2}}\right)\right)\right].$$

It can be seen that as long as $mp \gg \log np$, the dominant term on the right hand side is $\frac{18L^2 \exp\left(\frac{2}{9L^2}\right)}{(n-1)p}$ which scales as $O\left(\frac{1}{np}\right)$. And $MSE\left(\breve{\varphi}\right) \to 0$ as long as $p = \omega\left(\frac{1}{m}, \frac{1}{n}\right)$.

Proof [Proof of Theorem 18] We can prove the MSE upper bound by integrate the probabilistic tail bound in Theorem 17. We first observe that for $c > 0$, $1 - e^{-cu^2} \geq ce^{-c}u^2$ for $0 \leq u \leq 1$; and for $u \geq 1$, $1 - e^{-cu^2} \geq 1 - e^{-c}$.

Plugging in Eq. (25) to Eq. (28) leads to (with notation $c = \frac{2}{9L^2}$ below)

$$\begin{aligned}
MSE\left(\breve{\varphi}\right) &= \int_0^D 2u\mathbb{P}\left(\left|A(i,j) - \breve{A}(i,j)\right| > u\right) du \\
&\leq \int_0^D 4u \exp\left(-mp\right) du \\
&\quad + \int_0^D 8u \exp\left(-(n-1)p\left(1 - \exp\left(-\frac{2u^2}{9L^2}\right)\right)\right) du \\
&\leq 2D^2 \exp\left(-mp\right) + \int_0^1 8u \exp\left(-(n-1)pce^{-c}u^2\right) du \\
&\quad + \int_1^D 8u \exp\left(-(n-1)p\left(1 - e^{-c}\right)\right) du \\
&\leq \frac{18L^2 \exp\left(\frac{2}{9L^2}\right)}{(n-1)p} \\
&\quad + D^2\left[2\exp(-mp) + 4\exp\left(-(n-1)p\left(1 - e^{-\frac{2}{9L^2}}\right)\right)\right].
\end{aligned}$$

■

Appendix D. Proof of Theorem 2

In this section, we shall establish Theorem 2 bounding Mean-Squared-Error for an estimator in a noisy setting with the known noise distribution. We shall start by describing the estimation algorithm followed by its analysis that will lead to the desired bound.

D.1. Algorithm Description

The generic algorithm remains the same as that described in Section C.1.1. However, the details of step 1 (estimating $\theta^{(j)}_{col},\ j \in [n]$) and step 2 (estimating $F^{(i)},\ i \in [m]$) of the algorithm change due to presence of noise. We shall explicitly use the knowledge of noise distribution in step 2.

1. $\hat{q}_{\text{marg}}(j)$: Estimate of $\theta^{(j)}_{col},\ j \in [n]$. Unlike noiseless case, we can not simply use empirical quantile along a given row i, $\hat{q}_i(j)$ as a proxy since noise in data can non-trivially corrupt the estimation.

Instead, we need to overcome the effect of noise by "averaging" it out. To that end, we shall use empirical quantile estimation with respect to "column average" value rather than simply with respect to a given row. Formally, we define this below.

Let $g_{\text{marg}}(y) \equiv \int_0^1 g(x,y)dx$. Then $g_{\text{marg}}(\cdot)$ is increasing since $g(x,\cdot)$ is. Given observations $Z \in \mathbb{R}^{m\times n}$, define

$$Z_{\text{marg}}(j) = \begin{cases} \frac{\sum_{i=1}^m M(i,j)Z(i,j)}{\sum_{i=1}^m M(i,j)}, & \text{if } \mathcal{B}^j \neq \emptyset \\ \frac{1}{2}, & \text{if } \mathcal{B}^j = \emptyset. \end{cases} \tag{31}$$

Then, we estimate the column feature of $j \in [n]$ as

$$\hat{q}_{\text{marg}}(j) = \frac{1}{n}\sum_{j'=1}^{n} H\left(Z_{\text{marg}}(j) - Z_{\text{marg}}(j')\right), \tag{32}$$

where H is the Heaviside step function cf. (20).

2. $\tilde{F}^{(i)}$: Estimate of $F^{(i)} = g^{-1}_{x=\theta^{(i)}_{row}}$, $i \in [m]$. In the noiseless setting, we simply used the empirical CDF as the estimation for $F^{(i)}$ by using observations along row i in matrix Z. Since there is noise added in each entry, such an estimator will provide estimate that is corrupted by additive noise.

Effectively, each entry in the row i can be viewed as summation of two independent random variables: the first random variable is $g(\theta^{(i)}_{row}, \theta^{(j)}_{col})$ with the randomness induced due to that in the column parameter $\theta^{(j)}_{col}$ that are sample uniformly from $[0,1]$; the second random variable is the additive noise. Therefore, the empirical CDF of the observations gives good estimation of distribution of the summation of these two random variables. However, the interest is to recover the distribution of the first random variable. And we do know the distribution of the second random variable.

Some Background. Putting it other way, we wish to recover distribution of random variable X, but we observe samples of $Z = X + N$ instead of X. And we do know distribution of N. Due to independence, we know that $\phi_Z(t) = \phi_X(t)\phi_N(t)$ for all $t \in \mathbb{R}$, where ϕ_Z, ϕ_X, ϕ_N denote the characteristic function of random variable Z, X and N respectively.

Since we know noise distribution, equivalently $\phi_N(\cdot)$, if we can estimate $\phi_Z(\cdot)$ from observations, say $\hat{\phi}_Z(\cdot)$, then we can "de-convolve" it to obtain estimation $\phi_X(\cdot)$ as

$$\hat{\phi}_X(t) = \frac{\hat{\phi}_Z(t)}{\phi_N(t)}, \quad t \in \mathbb{R}.$$

Now to produce estimate $\hat{\phi}_Z(\cdot)$, the first step is a non-parametric estimator of distribution of Z. The Kernel smoothing is a well-studied non-parametric approach which would attempt to estimate the density (which exists in our setting) through interpolation. Precisely, given a kernel $K : \mathbb{R} \to \mathbb{R}_{\geq 0}$ and bandwidth parameter $h > 0$, the density of Z is estimated as

$$\hat{f}_Z(z) = \frac{1}{hn}\sum_{i=1}^{n} K\left(\frac{z - Z_i}{h}\right), \quad z \in \mathbb{R}. \tag{33}$$

Denote Fourier transformation operator $\mathcal{F} : L^1(\mathbb{R}) \to C_b(\mathbb{R})$ which maps the space of absolutely integrable functions $L^1(\mathbb{R})$ to the space of continuous bounded functions. Recall that $\mathcal{F}$ maps $f \in L^1(\mathbb{R})$ to $\mathcal{F}\{f\} \in C_b(\mathbb{R})$ where for all $t \in \mathbb{R}$,

$$\mathcal{F}\Big\{f\Big\}(t) = \int_{-\infty}^{\infty} \exp(\mathbf{i}\, ts) f(s) ds.$$

We use notation $\mathbf{i} \equiv \sqrt{-1}$. Similarly, for any absolutely integrable function $g \in L^1(\mathbb{R})$ and for all $s \in \mathbb{R}$, it is possible to define an operator $\mathcal{F}^{-1} : L^1(\mathbb{R}) \to C_b(\mathbb{R})$ as

$$\mathcal{F}^{-1}\Big\{g\Big\}(s) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-\mathbf{i}\, ts) g(t) dt.$$

The Fourier inversion theorem ensures that $\mathcal{F}^{-1}\mathcal{F}f = f$ if f satisfies certain conditions. For example, if the function is absolutely integrable and piecewise continuous (which is the case in our model), then $\mathcal{F}^{-1}\left(\mathcal{F}f\right)(s) = \frac{1}{2}\left(f(s_-) + f(s_+)\right)$.

Applying Fourier operator to (33) and using linearity of $\mathcal{F}$, we obtain

$$\begin{aligned}\hat{\phi}_Z(t) &= \mathcal{F}\left\{\hat{f}_Z\right\} \\ &= \frac{1}{hn}\sum_{i=1}^{n} \mathcal{F}\left\{K\left(\frac{\cdot - Z_i}{h}\right)\right\}.\end{aligned}$$

Now, applying inverse Fourier operator, $\mathcal{F}^{-1}$, to $\hat{\phi}_Z/\phi_N$ we obtain

$$\begin{aligned}\hat{f}_X &= \mathcal{F}^{-1}\left\{\frac{\hat{\phi}_Z}{\phi_N}\right\} \\ &= \frac{1}{hn}\sum_{i=1}^{n} \mathcal{F}^{-1}\left\{\frac{\mathcal{F}\left\{K\left(\frac{\cdot - Z_i}{h}\right)\right\}}{\phi_N}\right\} \\ &= \frac{1}{hn}\sum_{i=1}^{n} \mathcal{F}^{-1}\left\{\frac{h\exp(\mathbf{i}\, Z_i\, \cdot)\phi_K(h\,\cdot)}{\phi_N}\right\},\end{aligned} \tag{34}$$

where we used the following properties of Fourier operator:

$$\begin{aligned}\mathcal{F}\left\{f(\cdot - a)\right\}(t) &= \exp(\mathbf{i}\, a\, t)\mathcal{F}\{f\}(t) \\ \mathcal{F}\left\{f(b\,\cdot)\right\}(t) &= \frac{1}{|b|}\mathcal{F}\left\{f(\cdot)\right\}\left(\frac{t}{b}\right).\end{aligned}$$

Applying similar properties to inverse Fourier operator, $\mathcal{F}^{-1}$, we obtain

$$\mathcal{F}^{-1}\left\{\frac{h\exp(\mathbf{i}\, Z_i\, \cdot)\phi_K(h\,\cdot)}{\phi_N(\,\cdot\,)}\right\}(x) = \mathcal{F}^{-1}\left\{\frac{\phi_K(\,\cdot\,)}{\phi_N(\,\cdot\, h^{-1})}\right\}\left(\frac{x - Z_i}{h}\right). \tag{35}$$

Define function L as

$$L \equiv \mathcal{F}^{-1}\left\{\frac{\phi_K(\,\cdot\,)}{\phi_N(\,\cdot\, h^{-1})}\right\}, \quad \text{i.e.,} \quad L(z) = \frac{1}{2\pi}\int \exp(-\mathbf{i}\, tz)\frac{\phi_K(t)}{\phi_N\left(\frac{t}{h}\right)}dt, \;\; z \in \mathbb{R}. \tag{36}$$

From (34) and (35), and definition of L, we obtain

$$\hat{f}_X(x) = \frac{1}{hn}\sum_{i=1}^{n} L\Big(\frac{x - Z_i}{h}\Big). \tag{37}$$

Indeed, this is known as deconvolution kernel density estimator in literature. We shall adopt prior results Carroll and Hall (1988); Fan (1991); Delaigle et al. (2008) on its consistency to establish our results. Appendix L provides their summary.

Summary of Estimator. Recall $\mathcal{B}_i = \{j \in [n] : M(i,j) = 1\}$. Let ϕ_N be Fourier transform of density of noise which is known. Let K be symmetric Kernel with ϕ_K being its Fourier transform. We define $\tilde{F}^{(i)}$, estimate of $F^{(i)}$ as follows: for any choice of constants D_1, D_2 such that $D_1 \leq D_{min} \leq D_{max} \leq D_2$,

$$\tilde{F}^{(i)}(z) = \begin{cases} \int_{D_1}^{z} \tilde{f}^{(i)}(w)dw, & \text{if } z < D_2, \\ 1, & \text{if } z \geq D_2. \end{cases} \tag{38}$$

where following (37) we define

$$\tilde{f}^{(i)}(z) = \frac{1}{h|\mathcal{B}_i|} \sum_{j \in \mathcal{B}_i} L\left(\frac{z - Z(i,j)}{h}\right). \tag{39}$$

The kernel bandwidth parameter $h = (4\gamma)^{\frac{1}{\beta}} (\log |\mathcal{B}_i|)^{-\frac{1}{\beta}}$ where β and γ are smoothness parameters for the noise N (see Eq. (5)).

Remark 19 *[Constraints on kernel K.] We choose kernel K to satisfy the following conditions:*

1. *It is symmetric, i.e. $K(x) = K(-x)$ for all $x \in \mathbb{R}$.*
2. $\sup_{t \in \mathbb{R}} |\phi_K(t)| < \infty$.
3. *Support of ϕ_K is assumed to be within $[-1, 1]$. For $K \in L_1(\mathbb{R})$, $\mathcal{F}\{K\}$ is uniformly continuous, so there exists $K_{max} = \max_{t \in [-1,1]} |\phi_K(t)| < \infty$.*

3. $\tilde{A}(i,j)$: Estimate of $A(i,j)$, $i \in [m], j \in [n]$ For each $i \in [m]$, let $\tilde{g}^{(i)} = \left(\tilde{F}^{(i)}\right)^{-1}$ denote the quantile function (right pseudo-inverse) associated with $\tilde{F}^{(i)}$. Plugging Eq. (32) into it leads to the estimate of matrix entry:

$$\tilde{A}(i,j) = \tilde{g}^{(i)}\left(\hat{q}_{\text{marg}}(j)\right). \tag{40}$$

D.2. Algorithm Analysis

Similar to Section C.2, we shall establish proof of Theorem 2 by establishing concentration of quantile estimation, $\hat{q}_{\text{marg}}(j)$ around $\theta_{col}^{(j)}$ for $j \in [n]$ in Lemma 20 and concentration of CDF estimator $\tilde{F}^{(i)}$ around $F^{(i)}$ for $i \in [m]$ in Lemma 21 to se tup key results needed to conclude the desired Mean-Squared-Error bound on the eventual estimator.

D.2.1. CONCENTRATION OF $\hat{q}_{\text{MARG}}(j)$ AROUND $\theta_{col}^{(j)}$, $j \in [n]$

The quantile estimator, $\hat{q}_{\text{marg}}(j)$ as defined in (32) is shown to be concentrated around $\theta_{col}^{(j)}$ under the assumption on the noise as stated in Section 2.3.2. We define function $Q^* : \mathbb{R}_+ \to \mathbb{R}_+$ as

$$Q^*(x) = 2\sqrt{\pi}\left(\frac{1}{\sqrt{C_1 x}} + \frac{1}{\sqrt{C_2 x}} + \frac{1}{\sqrt{mpC_1 e^{-C_1}}} + \frac{1}{\sqrt{mpC_2 e^{-C_2}}}\right), \tag{41}$$

where $C_1 = \frac{l^2}{2(D_{max} - D_{min})^2}$ and $C_2 = \frac{l^2}{8\sigma^2}$ are model dependent constants.

Lemma 20 *For any* $t \geq 4Q^*(\frac{mp}{2}) = \Theta\Big(\frac{1}{\sqrt{mp}}\Big)$,

$$\mathbb{P}\left(\left|\hat{q}_{marg}(j) - \theta^{(j)}_{col}\right| > t\right) \leq \exp\left(-\frac{nt^2}{2}\right) + \exp\left(-\frac{n(\frac{t}{2} - Q^*\left(\frac{mp}{2}\right))}{3}\right) + \exp\left(-\frac{mp}{8}\right).$$

In the main text, we defined $t^*_q = Q^*\left(\frac{mp}{2}\right)$ for simplicity. Proof can be found in Section G.

D.2.2. CONCENTRATION OF $\tilde{F}^{(i)}$ AROUND $F^{(i)}$, $i \in [m]$

Here we shall establish that $\tilde{F}^{(i)}$ converges uniformly to $F^{(i)}$ in large sample limit. Specifically, we obtain the following Lemma that provides an exponentially decaying probabilistic tail bound for this uniform convergence.

Before stating the lemma, we recall $C_3 = C(l)$ (see Lemma 29) is an absolute constant which depends only on the parameter l and define a new constant $C_4 = \frac{BK_{max}(D_2 - D_1)}{\pi(4\gamma)^{\frac{1}{\beta}}}$ which also depends only on the model parameter. Let $C = C_3 + C_4$ denote the sum of those two constants.

Lemma 21 *For any* $i \in [m]$, *and for any* $t > C\left(\log|\mathcal{B}_i|\right)^{-1/\beta}$,

$$\mathbb{P}\left(\sup_{z\in[D_1,D_2]}\left|\tilde{F}^{(i)}(z) - F^{(i)}(z)\right| > t\right) \leq 2\left|\mathcal{B}_i\right|^{\frac{1}{4}}\left(\log|\mathcal{B}_i|\right)^{\frac{2}{\beta}} \exp\left(\frac{-\left|\mathcal{B}_i\right|^{1/2}}{2C_4^2\left(\log|\mathcal{B}_i|\right)^{\frac{2}{\beta}}}\left(t - C\left(\log|\mathcal{B}_i|\right)^{-1/\beta}\right)^2\right).$$

We state a useful consequence of the above result. To that end, for any $i \in [m]$, define

$$E_{row,(i)} \equiv \left\{|\mathcal{B}_i| \geq \frac{np}{2}\right\}, \text{ and } E'_{row,(i)} \equiv \left\{|\mathcal{B}_i| \leq 2np\right\}. \tag{42}$$

We define another constant for brevity

$$c_{n,p} = 2(2np)^{\frac{1}{4}}\left(\log(2np)\right)^{\frac{2}{\beta}}.$$

Corollary 22 *For any* $i \in [m]$, *and any* $t > C\left(\log\frac{np}{2}\right)^{-1/\beta}$,

$$\mathbb{P}\left(\sup_{z\in[D_1,D_2]}\left|\tilde{F}^{(i)}(z) - F^{(i)}(z)\right| > t \,\middle|\, E_{row,(i)}, E'_{row,(i)}\right) \leq c_{n,p}\exp\left(\frac{-\left(\frac{np}{2}\right)^{1/2}}{2C_4^2\left(\log(2np)\right)^{\frac{2}{\beta}}}\left(t - C\left(\log\frac{np}{2}\right)^{-1/\beta}\right)^2\right).$$

D.3. Completing Proof of Theorem 2

In this section, we complete the proof of Theorem 2 by using Lemma 20 and Corollary 22. The proof follows similar structure as that of Theorem 1. First, we establish tail bound on $|\tilde{A}(i,j) - A(i,j)|$ and then integrate it to obtain bound on Mean-Squared-Error (MSE). The details differ due to extra care required to handle noisy setting.

D.3.1. TAIL BOUND ON $|\tilde{A}(i,j) - A(i,j)|$

For given choice of parameters $t > 0$ and L, β, Q^*, m, n and p as defined before along with a universal constant C, define conditions

$$E_1 = \left\{ t \leq 8LQ^* \left(\frac{mp}{2}\right) \right\} \qquad \text{and} \qquad E_2 = \left\{ t \leq 4LC \left(\log \frac{np}{2}\right)^{-1/\beta} \right\}. \tag{43}$$

Theorem 23 *For each $(i,j) \in [m] \times [n]$, for any $t \geq 0$,*

$$\begin{aligned}
&\mathbb{P}\left(\left|\tilde{A}(i,j) - A(i,j)\right| > t\right) \\
&\quad \leq \mathbb{I}\{E_1\} + \mathbb{I}\{E_2\} + \exp\left(-\frac{n(\frac{t}{4L} - Q^*\left(\frac{mp}{2}\right))}{3}\right)\mathbb{I}\{E_1^c\} \\
&\qquad + c_{n,p}\exp\left(\frac{-\left(\frac{np}{2}\right)^{1/2}}{2C_4^2\left(\log(2np)\right)^{\frac{2}{\beta}}}\left(\frac{t}{2L} - C\left(\log\frac{np}{2}\right)^{-1/\beta}\right)^2\right)\mathbb{I}\{E_2^c\} \\
&\qquad + \exp\left(-\frac{nt^2}{8L^2}\right) + \exp\left(-\frac{mp}{8}\right) + 2\exp\left(-\frac{np}{8}\right).
\end{aligned}$$

Note that the terms in the last line which are independent of t, decays to 0 as $n \to \infty$ at the exponential rate of np as long as the sampling probability is sufficiently large, i.e., $p = \omega\left(\frac{1}{n}\right)$, .

Proof Let $\theta^* \equiv F^{(i)}\left(\tilde{A}(i,j)\right) = F^{(i)}\left(\tilde{g}^{(i)}\left(\hat{q}_{\text{marg}}(j)\right)\right)$. Since $\tilde{F}^{(i)}$ is continuous, $\left|\theta^* - \hat{q}_{\text{marg}}(j)\right| \leq \left\|\tilde{F}^{(i)} - F^{(i)}\right\|_\infty$. By the same line of argument as in the proof of Theorem 17, since $\tilde{A}(i,j) = \tilde{g}^{(i)}\left(\hat{q}_{\text{marg}}(j)\right) = g\left(\theta^{(i)}_{row}, \theta^*\right)$, and g is (l,L)-biLipschitz,

$$\begin{aligned}
\left|\tilde{A}(u,i) - A(i,j)\right| &= \left|g\left(\theta^{(i)}_{row}, \theta^{(j)}_{col}\right) - g\left(\theta^{(i)}_{row}, \theta^*\right)\right| \\
&\leq L\left|\theta^{(j)}_{col} - \theta^*\right| \\
&\leq L\left(\left|\theta^{(j)}_{col} - \hat{q}_{\text{marg}}(j)\right| + \left|\hat{q}_{\text{marg}}(j) - \theta^*\right|\right) \\
&\leq L\left(\left|\theta^{(j)}_{col} - \hat{q}_{\text{marg}}(j)\right| + \left\|\tilde{F}^{(i)} - F^{(i)}\right\|_\infty\right).
\end{aligned}$$

Again, if both $\left|\theta^{(j)}_{col} - \hat{q}_{\text{marg}}(j)\right| \leq \frac{t}{2L}$ and $\left\|\tilde{F}^{(i)} - F^{(i)}\right\|_\infty \leq \frac{t}{2L}$ are satisfied, then $\left|\tilde{A}(u,i) - A(i,j)\right| \leq t$. We can achieve the following upper bound by applying the union bound on the contraposition. We let $E_{(i)} := E_{row,(i)} \cap E'_{row,(i)}$ in this proof. Then it follows that

$$\begin{aligned}
&\mathbb{P}\left(\left|\tilde{A}(i,j) - A(i,j)\right| > t\right) \qquad (44) \\
&\quad \leq \mathbb{P}\left(\left|\hat{q}_{\text{marg}}(j) - \theta^{(j)}_{col}\right| > \frac{t}{2L}\right) + \mathbb{P}\left(\sup_{z\in\mathbb{R}}\left|\tilde{F}^{(i)}(z) - F^{(i)}(z)\right| > \frac{t}{2L}\right) \\
&\quad \leq \mathbb{P}\left(\left|\hat{q}_{\text{marg}}(j) - \theta^{(j)}_{col}\right| > \frac{t}{2L}\right) \\
&\qquad + \mathbb{P}\left(\sup_{z\in\mathbb{R}}\left|\tilde{F}^{(i)}(z) - F^{(i)}(z)\right| > \frac{t}{2L}\middle| E_{(i)}\right) + \mathbb{P}\left(E^c_{(i)}\right).
\end{aligned}$$

Because we have a trivial upper bound 1 on probability, it follows from Lemma 20 that

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{q}_{\text{marg}}(j) - \theta_{col}^{(j)}\right| > \frac{t}{2L}\right) \\
&\quad \leq \mathbb{I}\left\{t \leq 8LQ^*\left(\frac{mp}{2}\right)\right\} \\
&\qquad + \mathbb{I}\left\{t \geq 8LQ^*\left(\frac{mp}{2}\right)\right\} \\
&\qquad\quad \times \left[\exp\left(-\frac{nt^2}{8L^2}\right) + \exp\left(-\frac{n(\frac{t}{4L} - Q^*\left(\frac{mp}{2}\right))}{3}\right) + \exp\left(-\frac{mp}{8}\right)\right].
\end{aligned}$$

In a similar manner, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{z\in\mathbb{R}}\left|\tilde{F}^{(i)}(z) - F^{(i)}(z)\right| > \frac{t}{2L}\middle| E_{(i)}\right) \\
&\quad \leq \mathbb{I}\left\{t \leq 4LC\left(\log\frac{np}{2}\right)^{-1/\beta}\right\} \\
&\qquad + \mathbb{I}\left\{t \geq 4LC\left(\log\frac{np}{2}\right)^{-1/\beta}\right\} \\
&\qquad\quad \times c_{n,p}\exp\left(\frac{-\left(\frac{np}{2}\right)^{1/2}}{2C_4^2\left(\log\left(2np\right)\right)^{\frac{2}{\beta}}}\left(\frac{t}{2L} - C\left(\log\frac{np}{2}\right)^{-1/\beta}\right)^2\right).
\end{aligned}$$

Note that $t \geq 4LC\left(\log\frac{np}{2}\right)^{-1/\beta}$ implies that $\frac{t}{2L} \geq C\left(\log\frac{np}{2}\right)^{-1/\beta}$.

We used an upper bound on $\mathbb{P}\left(E_{(i)}^c\right)$ obtained from the binomial Chernoff bound:

$$\begin{aligned}
\mathbb{P}\left(E_{(i)}^c\right) &= \mathbb{P}\left(|\mathcal{B}_i| < \frac{np}{2} \text{ or } |\mathcal{B}_i| > 2np\right) \\
&\leq \mathbb{P}\left(|\mathcal{B}_i| < \frac{np}{2}\right) + \mathbb{P}\left(|\mathcal{B}_i| > 2np\right) \\
&\leq \exp\left(-\frac{np}{8}\right) + \exp\left(-\frac{np}{3}\right) \\
&\leq 2\exp\left(-\frac{np}{8}\right).
\end{aligned}$$

Substituting these three upper bounds back to Eq. (44), we can conclude that

$$\begin{aligned}
&\mathbb{P}\left(\left|\tilde{A}(i,j)-A(i,j)\right|>t\right)\\
&\quad\leq \mathbb{I}\left\{t\leq 8LQ^*\left(\frac{mp}{2}\right)\right\}+\mathbb{I}\left\{t\leq 4LC\left(\log\frac{np}{2}\right)^{-1/\beta}\right\}\\
&\qquad+\exp\left(-\frac{n(\frac{t}{4L}-Q^*\left(\frac{mp}{2}\right))}{3}\right)\mathbb{I}\left\{t\geq 8LQ^*\left(\frac{mp}{2}\right)\right\}\\
&\qquad+\mathbb{I}\left\{t\geq 4LC\left(\log\frac{np}{2}\right)^{-1/\beta}\right\}\\
&\qquad\quad\times c_{n,p}\exp\left(\frac{-\left(\frac{np}{2}\right)^{1/2}}{2C_4^2\left(\log\left(2np\right)\right)^{\frac{2}{\beta}}}\left(\frac{t}{2L}-C\left(\log\frac{np}{2}\right)^{-1/\beta}\right)^2\right)\\
&\qquad+\exp\left(-\frac{nt^2}{8L^2}\right)+\exp\left(-\frac{mp}{8}\right)+2\exp\left(-\frac{np}{8}\right).
\end{aligned}$$

■

D.3.2. MEAN SQUARED ERROR

Let $\tilde{\varphi}$ denote the estimator which maps Z to $\tilde{A}$. By the same line of arguments as in Eq. (28), the mean squared error of estimator $\tilde{\varphi}$ is given as

$$MSE\left(\tilde{\varphi}\right)=\int_0^\infty 2u\mathbb{P}\left(\left|\tilde{A}(i,j)-A(i,j)\right|>u\right)du \tag{45}$$

Also, from the model assumption and the construction of the estimators, the estimation error is bounded above:

$$\left|\tilde{A}(i,j)-A(i,j)\right|\leq D_2-D_1,$$

Let $D=D_2-D_1$ denote the upper bound. Note that D is a constant independent of m,n.

Theorem 24 (The Full Version of Main theorem 2; MSE with known noise) *The mean squared error of the deconvolution kernel estimator $\tilde{\varphi}$ is bounded above as follows:*

$$\begin{aligned}
MSE\left(\tilde{\varphi}\right)\leq\ & 16L^2C^2\left(\log\frac{np}{2}\right)^{-2/\beta}+64\sqrt[4]{8}L^2C_4^2\frac{\left(\log\left(2np\right)\right)^{\frac{2}{\beta}}}{(np)^{\frac{1}{4}}}+64L^2Q^*\left(\frac{mp}{2}\right)^2\\
&+\frac{8L^2}{n}+\frac{288L^2}{n^2}+8LQ^*\left(\frac{mp}{2}\right)\sqrt{\frac{3L\pi}{n}}+D^2\left[\exp\left(-\frac{mp}{8}\right)+2\exp\left(-\frac{np}{8}\right)\right].
\end{aligned}$$

First of all, we note that $MSE\left(\tilde{\varphi}\right)\to 0$ as $m,n\to\infty$ as long as the sample complexity satisfies $p=\omega\Big(\max\left\{\frac{1}{m},\frac{1}{n}\right\}\Big)$.

Recall from Eq. (41) that $Q^*\left(\frac{mp}{2}\right)=\Theta\left(\frac{1}{\sqrt{mp}}\right)$. We can observe that the term $16L^2C^2\left(\log\frac{np}{2}\right)^{-\frac{2}{\beta}}$ dominates in MSE, while the other terms decay faster unless the matrix is highly imbalanced so that

$mp = O(\log np)$. This MSE bound achieves the asymptotically optimal rate of convergence as long as $mp = \omega(\log np)$.

Proof [Proof of Theorem 24] In order to achieve an upper bound on the MSE for the kernel density estimator with known noise, $\tilde{\varphi}$, we integrate the tail probability bound from Theorem 23.

First of all, we recall from Eqs. (29) and (30) that

$$\int_0^\infty ue^{-au^2}du = \frac{1}{2a}, \quad \text{and} \quad \int_0^\infty ue^{-au}du = \frac{1}{a^2}.$$

Now, the mean squared error can be written in the following form:

$$\begin{aligned}
MSE(\tilde{\varphi}) &= \int_0^D 2u\mathbb{P}\left(\left|\tilde{A}(i,j) - A(i,j)\right| > u\right)du \\
&\leq \int_0^D 2u\left[\exp\left(-\frac{mp}{8}\right) + 2\exp\left(-\frac{np}{8}\right)\right]du \\
&\quad + \int_0^{8LQ^*\left(\frac{mp}{2}\right)} 2u\,du + \int_0^{4LC\left(\log\frac{np}{2}\right)^{-1/\beta}} 2u\,du \\
&\quad + \int_0^D 2u\exp\left(-\frac{nu^2}{8L^2}\right)du \qquad (46) \\
&\quad + \int_{8LQ^*\left(\frac{mp}{2}\right)}^D 2u\exp\left(-\frac{n(\frac{u}{4L} - Q^*\left(\frac{mp}{2}\right))}{3}\right)du \qquad (47) \\
&\quad + \int_{4LC\left(\log\frac{np}{2}\right)^{-1/\beta}}^D 4c_{n,p}u \\
&\qquad \times \exp\left(\frac{-\left(\frac{np}{2}\right)^{1/2}}{2C_4^2\left(\log(2np)\right)^{\frac{2}{\beta}}}\left(\frac{u}{2L} - C\left(\log\frac{np}{2}\right)^{-1/\beta}\right)^2\right)du. \qquad (48)
\end{aligned}$$

Recall that $Q^* : \mathbb{R}_+ \to \mathbb{R}_+$ is the monotone decreasing function defined in front of Lemma 20: $Q^*(x) = 2\sqrt{\pi}\left(\frac{1}{\sqrt{C_1 x}} + \frac{1}{\sqrt{C_2 x}} + \frac{1}{\sqrt{mpC_1 e^{-C_1}}} + \frac{1}{\sqrt{mpC_2 e^{-C_2}}}\right)$, where $C_1 = \frac{l^2}{2(D_{max}-D_{min})^2}$ and $C_2 = \frac{l^2}{8\sigma^2}$ are some constants which depend only on model parameters. $C = C_3 + C_4$ is the sum of two model dependent constants, where $C_3 = C(l)$ (see Lemma 29) and $C_4 = \frac{BK_{max}(D_2-D_1)}{\pi(4\gamma)^{\frac{1}{\beta}}}$. We also recall $c_{n,p} = 2(2np)^{\frac{1}{4}}\left(\log(2np)\right)^{\frac{2}{\beta}}$.

First of all, Eq. (46) is bounded above by

$$Eq.(46) \leq \int_0^\infty 2u\exp\left(-\frac{nu^2}{8L^2}\right)du = \frac{8L^2}{n}.$$

Next, we can achieve the following upper bound on Eq. (47):

$$\begin{aligned}
Eq.(47) &= \int_{8LQ^*\left(\frac{mp}{2}\right)}^{D} 2u \exp\left(-\frac{n\left(u-4LQ^*\left(\frac{mp}{2}\right)\right)}{12L}\right) du \\
&= \int_0^D 2\left(u'+8LQ^*\left(\frac{mp}{2}\right)\right)\exp\left(-\frac{n\left(u'+4LQ^*\left(\frac{mp}{2}\right)\right)}{12L}\right) du' \\
&\leq \int_0^D 2\left(u'+8LQ^*\left(\frac{mp}{2}\right)\right)\exp\left(-\frac{nu'}{12L}\right) du' \qquad \because Q^*\left(\frac{mp}{2}\right)\geq 0 \\
&\leq \int_0^\infty 2\left(u'+8LQ^*\left(\frac{mp}{2}\right)\right)\exp\left(-\frac{nu'}{12L}\right) du' \\
&= \frac{288L^2}{n^2}+8LQ^*\left(\frac{mp}{2}\right)\sqrt{\frac{3L\pi}{n}}.
\end{aligned}$$

Lastly, we compute an upper bound of the term Eq. (48). For brevity's sake, we let $c_1 = \frac{1}{2C_4^2(\log(2np))^{\frac{2}{\beta}}}\left(\frac{np}{2}\right)^{\frac{1}{2}}$, and $c_2 = C\left(\log\frac{np}{2}\right)^{-1/\beta}$ and divide the region of integration into two parts pivoting on $u = 2Lc_2$:

$$\begin{aligned}
Eq.(48) &= \int_{4Lc_2}^{D} 4c_{n,p}u\exp\left(-c_1\left(\frac{u}{2L}-c_2\right)^2\right) du \\
&\leq \int_{4Lc_2}^{D} 4c_{n,p}u\exp\left(-c_1\left(\frac{u}{4L}\right)^2\right) du \qquad \because \frac{u}{2L}-c_2\geq\frac{u}{4L}, \forall u\geq 4Lc_2 \\
&\leq \int_0^\infty 4c_{n,p}u\exp\left(-\frac{c_1}{16L^2}u^2\right) du \qquad \because u\exp\left(-\frac{c_1}{16L^2}u^2\right)\geq 0, \forall u\geq 0 \\
&= \frac{32c_{n,p}L^2}{c_1}.
\end{aligned}$$

Plugging these upper bounds back into Eqs. (46), (47) and (48) , we can obtain the following upper bound

$$\begin{aligned}
MSE\left(\tilde{\varphi}\right) \leq D^2\left[\exp\left(-\frac{mp}{8}\right)+2\exp\left(-\frac{np}{8}\right)\right] + \left[8LQ^*\left(\frac{mp}{2}\right)\right]^2 + \left[4LC\left(\log\frac{np}{2}\right)^{-1/\beta}\right]^2 \\
+\frac{8L^2}{n}+\frac{288L^2}{n^2}+8LQ^*\left(\frac{mp}{2}\right)\sqrt{\frac{3L\pi}{n}}+64\sqrt[4]{8}L^2C_4^2\frac{\left(\log\left(2np\right)\right)^{\frac{4}{\beta}}}{(np)^{\frac{1}{4}}}.
\end{aligned}$$

Rearranging the terms in the increasing order of convergence rates concludes the proof.

■

Appendix E. Proof of Theorem 3

In the previous section, we proposed an estimation procedure assuming the noise distribution is known. Here, we discuss consistent estimation procedure for similar setting with the only difference that noise distribution is *unknown*. Specifically, we shall establish Theorem 3. The structure of the section is the same with the preceding sections.

E.1. Algorithm Description

In the absence of knowledge of noise distribution, the CDF estimation algorithm presented in the previous section is no longer valid because the noise characteristic function ϕ_N in Eq. (36) is not available. To overcome the challenge of unknown noise distribution, we estimate the noise characteristic function first and then estimate the CDF using kernel deconvolution in a similar manner, but with an additional ridge parameter to avoid division by zero. It is important to recall that knowledge of noise distribution was not used for the column feature estimation in D.1. And hence it still remains valid.

The generic algorithm remains the same as that described in Section C.1.1. Step 1 (estimating $\theta_{col}^{(j)},\ j \in [n]$) remains the same as in Section D.1, but Step 2 (estimating $F^{(i)},\ i \in [m]$) of the algorithm requires an additional procedure of estimating the noise density because ϕ_N is unknown.

1. $\hat{q}_{\text{marg}}(j)$: Estimate of $\theta_{col}^{(j)}$, $j \in [n]$ The same as in Section D.1: see Eqs. (31) and (32).

2. $\hat{F}^{(i)}$: Estimate of $F^{(i)} = g^{-1}_{x=\theta_{row}^{(i)}}$, $i \in [m]$ We estimate the distribution over each row by essentially same procedure as in section D.1. Recall that the characteristic function of the additive noise, ϕ_N, is unknown and has to be estimated from data which we describe next.

2-1.* $\hat{\phi}_N(t)$*: Estimate for $\phi_N(t)$ Since the noise distribution is unknown, we need an auxiliary procedure to estimate the noise density. Here we explain an algorithm to estimate the noise characteristic function $\hat{\phi}_N(t)$.

Some Background. Before presenting the noise estimation procedure, we provide intuition behind that. Suppose that we can repeatedly observe the same instance X_i of target random variable up to independent additive noise, i.e., $Z_{ij} = X_i + N_{ij}$ with N_{ij} independent. Although we don't know the value of X_i, we can see that the difference in the observed data entries is equal to the difference between two independent noise instances: $Z_{i1} - Z_{i2} = (X_i + N_{i1}) - (X_i + N_{i2}) = N_{i1} - N_{i2}$. Assuming symmetry in the noise distribution, $N \equiv -N$ in distribution, and $N_{i1} - N_{i2}$ follows the same distribution with the sum of two independent copies of noise: $N_{i1} - N_{i2} \equiv N_{i1} + N_{i2}$. Therefore, $\phi_{N_{i1}-N_{i2}}(t) = \phi_N(t)^2$.

From symmetry of N, we know that $\phi_N(t)$, the Fourier transform of the noisy density is real-valued. In fact, we know $\phi_N(t)$ is not only real-valued but positive from the model assumption of the supersmooth noise (see Eq. (5)). It implies $\phi_{N_1-N_2}(t) = \phi_N(t)^2$ is also positive real-valued. Hence,

$$\begin{aligned}
\phi_{N_1-N_2}(t) &= \mathbb{E}\left[e^{\mathbf{i}t(N_1-N_2)}\right] \\
&= \mathbb{E}\left[\frac{e^{\mathbf{i}t(N_1-N_2)} + e^{-it(N_1-N_2)}}{2}\right] \\
&= \mathbb{E}\left[\cos t(N_1 - N_2)\right].
\end{aligned}$$

Therefore, we can estimate $\phi_N(t)$ by taking square root of the (the absolute value of) estimate $\hat{\phi}_{N_1-N_2}(t)$, which is computed as the sample-analog estimator with n independent copies of noise difference $\{N_{i1} - N_{i2}\}_{i=1}^n$. Specifically,

$$\hat{\phi}_N(t) = \hat{\phi}_{N_1-N_2}(t)^{\frac{1}{2}} = \left|\frac{1}{n}\sum_{i=1}^{n} \cos\left[t\left(N_{i1} - N_{i2}\right)\right]\right|^{\frac{1}{2}}.$$

However, the repeated measurement assumption may not be realistic, because we may not be allowed to measure the same entry multiple times. Therefore, we imitate the setup of repeated measurements by considering two columns $j_1, j_2 \in [n]$ with similar column features $\theta_{col}^{(j_1)} \approx \theta_{col}^{(j_2)}$ so that

$$\begin{aligned} Z(i,j_1) - Z(i,j_2) &= [A(i,j_1) + N(i,j_1)] - [A(i,j_2) + N(i,j_2)] \\ &= \underbrace{[A(i,j_1) - A(i,j_2)]}_{\approx 0,\ \because \theta_{col}^{(j_1)} \approx \theta_{col}^{(j_2)}} + [N(i,j_1) - N(i,j_2)] \\ &\approx N(i,j_1) - N(i,j_2). \end{aligned}$$

Summary of the Noise Density Estimation Procedure.

1. Construct $\mathcal{T} := \left\{(i,j_1,j_2) \in [m] \times [n]^2 : M(i,j_1) = M(i,j_2) = 1 \text{and } \hat{q}_{\text{marg}}(j_1) \approx \hat{q}_{\text{marg}}(j_2)\right\}$ as described in Algorithm 2.
2. For each $i \in [n]$, define $\mathcal{T}_i$ as $\mathcal{T}_i := \left\{(i',j_1,j_2) \in \mathcal{T} : i' \neq i\right\}$.
3. For each $i \in [n]$, estimate the noise characteristion function ϕ_N with the triples in $\mathcal{T}_i$ as

$$\hat{\phi}_{N,i}(t) = \left| \frac{1}{|\mathcal{T}_i|} \sum_{(i,j_1,j_2) \in \mathcal{T}_i} \cos\left[t \left(Z(i,j_1) - Z(i,j_2) \right) \right] \right|^{1/2}, \tag{49}$$

Roughly speaking, $\mathcal{T}$ is the set of index triples to mimic the repeated measurements. For row i, we use $\mathcal{T}_i$, which is a subset of $\mathcal{T}$ tailored to exclude the data from row i. This refinement of $\mathcal{T}$ to $\mathcal{T}_i$ for each row i is done for the convenience in analysis.

2-2. Computing $\hat{F}^{(i)}$ If we blindly replace ϕ_N with $\hat{\phi}_{N,i}$ in Eq. (36), it might happen that $\hat{\phi}_{N,i}\left(\frac{t}{h}\right) = 0$ while $\phi_K(t) \neq 0$ for some t. To avoid the division-by-zero problem, we introduce a ridge parameter ρ in the denominator of deconvolution kernel. By choosing an appropriate value of ρ, it vanishes fast enough as the number of samples increases so that we can achieve a consistent CDF estimator even when the noise distribution is unknown.

Summary of Estimator. Recall that $\mathcal{B}_i$ is the set of column indices j for which $Z(i,j)$ is observed; $\mathcal{B}_i = \{j \in [n] : M(i,j) = 1\}$ (see Eq. (18)). We define the kernel smoothed CDF estimator with unknown noise density as follows: for any choice of constants D_1, D_2 such that $D_1 \leq D_{min}$ and $D_2 \geq D_{max}$,

$$\hat{F}^{(i)}(z) = \begin{cases} \int_{D_1}^{z} \hat{f}^{(i)}(w)dw, & \text{if } z < D_2, \\ 1, & \text{if } z \geq D_2, \end{cases} \tag{50}$$

where

$$\hat{f}^{(i)}(z) = \frac{1}{h|\mathcal{B}_i|} \sum_{j \in \mathcal{B}_i} \hat{L}\left(\frac{z - Z(i,j)}{h} \right) \text{ and} \tag{51}$$

$$\hat{L}(z) = \frac{1}{2\pi} \int e^{-\mathbf{i}tz} \frac{\phi_K(t)}{\hat{\phi}_{N,i}\left(\frac{t}{h}\right) + \rho} dt. \tag{52}$$

Algorithm 2: Construction of the set $\mathcal{T}$ for noise density estimation.

Result: Return the set of triples $\mathcal{T}$ for noise density estimation

$J \leftarrow \left\{j \in [n] : |\mathcal{B}^j| \geq \frac{mp}{2}\right\}$;

$I \leftarrow \left\{i \in [m] : |\mathcal{B}_i \cap J| \geq \frac{|J|p}{2}\right\}$;

$\mathcal{T} \leftarrow \emptyset$;

Sort $j \in [n]$ in the increasing order of $\hat{q}_{\text{marg}}(j)$, i.e., find a permutation π such that $\hat{q}_{\text{marg}}(j) \leq \hat{q}_{\text{marg}}(j')$ if $\pi(j) < \pi(j')$;

for $i \in I$ **do**

  Renumber $j \in \mathcal{B}_i \cap J$ with $j' \in [|\mathcal{B}_i \cap J|]$ in the increasing order of $\hat{q}_{\text{marg}}(j)$;

  (let $\sigma_i : \mathcal{B}_i \cap J \subseteq [n] \to [|\mathcal{B}_i \cap J|]$; this map can be induced from π)

  $j' \leftarrow 0$;

  while $j' \leq |\mathcal{B}_i \cap J| - 1$ **do**

    if $\hat{q}_{marg}\left(\sigma_i^{-1}(j'+1)\right) - \hat{q}_{marg}\left(\sigma_i^{-1}(j')\right) \leq \frac{1}{\sqrt{|\mathcal{B}_i \cap J|}}$ **then**

      $\mathcal{T} \leftarrow \mathcal{T} \cup \left\{\left(i, \sigma_i^{-1}(j'), \sigma_i^{-1}(j'+1)\right)\right\}$;

      $j' \leftarrow j' + 2$;

    else

      $j' \leftarrow j' + 1$;

    end

  end

end

The kernel bandwidth parameter $h = (4\gamma)^{\frac{1}{\beta}} (\log |\mathcal{B}_i|)^{-\frac{1}{\beta}}$ where β and γ are smoothness parameters for the noise (see Eq. (5)) though the exact density of noise is unknown. In this paper, we choose the ridge parameter $\rho = |\mathcal{B}_i|^{-7/24}$.

3. $\hat{A}(i,j)$: Estimate of $A(i,j)$, $i \in [m], j \in [n]$ For each $i \in [m]$, let $\hat{g}^{(i)} = \left(\hat{F}^{(i)}\right)^{-1}$ denote the quantile function (right pseudo-inverse) associated with $\hat{F}^{(i)}$. Plugging Eq. (32) into it leads to the estimate of matrix entry:

$$\hat{A}(i,j) = \hat{g}^{(i)}\left(\hat{q}_{\text{marg}}(j)\right). \tag{53}$$

E.2. Algorithm Analysis

The analysis is done in parallel to those in sections C.2 and D.2. Since the quantile estimator is the same as before, we can reuse Lemma 20 to show that the quantile estimates for all $j \in [n]$ concentrate to the true values (the column features in our model) with high probability. It suffices to show the regularized deconvolution kernel ECDF consistently estimates the true CDF even when the distribution of the additive noise is unknown. Lemma 25 ensures the deconvolution kernel ECDF $\hat{F}^{(i)}$ uniformly converges to $F^{(i)}$ with high probability.

E.2.1. CONCENTRATION OF $\hat{q}_{\text{MARG}}(j)$ AROUND $\theta_{col}^{(j)}$, $j \in [n]$

The quantile estimator, $\hat{q}_{\text{marg}}(j)$ as defined in (32) is shown to be concentrated around $\theta_{col}^{(j)}$ under the assumption on the noise as stated in Section 2.3.2. See Lemma 20 for detailed statement.

E.2.2. CONCENTRATION OF $\hat{F}^{(i)}$ AROUND $F^{(i)}$, $i \in [m]$

Here we shall establish that $\hat{F}^{(i)}$ converges uniformly to $F^{(i)}$ in large sample limit. Specifically, we obtain a Lemma (Lemma 25) that provides an exponentially decaying probabilistic tail bound for this uniform convergence (See Lemma 21 for comparison with the known noise case).

Notation. We recall absolute constants $C_1 \equiv \frac{l^2}{2(D_{max}-D_{min})^2}$, $C_2 \equiv \frac{l^2}{8\sigma^2}$ and define

$$c_{\Delta A} \equiv 8\sqrt{\pi}\left(\frac{\sqrt{e^{C_1}}+\sqrt{2}}{\sqrt{C_1}} + \frac{\sqrt{e^{C_2}}+\sqrt{2}}{\sqrt{C_2}}\right).$$

Define a monotone increasing function $s_\phi : \mathbb{Z}_+ \to \mathbb{R}_+$ with $c_{\Delta A}$ as (see item 6 in section I.6 for reasons behind this definition)

$$s_\phi(x) = \frac{8\sigma(\log x)^{\frac{1}{\beta}}}{(4\gamma)^{\frac{1}{\beta}}}\frac{\sqrt{\log(4mnp)}}{(mnp)^{\frac{1}{4}}} + \frac{2(\log x)^{\frac{1}{\beta}}}{(4\gamma)^{\frac{1}{\beta}}}\left[\frac{c_{\Delta A}}{\sqrt{mp}} + \frac{2L\sqrt{2}}{\sqrt{np}}(1+\sqrt[4]{np})\right]. \tag{54}$$

Note that σ, β, γ are model parameters for noise, and L, l are lipschitz constants for the class of latent functions. The absolute constant $C_3 = C_3(l)$ (see Lemma 29) depends only on l. The bandwidth parameter h is chosen as $h = (4\gamma)^{\frac{1}{\beta}}(\log|\mathcal{B}_i|)^{-\frac{1}{\beta}}$ and the ridge parameter $\rho = |\mathcal{B}_i|^{-\frac{7}{24}}$. $K_{max} = \max_{t\in[-1,1]}|\phi_K(t)| < \infty$ is the maximum modulus of the kernel used.

Error thresholds. Our objective in this section is to obtain a probabilistic tail bound on the uniform convergence of $\hat{F}^{(i)}$ to $F^{(i)}$. However, we cannot expect convergence up to arbitrary precision, but there exists a fundamental limit. We define thresholding values for the error in CDF estimation for the convenience in presenting the results. For $i \in [m]$, we let

$$t_0^{(i)} \equiv C_3\left(\log|\mathcal{B}_i|\right)^{-1/\beta} + \frac{2K_{max}(D_2-D_1)}{\pi h}\left(s_\phi(|\mathcal{B}_i|)+\rho\right), \quad \text{and} \tag{55}$$

$$T_0^{(i)} \equiv t_0^{(i)} + \frac{4K_{max}(D_2-D_1)}{\pi(4\gamma)^{\frac{1}{\beta}}}|\mathcal{B}_i|^{-\frac{5}{24}}(\log|\mathcal{B}_i|)^{\frac{1}{\beta}}. \tag{56}$$

Note that these are not constants but functions which depend on $|\mathcal{B}_i|$. We also remark that $C_3\left(\log|\mathcal{B}_i|\right)^{-1/\beta}$ is the essential limit for the convergence, while the other slack terms are introduced for the convenience of analysis.

Recall we defined the following conditioning events (see Eq. (42)) to make the probabilistic tail bound more amenable for the analysis: for any $i \in [m]$,

$$E_{row,(i)} \equiv \left\{|\mathcal{B}_i| \geq \frac{np}{2}\right\}, \quad \text{and} \quad E'_{row,(i)} \equiv \left\{|\mathcal{B}_i| \leq 2np\right\}.$$

We define t_0^* (resp. T_0^*) as the supremum of $t_0^{(i)}$ (resp. $T_0^{(i)}$) under $E_{row,(i)} \cap E'_{row,(i)}$:

$$t_0^* \equiv C_3\left(\log\left(\frac{np}{2}\right)\right)^{-1/\beta} + \frac{2K_{max}(D_2-D_1)}{\pi h}\left(s_\phi(2np)+\rho\right), \quad \text{and} \tag{57}$$

$$T_0^* \equiv t_0^* + \frac{4K_{max}(D_2-D_1)}{\pi(4\gamma)^{\frac{1}{\beta}}}\left(\frac{np}{2}\right)^{-\frac{5}{24}}(\log(2np))^{\frac{1}{\beta}}. \tag{58}$$

Lemma statements. We define a function $\tilde{\Psi}_{m,n,p} : \mathbb{Z}_+ \to \mathbb{R}_+$ as

$$\begin{aligned}\tilde{\Psi}_{m,n,p}(x) = &\exp\left(-\frac{n}{16}\right) + \exp\left(-\frac{m}{16}\right) + \exp\left(-\frac{mnp}{3}\right) \\ &+ n\exp\left(-n^{\frac{1}{2}}\right) + n\exp\left(-\frac{1}{3\sqrt{2}}n^{\frac{3}{4}}\right) + \frac{128}{mnp} \\ &+ \exp\left(-\frac{\sigma^4(\log x)^{\frac{4}{\beta}}}{(4\gamma)^{\frac{4}{\beta}}}\log^2(4mnp) + \log(4mnp)\right) \\ &+ \exp\left(-\frac{(\log x)^{\frac{2}{\beta}}}{256(4\gamma)^{\frac{2}{\beta}}}\left[c_{\Delta A}\sqrt{n} + 2L\sqrt{2m}\right]^2\right. \\ &\qquad \left. + \frac{1}{2}\Big(\log mnp + \log\log(4mnp)\Big) + \log\frac{16\sigma}{c_{\Delta A} + 2L\sqrt{2}}\right). \end{aligned} \tag{59}$$

In the following lemma, we will let $\tilde{\Psi}_{m,n,p}(|\mathcal{B}_i|)$ denote the remainder term which does not depend on the error level t. For completeness, we note that the remainder term is the sum of upper bounds in Eq. (92) - (98), which vanishes as $mp, np \to \infty$. Recall that $C_4 = \frac{BK_{max}(D_2 - D_1)}{\pi(4\gamma)^{\frac{1}{\beta}}}$, where $B \geq 1$.

Lemma 25 *For any $i \in [m]$, and for any $t \geq T_0^{(i)}$,*

$$\mathbb{P}\left(\sup_{z\in[D_1,D_2]}\left|\hat{F}^{(i)}(z) - F^{(i)}(z)\right| > t\right) \leq |\mathcal{B}_i|^{\frac{1}{6}}\exp\left(\frac{-|\mathcal{B}_i|^{5/12}}{8C_4^2(\log|\mathcal{B}_i|)^{\frac{2}{\beta}}}(t - t_0^{(i)})^2\right) + \tilde{\Psi}_{m,n,p}(|\mathcal{B}_i|).$$

We state a useful consequence of the above result with conditioning events $E_{row,(i)}, E'_{row,(i)}$. Note that $\tilde{\Psi}_{m,n,p}\left(\frac{np}{2}\right)$ sets an upper bound on $\tilde{\Psi}_{m,n,p}(|\mathcal{B}_i|)$ under $E_{row,(i)} \cap E'_{row,(i)}$.

Corollary 26 *For any $i \in [m]$, and any $t \geq T_0^*$,*

$$\begin{aligned}&\mathbb{P}\left(\sup_{z\in[D_1,D_2]}\left|\tilde{F}^{(i)}(z) - F^{(i)}(z)\right| > t \,\middle|\, E_{row,(i)}, E'_{row,(i)}\right) \\ &\qquad \leq (2np)^{\frac{1}{6}}\exp\left(\frac{-\left(\frac{np}{2}\right)^{5/12}}{8C_4^2(\log(2np))^{\frac{2}{\beta}}}(t - t_0^*)^2\right) + \tilde{\Psi}_{m,n,p}\left(\frac{np}{2}\right).\end{aligned}$$

E.3. Proof of Theorem 3

In this section, we complete the proof of Theorem 3. The proof follows similar structure as that of Theorem 2. First, we establish tail bound on $|\tilde{A}(i,j) - A(i,j)|$ and then integrate it to obtain bound on Mean-Squared-Error (MSE). The main difference is that we use Lemma 25 (Corollary 26) in place of Lemma 21 due to the lack of knowledge on noise distribution $\phi_N(t)$.

E.3.1. TAIL BOUND ON $|\hat{A}(i,j) - A(i,j)|$

For given choice of parameters $t > 0$ and L, Q^*, m, n, p and T_0^* as defined before , we define conditions in the same manner as in Eq. (60) (we newly define E_3 instead of E_2 there):

$$E_1 = \left\{t \leq 8LQ^*\left(\frac{mp}{2}\right)\right\} \qquad \text{and} \qquad E_3 = \left\{t \leq 2LT_0^*\right\}. \tag{60}$$

Theorem 27 *For each $(i,j) \in [m] \times [n]$, for any $t \geq 0$,*

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{A}(i,j) - A(i,j)\right| > t\right) \\
&\quad \leq \mathbb{I}\{E_1\} + \exp\left(-\frac{n(\frac{t}{4L} - Q^*\left(\frac{mp}{2}\right))}{3}\right)\mathbb{I}\{E_1^c\} \\
&\qquad + \mathbb{I}\{E_3\} + (2np)^{\frac{1}{6}}\exp\left(\frac{-\left(\frac{np}{2}\right)^{5/12}}{8C_4^2\left(\log(2np)\right)^{\frac{2}{\beta}}}(t - t_0^*)^2\right)\mathbb{I}\{E_3^c\} \\
&\qquad + \exp\left(-\frac{nt^2}{8L^2}\right) + \exp\left(-\frac{mp}{8}\right) + 2\exp\left(-\frac{np}{8}\right) + \tilde{\Psi}_{m,n,p}\left(\frac{np}{2}\right),
\end{aligned}$$

where t_0^*, T_0^* and $\tilde{\Psi}_{m,n,p}\left(\frac{np}{2}\right)$ are as defined previously.

Note that the term in the last line, which are independent of t, decays to 0 at an exponential rate as $mp, np \to \infty$.

Proof The proof follows the same logic as in the proof of Theorem 23, while we use the upper bound from Corollary 26 in lieu of Corollary 22. Let $\theta^* \equiv F^{(i)}\left(\hat{A}(i,j)\right) = F^{(i)}\left(\hat{g}^{(i)}\left(\hat{q}_{\text{marg}}(j)\right)\right)$. Since $\hat{F}^{(i)}$ is continuous, $\left|\theta^* - \hat{q}_{\text{marg}}(j)\right| \leq \left\|\hat{F}^{(i)} - F^{(i)}\right\|_\infty$. By the same line of argument as in the proof of Theorem 23, since $\hat{A}(i,j) = \hat{g}^{(i)}\left(\hat{q}_{\text{marg}}(j)\right) = g\left(\theta_{row}^{(i)}, \theta^*\right)$, and g is (l,L)-biLipschitz,

$$\begin{aligned}
\left|\hat{A}(u,i) - A(i,j)\right| &= \left|g\left(\theta_{row}^{(i)}, \theta_{col}^{(j)}\right) - g\left(\theta_{row}^{(i)}, \theta^*\right)\right| \leq L\left|\theta_{col}^{(j)} - \theta^*\right| \\
&\leq L\left(\left|\theta_{col}^{(j)} - \hat{q}_{\text{marg}}(j)\right| + \left|\hat{q}_{\text{marg}}(j) - \theta^*\right|\right) \\
&\leq L\left(\left|\theta_{col}^{(j)} - \hat{q}_{\text{marg}}(j)\right| + \left\|\hat{F}^{(i)} - F^{(i)}\right\|_\infty\right).
\end{aligned}$$

If $\left|\theta_{col}^{(j)} - \hat{q}_{\text{marg}}(j)\right| \leq \frac{t}{2L}$, $\left\|\hat{F}^{(i)} - F^{(i)}\right\|_\infty \leq \frac{t}{2L}$ then $\left|\hat{A}(u,i) - A(i,j)\right| \leq t$. We can achieve the following upper bound by applying the union bound on the contraposition. We let $E_{(i)} := E_{row,(i)} \cap E'_{row,(i)}$ in this proof. Then it follows that

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{A}(i,j) - A(i,j)\right| > t\right) \\
&\quad \leq \mathbb{P}\left(\left|\hat{q}_{\text{marg}}(j) - \theta_{col}^{(j)}\right| > \frac{t}{2L}\right) + \mathbb{P}\left(\sup_{z\in\mathbb{R}}\left|\hat{F}^{(i)}(z) - F^{(i)}(z)\right| > \frac{t}{2L}\right) \\
&\quad \leq \mathbb{P}\left(\left|\hat{q}_{\text{marg}}(j) - \theta_{col}^{(j)}\right| > \frac{t}{2L}\right) \\
&\qquad + \mathbb{P}\left(\sup_{z\in\mathbb{R}}\left|\hat{F}^{(i)}(z) - F^{(i)}(z)\right| > \frac{t}{2L}\middle| E_{(i)}\right) + \mathbb{P}\left(E_{(i)}^c\right).
\end{aligned}$$

Because we have a trivial upper bound 1 on probability, it follows from Lemma 20 that

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{q}_{\text{marg}}(j) - \theta^{(j)}_{col}\right| > \frac{t}{2L}\right) \\
&\quad \leq \mathbb{I}\left\{t \leq 8LQ^*\left(\frac{mp}{2}\right)\right\} \\
&\qquad + \mathbb{I}\left\{t \geq 8LQ^*\left(\frac{mp}{2}\right)\right\} \\
&\qquad\quad \times \left[\exp\left(-\frac{nt^2}{8L^2}\right) + \exp\left(-\frac{n(\frac{t}{4L} - Q^*\left(\frac{mp}{2}\right))}{3}\right) + \exp\left(-\frac{mp}{8}\right)\right].
\end{aligned}$$

In a similar manner, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{z\in\mathbb{R}}\left|\hat{F}^{(i)}(z) - F^{(i)}(z)\right| > \frac{t}{2L}\,\middle|\, E_{(i)}\right) \\
&\quad \leq \mathbb{I}\left\{t \leq 2LT_0^*\right\} \\
&\qquad + \mathbb{I}\left\{t \geq 2LT_0^*\right\}(2np)^{\frac{1}{6}}\exp\left(\frac{-\left(\frac{np}{2}\right)^{5/12}}{8C_4^2\left(\log(2np)\right)^{\frac{2}{\beta}}}(t - t_0^*)^2\right) \\
&\qquad + \mathbb{I}\left\{t \geq 2LT_0^*\right\}\tilde{\Psi}_{m,n,p}\left(\frac{np}{2}\right).
\end{aligned}$$

Note that $t \geq 2LT_0^*$ implies that $\frac{t}{2L} \geq t_0^*$.

We used an upper bound on $\mathbb{P}\left(E^c_{(i)}\right)$ obtained from the binomial Chernoff bound:

$$\begin{aligned}
\mathbb{P}\left(E^c_{(i)}\right) &= \mathbb{P}\left(|\mathcal{B}_i| < \frac{np}{2} \text{ or } |\mathcal{B}_i| > 2np\right) \\
&\leq \mathbb{P}\left(|\mathcal{B}_i| < \frac{np}{2}\right) + \mathbb{P}\left(|\mathcal{B}_i| > 2np\right) \\
&\leq \exp\left(-\frac{np}{8}\right) + \exp\left(-\frac{np}{3}\right) \\
&\leq 2\exp\left(-\frac{np}{8}\right).
\end{aligned}$$

Substituting these three upper bounds back to Eq. (44), we can conclude that

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{A}(i,j) - A(i,j)\right| > t\right) \\
&\quad \leq \mathbb{I}\left\{t \leq 8LQ^*\left(\frac{mp}{2}\right)\right\} + \mathbb{I}\left\{t \leq 2LT_0^*\right\} \\
&\qquad + \exp\left(-\frac{n(\frac{t}{4L} - Q^*\left(\frac{mp}{2}\right))}{3}\right)\mathbb{I}\left\{t \geq 8LQ^*\left(\frac{mp}{2}\right)\right\} \\
&\qquad + \mathbb{I}\left\{t \geq 2LT_0^*\right\}(2np)^{\frac{1}{6}}\exp\left(\frac{-\pi^2\left(4\gamma\right)^{\frac{2}{\beta}}\left(\frac{np}{2}\right)^{5/12}}{8K_{max}^2(D_2 - D_1)^2\left(\log(2np)\right)^{\frac{2}{\beta}}}(t - t_0^*)^2\right) \\
&\qquad + \exp\left(-\frac{nt^2}{8L^2}\right) + \exp\left(-\frac{mp}{8}\right) + 2\exp\left(-\frac{np}{8}\right) + \tilde{\Psi}_{m,n,p}\left(\frac{np}{2}\right).
\end{aligned}$$

■

E.3.2. MEAN SQUARED ERROR

Let $\hat{\varphi}$ denote the estimator which maps Z to $\hat{A}$. By the same line of arguments as in Eq. (28), the mean squared error of estimator $\hat{\varphi}$ is given as

$$MSE\left(\hat{\varphi}\right) = \int_0^\infty 2u\mathbb{P}\left(\left|\hat{A}(i,j) - A(i,j)\right| > u\right) du \tag{61}$$

Also, from the model assumption and the construction of the estimators, the estimation error is bounded above:

$$\left|\hat{A}(i,j) - A(i,j)\right| \leq D_2 - D_1,$$

Let $D = D_2 - D_1$ denote the upper bound. Note that D is a constant independent of m, n.

For brevity's sake, we introduce some notations for abbreviation. We let

$$c_3 \equiv \frac{\left(\frac{np}{2}\right)^{5/12}}{8C_4^2\left(\log(2np)\right)^{\frac{2}{\beta}}}.$$

We define $\Psi(m,n,p)$ to capture all constant terms in the probabilistic bound of Theorem 27. That is to say,

$$\begin{aligned}
\Psi(m,n,p) &\equiv \exp\left(-\frac{mp}{8}\right) + 2\exp\left(-\frac{np}{8}\right) + \tilde{\Psi}_{m,n,p}\left(\frac{np}{2}\right) \\
&= \exp\left(-\frac{mp}{8}\right) + 2\exp\left(-\frac{np}{8}\right) \\
&\quad + \exp\left(-\frac{n}{16}\right) + \exp\left(-\frac{m}{16}\right) + \exp\left(-\frac{mnp}{3}\right) \\
&\quad + n\exp\left(-n^{\frac{1}{2}}\right) + n\exp\left(-\frac{1}{3\sqrt{2}}n^{\frac{3}{4}}\right) + \frac{128}{mnp} \\
&\quad + \exp\left(-\frac{\sigma^4(\log\frac{np}{2})^{\frac{4}{\beta}}}{(4\gamma)^{\frac{4}{\beta}}}\log^2(4mnp) + \log(4mnp)\right) \\
&\quad + \exp\left(-\frac{(\log\frac{np}{2})^{\frac{2}{\beta}}}{256(4\gamma)^{\frac{2}{\beta}}}\left[c_{\Delta A}\sqrt{n} + 2L\sqrt{2m}\right]^2\right. \\
&\qquad \left. + \frac{1}{2}\Big(\log mnp + \log\log(4mnp)\Big) + \log\frac{16\sigma}{c_{\Delta A} + 2L\sqrt{2}}\right).
\end{aligned} \tag{62, 63}$$

Theorem 28 (The Full Version of Main theorem 3; MSE with unknown noise) *The mean squared error of the deconvolution kernel estimator $\hat{\varphi}$ is bounded above as follows:*

$$\begin{aligned}
MSE\left(\hat{\varphi}\right) &\leq 4L^2T_0^{*2} + 64L^2Q^*\left(\frac{mp}{2}\right)^2 \\
&\quad + 8LQ^*\left(\frac{mp}{2}\right)\sqrt{\frac{3L\pi}{n}} + 4L^2(2np)^{\frac{1}{6}}\left[\frac{1}{c_3} + t_0^*\sqrt{\frac{\pi}{c_3}}\right] \\
&\quad + \frac{8L^2}{n} + \frac{288L^2}{n^2} + D^2\Psi(m,n,p).
\end{aligned}$$

The upper bound diminishes to 0 *as* $mp, np \to \infty$ *at the rate of* $(\log np)^{-\frac{2}{\beta}}$.

We remark that $4L^2T_0^{*2}$ is the asymptotically dominant term, which scales as $O\left((\log np)^{-\frac{2}{\beta}}\right)$ (see Eq. (56) for definition of T_0^*). All the other terms decay at polynomial rate in the least. For example, $Q^*\left(\frac{mp}{2}\right) = O\left(\frac{1}{\sqrt{mp}}\right)$ (see Eq. (41)). To see the polynomial convergence of $4L^2(2np)^{\frac{1}{6}}\left[\frac{1}{c_3} + t_0^*\sqrt{\frac{\pi}{c_3}}\right]$, recall from Eqs. (54) and (56) that

$$t_0^* \equiv C_3\left(\log\left(\frac{np}{2}\right)\right)^{-1/\beta} + \frac{2K_{max}(D_2 - D_1)}{\pi h}\Big(s_\phi(2np) + \rho\Big), \quad \text{where}$$

$$s_\phi(2np) = \frac{8\sigma(\log(2np))^{\frac{1}{\beta}}}{(4\gamma)^{\frac{1}{\beta}}}\frac{\sqrt{\log(4mnp)}}{(mnp)^{\frac{1}{4}}} + \frac{2(\log(2np))^{\frac{1}{\beta}}}{(4\gamma)^{\frac{1}{\beta}}}\left[\frac{c_{\Delta A}}{\sqrt{mp}} + \frac{2L\sqrt{2}}{\sqrt{np}}(1 + \sqrt[4]{np})\right].$$

We can see that $\left[\frac{1}{c_3} + t_0^*\sqrt{\frac{\pi}{c_3}}\right] = O\left((np)^{-\frac{5}{24}}\right)$ because $t_0^*\sqrt{\frac{\pi}{c_3}}$ dominates asymptotically.

Proof [Proof of Theorem 28] In order to achieve an upper bound on the MSE for the kernel density estimator with known noise, $\hat{\varphi}$, we integrate the tail probability bound from Theorem 27.

First of all, we recall from Eqs. (29) and (30) that

$$\int_0^\infty ue^{-au^2}du = \frac{1}{2a}, \quad \text{and} \quad \int_0^\infty ue^{-au}du = \frac{1}{a^2}.$$

Also, we know that

$$\int_0^\infty e^{-au^2}du = \frac{1}{2}\sqrt{\frac{\pi}{a}}. \tag{64}$$

Now, the mean squared error can be written in the following form:

$$\begin{aligned}
MSE\left(\hat{\varphi}\right) &= \int_0^D 2u\mathbb{P}\left(\left|\hat{A}(i,j) - A(i,j)\right| > u\right)du \\
&\leq \int_0^D 2u\Psi(m,n,p)du + \int_0^{8LQ^*\left(\frac{mp}{2}\right)} 2udu + \int_0^{2LT_0^*} 2udu \\
&\quad + \int_0^D 2u\exp\left(-\frac{nu^2}{8L^2}\right)du \qquad (65) \\
&\quad + \int_{8LQ^*\left(\frac{mp}{2}\right)}^D 2u\exp\left(-\frac{n(\frac{u}{4L} - Q^*\left(\frac{mp}{2}\right))}{3}\right)du \qquad (66) \\
&\quad + \int_{2LT_0^*}^D 2u(2np)^{\frac{1}{6}}\exp\left(\frac{-\left(\frac{np}{2}\right)^{5/12}}{8C_4^2\left(\log(2np)\right)^{\frac{2}{\beta}}}\left(\frac{u}{2L} - t_0^*\right)^2\right)du. \qquad (67)
\end{aligned}$$

We can reuse some calculations from the proof of Theorem 24. Note that the term in Eq. 65 is the same with that in Eq. (46), and Eq. (66) is the same with Eq. (47). Therefore,

$$\begin{aligned}
Eq.(65) &\leq \int_0^\infty 2u\exp\left(-\frac{nu^2}{8L^2}\right)du = \frac{8L^2}{n}, \\
Eq.(66) &\leq \frac{288L^2}{n^2} + 8LQ^*\left(\frac{mp}{2}\right)\sqrt{\frac{3L\pi}{n}}.
\end{aligned}$$

It remains to compute an upper bound of the term Eq. (67).

$$\begin{aligned} Eq.(67) &= 2(2np)^{\frac{1}{6}} \int_{2LT_0^*}^{D} u \exp\left(-c_3\left(\frac{u}{2L} - t_0^*\right)^2\right) du \\ &= 2(2np)^{\frac{1}{6}} \int_{T_0^* - t_0^*}^{\frac{D}{2L} - t_0^*} (2L)^2 (v + t_0^*) \exp\left(-c_3 v^2\right) dv \\ &\leq 8L^2(2np)^{\frac{1}{6}} \left[\int_0^{\infty} v \exp\left(-c_3 v^2\right) dv + t_0^* \int_0^{\infty} \exp\left(-c_3 v^2\right) dv\right] \\ &= 4L^2(2np)^{\frac{1}{6}} \left[\frac{1}{c_3} + t_0^* \sqrt{\frac{\pi}{c_3}}\right]. \end{aligned}$$

The second line follows by substituting $v = \frac{u}{2L} - t_0^*$ and the third line follows from that $T_0^* - t_0^* \geq 0$.

Plugging these upper bounds back into Eqs. (46), (47) and (48) , we can obtain the following upper bound

$$\begin{aligned} MSE\left(\hat{\varphi}\right) \leq D^2 \Psi(m,n,p) + \left[8LQ^*\left(\frac{mp}{2}\right)\right]^2 + \left[2LT_0^*\right]^2 + \frac{8L^2}{n} + \frac{288L^2}{n^2} \\ + 8LQ^*\left(\frac{mp}{2}\right)\sqrt{\frac{3L\pi}{n}} + 4L^2(2np)^{\frac{1}{6}}\left[\frac{1}{c_3} + t_0^*\sqrt{\frac{\pi}{c_3}}\right]. \end{aligned}$$

Rearranging the terms in the increasing order of convergence rates concludes the proof. ■

Appendix F. Proof of Lemma 15

Proof [Proof of Lemma 15] Recall from Eq. (21) that the quantile of j estimated from row i is a function of $|\mathcal{B}_i| = \sum_{j'=1}^{n} M(i,j')$ many independent random variables, $H\big(Z(i,j) - Z(i,j')\big)$:

$$\hat{q}_i(j) = \frac{\sum_{j'=1}^{n} M(i,j') H\big(Z(i,j) - Z(i,j')\big)}{\sum_{j'=1}^{n} M(i,j')}.$$

Since $H\big(Z(i,j_1) - Z(i,j_2)\big)$ takes value in $\{0, \frac{1}{2}, 1\}$, it satisfies the bounded difference condition. To be more specific, let's consider a perturbation on the column feature associated with one index. For any $j_0 \in [n]$, if $j_0 \in \mathcal{B}_i$ (i.e., if $M(i,j_0) = 1$), then

$$\left|\hat{q}_i(j)\big|_{\theta_{col}^{(j_0)}=a} - \hat{q}_i(j)\big|_{\theta_{col}^{(j_0)}=b}\right| \leq \frac{1}{|\mathcal{B}_i|},$$

for any value $a, b \in [0,1]$, while if $j_0 \notin \mathcal{B}_i$ (i.e., if $M(i,j_0) = 0$), then obviously

$$\left|\hat{q}_i(j)\big|_{\theta_{col}^{(j_0)}=a} - \hat{q}_i(j)\big|_{\theta_{col}^{(j_0)}=b}\right| = 0.$$

Since $\mathbb{E}\left[\hat{q}_i(j)\right] = \theta_{col}^{(j)}$, we can achieve the following probabilistic tail bound by an application of McDiarmid's inequality

$$\mathbb{P}\left(\left|\hat{q}_i(j) - \theta_{col}^{(j)}\right| \geq t\right) \leq 2\exp\left(-2|\mathcal{B}_i|t^2\right).$$

■

Appendix G. Proof of Lemma 20

When there is nontrivial noise present, the indicator may not be reliable any more. Hence, we need a way to control the effect of noise. We assume the additive noise is sub-Gaussian.

In addition to condition defined in (42), we will use the following notation.

$$E_{col,(j)} \equiv \left\{ |\mathcal{B}^j| \geq \frac{mp}{2} \right\}. \tag{68}$$

Proof [Proof of Lemma 20] Recall from section D.1 (see Eqs. (32) and (31)) that the quantile estimator is defined as

$$\hat{q}_{\text{marg}}(j) = \frac{1}{n} \sum_{j'=1}^{n} H\left(Z_{\text{marg}}(j) - Z_{\text{marg}}(j') \right),$$

where

$$Z_{\text{marg}}(j) = \begin{cases} \frac{\sum_{i=1}^{m} M(i,j) Z(i,j)}{\sum_{i=1}^{m} M(i,j)}, & \text{if } \mathcal{B}^j \neq \emptyset \\ \frac{1}{2}, & \text{if } \mathcal{B}^j = \emptyset. \end{cases}$$

We also note that since the marginalization of the latent function $g_{\text{marg}}(y) := \int_0^1 g(x,y)dx$ is strictly increasing and (l, L)-biLipschitz, hence, invertible. We let $\zeta^{(j)} = g_{\text{marg}}^{-1}\left(Z_{\text{marg}}(j)\right)$ for the purpose of analysis. We also define an imaginary estimator

$$\hat{q}_*(j) = \frac{1}{n} \sum_{j'=1}^{n} H\left(\theta_{col}^{(j)} - \theta_{col}^{(j')} \right),$$

which will be used solely for analysis.

By triangle inequality, the error in quantile estimation is upper bounded as

$$\left| \hat{q}_{\text{marg}}(j) - \theta_{col}^{(j)} \right| \leq \left| \hat{q}_{\text{marg}}(j) - \hat{q}_*(j) \right| + \left| \hat{q}_*(j) - \theta_{col}^{(j)} \right|.$$

If both $\left| \hat{q}_{\text{marg}}(j) - \hat{q}_*(j) \right| \leq t_1$ and $\left| \hat{q}_*(j) - \theta_{col}^{(j)} \right| \leq t_2$ are satisfied, then $\left| \hat{q}_{\text{marg}}(j) - \theta_{col}^{(j)} \right| \leq t_1 + t_2$. Therefore, for any $t_1, t_2 > 0$ and $t = t_1 + t_2$,

$$\begin{aligned} &\mathbb{P}\left(\left| \hat{q}_{\text{marg}}(j) - \theta_{col}^{(j)} \right| > t \right) \\ &\quad \leq \mathbb{P}\left(\left| \hat{q}_{\text{marg}}(j) - \hat{q}_*(j) \right| > t_1 \right) + \mathbb{P}\left(\left| \hat{q}_*(j) - \theta_{col}^{(j)} \right| > t_2 \right). \end{aligned} \tag{69}$$

Note that $\hat{q}_*(j)$ exponentially concentrates to $\theta_{col}^{(j)}$ as $n \to \infty$ by McDiarmid's inequality, for example. Therefore, it suffices to find a probabilistic tail upper bound for $\left| \hat{q}_{\text{marg}}(j) - \hat{q}_*(j) \right|$:

$$\begin{aligned} \left| \hat{q}_{\text{marg}}(j) - \hat{q}_*(j) \right| &= \left| \frac{1}{n} \sum_{j'=1}^{n} \left[H\left(Z_{\text{marg}}(j) - Z_{\text{marg}}(j') \right) - H\left(\theta_{col}^{(j)} - \theta_{col}^{(j')} \right) \right] \right| \\ &\leq \frac{1}{n} \sum_{j'=1}^{n} \left| \left[H\left(Z_{\text{marg}}(j) - Z_{\text{marg}}(j') \right) - H\left(\theta_{col}^{(j)} - \theta_{col}^{(j')} \right) \right] \right|. \end{aligned}$$

For $j' \neq j$, $\left|\left[H\left(Z_{\text{marg}}(j) - Z_{\text{marg}}(j')\right) - H\left(\theta_{col}^{(j)} - \theta_{col}^{(j')}\right)\right]\right| = 1$ with probability p_{fail}, and 0 otherwise (it is uniformly 0 for $j' = j$). Now, if we can find an upper bound $p_{\text{fail}}^* \geq p_{\text{fail}}$, then for $t > p_{\text{fail}}^*$,

$$\mathbb{P}\Big(\left|\hat{q}_{\text{marg}}(j) - \hat{q}_*(j)\right| > t\Big) \leq \mathbb{P}\Big(Y > nt\Big) \leq \exp\left(-\frac{n(t - p_{\text{fail}}^*)^2}{t + p_{\text{fail}}^*}\right),$$

where $Y \sim Binomial(n, p_{\text{fail}}^*)$.

We define a monotone decreasing function $Q^* : \mathbb{Z}_+ \to \mathbb{R}_+$ as

$$Q^*(x) = 2\sqrt{\pi}\left(\frac{1}{\sqrt{C_1 x}} + \frac{1}{\sqrt{C_2 x}} + \frac{1}{\sqrt{mpC_1 e^{-C_1}}} + \frac{1}{\sqrt{mpC_2 e^{-C_2}}}\right),$$

where $C_1 = \frac{l^2}{2(D_{max} - D_{min})^2}$ and $C_2 = \frac{l^2}{8\sigma^2}$ are some model-dependent constants.

Claim 1. We show that $p_{\text{fail}} \leq Q^*\left(\left|\mathcal{B}^j\right|\right)$, i.e., p_{fail} is bounded above by a function of the number of revealed entries on column j, $|\mathcal{B}^j|$.

The estimator $\hat{q}_{\text{marg}}(j)$ exploits the pairwise ordering information of column pair (j, j') by taking the sign of $Z_{\text{marg}}(j) - Z_{\text{marg}}(j')$, which might be different from the true ordering $sign\left(\theta_{col}^{(j)} - \theta_{col}^{(j')}\right)$ due to the presence of noise. We analyze the probability of the order to be distrubed. Note that $sign\Big(Z_{\text{marg}}(j) - Z_{\text{marg}}(j')\Big) = sign\left(\zeta^{(j)} - \zeta^{(j')}\right)$ because g_{marg} is strictly monotone increasing.

Let $X_j := \zeta^{(j)} - \theta_{col}^{(j)}$. Then since g_{marg} is (l, L)-biLipschitz, for any $s > 0$,

$$\begin{aligned}
\mathbb{P}\left(X_j \geq s\right) &\leq \mathbb{P}\left(Z_{\text{marg}}(j) - g_{\text{marg}}\left(\theta_{col}^{(j)}\right) \geq ls\right) \\
&= \mathbb{P}\left(\frac{1}{|\mathcal{B}^j|}\sum_{i' \in \mathcal{B}^j} Z(i', j) - g_{\text{marg}}\left(\theta_{col}^{(j)}\right) \geq ls\right) \\
&\leq \mathbb{P}\left(\frac{1}{|\mathcal{B}^j|}\sum_{i' \in \mathcal{B}^j} A(i', j) - g_{\text{marg}}\left(\theta_{col}^{(j)}\right) \geq \frac{ls}{2}\right) \\
&\quad + \mathbb{P}\left(\frac{1}{|\mathcal{B}^j|}\sum_{i' \in \mathcal{B}^j} N(i', j) \geq \frac{ls}{2}\right) \\
&\leq \exp\left(-\frac{\left|\mathcal{B}^j\right| l^2 s^2}{2(D_{max} - D_{min})^2}\right) + \exp\left(-\frac{\left|\mathcal{B}^j\right| l^2 s^2}{8\sigma^2}\right).
\end{aligned}$$

For the brevity, we will let $C_1 = \frac{l^2}{2(D_{max} - D_{min})^2}$ and $C_2 = \frac{l^2}{8\sigma^2}$ throughout the rest of the proof. Also, we can achieve the same upper bound for $\mathbb{P}\left(X_j \leq -s\right)$. Since $X_j - X_{j'} = \left(\zeta^{(j)} - \zeta^{(j')}\right) - \left(\theta_{col}^{(j)} - \theta_{col}^{(j')}\right)$, the pairwise order is conserved unless

$$\begin{cases} X_j - X_{j'} < -\left(\theta_{col}^{(j)} - \theta_{col}^{(j')}\right), & \text{when } \theta_{col}^{(j)} - \theta_{col}^{(j')} \geq 0, \\ X_j - X_{j'} > \theta_{col}^{(j')} - \theta_{col}^{(j)}, & \text{when } \theta_{col}^{(j)} - \theta_{col}^{(j')} < 0. \end{cases}$$

Given $\theta^{(j)}_{col}$, the probability of $\theta^{(j')}_{col}$ be smaller than $\theta^{(j)}_{col}$ is equal to $\theta^{(j)}_{col}$, i.e., $\mathbb{P}\left(\theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0\right) = \theta^{(j)}_{col}$. Therefore, the probability of the problematic event can be partitioned as

$$\begin{aligned}
&\mathbb{P}\left(sign\left(\zeta^{(j)} - \zeta^{(j')}\right) \neq sign\left(\theta^{(j)}_{col} - \theta^{(j')}_{col}\right)\right) \\
&\quad = \mathbb{P}\left(X_j - X_{j'} < -\left(\theta^{(j)}_{col} - \theta^{(j')}_{col}\right) \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0\right) \mathbb{P}\left(\theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0\right) \\
&\qquad + \mathbb{P}\left(X_j - X_{j'} > \theta^{(j')}_{col} - \theta^{(j)}_{col} \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} < 0\right) \mathbb{P}\left(\theta^{(j)}_{col} - \theta^{(j')}_{col} < 0\right).
\end{aligned}$$

The first conditional probability can be upper bounded by

$$\begin{aligned}
&\mathbb{P}\left(X_j - X_{j'} < -\left(\theta^{(j)}_{col} - \theta^{(j')}_{col}\right) \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0\right) \\
&\quad \leq \mathbb{P}\left(X_j < -\frac{\theta^{(j)}_{col} - \theta^{(j')}_{col}}{2} \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0\right) \\
&\qquad + \mathbb{P}\left(X_{j'} > \frac{\theta^{(j)}_{col} - \theta^{(j')}_{col}}{2} \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0\right).
\end{aligned}$$

Meanwhile, if we define a new random variable $T := \frac{\theta^{(j)}_{col} - \theta^{(j')}_{col}}{2}$ and let τ denote its realization, we can see that $f_T(\tau) = \frac{2}{\theta^{(j)}_{col}} \mathbb{I}\left\{0 \leq T \leq \frac{\theta^{(j)}_{col}}{2}\right\}$, conditioned on $\theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0$.

$$\begin{aligned}
&\mathbb{P}\left(X_j < -\tau \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0\right) \\
&\quad = \sum_{k=0}^{m} \mathbb{P}\left(|\mathcal{B}^j| = k\right) \mathbb{P}\left(X_j < -\tau \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0, |\mathcal{B}^j| = k\right) \\
&\quad \leq \sum_{k=0}^{m} \binom{m}{k} p^k (1-p)^{m-k} \left[\exp\left(-C_1 k \tau^2\right) + \exp\left(-C_2 k \tau^2\right)\right] \\
&\quad = \left[p e^{-C_1 \tau^2} + (1-p)\right]^m + \left[p e^{-C_2 \tau^2} + (1-p)\right]^m \\
&\quad = \left[1 - \frac{mp\left(1 - e^{-C_1 \tau^2}\right)}{m}\right]^m + \left[1 - \frac{mp\left(1 - e^{-C_2 \tau^2}\right)}{m}\right]^m \\
&\quad \leq \exp\left[-mp\left(1 - e^{-C_1 \tau^2}\right)\right] + \exp\left[-mp\left(1 - e^{-C_2 \tau^2}\right)\right].
\end{aligned}$$

As a result,

$$\begin{aligned}
&\mathbb{P}\left(sign\left(\zeta^{(j)} - \zeta^{(j')}\right) \neq sign\left(\theta^{(j)}_{col} - \theta^{(j')}_{col}\right) \middle| |\mathcal{B}^j| = k\right) \\
&\quad = \mathbb{P}\left(\theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0\right) \qquad (70) \\
&\qquad \times \mathbb{P}\left(X_j - X_{j'} < -\left(\theta^{(j)}_{col} - \theta^{(j')}_{col}\right) \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0, |\mathcal{B}^j| = k\right) \\
&\qquad + \mathbb{P}\left(\theta^{(j)}_{col} - \theta^{(j')}_{col} < 0\right) \qquad (71) \\
&\qquad \times \mathbb{P}\left(X_j - X_{j'} > \theta^{(j')}_{col} - \theta^{(j)}_{col} \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} < 0, |\mathcal{B}^j| = k\right).
\end{aligned}$$

Note that $X_j < -\tau$ and $X_{j'} > \tau$ implies $X_j - X_{j'} < -2\tau$ for any $\tau \in \mathbb{R}$. Therefore, for any $\tau \in \mathbb{R}$, it follows that $\mathbb{P}\left(X_j - X_{j'} < -2\tau\right) \leq \mathbb{P}\left(X_j < -\tau\right) + \mathbb{P}\left(X_{j'} > \tau\right)$. Now we will obtain an upper bound on Eq. (70) by finding upper bounds on each terms and then taking the union bound. Note that $\frac{d\mathbb{P}\left(\theta^{(j)}_{col} - \theta^{(j')}_{col} = 2\tau \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0\right)}{d\tau} = \frac{2}{\theta^{(j)}_{col}} \mathbb{I}\left\{0 \leq \tau \leq \frac{\theta^{(j)}_{col}}{2}\right\}$ and $\mathbb{P}\left(\theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0\right) = \theta^{(j)}_{col}$.

$$
\begin{aligned}
&\mathbb{P}\left(X_j < -\frac{\theta^{(j)}_{col} - \theta^{(j')}_{col}}{2} \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0, |\mathcal{B}^j| = k\right) \mathbb{P}\left(\theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0\right) \\
&\quad = \int_{\tau} \mathbb{P}\left(X_j < -\tau \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} = 2\tau, |\mathcal{B}^j| = k\right) \times \\
&\qquad \frac{d\mathbb{P}\left(\theta^{(j)}_{col} - \theta^{(j')}_{col} = 2\tau \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0\right)}{d\tau} \mathbb{P}\left(\theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0\right) d\tau \\
&\quad = 2 \int_0^{\frac{\theta^{(j)}_{col}}{2}} \mathbb{P}\left(X_j < -\tau \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} = 2\tau, |\mathcal{B}^j| = k\right) d\tau \\
&\quad \leq 2 \int_0^{\frac{\theta^{(j)}_{col}}{2}} \exp\left(-C_1 k \tau^2\right) + \exp\left(-C_2 k \tau^2\right) d\tau \\
&\quad \leq 2 \int_0^{\infty} \exp\left(-C_1 k \tau^2\right) + \exp\left(-C_2 k \tau^2\right) d\tau \\
&\quad = \sqrt{\pi} \left(\frac{1}{\sqrt{C_1 k}} + \frac{1}{\sqrt{C_2 k}}\right).
\end{aligned}
$$

Similarly, we can obtain an upper bound for $X_{j'}$. Note that column j and j' are independent and that for $c > 0$, $1 - e^{-cu^2} \geq ce^{-c}u^2$, $\forall u \in [0,1]$.

$$\begin{aligned}
&\mathbb{P}\left(X_{j'} > \frac{\theta^{(j)}_{col} - \theta^{(j')}_{col}}{2} \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0, |\mathcal{B}^j| = k \right) \mathbb{P}\left(\theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0 \right) \\
&= \int_\tau \mathbb{P}\left(X_{j'} > \tau \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} = 2\tau, |\mathcal{B}^j| = k \right) \times \\
&\quad \frac{d\mathbb{P}\left(\theta^{(j)}_{col} - \theta^{(j')}_{col} = 2\tau \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0 \right)}{d\tau} \mathbb{P}\left(\theta^{(j)}_{col} - \theta^{(j')}_{col} \geq 0 \right) d\tau \\
&= 2\int_0^{\frac{\theta^{(j)}_{col}}{2}} \mathbb{P}\left(X_{j'} > \tau \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} = 2\tau, |\mathcal{B}^j| = k \right) d\tau \\
&= 2\int_0^{\frac{\theta^{(j)}_{col}}{2}} \mathbb{P}\left(X_{j'} > \tau \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} = 2\tau \right) d\tau \\
&\leq 2\int_0^{\frac{\theta^{(j)}_{col}}{2}} \exp\left[-mp\left(1 - e^{-C_1\tau^2} \right) \right] + \exp\left[-mp\left(1 - e^{-C_2\tau^2} \right) \right] d\tau \\
&\leq 2\int_0^{\frac{\theta^{(j)}_{col}}{2}} \exp\left(-mpC_1 e^{-C_1}\tau^2 \right) + \exp\left(-mpC_2 e^{-C_2}\tau^2 \right) d\tau \\
&\leq 2\int_0^{\infty} \exp\left(-mpC_1 e^{-C_1}\tau^2 \right) + \exp\left(-mpC_2 e - C_2\tau^2 \right) d\tau \\
&= \sqrt{\pi}\left(\frac{1}{\sqrt{mpC_1 e^{-C_1}}} + \frac{1}{\sqrt{mpC_2 e^{-C_2}}} \right).
\end{aligned}$$

We used the fact (see Eq. (64) that

$$\int_0^\infty e^{-ax^2} dx = \frac{1}{2}\sqrt{\frac{\pi}{a}}.$$

From these, we can conclude that

$$Eq.(70) \leq \sqrt{\pi}\left(\frac{1}{\sqrt{C_1 k}} + \frac{1}{\sqrt{C_2 k}} + \frac{1}{\sqrt{mpC_1 e^{-C_1}}} + \frac{1}{\sqrt{mpC_2 e^{-C_2}}} \right).$$

In the same vein, a similar upper bound can be derived for Eq. (71). It suffices to remark that

$$\frac{d\mathbb{P}\left(\theta^{(j)}_{col} - \theta^{(j')}_{col} = -2\tau \middle| \theta^{(j)}_{col} - \theta^{(j')}_{col} < 0 \right)}{d\tau} = \frac{2}{1 - \theta^{(j)}_{col}} \mathbb{I}\left\{ 0 \leq \tau \leq \frac{1 - \theta^{(j)}_{col}}{2} \right\},$$

$$\mathbb{P}\left(\theta^{(j)}_{col} - \theta^{(j')}_{col} < 0 \right) = 1 - \theta^{(j)}_{col}.$$

Then by the same logic,

$$Eq.(71) \leq \sqrt{\pi}\left(\frac{1}{\sqrt{C_1 k}} + \frac{1}{\sqrt{C_2 k}} + \frac{1}{\sqrt{mpC_1 e^{-C_1}}} + \frac{1}{\sqrt{mpC_2 e^{-C_2}}} \right).$$

Consequently, we can conclude our claim 1:

$$\begin{aligned} q &= \mathbb{P}\left(sign\left(\zeta^{(j)} - \zeta^{(j')}\right) \neq sign\left(\theta_{col}^{(j)} - \theta_{col}^{(j')}\right)\right) \\ &\leq 2\sqrt{\pi}\left(\frac{1}{\sqrt{C_1\left|\mathcal{B}^j\right|}} + \frac{1}{\sqrt{C_2\left|\mathcal{B}^j\right|}} + \frac{1}{\sqrt{mpC_1e^{-C_1}}} + \frac{1}{\sqrt{mpC_2e^{-C_2}}}\right) \\ &=: Q^*\left(\left|\mathcal{B}^j\right|\right). \end{aligned}$$

Claim 2. Next, we can observe that for $t \geq Q^*\left(\frac{mp}{2}\right)$,

$$\mathbb{P}\left(\left|\hat{q}_{\text{marg}}(j) - \hat{q}_*(j)\right| > t\right) \leq \exp\left(-\frac{n(t - p^*_{\text{fail}})}{3}\right)\Big|_{p^*_{\text{fail}} = Q^*\left(\frac{mp}{2}\right)} + \exp\left(-\frac{mp}{8}\right).$$

It follows from the usual union bound trick with conditioning event $E_{col,(j)}$ (see Eq. (68)) :

$$\begin{aligned} &\mathbb{P}\left(\left|\hat{q}_{\text{marg}}(j) - \hat{q}_*(j)\right| > t\right) \\ &\quad\leq \mathbb{P}\left(Y > nt\right) \\ &\quad= \mathbb{P}\left(Y > nt \,\middle|\, E_{col,(j)}\right)\mathbb{P}\left(E_{col,(j)}\right) + \mathbb{P}\left(Y > nt \,\middle|\, E^c_{col,(j)}\right)\mathbb{P}\left(E^c_{col,(j)}\right) \\ &\quad\leq \mathbb{P}\left(Y > nt \,\middle|\, E_{col,(j)}\right) + \mathbb{P}\left(E^c_{col,(j)}\right) \\ &\quad\leq \exp\left(-\frac{n(t - p^*_{\text{fail}})^2}{t + p^*_{\text{fail}}}\right)\Big|_{p^*_{\text{fail}} = Q^*\left(\frac{mp}{2}\right)} + \exp\left(-\frac{mp}{8}\right). \end{aligned}$$

We respectively used the fact that Q^* is monotone decreasing and the Binomial Chernoff bound to bound the terms.

For $t \geq 2p^*_{\text{fail}}$, $\frac{t - p^*_{\text{fail}}}{t + p^*_{\text{fail}}} \geq \frac{1}{3}$ and hence,

$$\mathbb{P}\left(\left|\hat{q}_{\text{marg}}(j) - \hat{q}_*(j)\right| > t\right) \leq \exp\left(-\frac{n(t - p^*_{\text{fail}})}{3}\right)\Big|_{p^*_{\text{fail}} = Q^*\left(\frac{mp}{2}\right)} + \exp\left(-\frac{mp}{8}\right).$$

Combining the results in Claims 1 and 2 back to Eq. (69) with the choice of $t_1 = t_2 = \frac{t}{2}$, we have for any $t \geq 4Q^*(\frac{mp}{2}) = \frac{8\sqrt{\pi}}{\sqrt{mp}}\left(\frac{\sqrt{2}+e^{C_1/2}}{\sqrt{C_1}} + \frac{\sqrt{2}+e^{C_2/2}}{\sqrt{C_2}}\right)$,

$$\begin{aligned} &\mathbb{P}\left(\left|\hat{q}_{\text{marg}}(j) - \theta_{col}^{(j)}\right| > t\right) \\ &\quad\leq \mathbb{P}\left(\left|\hat{q}_{\text{marg}}(j) - \hat{q}_*(j)\right| > \frac{t}{2}\right) + \mathbb{P}\left(\left|\hat{q}_*(j) - \theta_{col}^{(j)}\right| > \frac{t}{2}\right) \\ &\quad\leq \exp\left(-\frac{n(\frac{t}{2} - p^*_{\text{fail}})}{3}\right)\Big|_{p^*_{\text{fail}} = Q^*\left(\frac{mp}{2}\right)} + \exp\left(-\frac{nt^2}{2}\right) \\ &\qquad+ \exp\left(-\frac{mp}{8}\right). \end{aligned}$$

■

Appendix H. Proof of Lemma 21 and Auxiliary Lemmas

H.1. Lemmas to Control the Bias and Concentration of $\tilde{F}^{(i)}$

We show that the estimated CDF $\tilde{F}^{(i)}$ is close to the true CDF $F^{(i)}$ by showing both the bias $\left|\mathbb{E}\left[\tilde{F}^{(i)}(z)\right] - F^{(i)}(z)\right|$ and the variance of $\tilde{F}^{(i)}(z)$ are small. The following two lemmas assert these claims, based on consistency results for deconvolution (see Appendix L for detail).

Lemma 29 (Bias is small) *For every $i \in [m]$, the expectation of the kernel smoothed ECDF $\tilde{F}^{(i)}$ defined as in Eq. (38) uniformly converges to the true CDF $F^{(i)}$, and the convergence rate is given as $(\log|\mathcal{B}_i|)^{1/\beta}$, i.e., there exists a constant $C_3 = C(l) > 0$ such that*

$$\sup_{z\in\mathbb{R}}\left|\mathbb{E}\left[\tilde{F}^{(i)}(z)\right] - F^{(i)}(z)\right| \le C_3\left(\log|\mathcal{B}_i|\right)^{-1/\beta}, \quad \forall i \in [m].$$

Here, β is the smoothness parameter of the supersmooth noise.

Proof [Proof of Lemma 29] The expectation in the lemma statement is taken with respect to the randomness in data, i.e., realization of the samples which play the role of pivot points for kernel density estimation. Hence,

$$\begin{aligned}\left|\mathbb{E}\left[\tilde{F}^{(i)}(z)\right] - F^{(i)}(z)\right| &= \left|\mathbb{E}\left[\tilde{F}^{(i)}(z) - F^{(i)}(z)\right]\right| \\ &\le \mathbb{E}\left[\left(\tilde{F}^{(i)}(z) - F^{(i)}(z)\right)^2\right]^{1/2}, \end{aligned} \tag{72}$$

since $\mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2 \ge 0$. We will control the term in the right hand side of Eq. (72) by applying Theorem 57. For that purpose, we need to ensure that our density $f^{(i)}(z) = \frac{d}{dz}F^{(i)}(z)$ is in Fan's class for some m, a, and B (see Eq. (104) for the definition of Fan's class).

Note that $F^{(i)}$ is the inverse function of a slice of the latent function with a fixed row feature, $g_{x=\theta^{(i)}_{row}}$, in our model. We assume it admits a probability density $f^{(i)}$. It is easy to see that $\frac{1}{L} \le f^{(i)}(z) \le \frac{1}{l}$ for all $z \in supp\ f^{(i)}$ (and $f^{(i)}(z) = 0$ outside the support) because the inverse of $F^{(i)}$ is assumed (l, L) bi-Lipschitz in our model. This $f^{(i)}$ belongs to Fan's class

$$\mathcal{C}_{m,\alpha,B} = \left\{f(x) : \left|f^{(m)}(x) - f^{(m)}\left(x+\delta\right)\right| \le B\delta^{\alpha}\right\},$$

with the choice of $m = 0, \alpha = 0$, and $B = \frac{1}{l}$.

Therefore, for all $i \in [m]$, the density corresponding to $F^{(i)}$ belongs to a Fan's class, i.e., $f^{(i)} \in \mathcal{C}_{0,0,\frac{1}{l}}$. As a result, we can apply Theorem 57 on Eq. (72) to conclude that for any $i \in [m]$,

$$\begin{aligned}\sup_{z\in\mathbb{R}}\left|\mathbb{E}\left[\tilde{F}^{(i)}(z)\right] - F^{(i)}(z)\right| &\le \sup_{z\in\mathbb{R}}\mathbb{E}\left[\left(\tilde{F}^{(i)}(z) - F^{(i)}(z)\right)^2\right]^{1/2} \\ &\le \sup_{f\in\mathcal{C}_{0,0,\frac{1}{l}}}\sup_{z\in\mathbb{R}}\mathbb{E}\left[\left(\tilde{F}_{|\mathcal{B}_i|}(z) - F(z)\right)^2\right]^{1/2} \\ &= O\left(\left(\log|\mathcal{B}_i|\right)^{-1/\beta}\right).\end{aligned}$$

$\tilde{F}_{|\mathcal{B}_i|}$ denotes an estimate of F with $|\mathcal{B}_i|$ number of samples. Moreover, the constant C_3 hidden in the big O notation is dependent on the class $\mathcal{C}_{0,0,\frac{1}{l}}$, hence, only on the model parameter l, because Fan's original result holds uniformly over the whole class $\mathcal{C}_{0,0,\frac{1}{l}}$. ■

Lemma 30 (Variance is small) *For each $i \in [m]$, the kernel smoothed ECDF $\tilde{F}^{(i)}$ defined as in Eq. (38) concentrates to its expectation, i.e.,*

$$\mathbb{P}\left(\left| \tilde{F}^{(i)}(z) - \mathbb{E}\left[\tilde{F}^{(i)}(z) \right] \right| \geq t \right) \leq 2 \exp\left(\frac{-|\mathcal{B}_i|^{1/2}}{2C_4^2 \left(\log |\mathcal{B}_i|\right)^{\frac{2}{\beta}}} t^2 \right).$$

Recall we defined the constant $C_4 = \frac{BK_{max}(D_2 - D_1)}{\pi(4\gamma)^{\frac{1}{\beta}}}$ where $\beta, \gamma > 0$ are smoothness parameters for the noise, and $K_{max} = \max_{t\in[-1,1]} |\phi_K(t)|$.

Proof [Proof of Lemma 30] Recall that the kernel smoothed ECDF $\tilde{F}^{(i)}$ evaluated at z is a function of $|\mathcal{B}_i|$ independent random variables $\{Z(i,j)\}_{j\in\mathcal{B}_i}$, i.e., when z is fixed, $\tilde{F}^{(i)}(z) : \mathbb{R}^{|\mathcal{B}_i|} \to \mathbb{R}$ such that

$$\tilde{F}^{(i)}(z)\left[Z(i,j_1), \ldots, Z(i,j_{|\mathcal{B}_i|})\right] = \int_{D_1}^{z\wedge D_2} \frac{1}{h\,|\mathcal{B}_i|} \sum_{j\in\mathcal{B}_i} L\left(\frac{w - Z(i,j)}{h}\right) dw,$$

where $L(z) = \frac{1}{2\pi}\int e^{-itz} \frac{\phi_K(t)}{\phi_N\left(\frac{t}{h}\right)} dt$ and h is the bandwidth parameter for kernel K.

We will first show that $\tilde{F}^{(i)}(z)$ satisfies the bounded difference condition (see Eq. (102)).

Let $\zeta^n = (\zeta_1, \ldots, \zeta_n)$ and $\zeta_j^n = (\zeta_1, \ldots, \zeta_j', \ldots, \zeta_n)$ be two n-tuples of real numbers, which differ only at the j-th position. Then

$$\begin{aligned}
&\tilde{F}^{(i)}(z)[\zeta^n] - \tilde{F}^{(i)}(z)[\zeta_j^n] \qquad (73)\\
&= \frac{1}{hn}\int_{D_1}^{z\wedge D_2} L\left(\frac{w-\zeta_j}{h}\right) - L\left(\frac{w-\zeta_j'}{h}\right) dw\\
&= \frac{1}{hn}\int_{D_1}^{z\wedge D_2} \frac{1}{2\pi}\int \left(e^{-it\frac{w-\zeta_j}{h}} - e^{-it\frac{w-\zeta_j'}{h}}\right) \frac{\phi_K(t)}{\phi_N\left(\frac{t}{h}\right)} dt dw\\
&\leq \frac{1}{2\pi hn}\int_{D_1}^{z\wedge D_2} \int \left| e^{-it\frac{w-\zeta_j}{h}} - e^{-it\frac{w-\zeta_j'}{h}} \right| \left| \frac{\phi_K(t)}{\phi_N\left(\frac{t}{h}\right)} \right| dt dw. \qquad (74)
\end{aligned}$$

Because e^{-itz} is on the unit circle in the complex plane for any real numbers t and z, we have

$$\left| e^{-it\frac{w-\zeta_j}{h}} - e^{-it\frac{w-\zeta_j'}{h}} \right| \leq \left| e^{-it\frac{w-\zeta_j}{h}} \right| + \left| e^{-it\frac{w-\zeta_j'}{h}} \right| = 2.$$

Since ϕ_K is assumed to have compact support (see Appendix L.2) within $[-1,1]$, and a Fourier transform of L^1 function is uniformly continuous, there exists $K_{max} = \max_{t\in[-1,1]} |\phi_K(t)| < \infty$ such that $|\phi_K(t)| \leq K_{max}, \forall t$. From the supersmoothness assumption on the noise (Eq. (5)), we have $\left|\phi_N\left(\frac{t}{h}\right)\right| \geq B^{-1}\exp\left(-\gamma\left|\frac{t}{h}\right|^{\beta}\right)$.

We choose the bandwidth parameter $h = (4\gamma)^{\frac{1}{\beta}} (\log n)^{-\frac{1}{\beta}}$ following Fan (Theorems 56, 57). Plugging these expresions into Eq. (74) leads to

$$\begin{aligned}
Eq.(74) &\leq \frac{(\log n)^{\frac{1}{\beta}}}{2\pi (4\gamma)^{\frac{1}{\beta}} n} \int_{D_1}^{z \wedge D_2} \int_{-1}^{1} 2BK_{max} \exp\left(\frac{1}{4}|t|^{\beta} \log n\right) dt dw \\
&\leq \frac{BK_{max} (\log n)^{\frac{1}{\beta}}}{\pi (4\gamma)^{\frac{1}{\beta}} n} \int_{D_1}^{z \wedge D_2} (1 - (-1)) \max_{t \in [-1,1]} \exp\left(\frac{1}{4}|t|^{\beta} \log n\right) dw \\
&= \frac{BK_{max} (\log n)^{\frac{1}{\beta}}}{\pi (4\gamma)^{\frac{1}{\beta}} n} ((z \wedge D_2) - D_1) 2n^{\frac{1}{4}} \\
&\leq \frac{2BK_{max}(D_2 - D_1) (\log n)^{\frac{1}{\beta}}}{\pi (4\gamma)^{\frac{1}{\beta}} n^{\frac{3}{4}}} \\
&= \frac{2C_4 (\log n)^{\frac{1}{\beta}}}{n^{\frac{3}{4}}}, \qquad \text{for any } z \in [D_1, D_2].
\end{aligned}$$

Applying McDiarmid's inequality (Lemma 55), we can conclude that,

$$\mathbb{P}\left(\left| \tilde{F}^{(i)}(z)[\zeta^n] - \mathbb{E}_{\zeta^n} \tilde{F}^{(i)}(z)[\zeta^n] \right| \geq t \right) \leq 2 \exp\left(\frac{-n^{1/2}}{2C_4^2 (\log n)^{\frac{2}{\beta}}} t^2 \right).$$

This argument holds for every $i \in [m]$, with replacing generic variable n with corresponding $|\mathcal{B}_i|$. ■

Lemma 31 (Variance is uniformly small) *For each $i \in [m]$, the kernel smoothed ECDF $\tilde{F}^{(i)}$ defined as in Eq. (38) uniformly concentrates to its expectation, i.e., for any nonnegative integer N and for any $t \geq \frac{\Delta^{(i)}(D_2 - D_1)}{N}$ (we define $\Delta^{(i)} := \frac{BK_{max}}{\pi(4\gamma)^{\frac{1}{\beta}}} |\mathcal{B}_i|^{\frac{1}{4}} (\log |\mathcal{B}_i|)^{\frac{1}{\beta}}$),*

$$\mathbb{P}\left(\sup_{z \in [D_1, D_2]} \left| \tilde{F}^{(i)}(z) - \mathbb{E}\left[\tilde{F}^{(i)}(z) \right] \right| \geq t \right) \leq 2N \exp\left(\frac{-|\mathcal{B}_i|^{1/2}}{2C_4^2 (\log |\mathcal{B}_i|)^{\frac{2}{\beta}}} \left(t - \frac{\Delta^{(i)} (D_2 - D_1)}{N} \right)^2 \right),$$

where $\beta, \gamma > 0$ are smoothness parameters for the noise, and $K_{max} = \max_{t \in [-1,1]} |\phi_K(t)|$.

Proof [Proof of Lemma 31] First, we discretize the interval interval $[D_1, D_2]$ by constructing a finite ε-net. For any $N \geq 1$, define the set

$$\mathcal{T}_N := \left\{ D_{min} + \frac{2k-1}{2N} (D_2 - D_1), \ \forall k \in [N] \right\}.$$

Then for any $N > 0$, $\mathcal{T}_N \subset [D_1, D_2]$ and it forms a $\frac{(D_2 - D_1)}{2N}$-net with $|\mathcal{T}_N| = N$, i.e., for any $z \in [D_1, D_2]$, there exists $k \in [N]$ such that $\left| z - \frac{2k-1}{2N} (D_2 - D_1) \right| \leq \frac{(D_2 - D_1)}{2N}$.

We can observe that

$$\begin{aligned}
\left\|\tilde{f}^{(i)}\right\|_\infty &= \left\|\frac{1}{h\,|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} L\left(\frac{z-Z(i,j)}{h}\right)\right\|_\infty \\
&\le \frac{1}{h}\,\|L\|_\infty \\
&= \frac{1}{2\pi h}\left\|\int_{-\infty}^{\infty} e^{-\mathbf{i}tz}\frac{\phi_K(t)}{\phi_N\left(\frac{t}{h}\right)}dt\right\|_\infty \\
&\le \frac{1}{2\pi h}\int_{-\infty}^{\infty}\left|e^{-\mathbf{i}tz}\frac{\phi_K(t)}{\phi_N\left(\frac{t}{h}\right)}\right|dt \\
&\le \frac{1}{2\pi h}\int_{-1}^{1}\frac{K_{max}}{B^{-1}\exp\left(-\gamma\left|\frac{t}{h}\right|^\beta\right)}dt \\
&\le \frac{BK_{max}\,(\log|\mathcal{B}_i|)^{\frac{1}{\beta}}}{2\pi\,(4\gamma)^{\frac{1}{\beta}}}\int_{-1}^{1}\exp\left(\frac{1}{4}\,|t|^\beta\log|\mathcal{B}_i|\right)dt \\
&\le \frac{BK_{max}\,(\log|\mathcal{B}_i|)^{\frac{1}{\beta}}}{2\pi\,(4\gamma)^{\frac{1}{\beta}}}\int_{-1}^{1}|\mathcal{B}_i|^{\frac{1}{4}}\,dt \\
&= \frac{BK_{max}}{\pi\,(4\gamma)^{\frac{1}{\beta}}}\,|\mathcal{B}_i|^{\frac{1}{4}}\,(\log|\mathcal{B}_i|)^{\frac{1}{\beta}}\,.
\end{aligned}$$

Let $\Delta^{(i)}$ denote the upper bound in the last line. Since this upper bound is universal for all realization of samples, $\left\|\mathbb{E}\left[\tilde{f}^{(i)}\right]\right\|_\infty \le \Delta^{(i)}$, too. Then $\left\|\tilde{f}^{(i)}-\mathbb{E}\left[\tilde{f}^{(i)}\right]\right\|_\infty \le 2\Delta^{(i)}$ and it follows from the definition of $\tilde{F}^{(i)}$ (see Eq. (38)) that

$$\sup_{z\in[D_1,D_2]}\left|\tilde{F}^{(i)}(z)-\mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right| \le \sup_{z\in\mathcal{T}_N}\left|\tilde{F}^{(i)}(z)-\mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right| + \frac{\Delta^{(i)}\,(D_2-D_1)}{N}.$$

Therefore, if $\left|\tilde{F}^{(i)}(z)-\mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right| \le \varepsilon$ for all $z\in\mathcal{T}_n$, the supremum over the whole domain is bounded above up to an additional term, that is to say, $\sup_{z\in[D_1,D_2]}\left|\tilde{F}^{(i)}(z)-\mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right| \le \varepsilon+\frac{\Delta^{(i)}(D_2-D_1)}{N}$. An application of the union bound on the contraposition of the previous statement yields

$$\begin{aligned}
\mathbb{P}\left(\sup_{z\in[D_1,D_2]}\left|\tilde{F}^{(i)}(z)-\mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right|\ge t\right) &\le \mathbb{P}\left(\sup_{z\in\mathcal{T}_N}\left|\tilde{F}^{(i)}(z)-\mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right|\ge t-\frac{\Delta^{(i)}\,(D_2-D_1)}{N}\right) \\
&\le \sum_{z\in\mathcal{T}_N}\mathbb{P}\left(\left|\tilde{F}^{(i)}(z)-\mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right|\ge t-\frac{\Delta^{(i)}\,(D_2-D_1)}{N}\right) \\
&\le 2N\exp\left(\frac{-\,|\mathcal{B}_i|^{1/2}}{2C_4^2\,(\log|\mathcal{B}_i|)^{\frac{2}{\beta}}}\left(t-\frac{\Delta^{(i)}\,(D_2-D_1)}{N}\right)^2\right).
\end{aligned}$$

■

H.2. Proof of Lemma 21

Proof [Proof of Lemma 21] By Lemma 29, we have a universal upper bound: for any $i \in [m]$, $\sup_{z\in\mathbb{R}} \left|\mathbb{E}\left[\tilde{F}^{(i)}(z)\right] - F^{(i)}(z)\right| = O\left((\log|\mathcal{B}_i|)^{-1/\beta}\right)$. Actually this bound is uniform over all possible realizations of $\theta^{(i)}_{row} \in [0,1]$. Therefore, we can explicitly introduce a constant $C_3 = C(l)$, which does not depend on $i \in [m]$, to write

$$\sup_i \sup_{z\in\mathbb{R}} \left|\mathbb{E}\left[\tilde{F}^{(i)}(z)\right] - F^{(i)}(z)\right| \leq C_3 \left(\log|\mathcal{B}_i|\right)^{-1/\beta}. \tag{75}$$

The concentration rate obtained in Lemma 31 is stronger than $(\log|\mathcal{B}_i|)^{1/\beta}$ as long as N is a subexponential function of $|\mathcal{B}_i|$:

$$\mathbb{P}\left(\sup_{z\in[D_1,D_2]} \left|\tilde{F}^{(i)}(z) - \mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right| \geq t\right) \leq 2N \exp\left(\frac{-|\mathcal{B}_i|^{1/2}}{2C_4^2 \left(\log|\mathcal{B}_i|\right)^{\frac{2}{\beta}}} \left(t - \frac{\Delta^{(i)}(D_2 - D_1)}{N}\right)^2\right),$$

Therefore, it is the bias which dominates the discrepancy between the kernel smoothed ECDF $\tilde{F}^{(i)}$ and the true CDF $F^{(i)} = g^{-1}_{x=\theta^{(i)}_{row}}$.

Now we will combine these two inequality by applying the union bound. For any $\delta_1, \delta_2 > 0$, suppose that both $\left|F^{(i)}(z) - \mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right| \leq \delta_1$ and $\left|\tilde{F}^{(i)}(z) - \mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right| \leq \delta_2$ are satisfied. Then $\left|\tilde{F}^{(i)}(z) - F^{(i)}\right| \leq \delta_1 + \delta_2$ follows by triangle inequality. We can obtain the desired concentration inequality by applying the union bound on the contraposition of this statement with the particular choice of $\delta_1 = C_3 \left(\log|\mathcal{B}_i|\right)^{-1/\beta}$ and $\delta_2 = t - \delta_1$. To be more specific, for any nonnegative integer N and for any $t > \frac{\Delta^{(i)}(D_2-D_1)}{N} + C_3 \left(\log|\mathcal{B}_i|\right)^{-1/\beta}$ (where C_3 is the constant as in Eq. (75)),

$$\begin{aligned}
&\mathbb{P}\left(\sup_{z\in[D_1,D_2]} \left|\tilde{F}^{(i)}(z) - F^{(i)}\right| > t\right) \\
&\quad\leq \mathbb{P}\left(\sup_{z\in[D_1,D_2]} \left|F^{(i)}(z) - \mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right| > C_3 \left(\log|\mathcal{B}_i|\right)^{-1/\beta}\right) \\
&\qquad + \mathbb{P}\left(\sup_{z\in[D_1,D_2]} \left|\tilde{F}^{(i)}(z) - \mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right| > t - C_3 \left(\log|\mathcal{B}_i|\right)^{-1/\beta}\right) \\
&\quad\leq 2N \exp\left(\frac{-|\mathcal{B}_i|^{1/2}}{2C_4^2 \left(\log|\mathcal{B}_i|\right)^{\frac{2}{\beta}}} \left(t - \frac{\Delta^{(i)}(D_2 - D_1)}{N} - C_3 \left(\log|\mathcal{B}_i|\right)^{-1/\beta}\right)^2\right).
\end{aligned}$$

Finally, letting $N = |\mathcal{B}_i|^{\frac{1}{4}} \left(\log|\mathcal{B}_i|\right)^{\frac{2}{\beta}}$ leads to $\frac{\Delta^{(i)}(D_2-D_1)}{N} = C_4 \left(\log|\mathcal{B}_i|\right)^{-\frac{1}{\beta}}$. ■

Proof [Proof of Corollary 22] Conditioned on event $E_{row,(i)}$, it holds for all $i \in [m]$ that $|\mathcal{B}_i| \geq \frac{np}{2}$. Similarly, $|\mathcal{B}_i| \leq 2np$ for all $i \in [m]$, when conditioned on event $E'_{row,(i)}$. Therefore, for any

$i \in [m]$,

$$\mathbb{P}\left(\sup_{z\in[D_1,D_2]}\left|\tilde{F}^{(i)}(z)-F^{(i)}\right|>t\middle|E_{row,(i)},E'_{row,(i)}\right)$$
$$\leq c_{n,p}\exp\left(\frac{-\left(\frac{np}{2}\right)^{1/2}}{2C_4^2\left(\log\left(2np\right)\right)^{\frac{2}{\beta}}}\left(t-C\left(\log\frac{np}{2}\right)^{-1/\beta}\right)^2\right).$$

where $c_{n,p}=2(2np)^{\frac{1}{4}}\left(\log\left(2np\right)\right)^{\frac{2}{\beta}}$. ■

Appendix I. Proof of Lemma 25 and Auxiliary Lemmas

The purpose of this section is to prove Lemma 25, which provides a probabilistic uniform bound on the CDF estimate $\hat{F}^{(i)}$. We will prove the desired result by showing (1) $\hat{F}^{(i)}$ concentrates around its expectation; (2) the expectation of $\hat{F}^{(i)}$ is close to that of $\tilde{F}^{(i)}$ under a high-probability conditioning event; and (3) the expectation of $\tilde{F}^{(i)}$ is uniformly close to the true CDF as shown in Lemma 29.

Claim (1) is proved in section I.5 by essentially the same argument as the known noise case (see Lemma 31) and (3) is already shown. It is the proof of claim (2), for which most of this section is spared.

Throughout the first three subsections (I.1, I.2, I.3) we show that the size of the set for noise density estimation, $\mathcal{T}_i$, is neither too big nor too small. With aid of auxiliary lemmas, we show the estimated characteristic function of the noise is sufficiently accurate so that the modified kernel estimator is sufficiently precise. The summarized result can be bound in section I.4, which characterizes the bias between $\tilde{F}^{(i)}$ and $\hat{F}^{(i)}$.

In section I.6, we introduce appropriate conditioning events which are used to prove claim (2), all of which are high probability events according to the lemmas proved. In the end, Lemma 25 is proved by applying union bound.

I.1. The size of the base set $\mathcal{T}_i$ for noise density estimation

We defined the set $\mathcal{T}_i$ to estimate the distribution of additive noise by emulating the setup of repeated measurements. In this section, we present two lemmas: Lemma 33 shows there are a plenty of triples in $\mathcal{T}_i$ enabling the estimation; on the other hand, Lemma 34 claims that there are not too many triples in $\mathcal{T} \supset \mathcal{T}_i$. the discrepancy between the presented procedure with formula in Eq. (**??**) and the original estimation procedure with repeated measurements is small with high probability.

Lemma 32 *The sets J and I defined in Algorithm 2 are sufficiently large with high probability. Specifically,*

$$\mathbb{P}\left(|J|\leq\frac{n\left[1-\exp\left(-\frac{mp}{8}\right)\right]}{2}\right)\leq\exp\left(-\frac{n\left[1-\exp\left(-\frac{mp}{8}\right)\right]}{8}\right),$$
$$\mathbb{P}\left(|I|\leq\frac{m\left[1-\exp\left(-\frac{|J|p}{8}\right)\right]}{2}\right)\leq\exp\left(-\frac{m\left[1-\exp\left(-\frac{|J|p}{8}\right)\right]}{8}\right).$$

Proof Recall the construction procedure of the set $\mathcal{T}$ (see Algorithm 2). The number of column indices in J is given as the sum of indicator variables

$$|J| := \sum_{j\in[n]} \mathbb{I}\left\{\left|\mathcal{B}^j\right| \geq \frac{mp}{2}\right\}. \tag{76}$$

Note that $|\mathcal{B}^j| = \sum_{i\in[m]} M(i,j)$ is distributed as $Binomial(m,p)$. It follows from the binomial Chernoff bound that

$$\mathbb{P}\left(\left|\mathcal{B}^j\right| \geq \frac{mp}{2}\right) \geq 1 - \exp\left(-\frac{mp}{8}\right).$$

Therefore, n indicator variables in Eq. (76) are independent Bernoulli variables, each of which takes value 1 with probability greater than $1 - \exp\left(-\frac{mp}{8}\right)$.

Therefore, $|J| \sim Binomial(n, p_2)$ with $p_2 \geq 1 - \exp\left(-\frac{mp}{8}\right)$. It follows that

$$\begin{aligned}\mathbb{P}\left(|J| \leq \frac{n\left[1-\exp\left(-\frac{mp}{8}\right)\right]}{2}\right) &\leq \mathbb{P}\left(|J| \leq \frac{np_2}{2}\right)\\ &\leq \exp\left(-\frac{np_2}{8}\right)\\ &\leq \exp\left(-\frac{n\left[1-\exp\left(-\frac{mp}{8}\right)\right]}{8}\right).\end{aligned}$$

In the same vein, the number of column indices in I is given as the sum of indicator variables

$$|I| := \sum_{i\in[m]} \mathbb{I}\left\{|\mathcal{B}_i \cap J| \geq \frac{|J|p}{2}\right\}.$$

Now $|\mathcal{B}_i \cap J| = \sum_{j\in J} M(i,j)$ is distributed as $Binomial(m,p')$ with $p' \geq p$, because $p' = \mathbb{P}\left(M(i,j) = 1 \,|\, j \in J\right) \geq \mathbb{P}\left(M(i,j) = 1\right) = p$. These m indicator variables are independent Bernoulli variables, each of which takes value 1 with probability greater than

$$\mathbb{P}\left(|\mathcal{B}_i \cap J| \geq \frac{|J|p}{2}\right) \geq 1 - \exp\left(-\frac{|J|p}{8}\right).$$

Therefore, $|I| \sim Binomial(m, p_3)$ with $p_3 \geq 1 - \exp\left(-\frac{|J|p}{8}\right)$. It follows that

$$\begin{aligned}\mathbb{P}\left(|I| \leq \frac{m\left[1-\exp\left(-\frac{|J|p}{8}\right)\right]}{2}\right) &\leq \mathbb{P}\left(|I| \leq \frac{mp_3}{2}\right)\\ &\leq \exp\left(-\frac{mp_3}{8}\right)\\ &\leq \exp\left(-\frac{m\left[1-\exp\left(-\frac{|J|p}{8}\right)\right]}{8}\right).\end{aligned}$$

■

Lemma 33 *For any* $i \in [m]$,

$$|\mathcal{T}_i| \geq \left(\frac{m\left[1-\exp\left(-\frac{|J|p}{8}\right)\right]}{2} - 1 \right) \left\lceil \frac{\frac{|J|p}{2} - 1 - \lfloor\sqrt{\frac{|J|p}{2}}\rfloor}{2} \right\rceil,$$

with probability at least $1 - \exp\left(- \frac{m\left[1-\exp\left(-\frac{|J|p}{8}\right)\right]}{8} \right)$.

Proof Recall the construction procedure of the set $\mathcal{T}$ and $\mathcal{T}_i$ (see Algorithm 2).

Given $i' \in I$, we let $\sigma_{i'} : \mathcal{B}_{i'} \cap J \to [|\mathcal{B}_{i'} \cap J|]$ denote a map which maps the column index in $\mathcal{B}_{i'} \cap J \subseteq [n]$ to integers $1, 2, \ldots, |\mathcal{B}_{i'} \cap J|$ such that $\sigma(j_1) < \sigma(j_2)$ implies that $\hat{q}_{\text{marg}}(j_1) \leq \hat{q}_{\text{marg}}(j_2)$. Note that $\sigma_{i'}$ is a bijection and is invertible where its inverse $\sigma_{i'}^{-1} : [|\mathcal{B}_{i'} \cap J|] \to \mathcal{B}_{i'} \cap J \subseteq [n]$.

First of all, we show that there cannot exist more than $\left\lfloor \sqrt{|\mathcal{B}_{i'} \cap J|} \right\rfloor$ k's (where $k \in [|\mathcal{B}_{i'} \cap J| - 1]$) such that

$$\left| \hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k+1)\right) - \hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k)\right) \right| > \frac{1}{\sqrt{|\mathcal{B}_{i'} \cap J|}}. \tag{77}$$

Let $[a, b)$ denote the half-open interval, that is to say, $[a, b) := \{x \in \mathbb{R} : a \leq x < b\}$. If $k_1 \neq k_2$,

$$\left[\hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k_1)\right), \hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k_1+1)\right)\right) \cap \left[\hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k_2)\right), \hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k_2+1)\right)\right) = \emptyset,$$

and hence,

$$\begin{aligned}
&\mu\left(\left[\hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k_1)\right), \hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k_1+1)\right)\right) \cup \left[\hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k_2)\right), \hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k_2+1)\right)\right)\right) \\
&\quad = \mu\left(\left[\hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k_1)\right), \hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k_1+1)\right)\right)\right) \\
&\qquad + \mu\left(\left[\hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k_2)\right), \hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k_2+1)\right)\right)\right),
\end{aligned}$$

where μ is the Lebesgue measure for $\mathbb{R}$, and $\mu([a, b)) = (b - a)\mathbb{I}\{b \geq a\}$. Let $\mathcal{S}_{i'}$ denote the set of k's in $[|\mathcal{B}_{i'} \cap J| - 1]$, which satisfies Eq. (77).

Let's Assume that $|\mathcal{S}_{i'}| \geq \left\lfloor \sqrt{|\mathcal{B}_{i'} \cap J|} \right\rfloor + 1$. Since $\hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k)\right) \in [0, 1], \forall k \in \left[|\mathcal{B}_{i'} \cap J|\right]$,

$$\begin{aligned}
1 = \mu([0,1)) &\geq \mu\left(\bigcup_{k \in |\mathcal{B}_{i'}|-1} \left[\hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k)\right), \hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k+1)\right)\right) \right) \\
&\geq \mu\left(\bigcup_{k \in \mathcal{S}_{i'}} \left[\hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k)\right), \hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k+1)\right)\right) \right) \\
&= \sum_{k \in \mathcal{S}_{i'}} \left(\hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k+1)\right) - \hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k)\right) \right) \\
&\geq \left(\left\lfloor \sqrt{|\mathcal{B}_{i'} \cap J|} \right\rfloor + 1 \right) \left(\frac{1}{\sqrt{|\mathcal{B}_{i'} \cap J|}} \right) \\
&> 1,
\end{aligned}$$

which is a contradiction. Therefore, it is proved that $|\mathcal{S}_{i'}| \leq \left\lfloor \sqrt{|\mathcal{B}_{i'} \cap J|} \right\rfloor$. For those $k \in [|\mathcal{B}_{i'} \cap J| - 1] \setminus \mathcal{S}_{i'}$, we have

$$\hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k+1)\right) - \hat{q}_{\text{marg}}\left(\sigma_{i'}^{-1}(k)\right) \leq \frac{1}{\sqrt{|\mathcal{B}_{i'} \cap J|}}.$$

In case both $k, k+1 \in [|\mathcal{B}_{i'} \cap J| - 1] \setminus \mathcal{S}_{i'}$, either $\left(i', \sigma_{i'}^{-1}(k), \sigma_{i'}^{-1}(k+1)\right) \in \mathcal{T}$ or $\left(i', \sigma_{i'}^{-1}(k+1), \sigma_{i'}^{-1}(k+2)\right) \in \mathcal{T}$, but not both. However, no more than half of $k \in \left[\, |\mathcal{B}_{i'} \cap J| - 1\right] \setminus \mathcal{S}_{i'}$ is excluded and there exist at least $\left\lceil \frac{|\mathcal{B}_{i'} \cap J| - 1 - \left\lfloor \sqrt{|\mathcal{B}_{i'} \cap J|} \right\rfloor}{2} \right\rceil$ number of k's such that $\left(i', \sigma_{i'}^{-1}(k), \sigma_{i'}^{-1}(k+1)\right) \in \mathcal{T}$.

From Lemma 32, we know that $|I| > \frac{m\left[1 - \exp\left(-\frac{|J|p}{8}\right)\right]}{2}$ with high probability (i might be also in I). We also know from the argument above that for each $i' \in I$, there exist at least $\left\lceil \frac{|\mathcal{B}_{i'} \cap J| - 1 - \left\lfloor \sqrt{|\mathcal{B}_{i'} \cap J|} \right\rfloor}{2} \right\rceil \geq \left\lceil \frac{\frac{|J|p}{2} - 1 - \lfloor \sqrt{\frac{|J|p}{2}} \rfloor}{2} \right\rceil$ number of k's such that $\left(i', \sigma_{i'}^{-1}(k), \sigma_{i'}^{-1}(k+1)\right) \in \mathcal{T}$. All in all, we can conclude that

$$|\mathcal{T}_i| \geq \left(\frac{m\left[1 - \exp\left(-\frac{|J|p}{8}\right)\right]}{2} - 1 \right) \left\lceil \frac{\frac{|J|p}{2} - 1 - \lfloor \sqrt{\frac{|J|p}{2}} \rfloor}{2} \right\rceil,$$

with probability at least $1 - \exp\left(- \frac{m\left[1 - \exp\left(-\frac{|J|p}{8}\right)\right]}{8} \right)$. ■

We have shown that $\mathcal{T}_i$ is sufficiently large with high probability. On the other hand, we can also show that $\mathcal{T}$ is not too large compared to the total number of observed entries in the matrix ($= mnp$) with high probability.

Lemma 34 *The set $\mathcal{T}$ is not too large with high probability. Specifically,*

$$\mathbb{P}\left(|\mathcal{T}| > mnp\right) \leq \exp\left(-\frac{mnp}{3}\right).$$

Proof It is clear from the description of algorithm (see Algorithm 2) that for each (i, j), there can exist at most one element $(i', j_1, j_2) \in \mathcal{T}$ such that either $(i, j) = (i', j_1)$ or $(i, j) = (i', j_2)$. Moreover, if there exists (i', j_1, j_2) satisfying either of those two conditions, $M(i, j) = 1$. As a result, $|\mathcal{T}| \leq \frac{1}{2} \sum_{i,j} M(i, j)$, which is the sum of mn independent and identically distributed Bernoulli random variable with probability p. Applying the binomial Chernoff bound yields

$$\mathbb{P}\left(|\mathcal{T}| > mnp\right) \leq \mathbb{P}\left(\sum_{i,j} M(i, j) > 2mnp \right) \leq \exp\left(-\frac{mnp}{3}\right).$$

■

I.2. Useful properties for noise density estimation

The set $\mathcal{T}$ is carefully constructed for estimating the noise distribution. To analyze the quality of estimated characteristic function of the noise, we introduce the following notations:

$$\|\Delta A\|_{\infty,(i)} = \max_{(i',j_1,j_2)\in\mathcal{T}_i} \left|A(i,j_1) - A(i,j_2)\right|, \quad \text{and} \tag{78}$$

$$\|\Delta N\|_{\infty,(i)} = \max_{(i',j_1,j_2)\in\mathcal{T}_i} \left|N(i,j_1) - N(i,j_2)\right|. \tag{79}$$

The following two lemmas show that these two quantities are not too large with high probability. In particular, Lemma 35 shows that $\|\Delta A\|_{\infty,(i)}$ is vanishingly small as $m, n \to \infty$, while Lemma 36 shows that $\|\Delta N\|_{\infty,(i)}$ scales only logarithmically with respect to m, n and p.

Lemma 35 *For* $t > L\sqrt{\frac{2}{|J|p}} + 4LQ^*\left(\frac{mp}{2}\right)$,

$$\mathbb{P}\left(\|\Delta A\|_{\infty,(i)} > t\right) \le |J| \exp\left(-\frac{n}{8L^2}\left(t - L\sqrt{\frac{2}{|J|\,p}}\right)^2\right) + |J| \exp\left(-\frac{n}{12L}\left(t - L\sqrt{\frac{2}{|J|\,p}} - 4LQ^*\left(\frac{mp}{2}\right)\right)\right).$$

Proof From the Lipschitz assumption on the latent function, we have

$$\left|A(i',j_1) - A(i',j_2)\right| \le L\left|\theta_{col}^{(j_1)} - \theta_{col}^{(j_2)}\right|.$$

It suffices to find an upper bound on $\left|\theta_{col}^{(j_1)} - \theta_{col}^{(j_2)}\right|$ to control $\|\Delta A\|_{\infty,(i)}$. However, this is a latent quantity, which is not observable from data. Instead, we take a detour using triangle inequality:

$$\left|\theta_{col}^{(j_1)} - \theta_{col}^{(j_2)}\right| \le \left|\theta_{col}^{(j_1)} - \hat{q}_{\text{marg}}(j_1)\right| + \left|\hat{q}_{\text{marg}}(j_1) - \hat{q}_{\text{marg}}(j_2)\right| + \left|\hat{q}_{\text{marg}}(j_2) - \theta_{col}^{(j_2)}\right|.$$

We will show $\left|\hat{q}_{\text{marg}}(j_1) - \hat{q}_{\text{marg}}(j_2)\right|$ is small for all $(i', j_1, j_2) \in \mathcal{T}$ by careful construction of $\mathcal{T}$, and $\left|\theta_{col}^{(j)} - \hat{q}_{\text{marg}}(j)\right|$ is small for all $j \in J$ due to the concentration of quantile estimates (see Lemma 20).

First of all, note that $|\mathcal{B}_i \cap J| \ge \frac{|J|p}{2}$ for any $i \in I$ by construction of $\mathcal{T}$. Therefore, for any $(i', j_1, j_2) \in \mathcal{T}$,

$$\hat{q}_{\text{marg}}(j_1) - \hat{q}_{\text{marg}}(j_2) \le \frac{1}{\sqrt{|\mathcal{B}_i \cap J|}} \le \sqrt{\frac{2}{|J|\,p}}.$$

In other words,

$$\mathbb{P}\left(\bigcup_{(i',j_1,j_2)\in\mathcal{T}_i} \left\{\left|\hat{q}_{\text{marg}}(j_1) - \hat{q}_{\text{marg}}(j_2)\right| > \sqrt{\frac{2}{|J|\,p}}\right\}\right) = 0.$$

Next, recall that we defined function $Q^* : \mathbb{R}_+ \to \mathbb{R}_+$ as (see Eq. (41))

$$Q^*(x) = 2\sqrt{\pi}\left(\frac{1}{\sqrt{C_1 x}} + \frac{1}{\sqrt{C_2 x}} + \frac{1}{\sqrt{mpC_1 e^{-C_1}}} + \frac{1}{\sqrt{mpC_2 e^{-C_2}}}\right),$$

where $C_1 = \frac{l^2}{2(D_{max}-D_{min})^2}$ and $C_2 = \frac{l^2}{8\sigma^2}$ are model dependent constants. Note that the set J is defined as $J = \left\{j \in [n] : |\mathcal{B}^j| \geq \frac{mp}{2}\right\}$ (see Algorithm 2 in section E.1). By Lemma 20, for any $t \geq 4Q^*(\frac{mp}{2}) = \Theta\left(\frac{1}{\sqrt{mp}}\right)$,

$$\mathbb{P}\left(\left|\hat{q}_{\text{marg}}(j) - \theta^{(j)}_{col}\right| > t \,\middle|\, j \in J\right) \leq \exp\left(-\frac{nt^2}{2}\right) + \exp\left(-\frac{n(\frac{t}{2} - Q^*\left(\frac{mp}{2}\right))}{3}\right).$$

It is worthwhile to remark that $\exp\left(-\frac{mp}{8}\right)$ term is removed from the original statement of Lemma 20. That term was originally coming from $\mathbb{P}\left(E^c_{col,(j)}\right)$ (see the Claim 2 in the proof of the lemma), however, that term disappears once $j \in J$. By applying the union bound, it follows that

$$\begin{aligned}\mathbb{P}\left(\left|\hat{q}_{\text{marg}}(j) - \theta^{(j)}_{col}\right| > t,\ \forall j \in J\right) &\leq \sum_{j\in J}\mathbb{P}\left(\left|\hat{q}_{\text{marg}}(j) - \theta^{(j)}_{col}\right| > t \,\middle|\, j \in J\right)\\ &\leq |J|\left[\exp\left(-\frac{nt^2}{2}\right) + \exp\left(-\frac{n(\frac{t}{2} - Q^*\left(\frac{mp}{2}\right))}{3}\right)\right].\end{aligned}$$

From the argument above, if $\left|\hat{q}_{\text{marg}}(j) - \theta^{(j)}_{col}\right| \leq t_1$ for all $j \in J$ and $\left|\hat{q}_{\text{marg}}(j_1) - \hat{q}_{\text{marg}}(j_2)\right| \leq t_2$ for all triple $(i', j_1, j_2) \in \mathcal{T}$, then $\left|A(i', j_1) - A(i', j_2)\right| \leq L(2t_1 + t_2)$ for all $(i', j_1, j_2) \in \mathcal{T}$. Consequently, for $t > L\sqrt{\frac{2}{|J|p}} + 4LQ^*\left(\frac{mp}{2}\right)$,

$$\begin{aligned}\mathbb{P}\left(\|\Delta A\|_{\infty,(i)} > t\right) &= \mathbb{P}\left(\max_{(i',j_1,j_2)\in\mathcal{T}_i}\left|A(i', j_1) - A(i', j_2)\right| > t\right)\\ &\leq \mathbb{P}\left(\bigcup_{(i',j_1,j_2)\in\mathcal{T}_i}\left\{\left|\theta^{(j_1)}_{col} - \theta^{(j_2)}_{col}\right| > \frac{t}{L}\right\}\right)\\ &\leq \mathbb{P}\left(\bigcup_{(i',j_1,j_2)\in\mathcal{T}_i}\left\{\left|\hat{q}_{\text{marg}}(j_1) - \hat{q}_{\text{marg}}(j_2)\right| > \sqrt{\frac{2}{|J|p}}\right\}\right)\\ &\quad + \mathbb{P}\left(\left|\hat{q}_{\text{marg}}(j) - \theta^{(j)}_{col}\right| > \frac{1}{2}\left[\frac{t}{L} - \sqrt{\frac{2}{|J|p}}\right],\ \forall j \in J\right)\\ &\leq |J|\exp\left(-\frac{n}{8L^2}\left(t - L\sqrt{\frac{2}{|J|p}}\right)^2\right)\\ &\quad + |J|\exp\left(-\frac{n}{12L}\left(t - L\sqrt{\frac{2}{|J|p}} - 4LQ^*\left(\frac{mp}{2}\right)\right)\right).\end{aligned}$$

■

Lemma 36 *$\|\Delta N\|_{\infty,(i)}$ does not exceed $4\sigma\sqrt{\log 4|\mathcal{T}|}$ with high probability. Specifically,*

$$\mathbb{P}\left(\|\Delta N\|_{\infty,(i)} > 4\sigma\sqrt{\log 4|\mathcal{T}|}\right) \le \frac{1}{4|\mathcal{T}|}.$$

Combined with Lemmas 33 and 34, this lemma asserts that $\|\Delta N\|_{\infty,(i)} < 4\sigma\sqrt{\log 4mnp}$ with high probability, i.e. $1 - O\left(\frac{1}{mnp}\right)$.

Proof For any $t > 0$, if $\left|N(i', j_1)\right|, \left|N(i', j_2)\right| \le \frac{t}{2}$ for all $(i', j_1, j_2) \in \mathcal{T}$, then $\|\Delta N\|_{\infty,(i)} \le t$. Considering its contrapositive,

$$\begin{aligned}
\mathbb{P}\left(\|\Delta N\|_{\infty,(i)} > t\right) &\le \mathbb{P}\left(\exists (i', j_1, j_2) \in \mathcal{T} : \left|N(i', j_1)\right| \ge \frac{t}{2} \text{ or } \left|N(i', j_2)\right| \ge \frac{t}{2}\right) \\
&\le \sum_{(i', j_1, j_2) \in \mathcal{T}} \left[\mathbb{P}\left(\left|N(i', j_1)\right| \ge \frac{t}{2}\right) + \mathbb{P}\left(\left|N(i', j_2)\right| \ge \frac{t}{2}\right)\right] \\
&\le 2|\mathcal{T}|\mathbb{P}\left(\left|N(i, j)\right| \ge \frac{t}{2}\right) \\
&\le 4|\mathcal{T}| \exp\left(-\frac{t^2}{8\sigma^2}\right).
\end{aligned}$$

The last line follows from the sub-Gaussian assumption on the noise and the Chernoff bound.

With the choice of $t = 4\sigma\sqrt{\log 4|\mathcal{T}|}$,

$$\mathbb{P}\left(\|\Delta N\|_{\infty,(i)} > 4\sigma\sqrt{\log 4|\mathcal{T}|}\right) \le 4|\mathcal{T}| \exp\left(-2\log 4|\mathcal{T}|\right) = \frac{1}{4|\mathcal{T}|}.$$

■

I.3. Uniform convergence of $\hat{\phi}(t)$ to $\phi(t)$: step 2-1 in section E.1

Recall that the estimator $\hat{F}$ of interest differs from $\tilde{F}$ already analyzed only in one sense; $\hat{L}$ is defined with estimated characteristic function of the noise $\hat{\phi}_{N,i}$ with ridge parameter to avoid division-by-zero (see Eqs. (51), (52)), while L is defined with true noise characteristic function ϕ_N.

$$\hat{f}^{(i)}(z) = \frac{1}{h|\mathcal{B}_i|} \sum_{j \in \mathcal{B}_i} \hat{L}\left(\frac{z - Z(i,j)}{h}\right), \text{ where } \hat{L}(z) = \frac{1}{2\pi} \int e^{-\mathrm{i}tz} \frac{\phi_K(t)}{\hat{\phi}_{N,i}\left(\frac{t}{h}\right) + \rho} dt.$$

The goal of this section is to show for any $i \in [m]$, $\hat{\phi}_{N,i} \approx \phi_N$, thereby having $\hat{f} \approx \tilde{f}$, which will be shown in the next section.

Recall that the noise density is estimated from the base set $\mathcal{T}_i$ as per described in Algorithm 2 and that the estimated characteristic function is defined as follows (see Eq. (49)):

$$\hat{\phi}_{N,i}(t) = \left| \frac{1}{|\mathcal{T}_i|} \sum_{(i, j_1, j_2)) \in \mathcal{T}_i} \cos\left[t\left(Z(i, j_1) - Z(i, j_2)\right)\right] \right|^{1/2}.$$

For analytical purpose, we define an imaginary estimator of the characteristic function of noise as

$$\hat{\phi}^*_{N,i}(t) = \left| \frac{1}{|\mathcal{T}_i|} \sum_{(i,j_1,j_2))\in\mathcal{T}_i} \cos\left[t\left(N(i,j_1) - N(i,j_2)\right)\right] \right|^{1/2}. \tag{80}$$

We label the argument inside the absolute value bracket as follows so that $\hat{\phi}^*_{N,i}(t) = \left|\hat{\Phi}^*_{N,i}(t)\right|^{\frac{1}{2}}$:

$$\hat{\Phi}^*_{N,i}(t) = \frac{1}{|\mathcal{T}_i|} \sum_{(i,j_1,j_2))\in\mathcal{T}_i} \cos\left[t\left(N(i,j_1) - N(i,j_2)\right)\right]. \tag{81}$$

Lemma 37 *For any $i \in [m]$, $\hat{\phi}^*_{N,i}$ is close to ϕ_N with high probability. Specifically, for any $t \in \mathbb{R}$ and for any $s > 0$,*

$$\begin{aligned}\mathbb{P}\left(\left|\hat{\phi}^*_{N,i}(t) - \phi_N(t)\right| > s\right) &\le \mathbb{P}\left(\left|\hat{\Phi}^*_{N,i}(t) - \phi_N(t)^2\right| > s^2\right) \\ &\le 2\exp\left(-\frac{|\mathcal{T}_i|s^4}{2}\right).\end{aligned}$$

Proof By the assumption of supersmooth noise (see Eq. (5)), $\phi_N(t) \ge B^{-1}\exp\left(-\gamma|t|^\beta\right) > 0$ for all $t \in \mathbb{R}$. Also, by definition of the estimator (see Eq. (80)), $\hat{\phi}^*_{N,i}(t) \ge 0$ for all $t \in \mathbb{R}$. Since $|a-b| \le |a+b|$ for $a, b \ge 0$, we have for any $t \in \mathbb{R}$,

$$\begin{aligned}\left|\hat{\phi}^*_{N,i}(t) - \phi_N(t)\right| &\le \left(\left|\hat{\phi}^*_{N,i}(t) - \phi_N(t)\right|\left|\hat{\phi}^*_{N,i}(t) + \phi_N(t)\right|\right)^{\frac{1}{2}} \\ &= \left|\hat{\phi}^*_{N,i}(t)^2 - \phi_N(t)^2\right|^{\frac{1}{2}} \\ &\le \left|\hat{\Phi}^*_{N,i}(t) - \phi_N(t)^2\right|^{\frac{1}{2}}.\end{aligned}$$

The last inequality follows from $\left|\left|\hat{\Phi}^*_{N,i}(t)\right| - \phi_N(t)^2\right| \le \left|\hat{\Phi}^*_{N,i}(t) - \phi_N(t)^2\right|$, because $\phi_N(t) > 0$.

From the symmetry of the noise distribution and the independence between $N(i,j_1)$ and $N(i,j_2)$ for $(i,j_1,j_2) \in \mathcal{T}_i$,

$$\begin{aligned}&\mathbb{E}\left[\cos\left[t\left(N(i,j_1) - N(i,j_2)\right)\right]\right. \\ &\quad= \mathbb{E}\left[\frac{1}{2}\exp\left(t\left(N(i,j_1) - N(i,j_2)\right)\right) + \frac{1}{2}\exp\left(-t\left(N(i,j_1) - N(i,j_2)\right)\right)\right] \\ &\quad= \frac{1}{2}\mathbb{E}\left[tN(i,j_1)\right]\mathbb{E}\left[-tN(i,j_2)\right] + \mathbb{E}\left[-tN(i,j_1)\right]\mathbb{E}\left[tN(i,j_2)\right] \\ &\quad= \phi_N(t)^2.\end{aligned}$$

Therefore, $\mathbb{E}\left[\hat{\Phi}^*_{N,i}(t)\right] = \phi_N(t)^2$ for all $t \in \mathbb{R}$.

Since $\left|\cos\left[t\left(N(i,j_1) - N(i,j_2)\right)\right|\right. \le 1$, we can apply Hoeffding's inequality to achieve

$$\mathbb{P}\left(\left|\hat{\Phi}^*_{N,i}(t) - \phi_N(t)^2\right| > s\right) \le 2\exp\left(-\frac{|\mathcal{T}_i|s^2}{2}\right), \quad \text{for all } t \in \mathbb{R}.$$

All in all, for any $t \in \mathbb{R}$ and for any $s > 0$,

$$\begin{aligned}\mathbb{P}\left(\left|\hat{\phi}^*_{N,i}(t) - \phi_N(t)\right| > s\right) &\leq \mathbb{P}\left(\left|\hat{\phi}^*_{N,i}(t)^2 - \phi_N(t)^2\right| > s^2\right) \\ &\leq \mathbb{P}\left(\left|\hat{\Phi}^*_{N,i}(t) - \phi_N(t)^2\right| > s^2\right) \\ &\leq 2\exp\left(-\frac{|\mathcal{T}_i|s^4}{2}\right).\end{aligned}$$

■

Lemma 38 *For any $i \in [m]$, $\hat{\phi}^*_{N,i}$ is uniformly close to ϕ_N with high probability. Specifically, for any $\Lambda > 0$, any $N \in \mathbb{N}$ and any $s > \left\|\Delta^{*(i)}_{N,\Lambda}\right\|^{\frac{1}{2}}_{\infty}$,*

$$\mathbb{P}\left(\sup_{t\in[-\Lambda,\Lambda]}\left|\hat{\phi}^*_{N,i}(t) - \phi_N(t)\right| > s\right) \leq 2N\exp\left(-\frac{|\mathcal{T}_i|}{2}\left(s^2 - \left\|\Delta^{*(i)}_{N,\Lambda}\right\|_{\infty}\right)^2\right), \tag{82}$$

where $\left\|\Delta^{*(i)}_{N,\Lambda}\right\|_{\infty} = \frac{\Lambda}{N}\left[|\Lambda|\|\Delta N\|^2_{\infty,(i)} + 2\sigma B\right]$.

Proof

First, we discretize the interval interval $[-\Lambda, \Lambda]$ by constructing a finite ε-net. For any $N \geq 1$, define the set

$$\mathcal{T}_N := \left\{\frac{(2k-1-N)\Lambda}{2N},\ \forall k \in [N]\right\}.$$

Then for any $N > 0$, $\mathcal{T}_N \subset [-\Lambda, \Lambda]$ and it forms a $\frac{\Lambda}{N}$-net with $|\mathcal{T}_N| = N$, i.e., for any z with $|z| \leq \Lambda$, there exists $z' \in \mathcal{T}_N$ such that $|z - z'| \leq \frac{\Lambda}{N}$.

Next, we consider the maximum rate of change of the function $\hat{\Phi}^*_{N,i}(t) - \phi^2_N(t)$ to determine the resolution of the net. For brevity, we let $\Delta N \equiv N(i, j_1) - N(i, j_2)$. We can observe that

$$\begin{aligned}\left|\frac{d}{dt}\hat{\Phi}^*_{N,i}(t)\right| &= \left|\frac{1}{|\mathcal{T}_i|}\sum_{(i,j_1,j_2))\in\mathcal{T}_i}\frac{d}{dt}\cos\left[t\left(N(i,j_1) - N(i,j_2)\right)\right]\right| \\ &= \left|\frac{-1}{|\mathcal{T}_i|}\sum_{(i,j_1,j_2))\in\mathcal{T}_i}\sin\left[t\left(N(i,j_1) - N(i,j_2)\right)\right]\left(N(i,j_1) - N(i,j_2)\right)\right| \\ &\leq \max_{(i,j_1,j_2))\in\mathcal{T}_i}\left|t\right|\left|N(i,j_1) - N(i,j_2)\right|^2 \\ &= |t|\|\Delta N\|^2_{\infty,(i)}.\end{aligned}$$

and

$$\begin{aligned}\left|\frac{d}{dt}\phi_N^2(t)\right| &= 2\left|\phi_N(t)\frac{d}{dt}\phi_N(t)\right| \\ &\le 2\big|\phi_N(t)\big|\left|\frac{d}{dt}\int_{-\infty}^{\infty} e^{\mathbf{i}tx}dF_N(x)\right| \\ &\le 2\big|\phi_N(t)\big|\left|\int_{-\infty}^{\infty} \mathbf{i}xe^{\mathbf{i}tx}dF_N(x)\right| \\ &\le 2\big|\phi_N(t)\big|\int_{-\infty}^{\infty} \big|x\big|dF_N(x) \\ &\le 2\sigma B\exp\left(-\gamma|t|^\beta\right).\end{aligned}$$

The last line follows from the sub-Gaussian noise assumption:

$$\int_{-\infty}^{\infty}\big|x\big|dF_N(x) = \mathbb{E}\left[\big|N\big|\right] \le \mathbb{E}\left[N^2\right]^{\frac{1}{2}} \le \sigma.$$

Therefore,

$$\begin{aligned}\sup_{t\in[-\Lambda,\Lambda]}\left|\frac{d}{dt}\left(\hat{\Phi}_{N,i}^*(t)-\phi_N^2(t)\right)\right| &\le \sup_{t\in[-\Lambda,\Lambda]}\left|\frac{d}{dt}\hat{\Phi}_{N,i}^*(t)\right| + \sup_{t\in[-\Lambda,\Lambda]}\left|\frac{d}{dt}\phi_N^2(t)\right| \\ &\le |\Lambda|\|\Delta N\|_{\infty,(i)}^2 + 2\sigma B.\end{aligned}$$

Then it follows from the continuity of $\hat{\Phi}_{N,i}^*(t)-\phi_N^2(t)$ that

$$\sup_{t\in[-\Lambda,\Lambda]}\left|\hat{\Phi}_{N,i}^*(t)-\phi_N^2(t)\right| \le \sup_{t\in\mathcal{T}_N}\left|\hat{\Phi}_{N,i}^*(t)-\phi_N^2(t)\right| + \frac{\Lambda}{N}\left[|\Lambda|\|\Delta N\|_{\infty,(i)}^2 + 2\sigma B\right].$$

We let $\left\|\Delta_{N,\Lambda}^{*(i)}\right\|_\infty$ denote the upper bound on the error term, i.e.,

$$\left\|\Delta_{N,\Lambda}^{*(i)}\right\|_\infty := \frac{\Lambda}{N}\left[|\Lambda|\|\Delta N\|_{\infty,(i)}^2 + 2\sigma B\right].$$

Therefore, if $\left|\hat{\Phi}_{N,i}^*(t)-\phi_N^2(t)\right| \le s$ for all $t\in\mathcal{T}_N$, the supremum over the entire domain $[-\Lambda,\Lambda]$ is bounded above up to an additional term as $\sup_{z\in[-\Lambda,\Lambda]}\left|\hat{\Phi}_{N,i}^*(t)-\phi_N^2(t)\right| \le s + \Delta_*^{(i)}\frac{\Lambda}{N}$.

An application of the union bound on the contraposition of the previous statement yields

$$\begin{aligned}
\mathbb{P}\left(\sup_{t\in[-\Lambda,\Lambda]}\left|\hat{\phi}^*_{N,i}(t)-\phi_N(t)\right|>s\right) &\le \mathbb{P}\left(\sup_{t\in[-\Lambda,\Lambda]}\left|\hat{\Phi}^*_{N,i}(t)-\phi^2_N(t)\right|>s^2\right)\\
&\le \mathbb{P}\left(\sup_{t\in\mathcal{T}_N}\left|\hat{\Phi}^*_{N,i}(t)-\phi^2_N(t)\right|>s^2-\left\|\Delta^{*(i)}_{N,\Lambda}\right\|_\infty\right)\\
&\le \sum_{t\in\mathcal{T}_N}\mathbb{P}\left(\left|\hat{\Phi}^*_{N,i}(t)-\phi^2_N(t)\right|>s^2-\left\|\Delta^{*(i)}_{N,\Lambda}\right\|_\infty\right)\\
&\le 2\sum_{t\in\mathcal{T}_N}\exp\left(-\frac{|\mathcal{T}_i|}{2}\left(s^2-\left\|\Delta^{*(i)}_{N,\Lambda}\right\|_\infty\right)^2\right)\\
&\le 2N\exp\left(-\frac{|\mathcal{T}_i|}{2}\left(s^2-\left\|\Delta^{*(i)}_{N,\Lambda}\right\|_\infty\right)^2\right).
\end{aligned}$$

■

As in Eq. (81), we let

$$\hat{\Phi}_{N,i}(t)=\frac{1}{|\mathcal{T}_i|}\sum_{(i,j_1,j_2))\in\mathcal{T}_i}\cos\left[t\left(Z(i,j_1)-Z(i,j_2)\right)\right], \tag{83}$$

so that $\hat{\phi}_{N,i}(t)=\left|\hat{\Phi}_{N,i}(t)\right|^{\frac{1}{2}}$.

Lemma 39 *For any $i\in[m]$, $\hat{\phi}_{N,i}$ is close to $\hat{\phi}^*_{N,i}$ with high probability. Specifically, for any $t\in\mathbb{R}$ and for any $s>\frac{|t|}{\sqrt{2}}\|\Delta A\|_{\infty,(i)}$,*

$$\begin{aligned}
\mathbb{P}\left(\left|\hat{\phi}_{N,i}(t)-\hat{\phi}^*_{N,i}(t)\right|>s\right) &\le \mathbb{P}\left(\left|\hat{\Phi}_{N,i}(t)-\hat{\Phi}^*_{N,i}(t)\right|>s^2\right)\\
&\le 2\exp\left(-\frac{|\mathcal{T}_i|}{2t^2\|\Delta A\|^2_{\infty,(i)}}\left(s^2-\frac{t^2\|\Delta A\|^2_{\infty,(i)}}{2}\right)^2\right).
\end{aligned}$$

Proof We know that $\hat{\phi}_{N,i}(t),\hat{\phi}^*_{N,i}(t)\ge 0$ for all $t\in\mathbb{R}$ (see Eqs. (49), (80)). By the same argument as in the proof of Lemma 37, for any $t\in\mathbb{R}$,

$$\begin{aligned}
\left|\hat{\phi}_{N,i}(t)-\hat{\phi}^*_{N,i}(t)\right| &\le \left(\left|\hat{\phi}_{N,i}(t)-\hat{\phi}^*_{N,i}(t)\right|\left|\hat{\phi}_{N,i}(t)+\hat{\phi}^*_{N,i}(t)\right|\right)^{\frac{1}{2}}\\
&= \left|\hat{\phi}_{N,i}(t)^2-\hat{\phi}^*_{N,i}(t)^2\right|^{\frac{1}{2}}.
\end{aligned}$$

Note that for any $a,b\in\mathbb{R}$, $\left||a|-|b|\right|\le|a-b|$.

$$\left|\hat{\phi}_{N,i}(t)^2-\hat{\phi}^*_{N,i}(t)^2\right|=\left|\left|\hat{\Phi}_{N,i}(t)\right|-\left|\hat{\Phi}^*_{N,i}(t)\right|\right|\le\left|\hat{\Phi}_{N,i}(t)-\hat{\Phi}^*_{N,i}(t)\right|.$$

By the model assumption, $Z(i,j) = A(i,j) + N(i,j)$. Changing the perspective, we now consider $Z(i,j_1) - Z(i,j_2)$ as a perturbed instance of the noise $N(i,j_1) - N(i,j_2)$ by the signal difference $A(i,j_1) - A(i,j_2)$, which is assumed to be small for $(i,j_1,j_2) \in \mathcal{T}_i$.

For brevity, we let $\Delta N \equiv N(i,j_1) - N(i,j_2)$, $\Delta A \equiv A(i,j_1) - A(i,j_2)$ and $\Delta Z \equiv Z(i,j_1) - Z(i,j_2)$. Since it is known that $\cos a - \cos b = -2 \sin \frac{a+b}{2} \sin \frac{a-b}{2}$,

$$\begin{aligned}\left|\hat{\Phi}_{N,i}(t) - \hat{\Phi}^*_{N,i}(t)\right| &= \left|\frac{1}{|\mathcal{T}_i|} \sum_{(i,j_1,j_2))\in\mathcal{T}_i} \left\{ \cos\left[t\Delta Z\right] - \cos\left[t\Delta N\right] \right\}\right| \\ &= \left|\frac{-2}{|\mathcal{T}_i|} \sum_{(i,j_1,j_2))\in\mathcal{T}_i} \sin\left(t\Delta N + \frac{t\Delta A}{2}\right) \sin\left(\frac{t\Delta A}{2}\right)\right|.\end{aligned}$$

We will find an upper bound on this last term by showing that it sharply concentrates to its expectation, which is small.

Note that the distribution of ΔN is governed by the randomness in $\left\{N(i',j_1), N(i',j_2)\right\}_{(i',j_1,j_2)\in\mathcal{T}_i}$ and that of ΔA is by $\left\{\theta^{(i')}_{row}, \theta^{(j_1)}_{col}, \theta^{(j_2)}_{col}\right\}_{(i',j_1,j_2)\in\mathcal{T}_i}$. Conditioned on $\left\{\theta^{(i')}_{row}, \theta^{(j_1)}_{col}, \theta^{(j_2)}_{col}\right\}_{(i',j_1,j_2)\in\mathcal{T}_i}$, the summands, $\sin\left(t\Delta N + \frac{t\Delta A}{2}\right) \times \sin\left(\frac{t\Delta A}{2}\right)$, are independent from each other so that we can apply the Hoeffding's inequality.

Let $\Delta\hat{\Phi}_{N,i}(t) \equiv \hat{\Phi}_{N,i}(t) - \hat{\Phi}^*_{N,i}(t)$ and note that $\left|\sin x\right| \le \left|x\right|$ for $x \in \mathbb{R}$. Then for any $t \in \mathbb{R}$ and any $s > 0$,

$$\begin{aligned}\mathbb{P}\Big(\Big|\Delta\hat{\Phi}_{N,i}(t) - \mathbb{E}\left[\Delta\hat{\Phi}_{N,i}(t)\right]\Big| > s\Big) &\le 2\exp\left(-\frac{2\left(\frac{|\mathcal{T}_i|s}{2}\right)^2}{\sum_{(i,j_1,j_2)\in\mathcal{T}_i}\left(t\Delta A\right)^2}\right) \\ &\le 2\exp\left(-\frac{|\mathcal{T}_i|s^2}{2\max_{(i,j_1,j_2)\in\mathcal{T}_i}\left(t\Delta A\right)^2}\right) \\ &= 2\exp\left(-\frac{|\mathcal{T}_i|s^2}{2t^2\|\Delta A\|^2_{\infty,(i)}}\right). \qquad (84)\end{aligned}$$

Now we consider the expectation $\mathbb{E}\left[\Delta\hat{\Phi}_{N,i}(t)\right]$, where the expectation is with respect to the first source of randomness, $\left\{N(i',j_1), N(i',j_2)\right\}_{(i',j_1,j_2)\in\mathcal{T}_i}$. From the symmetry in the noise distribution,

$$\begin{aligned}\mathbb{E}\big[\Delta\hat{\Phi}_{N,i}(t)\big] &= \mathbb{E}\left[\frac{-2}{|\mathcal{T}_i|} \sum_{(i,j_1,j_2))\in\mathcal{T}_i} \sin\left(t\Delta N + \frac{t\Delta A}{2}\right) \sin\left(t\frac{\Delta A}{2}\right)\right] \\ &= \mathbb{E}\left[\frac{-1}{|\mathcal{T}_i|} \sum_{(i,j_1,j_2))\in\mathcal{T}_i} \left[\sin\left(t\Delta N + \frac{t\Delta A}{2}\right) + \sin\left(-t\Delta N + \frac{t\Delta A}{2}\right)\right] \sin\left(t\frac{\Delta A}{2}\right)\right] \\ &= \mathbb{E}\left[\frac{-2}{|\mathcal{T}_i|} \sum_{(i,j_1,j_2))\in\mathcal{T}_i} \cos\left(t\Delta N\right) \sin^2\left(\frac{t\Delta A}{2}\right)\right].\end{aligned}$$

We used the fact that $\sin(a+b)+\sin(a-b) = 2\sin\frac{a+b}{2}\cos\frac{a-b}{2}$. Since $\left|\cos\left(t\Delta N\right)\right| \le 1$ and $\left|\sin\left(\frac{t\Delta A}{2}\right)\right| \le \left|\frac{t\Delta A}{2}\right|$,

$$\left|\mathbb{E}\left[\Delta\hat{\Phi}_{N,i}(t)\right]\right| \le \frac{2}{|\mathcal{T}_i|}\sum_{(i,j_1,j_2))\in\mathcal{T}_i}\left|\frac{t\Delta A}{2}\right|^2 \le \max_{(i,j_1,j_2))\in\mathcal{T}_i}\frac{\left(t\Delta A\right)^2}{2} = \frac{t^2}{2}\|\Delta A\|^2_{\infty,(i)}. \tag{85}$$

Combining the upper bound on $\left|\mathbb{E}\left[\Delta\hat{\Phi}_{N,i}(t)\right]\right|$ in Eq. (85) together with the concentration inequality Eq. (84) yields the following result: for any $t\in\mathbb{R}$ and any $s > \frac{t^2}{2}\|\Delta A\|^2_{\infty,(i)}$,

$$\mathbb{P}\left(\left|\Delta\hat{\Phi}_{N,i}(t)\right| > s\right) \le 2\exp\left(-\frac{|\mathcal{T}_i|}{2t^2\|\Delta A\|^2_{\infty,(i)}}\left(s - \frac{t^2\|\Delta A\|^2_{\infty,(i)}}{2}\right)^2\right).$$

All in all, for any $t\in\mathbb{R}$ and for any $s > \frac{t}{\sqrt{2}}\|\Delta A\|_{\infty,(i)}$,

$$\begin{aligned}\mathbb{P}\left(\left|\hat{\phi}_{N,i}(t) - \hat{\phi}^*_{N,i}(t)\right| > s\right) &\le \mathbb{P}\left(\left|\hat{\Phi}_{N,i}(t) - \hat{\Phi}^*_{N,i}(t)\right| > s^2\right)\\ &\le 2\exp\left(-\frac{|\mathcal{T}_i|}{2t^2\|\Delta A\|^2_{\infty,(i)}}\left(s^2 - \frac{t^2\|\Delta A\|^2_{\infty,(i)}}{2}\right)^2\right).\end{aligned}$$

■

We can refine the result obtained so far to get a uniform upper bound with the ε-net argument. Recall that

$$\left|\hat{\phi}_{N,i}(t) - \hat{\phi}^*_{N,i}(t)\right| \le \left|\hat{\phi}_{N,i}(t)^2 - \hat{\phi}^*_{N,i}(t)^2\right|^{\frac{1}{2}} \le \left|\hat{\Phi}_{N,i}(t) - \hat{\Phi}^*_{N,i}(t)\right|^{\frac{1}{2}}.$$

It suffices to find a uniform upper bound on $\left|\hat{\Phi}_{N,i}(t) - \hat{\Phi}^*_{N,i}(t)\right|$.

Lemma 40 (Uniform convergence of the noise estimate) *For any $i\in[m]$, $\hat{\phi}_{N,i}$ is uniformly close to $\hat{\phi}^*_{N,i}$ with high probability. Specifically, for any $\Lambda > 0$, any $N\in\mathbb{N}$ and $s > \left\|\Delta^{(i)}_{N,\Lambda}\right\|^{\frac{1}{2}}_{\infty}$,*

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in[-\Lambda,\Lambda]}\left|\hat{\phi}_{N,i}(t) - \hat{\phi}^*_{N,i}(t)\right| > s\right) &\le \mathbb{P}\left(\sup_{t\in[-\Lambda,\Lambda]}\left|\hat{\Phi}_{N,i}(t) - \hat{\Phi}^*_{N,i}(t)\right| > s^2\right)\\ &\le 2N\exp\left(-\frac{|\mathcal{T}_i|}{2\Lambda^2\|\Delta A\|^2_{\infty,(i)}}\left(s^2 - \left\|\Delta^{(i)}_{N,\Lambda}\right\|_{\infty}\right)^2\right),\end{aligned}$$

where $\left\|\Delta^{(i)}_{N,\Lambda}\right\|_{\infty} = \frac{\Lambda^2\|\Delta A\|_{\infty,(i)}}{2N}\left[(N+2)\|\Delta A\|_{\infty,(i)} + 4\|\Delta N\|_{\infty,(i)}\right]$.

We note that, as we refine the net by letting $N\to\infty$, $\left\|\Delta^{(i)}_{N}\right\|_{\infty} \to \frac{\Lambda^2\|\Delta A\|^2_{\infty,(i)}}{2}$, which sets the fundamental lower bound on $\sup_{t\in[-\Lambda,\Lambda]}\left|\hat{\phi}_{N,i}(t) - \hat{\phi}^*_{N,i}(t)\right|$. That is to say, $\left\|\hat{\phi}_{N,i}(t) - \hat{\phi}^*_{N,i}(t)\right\|_{\infty} \approx$

$\Lambda\|\Delta A\|_{\infty,(i)}$. Indeed, such is a limit on the deconvolution obtained due to the inherent noise represented by term $\|\Delta A\|_{\infty,(i)}$ and some such limit is naturally expected.

Proof [Proof of Lemma 40] First, we discretize the interval interval $[-\Lambda, \Lambda]$ by constructing a finite ε-net. For any $N \geq 1$, define the set

$$\mathcal{T}_N := \left\{ \frac{(2k-1-N)\Lambda}{2N},\ \forall k \in [N] \right\}.$$

Then for any $N > 0$, $\mathcal{T}_N \subset [-\Lambda, \Lambda]$ and it forms a $\frac{\Lambda}{N}$-net with $|\mathcal{T}_N| = N$, i.e., for any z with $|z| \leq \Lambda$, there exists $z' \in \mathcal{T}_N$ such that $|z - z'| \leq \frac{\Lambda}{N}$.

Next, we consider the maximum rate of change of the function $\Delta\hat{\Phi}_N(t) \equiv \hat{\Phi}_{N,i}(t) - \hat{\Phi}^*_{N,i}(t)$ to determine the resolution of the net. We can observe that

$$\begin{aligned}
\frac{d}{dt}\Delta\hat{\Phi}_N(t) &= \frac{d}{dt}\hat{\Phi}_{N,i}(t) - \hat{\Phi}^*_{N,i}(t) \\
&= \frac{d}{dt}\left[\frac{-2}{|\mathcal{T}_i|}\sum_{(i,j_1,j_2))\in\mathcal{T}_i} \sin\left(t\left(\Delta N + \frac{\Delta A}{2}\right)\right)\sin\left(\frac{t\Delta A}{2}\right)\right] \\
&= \frac{-2}{|\mathcal{T}_i|}\sum_{(i,j_1,j_2))\in\mathcal{T}_i}\Bigg[\left(\Delta N + \frac{\Delta A}{2}\right)\cos\left(t\left(\Delta N + \frac{\Delta A}{2}\right)\right)\sin\left(\frac{t\Delta A}{2}\right) \\
&\qquad + \frac{\Delta A}{2}\sin\left(t\left(\Delta N + \frac{\Delta A}{2}\right)\right)\cos\left(\frac{t\Delta A}{2}\right)\Bigg],
\end{aligned}$$

and hence,

$$\begin{aligned}
&\sup_{t\in[-\Lambda,\Lambda]}\left|\frac{d}{dt}\Delta\hat{\Phi}_N(t)\right| \\
&\leq \sup_{t\in[-\Lambda,\Lambda]}\frac{2}{|\mathcal{T}_i|}\sum_{(i,j_1,j_2))\in\mathcal{T}_i}\Bigg|\left(\Delta N + \frac{\Delta A}{2}\right)\cos\left(t\left(\Delta N + \frac{\Delta A}{2}\right)\right)\sin\left(\frac{t\Delta A}{2}\right) \\
&\qquad + \frac{\Delta A}{2}\sin\left(t\left(\Delta N + \frac{\Delta A}{2}\right)\right)\cos\left(\frac{t\Delta A}{2}\right)\Bigg| \\
&\leq \sup_{t\in[-\Lambda,\Lambda]} 2 \max_{(i,j_1,j_2))\in\mathcal{T}_i}\Bigg[\left|\Delta N + \frac{\Delta A}{2}\right|\left|\cos\left(t\left(\Delta N + \frac{\Delta A}{2}\right)\right)\right|\left|\sin\left(\frac{t\Delta A}{2}\right)\right| \\
&\qquad + \left|\frac{\Delta A}{2}\right|\left|\sin\left(t\left(\Delta N + \frac{\Delta A}{2}\right)\right)\right|\left|\cos\left(\frac{t\Delta A}{2}\right)\right|\Bigg] \\
&\leq \sup_{t\in[-\Lambda,\Lambda]} 2 \max_{(i,j_1,j_2))\in\mathcal{T}_i}\left|\Delta N + \frac{\Delta A}{2}\right|\left|\frac{t\Delta A}{2}\right| + \left|\frac{\Delta A}{2}\right|\left|t\left(\Delta N + \frac{\Delta A}{2}\right)\right| \\
&\leq \sup_{t\in[-\Lambda,\Lambda]}|t|\Big(2\|\Delta N\|_{\infty,(i)} + \|\Delta A\|_{\infty,(i)}\Big)\|\Delta A\|_{\infty,(i)} \\
&\leq |\Lambda|\Big(2\|\Delta N\|_{\infty,(i)} + \|\Delta A\|_{\infty,(i)}\Big)\|\Delta A\|_{\infty,(i)}.
\end{aligned}$$

Let $\Delta^{(i)} = \big|\Lambda\big|\Big(2\|\Delta N\|_{\infty,(i)} + \|\Delta A\|_{\infty,(i)}\Big)\|\Delta A\|_{\infty,(i)}$, the upper bound in the last line. Then it follows from the continuity of $\Delta\hat{\Phi}_N(t)$ that

$$\sup_{t\in[-\Lambda,\Lambda]}\left|\Delta\hat{\Phi}_N(t)\right| \le \sup_{t\in\mathcal{T}_N}\left|\Delta\hat{\Phi}_N(t)\right| + \Delta^{(i)}\frac{\Lambda}{N}.$$

Therefore, if $\left|\Delta\hat{\Phi}_N(t)\right| \le s$ for all $t \in \mathcal{T}_N$, the supremum over the entire domain $[-\Lambda,\Lambda]$ is bounded above up to an additional term as $\sup_{z\in[-\Lambda,\Lambda]}\left|\Delta\hat{\Phi}_N(t)\right| \le s + \Delta^{(i)}\frac{\Lambda}{N}$. An application of the union bound on the contraposition of the previous statement yields

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in[-\Lambda,\Lambda]}\left|\hat{\phi}_{N,i}(t) - \hat{\phi}^*_{N,i}(t)\right| > s\right)\\
&\quad\le \mathbb{P}\left(\sup_{t\in[-\Lambda,\Lambda]}\left|\hat{\Phi}_{N,i}(t) - \hat{\Phi}^*_{N,i}(t)\right| > s^2\right)\\
&\quad\le \mathbb{P}\left(\sup_{t\in\mathcal{T}_N}\left|\hat{\Phi}_{N,i}(t) - \hat{\Phi}^*_{N,i}(t)\right| > s^2 - \Delta^{(i)}\frac{\Lambda}{N}\right)\\
&\quad\le \sum_{t\in\mathcal{T}_N}\mathbb{P}\left(\left|\hat{\Phi}_{N,i}(t) - \hat{\Phi}^*_{N,i}(t)\right| > s^2 - \Delta^{(i)}\frac{\Lambda}{N}\right)\\
&\quad\le 2\sum_{t\in\mathcal{T}_N}\exp\left(-\frac{|\mathcal{T}_i|}{2t^2\|\Delta A\|^2_{\infty,(i)}}\left(s^2 - \Delta^{(i)}\frac{\Lambda}{N} - \frac{t^2\|\Delta A\|^2_{\infty,(i)}}{2}\right)^2\right)\\
&\quad\le 2N\exp\left(-\frac{|\mathcal{T}_i|}{2\Lambda^2\|\Delta A\|^2_{\infty,(i)}}\left(s^2 - \Delta^{(i)}\frac{\Lambda}{N} - \frac{\Lambda^2\|\Delta A\|^2_{\infty,(i)}}{2}\right)^2\right).
\end{aligned}$$

We can simplify the last line by defining

$$\left\|\Delta^{(i)}_{N,\Lambda}\right\|_\infty = \Delta^{(i)}\frac{\Lambda}{N} + \frac{\Lambda^2\|\Delta A\|^2_{\infty,(i)}}{2} = \frac{\Lambda^2\|\Delta A\|_{\infty,(i)}}{2N}\Big[(N+2)\|\Delta A\|_{\infty,(i)} + 4\|\Delta N\|_{\infty,(i)}\Big],$$

because $\Delta^{(i)} = \big|\Lambda\big|\Big(2\|\Delta N\|_{\infty,(i)} + \|\Delta A\|_{\infty,(i)}\Big)\|\Delta A\|_{\infty,(i)}$. ■

Lemma 41 *For any $i \in [m]$, $\hat{\phi}_{N,i}$ is uniformly close to ϕ_N with high probability. Specifically, for any $\Lambda > 0$, any $N_1, N_2 \in \mathbb{N}$ and for any $s_1 > \left\|\Delta^{*(i)}_{N_1,\Lambda}\right\|^{\frac{1}{2}}_\infty$ and $s_2 > \left\|\Delta^{(i)}_{N_2,\Lambda}\right\|^{\frac{1}{2}}_\infty$,*

$$\begin{aligned}
\mathbb{P}\left(\sup_{t\in[-\Lambda,\Lambda]}\left|\hat{\phi}_{N,i}(t) - \phi_N(t)\right| > s_1 + s_2\right) &\le 2N_1\exp\left(-\frac{|\mathcal{T}_i|}{2}\left(s_1^2 - \left\|\Delta^{*(i)}_{N_1,\Lambda}\right\|_\infty\right)^2\right)\\
&\quad + 2N_2\exp\left(-\frac{|\mathcal{T}_i|}{2\Lambda^2\|\Delta A\|^2_{\infty,(i)}}\left(s_2^2 - \left\|\Delta^{(i)}_{N_2,\Lambda}\right\|_\infty\right)^2\right),
\end{aligned}$$

where

$$\left\|\Delta^{*(i)}_{N_1,\Lambda}\right\|_\infty = \frac{\Lambda}{N_1}\left[|\Lambda|\|\Delta N\|^2_{\infty,(i)} + 2\sigma B\right] \quad \textit{and}$$
$$\left\|\Delta^{(i)}_{N_2,\Lambda}\right\|_\infty = \frac{\Lambda^2\|\Delta A\|_{\infty,(i)}}{2N_2}\left[(N_2+2)\|\Delta A\|_{\infty,(i)} + 4\|\Delta N\|_{\infty,(i)}\right].$$

Proof If $\sup_{t\in[-\Lambda,\Lambda]}\left|\hat{\phi}^*_{N,i}(t) - \phi_N(t)\right| \le s_1$ and $\sup_{t\in[-\Lambda,\Lambda]}\left|\hat{\phi}_{N,i}(t) - \hat{\phi}^*_{N,i}(t)\right| \le s_2$, then $\sup_{t\in[-\Lambda,\Lambda]}\left|\hat{\phi}_{N,i}(t) - \phi_N(t)\right| \le s_1 + s_2$ by triangle inequality. Therefore,

$$\mathbb{P}\left(\sup_{t\in[-\Lambda,\Lambda]}\left|\hat{\phi}_{N,i}(t) - \phi_N(t)\right| > s_1 + s_2\right) \le \mathbb{P}\left(\sup_{t\in[-\Lambda,\Lambda]}\left|\hat{\phi}^*_{N,i}(t) - \phi_N(t)\right| > s_1\right) + \mathbb{P}\left(\sup_{t\in[-\Lambda,\Lambda]}\left|\hat{\phi}_{N,i}(t) - \hat{\phi}^*_{N,i}(t)\right| > s_2\right).$$

Applying Lemma 38 and 40 concludes the proof. ■

I.4. Bias from $\tilde{F}$ to $\hat{F}$

We show that the CDF estimated by the modified kernel estimator is uniformly close to that estimated by the traditional kernel estimator. For simplicity of the lemma statement, we introduce a conditioning event indexed by $i \in [m]$ as

$$E_{\phi,i} \equiv \left\{\sup_{t\in[-\frac{1}{h},\frac{1}{h}]}\left|\hat{\phi}_{N,i}(t) - \phi_N(t)\right| \le s_\phi\right\}.$$

We will show this event is a high probability event later in appendix I.6.

Lemma 42 (Bias is small) *The expectation of $\hat{F}$ is close to the expectation of $\tilde{F}$. Specifically, for any $i \in [m]$, conditioned on the event $E_{\phi,i}$,*

$$\sup_{z\in\mathbb{R}}\left|\mathbb{E}\left[\hat{F}^{(i)}(z)\right] - \mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right| \le \frac{2K_{max}(D_2 - D_1)}{\pi h}\left(\max_{t\in[-\frac{1}{h},\frac{1}{h}]}\left|\phi_N(t) - \hat{\phi}_{N,i}(t)\right| + \rho\right).$$

Recall that the kernel bandwidth parameter $h = (4\gamma)^{\frac{1}{\beta}}(\log|\mathcal{B}_i|)^{-\frac{1}{\beta}}$.

Proof [Proof of Lemma 42] We want to show that

$$\sup_{z\in[D_1,D_2]}\left|\mathbb{E}\left[\hat{F}^{(i)}(z) - \tilde{F}^{(i)}(z)\right]\right|$$

is small. Here, expectation is taken with respect to data generation process, which can subdivided to the generation of $\{Z(i,j)\}_{j\in\mathcal{B}_i}$ and $\{N(i',j_1) - N(i',j_2)\}_{(i',j_1,j_2)\in\mathcal{T}_i}$, which are independent from each other (see the construction of the set $\mathcal{T}_i$).

$$
\begin{aligned}
&\mathbb{E}\left[\hat{F}^{(i)}(z) - \tilde{F}^{(i)}(z)\right] \\
&\quad = \mathbb{E}\left[\int_{D_1}^{z\wedge D_2} \hat{f}^{(i)}(w) - \tilde{f}^{(i)}(w) dw\right] \\
&\quad = \mathbb{E}\left[\int_{D_1}^{z\wedge D_2} \frac{1}{h|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} \hat{L}\left(\frac{w - Z(i,j)}{h}\right) - L\left(\frac{w - Z(i,j)}{h}\right) dw\right] \\
&\quad = \mathbb{E}\left[\int_{D_1}^{z\wedge D_2} \frac{1}{h|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\mathbf{i}t\frac{w-Z(i,j)}{h}}\left[\frac{\phi_K(t)}{\hat{\phi}_{N,i}(\frac{t}{h}) + \rho} - \frac{\phi_K(t)}{\phi_N(\frac{t}{h})}\right] dt\, dw\right] \\
&\quad = \mathbb{E}\left[\int_{D_1}^{z\wedge D_2} \frac{1}{h|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\mathbf{i}t\frac{w-Z(i,j)}{h}}\phi_K(t)\frac{\phi_N(\frac{t}{h}) - \left[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\right]}{\phi_N(\frac{t}{h})\left[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\right]} dt\, dw\right]. \quad (86)
\end{aligned}
$$

Noting that the support of ϕ_K is contained in $[-1, 1]$ and that the integrand is a bounded continuous function, we exchange the order of integrals.

$$
\begin{aligned}
Eq.(86) &= \int_{D_1}^{z\wedge D_2} \mathbb{E}\left[\frac{1}{h|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\mathbf{i}t\frac{w-Z(i,j)}{h}}\phi_K(t)\frac{\phi_N(\frac{t}{h}) - \left[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\right]}{\phi_N(\frac{t}{h})\left[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\right]} dt\right] dw \\
&= \int_{D_1}^{z\wedge D_2} \frac{1}{h|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} \frac{1}{2\pi}\mathbb{E}\left[\int_{-\infty}^{\infty} e^{-\mathbf{i}t\frac{w-Z(i,j)}{h}}\phi_K(t)\frac{\phi_N(\frac{t}{h}) - \left[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\right]}{\phi_N(\frac{t}{h})\left[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\right]} dt\right] dw \\
&= \int_{D_1}^{z\wedge D_2} \frac{1}{h|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathbb{E}\left[e^{-\mathbf{i}t\frac{w-Z(i,j)}{h}}\phi_K(t)\frac{\phi_N(\frac{t}{h}) - \left[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\right]}{\phi_N(\frac{t}{h})\left[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\right]}\right] dt\, dw \\
&= \int_{D_1}^{z\wedge D_2} \frac{1}{h|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\mathbf{i}t\frac{w}{h}}\mathbb{E}\left[e^{\mathbf{i}\frac{t}{h}Z(i,j)}\right]\phi_K(t)\frac{\phi_N(\frac{t}{h}) - \left[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\right]}{\phi_N(\frac{t}{h})\left[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\right]} dt\, dw.
\end{aligned} \quad (87)
$$

Recall that $\hat{\phi}_{N,i}$ estimates ϕ_N using data other than those from the i-th row, and hence, $Z(i,j)$ is independent of $\hat{\phi}_{N,i}$. $\mathbb{E}\left[e^{\mathbf{i}\frac{t}{h}Z(i,j)}\right]$ is the moment generating function of $Z(i,j)$ evaluated at $\frac{t}{h}$. Since $Z = A + N$ is the independent sum of $A \sim F^{(i)}$ and N, the moment generating function of Z is equal to the product of those, i.e.,

$$
\mathbb{E}\left[e^{\mathbf{i}\frac{t}{h}Z(i,j)}\right] = \phi_{Z(i,j)}\left(\frac{t}{h}\right) = \phi_{F^{(i)}}\left(\frac{t}{h}\right)\phi_N\left(\frac{t}{h}\right).
$$

Therefore,

$$
\begin{aligned}
&Eq.(87)\\
&= \int_{D_1}^{z\wedge D_2} \frac{1}{h|\mathcal{B}_i|} \sum_{j\in\mathcal{B}_i} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\mathbf{i}t\frac{w}{h}} \phi_{X^{(i)}}\Big(\frac{t}{h}\Big)\phi_N\Big(\frac{t}{h}\Big)\phi_K(t) \frac{\phi_N(\frac{t}{h}) - \Big[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\Big]}{\phi_N(\frac{t}{h})\Big[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\Big]} dt\, dw\\
&= \int_{D_1}^{z\wedge D_2} \frac{1}{h|\mathcal{B}_i|} \sum_{j\in\mathcal{B}_i} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\mathbf{i}t\frac{w}{h}} \phi_{X^{(i)}}\Big(\frac{t}{h}\Big)\phi_K(t) \frac{\phi_N(\frac{t}{h}) - \Big[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\Big]}{\hat{\phi}_{N,i}(\frac{t}{h}) + \rho} dt\, dw
\end{aligned}
$$

In short,

$$
\begin{aligned}
&\Big|\sup_{z\in\mathbb{R}} \mathbb{E}\Big[\hat{F}^{(i)}(z) - \tilde{F}^{(i)}(z)\Big]\Big|\\
&\quad\le \left|\sup_{z\in\mathbb{R}} \int_{D_1}^{z\wedge D_2} \frac{1}{h|\mathcal{B}_i|} \sum_{j\in\mathcal{B}_i} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\mathbf{i}t\frac{w}{h}} \phi_{X^{(i)}}\Big(\frac{t}{h}\Big)\phi_K(t) \frac{\phi_N(\frac{t}{h}) - \Big[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\Big]}{\hat{\phi}_{N,i}(\frac{t}{h}) + \rho} dt\, dw\right|\\
&\quad\le \frac{D_2 - D_1}{2\pi h} \int_{-\infty}^{\infty} \left|\phi_{X^{(i)}}\Big(\frac{t}{h}\Big)\phi_K(t) \frac{\phi_N(\frac{t}{h}) - \Big[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\Big]}{\hat{\phi}_{N,i}(\frac{t}{h}) + \rho}\right| dt\\
&\quad\le \frac{D_2 - D_1}{2\pi h} \int_{-\infty}^{\infty} \left|\phi_{X^{(i)}}\Big(\frac{t}{h}\Big)\right| \Big|\phi_K(t)\Big| \left|\frac{\phi_N(\frac{t}{h}) - \Big[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\Big]}{\hat{\phi}_{N,i}(\frac{t}{h}) + \rho}\right| dt\\
&\quad\le \frac{D_2 - D_1}{2\pi h} \int_{-1}^{1} K_{max} \left|\frac{\phi_N(\frac{t}{h}) - \Big[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\Big]}{\hat{\phi}_{N,i}(\frac{t}{h}) + \rho}\right| dt \qquad (88)\\
&\quad\le \frac{K_{max}(D_2 - D_1)}{\pi h} \max_{t\in[-1,1]} \left|\frac{\phi_N(\frac{t}{h}) - \Big[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\Big]}{\hat{\phi}_{N,i}(\frac{t}{h}) + \rho}\right|. \qquad (89)
\end{aligned}
$$

Eq. (88) follows from our assumption that the support of ϕ_K is contained within $[-1, 1]$ and that there exists $K_{max} = \max_{t\in[-1,1]} |\phi_K(t)| < \infty$.

To further simplify the upper bound in Eq. (89), we remark that

$$
\frac{\phi_N(\frac{t}{h}) - \Big[\hat{\phi}_{N,i}(\frac{t}{h}) + \rho\Big]}{\hat{\phi}_{N,i}(\frac{t}{h}) + \rho} = \frac{\phi_N(\frac{t}{h}) - \hat{\phi}_{N,i}(\frac{t}{h}) - \rho}{\phi_N(\frac{t}{h}) - \Big[\phi_N(\frac{t}{h}) - \hat{\phi}_{N,i}(\frac{t}{h}) - \rho\Big]}.
$$

From the supersmooth assumption on the noise, for any $t \in [-1, 1]$,

$$
\begin{aligned}
\phi_N\left(\frac{t}{h}\right) &\ge \frac{1}{B}\exp\left(-\gamma\left|\frac{t}{h}\right|^{\beta}\right) = \frac{1}{B}\exp\left(-\frac{1}{4}t^{\beta}\log|\mathcal{B}_i|\right)\\
&= \frac{1}{B}|\mathcal{B}_i|^{-\frac{1}{4}t^{\beta}} \ge \frac{1}{B}|\mathcal{B}_i|^{-\frac{1}{4}}. \qquad (90)
\end{aligned}
$$

The kernel bandwidth parameter is chosen as $h = (4\gamma)^{\frac{1}{\beta}} (\log |\mathcal{B}_i|)^{-\frac{1}{\beta}}$.

The ridge parameter $\rho = |\mathcal{B}_i|^{-\frac{7}{24}}$ and $\left|\phi_N(\frac{t}{h}) - \hat{\phi}_N(\frac{t}{h})\right|$ is sufficiently small when conditioned on $E_{\phi,i}$ (see Appendix I.6 for the definition of the event $E_{\phi,i}$). Since $\left|\frac{\delta}{1-\delta}\right| \leq 2|\delta|$ given that $|\delta| \leq \frac{1}{2}$,

$$\begin{aligned}
\max_{t\in[-1,1]} \left| \frac{\phi_N(\frac{t}{h}) - \hat{\phi}_{N,i}(\frac{t}{h}) - \rho}{\hat{\phi}_{N,i}(\frac{t}{h}) + \rho} \right| &\leq 2 \max_{t\in[-1,1]} \left| \phi_N\left(\frac{t}{h}\right) - \hat{\phi}_{N,i}\left(\frac{t}{h}\right) - \rho \right| \\
&\leq 2 \max_{t\in[-1,1]} \left| \phi_N\left(\frac{t}{h}\right) - \hat{\phi}_{N,i}\left(\frac{t}{h}\right) \right| + 2\rho \\
&= 2 \max_{t\in[-\frac{1}{h},\frac{1}{h}]} \left| \phi_N\left(t\right) - \hat{\phi}_{N,i}\left(t\right) \right| + 2\rho
\end{aligned}$$

Plugging in this expression to Eq. (89) concludes the proof. ■

I.5. Concentration of $\hat{F}$

Lemma 43 *For each $i \in [m]$, the kernel smoothed ECDF $\hat{F}^{(i)}$ defined as in Eq. (50) uniformly concentrates to its expectation, i.e., $\forall z \in [D_1, D_2]$,*

$$\mathbb{P}\left(\left| \hat{F}^{(i)}(z) - \mathbb{E}\left[\hat{F}^{(i)}(z)\right] \right| > t \right) \leq 2\exp\left(\frac{-|\mathcal{B}_i|^{5/12}}{2C_4^2 \left(\log|\mathcal{B}_i|\right)^{\frac{2}{\beta}}} t^2 \right).$$

Proof [Proof of Lemma 43] Recall that the kernel smoothed ECDF $\hat{F}^{(i)}$ evaluated at z is a function of n_i independent random variables $\{Z(i,j)\}_{j\in\mathcal{B}_i}$, i.e., when z is fixed, $\hat{F}^{(i)}(z) : \mathbb{R}^{|\mathcal{B}_i|} \to \mathbb{R}$ such that

$$\hat{F}^{(i)}(z)\left[Z(i,j_1), \ldots, Z(i,j_{n_i})\right] = \int_{D_1}^{z\wedge D_2} \frac{1}{hn_i} \sum_{j\in\mathcal{B}_i} \hat{L}\left(\frac{w - Z(i,j)}{h}\right) dw,$$

where $\hat{L}(z) = \frac{1}{2\pi}\int e^{-\mathbf{i}tz} \frac{\phi_K(t)}{\hat{\phi}_N(\frac{t}{h})+\rho} dt$. We can show that $\hat{F}^{(i)}(z)$ considered as a function of measurements $\{Z(i,j_1), \ldots, Z(i,j_{n_i})\}$ satisfies the bounded difference condition (see Eq. (102)) as in the proof of Lemma 30.

We take a similar approach as in the proof of Lemma 30. Let $\zeta^n = (\zeta_1, \ldots, \zeta_n)$ and $\zeta_j^n = (\zeta_1, \ldots, \zeta_j', \ldots, \zeta_n)$ be two n-tuples of real numbers, which differ only at the j-th position. Then

$$\begin{aligned}
&\hat{F}^{(i)}(z)[\zeta^n] - \hat{F}^{(i)}(z)[\zeta_j^n] \\
&\quad= \frac{1}{hn} \int_{D_1}^{z\wedge D_2} \hat{L}\left(\frac{w-\zeta_j}{h}\right) - \hat{L}\left(\frac{w-\zeta_j'}{h}\right) dw \\
&\quad= \frac{1}{hn} \int_{D_1}^{z\wedge D_2} \frac{1}{2\pi} \int \left(e^{-\mathbf{i}t\frac{w-\zeta_j}{h}} - e^{-\mathbf{i}t\frac{w-\zeta_j'}{h}} \right) \frac{\phi_K(t)}{\hat{\phi}_N\left(\frac{t}{h}\right) + \rho} dt dw \\
&\quad\leq \frac{1}{2\pi hn} \int_{D_1}^{z\wedge D_2} \int \left| e^{-\mathbf{i}t\frac{w-\zeta_j}{h}} - e^{-\mathbf{i}t\frac{w-\zeta_j'}{h}} \right| \left| \frac{\phi_K(t)}{\hat{\phi}_N\left(\frac{t}{h}\right) + \rho} \right| dt dw. \qquad (91)
\end{aligned}$$

Because $e^{-\mathbf{i}tz}$ is on the unit circle in the complex plane for any real numbers t and z, we have

$$\left|e^{-\mathbf{i}t\frac{w-\zeta_j}{h}} - e^{-\mathbf{i}t\frac{w-\zeta_j'}{h}}\right| \leq \left|e^{-\mathbf{i}t\frac{w-\zeta_j}{h}}\right| + \left|e^{-\mathbf{i}t\frac{w-\zeta_j'}{h}}\right| = 2.$$

Since ϕ_K is assumed to have compact support (see Appendix L.2) within $[-1,1]$, and a Fourier transform of L^1 function is uniformly continuous, there exists $K_{max} = \max_{t\in[-1,1]} |\phi_K(t)| < \infty$ such that $|\phi_K(t)| \leq K_{max}, \forall t$. From the algorithm description in Section E.1, $\rho = n^{-7/24}$ (here, $n = |\mathcal{B}_i|$ is the generic variable which stands for the number of samples in a row). By definition, $\hat{\phi}_N\left(\frac{t}{h}\right) \geq 0,\ \forall t$, and hence, $\hat{\phi}_N\left(\frac{t}{h}\right) + \rho \geq \rho,\ \forall t$.

We choose the bandwidth parameter $h = (4\gamma)^{\frac{1}{\beta}} (\log n)^{-\frac{1}{\beta}}$ following Fan (Theorems 56, 57). Plugging these expresions into Eq. (91) leads to

$$\begin{aligned}
Eq.(91) &\leq \frac{(\log n)^{\frac{1}{\beta}}}{2\pi\,(4\gamma)^{\frac{1}{\beta}}\,n} \int_{D_1}^{z\wedge D_2} \int_{-1}^{1} 2K_{max} n^{7/24} dt dw \\
&\leq \frac{K_{max}\,(\log n)^{\frac{1}{\beta}}}{\pi\,(4\gamma)^{\frac{1}{\beta}}\,n^{17/24}} \int_{D_1}^{z\wedge D_2} (1-(-1))\,dw \\
&\leq \frac{2K_{max}\,(D_2-D_1)\,(\log n)^{\frac{1}{\beta}}}{\pi\,(4\gamma)^{\frac{1}{\beta}}\,n^{17/24}} \\
&\leq \frac{2C_4\,(\log n)^{\frac{1}{\beta}}}{n^{17/24}}, \quad \text{for any } z \in [D_1, D_2]\,.
\end{aligned}$$

The last line follows from the definition of C_4 and the fact that $B \geq 1$ in our model.

Applying McDiarmid's inequality (Lemma 55), we can conclude that for any $z \in [D_1, D_2]$,

$$\mathbb{P}\left(\left|\hat{F}^{(i)}(z)[\zeta^n] - \mathbb{E}_{\zeta^n}\hat{F}^{(i)}(z)[\zeta^n]\right| \geq t\right) \leq 2\exp\left(\frac{-n^{5/12}}{2C_4^2\,(\log n)^{\frac{2}{\beta}}} t^2\right).$$

This argument holds for every $i \in [m]$, with replacing generic variable n with corresponding $|\mathcal{B}_i|$. ■

Lemma 44 (Variance is uniformly small) *For each $i \in [m]$, the kernel smoothed ECDF $\hat{F}^{(i)}$ defined as in Eq. (50) uniformly concentrates to its expectation, i.e., for any nonnegative integer N and for any $t \geq \frac{\Delta^{(i)}(D_2-D_1)}{N}$ (we define $\Delta^{(i)} := \frac{K_{max}}{\pi(4\gamma)^{\frac{1}{\beta}}} |\mathcal{B}_i|^{\frac{7}{24}} (\log|\mathcal{B}_i|)^{\frac{1}{\beta}}$),*

$$\begin{aligned}
&\mathbb{P}\left(\sup_{z\in[D_1,D_2]} \left|\hat{F}^{(i)}(z) - \mathbb{E}\left[\hat{F}^{(i)}(z)\right]\right| \geq t\right) \\
&\qquad \leq 2N\exp\left(\frac{-|\mathcal{B}_i|^{5/12}}{2C_4^2\,(\log|\mathcal{B}_i|)^{\frac{2}{\beta}}} \left(t - \frac{\Delta^{(i)}\,(D_2-D_1)}{N}\right)^2\right),
\end{aligned}$$

where $\beta, \gamma > 0$ are smoothness parameters for the noise, and $K_{max} = \max_{t\in[-1,1]} |\phi_K(t)|$.

Proof [Proof of Lemma 44] First, we discretize the interval interval $[D_1, D_2]$ by constructing a finite ε-net. For any $N \geq 1$, define the set

$$\mathcal{T}_N := \left\{ D_{min} + \frac{2k-1}{2N} (D_2 - D_1), \ \forall k \in [N] \right\}.$$

Then for any $N > 0$, $\mathcal{T}_N \subset [D_1, D_2]$ and it forms a $\frac{(D_2-D_1)}{2N}$-net with $|\mathcal{T}_N| = N$, i.e., for any $z \in [D_1, D_2]$, there exists $k \in [N]$ such that $\left| z - \frac{2k-1}{2N} (D_2 - D_1) \right| \leq \frac{(D_2-D_1)}{2N}$.

We can observe that

$$\begin{aligned}
\left\| \hat{f}^{(i)} \right\|_\infty &= \left\| \frac{1}{h \left| \mathcal{B}_i \right|} \sum_{j \in \mathcal{B}_i} \hat{L} \left(\frac{z - Z(i,j)}{h} \right) \right\|_\infty \\
&\leq \frac{1}{h} \left\| \hat{L} \right\|_\infty \\
&= \frac{1}{2\pi h} \left\| \int_{-\infty}^{\infty} e^{-\mathbf{i}tz} \frac{\phi_K(t)}{\hat{\phi}_{N,i}\left(\frac{t}{h}\right) + \rho} dt \right\|_\infty \\
&\leq \frac{1}{2\pi h} \int_{-\infty}^{\infty} \left| e^{-\mathbf{i}tz} \frac{\phi_K(t)}{\hat{\phi}_{N,i}\left(\frac{t}{h}\right) + \rho} \right| dt \\
&\leq \frac{1}{2\pi h} \int_{-\infty}^{\infty} \left| e^{-\mathbf{i}tz} \frac{\phi_K(t)}{\rho} \right| dt \\
&\leq \frac{1}{2\pi h} \int_{-1}^{1} K_{max} \left| \mathcal{B}_i \right|^{\frac{7}{24}} dt \\
&\leq \frac{(\log |\mathcal{B}_i|)^{\frac{1}{\beta}}}{2\pi (4\gamma)^{\frac{1}{\beta}}} \int_{-1}^{1} K_{max} \left| \mathcal{B}_i \right|^{\frac{7}{24}} dt \qquad \because h = (4\gamma)^{\frac{1}{\beta}} (\log |\mathcal{B}_i|)^{-\frac{1}{\beta}} \\
&\leq \frac{K_{max}}{\pi (4\gamma)^{\frac{1}{\beta}}} \left| \mathcal{B}_i \right|^{\frac{7}{24}} (\log |\mathcal{B}_i|)^{\frac{1}{\beta}}.
\end{aligned}$$

Let $\Delta^{(i)}$ denote the upper bound in the last line. Since this upper bound is universal for all realization of samples, $\left\| \mathbb{E}\left[\hat{f}^{(i)} \right] \right\|_\infty \leq \Delta^{(i)}$, too. Then $\left\| \hat{f}^{(i)} - \mathbb{E}\left[\hat{f}^{(i)} \right] \right\|_\infty \leq 2\Delta^{(i)}$ and it follows from the definition of $\hat{F}^{(i)}$ (see Eq. (50)) that

$$\sup_{z \in [D_1, D_2]} \left| \hat{F}^{(i)}(z) - \mathbb{E}\left[\hat{F}^{(i)}(z) \right] \right| \leq \sup_{z \in \mathcal{T}_N} \left| \hat{F}^{(i)}(z) - \mathbb{E}\left[\hat{F}^{(i)}(z) \right] \right| + \frac{\Delta^{(i)} (D_2 - D_1)}{N}.$$

Therefore, if $\left| \hat{F}^{(i)}(z) - \mathbb{E}\left[\hat{F}^{(i)}(z) \right] \right| \leq \varepsilon$ for all $z \in \mathcal{T}_n$, the supremum over the whole domain is bounded above up to an additional term as

$$\sup_{z \in [D_1, D_2]} \left| \hat{F}^{(i)}(z) - \mathbb{E}\left[\hat{F}^{(i)}(z) \right] \right| \leq \varepsilon + \frac{\Delta^{(i)} (D_2 - D_1)}{N}.$$

An application of the union bound on the contraposition of the previous statement yields

$$\begin{aligned}
&\mathbb{P}\left(\sup_{z\in[D_1,D_2]}\left|\hat{F}^{(i)}(z)-\mathbb{E}\left[\hat{F}^{(i)}(z)\right]\right|\geq t\right)\\
&\quad\leq\mathbb{P}\left(\sup_{z\in\mathcal{T}_N}\left|\hat{F}^{(i)}(z)-\mathbb{E}\left[\hat{F}^{(i)}(z)\right]\right|\geq t-\frac{\Delta^{(i)}\left(D_2-D_1\right)}{N}\right)\\
&\quad\leq\sum_{z\in\mathcal{T}_N}\mathbb{P}\left(\left|\hat{F}^{(i)}(z)-\mathbb{E}\left[\hat{F}^{(i)}(z)\right]\right|\geq t-\frac{\Delta^{(i)}\left(D_2-D_1\right)}{N}\right)\\
&\quad\leq 2N\exp\left(\frac{-\left|\mathcal{B}_i\right|^{5/12}}{2C_4^2\left(\log\left|\mathcal{B}_i\right|\right)^{\frac{2}{\beta}}}\left(t-\frac{\Delta^{(i)}\left(D_2-D_1\right)}{N}\right)^2\right).
\end{aligned}$$

■

I.6. Conditioning Events

For analysis, we define some conditioning events.

$$\begin{aligned}
E_J &\equiv \left\{|J|\geq\frac{1}{4}n\right\},\\
E_{\mathcal{T}_i} &\equiv \left\{|\mathcal{T}_i|\geq\frac{1}{512}mnp\right\},\\
E_{\mathcal{T}} &\equiv \left\{|\mathcal{T}|\leq mnp\right\},\\
E_{\Delta A} &\equiv \left\{\left|A(i,j_1)-A(i,j_2)\right|\leq\frac{c_{\Delta A}}{\sqrt{mp}}+\frac{2L\sqrt{2}}{\sqrt{np}}(1+\sqrt[4]{np}),\ \forall(i,j_1,j_2)\in\mathcal{T}\right\},\\
E_{\Delta N} &\equiv \left\{\left|N(i,j_1)-N(i,j_2)\right|\leq 4\sigma\sqrt{\log(4mnp)},\ \forall(i,j_1,j_2)\in\mathcal{T}\right\},\\
E_{\phi,i} &\equiv \left\{\sup_{t\in[-\frac{1}{h},\frac{1}{h}]}\left|\hat{\phi}_{N,i}(t)-\phi_N(t)\right|\leq s_\phi\right\}.
\end{aligned}$$

Here, $c_{\Delta A}=8\sqrt{\pi}\left(\frac{\sqrt{e^{C_1}}+\sqrt{2}}{\sqrt{C_1}}+\frac{\sqrt{e^{C_2}}+\sqrt{2}}{\sqrt{C_2}}\right)$ and $s_\phi=s_1+s_2$ where $s_1=\frac{8\sigma(\log|\mathcal{B}_i|)^{\frac{1}{\beta}}}{(4\gamma)^{\frac{1}{\beta}}}\frac{\sqrt{\log(4mnp)}}{(mnp)^{\frac{1}{4}}}$ and $s_2=\frac{2(\log|\mathcal{B}_i|)^{\frac{1}{\beta}}}{(4\gamma)^{\frac{1}{\beta}}}\left[\frac{c_{\Delta A}}{\sqrt{mp}}+\frac{2L\sqrt{2}}{\sqrt{np}}(1+\sqrt[4]{np})\right]$.

We analyze probabilities of these conditioning events, which will be used in the proof of Lemma 25 in the next section. We may assume $m,n\gg 1$ so that $mp\geq 8\ln 2$ and $np\geq 48>32\ln 2$. These assumptions are arbitrary and can be removed; the only purpose of these assumptions are to simplify the following probabilistic bounds.

1. E_J: $\mathbb{P}(E_J^c)$ is small. Since $mp \geq 8 \ln 2$, $\exp\left(-\frac{mp}{8}\right) \leq \frac{1}{2}$. By Lemma 32,

$$\begin{aligned}\mathbb{P}(E_J^c) &\leq \mathbb{P}\left(|J| \leq \frac{n\left[1-\exp\left(-\frac{mp}{8}\right)\right]}{2}\right) \\ &\leq \exp\left(-\frac{n\left[1-\exp\left(-\frac{mp}{8}\right)\right]}{8}\right) \\ &\leq \exp\left(-\frac{n}{16}\right). \end{aligned} \tag{92}$$

2. $E_{\mathcal{T}_i}$: $\mathbb{P}\left(E_{\mathcal{T}_i}^c | E_J\right)$ is small. Conditioned on E_J, $|J| \geq \frac{1}{4}n$ and $|J|p \geq \frac{np}{4}$. Therefore,

$$\begin{aligned}&\frac{m\left[1-\exp\left(-\frac{|J|p}{8}\right)\right]}{2} - 1 \geq \frac{m}{4} - 1 \geq \frac{m}{8}, \quad \text{and} \\ &\left\lceil \frac{\frac{|J|p}{2} - 1 - \lfloor\sqrt{\frac{|J|p}{2}}\rfloor}{2} \right\rceil \geq \left\lceil \frac{\frac{np}{8} - 1 - \lfloor\sqrt{\frac{np}{8}}\rfloor}{2} \right\rceil \geq \frac{1}{4}\left(\frac{np}{8} - 1\right) \geq \frac{np}{64}.\end{aligned}$$

For any $i \in [m]$ Lemma 33 asserts that

$$\begin{aligned}\mathbb{P}\left(E_{\mathcal{T}_i}^c | E_J\right) &\leq \mathbb{P}\left(|\mathcal{T}_i| < \left(\frac{m\left[1-\exp\left(-\frac{|J|p}{8}\right)\right]}{2} - 1\right)\left\lceil \frac{\frac{|J|p}{2} - 1 - \lfloor\sqrt{\frac{|J|p}{2}}\rfloor}{2} \right\rceil \middle| E_J\right) \\ &\leq \exp\left(-\frac{m\left[1-\exp\left(-\frac{|J|p}{8}\right)\right]}{8}\right)\Bigg|_{|J|\geq\frac{1}{4}n} \\ &\leq \exp\left(-\frac{m}{16}\right). \end{aligned} \tag{93}$$

3. $E_{\mathcal{T}}$: $\mathbb{P}(E_{\mathcal{T}}^c)$ is small. Lemma 34 ensure that

$$\mathbb{P}(E_{\mathcal{T}}^c) \leq \exp\left(-\frac{mnp}{3}\right). \tag{94}$$

4. $E_{\Delta A}$: $\mathbb{P}(E_{\Delta A}^c | E_J)$ is small. Conditioned on E_J, $|J| \geq \frac{n}{4}$. Hence,

$$\begin{aligned}&L\sqrt{\frac{2}{|J|\,p}} + 4LQ^*\left(\frac{mp}{2}\right) \\ &\leq \frac{2L\sqrt{2}}{\sqrt{np}} + 8L\sqrt{\pi}\left(\sqrt{\frac{2}{C_1 mp}} + \sqrt{\frac{2}{C_2 mp}} + \sqrt{\frac{e^{C_1}}{C_1 mp}} + \sqrt{\frac{e^{C_2}}{C_2 mp}}\right) \\ &= \frac{2L\sqrt{2}}{\sqrt{np}} + \frac{c_{\Delta A}}{\sqrt{mp}}.\end{aligned}$$

Then $\frac{c_{\Delta A}}{\sqrt{mp}} + \frac{2L\sqrt{2}}{\sqrt{np}}(1+\sqrt[4]{np}) - \left[L\sqrt{\frac{2}{|J|p}} + 4LQ^*\left(\frac{mp}{2}\right)\right] \geq \frac{2L\sqrt{2}}{\sqrt[4]{np}}$. Note that $Q^*\left(\frac{mp}{2}\right) > 0$ and hence, by Lemma 35,

$$\begin{aligned}\mathbb{P}\left(E_{\Delta A}^c | E_J\right) &\leq n\exp\left(-\frac{n}{8L^2}\left(\frac{2L\sqrt{2}}{\sqrt[4]{np}}\right)^2\right) + n\exp\left(-\frac{n}{12L}\left(\frac{2L\sqrt{2}}{\sqrt[4]{np}}\right)\right) \\ &\leq n\exp\left(-n^{\frac{1}{2}}\right) + n\exp\left(-\frac{1}{3\sqrt{2}}n^{\frac{3}{4}}\right). \end{aligned} \tag{95}$$

We used the fact $J \subset [n]$ implies $|J| \leq n$ and $|J| \geq \frac{n}{4}$ when conditioned on E_J and that $p \leq 1$.

5. $E_{\Delta N}$: $\mathbb{P}\left(E_{\phi,i}^c | E_{\mathcal{T}_i}, E_{\Delta A}, E_{\Delta N}\right)$ is small. Conditioned on $E_{\mathcal{T}_i}$, $|\mathcal{T}| \geq |\mathcal{T}_i| \geq \frac{mnp}{512}$, while $E_{\mathcal{T}}$ ensures $|\mathcal{T}| \leq mnp$. Recall that Lemma 36 ascertains $\|\Delta N\|_{\infty,(i)}$ does not exceed $4\sigma\sqrt{\log 4|\mathcal{T}|}$ with high probability as

$$\mathbb{P}\left(\|\Delta N\|_{\infty,(i)} > 4\sigma\sqrt{\log 4|\mathcal{T}|}\right) \leq \frac{1}{4|\mathcal{T}|}.$$

If we combine this probabilistic bound with the conditioning events, then the following upper bound can be achieved:

$$\begin{aligned}\mathbb{P}\left(E_{\Delta N}^c | E_{\mathcal{T}_i}, E_{\mathcal{T}}\right) &\leq \mathbb{P}\left(\|\Delta N\|_{\infty,(i)} > 4\sigma\sqrt{\log 4|\mathcal{T}|} | E_{\mathcal{T}_i}, E_{\mathcal{T}}\right) \\ &\leq \frac{1}{4|\mathcal{T}|}\Bigg|_{E_{\mathcal{T}_i}, E_{\mathcal{T}}} \\ &\leq \frac{128}{mnp}. \end{aligned} \tag{96}$$

6. $E_{\phi,i}$: $\mathbb{P}\left(E_{\phi,i}^c | E_{\mathcal{T}_i}, E_{\Delta A}, E_{\Delta N}\right)$ is small. Conditioned on $E_{\mathcal{T}_i}, E_{\Delta A}, E_{\Delta N}$,

$$\begin{aligned}|\mathcal{T}_i| &\geq \frac{mnp}{512} \\ \|\Delta A\|_{\infty,(i)} &\leq \frac{c_{\Delta A}}{\sqrt{mp}} + \frac{2L\sqrt{2}}{\sqrt{np}}(1+\sqrt[4]{np}) \\ \|\Delta N\|_{\infty,(i)} &\leq 4\sigma\sqrt{\log(4mnp)}. \end{aligned}$$

Now the length of our interval $\Lambda = \frac{1}{h} = (\frac{\log|\mathcal{B}_i|}{4\gamma})^{\frac{1}{\beta}}$.

If $mnp \gg 1$ so that $\log(4mnp)(\log|\mathcal{B}_i|)^{\frac{1}{\beta}} \geq \frac{B(4\gamma)^{\frac{1}{\beta}}}{4\sigma}$, then

$$\left\|\Delta^{*(i)}_{N_1,\frac{1}{h}}\right\|_\infty = \frac{1}{N_1}\left[\frac{1}{h^2}\|\Delta N\|^2_{\infty,(i)} + \frac{2}{h}\sigma B\right] \leq \frac{32\sigma^2(\log|\mathcal{B}_i|)^{\frac{2}{\beta}}}{N_1(4\gamma)^{\frac{2}{\beta}}}\log(4mnp).$$

Let $N_1 = \sqrt{mnp}$, and $s_1 = \frac{8\sigma(\log|\mathcal{B}_i|)^{\frac{1}{\beta}}}{(4\gamma)^{\frac{1}{\beta}}}\frac{\sqrt{\log(4mnp)}}{(mnp)^{\frac{1}{4}}}$.

$$\begin{aligned}
&2N_1 \exp\left(-\frac{|\mathcal{T}_i|}{2}\left(s_1^2 - \left\|\Delta^{*(i)}_{N_1,\Lambda}\right\|_\infty\right)^2\right) \\
&\quad\le 2\sqrt{mnp}\exp\left(-\frac{mnp}{1024}\left(\frac{32\sigma^2(\log|\mathcal{B}_i|)^{\frac{2}{\beta}}}{(4\gamma)^{\frac{2}{\beta}}}\frac{\log(4mnp)}{\sqrt{mnp}}\right)^2\right) \\
&\quad= \exp\left(-\frac{\sigma^4(\log|\mathcal{B}_i|)^{\frac{4}{\beta}}}{(4\gamma)^{\frac{4}{\beta}}}\log^2(4mnp) + \log(4mnp)\right).
\end{aligned}$$

Similarly, (assume $N_2 \ge 2$)

$$\begin{aligned}
\left\|\Delta^{(i)}_{N_2,\frac{1}{h}}\right\|_\infty &= \frac{N_2+2}{2N_2h^2}\|\Delta A\|^2_{\infty,(i)} + \frac{2}{N_2h^2}\|\Delta A\|_{\infty,(i)}\|\Delta N\|_{\infty,(i)} \\
&\le \frac{(\log|\mathcal{B}_i|)^{\frac{2}{\beta}}}{(4\gamma)^{\frac{2}{\beta}}}\left[\frac{c_{\Delta A}}{\sqrt{mp}} + \frac{2L\sqrt{2}}{\sqrt{np}}(1+\sqrt[4]{np})\right]^2 \\
&\quad + \frac{8\sigma}{N_2}\frac{(\log|\mathcal{B}_i|)^{\frac{2}{\beta}}}{(4\gamma)^{\frac{2}{\beta}}}\left[\frac{c_{\Delta A}}{\sqrt{mp}} + \frac{2L\sqrt{2}}{\sqrt{np}}(1+\sqrt[4]{np})\right]\sqrt{\log(4mnp)}.
\end{aligned}$$

Let

$$\begin{aligned}
N_2 &= 8\sigma\sqrt{\log(4mnp)}\left[\frac{c_{\Delta A}}{\sqrt{mp}} + \frac{2L\sqrt{2}}{\sqrt{np}}(1+\sqrt[4]{np})\right]^{-1} \\
&\le \frac{8\sigma}{c_{\Delta A} + 2L\sqrt{2}}\sqrt{mnp}\sqrt{\log(4mnp)},
\end{aligned}$$

and

$$s_2 = \frac{2(\log|\mathcal{B}_i|)^{\frac{1}{\beta}}}{(4\gamma)^{\frac{1}{\beta}}}\left[\frac{c_{\Delta A}}{\sqrt{mp}} + \frac{2L\sqrt{2}}{\sqrt{np}}(1+\sqrt[4]{np})\right].$$

Then,

$$
\begin{aligned}
&2N_2 \exp\left(-\frac{|\mathcal{T}_i|}{2\Lambda^2\|\Delta A\|^2_{\infty,(i)}}\left(s_2^2 - \left\|\Delta^{(i)}_{N_2,\Lambda}\right\|_\infty\right)^2\right) \\
&\quad\le 2N_2 \exp\left(-\frac{|\mathcal{T}_i|}{2\Lambda^2\|\Delta A\|^2_{\infty,(i)}}\frac{4(\log|\mathcal{B}_i|)^{\frac{4}{\beta}}}{(4\gamma)^{\frac{4}{\beta}}}\left[\frac{c_{\Delta A}}{\sqrt{mp}} + \frac{2L\sqrt{2}}{\sqrt{np}}(1+\sqrt[4]{np})\right]^4\right) \\
&\quad\le 2N_2 \exp\left(-2|\mathcal{T}_i|\frac{(\log|\mathcal{B}_i|)^{\frac{2}{\beta}}}{(4\gamma)^{\frac{2}{\beta}}}\left[\frac{c_{\Delta A}}{\sqrt{mp}} + \frac{2L\sqrt{2}}{\sqrt{np}}(1+\sqrt[4]{np})\right]^2\right) \\
&\quad\le 2N_2 \exp\left(-\frac{mnp}{256}\frac{(\log|\mathcal{B}_i|)^{\frac{2}{\beta}}}{(4\gamma)^{\frac{2}{\beta}}}\left[\frac{c_{\Delta A}}{\sqrt{mp}} + \frac{2L\sqrt{2}}{\sqrt{np}}(1+\sqrt[4]{np})\right]^2\right) \\
&\quad\le \exp\Bigg(-\frac{(\log|\mathcal{B}_i|)^{\frac{2}{\beta}}}{256(4\gamma)^{\frac{2}{\beta}}}\left[c_{\Delta A}\sqrt{n} + 2L\sqrt{2m}\right]^2 \\
&\qquad\qquad + \frac{1}{2}\Big(\log mnp + \log\log(4mnp)\Big) + \log\frac{16\sigma}{c_{\Delta A} + 2L\sqrt{2}}\Bigg).
\end{aligned}
$$

All in all,

$$
\begin{aligned}
&\mathbb{P}\left(E^c_{\phi,i}\middle|E_{\mathcal{T}_i}, E_{\Delta A}, E_{\Delta N}\right) \qquad (97)\\
&\quad\le \exp\left(-\frac{\sigma^4(\log|\mathcal{B}_i|)^{\frac{4}{\beta}}}{(4\gamma)^{\frac{4}{\beta}}}\log^2(4mnp) + \log(4mnp)\right) \\
&\qquad + \exp\Bigg(-\frac{(\log|\mathcal{B}_i|)^{\frac{2}{\beta}}}{256(4\gamma)^{\frac{2}{\beta}}}\left[c_{\Delta A}\sqrt{n} + 2L\sqrt{2m}\right]^2 \qquad (98)\\
&\qquad\qquad + \frac{1}{2}\Big(\log mnp + \log\log(4mnp)\Big) + \log\frac{16\sigma}{c_{\Delta A} + 2L\sqrt{2}}\Bigg).
\end{aligned}
$$

I.7. Proof of Lemma 25

Proof [Proof of Lemma 25] By the usual trick of applying triangle inequality, we have

$$
\begin{aligned}
&\sup_{z\in[D_1,D_2]}\left|\hat{F}^{(i)}(z) - F^{(i)}\right| \\
&\quad= \sup_{z\in[D_1,D_2]}\left|\hat{F}^{(i)}(z) - \mathbb{E}\left[\hat{F}^{(i)}(z)\right] + \mathbb{E}\left[\hat{F}^{(i)}(z)\right] - \mathbb{E}\left[\tilde{F}^{(i)}(z)\right] + \mathbb{E}\left[\tilde{F}^{(i)}(z)\right] - F^{(i)}\right| \\
&\quad\le \sup_{z\in[D_1,D_2]}\left|\hat{F}^{(i)}(z) - \mathbb{E}\left[\hat{F}^{(i)}(z)\right]\right| + \sup_{z\in[D_1,D_2]}\left|\mathbb{E}\left[\hat{F}^{(i)}(z)\right] - \mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right| \\
&\qquad + \sup_{z\in[D_1,D_2]}\left|\mathbb{E}\left[\tilde{F}^{(i)}(z)\right] - F^{(i)}\right|.
\end{aligned}
$$

If $\sup_{z\in[D_1,D_2]}\left|\hat{F}^{(i)}(z)-\mathbb{E}\left[\hat{F}^{(i)}(z)\right]\right|\le t_1$, $\sup_{z\in[D_1,D_2]}\left|\mathbb{E}\left[\hat{F}^{(i)}(z)\right]-\mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right|\le t_2$, and $\sup_{z\in[D_1,D_2]}\left|\mathbb{E}\left[\tilde{F}^{(i)}(z)\right]-F^{(i)}\right|\le t_3$, then $\sup_{z\in[D_1,D_2]}\left|\hat{F}^{(i)}(z)-F^{(i)}\right|\le t_1+t_2+t_3$. Applying union bound on the contrapositive yields

$$\begin{aligned}
&\mathbb{P}\left(\sup_{z\in[D_1,D_2]}\left|\hat{F}^{(i)}(z)-F^{(i)}\right|>t_1+t_2+t_3\right)\\
&\quad\le\mathbb{P}\left(\sup_{z\in[D_1,D_2]}\left|\hat{F}^{(i)}(z)-\mathbb{E}\left[\hat{F}^{(i)}(z)\right]\right|>t_1\right) &(99)\\
&\quad+\mathbb{P}\left(\sup_{z\in[D_1,D_2]}\left|\mathbb{E}\left[\tilde{F}^{(i)}(z)\right]-F^{(i)}(z)\right|>t_2\right) &(100)\\
&\quad+\mathbb{P}\left(\sup_{z\in[D_1,D_2]}\left|\mathbb{E}\left[\hat{F}^{(i)}(z)\right]-\mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right|>t_3\right). &(101)
\end{aligned}$$

1. Eq. (99): Eq. (99) is bounded by Lemma 44. We take integer $N=\frac{1}{2}|\mathcal{B}_i|^{\frac{1}{6}}$. Then for any $t_1\ge\frac{2\Delta^{(i)}(D_2-D_1)}{N}=\frac{4K_{max}(D_2-D_1)}{\pi(4\gamma)^{\frac{1}{\beta}}}|\mathcal{B}_i|^{-\frac{5}{24}}(\log|\mathcal{B}_i|)^{\frac{1}{\beta}}$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{z\in[D_1,D_2]}\left|\hat{F}^{(i)}(z)-\mathbb{E}\left[\hat{F}^{(i)}(z)\right]\right|\ge t_1\right)\\
&\quad\le 2N\exp\left(\frac{-|\mathcal{B}_i|^{5/12}}{2C_4^2(\log|\mathcal{B}_i|)^{\frac{2}{\beta}}}\left(t_1-\frac{\Delta^{(i)}(D_2-D_1)}{N}\right)^2\right)\\
&\quad\le|\mathcal{B}_i|^{\frac{1}{6}}\exp\left(\frac{-|\mathcal{B}_i|^{5/12}}{8C_4^2(\log|\mathcal{B}_i|)^{\frac{2}{\beta}}}t_1^2\right),
\end{aligned}$$

where $\beta,\gamma>0$ are smoothness parameters for the noise, and $K_{max}=\max_{t\in[-1,1]}|\phi_K(t)|$.

2. Eq. (100): If we take $t_2=C_3(\log|\mathcal{B}_i|)^{-1/\beta}$, the probability in Eq. (100) becomes 0 by Lemma 29.

3. Eq. (101): We further partition the probability in Eq. (101) by conditioning events defined in Section I.6.

$$Eq.(101)\le\mathbb{P}\left(\sup_{z\in[D_1,D_2]}\left|\mathbb{E}\left[\hat{F}^{(i)}(z)\right]-\mathbb{E}\left[\tilde{F}^{(i)}(z)\right]\right|>t_3\middle|E_{\phi,i}\right)+\mathbb{P}\left(E_{\phi,i}^c\right).$$

The first term is bounded by Lemma 42: the conditional probability becomes 0 if we choose $t_3=\frac{2K_{max}(D_2-D_1)}{\pi h}(s_\phi+\rho)$.

It remains to analyze $\mathbb{P}\left(E_{\phi,i}^c\right)$.

$$\begin{aligned}
\mathbb{P}\left(E_{\phi,i}^c\right)&\le\mathbb{P}\left(E_{\phi,i}^c\middle|E_{\mathcal{T}_i}\cap E_{\Delta A}\cap E_{\Delta N}\right)+\mathbb{P}\left(E_{\mathcal{T}_i}^c\cup E_{\Delta A}^c\cup E_{\Delta N}^c\right)\\
&=\mathbb{P}\left(E_{\phi,i}^c\middle|E_{\mathcal{T}_i}\cap E_{\Delta A}\cap E_{\Delta N}\right)+\mathbb{P}\left(E_{\mathcal{T}_i}^c\cup E_{\Delta A}^c\right)+\mathbb{P}\left(E_{\Delta N}^c\cap E_{\mathcal{T}_i}\right).
\end{aligned}$$

The first term is small (see Eq. (98)).
The second term:

$$\begin{aligned}\mathbb{P}\left(E_{\mathcal{T}_i}^c \cup E_{\Delta A}^c\right) &\leq \mathbb{P}\left(E_J^c\right) + \mathbb{P}\left(E_{\mathcal{T}_i}^c \cup E_{\Delta A}^c \middle| E_J\right) \\ &\leq \mathbb{P}\left(E_J^c\right) + \mathbb{P}\left(E_{\mathcal{T}_i}^c \middle| E_J\right) + \mathbb{P}\left(E_{\Delta A}^c \middle| E_J\right).\end{aligned}$$

See Eqs. (92), (93), (95).
The third term:

$$\begin{aligned}\mathbb{P}\left(E_{\Delta N}^c \cap E_{\mathcal{T}_i}\right) &\leq \mathbb{P}\left(E_{\Delta N}^c \cap E_{\mathcal{T}_i} \cap E_{\mathcal{T}}\right) + \mathbb{P}\left(E_{\mathcal{T}}^c\right) \\ &= \mathbb{P}\left(E_{\Delta N}^c \middle| E_{\mathcal{T}_i} \cap E_{\mathcal{T}}\right) \mathbb{P}\left(E_{\mathcal{T}_i} \cap E_{\mathcal{T}}\right) + \mathbb{P}\left(E_{\mathcal{T}}^c\right) \\ &\leq \mathbb{P}\left(E_{\Delta N}^c \middle| E_{\mathcal{T}_i} \cap E_{\mathcal{T}}\right) + \mathbb{P}\left(E_{\mathcal{T}}^c\right).\end{aligned}$$

See Eqs. (96) and (94).

To sum up, let $t_0 = C_3 \left(\log |\mathcal{B}_i|\right)^{-1/\beta} + \frac{2K_{max}(D_2 - D_1)}{\pi h} \left(s_\phi + \rho\right)$. Then we can conclude that for any $i \in [m]$, and for any $t \geq t_0 + \frac{4K_{max}(D_2 - D_1)}{\pi (4\gamma)^{\frac{1}{\beta}}} |\mathcal{B}_i|^{-\frac{5}{24}} \left(\log |\mathcal{B}_i|\right)^{\frac{1}{\beta}}$,

$$\begin{aligned}&\mathbb{P}\left(\sup_{z \in [D_1, D_2]} \left|\tilde{F}^{(i)}(z) - F^{(i)}(z)\right| > t + t_0\right) \\ &\quad \leq |\mathcal{B}_i|^{\frac{1}{6}} \exp\left(\frac{-|\mathcal{B}_i|^{5/12}}{8C_4^2 \left(\log |\mathcal{B}_i|\right)^{\frac{2}{\beta}}} (t - t_0)^2\right) + \tilde{\Psi}_{m,n,p}\left(|\mathcal{B}_i|\right).\end{aligned}$$

For completeness, we note that the Remainder term, $\tilde{\Psi}_{m,n,p}\left(|\mathcal{B}_i|\right)$ (see Eq. (63)), is the sum of upper bounds in Eq. (92) - (98), which vanishes as $m, n \to \infty$. ■

Proof [Proof of Corollary 26] Conditioned on event $E_{row,(i)}$, it holds for all $i \in [m]$ that $|\mathcal{B}_i| \geq \frac{np}{2}$. Similarly, $|\mathcal{B}_i| \leq 2np$ for all $i \in [m]$, when conditioned on event $E'_{row,(i)}$. Therefore, for any $i \in [m]$, and any $t \geq T_0^*$,

$$\begin{aligned}&\mathbb{P}\left(\sup_{z \in [D_1, D_2]} \left|\tilde{F}^{(i)}(z) - F^{(i)}(z)\right| > t \middle| E_{row,(i)}, E'_{row,(i)}\right) \\ &\quad \leq (2np)^{\frac{1}{6}} \exp\left(\frac{-\left(\frac{np}{2}\right)^{5/12}}{8C_4^2 \left(\log(2np)\right)^{\frac{2}{\beta}}} (t - t_0^*)^2\right) + \tilde{\Psi}_{m,n,p}\left(\frac{np}{2}\right).\end{aligned}$$

■

Appendix J. Known Facts about Distribution

J.1. Basic Definitions

In this section, we briefly restate some basic facts and functions related to a random variable. We let $(\Omega, \mathcal{F}, P)$ denote our probability space.

Definition 45 (Random variable) *A random variable $X : \Omega \to E$ is a measurable function from a set of possible outcomes Ω to a measurable space E. When $E = \mathbb{R}$, we call X a real-valued random variable.*

For a real-valued random variable X, we can define its distribution function, whose evaluation at x is the probability that X will take a value less than or equal to x.

Definition 46 (Cumulative distribution function (CDF)) *The cumulative distribution function of a real-valued random variable X is defined as a function $F_X : \mathbb{R} \to [0, 1]$ such that*

$$F_X(x) = \mathbb{P}(X \leq x).$$

Every cumulative distribution function F is non-decreasing, right-continuous, $\lim_{x \to -\infty} F(x) = 0$, and $\lim_{x \to \infty} F(x) = 1$. Conversely, every function with these four properties is a CDF, i.e., a random variable can be defined so that the function is the CDF of that random variable.

We can define a pseudo-inverse of the distribution function, which returns a threshold value x below which random draws from the given CDF would fall with given input probability p.

Definition 47 (Quantile function) *Given a distribution function $F : \mathbb{R} \to [0, 1]$, the associated quantile function $Q : (0, 1) \to \mathbb{R}$ is defined as*

$$Q(p) = \inf \{x \in \mathbb{R} : p \leq F(x)\}.$$

If the function F is continuous and strictly monotone increasing, then the infimum can be replaced by the minimum and $Q = F^{-1}$.

When F is absolutely continuous, then there exists a Lebesgue-integrable function $f(x)$ such that

$$F(b) - F(a) = \mathbb{P}(a < X \leq b) = \int_a^b f(x)dx,$$

for all real numbers a and b. The function f is the (Radon-Nikodym) derivative of F, and it is called the probability density function of distribution of X.

Note that the CDF can be expressed as the expectation of an indicator function, $F_X(x) = \mathbb{E}[\mathbb{I}\{X \leq x\}]$. There is an alternative way to describe a random variable.

Definition 48 (Characteristic function) *The characteristic function $\phi_X : \mathbb{R} \to \mathbb{C}$ for a real-valued random variable is defined as the expected value of e^{itX}, where i is the imaginary unit, and $t \in \mathbb{R}$ is the argument of the characteristic function:*

$$\begin{aligned}
\phi_X(t) &= \mathbb{E}\left[e^{itX}\right] \\
&= \int_{\mathbb{R}} e^{itx} dF_X(x) \\
&= \int_{\mathbb{R}} e^{itx} f_X(x)dx \\
&= \int_0^1 e^{itQ_X(p)} dp.
\end{aligned}$$

If random variable X has a probability density function f_X, then the characteristic function is the Fourier transform with sign reversal in the complex exponential (note that the constant differs from the usual convention for the Fourier transform).

J.2. Empirical CDF and Empirical Characteristic Function

Given $X_1, \ldots, X_n$ (n is a natural number) be real-valued independent and identically distributed random variables with common cumulative distribution function F. We let F_n denote the empirical distribution function associated with $\{X_1, \ldots, X_n\}$, which is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i \leq x\}, \quad , \forall x \in \mathbb{R}.$$

$F_n(x)$ is the average number of random variables among $\{X_1, \ldots, X_n\}$ which take value smaller than x.

It is knwon that the empirical distribution function converges to the distribution function from which the samples are drawn. The following concentration results known as the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality quantifies the rate of convergence of F_n to F with respect to the uniform norm as n tends to infinity. It is named after Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz, who proved the inequality in 1956 with an unspecified multiplicative constant C. Later in 1990, Pascal Massart proved the inequality with the sharp constant $C = 2$. This result strengthens the Glivenko-Cantelli theorem.

Lemma 49 (Dvoretzky-Kiefer-Wolfowitz) *Given a natural number n, let $X_1, \ldots, X_n$ be real-valued independent and identically distributed random variables with common cumulative distribution function F. Then for every $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}.$$

Appendix K. Sub-Gaussian Random Variable and the Chernoff Bound

First of all, we recall the Markov's inequality.

Theorem 50 (Markov's inequality) *Given a nonnegative random variable X, for all $t > 0$,*

$$\mathbb{P}(X \geq t) \frac{\leq \mathbb{E}[X]}{t}.$$

Proof For all $t > 0$, $t\mathbb{I}\{X \geq t\} \leq X\mathbb{I}\{X \geq t\} \leq X$. Taking expectation, $t\mathbb{P}(X \geq t) \leq \mathbb{E}[X]$, and hence, $\mathbb{P}(X \geq t) \frac{\leq \mathbb{E}[X]}{t}$. ■

Now let X be a real-valued random variable. Applying Markov's inequality with an exponential function, it follows that for $\lambda \geq 0$,

$$\mathbb{P}(X \geq t) = \mathbb{P}\left(e^{\lambda X} \geq e^{\lambda t}\right) \leq \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda t}}.$$

Since this inequality holds for all values of $\lambda \geq 0$, one may optimize λ to obtain the tightest tail bound.

Next, we define a class of random variables, whose tail behavior is easy to control.

Definition 51 (Sub-Gaussian random variable) *A random variable X with mean $\mu = \mathbb{E}[X]$ is called sub-Gaussian if there is a positive constant σ such that*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}, \quad \forall\lambda \in \mathbb{R}.$$

We will call σ the sub-Gaussian parameter of X.

An application of the Chernoff bound leads to

$$\mathbb{P}(X - \mu \geq t) \leq \inf_{\lambda} \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}},$$

where λ is optimized over the interval $[0, \lambda^*]$ in which the moment generating function of X exists. It is possible to achieve the same upper bound for $\mathbb{P}(X - \mu \leq -t) = \mathbb{P}(-(X - \mu) \geq t)$. We can conclude that a sub-Gaussian random variable satisfies that for all $t \in \mathbb{R}$,

$$\mathbb{P}(|X - \mu| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}.$$

The class of sub-Gaussian random variables subsumes Gaussian random variable and any bounded random variables.

Hoeffding-type Inequalities Now, we present several forms of concentration inequalities for the sum of independent random variables. Essentially they are all Chernoff bounds, tailored to specific random variable assumptions. We present three lemmas in the increasing order of generality, starting from the bound for a sum of independent Bernoulli trials.

Lemma 52 (Binomial Chernoff bound) *Let $X = \sum_{i=1}^n X_i$, where $X_i = 1$ with probability p_i, and $X_i = 0$ with probability $1 - p_i$, and X_i's are independent. Let $\mu = \mathbb{E}[X] = \sum_{i=1}^n p_i$. Then*

1. *Upper tail: $\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2}{2+\delta}\mu\right)$ for all $\delta > 0$.*
2. *Lower tail: $\mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(-\frac{\delta^2}{2}\mu\right)$ for all $0 < \delta < 1$.*

Hoeffding derived a more general result for bounded random variables, which is known as (Azuma-) Hoeffding's inequality.

Lemma 53 (Hoeffding's inequality for bounded ranom variables) *Let $X_1, \ldots, X_n$ be n independent random variables such that almost surely $X_i \in [a_i, b_i], \forall i$. Let $X = \sum_{i=1}^n X_i$, then for any $t > 0$,*

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

and

$$\mathbb{P}(X - \mathbb{E}[X] \leq -t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Although Hoeffding's inequality is often presented only for the special case of bounded random variables, the same idea applies to sub-Gaussian random variables.

Lemma 54 (Hoeffding's inequality for sub-Gaussian ranom variables) *Let $X_1, \ldots, X_n$ be n independent random variables such that X_i has mean μ_i and sub-Gaussian parameter σ_i. Let $X = \sum_{i=1}^n X_i$, then for any $t > 0$,*

$$\mathbb{P}\left(X - \mathbb{E}\left[X\right] \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right),$$

and

$$\mathbb{P}\left(X - \mathbb{E}\left[X\right] \leq -t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

Bounded Difference Condition While the previous inequalities showed concentration for the sum of independent random variables whose tail probability behavior is well-controlled, McDiarmid's inequality provides concentration results for general class of functions which depend on independent random variables, but in a limited way, satisfying the so-called "bounded difference" condition.

Lemma 55 (McDiarmid's inequality) *Let $X_1, \ldots, X_n$ be independent random variables such that for each $i \in [n]$, $X_i \in X$. Let $\xi : \prod_{i=1}^n X_i \to \mathbb{R}$ be a function of $(X_1, \ldots, X_n)$ that satisfies $\forall i, \forall x_1, \ldots, x_n, \forall x_i' \in X_i$,*

$$\left|\xi\left(x_1, \ldots, x_i, \ldots, x_n\right) - \xi\left(x_1, \ldots, x_i', \ldots, x_n\right)\right| \leq c_i. \tag{102}$$

Then for all $t > 0$,

$$\mathbb{P}\left(\xi - \mathbb{E}\left[\xi\right] \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right).$$

By considering the negation of the function $-\xi$ in lieu of ξ, one can obtain the same tail bound for the opposite direction.

Appendix L. Some Known Results from Deconvolution Literature

In this section, we introduce some known results for estimating the unknown density f_X of random variable X by deconvolution techniques. Suppose that $Z = X + N$ is a measurement of X with additive noise N and we have n i.i.d. observations $Z_1, \ldots, Z_n$. Fan (1991) reported that we can achieve an asymptotically consistent density estimate when the noise density is known and f_X satisfies certain smoothness conditions. Later, Delaigle et al. (2008) showed that consistent estimation is possible even when the noise distribution is unknown, with aid of repeated measurements.

Their estimators and proof techniques rely on the kernel smoothing method. Here we only present the abbreviated version of the concepts, the estimator, and the results to the minimum amount we need. We would refer interested readers to relevant references for more detail; for example, Carroll and Hall (1988); Fan (1991); Delaigle et al. (2008).

L.1. Deconvolution Kernel Density Estimator

We provide a summary for deconvolution kernel density estimator, which we already discussed in detail to provide intuition for our algorithm in Appendix D.1. For more detailed explanations to see how and why it works, please see that discussion.

Our goal is to recover distribution of random variable X, but we observe samples of $Z = X+N$ instead of X. We assume we know the distribution of N. Due to independence, we know that $\phi_Z(t) = \phi_X(t)\phi_N(t)$ for all $t \in \mathbb{R}$, where ϕ_Z, ϕ_X, ϕ_N denote the characteristic function of random variable Z, X and N respectively.

Let $\mathcal{F}$ denote Fourier transformation operator and $\mathcal{F}^{-1}$ denote the inverse Fourier transformation operator. By applying these operators, we obtain

$$\hat{f}_X(x) = \mathcal{F}^{-1}\left\{\frac{\mathcal{F}\{\hat{f}_Z(x)\}(t)}{\phi_N(t)}\right\} = \frac{1}{hn}\sum_{i=1}^{n} L\Big(\frac{x - Z_i}{h}\Big), \tag{103}$$

where

$$L \equiv \mathcal{F}^{-1}\left\{\frac{\phi_K(\,\cdot\,)}{\phi_N(\,\cdot\, h^{-1})}\right\}, \quad \text{i.e.,} \quad L(z) = \frac{1}{2\pi}\int \exp(-\mathbf{i}\, tz)\frac{\phi_K(t)}{\phi_N\left(\frac{t}{h}\right)}dt, \;\; z \in \mathbb{R}.$$

A more detailed description of the derivation can be found in Appendix D.1.

Indeed, this is known as deconvolution kernel density estimator in literature. We shall adopt prior results of Fan (1991) on its consistency to establish our results. We refer interested readers to Wand and Jones (1994) for more details and properties of kernel density estimation.

L.2. Consistency Results for Deconvolution

L.2.1. ASSUMPTIONS

Assumptions on the signal density For constants $m, B \geq 0$, and $\alpha \in [0, 1)$, Fan defined a class of densities as

$$\mathcal{C}_{m,\alpha,B} = \{f_X(x) : \left|f_X^{(m)}(x) - f_X^{(m)}(x+\delta)\right| \leq B\delta^{\alpha}\}. \tag{104}$$

Intuitively, that implies that f_X is slowly varying, i.e., the density is sufficiently "smooth' so that there is a hope to reconstruct it from a finite number of samples by interpolating the empirical density.

Assumptions on the noise Fan (1991) showed that the difficulty of deconvolution depends on the smoothness of the noise distribution and that of the density to be estimated. Here, the term 'smoothness' means the order of the characteristic function as $t \to \infty$. In short, the deconvolution becomes more difficult as it is corrupted by smoother additive noise. Following Fan (1991), we call the distribution of a random variable N smooth of order β if its characteristic function ϕ_N satisfies

$$B^{-1}\left(1 + |t|\right)^{-\beta} \leq |\phi_N(t)| \leq B\left(1 + |t|\right)^{-\beta}, \tag{105}$$

for some positive constants $\beta > 0$ and $B > 0$, and for all real t. This class of densities with polynomially decaying tails in the Fourier domain is called ordinary-smooth. Some examples of this ordinary-smooth error distributions include symmetric Gamma and double exponential distributions.

There is another interesting class of error distributions, whose tails decay much faster in the Fourier domain. We will call the distribution of a random variable N super-smooth of order β if its characteristic function ϕ_N satisfies

$$B^{-1}\exp\left(-\gamma|t|^{\beta}\right) \leq |\phi_N(t)| \leq B\exp\left(-\gamma|t|^{\beta}\right), \tag{106}$$

for some positive constants $\beta, \gamma > 0$ and $B > 1$, and for all real t. Normal, mixture normal, Cauchy distributions belong to the super-smooth class.

Assumptions on the Kernel We summarize some required properties of kernel used in the density estimator and the smoothness of noise before stating the results of Fan (1991).

(K1) $\phi_K(t)$ is a symmetric function, which has bounded integrable derivatives up to order $m+2$ on $\mathbb{R}$;

(K2) $\phi_K(t) = 1 + O\left(|t|^m\right)$ as $t \to 0$;

(K3) $\phi_K(t) = 0$, for $|t| \geq 1$.

(N1) $\phi_N(t)$ is supersmooth of order β; see Eq. (106)

Note that $\phi_N(t) \neq 0, \forall t$ is subsumed in (N1).

L.2.2. SOME DECONVOLUTION RESULTS

The following theorem provides the consistency and the convergence rate of the kernel density estimator with known noise density (Eq. (103)) when the error distribution is supersmooth. We use subscript n in $\hat{f}_n$ to emphasize that $\hat{f}_X$ is an estimator for f_X based on n samples.

Theorem 56 (Fan (1991), Theorem 1) *Let the kernel satisfies (K1), (K2), (K3), and the distribution of error satisfies (N1). With the choice of kernel bandwidth parameter* $h_n = (4\gamma)^{\frac{1}{\beta}} (\log n)^{-\frac{1}{\beta}}$, *we have*

$$\sup_{f \in \mathcal{C}_{m,\alpha,B}} \sup_{x \in \mathbb{R}} \mathbb{E}\left[\left(\hat{f}_n(x) - f(x)\right)^2\right] = O\left((\log n)^{-2(m+\alpha)/\beta}\right).$$

There is another result (which is actually a corollary of the above theorem) in the same paper, which serves better for our purpose. With $\hat{f}_n$, it is possible to define an estimator of the CDF, F, of the random variable X by integration:

$$\hat{F}_n(x) = \int_{-M_n}^{x} \hat{f}_n(z) dz. \tag{107}$$

M_n is a sequence of constants, which tends to $-\infty$ as $n \to \infty$. The following theorem provides a convergence rate, which is better than naïvely integrating that bound from Theorem 56.

Theorem 57 (Fan (1991), Theorem 3) *Let the same assumptions with Theorem 56 except for that* m *is replaced with* $m+1$ *in (K1) and (K2). Then by choosing the same bandwidth parameter* $h_n = (4\gamma)^{\frac{1}{\beta}} (\log n)^{-\frac{1}{\beta}}$ *and* $M_n = n^{\frac{1}{6}}$, *we have*

$$\sup_{f \in \mathcal{C}'_{m,\alpha,B}} \sup_{x \in \mathbb{R}} \mathbb{E}\left[\left(\tilde{F}_n(x) - F(x)\right)^2\right] = O\left((\log n)^{-2(m+\alpha+1)/\beta}\right).$$

where $\mathcal{C}'_{m,\alpha,B} = \left\{f \in \mathcal{C}_{m,\alpha,B} : F(-n) \leq D\left(\log n\right)^{-(m+2)/\beta}\right\}$.

In the original paper, $M_n = n^{\frac{1}{3}}$ is used. However, the theorem still remains valid with the modificatio to $M_n = n^{\frac{1}{6}}$ (see Fan (1991), the proof of Theorem 3).