Multi-Item Mechanisms without Item-Independence: Learnability via Robustness

JOHANNES BRUSTLE*, McGill University, Canada
YANG CAI†, Yale University
CONSTANTINOS DASKALAKIS‡, MIT

We study the sample complexity of learning revenue-optimal multi-item auctions. We obtain the first set of positive results that go beyond the standard but unrealistic setting of item-independence. In particular, we consider settings where bidders' valuations are drawn from correlated distributions that can be captured by Markov Random Fields or Bayesian Networks – two of the most prominent graphical models. We establish parametrized sample complexity bounds for learning an up-to-ε optimal mechanism in both models, which scale polynomially in the size of the model, i.e. the number of items and bidders, and only exponential in the natural complexity measure of the model, namely either the largest in-degree (for Bayesian Networks) or the size of the largest hyper-edge (for Markov Random Fields).

We obtain our learnability results through a novel and modular framework that involves first proving a robustness theorem. We show that, given only "approximate distributions" for bidder valuations, we can learn a mechanism whose revenue is nearly optimal simultaneously for all "true distributions" that are close to the ones we were given in Prokhorov distance. Thus, to learn a good mechanism, it suffices to learn approximate distributions. When item values are independent, learning in Prokhorov distance is immediate, hence our framework directly implies the main result of Gonczarowski and Weinberg [36]. When item values are sampled from more general graphical models, we combine our robustness theorem with novel sample complexity results for learning Markov Random Fields or Bayesian Networks in Prokhorov distance, which may be of independent interest. Finally, in the single-item case, our robustness result can be strengthened to hold under an even weaker distribution distance, the Lévy distance.

CCS Concepts: • **Theory of computation → Algorithmic mechanism design**; **Computational pricing and auctions**; • **Computing methodologies → Learning in probabilistic graphical models**.

Additional Key Words and Phrases: revenue maximization, multi-item auctions, beyond item-independence, robustness, sample complexity, Markov random fields, Bayesian networks

ACM Reference Format:
Johannes Brustle, Yang Cai, and Constantinos Daskalakis. 2020. Multi-Item Mechanisms without Item-Independence: Learnability via Robustness. In *Proceedings of the 21st ACM conference on Economics and Computation (EC '20), July 13–17, 2020, Virtual Event, Hungary.* ACM, New York, NY, USA, 47 pages. https://doi.org/10.1145/3391403.3399541

*Supported by the NSERC Discovery Award RGPIN-2015-06127 and FRQNT Award 2017-NC-198956.
†Supported by the NSF Award CCF-1942583 (CAREER) and a Sloan Foundation Research Fellowship. Part of Cai's work was done under the support of the NSERC Discovery Award RGPIN-2015-06127 and FRQNT Award 2017-NC-198956.
‡Supported by NSF Awards IIS-1741137, CCF-1617730 and CCF-1901292, by a Simons Investigator Award, by the DOE PhILMs project (No. DE-AC05-76RL01830), by the DARPA award HR00111990021, by a Google Faculty award, by the MIT Frank Quick Faculty Research and Innovation Fellowship, and by the MIT-IBM Watson AI Lab.

EC '20, July 13–17, 2020, Virtual Event, Hungary

ACM ISBN 978-1-4503-7975-5/20/07…$15.00
https://doi.org/10.1145/3391403.3399541

1 INTRODUCTION

A central problem in Economics and Computer Science is the design of revenue-optimal auctions. The problem involves a seller who wants to sell one or more items to one or more strategic bidders. As bidders' valuation functions are private, no meaningful revenue guarantee can be achieved without any information about these functions. To remove this impossibility, it is standard to make a *Bayesian assumption,* whereby a joint distribution from which bidders' valuations are drawn is assumed common knowledge, and the goal is to design an auction that maximizes expected revenue with respect to this distribution.

In the *single-item setting*, a celebrated result by Myerson characterizes the optimal auction when bidder values are independent [49]. The quest for optimal *multi-item auctions* has been quite more challenging. It has been recognized that revenue-optimal multi-item auctions can be really complex and may exhibit counter-intuitive properties [9, 22, 23, 39, 40]. As such, it is doubtful that there is a clean characterization similar to Myerson's for the optimal multi-item auction. On the other hand, there has been significant recent progress in efficient computation of revenue-optimal auctions [2–4, 8, 10, 12–14, 16, 18, 19, 24]. This progress has enabled the identification of *simple auctions* (mostly variants of sequential posted pricing mechanisms) that achieve constant factor approximations to the optimum revenue [6, 15, 17, 20, 56], under *item-independence* assumptions.[1]

Making Bayesian assumptions in the study of revenue-optimal auctions is both crucial and fruitful. However, to apply the theory to practice, we would need to know the underlying distributions. Where does such knowledge come from? A common answer is that we estimate the distributions through market research or observation of bidder behavior in previously run auctions. Unavoidably, errors will creep in to the estimation, and a priori it seems possible that the performance of our mechanisms may be fragile to such errors. This has motivated a quest for optimal or approximately optimal mechanisms under imperfect knowledge of the underlying distributions.

This problem has received lots of attention from Theory of Computation recently. The focus has been on whether optimal or approximately optimal mechanisms are learnable given sample access to the true distributions. In single-item settings, where Myerson's characterization result applies, it is possible to learn up-to-ε optimal auctions [21, 28, 32, 35, 41, 46, 48, 52].[2] A recent paper by Guo et al. [37] provides upper and lower bounds on the sample complexity, which are tight up to logarithmic factors, thereby rendering a nearly complete picture for the single-item case.

In multi-item settings, largely due to the lack of simple characterizations of optimal mechanisms, results have been sparser. Recent work [11, 34, 48, 55] has shown how to learn simple mechanisms which attain a constant factor of the optimum revenue using polynomially many samples in the number of bidders and items. Last year, a surprising result by Gonczarowski and Weinberg [36] shows that the sample complexity of learning an up-to-ε optimal mechanism is also polynomial.[3] However, all these results rely on the *item-independence* assumption mentioned earlier, which limits their applicability. A main goal of our work is the following:

Goal I: Push the boundary of learning (approximately) optimal multi-item auctions to the important setting of **item dependence**.

Unfortunately, it is impossible to learn approximately optimal auctions from polynomially many samples under general item dependence. Indeed, an exponential sample complexity lower bound

[1] Intuitively, item independence means that each bidder's value for each item is independently distributed, and this definition has been suitably generalized to set value functions such as submodular or subadditive functions [53].

[2] The term "up-to-ε optimal" introduced in [36] means an additive $\varepsilon \cdot H$ approximation for distributions supported on $[0, H]$. Under tail assumption on the distribution, it is also possible to obtain $(1 - \varepsilon)$-multiplicative approximations.

[3] In particular, they learn a mechanism that is $O(\varepsilon)$-truthful and has up-to-ε optimal revenue.

has been established by Dughmi et al. [31] for even a single unit-demand buyer. Arguably, however, in auction settings, as well as virtually any high-dimensional setting, the distributions that arise are not arbitrary. Arbitrary high-dimensional distributions cannot be represented efficiently, and are known to require exponentially many samples to learn or even perform the most basic statistical tests on them; see e.g. [25] for a discussion. Accordingly a large focus of Statistics and Machine Learning has been on identifying structural properties of high-dimensional distributions, which enable succinct representation, efficient learning, and efficient statistical inference. In line with this literature, we propose *learning multi-item auctions under the assumption that item values are jointly sampled from a high-dimensional distribution with structure.*

There are several widely-studied probabilistic frameworks which allow modeling structure in a high-dimensional distribution. In this work we consider two of the most prominent ones: *Markov Random Fields (MRFs)* and *Bayesian Networks, a.k.a. Bayesnets,* which are the two most common types of *graphical models*. Both MRFs and Bayesnets have been studied in Machine Learning and Statistics for decades. Both frameworks can be used to express arbitrary high-dimensional distributions. Their advantage, however, is that they are associated with natural complexity parameters which allow tuning the dependence structure in the distributions they model, from product measures all the way up to arbitrary distributions. In Figure 1, we show a very simple example illustrating how naturally these models express dependence structure in a distribution. The figure shows a Bayesnet, which samples the values of a buyer for four items. The structure of the Bayesnet implies (see Definition 11) that these values are sampled conditionally independently, conditioning on the value of the variable at the root of the Bayesnet which is the state of the buyer's residence. The node is shaded because we assume that the corresponding variable is not observable. The pertinent question is how we might exploit the structure of the distribution, as captured by the natural complexity parameter of an MRF or a Bayesnet, to efficiently learn a good mechanism. At a high level, there are two components to the problem of learning approximately optimal auctions. One is *inference from samples*, i.e. extracting information about the distribution using samples. The other is *mechanism design*, i.e. constructing a good mechanism using the information extracted. A main goal of our work is:

Goal II: Provide a modular approach for learning multi-item auctions which decouples the Inference and Mechanism Design components, so that one may leverage all techniques from Machine Learning and Statistics to tackle the first and, independently, leverage all techniques from Mechanism Design to address the second.

Unfortunately, the Statistical and Mechanism design components are complexly intertwined in prior work on learning multi-item auctions. Specifically, [11, 36, 47, 55] are PAC-learning approaches, which require a fine balance between (i) selecting a class of mechanisms that is rich enough to contain an approximately optimal one for a class of distributions; and (ii) having small enough statistical complexity so that the performance of all mechanisms in the class on a small sample is representative of their performance with respect to the whole distribution, and so that a small sample suffices to select a good mechanism in the class. See the related work section for a detailed discussion of these works and their natural limitations. Our goal in this work is to avoid a joint consideration of (i) and (ii). Rather we want to obtain a learning framework that separates Mechanism Design from Statistical Inference, based on the following:

(i)' find an algorithm $\mathcal{M}$, which given a distribution F in some family of distributions $\mathcal{F}$, computes an (approximately) optimal mechanism $\mathcal{M}(F)$ when bidders' valuations are drawn from F;

(ii)' find an algorithm $\mathcal{L}$, which given sample access to a distribution F from the family of distributions $\mathcal{F}$ learns a distribution $\mathcal{L}(F)$ that is close to F in some distribution distance d.

Achieving (i)' and (ii)' is of course not enough, unless we also guarantee the following:

(iii)' given an (approximately) optimal mechanism M for some F there is a way to transform M to some M' that is approximately optimal for *any distribution* F' that is close to F under distribution distance d.

Given (i)'–(iii)', the learnability of (approximately) optimal mechanisms for a family of distributions $\mathcal{F}$ can be established as follows: (a) Given sample access to some distribution $F \in \mathcal{F}$ we use $\mathcal{L}$ to learn some distribution F' that is close to F under d; (b) we then use $\mathcal{M}$ to compute an (approximately) optimal mechanism M' for F'; and (c) finally, we use (iii)' to argue that M' can be converted to a mechanism M that is (approximately) optimal for F because M is (approximately) optimal for any distribution that is close to F'.

Clearly, (iii)' is important for decoupling (i)'—i.e. computing (approximately) optimal mechanisms for a family of distributions $\mathcal{F}$, and (ii)'—i.e. learning distributions in $\mathcal{F}$. At the same time, it is important in its own right:

Goal III: Develop robust mechanism design tools, allowing to transform a mechanism M designed for some distribution F into a mechanism M_{robust} which attains similar performance simultaneously for all distributions that are close to F in some distribution distance of interest.

The reason Goal III is interesting in its own right is that oftentimes we actually have no sample access to the underlying distribution over valuations. It is possible that we estimate that distribution through market research or econometric analysis in related settings, so we only know some approximate distribution. In other settings, we may have sample access to the true distribution but there might be errors in measuring or recording those samples. In both cases, we would know some approximate distribution F that is close to the true distribution under some distribution distance, and we would want to use F to identify a good mechanism for the unknown distribution that is close to F. Clearly, outputting a mechanism M that attains good performance under F might be a terrible idea as this mechanism might very well overfit the details of F. So we need to "robustify" M. A similar goal was pursued in the work of Bergemann and Schlag [7], for single-item and single-bidder settings, and in the work of Cai and Daskalakis [11], for robustifying a specific class of mechanisms under item-independence. Our goal here is to provide a very general robustification result.

1.1 Our Results

We discuss our contributions in the setting of additive bidders, whose values for the items are not necessarily independent. Our results hold for quite more general valuations, including constrained additive and any family of Lipschitz valuations (Definition 2), but we do not discuss these here to avoid overloading our notation. We will denote by n the number of bidders, and by m the number of items. We will also assume that the bidders' values for the items lie in some bounded interval $[0, H]$.

Our Robustness Results (cf. Goal III above). The setting we consider is the following. We are given a collection of model distributions $\mathcal{D} = \{\mathcal{D}_i\}_{i\in[n]}$, one for each bidder $i = 1, \ldots, n$. We do not know the true distributions $\widehat{\mathcal{D}} = \{\widehat{\mathcal{D}}_i\}_i$ sampling the valuations of the bidders, and the only information we have about each $\widehat{\mathcal{D}}_i$ is that $d(\mathcal{D}_i, \widehat{\mathcal{D}}_i) < \varepsilon$, under some distribution distance $d(\cdot,\cdot)$—we will discuss distances shortly.

Our goal is to design a mechanism that performs well under any possible collection of true distributions $\{\widehat{\mathcal{D}}_i\}_i$ that are close to their corresponding distributions $\{\mathcal{D}_i\}_i$ under d. We show that

Setting	Distance d	Robustness	Continuity
Single Item	Kolmogrov	$\textsc{Rev}\left(\widehat{M}, \widehat{\mathcal{D}}\right) \geq \text{OPT}\left(\widehat{\mathcal{D}}\right) - O\left(nH\varepsilon\right)$ $\widehat{M}$ is IR and DSIC (Theorem 5)	$\left\lvert\text{OPT}\left(\widehat{\mathcal{D}}\right) - \text{OPT}\left(\mathcal{D}\right)\right\rvert \leq O(nH\varepsilon)$ (Theorem 5)
	Lévy	$\textsc{Rev}\left(\widehat{M}, \widehat{\mathcal{D}}\right) \geq \text{OPT}\left(\widehat{\mathcal{D}}\right) - O\left(nH\varepsilon\right)$ $\widehat{M}$ is IR and DSIC (Theorem 4)	$\left\lvert\text{OPT}\left(\widehat{\mathcal{D}}\right) - \text{OPT}\left(\mathcal{D}\right)\right\rvert \leq O(nH\varepsilon)$ (Corollary 1)
Multiple Items	TV	$\textsc{Rev}\left(\widehat{M}, \widehat{\mathcal{D}}\right) \geq \text{OPT}_\eta\left(\widehat{\mathcal{D}}\right) - O\left(n^2mH\varepsilon + nmH\sqrt{n\varepsilon}\right)$ $\widehat{M}$ is IR and η- BIC w.r.t. $\widehat{\mathcal{D}}$, where $\eta = O\left(n^2mH\varepsilon\right)$ (Theorem 7)	$\left\lvert\text{OPT}\left(\widehat{\mathcal{D}}\right) - \text{OPT}\left(\mathcal{D}\right)\right\rvert \leq O\left(n^2mH\varepsilon + nmH\sqrt{n\varepsilon}\right)$ (Theorem 6)
	Prokhorov	$\textsc{Rev}\left(\widehat{M}, \widehat{\mathcal{D}}\right) \geq \text{OPT}_\eta\left(\widehat{\mathcal{D}}\right) - O\left(n\eta + n\sqrt{mH\eta}\right)$ $\widehat{M}$ is IR and η- BIC w.r.t. $\widehat{\mathcal{D}}$, where $\eta = O\left(nmH\varepsilon + m\sqrt{nH\varepsilon}\right)$ (Theorem 7)	$\left\lvert\text{OPT}(\mathcal{D}) - \text{OPT}(\widehat{\mathcal{D}})\right\rvert \leq O\left(n\xi + n\sqrt{mH\xi}\right)$ where $\xi = O\left(nmH\varepsilon + m\sqrt{nH\varepsilon}\right)$ (Theorem 6)

Table 1. Summary of Our Robustness and Revenue Continuity Results. Recall that the true bidder distributions $\widehat{\mathcal{D}}$ are unknown, and that $\widehat{M}$ is the robustified mechanism returned by our algorithm given an optimal mechanism M for a collection of bidder distributions $\mathcal{D}$ that are ε-close to $\widehat{\mathcal{D}}$ under distribution distance d. $\textsc{Rev}(\widehat{M}, \widehat{\mathcal{D}})$ denotes the revenue of $\widehat{M}$ when the bidder distributions are $\widehat{\mathcal{D}}$. For a collection of bidder distributions $\mathcal{F}$, $\text{OPT}(\mathcal{F})$ is the optimal revenue attainable by any BIC and IR mechanism under distributions $\mathcal{F}$, and $\text{OPT}_\eta(\mathcal{F})$ denotes the optimum revenue attainable by any η-BIC and IR mechanism under $\mathcal{F}$. Not included in the table are approximation preserving robustification results under TV and Prokhorov closeness. We show that we can transform any c-approximation M w.r.t. $\mathcal{D}$ to a robust mechanism $\widehat{M}$, so that $\widehat{M}$ is almost a c-approximation w.r.t. $\widehat{\mathcal{D}}$. The results included in the table are corollaries of this more general result when $c = 1$. See our theorem statements for the complete details. Moreover, if there is only a single bidder, we can strengthen our robustness results in multi-item settings so that $\widehat{M}$ is IC instead of η-IC (see Theorem 8). Our continuity results hold for any $\mathcal{D}$ and $\widehat{\mathcal{D}}$ as long as $d(\mathcal{D}_i, \widehat{\mathcal{D}}_i) \leq \varepsilon$ for each bidder i.

there are robustification algorithms, which transform a mechanism M into a robust mechanism $\widehat{M}$ that attains similar revenue to that of M under $\mathcal{D}$, except that $\widehat{M}$'s revenue guarantee holds *simultaneously* for any collection $\widehat{\mathcal{D}}$ that is close to $\mathcal{D}$. Applying our robustification algorithm to the optimum mechanism for $\mathcal{D}$ allows us to obtain the results reported in the first three columns of Table 1. DSIC and BIC refer to the standard properties of Dominant Strategy and Bayesian Incentive Compatibility of mechanisms, IR refers to the standard notion of Individual Rationality, and η-BIC is the standard notion of approximate Bayesien Incentive Compatibility. For completeness these notions are reviewed in Appendix B.

Some remarks are in order. First, in multi-item settings, it is unavoidable that our robustified mechanism is only approximately BIC, as we do not know the true distributions. In single-item settings, the optimal mechanism is DSIC, and we can indeed robustify it into a mechanism $\widehat{M}$ that is DSIC. In the multi-item case, however, it is known that DSIC mechanisms sometimes can extract at most a constant fraction of the optimal revenue [57], so it is necessary to consider BIC mechanisms and the BIC property is fragile to errors in the distributions.

Second, we consider several natural distribution distances. In multi-item settings, we consider both the Prokhorov and the Total Variation distance. In single-item settings, we consider both the Lévy and the Kolmogorov distance. Please see Section 2 for formal definitions of these distances and a discussion of their relationships, and their relationship to other standard distribution distances. We note that the Lévy distance for single-dimensional distributions, and the Prokhorov distance for multi-dimensional distributions are quite permissive notions of distribution distance. This makes our robustness results for these distances stronger, automatically implying robustness results under several other common distribution distances.

Finally, en route to proving our robustness results, we show a result of independent interest, namely that *the optimal revenue is continuous with respect to the distribution distances that we consider.* Our continuity results are summarized in the last column of Table 1. Note that the continuity results are substantially easier to establish than the robustness results, please see Section 1.2 for details.

Learning Multi-Item Auctions Under Item Dependence (cf. Goal I above). In view of our robustness results, presented above, the challenge of learning near-optimal auctions given sample access to the bidders' valuation distributions, becomes a matter of estimating these distributions in the required distribution distance, depending on which robustification result we want to apply.

When the item values are independent, learning bidders' type distributions in our desired distribution distances is immediate. So we easily recover the guarantees of the main theorem of [36]. These guarantees are summarized in the second row of Table 2, and are expanded upon in Theorem 9.

But a main goal of our work (namely Goal I from earlier) is to push the learnability of auctions well beyond item-independence. As stated earlier, it is impossible to attain learnability from polynomially many samples for arbitrary joint distributions over item values so we consider the well-studied frameworks of MRFs and Bayesnets. These frameworks are flexible and can model *any* distribution, but they have a tunable complexity parameter whose value controls the dependence structure. This parameter is the maximum clique size of an MRF and maximum in-degree of a Bayesnet. We will denote this complexity parameter d in both cases. Recall that we also used $d(\cdot, \cdot)$ to denote distribution distances. To disambiguate, whenever we study MRFs or Bayesnets, we make sure to use $d(\cdot, \cdot)$, *with parentheses*, to denote distribution distances. Note that a small value of the complexity parameter d *does not mean* that the corresponding MRF or Bayesnet does not have correlations among every pair of item values. Many natural MRF structures, with $d = 2$, and Bayesnet structures, with $d = 1$, permit distributions where all the variables are correlated, and indeed any pair of variables remain correlated even after conditioning on the values of all the other variables. In Figure 1, we show a simple such example where the values of a bidder on four items are sampled from a Naive Bayes Model, which is a very simple type of Bayesnet with $d = 1$. While even small values of d allow all pairs of variables to be correlated even conditioning on everything else, the complexity parameter d forbids *arbitrary* dependence structures. Indeed, this is the reason why MRFs and Bayesnets are so prevalent. They allow rich dependent structures but not arbitrary ones, unless their complexity parameter d is tuned up to its maximum possible value, i.e. equal to the total number of variables, in which case they can express any dependence structure. In particular, a model of complexity d can express arbitrary dependence on subsets of d (for MRFs) or $d + 1$ (for Bayesnets) variables, and it allows some dependence structures on larger subsets of variables depending on the graphical structure of the model.

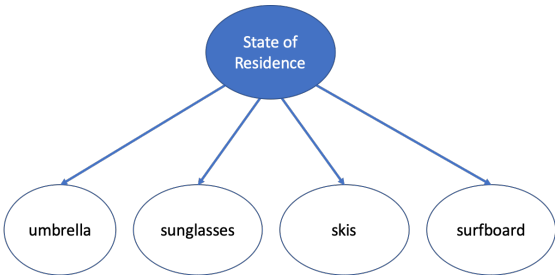

Fig. 1. The values of a buyer for an umbrella, a pair of sunglasses, a pair of skis, and a surfboard are sampled from a Naive Bayes model. These values are sampled conditionally independently conditioning on the value of the variable at the root of the network, which is the state of the buyer's residence. This variable is latent, i.e. non-observable, and this is why the corresponding node of the network is shaded blue. The distribution over $(v_{\text{umbrella}}, v_{\text{sunglasses}}, v_{\text{skis}}, v_{\text{surfboard}})$ has the property that any pair of values remain correlated even conditioning on all the other values, unless the conditional distributions in the Bayesnet have special structure.

Now, in order to learn near-optimal mechanisms when item values for each bidder are sampled from an MRF or a Bayesnet of certain complexity d, our robustness results reassure us that it

suffices to learn MRFs and Bayesnets under Total Variation or Prokhorov distance, depending on which multi-item robustenss theorem we seek to apply. So we need an upper bound on the sample complexity necessary to learn MRFs and Bayesnets. One of the contributions of our paper is to provide very general sample complexity bounds for learning these distributions, as summarized in Theorems 12 and 13 for MRFs and Bayesnets respectively. In both theorems, V is the set of variables, d is the complexity measure of the underlying distribution, and ε is the distance within which we are seeking to learn the distribution. Each theorem has a version when the variables take values in a finite alphabet Σ, and a version when the variables take values in some interval $\Sigma = [0, H]$. In the first case, we provide bounds for learning in the stronger notion of Total Variation distance. In the second case, since we are learning from finitely many samples, we need to settle for the weaker notion of Prokhorov distance. For the same reason, we need to make some Lipschitzness assumption on the density, so our sample bounds depend on the Lipschitzness C of the MRF's potential functions and the Bayesnet's conditional distributions.

The sample bounds we obtain for learning MRFs and Bayesnets are directly reflected in the sample bounds we obtain for learning multi-item auctions when the item-values are sampled from an MRF or a Bayesnet respectively, as summarized in the last two rows of Table 2. Indeed, the sample complexity for learning auctions is entirely due to the sample complexity needed to learn the underlying item-distribution. In all cases we consider, the complexity is polynomial in number of variables $n = |V|$ and only depends exponentially in d, the complexity of the distribution, and this is unavoidable. [4]

Setting	Revenue Guarantee and Sample Complexity	Prior Result	Technique
Item Independence	up-to-ε optimal, η-BIC (Theorem 9) poly $(n, m, H, 1/\varepsilon, 1/\eta, \log(1/\delta))$	recovers main result of [36]	Prokhorov Robustness + Learnability of Product Dist. (Folklore)
MRF	up-to-ε optimal, η-BIC (Theorem 10) poly $\left(n, m^d, \lvert\Sigma\rvert^d, H, 1/\eta, 1/\varepsilon, \log(1/\delta)\right)$ (Finite Σ) poly $\left(n, m^{d^2}, (\frac{CH}{\varepsilon})^d, 1/\eta, \log(1/\delta)\right)$ ($\Sigma = [0, H]$)	unknown	Prokhorov Robustness + Learnability of MRFs (Theorem 12)
Bayesnet	up-to-ε optimal, η-BIC (Theorem 11) poly $\left(n, d, m, \lvert\Sigma\rvert^{d+1}, H, 1/\eta, 1/\varepsilon, \log(1/\delta)\right)$ (Finite Σ) poly $\left(n, d^{d+1}, m^{d+1}, (\frac{CH}{\varepsilon})^{d+1}, 1/\eta, \log(1/\delta)\right)$ ($\Sigma = [0, H]$)	unknown	Prokhorov Robustness + Learnability of Bayesnets (Theorem 13)

Table 2. Summary of Our Sample-based Results. We denote by Σ the support of each item-marginal, taken to equal the interval $[0, H]$ in the continuous case, we use δ for the failure probability, and use d to denote the standard complexity measure of the graphical model used to model item dependence, namely the size of the maximum hyperedge in MRFs and the largest in-degree in Bayesnets. For both MRFs and Bayesnets we allow latent variables and we also do not need to know the underlying graphical structure. Moreover, for continuous distributions, our results require Lipschitzness of potential functions in MRFs and conditional distributions in Bayesnets, which we denote with C. Finally, if there is only a single bidder, the mechanism we learnt is strengthened to be IC instead of η-IC. See our theorem statements for our complete results.

Our sample bounds improve if the underlying graph of the MRF or Bayesnet are known and, importantly, without any essential modifications *our sample bounds hold even when there are latent, i.e. unobserved, variables in the distribution.* This makes both our auction and our distribution learning results much more richly applicable. As a simple example of the modeling power of latent variables, situations can be captured where an unobserved random variable determines the type of a bidder, and conditioning on this type the observable values of the bidder for different items are sampled.

[4]Note that the example by Dughmi et al. [31] can be captured by an MRF or Bayesnet with $d = O(m)$, and it is shown in [31] that the sample complexity for learning a mechanism that is a constant factor approximation to the optimal revenue in this example is at least $2^{\Omega(m)}$.

Finally, it is worth noting that our sample bounds for learning MRFs (i.e. Theorem 12) provide broad generalizations of the bounds for learning Ising models and Gaussian MRFs presented in recent work of Devroye et al [30]. Their bounds are obtained by bounding the VC complexity of the Yatracos class induced by the distributions of interest, while our bounds are obtained by constructing ε-nets of the distributions of interest, and running a tournament-style hypothesis selection algorithm [1, 26, 29] to select one distribution from the net. Since the distribution families we consider are non-parametric, our main technical contribution is to bound the size of an ε-net sufficient to cover the distributions of interest. Interestingly, we use properties of linear programs to argue through a sequence of transformations that the net size can be upper bounded in terms of the bit complexity of solutions to a linear program that we construct.

1.2 Roadmap and Technical Ideas

In this section, we provide a roadmap to the paper and survey some of our technical ideas.

Single-item Robustness (Appendix C). We consider first the setting where the model distribution $\mathcal{D}$ is ε-close to the true, but unknown distribution $\widehat{\mathcal{D}}$ in Kolmogorov distance. In this case, we argue directly that Myerson's optimal mechanism [49] for $\mathcal{D}$ is approximately optimal for any distribution that is in the ε-Kolmogorov-ball around $\mathcal{D}$, which includes $\widehat{\mathcal{D}}$ (Theorem 5). The idea is that the revenue of the optimal mechanism can be written as an integral over probabilities of events of the form: does v_i lie in a certain interval $[a, b]$? Since $\mathcal{D}$ and $\widehat{\mathcal{D}}$ are ε-close in Kolmogorov distance, the probabilities of all such events are within ε of each other, which implies that the revenues under $\mathcal{D}$ and $\widehat{\mathcal{D}}$ are also close. Finally, note that Myerson's optimal mechanism is DSIC and IR, so it is truthful and IR w.r.t. any distribution.

Unfortunately, the same idea fails for Lévy distance, as the difference in the probabilities of the event that a certain v_i lies in some interval $[a, b]$ under $\mathcal{D}$ and $\widehat{\mathcal{D}}$ can be as large as 1 even when $\mathcal{D}$ and $\widehat{\mathcal{D}}$ are ε-close in Lévy distance. (Indeed, consider two single point distributions: a point mass at A and a point mass at $A - \varepsilon$; their probabilities of falling in the interval $[A - \varepsilon/2, A + \varepsilon/2]$ are respectively 1 and 0.) We thus prove our robustness result for Lévy distance via a different route. Given any model distribution $\mathcal{D}$, we first construct the "worst" distribution $\underline{\mathcal{D}}$ and the "best" distribution $\overline{\mathcal{D}}$ in the ε-Lévy ball around $\mathcal{D}$: this means that, for any $\widehat{\mathcal{D}}$ that lies in the ε-Lévy ball around $\mathcal{D}$, $\widehat{\mathcal{D}}$ first-order stochastically dominates $\underline{\mathcal{D}}$ and is dominated by $\overline{\mathcal{D}}$ (see Definition 7). We choose our robust mechanism $\widehat{M}$ to be Myerson's optimal mechanism for $\underline{\mathcal{D}}$. It is not hard to argue that $\widehat{M}$'s revenue under $\widehat{\mathcal{D}}$ is at least $\mathrm{OPT}(\underline{\mathcal{D}})$, the optimal revenue under the "worst" distribution (Lemma 8), due to the revenue monotonicity lemma (Lemma 7) shown in [28]. The statement provides a lower bound of $\widehat{M}$'s revenue under the unknown true distribution $\widehat{\mathcal{D}}$. To complete the argument, we need to argue that $\mathrm{OPT}(\widehat{\mathcal{D}})$ cannot be too much larger than $\mathrm{OPT}(\underline{\mathcal{D}})$. Indeed, we relax $\mathrm{OPT}(\widehat{\mathcal{D}})$ to $\mathrm{OPT}(\overline{\mathcal{D}})$, and show that even the optimal revenue under the "best" distribution $\mathrm{OPT}(\overline{\mathcal{D}}) \approx \mathrm{OPT}(\underline{\mathcal{D}})$. To do so, we construct two auxiliary distributions P and Q, such that (i) $\mathrm{OPT}(P) \approx \mathrm{OPT}(Q)$; and (ii) P and $\underline{\mathcal{D}}$ are ε-close in *Kolmogorov* distance, and Q and $\overline{\mathcal{D}}$ are ε-close also in *Kolmogorov* distance. Our robustness theorem under Kolmogorov distance (Theorem 5) implies then that $\mathrm{OPT}(P) \approx \mathrm{OPT}(\underline{\mathcal{D}})$ and $\mathrm{OPT}(Q) \approx \mathrm{OPT}(\overline{\mathcal{D}})$. Hence, $\mathrm{OPT}(\overline{\mathcal{D}}) \approx \mathrm{OPT}(\underline{\mathcal{D}})$, which completes our proof.

Multi-item Robustness (Section 3). We first discuss our result for total variation distance. Unfortunately, our approach for Lévy distance—of simply choosing the optimal mechanism for the "worst," in the first-order stochastic dominance sense, distribution in the ε-TV-ball around $\mathcal{D}$ to be our robust mechanism—no longer applies. Indeed, it is known that the optimal revenue in multi-item auctions

may be non-monotone with respect to first-order stochastic dominance [40], i.e. a distribution may be stochastically dominated by another but result in higher revenue. However, if $\mathcal{D}$ and $\widehat{\mathcal{D}}$ are ε-close in total variation distance, this means that there is a coupling between $\mathcal{D}$ and $\widehat{\mathcal{D}}$ under which the valuation profiles are almost always sampled the same. If we take the optimal mechanism M for $\mathcal{D}$, and apply to bidders from $\widehat{\mathcal{D}}$, it will produce almost the same revenue under $\widehat{\mathcal{D}}$, and vice versa. Indeed, the only event under which M may generate different revenue under the two distributions is when the coupling samples different profiles, but this happens with small probability. Similarly, the BIC and IR properties of M under $\mathcal{D}$ become slightly approximate under $\widehat{\mathcal{D}}$. We claim that we can massage M, in a way *oblivious to* $\widehat{\mathcal{D}}$, to produce a $(\text{poly}(n, m, H) \cdot \varepsilon)$-truthful and *exactly* IR mechanism $\widehat{M}$ for $\widehat{\mathcal{D}}$, which achieves an up-to-$(\text{poly}(n, m, H) \cdot \varepsilon)$ revenue (Theorem 1).

The main challenge is when $\mathcal{D}$ and $\widehat{\mathcal{D}}$ are only ε-close in Prokhorov distance. Note that two distributions within Prokhorov distance ε may have total variation distance 1. Just imagine two point masses: one at A and another at $A - \varepsilon$. So Prokhorov robustness is not directly implied by TV robustness.

Why Standard Discretization Arguments are Insufficient? Unlike standard algorithmic problems, discretization is subtle in mechanism design. Due to the presence of incentives, a small change in the bidders' value distributions may change the distribution of outcomes of the mechanism dramatically. To perform discretization in mechanism design, a standard procedure goes as follows [5, 36, 45]: let $\widehat{\mathcal{D}}$ be the true distribution, and $\mathcal{D}$ be the distribution after discretization; design the optimal mechanism M for $\mathcal{D}$; to run M on a bid vector b from $\widehat{\mathcal{D}}$, discretize it to $\gamma(b) = (\gamma_1(b_1), \ldots, \gamma_n(b_n))$ and apply mechanism M on $\gamma(b)$. This procedure can be generalized to any pair of distributions $\mathcal{D}$ and $\widehat{\mathcal{D}}$ as long as, we are given a coupling $\gamma(\cdot)$ between $\mathcal{D}$ and $\widehat{\mathcal{D}}$ that maps any bid vector b in the support of distribution $\widehat{\mathcal{D}}$ to a bit vector $\gamma(b)$ in the support of $\mathcal{D}$. If for every bidder i, b_i and $\gamma_i(b_i)$ are close with all but small probability, we can apply similar arguments as in the total variation robustness result to massage the mechanism above to be nearly-truthful and exactly IR for $\widehat{\mathcal{D}}$, and argue it is approximately revenue optimal. Clearly, in the context of discretization, b_i and $\gamma_i(b_i)$ are guaranteed to be close if the discretization is sufficiently fine.

At first glance, this procedure may seem applicable to our problem. A characterization of Prokhorov distance due to Strassen (Theorem 2) shows that: two distributions P and Q are ε-close in Prokhorov distance if and only if there exists a (potentially randomized) coupling γ such that if random variable s is distributed according to P, then $\gamma(s)$ is distributed according to Q and $\Pr\left[\|s - \gamma(s)\|_1 > \varepsilon\right] \leq \varepsilon$. If M is the optimal mechanism for the model distribution $\mathcal{D}$, and $\widehat{\mathcal{D}}$ is the true distribution that is ε-close to $\mathcal{D}$, why can't we combine the procedure above with the coupling γ to establish our Prokhorov robustness result?

Unfortunately, this approach is insufficient due to the following two issues: (i) The procedure relies on knowing the coupling γ. As we do not even know $\widehat{\mathcal{D}}$, how can we know the coupling? (ii) Even if we can identify the coupling γ between $\mathcal{D}$ and a specific $\widehat{\mathcal{D}}$, the procedure above constructs a mechanism that depends on the coupling γ. However, γ may change for every different $\widehat{\mathcal{D}}$ in the ε-Prokhorov-ball around $\mathcal{D}$, so the procedure generates a different mechanism for every possible true distribution. [5]

To satisfy our requirement for a robust mechanism in Goal III, we need to construct a *single mechanism* that is nearly truthful, IR, and near-optimal simultaneously for every distribution in the ε-Prokhorov-ball around $\mathcal{D}$. Our proof relies on a novel way to *"simultaneously couple"*

[5] It is worth noting that the procedure can indeed be employed to prove the Prokhorov continuity, as the the pure existence of a good coupling γ between $\mathcal{D}$ and $\widehat{\mathcal{D}}$ suffices.

$\mathcal{D}$ with every distribution $\widehat{\mathcal{D}}$ in the ε-Prokhorov-ball around $\mathcal{D}$. If we round both $\mathcal{D}$ and any $\widehat{\mathcal{D}}$ to a random grid G with width $\sqrt{\varepsilon}$, we can argue that the *expected total variation distance* (over the randomness of the grid) between the two rounded distributions $\mathcal{D}_G$ and $\widehat{\mathcal{D}}_G$ is $O(\sqrt{\varepsilon})$ (Lemma 2). Now consider the following mechanism: choose a random grid G, round the bids to the random grid, apply the optimal mechanism M_G that is designed for $\mathcal{D}_G$. Our robustness result under the total variation distance implies that for every realization of the random grid G, M_G is $O\left(\text{poly}(n,m,H)\cdot\left\|\mathcal{D}_G-\widehat{\mathcal{D}}_G\right\|_{TV}\right)$-truthful and up-to-$O\left(\text{poly}(n,m,H)\cdot\left\|\mathcal{D}_G-\widehat{\mathcal{D}}_G\right\|_{TV}\right)$ revenue optimal for any $\widehat{\mathcal{D}}_G$. Since the expected value (over the randomness of the grid) of $\left\|\mathcal{D}_G-\widehat{\mathcal{D}}_G\right\|_{TV}$ is $O(\sqrt{\varepsilon})$ for any $\widehat{\mathcal{D}}$ in the ε-Prokhorov-ball of $\mathcal{D}$, our randomized mechanism is simultaneously $O\left(\text{poly}(n,m,H)\cdot\sqrt{\varepsilon}\right)$-truthful and up-to-$O\left(\text{poly}(n,m,H)\cdot\sqrt{\varepsilon}\right)$ revenue optimal for all distributions in the ε-Prokhorov-ball around $\mathcal{D}$.[6]

Sample Complexity Results. In Section F, we apply our robustness theorem to obtain sample bounds for learning multi-item auctions under the item-independence assumption (Theorem 9). Our result provides an alternative proof of the main result of [36]. In Section H, we combine our robustness theorem with our sample bounds for learning Markov Random Fields and Bayesian Networks discussed earlier to derive new polynomial sample complexity results for learning multi-item auctions when the distributions have structured correlation over the items. Theorem 10 summarizes our results when item values are generated by an MRF, and Theorem 11 our results when item values are generated by a Bayesenet.

2 PRELIMINARIES

We first define a series statistical distances that we will use in the paper and discuss their relationships.

Definition 1 (Statistical Distance). *Let P and Q be two probability measures. We use $\|P-Q\|_{TV}$, $\|P-Q\|_K$, and $\|P-Q\|_L$ to denote the **total variational distance**, the **Kolmogorov distance**, and the **Lévy distance** between P and Q, respectively. See Appendix B for more details. **Prokhorov Distance** is a generalization of the Lévy Distance to high dimensional distributions. Let (U,d) be a metric space and $\mathcal{B}$ be a σ-algebra on U. For $A\in\mathcal{B}$, let $A^\varepsilon=\{x:\exists y\in A\ \ s.t\ \ d(x,y)<\varepsilon\}$. Then two measures P and Q on $\mathcal{B}$ have Prokhorov distance*

$$\inf\{\varepsilon>0: P(A)\le Q(A^\varepsilon)+\varepsilon,\ Q(A)\le P(A^\varepsilon)+\varepsilon\ \forall A\in\mathcal{B}\}$$

*We consider distributions supported on $\mathbb{R}^k$ for some $k\in\mathbb{N}$, so U will be the k-dimensional Euclidean Space, and we choose d **to be the ℓ_1-distance**. We denote the Prokhorov distance between distributions $\mathcal{F},\widehat{\mathcal{F}}$ by $\left\|\mathcal{F}-\widehat{\mathcal{F}}\right\|_P$.*

Relationships between the Statistical Distances. Among the four metrics, the Lévy distance and the Kolmogorov distance are only defined for single dimensional distributions, while the Prokhorov distance and the total variation distance are defined for general distributions. In the single dimensional case, the Lévy distance is a very liberal metric. In particular, for any two single dimensional distributions P and Q,

$$\|P-Q\|_L\le\|P-Q\|_K\le\|P-Q\|_{TV}.$$

Note that a robustness result for a more liberal metric is more general. For example, the robustness result for single-item auctions under the Lévy metric implies the robustness under the total variation

[6] Since we round the bids to a random grid, we will also need to accommodate the rounding error. Please see Theorem 3 for details.

and Kolmogorov metric, because the ε-ball in Lévy distance contains the ε-ball in total variation and Kolmogorov distance. An astute reader may wonder whether one can find a more liberal metric in the single dimensional case. Interestingly, for the most common metrics studied probability theory, including the Wasserstein distance, the Hellinger distance, and the relative entropy, the Lévy distance is the most liberal up to a polynomial factor. That is, if the Lévy distance is ε, the distance under any of these metrics is at least poly(ε). Indeed, the polynomial is simply the identity function or the quadratic function ε^2 in most cases. Please see the survey by Gibbs and Su [33] and the references therein for more details.

The Prokhorov distance, also known as Lévy-Prokhorov Distance, is the generalization of the Lévy distance to multi-dimensional distributions. It is also the standard metric in robust statistical decision theory, see Huber [42] and Hampel et al. [38]. The Prokhorov distance is almost as liberal as the Lévy distance. [7] First, for any two distributions P and Q,

$$\|P - Q\|_P \leq \|P - Q\|_{TV}.$$

Second, if we consider other well studied metrics such as the Wasserstein distance, the Hellinger distance, and the relative entropy, the Prokhorov distance is again the most liberal up to a polynomial factor.

Multi-item Auctions. We focus on revenue maximization in the combinatorial auction with n **bidders** and m **heterogenous items**. We use X to denote the set of possible allocations, and each bidder $i \in [n]$ has a valuation function/type $v_i(\cdot) : X \mapsto \mathbb{R}_{\geq 0}$. In this paper, we assume the function $v_i(\cdot)$ is parametrized by $(v_{i,1}, \ldots, v_{i,m})$, where $v_{i,j}$ is bidder i's value for item j. We assume that *bidder's types are distributed independently*. Throughout this paper, we assume all bidders types lie in $[0, H]^m$. We adopt the valuation model in Gonczarowski and Weinberg [36] and consider valuations that satisfy the following Lipschitz property.

Definition 2 (Lipschitz Valuations). *There exists an absolute constant $\mathcal{L}$ such that if type $\mathbf{v_i} = (v_{i,1}, \ldots, v_{i,m})$ and $\mathbf{v'_i} = (v'_{i,1}, \ldots, v'_{i,m})$ are within ℓ_1 distance ε, then for the corresponding valuations $v_i(\cdot)$ and $v'_i(\cdot)$, $|v_i(x) - v'_i(x)| \leq \mathcal{L} \cdot \varepsilon$ for all $x \in X$.*

This for example includes common settings such as additive and unit demand with Lipschitz constant $\mathcal{L} = 1$. More generally, $\mathcal{L} = 1$ holds for constrained additive valuations [8] and even in some settings with complementarities. Please see [36] for further discussion.

A mechanism M consists of an allocation rule $x(\cdot)$ and a payment rule $p(\cdot)$. For any input bids $b = (b_1, \ldots, b_n)$, the allocation rule outputs a distribution over allocations $x(b) \in \Delta(X)$ and payments $p(b) = (p_1(b), \ldots, p_n(b))$. If bidder i's type is v_i, her utility under input b is $u_i(v_i, M(b)) = \mathbb{E}[v_i(x(b)) - p_i(b)]$.

Truthfulness and Revenue: We use the standard notion ε-BIC and IR (see Appendix B for details). If M is a ε-BIC mechanism w.r.t. some distribution $\mathcal{D}$, we use $\textsc{Rev}_T(M, \mathcal{D})$ to denote the revenue of mechanism M under distribution $\mathcal{D}$ assuming bidders are bidding truthfully. Clearly, $\textsc{Rev}_T(M, \mathcal{D}) = \textsc{Rev}(M, \mathcal{D})$ when M is BIC w.r.t. $\mathcal{D}$. We denote the optimal revenue achievable by any ε-BIC (or BIC) mechanism by $\text{OPT}_\varepsilon(\mathcal{D})$ (or $\text{OPT}(\mathcal{D})$). Although it is conceivable that permitting mechanisms to be ε-BIC allows for much greater expected revenue than if they were restricted to be BIC, past results show that this is not the case.

[7] Note that for single dimensional distributions, the Prokhorov distance is not equivalent to Lévy distance. In particular, $\|P - Q\|_L \leq \|P - Q\|_P$ for any single dimensional distributions P and Q.

[8] $v_i(\cdot)$ is constrained additive if $v_i(X) = \max_{R \subseteq S, R \in \mathcal{I}} \sum_{j \in R} v_{i,j}$, for some downward closed set system $\mathcal{I} \subseteq 2^{[m]}$ and $S = \{j : x_{i,j} = 1\}$.

Lemma 1. *[27, 53] In any n-bidder m-item auction, let $\mathcal{D}$ be any joint distribution over arbitrary $\mathcal{L}$-Lipschitz valuations, where the valuations of different bidders are independent. The maximum revenue attainable by any IR and ε-BIC auction for a given product distribution is at most $2n\sqrt{m\mathcal{L}H\varepsilon}$ greater than the maximum revenue attainable by any IR and BIC auction for that distribution.*

Notations: We allow the bidders to submit a special type $\perp$, which represents not participating the auction. If anyone submits $\perp$, the mechanism terminates immediately, and does not allocate any item to any bidder or charge any bidder. A bidder's utility for submitting type $\perp$ is 0. We will sometimes refer to $\perp$ as the **IR type**.Throughout the paper, we use $\widehat{\mathcal{D}} = \bigtimes_{i=1}^{n} \widehat{\mathcal{D}}_i$ to denote the true type distributions of the bidders. We use $\mathcal{D} = \bigtimes_{i=1}^{n} \mathcal{D}_i$ to denote the model type distributions or our learned type distributions from samples. We use $\mathcal{D}_i \mid \bigtimes_{j=1}^{m}[w_{ij}, w_{ij} + \delta)$ to denote the distribution induced by $\mathcal{D}_i$ conditioned on being in the m-dimensional cube $\bigtimes_{j=1}^{m}[w_{ij}, w_{ij} + \delta)$, and $\mathbf{supp}(\mathcal{F})$ to denote the support of distribution $\mathcal{F}$.

3 ROBUSTNESS FOR MULTI-ITEM AUCTIONS

In this section, we prove our robustness results under the total variation distance and the Prokhorov distance in multi-item settings. As discussed in Section 1.2, the proof strategy for single-item auctions fails miserably in multi-item settings due to the lack of structure of the optimal mechanism. In particular, one of the crucial tools we relied on in single-item settings, the revenue monotonicity, no longer holds in multi-item settings [40]. Nevertheless, we still manage to provide robustness guarantees in multi-item auctions. The plan is to first prove the robustness result under the total variation distance in Section 3.1, then we show show to relate the Prokhorov distance with the total variation distance using randomized rounding in Section 3.2, and reduce the robustness under the Prokhorov distance to the robustness under the total variation distance in Section 3.3.

3.1 TV-Robustness for Multi-item Auctions

Theorem 1 (TV-Robustness for Multi-item Auctions). *Given any distribution $\mathcal{D} = \bigtimes_{i=1}^{n} \mathcal{D}_i$, where each $\mathcal{D}_i$ is a distribution supported on $[0, H]^m$, and a η-BIC and IR mechanism M w.r.t. $\mathcal{D}$, we can construct a mechanism $\widehat{M}$ such that for any distribution $\widehat{\mathcal{D}} = \bigtimes_{i=1}^{n} \widehat{\mathcal{D}}_i \in [0, H]^{nm}$, if we let $\varepsilon_i = \left\|\widehat{\mathcal{D}}_i - \mathcal{D}_i\right\|_{TV}$ for all $i \in [n]$ and $\rho = \sum_{i\in[n]} \varepsilon_i$, then $\widehat{M}$ is $2m\mathcal{L}H\rho + \eta$-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR. Moreover, $\textsc{Rev}_T(\widehat{M}, \widehat{\mathcal{D}}) \geq \textsc{Rev}_T(M, \mathcal{D}) - nm\mathcal{L}H\rho$. Note that our construction of $\widehat{M}$ only depends on $\mathcal{D}$ and does not require any knowledge of $\widehat{\mathcal{D}}$.*

We briefly describe the ideas behind the proof. If $\widehat{\mathcal{D}}$ and $\mathcal{D}$ share the same support, it is not hard to see that M is already $(2m\mathcal{L}H\rho + \eta)$-BIC w.r.t. $\widehat{\mathcal{D}}$. The reason is that for any bidder i and any type v_i, her expected utility under any report can change by at most $m\mathcal{L}H\rho$ when the other bidders' bids are drawn from $\widehat{\mathcal{D}}_{-i}$ rather than $\mathcal{D}_{-i}$, as $\left\|\widehat{\mathcal{D}}_j - \mathcal{D}_j\right\|_{TV} = \varepsilon_j$ for all $j \in [n]$. The bulk of the proof is dedicated to the case, where $\widehat{\mathcal{D}}$ and $\mathcal{D}$ have different supports. We construct mechanism $\widehat{M}$, which first takes each bidder i's report and maps it to the "best" possible report from $\text{supp}(\mathcal{D}_i)$, then runs essentially M on the transformed reports. We show that $\widehat{M}$ is $2m\mathcal{L}H\rho + \eta$-BIC w.r.t. $\widehat{\mathcal{D}}$ and generates at most $nm\mathcal{L}H\rho$ less revenue. The proof of Theorem 1 is postponed to Appendix E.1.

3.2 Connecting the Prokhorov Distance with the Total Variation Distance

In this section, we provide a randomized rounding scheme that relates the Prokhorov distance to the total variation distance. We first state a characterization of the Prokhorov distance due to Strassen [54] that is useful for our analysis.

Theorem 2 (Characterization of the Prokhorov Metric [54]). *Let $\mathcal{F}$ and $\widehat{\mathcal{F}}$ be two distributions supported on $\mathbb{R}^k$. $\left\|\mathcal{F}-\widehat{\mathcal{F}}\right\|_P \le \varepsilon$ if and only if there exists a coupling γ of $\mathcal{F}$ and $\widehat{\mathcal{F}}$, such that $\Pr_{(x,y)\sim\gamma}[d(x,y)>\varepsilon] \le \varepsilon$, where $d(\cdot,\cdot)$ is the ℓ_1 distance.*

Theorem 2 states that $\mathcal{F}$ and $\widehat{\mathcal{F}}$ are within Prokhorov distance ε of each other if and only if there exists a coupling between the two distributions such that the two random variables are within ε of each other with probability at least $1-\varepsilon$. Next, we show that if $\mathcal{F}$ and $\widehat{\mathcal{F}}$ are close to each other in Prokhorov distance, then one can use a randomized rounding scheme to round both $\mathcal{F}$ and $\widehat{\mathcal{F}}$ to discrete distributions so that the two rounded distributions are close in total variation distance with high probability.

First, let us fix some notations.

Definition 3 (Rounded Distribution). *Let $\mathcal{F}$ be a distribution supported on $\mathbb{R}^k_{\ge 0}$. For any $\delta>0$ and $\ell\in[0,\delta]^k$, we define function $r^{(\ell,\delta)}:\mathbb{R}^k_{\ge0}\mapsto\mathbb{R}^k$ as follows: $r_i^{(\ell,\delta)}(x)=\max\left\{\left\lfloor\frac{x_i-\ell_i}{\delta}\right\rfloor\cdot\delta+\ell_i,0\right\}$ for all $i\in[k]$. Let X be a random variable sampled from distribution $\mathcal{F}$. We define $\lfloor\mathcal{F}\rfloor_{\ell,\delta}$ as the distribution for the random variable $r^{(\ell,\delta)}(X)$, and we call $\lfloor\mathcal{F}\rfloor_{\ell,\delta}$ as the rounded distribution of $\mathcal{F}$.*

Lemma 2. *Let $\mathcal{F}$ and $\widehat{\mathcal{F}}$ be two distributions supported on $\mathbb{R}^k$, and $\left\|\mathcal{F}-\widehat{\mathcal{F}}\right\|_P\le\varepsilon$. For any $\delta>0$, sample ℓ from the uniform distribution over $[0,\delta]^k$, $\mathbb{E}_{\ell\sim U[0,\delta]^k}\left[\left\|\lfloor\mathcal{F}\rfloor_{\ell,\delta}-\left\lfloor\widehat{\mathcal{F}}\right\rfloor_{\ell,\delta}\right\|_{TV}\right]\le\left(1+\frac{1}{\delta}\right)\varepsilon$.*

We only sketch the idea and postpone the formal proof to Appendix E.2. Let x be a random variable sampled from $\mathcal{F}$ and y be a random variable sampled from $\widehat{\mathcal{F}}$. Since $\mathcal{F}$ and $\widehat{\mathcal{F}}$ are close in Prokhorov distance, we can couple x and y according to Theorem 2 such that they are within ε of each other with probability at least $1-\varepsilon$. The rounding scheme chooses a random origin ℓ from $[0,\delta]^k$ and rounds $\mathcal{F}$ and $\widehat{\mathcal{F}}$ to the corresponding random grid with width δ. More specifically, we round $\mathcal{F}$ and $\widehat{\mathcal{F}}$ to $\lfloor\mathcal{F}\rfloor_{\ell,\delta}$ and $\left\lfloor\widehat{\mathcal{F}}\right\rfloor_{\ell,\delta}$ respectively. For simplicity, consider $\delta=\Theta(\sqrt{\varepsilon})$. The key observation is that when x and y are within ℓ_1-distance ε of each other, they lie in the same grid with probability at least $1-O(\sqrt{\varepsilon})$ over the randomness of ℓ. If x and y are in the same grid, they will be rounded to the same point. In other words, the coupling between x and y induces a coupling between $\lfloor\mathcal{F}\rfloor_{\ell,\delta}$ and $\left\lfloor\widehat{\mathcal{F}}\right\rfloor_{\ell,\delta}$ such that, in expectation over the choice of ℓ, the event that the corresponding two rounded random variables have different values happens with probability at most $\varepsilon+(1-\varepsilon)\cdot O(\sqrt{\varepsilon})=O(\sqrt{\varepsilon})$. By the definition of total variation distance, this implies that the expected total variation distance between $\lfloor\mathcal{F}\rfloor_{\ell,\delta}$ and $\left\lfloor\widehat{\mathcal{F}}\right\rfloor_{\ell,\delta}$ is also at most $O(\sqrt{\varepsilon})$. A similar argument applies to other choices of δ.

3.3 Prokhorov-Robustness for Multi-item Auctions

In this section, we show that even in multi-item settings, if every bidder's approximate type distribution $\mathcal{D}_i$ is within Prokhorov distance ε of her true type distribution $\widehat{\mathcal{D}}_i$, given any BIC and IR mechanism M for $\mathcal{D}=\bigtimes_{i=1}^n\mathcal{D}_i$, we can construct a mechanism $\widehat{M}$ that is $O(\mathrm{poly}(n,m,\mathcal{L},H,\varepsilon))$-BIC w.r.t. $\widehat{\mathcal{D}}=\bigtimes_{i=1}^n\widehat{\mathcal{D}}_i$, IR, and its revenue under truthful bidding $\textsc{Rev}_T(\widehat{M},\widehat{\mathcal{D}})$ is at most $O(\mathrm{poly}(n,m,\mathcal{L},H,\varepsilon))$ worse than $\textsc{Rev}(M,\mathcal{D})$.

Theorem 3. *Suppose we are given $\mathcal{D}=\bigtimes_{i=1}^n\mathcal{D}_i$, where $\mathcal{D}_i$ is an m-dimensional distribution for each $i\in[n]$, and a BIC and IR mechanism M w.r.t. $\mathcal{D}$. Suppose $\widehat{\mathcal{D}}=\bigtimes_{i=1}^n\widehat{\mathcal{D}}_i$ is the true but unknown*

type distribution such that $\left\|\mathcal{D}_i - \widehat{\mathcal{D}}_i\right\|_P \le \varepsilon$ *for all* $i \in [n]$. *We can construct a randomized mechanism* $\widehat{M}$, *oblivious to the true distribution* $\widehat{\mathcal{D}}$, *such that for any* $\widehat{\mathcal{D}}$ *the followings hold:*

(1) $\widehat{M}$ *is* κ*-BIC w.r.t.* $\widehat{\mathcal{D}}$ *and IR, where* $\kappa = O\left(nm\mathcal{L}H\varepsilon + m\mathcal{L}\sqrt{nH\varepsilon}\right)$;

(2) the expected revenue of $\widehat{M}$ *under truthful bidding is* $\textsc{Rev}_T\left(\widehat{M}, \widehat{\mathcal{D}}\right) \ge \textsc{Rev}(M, \mathcal{D}) - O(n\kappa)$.

We postpone the formal proof of Theorem 3 to Appendix E.3. We provide a complete sketch here. Our construction consist of the following five steps.

- **Step (1):** After receiving the bid profile, first sample ℓ from $U[0,\delta]^m$. For every realization of ℓ, we construct a mechanism $\widehat{M}^{(\ell)}$ and execute $\widehat{M}^{(\ell)}$ on the reported bids. In the next several steps, we show how to construct $\widehat{M}^{(\ell)}$ via two intermediate mechanisms $M_1^{(\ell)}$ and $M_2^{(\ell)}$ for every realization of ℓ based on M. Since ℓ is a random variable, $\widehat{M}$ is a randomized mechanism.
- **Step (2):** Round $\mathcal{D}_i$ to $\lfloor \mathcal{D}_i \rfloor_{\ell,\delta}$ for every bidder i. We construct mechanism $M_1^{(\ell)}$ based on M and show that $M_1^{(\ell)}$ is $O(m\mathcal{L}\delta)$-BIC w.r.t. $\bigtimes_{i=1}^n \lfloor \mathcal{D}_i \rfloor_{\ell,\delta}$ and IR. Moreover,
$$\textsc{Rev}_T\left(M_1^{(\ell)}, \bigtimes_{i=1}^n \lfloor \mathcal{D}_i \rfloor_{\ell,\delta}\right) \ge \textsc{Rev}(M, \mathcal{D}) - O(nm\mathcal{L}\delta).$$
Here is the idea behind the construction: for any bidder i and type w_i drawn from $\lfloor \mathcal{D}_i \rfloor_{\ell,\delta}$, we resample a type from $\mathcal{D}_i \mid \bigtimes_{j=1}^m [w_{ij}, w_{ij}+\delta)$, which is the distribution induced by $\mathcal{D}_i$ conditioned on being in the cube $\bigtimes_{j=1}^m [w_{ij}, w_{ij}+\delta)$. We use the allocation rule of M and a slightly modified payment rule on the resampled type profile. This guarantees that the new mechanism is $O(m\mathcal{L}\delta)$-BIC w.r.t. $\bigtimes_{i=1}^n \lfloor \mathcal{D}_i \rfloor_{\ell,\delta}$ and IR. The formal statement and analysis are shown in Lemma 3.
- **Step (3):** We use $\varepsilon_i^{(\ell)}$ to denote $\left\|\lfloor \mathcal{D}_i \rfloor_{\ell,\delta} - \left\lfloor \widehat{\mathcal{D}}_i \right\rfloor_{\ell,\delta}\right\|_{TV}$ for our sample ℓ and every $i \in [n]$, and $\rho^{(\ell)}$ to denote $\sum_{i\in[n]} \varepsilon_i^{(\ell)}$. We transform $M_1^{(\ell)}$ into a new mechanism $M_2^{(\ell)}$ using Theorem 1. In particular, $M_2^{(\ell)}$ is $O\left(m\mathcal{L}\delta + m\mathcal{L}H\cdot\rho^{(\ell)}\right)$-BIC w.r.t. $\bigtimes_{i=1}^n \left\lfloor \widehat{\mathcal{D}}_i \right\rfloor_{\ell,\delta}$ and IR. Importantly, the construction of $M_2^{(\ell)}$ is oblivious to $\bigtimes_{i=1}^n \left\lfloor \widehat{\mathcal{D}}_i \right\rfloor_{\ell,\delta}$ and $\left\{\varepsilon_i^{(\ell)}\right\}_{i\in[n]}$. Moreover, $\textsc{Rev}_T\left(M_2^{(\ell)}, \bigtimes_{i=1}^n \left\lfloor \widehat{\mathcal{D}}_i \right\rfloor_{\ell,\delta}\right) \ge \textsc{Rev}_T\left(M_1^{(\ell)}, \bigtimes_{i=1}^n \lfloor \mathcal{D}_i \rfloor_{\ell,\delta}\right) - O\left(nm\mathcal{L}H\cdot\rho^{(\ell)}\right)$.
- **Step (4):** We convert $M_2^{(\ell)}$ to $\widehat{M}^{(\ell)}$ so that it is $O\left(m\mathcal{L}\delta + m\mathcal{L}H\cdot\rho^{(\ell)}\right)$-BIC w.r.t. $\widehat{\mathcal{D}}$, IR and
$$\textsc{Rev}_T(\widehat{M}^{(\ell)}, \widehat{\mathcal{D}}) \ge \textsc{Rev}_T\left(M_2^{(\ell)}, \bigtimes_{i=1}^n \left\lfloor \widehat{\mathcal{D}}_i \right\rfloor_{\ell,\delta}\right) - nm\mathcal{L}\delta.$$
Here is the idea behind the construction of $\widehat{M}^{(\ell)}$: for every bidder i and her type w_i drawn from $\widehat{\mathcal{D}}_i$, round it to $r_i^{(\ell,\delta)}(w_i)$ (see Definition 3). We use the allocation rule of $M_2^{(\ell)}$ and a slightly modified payment rule on the rounded type profile. This guarantees that the new mechanism is $O\left(m\mathcal{L}\delta + m\mathcal{L}H\cdot\rho^{(\ell)}\right)$-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR. Note that our construction only requires knowledge of $M_2^{(\ell)}$, ℓ, and δ, and is completely oblivious to $\widehat{\mathcal{D}}$ and $\bigtimes_{i=1}^n \left\lfloor \widehat{\mathcal{D}}_i \right\rfloor_{\ell,\delta}$. The formal statement and analysis are shown in Lemma 4.
- **Step (5):** Since for every realization of ℓ, $\widehat{M}^{(\ell)}$ is $O\left(m\mathcal{L}\delta + m\mathcal{L}H\cdot\rho^{(\ell)}\right)$-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR, $\widehat{M}$ must be $O\left(m\mathcal{L}\delta + m\mathcal{L}H\cdot\mathbb{E}_{\ell\sim U[0,\delta]^m}\left[\rho^{(\ell)}\right]\right)$-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR. According to Lemma 2,

$\mathbb{E}_{\ell\sim U[0,\delta]^m}\left[\rho^{(\ell)}\right] = \sum_{i\in[n]}\mathbb{E}_{\ell\sim U[0,\delta]^m}\left[\varepsilon_i^{(\ell)}\right] = n\cdot\left(1+\frac{1}{\delta}\right)\varepsilon$. Therefore, $\widehat{M}$ is $O\left(m\mathcal{L}\delta + nm\mathcal{L}H\left(1+\frac{1}{\delta}\right)\epsilon\right)$-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR. Moreover, $\textsc{Rev}_T\left(\widehat{M},\widehat{\mathcal{D}}\right) \geq \textsc{Rev}(M,\mathcal{D}) - O\left(nm\mathcal{L}\delta + n^2m\mathcal{L}H\left(1+\frac{1}{\delta}\right)\varepsilon\right)$.

Lemma 3. *Given any $\delta > 0$, $\ell \in [0,\delta]^m$, and a BIC and IR mechanism M w.r.t. $\mathcal{D}$, we can construct a $\xi_1 = O(m\mathcal{L}\delta)$-BIC w.r.t. $\underline{\mathcal{D}} = \bigtimes_{i=1}^n \lfloor\mathcal{D}_i\rfloor_{\ell,\delta}$ and IR mechanism $M_1^{(\ell)}$, such that*

$$\textsc{Rev}_T\left(M_1^{(\ell)},\underline{\mathcal{D}}\right) \geq \textsc{Rev}(M,\mathcal{D}) - nm\mathcal{L}\delta.$$

The proof of Lemma 3 can be found in Appendix E.3. In the next Lemma, we make **Step (4)** formal.

Lemma 4. *For any $\delta > 0$, $\ell \in [0,\delta]^m$, and distribution $\widehat{\mathcal{D}}$, if $M_2^{(\ell)}$ is a ξ_2-BIC w.r.t. $\underline{\widehat{\mathcal{D}}} = \bigtimes_{i=1}^n \left\lfloor\widehat{\mathcal{D}}_i\right\rfloor_{\ell,\delta}$ and IR mechanism, we can transform $M_2^{(\ell)}$ into a mechanism $\widehat{M}^{(\ell)}$, so that $\widehat{M}$ is $(\xi_2 + 3m\mathcal{L}\delta)$-BIC w.r.t. $\widehat{\mathcal{D}}$, IR, and has revenue under truthful bidding $\textsc{Rev}_T\left(\widehat{M}^{(\ell)},\widehat{\mathcal{D}}\right) \geq \textsc{Rev}_T\left(M_2^{(\ell)},\underline{\widehat{\mathcal{D}}}\right) - nm\mathcal{L}\delta$. Moreover, the transformation does not rely on any knowledge of $\widehat{\mathcal{D}}$ or $\underline{\widehat{\mathcal{D}}}$.*

The proof of Lemma 4 is postpone to Appendix E.3.

3.4 Applications of Multi-Item Robustness

Lipschitz Continuity of the Optimal Revenue in Multi-item Auctions. Equipped with Theorem 1 and 3, we can easily argue the Lipschitz continuity of the optimal revenue in multi-item auctions (Theorem 6) as stated in the last column of the second half of Table 1. Due to Theorem 1 and 3, we know that the optimal revenue of a $O(\text{poly}(n,m,\mathcal{L},H,\varepsilon))$-BIC and IR mechanism w.r.t. distribution $\mathcal{F} = \bigtimes_{i\in[n]}\mathcal{F}_i$ is at least as large as the optimal revenue of a BIC and IR mechanism w.r.t. distribution $\widehat{\mathcal{F}} = \bigtimes_{i\in[n]}\widehat{\mathcal{F}}_i$, if $\left\|\mathcal{F}_i - \widehat{\mathcal{F}}_i\right\|_{TV} \leq \varepsilon, \forall i$ or $\left\|\mathcal{F}_i - \widehat{\mathcal{F}}_i\right\|_{P} \leq \varepsilon, \forall i$. According to Lemma 1, the optimal revenue of a $O(\text{poly}(n,m,\mathcal{L},H,\varepsilon))$-BIC and IR mechanism is at most $O(\text{poly}(n,m,\mathcal{L},H,\varepsilon))$ larger than the optimal revenue of a BIC and IR mechanism. Hence, $\text{OPT}(\mathcal{F}) \approx \text{OPT}(\widehat{\mathcal{F}})$. Please see Appendix E.4 for the formal statement and the proof of Theorem 6.

Approximation Preserving Transformation. One interesting implication of Theorem 6 is that the transformations of Theorems 1 and 3 are also approximation preserving. Given a a c-approximation mechanism M to the optimal revenue under distribution $\mathcal{D}$, applying the transformation in Theorem 3 (or Theorem 1) to M, we obtain a new mechanism $\widehat{M}$ that is $O(\text{poly}(n,m,\mathcal{L},H,\varepsilon))$-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR if $\left\|\mathcal{D}_i - \widehat{\mathcal{D}}_i\right\|_{P} \leq \varepsilon, \forall i$ (or if $\left\|\mathcal{D}_i - \widehat{\mathcal{D}}_i\right\|_{TV} \leq \varepsilon, \forall i$). Moreover, its revenue under truthful bidding is at least c fraction of the optimal $O(\text{poly}(n,m,\mathcal{L},H,\varepsilon))$-BIC revenue under $\widehat{\mathcal{D}}$ less a small additive term. The result is formally stated as Theorem 7 in Appendix E.5. Note that the third column of the second half of Table 1 is simply Theorem 7 with $c = 1$. Furthermore, if there is only a single bidder, the mechanism $\widehat{M}$ becomes exactly IC instead of approximately IC (Theorem 8).

Learning Multi-item Auctions under Item Independence. Since independent distributions are straightforward to learn within Prokhorov distance ε with polynomially many samples, the result of Gonczarowski and Weinberg [36] follows easily from our robustness result (see Theorem 9 in Appendix F).

Learning Multi-item Auctions under Structured Item Dependence. Going beyond product measures, we initiate the study of learning multi-item auctions when every bidder's item-values are dependent, but sampled from a joint distribution with structure. As we have already noted, arbitrary

joint distributions are both unnatural from a modeling perspective, as they require exponentially many bits to describe, and are also known to require exponentially many samples to even learn approximately optimal auctions [31]. We thus propose studying the learnability of auctions under the assumption that each bidder's item values are sampled from a Markov Random Field (MRF) or a Bayesian network (a.k.a. Bayeset). In fact, this is not really an assumption. These well-studied probabilistic frameworks, defined formally in Definitions 10 and 11 of Appendix H due to lack of space, are very flexible in that they can represent *any distribution*. The reason they are attractive from a modeling perspective is that they have a natural complexity parameter that controls how expressive they are, namely the maximum hyperedge size of an MRF and the maximum in-degree of a Bayesnet. Under the assumption that each bidder's item-values are drawn from an MRF or a Bayesnet of complexity d, we establish the results summarized in the last two rows of Table 2, whose main feature is that the sample complexity to learn an up-to-ϵ optimal auction is polynomial in the number of bidders n, the number of items m, the inverse approximation parameter $1/\varepsilon$, and other relevant parameters, and is only exponential in the complexity parameter d of the bidders' MRFs or Bayesian networks, as it should given the known lower bounds [31].

Our results for learning near-optimal auctions under MRF and Bayesnet assumptions are stated in more detail as Theorems 10 and 11 of Appendix H, and *can also accommodate unobservable variables* which makes their applicability very broad. In turn, these results are proven by combining our robustness result (Theorem 7) with new learnability results for MRFs and Bayesnets that we also establish, namely Theorems 12 and 13 of Appendix H respectively. These results are of independent interest and provide broad generalizations of the recent upper bounds of [30] for Gaussian MRFs and Ising models. While this recent work bounds the VC dimension of the Yatracos class of these families of distributions, for our more general families of non-parametric distributions we construct instead covers under either total variation distance or Prokhorov distance, and combine our cover-size upper bounds with generic tournament-style algorithms; see e.g. [1, 26, 29]. The details are provided in Appendix J. While there are many details, we illustrate one snippet of an idea used in constructing a ε-cover, in total variation distance, of the set of all MRFs with hyper-edges E of size at most d and a discrete alphabet Σ on every node. The proof argues that (i) the (appropriately normalized) log-potential functions of the MRF can be discretized to take values in the negative integers at a cost of ε in total variation distance; (ii) using properties of linear programming, it argues that using negative integers of bit complexity polynomial in $|E|$, $|\Sigma|^d$ and $\log(1/\varepsilon)$ suffices at another cost of ε in total variation distance. It thus argues that all MRFs can be covered by a set of MRFs of size exponential in poly $\left(|E|, |\Sigma|^d, \log(\frac{1}{\varepsilon})\right)$, which is sufficient to yield the required sample bounds using the tournament algorithm.

REFERENCES

[1] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. 2014. Sorting with adversarial comparators and application to density estimation. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*.

[2] Saeed Alaei. 2011. Bayesian Combinatorial Auctions: Expanding Single Buyer Mechanisms to Many Buyers. In *the 52nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*.

[3] Saeed Alaei, Hu Fu, Nima Haghpanah, Jason Hartline, and Azarakhsh Malekian. 2012. Bayesian Optimal Auctions via Multi- to Single-agent Reduction. In *the 13th ACM Conference on Electronic Commerce (EC)*.

[4] Saeed Alaei, Hu Fu, Nima Haghpanah, and Jason D. Hartline. 2013. The Simple Economics of Approximately Optimal Auctions. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*. 628–637. https://doi.org/10.1109/FOCS.2013.73

[5] Moshe Babaioff, Yannai A. Gonczarowski, and Noam Nisan. 2017. The menu-size complexity of revenue approximation. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, Hamed Hatami, Pierre McKenzie, and Valerie King (Eds.). ACM, 869–877. https://doi.org/10.1145/3055399.3055426

[6] Moshe Babaioff, Nicole Immorlica, Brendan Lucier, and S. Matthew Weinberg. 2014. A Simple and Approximately Optimal Mechanism for an Additive Buyer. In *the 55th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*.

[7] Dirk Bergemann and Karl Schlag. 2011. Robust monopoly pricing. *Journal of Economic Theory* 146, 6 (2011), 2527–2543.

[8] Anand Bhalgat, Sreenivas Gollapudi, and Kamesh Munagala. 2013. Optimal auctions via the multiplicative weight method. In *ACM Conference on Electronic Commerce, EC '13, Philadelphia, PA, USA, June 16-20, 2013*. 73–90. https://doi.org/10.1145/2482540.2482547

[9] Patrick Briest, Shuchi Chawla, Robert Kleinberg, and S. Matthew Weinberg. 2010. Pricing Randomized Allocations. In *the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*.

[10] Yang Cai and Constantinos Daskalakis. 2011. Extreme-Value Theorems for Optimal Multidimensional Pricing. In *the 52nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*.

[11] Yang Cai and Constantinos Daskalakis. 2017. Learning Multi-item Auctions with (or without) Samples. In *the 58th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*.

[12] Yang Cai, Constantinos Daskalakis, and S. Matthew Weinberg. 2012. An Algorithmic Characterization of Multi-Dimensional Mechanisms. In *the 44th Annual ACM Symposium on Theory of Computing (STOC)*.

[13] Yang Cai, Constantinos Daskalakis, and S. Matthew Weinberg. 2012. Optimal Multi-Dimensional Mechanism Design: Reducing Revenue to Welfare Maximization. In *the 53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*.

[14] Yang Cai, Constantinos Daskalakis, and S. Matthew Weinberg. 2013. Understanding Incentives: Mechanism Design becomes Algorithm Design. In *the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*.

[15] Yang Cai, Nikhil R. Devanur, and S. Matthew Weinberg. 2016. A Duality Based Unified Approach to Bayesian Mechanism Design. In *the 48th Annual ACM Symposium on Theory of Computing (STOC)*.

[16] Yang Cai and Zhiyi Huang. 2013. Simple and Nearly Optimal Multi-Item Auctions. In *the 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*.

[17] Yang Cai and Mingfei Zhao. 2017. Simple mechanisms for subadditive buyers via duality. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*. 170–183. https://doi.org/10.1145/3055399.3055465

[18] Shuchi Chawla, Jason D. Hartline, and Robert D. Kleinberg. 2007. Algorithmic Pricing via Virtual Valuations. In *the 8th ACM Conference on Electronic Commerce (EC)*.

[19] Shuchi Chawla, Jason D. Hartline, David L. Malec, and Balasubramanian Sivan. 2010. Multi-Parameter Mechanism Design and Sequential Posted Pricing. In *the 42nd ACM Symposium on Theory of Computing (STOC)*.

[20] Shuchi Chawla and J. Benjamin Miller. 2016. Mechanism Design for Subadditive Agents via an Ex-Ante Relaxation, In Proceedings of the 2016 ACM Conference on Economics and Computation, EC '16, Maastricht, The Netherlands, July 24-28, 2016. *CoRR*, 579–596. https://doi.org/10.1145/2940716.2940756

[21] Richard Cole and Tim Roughgarden. 2014. The sample complexity of revenue maximization. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*.

[22] Constantinos Daskalakis, Alan Deckelbaum, and Christos Tzamos. 2013. Mechanism design via optimal transport. In *ACM Conference on Electronic Commerce, EC '13, Philadelphia, PA, USA, June 16-20, 2013*. 269–286. https://doi.org/10.1145/2482540.2482593

[23] Constantinos Daskalakis, Alan Deckelbaum, and Christos Tzamos. 2014. The Complexity of Optimal Mechanism Design. In *the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*.

[24] Constantinos Daskalakis, Nikhil R. Devanur, and S. Matthew Weinberg. 2015. Revenue Maximization and Ex-Post Budget Constraints. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation, EC '15, Portland, OR, USA, June 15-19, 2015*. 433–447. https://doi.org/10.1145/2764468.2764521

[25] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. 2018. Testing ising models. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 1989–2007.

[26] Constantinos Daskalakis and Gautam Kamath. 2014. Faster and Sample Near-Optimal Algorithms for Proper Learning Mixtures of Gaussians. In *Proceedings of the 27th Conference on Learning Theory (COLT)*.

[27] Constantinos Daskalakis and S. Matthew Weinberg. 2012. Symmetries and Optimal Multi-Dimensional Mechanism Design. In *the 13th ACM Conference on Electronic Commerce (EC)*.

[28] Nikhil R Devanur, Zhiyi Huang, and Christos-Alexandros Psomas. 2016. The sample complexity of auctions with side information. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM, 426–439.

[29] Luc Devroye and Gábor Lugosi. 2012. *Combinatorial methods in density estimation*. Springer Science & Business Media.

[30] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. 2018. The minimax learning rate of normal and Ising undirected graphical models. *arXiv preprint arXiv:1806.06887* (2018).

[31] Shaddin Dughmi, Li Han, and Noam Nisan. 2014. Sampling and representation complexity of revenue maximization. In *Proceedings of the 10th International Conference on Web and Internet Economics (WINE)*.

[32] Edith Elkind. 2007. Designing and learning optimal finite support auctions. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 736–745.

[33] Alison L Gibbs and Francis Edward Su. 2002. On choosing and bounding probability metrics. *International statistical review* 70, 3 (2002), 419–435.

[34] Kira Goldner and Anna R Karlin. 2016. A prior-independent revenue-maximizing auction for multiple additive bidders. In *International Conference on Web and Internet Economics*. Springer, 160–173.

[35] Yannai A Gonczarowski and Noam Nisan. 2017. Efficient empirical revenue maximization in single-parameter auction environments. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*.

[36] Yannai A Gonczarowski and S Matthew Weinberg. 2018. The Sample Complexity of Up-to-ε Multi-Dimensional Revenue Maximization. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 416–426.

[37] Chenghao Guo, Zhiyi Huang, and Xinzhi Zhang. 2019. Settling the Sample Complexity of Single-parameter Revenue Maximization. In *the 51st Annual ACM Symposium on Theory of Computing (STOC)*.

[38] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. 2011. *Robust statistics: the approach based on influence functions*. Vol. 196. John Wiley & Sons.

[39] Sergiu Hart and Noam Nisan. 2013. The menu-size complexity of auctions. In *the 14th ACM Conference on Electronic Commerce (EC)*.

[40] Sergiu Hart and Philip J. Reny. 2012. Maximal Revenue with Multiple Goods: Nonmonotonicity and Other Observations. *Discussion Paper Series dp630, The Center for the Study of Rationality, Hebrew University, Jerusalem* (2012).

[41] Zhiyi Huang, Yishay Mansour, and Tim Roughgarden. 2015. Making the most of your samples. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*.

[42] Peter J Huber. 2011. *Robust statistics*. Springer.

[43] Finn V Jensen. 1996. *An introduction to Bayesian networks*. Vol. 210. UCL press London.

[44] Ross Kindermann and Laurie Snell. 1980. *Markov random fields and their applications*. American Mathematical Society.

[45] Pravesh Kothari, Sahil Singla, Divyarthi Mohan, Ariel Schvartzman, and S. Matthew Weinberg. 2019. Approximation Schemes for a Unit-Demand Buyer with Independent Items via Symmetries. In *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019*, David Zuckerman (Ed.). IEEE Computer Society, 220–232. https://doi.org/10.1109/FOCS.2019.00023

[46] Mehryar Mohri and Andres Munoz Medina. 2014. Learning Theory and Algorithms for revenue optimization in second price auctions with reserve.. In *ICML*. 262–270.

[47] Jamie Morgenstern and Tim Roughgarden. 2016. Learning simple auctions. In *Proceedings of the 30th Annual Conference on Learning Theory (COLT)*.

[48] Jamie H Morgenstern and Tim Roughgarden. 2015. On the pseudo-dimension of nearly optimal auctions. In *Proceedings of the the 29th Annual Conference on Neural Information Processing Systems (NIPS)*.

[49] Roger B. Myerson. 1981. Optimal Auction Design. *Mathematics of Operations Research* 6, 1 (1981), 58–73.

[50] Thomas Dyhre Nielsen and Finn Verner Jensen. 2009. *Bayesian networks and decision graphs*. Springer Science & Business Media.

[51] Judea Pearl. 2009. *Causality*. Cambridge university press.

[52] Tim Roughgarden and Okke Schrijvers. 2016. Ironing in the dark. In *Proceedings of the 2016 ACM Conference on Economics and Computation*. ACM, 1–18.

[53] Aviad Rubinstein and S. Matthew Weinberg. 2015. Simple Mechanisms for a Subadditive Buyer and Applications to Revenue Monotonicity. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation, EC '15, Portland, OR, USA, June 15-19, 2015*. 377–394. https://doi.org/10.1145/2764468.2764510

[54] Volker Strassen. 1965. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics* 36, 2 (1965), 423–439.

[55] Vasilis Syrgkanis. 2017. A sample complexity measure with applications to learning optimal auctions. In *Advances in Neural Information Processing Systems*. 5352–5359.

[56] Andrew Chi-Chih Yao. 2015. An n-to-1 Bidder Reduction for Multi-item Auctions and its Applications, In SODA. *CoRR* (2015). http://arxiv.org/abs/1406.3278

[57] Andrew Chi-Chih Yao. 2017. Dominant-Strategy versus Bayesian Multi-item Auctions: Maximum Revenue Determination and Comparison. In *Proceedings of the 2017 ACM Conference on Economics and Computation, EC '17, Cambridge, MA, USA, June 26-30, 2017*. 3–20. https://doi.org/10.1145/3033274.3085120

A FURTHER RELATED WORK

As described earlier, most prior work on learning multi-item auctions follows a PAC-learning approach, bounding the statistical complexity of classes of mechanisms that are (approximately) optimal for the setting of interest. The statistical complexity measures that are used for this purpose are the standard notions of pseudodimension, which generalizes VC dimension to real valued functions, and Rademacher complexity. In particular, Morgenstern and Roughgarden [47] and Syrgkanis [55] bound respectively the pseudodimension and Rademacher complexity of simple classes of mechanisms that have been shown in the literature to contain approximately optimal mechanisms in multi-item multi-bidder settings satisfying item-independence [6, 15, 17, 19, 56]. The classes of mechanisms studied by these works contain approximately optimal mechanisms in multi-item settings with item-independence and either multiple unit-demand/additive bidders, or a single subadditive bidder. More powerful classes of simple mechanisms are also known in the literature. The state-of-the-art is the *sequential two-part tariff mechanism* considered by Cai and Zhao [17], which is shown to approximate the optimal revenue in multi-item settings even with multiple bidders whose valuations are fractionally subadditive, again under item-independence. Unfortunately, both the pseudodimension and the empirical Rademacher complexity of sequential two-part tariff mechanisms are already exponential even in two bidder settings, making these measures unsuitable tools for showing the learnability of two-part tariff mechanisms.

An important feature of the afore-described works is that bounding the pseudo-dimension or empirical Rademacher complexity of mechanism classes is oblivious to the structure in the distribution. Hence, while the mechanisms considered in these works are only approximately optimal under item-independence, the independence cannot be exploited. In contrast to empirical Rademacher complexity, Rademacher complexity *is* sensitive to the underlying distribution, but bounds exploiting the structure of the distribution are not easy to obtain. This observation motivated another line of work which heavily exploits the structure of the distributions of interest to choose both the class of mechanisms *and* the statistical complexity measure to bound their learnability. So far, this approach has only been applied to settings satisfying item-independence. Indeed, Cai and Daskalakis [11] propose a statistical complexity measure that is tailored to product distributions, and use their new measure to establish learnability of sequential two-part tariff mechanisms under item-independence. Gonczarowski and Weinberg [36] choose a finite class of mechanisms so that an up-to-ε optimal mechanism is guaranteed to exist in the class. For item-independent distributions, the size of this class is only singly exponential implying polynomial sample learnability. Unfortunately, the size becomes doubly exponential for correlated items turning the sample complexity exponential.

Finally, Goldner and Karlin [34] do not use a PAC-learning based approach. They show how to learn approximately optimal auctions in the multi-item multi-bidder setting with additive bidders using only one sample from each bidder's distribution, assuming that it is regular and independent across items. Their approach is tailored for a mechanism designed by Yao [56] and does not apply to broader settings.

B ADDITIONAL PRELIMINARIES

Definition 4 (Total Variation Distance). *The* total variation distance *between two probability measures P and Q on a σ-algebra $\mathcal{F}$ of subsets of some sample space Ω, denoted $||P - Q||_{TV}$, is defined as*

$$\sup_{E \in \mathcal{F}} |P(E) - Q(E)| .$$

Definition 5 (Kolmogorov Distance). *The* Kolmogorov distance *between two distributions P and Q over $\mathbb{R}$, denoted $\|P-Q\|_K$, is defined as*

$$\sup_{x\in\mathbb{R}}\left|\Pr_{X\sim P}[X\le x]-\Pr_{X\sim Q}[X\le x]\right|.$$

Definition 6 (Lévy Distance). *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two probability distributions on $\mathbb{R}$ with cumulative distribution functions F and G respectively. Then we denote their Lévy distance by*

$$\|\mathcal{D}_1-\mathcal{D}_2\|_L=\inf\{\varepsilon>0: F(x-\varepsilon)-\varepsilon\le G(x)\le F(x+\varepsilon)+\varepsilon,\ \forall x\in\mathbb{R}\}$$

Multi-item Auctions: We focus on revenue maximization in the combinatorial auction with n **bidders** and m **heterogenous items**.

The outcomes of the auction lie in $X\subseteq\{0,1\}^{n\cdot m}$ such that for any allocation $x\in X$, $x_{i,j}$ is the probability that bidder i receives item j. Formally, $X=\left\{(x_{i,j})_{i\in[n],j\in[m]}\in\{0,1\}^{nm}\mid\forall j:\sum_{i=1}^{n}x_{i,j}\le 1\right\}$. Each bidder $i\in[n]$ has a valuation function $v_i(\cdot):X\to\mathbb{R}$ that maps an allocations of items to a real number. In this paper, we assume the function $v_i(\cdot)$ is parametrized by $(v_{i,1},\ldots,v_{i,m})$, where $v_{i,j}$ is bidder i's value for item j. We will refer to the vector $(v_{i,1},\ldots,v_{i,m})$ as bidder i's type, and we assume that *each bidder's type is drawn independently* from some distribution. [9] Throughout this paper, we assume all bidders types lie in $[0,H]^m$.

Mechanisms, Payments, and Utility: We use $p=(p_1,\ldots,p_n)$ to specify the payments for the bidders. Given some prices $p=(p_1,\ldots,p_n)$, allocation x and type v_i, denote the quasilinear utility of bidder $i\in[n]$ by $u_i(v_i,(x,p))=v_i(x)-p_i$. Let $M=(x(\cdot),p(\cdot))$ be a mechanism with allocation rule $x(\cdot)$ and payment rule $p(\cdot)$. For any input bid vector $b=(b_1,\ldots,b_n)$, the allocation rule outputs a distribution over allocations $x(b)\in\Delta(X)$ and payments $p(b)=(p_1(b),\ldots,p_n(b))$. Then $u_i(v_i,M(b))=v_i(x(b))-p_i(b)$. If bidder i's type is v_i, then her utility under input bid vector b is $u_i(v_i,M(b))=\mathbb{E}[v_i(x(b))-p_i(b)]$, where the expectation is over the randomness of the allocation and payment rule.

ε-Incentive Compatible and Individually Rational:

- *Ex-post Individually Rational (IR):* M is *IR* if for all types $v\in[0,H]^{n\cdot m}$ and all bidders $i\in[n]$,
$$u_i(v_i,M(v_i,v_{-i}))\ge 0.$$
- *ε-Dominant Strategy Incentive Compatible (ε-DSIC):* if for all $i\in[n]$, $v\in[0,H]^{n\cdot m}$ and potential misreports $v_i'\in[0,H]^m$ of bidder i, $u_i(v_i,M(v_i,v_{-i}))\ge u_i(v_i,M((v_i',v_{-i})))-\varepsilon$. A mechanism is DSIC if it is 0-DSIC.
- *ε-Bayesian Incentive Compatible (ε-BIC):* if bidders draw their values from some distribution $\mathcal{F}=(\mathcal{F}_1,\ldots,\mathcal{F}_n)$, then define M to be *ε-BIC with respect to $\mathcal{F}$* if
$$\mathbb{E}_{v_{-i}\sim\mathcal{F}_{-i}}[u_i(v_i,M(v_i,v_{-i}))]\ge\mathbb{E}_{v_{-i}\sim\mathcal{F}_{-i}}[u_i(v_i,M(v_i',v_{-i}))]-\varepsilon,$$
for all potential misreports v_i', in expectation over all other bidders bid v_{-i}. A mechanism is BIC if it is 0-BIC.

If there is only one bidder, the definition of DSIC coincides with the definition of BIC, and we simply use ε-IC to describe the incentive compatibility of single bidder mechanisms.

In single-bidder case, there is a well known transformation, Lemma 5, that maps any ε-IC mechanism to an IC mechanism with negligible revenue loss. To the best of our knowledge, the result is attributed Nisan in [18, 36, 39] and many other papers.

[9] We will not explicitly write bidder i's valuation as $v_{i,\mathbf{v_i}}(\cdot)$ where $\mathbf{v_i}=(v_{i,1},\ldots,v_{i,m})$.

Lemma 5 (Nisan, circa 2005). *Let M be an IR and ε-IC mechanism for a single bidder, and $\mathcal{D}$ be the bidder's type distribution. Modifying each possible allocation and payment pair by multiplying the payment by $1-\sqrt{\varepsilon}$ and letting the bidder choose the (modified) allocation and payment pair that maximizes her utility yields an IR and IC mechanism M' with expected revenue at least $(1-\sqrt{\varepsilon})(\textsc{Rev}_T(M,\mathcal{D})-\sqrt{\varepsilon})$. Importantly, the modification does not require any knowledge of $\mathcal{D}$.*

Interested readers can find a proof of Lemma 5 in [36].

Up-to-ε Optimal Mechanisms: We say a mechanism M is up-to-ε optimal under distribution $\mathcal{D}$, if

$$\textsc{Rev}_T(M,\mathcal{D}) \geq \text{OPT}(\mathcal{D}) - \varepsilon.$$

C LÉVY-ROBUSTNESS FOR SINGLE-ITEM AUCTIONS

In this section, we show the robustness result under the Lévy distance in the single-item setting. If we are given a model distribution $\mathcal{D}_i$ that is ε-close to the true distribution $\widehat{\mathcal{D}}_i$, in Lévy distance, for every bidder $i \in [n]$, we show how to design a mechanism M^* only based on $\mathcal{D} = \bigtimes_{i=1}^n \mathcal{D}_i$ and extracts revenue that is at most $O(nH\cdot\varepsilon)$ less than the optimal revenue under any possible true distribution $\widehat{\mathcal{D}} = \bigtimes_{i=1}^n \widehat{\mathcal{D}}_i$.

Theorem 4 (Lévy-Robustness for Single-item Auctions). *Given $\mathcal{D} = \bigtimes_{i=1}^n \mathcal{D}_i$, where $\mathcal{D}_i$ is an arbitrary distributions supported on $[0,H]$ for all $i \in [n]$. We can design a DSIC and IR mechanism M^* based on $\mathcal{D}$ such that for any product distribution $\widehat{\mathcal{D}} = \bigtimes_{i=1}^n \widehat{\mathcal{D}}_i$ satisfying $\left\|\mathcal{D}_i - \widehat{\mathcal{D}}_i\right\|_L \leq \varepsilon$ for all $i \in [n]$, we have:*

$$\textit{Rev}(M^*, \widehat{\mathcal{D}}) \geq \textit{OPT}(\widehat{\mathcal{D}}) - O(nH\cdot\varepsilon).$$

Let us sketch the proof of Theorem 4. We prove our statement in three steps.

- **Step (i):** We first identify the "best" and "worst" distributions (Definition 7), in terms of the first-order stochastic dominance (Definition 8), among all distributions in the ε-Lévy-ball around the model distribution $\mathcal{D}$. We construct the optimal mechanism M^* w.r.t. the "worst" distribution, and show that its revenue under any possible true distribution is at least M^*'s revenue under the "worst" distribution (Lemma 8). The statement provides a lower bound of M^*'s revenue under the unknown true distribution. Its proof follows from the revenue monotonicity lemma (Lemma 7) shown in [28].
- **Step (ii):** We use the revenue monotonicity lemma again to show the optimal revenue under the true distribution $\widehat{\mathcal{D}}$ is upper bounded by the optimal revenue under the "best" distribution(Lemma 9).
- **Step (iii):** We complete the proof by argueing that M^*'s revenue under the "worst" distribution can be at most $O(nH\cdot\varepsilon)$ worst than the optimal revenue under the "best" distribution (Lemma 10). The statement follows from a robustness theorem for single-item auctions under the Kolmogorov distance (Theorem 5).

We show Step (i) and (ii) in Section C.1 and Step (iii) in Section C.2.

C.1 Best and Worst Distributions in the ε-Lévy-Ball

We formally define the "best" and "worst" distributions in the ε-Lévy-ball around the model distribution.

Definition 7. *For every $i \in [n]$, we define $\overline{\mathcal{D}}_i$ and $\underline{\mathcal{D}}_i$ based on $\mathcal{D}_i$. $\overline{\mathcal{D}}_i$ is supported on $[0, H+\varepsilon]$, and its CDF is defined as $F_{\overline{\mathcal{D}}_i}(x) = \max\left\{F_{\mathcal{D}_i}(x-\varepsilon)-\varepsilon, 0\right\}$. $\underline{\mathcal{D}}_i$ is supported on $[-\varepsilon, H]$, and its CDF is defined as $F_{\underline{\mathcal{D}}_i}(x) = \min\left\{F_{\mathcal{D}_i}(x+\varepsilon)+\varepsilon, 1\right\}$.*

We provide a more intuitive interpretation of $\overline{\mathcal{D}}_i$ and $\underline{\mathcal{D}}_i$ here. To obtain $\overline{\mathcal{D}}_i$, we first shift all values in $\mathcal{D}_i$ to the right by ε, then we move the bottom ε probability mass to $H+\varepsilon$. To obtain $\underline{\mathcal{D}}_i$, we first shift all values in $\mathcal{D}_i$ to the left by ε, then we move the top ε probability mass to $-\varepsilon$. It is not hard to see that both $\overline{\mathcal{D}}_i$ and $\underline{\mathcal{D}}_i$ are still in the ε-ball around $\mathcal{D}_i$ in Lévy distance. More importantly, $\overline{\mathcal{D}}_i$ and $\underline{\mathcal{D}}_i$ are the "best" and "worst" distributions in the ε-Lévy-ball under first-order-stochastic-dominance.

Definition 8 (First-Order Stochastic Dominance). *We say distribution B* first-order stochastically dominates *A iff $F_B(x) \le F_A(x)$ for all $x \in \mathbb{R}$. We use $A \preccurlyeq B$ to denote that distribution B first-order stochastically dominates distribution A. If $\mathbf{A} = \times_{i=1}^n A_i$ and $\mathbf{B} = \times_{i=1}^n B_i$ are two product distributions, and $A_i \preccurlyeq B_i$ for all $i \in [n]$, we slightly abuse the notation $\preccurlyeq$ to write $\mathbf{A} \preccurlyeq \mathbf{B}$.*

Lemma 6. *For any $\widehat{\mathcal{D}}_i$, such that $\left\|\widehat{\mathcal{D}}_i - \mathcal{D}_i\right\|_L \le \varepsilon$, we have $\underline{\mathcal{D}}_i \preccurlyeq \widehat{\mathcal{D}}_i \preccurlyeq \overline{\mathcal{D}}_i$.*

Proof. It follows from the definition of Lévy distance and Definition 7. For any x,

$$F_{\widehat{\mathcal{D}}_i}(x) \in [F_{\mathcal{D}_i}(x-\varepsilon)-\varepsilon, F_{\mathcal{D}_i}(x+\varepsilon)+\varepsilon].$$

Clearly, $0 \le F_{\widehat{\mathcal{D}}_i}(x) \le 1$, so we have $F_{\overline{\mathcal{D}}_i}(x) \le F_{\widehat{\mathcal{D}}_i}(x) \le F_{\underline{\mathcal{D}}_i}(x)$ for all x. □

The plan is to construct the optimal mechanism for $\underline{\mathcal{D}} = \bigtimes_{i=1}^n \underline{\mathcal{D}}_i$ and show that this mechanism achieves up-to-$O(nH\cdot\varepsilon)$ optimal revenue under any possible true distribution $\mathcal{D}$.

Next, we state a revenue monotonicity lemma that will be useful. We first need the following definition.

Definition 9 (Extension of a Mechanism to All Values). *Suppose a mechanism $M = (x, p)$ is defined for all value profiles in $T = \times_{i=1}^n T_i$. Define its extension $M' = (x', p')$ to all values. We only specify x', as p' can be determined by the payment identity given x'. x' first rounds the bid of each bidder i down to the closest value in T_i, and then apply allocation rule x on the rounded bids. If some bidder i's bid is smaller than the lowest value in T_i, x' does not allocate the item to any bidder.*

Observe that the extension provides a DSIC and IR mechanism for all values if the original mechanism is DSIC and IR.

Lemma 7 (Strong Revenue Monotonicity [28]). *Let $\mathcal{F} = \bigtimes_{i=1}^n \mathcal{F}_i$ be a product distributions. There exists an optimal DSIC and IR mechanism M for $\mathcal{F}$ such that, for any product distribution $\mathcal{F}' = \bigtimes_{i=1}^n \mathcal{F}_i' \succcurlyeq \mathcal{F}$,*

$$\textsc{Rev}(M', \mathcal{F}') \ge \textsc{Rev}(M, \mathcal{F}) = OPT(\mathcal{F}).$$

M' is the extension of M. In particular, this implies $OPT(\mathcal{F}') \ge OPT(\mathcal{F})$.

Combining Lemma 6 and 7, we show that if M^* is the extension of the optimal mechanism for $\underline{\mathcal{D}}$, it achieves at least OPT($\underline{\mathcal{D}}$) under any distribution $\widehat{\mathcal{D}}$ with $\left\|\widehat{\mathcal{D}}_i - \mathcal{D}_i\right\|_L \le \varepsilon$.

Lemma 8. *Let M^* be the extension of the optimal DSIC and IR mechanism for $\underline{\mathcal{D}}$. For any product distribution $\widehat{\mathcal{D}} = \bigtimes_{i=1}^n \widehat{\mathcal{D}}_i$ with $\left\|\widehat{\mathcal{D}}_i - \mathcal{D}_i\right\|_L \le \varepsilon$ for all $i \in [n]$, we have the following:*

$$\textsc{Rev}(M^*, \widehat{\mathcal{D}}) \ge OPT(\underline{\mathcal{D}}).$$

Proof. Since $\widehat{\mathcal{D}} \succcurlyeq \underline{\mathcal{D}}$ (Lemma 6), the claim follows from Lemma 7. □

Lemma 8 shows that with only knowledge of the model distribution $\mathcal{D}$, we can design a mechanism whose revenue under any possible true distribution $\widehat{\mathcal{D}}$ is at least OPT($\underline{\mathcal{D}}$). Next, we upper bound the optimal revenue under $\widehat{\mathcal{D}}$ with the optimal revenue under $\overline{\mathcal{D}}$.

Lemma 9. *For any product distribution $\widehat{\mathcal{D}}$ with $\left\|\widehat{\mathcal{D}}_i - \mathcal{D}_i\right\|_L \le \varepsilon$ for all $i \in [n]$, we have the following:*

$$OPT(\overline{\mathcal{D}}) \ge OPT(\widehat{\mathcal{D}}).$$

Proof. Since $\overline{\mathcal{D}} \succcurlyeq \widehat{\mathcal{D}}$ (Lemma 6), the claim follows from Lemma 7. □

C.2 Comparing the Revenue of the Best and Worst Distributions

In this section, we show that our lower bound of M^*'s revenue under the true distribution $\widehat{\mathcal{D}}$ and our upper bound of the optimal revenue under $\widehat{\mathcal{D}}$ are at most $O(nH \cdot \varepsilon)$ away.

Lemma 10.

$$OPT(\underline{\mathcal{D}}) \ge OPT(\overline{\mathcal{D}}) - O(nH \cdot \varepsilon).$$

It is a priori not clear why Lemma 10 should be true, as $\overline{\mathcal{D}}$ is the "best" distribution and $\underline{\mathcal{D}}$ is the "worst" distribution in the ε-Lévy-ball around $\mathcal{D}$. We prove Lemma 10 by introducing another two auxiliary distributions $\underset{\sim}{\mathcal{D}}$ and $\widetilde{\mathcal{D}}$. In particular, we construct $\widetilde{\mathcal{D}}_i$ by shifting all values in $\mathcal{D}_i$ to the right by ε, and construct $\underset{\sim}{\mathcal{D}}_i$ by shifting all values in $\mathcal{D}_i$ to the left by ε. There are two important properties of these two new distributions: (i) one can couple $\underset{\sim}{\mathcal{D}}_i$ with $\widetilde{\mathcal{D}}_i$ so that the two random variables are always exactly 2ε away from each other; (ii) $\underset{\sim}{\mathcal{D}}_i$ and $\underline{\mathcal{D}}_i$ are within Kolmogorov distance ε, and $\widetilde{\mathcal{D}}_i$ and $\overline{\mathcal{D}}_i$ are also within Kolmogorov distance ε. Property (i) allows us to prove that $\left|\mathrm{OPT}(\underset{\sim}{\mathcal{D}}) - \mathrm{OPT}(\widetilde{\mathcal{D}})\right| \le 2\varepsilon$ (see Claim 2). To make use of property (ii), we prove the following robustness theorem w.r.t. the Kolmogorov distance.

Theorem 5. *For any buyer $i \in [n]$, let $\mathcal{D}_i$ and $\widehat{\mathcal{D}}_i$ be two arbitrary distributions supported on $(-\infty, H]$ such that $\left\|\mathcal{D}_i - \widehat{\mathcal{D}}_i\right\|_K \le \varepsilon$. We have the following:*

$$OPT\left(\widehat{\mathcal{D}}\right) \ge OPT(\mathcal{D}) - 3nH \cdot \varepsilon.$$

where $\mathcal{D} = \bigtimes_{i=1}^n \mathcal{D}_i$ and $\widehat{\mathcal{D}} = \bigtimes_{i=1}^n \widehat{\mathcal{D}}_i$.

The proof of Theorem 5 is postponed to Appendix D. Equipped with Theorem 5, we can immediately show that $\left|\mathrm{OPT}(\underset{\sim}{\mathcal{D}}) - \mathrm{OPT}(\underline{\mathcal{D}})\right| \le O(nH \cdot \varepsilon)$ and $\left|\mathrm{OPT}(\widetilde{\mathcal{D}}) - \mathrm{OPT}(\overline{\mathcal{D}})\right| \le O(nH \cdot \varepsilon)$. Lemma 10 follows quite easily from Claim 2 and the two inequalities above. The complete proof of Lemma 10 can be found in Appendix D.

We are now ready to prove Theorem 4.

Proof of Theorem 4: We first construct $\underline{\mathcal{D}}$ based on $\mathcal{D}$ and let M^* be the extension of the optimal mechanism for $\underline{\mathcal{D}}$. By Lemma 8, we know $\textsc{Rev}(M^*, \widehat{\mathcal{D}})$ is at least $\mathrm{OPT}(\underline{\mathcal{D}})$ for any $\widehat{\mathcal{D}}$. We also know that the optimal revenue under $\widehat{\mathcal{D}}$ is at most $\mathrm{OPT}(\overline{\mathcal{D}})$ by Lemma 9, and $\mathrm{OPT}(\overline{\mathcal{D}}) \le \mathrm{OPT}(\underline{\mathcal{D}}) + O(nH \cdot \varepsilon)$ by Lemma 10. Therefore,

$$\textsc{Rev}(M^*, \widehat{\mathcal{D}}) \ge \mathrm{OPT}(\overline{\mathcal{D}}) - O(nH \cdot \varepsilon) \ge \mathrm{OPT}(\widehat{\mathcal{D}}) - O(nH \cdot \varepsilon).$$

□

A simple corollary of Theorem 4 is the continuity of the optimal revenue under Lévy distance in single-item settings.

Corollary 1. *If $\mathcal{D}_i$ and $\widehat{\mathcal{D}}_i$ are supported on $[0, H]$, and $\left\|\mathcal{D}_i - \widehat{\mathcal{D}}_i\right\|_L \le \varepsilon$ for all $i \in [n]$, then*

$$\left|OPT(\mathcal{D}) - OPT(\widehat{\mathcal{D}})\right| \le O\left(nH \cdot \varepsilon\right),$$

where $\mathcal{D} = \bigtimes_{i=1}^{n} \mathcal{D}_i$ *and* $\widehat{\mathcal{D}} = \bigtimes_{i=1}^{n} \widehat{\mathcal{D}}_i$.

D MISSING PROOFS FROM SECTION C

Proof of Theorem 5: We prove the claim using a hybrid argument. We construct a collection of distributions, where $\mathcal{D}^{(0)} = \mathcal{D}$, $\mathcal{D}^{(i)} = \widehat{\mathcal{D}}_1 \times \cdots \times \widehat{\mathcal{D}}_i \times \mathcal{D}_{i+1} \times \cdots \times \mathcal{D}_n$ for all $1 \le i < n$, and $\mathcal{D}^{(n)} = \widehat{\mathcal{D}}$. We first show the following claim

Claim 1.

$$OPT\left(\mathcal{D}^{(i)}\right) \ge OPT\left(\mathcal{D}^{(i-1)}\right) - 3H \cdot \varepsilon,$$

for all $i \in [n]$.

Proof. W.l.o.g, we can assume the optimal mechanism for $\mathcal{D}^{(i-1)}$ is a deterministic. We use $M = (x, p)$ to denote it. In particular, there exists a collection of monotone non-decreasing functions $\{\mu_j(\cdot)\}_{\{j\in[n]\}}$ such that $\mu_j : \text{supp}\left(\mathcal{D}_j^{(i-1)}\right) \mapsto \mathbb{R}$. We extend the function $\mu_j(\cdot)$ to the whole interval $(-\infty, H]$. We slightly abuse notation and still call the extended function $\mu_j(\cdot)$. For any $z \in \text{supp}\left(\mathcal{D}_j^{(i-1)}\right)$, $\mu_j(x)$ remains the same. For any $z > \inf \text{supp}\left(\mathcal{D}_j^{(i-1)}\right)$, let

$$\mu_j(z) = \sup\left\{\mu_j(w) \mid w \le z \text{ and } w \in \text{supp}\left(\mathcal{D}_j^{(i-1)}\right)\right\}.$$

If $z \le \inf \text{supp}\left(\mathcal{D}_j^{(i-1)}\right)$ and $\notin \text{supp}\left(\mathcal{D}_j^{(i-1)}\right)$, let $\mu_j(z) = -\infty$.

Now we define a mechanism $M' = (x', p')$ for $\mathcal{D}^{(i)}$ based on the extended $\{\mu_j(\cdot)\}_{\{j\in[n]\}}$. For every profile v, let the bidder j^* with the highest positive $\mu_j(v_j)$ be the winner. If no bidder j has positive $\mu_j(v_j)$, the item is unallocated. When there are ties, break the tie in alphabetical order. Clearly, the allocation rule is monotone. According to Myerson's payment identity, if a bidder wins the item, she should pay $\inf\{z \mid z \text{ is a winning bid}\}$.

To prove the claim, we demonstrate the following two statements: for every fixed v_{-i} **(A1:)** bidder i's expected payments under $\mathcal{D}^{(i)}$ and $\mathcal{D}^{(i-1)}$ are within $O(H \cdot \varepsilon)$; **(A2:)** the total expected payments of all bidders except i under $\mathcal{D}^{(i)}$ and $\mathcal{D}^{(i-1)}$ are within $O(H \cdot \varepsilon)$. We first prove A1.

Proof of A1: For every fixed v_{-i}, let $\ell^* = \text{argmax}_{\ell \ne i} \mu_\ell(v_\ell)$. For bidder i to win the item, $\mu_i(v_i)$ needs to be greater than $\mu_{\ell^*}(v_{\ell^*})$. Therefore, there exists a threshold $\theta(v_{-i})$ for every fixed v_{-i}, such that bidder i wins the item iff $v_i \ge \theta(v_{-i})$. Clearly,

$$\mathbb{E}_{v_i\sim\widehat{\mathcal{D}}_i}[p_i'(v_i, v_{-i})] = \theta(v_{-i}) \cdot \Pr_{v_i\sim\widehat{\mathcal{D}}_i}[v_i \ge \theta(v_{-i})],$$

and

$$\mathbb{E}_{v_i\sim\mathcal{D}_i}[p_i(v_i, v_{-i})] = \theta(v_{-i}) \cdot \Pr_{v_i\sim\mathcal{D}_i}[v_i \ge \theta(v_{-i})].$$

Since $\left\|\mathcal{D}_i - \widehat{\mathcal{D}}_i\right\|_K \le \varepsilon$, $\left|\Pr_{v_i\sim\widehat{\mathcal{D}}_i}[v_i \ge \theta(v_{-i})] - \Pr_{v_i\sim\mathcal{D}_i}[v_i \ge \theta(v_{-i})]\right| \le \varepsilon$, which implies that

$$\begin{aligned}
&\left|\mathbb{E}_{v\sim\mathcal{D}^{(i)}}[p_i'(v)] - \mathbb{E}_{v\sim\mathcal{D}^{(i-1)}}[p_i(v)]\right| \\
&\le \mathbb{E}_{v_{-i}\sim\mathcal{D}_{-i}^{(i)}}\left[\left|\mathbb{E}_{v_i\sim\widehat{\mathcal{D}}_i}[p_i'(v_i, v_{-i})] - \mathbb{E}_{v_i\sim\mathcal{D}_i}[p_i(v_i, v_{-i})]\right|\right] \\
&\le \mathbb{E}_{v_{-i}\sim\mathcal{D}_{-i}^{(i)}}[\theta(v_{-i}) \cdot \varepsilon] \\
&\le H \cdot \varepsilon
\end{aligned}$$

This completes the argument for statement A1. Next, we prove statement A2.

Proof of A2: Since there is only one item, only the winner ℓ^* has non-zero payment and $\sum_{\ell \neq i} p_\ell(\mathbf{v}) = p_{\ell^*}(v)$ for any v_i. Our goal now is to bound the difference between $\mathbb{E}_{v_i \sim \mathcal{D}_i}[p_{\ell^*}(v)]$ and $\mathbb{E}_{v_i \sim \widehat{\mathcal{D}}_i}\left[p'_{\ell^*}(v)\right]$. Note that

$$\mathbb{E}_{v_i \sim \mathcal{D}_i}[p_{\ell^*}(v)] = \int_0^H \Pr_{v_i \sim \mathcal{D}_i}[p_{\ell^*}(v) > t]dt.$$

When $\mu_{\ell^*}(t) \geq \mu_{\ell^*}(v_{\ell^*})$, $\Pr_{v_i \sim \mathcal{D}_i}[p_{\ell^*}(v) > t] = 0$, so we only consider the case where $\mu_{\ell^*}(t) < \mu_{\ell^*}(v_{\ell^*})$. Let $\alpha = \max_{\ell \neq i \text{ or } \ell^*} \mu_\ell(v_\ell)$. $p_{\ell^*}(v) > t$ is equivalent to having $\max\{\alpha, \mu_i(v_i)\} > \mu_{\ell^*}(t)$ and $\mu_i(v_i) < \mu_{\ell^*}(v_{\ell^*})$ if $\ell^* > i$ (or $\mu_i(v_i) \leq \mu_{\ell^*}(v_{\ell^*})$ if $\ell^* < i$). Since $\mu_i(\cdot)$ is monotone, it is not hard to observe that this is equivalent to having v_i lying in some interval that only depends on v_{-i}. Let the lower bound of the interval be $a(v_{-i})$ and the upper bound be $b(v_{-i})$. Similarly, we know

$$\mathbb{E}_{v_i \sim \widehat{\mathcal{D}}_i}[p_{\ell^*}(v)] = \int_0^H \Pr_{v_i \sim \widehat{\mathcal{D}}_i}[p'_{\ell^*}(v) > t]dt,$$

and $\Pr_{v_i \sim \widehat{\mathcal{D}}_i}[p'_{\ell^*}(v) > t]$ is also the probability that v_i lies between $a(v_{-i})$ and $b(v_{-i})$. Since $\left\|\mathcal{D}_i - \widehat{\mathcal{D}}_i\right\|_K \leq \varepsilon$, $\left|\Pr_{v_i \sim \mathcal{D}_i}[p_{\ell^*}(v) > t] - \Pr_{v_i \sim \widehat{\mathcal{D}}_i}[p'_{\ell^*}(v) > t]\right| \leq 2\varepsilon$ for all $t \in [0, H]$, and

$$\left|\mathbb{E}_{v_i \sim \mathcal{D}_i}[p_{\ell^*}(v)] - \mathbb{E}_{v_i \sim \widehat{\mathcal{D}}_i}\left[p'_{\ell^*}(v)\right]\right| \leq H \cdot 2\varepsilon.$$

Combining statement (i) and (ii), we complete the proof. □

By Claim 1, it is clear that

$$\mathrm{OPT}\left(\widehat{\mathcal{D}}\right) = \mathrm{OPT}\left(\mathcal{D}^{(n)}\right) \geq \mathrm{OPT}\left(\mathcal{D}^{(0)}\right) - 3nH \cdot \varepsilon = \mathrm{OPT}\left(\mathcal{D}\right) - 3nH \cdot \varepsilon$$

□

Proof of Lemma 10: For every $i \in [n]$, we construct two extra distributions $\widetilde{\mathcal{D}}_i$ and $\underset{\sim}{\mathcal{D}}_i$ as follows. $\widetilde{\mathcal{D}}_i$ is supported on $[\varepsilon, H + \varepsilon]$, and its CDF is defined as $F_{\widetilde{\mathcal{D}}_i}(x) = F_{\mathcal{D}_i}(x - \varepsilon)$. $\underset{\sim}{\mathcal{D}}_i$ is supported on $[-\varepsilon, H - \varepsilon]$, and its CDF is defined as $F_{\underset{\sim}{\mathcal{D}}_i}(x) = F_{\mathcal{D}_i}(x + \varepsilon)$. In other words, $\widetilde{\mathcal{D}}_i$ is the distribution by shifting all values in $\mathcal{D}_i$ to the right by ε, and $\underset{\sim}{\mathcal{D}}_i$ is the distribution by shifting all values in $\mathcal{D}_i$ to the left by ε.

Claim 2. *Let M be any DSIC and IR mechanism for $\widetilde{\mathcal{D}} = \bigtimes_{i=1}^n \widetilde{\mathcal{D}}_i$, there exists a DSIC and IR mechanism M' for $\underset{\sim}{\mathcal{D}} = \bigtimes_{i=1}^n \underset{\sim}{\mathcal{D}}_i$ such that*

$$\textsc{Rev}(M', \underset{\sim}{\mathcal{D}}) \geq \textsc{Rev}(M, \widetilde{\mathcal{D}}) - 2\varepsilon.$$

Proof. Based on the construction of $\widetilde{\mathcal{D}}$ and $\underset{\sim}{\mathcal{D}}$, we can couple the two distributions so that whenever we draw a value profile $v = (v_1, \ldots, v_n)$ from $\widetilde{\mathcal{D}}$, we also draw a value profile $v - 2\varepsilon = (v_1 - 2\varepsilon, \ldots, v_n - 2\varepsilon)$ from $\underset{\sim}{\mathcal{D}}$. Given mechanism $M = (x, p)$, we construct mechanism M' as follows. For every bid profile v, we offer bidder i the item with probability $x_i(v + 2\varepsilon)$ and asks her to pay $p_i(v + 2\varepsilon) - 2\varepsilon \cdot x_i(v + 2\varepsilon)$. Why is M' a DSIC and IR mechanism? For any value profile v and any bidder i, her utility for reporting the true value is

$$(v_i + 2\varepsilon) \cdot x_i(v + 2\varepsilon) - p_i(v + 2\varepsilon),$$

and her utility for misreporting to v'_i is

$$(v_i + 2\varepsilon) \cdot x_i\left((v'_i, v_{-i}) + 2\varepsilon\right) - p_i((v'_i, v_{-i}) + 2\varepsilon).$$

Now consider a different scenario, where we run mechanism M and all the other bidders report $v_{-i} + 2\varepsilon$. The former is bidder i's utility in M when her true value is $v_i + 2\varepsilon$ and she reports truthfully. The latter is bidder i's utility in M when she lies and reports $v'_i + 2\varepsilon$. As M is a DSIC

and IR mechanism, $(v_i + 2\varepsilon) \cdot x_i(v + 2\varepsilon) - p_i(v + 2\varepsilon)$ is nonnegative and greater than $(v_i + 2\varepsilon) \cdot x_i\left((v_i', v_{-i}) + 2\varepsilon\right) - p_i((v_i', v_{-i}) + 2\varepsilon)$. Thus, M' is also a DSIC and IR mechanism. Since there is only one item for sale, $\sum_i x_i(v + 2\varepsilon) \le 1$. For every value profile v, the total payment in M' for this profile is at most 2ε smaller than the total payment in M for value profile $v + 2\varepsilon$. Therefore, $\textsc{Rev}(M', \underline{\mathcal{D}}) \ge \textsc{Rev}(M, \widetilde{\mathcal{D}}) - 2\varepsilon$. □

An easy corollary of Claim 2 is that

$$\mathrm{OPT}(\underline{\mathcal{D}}) \ge \mathrm{OPT}(\widetilde{\mathcal{D}}) - 2\varepsilon. \tag{1}$$

Next we will use this corollary and Theorem 5 to prove our claim. Note that $\left\|\underline{\mathcal{D}}_i - \underline{D}_i\right\|_K \le \varepsilon$ and $\left\|\widetilde{\mathcal{D}}_i - \overline{D}_i\right\|_K \le \varepsilon$ for all $i \in [n]$. Theorem 5 implies that

$$\mathrm{OPT}(\underline{\mathcal{D}}) \ge \mathrm{OPT}(\underline{\mathcal{D}}) - 3nH \cdot \varepsilon \tag{2}$$

and

$$\mathrm{OPT}(\widetilde{\mathcal{D}}) \ge \mathrm{OPT}(\overline{\mathcal{D}}) - 3n(H + \varepsilon) \cdot \varepsilon. \tag{3}$$

Chaining inequalities (2), (1), and (3), we have

$$\mathrm{OPT}(\underline{\mathcal{D}}) \ge \mathrm{OPT}(\overline{\mathcal{D}}) - (6nH + 3n\varepsilon + 2) \cdot \varepsilon.$$

□

E MISSING DETAILS FROM SECTION 3

E.1 Proof of Theorem 1

Proof of Theorem 1:

We first construct a mechanism M_2, and we show that M_2 is $(2m\mathcal{L}H\rho + \eta)$-BIC w.r.t. $\widehat{\mathcal{F}}$ and IR. We first define a mapping τ_i for every bidder i:

$$\tau_i(v_i) = \begin{cases} v_i, & \text{if } v_i \in \mathrm{supp}(\mathcal{F}_i) \\ \mathrm{argmax}_{z \in \mathrm{supp}(\mathcal{F}_i) \cup \perp} \mathbb{E}_{b_{-i} \sim \mathcal{F}_{-i}}\left[u_i(v_i, M_1(z, b_{-i}))\right], & \text{otherwise.} \end{cases} \tag{4}$$

Note that $\mathbb{E}_{b_{-i} \sim \mathcal{F}_{-i}}\left[u_i(v_i, M_1(\perp, b_{-i}))\right] = 0$. For any bid profile v, we use $\tau(v)$ to denote the vector $(\tau_1(v_1), \ldots, \tau_n(v_n))$. Let $x(\cdot)$ and $p(\cdot)$ be the allocation and payment rule for M_1. We now define M_2's allocation rule $x'(\cdot)$ and payment rule $p'(\cdot)$. For any bid profile v, $x'(v) = x(\tau(v))$. If $\tau_i(v_i) \ne v_i$ and $\tau_\ell(v_\ell) \ne \perp$ for all bidders $\ell \in [n]$, then

$$p_i'(v_i, v_{-i}) = v_i(x(\tau(v))) \cdot \frac{\mathbb{E}_{b_{-i} \sim \mathcal{F}_{-i}}\left[p_i(\tau_i(v_i), b_{-i})\right]}{\mathbb{E}_{b_{-i} \sim \mathcal{F}_{-i}}\left[v_i\left(x\left(\tau_i(v_i), b_{-i}\right)\right)\right]}.$$

Otherwise, $p_i'(v) = p_i(\tau(v))$.

An important property of $p'(\cdot)$ is that $\mathbb{E}_{b_{-i} \sim \mathcal{F}_{-i}}\left[p_i'(v_i, b_{-i})\right] = \mathbb{E}_{b_{-i} \sim \mathcal{F}_{-i}}\left[p_i(\tau_i(v_i), b_{-i})\right]$ for any v_i. We first argue that M_2 is IR.

*M_2 **is IR:**.* For any bidder i and any bid profile v, if any of $\tau_\ell(v_\ell) = \perp$ bidder i's utility is clearly 0. So we only need to consider the case where $\tau_\ell(v_\ell) \ne \perp$ for all $\ell \in [n]$.

- If $v_i = \tau_i(v_i)$, bidder i's utility is $v_i(x(v_i, \tau_{-i}(v_{-i}))) - p_i(v_i, \tau_{-i}(v_{-i})) = u_i\left(v_i, M_1(v_i, \tau_{-i}(v_{-i}))\right)$, which is non-negative as $v_i \in \mathrm{supp}(\mathcal{F}_i)$ and M_1 is IR.
- If $v_i \ne \tau_i(v_i)$, since $\tau_i(v_i) \ne \perp$ by our assumption,

$$\mathbb{E}_{b_{-i} \sim \mathcal{F}_{-i}}\left[v_i\left(x\left(\tau_i(v_i), b_{-i}\right)\right)\right] - \mathbb{E}_{b_{-i} \sim \mathcal{F}_{-i}}\left[p_i(\tau_i(v_i), b_{-i})\right] = \mathbb{E}_{b_{-i} \sim \mathcal{F}_{-i}}\left[u_i(v_i, M_1(\tau_i(v_i), b_{-i}))\right],$$

which is non-negative due to the definition of $\tau_i(\cdot)$. Equivalently, this means that

$$\frac{\mathbb{E}_{b_{-i}\sim\mathcal{F}_{-i}}\left[p_i(\tau_i(v_i),b_{-i})\right]}{\mathbb{E}_{b_{-i}\sim\mathcal{F}_{-i}}\left[v_i\left(x\left(\tau_i(v_i),b_{-i}\right)\right)\right]} \leq 1$$

and $p_i'(v_i,v_{-i}) \leq v_i(x(\tau(v))) = v_i(x'(v))$.

Next, we argue that M_2 is $(2m\mathcal{L}H\rho+\eta)$-BIC.

*M_2 **is** $(2m\mathcal{L}H\rho+\eta)$-**BIC:***. Consider any bidder i and any type v_i and t, we first bound the difference between $\mathbb{E}_{b_{-i}\sim\mathcal{F}_{-i}}\left[u_i(v_i,M_1(\tau_i(t),b_{-i}))\right]$ and $\mathbb{E}_{\hat{b}_{-i}\sim\widehat{\mathcal{F}}_{-i}}\left[u_i(v_i,M_2(t,\hat{b}_{-i}))\right]$. Note that

$$\mathbb{E}_{b_{-i}\sim\mathcal{F}_{-i}}\left[u_i(v_i,M_1(\tau_i(t),b_{-i}))\right] = \mathbb{E}_{b_{-i}\sim\mathcal{F}_{-i}}\left[u_i(v_i,M_2(t,b_{-i}))\right]. \tag{5}$$

This is because

$$x'(t,b_{-i}) = x(\tau_i(t),b_{-i})\ \forall b_{-i}\in\mathrm{supp}(\mathcal{F}_{-i})$$

and

$$\mathbb{E}_{b_{-i}\sim\mathcal{F}_{-i}}\left[p_i'(t,b_{-i})\right] = \mathbb{E}_{b_{-i}\sim\mathcal{F}_{-i}}\left[p_i(\tau_i(t),b_{-i})\right].$$

Since $\left\|\widehat{\mathcal{F}}_j-\mathcal{F}_j\right\|_{TV}=\varepsilon_j$, we can couple b_{-i} and $\hat{b}_{-i}$ so that

$$\Pr[b_{-i}\neq\hat{b}_{-i}]\leq\rho.$$

Clearly, when $b_{-i}=\hat{b}_{-i}$, $u_i(v_i,M_2(t,b_{-i}))=u_i(v_i,M_2(t,\hat{b}_{-i}))$. When $b_{-i}\neq\hat{b}_{-i}$,

$$\left|u_i(v_i,M_2(t,b_{-i}))-u_i(v_i,M_2(t,\hat{b}_{-i}))\right|\leq m\mathcal{L}H,$$

as $u_i(v_i,M_2(t,b'_{-i}))\in[0,m\mathcal{L}H]$ for any b'_{-i}. Hence, for any v_i and t

$$\left|\mathbb{E}_{b_{-i}\sim\mathcal{F}_{-i}}\left[u_i(v_i,M_2(t,b_{-i}))\right]-\mathbb{E}_{\hat{b}_{-i}\sim\widehat{\mathcal{F}}_{-i}}\left[u_i(v_i,M_2(t,\hat{b}_{-i}))\right]\right|\leq m\mathcal{L}H\rho. \tag{6}$$

Combining Inequality (5) and (6), we have the following inequality

$$\left|\mathbb{E}_{b_{-i}\sim\mathcal{F}_{-i}}\left[u_i(v_i,M_1(\tau_i(t),b_{-i}))\right]-\mathbb{E}_{\hat{b}_{-i}\sim\widehat{\mathcal{F}}_{-i}}\left[u_i(v_i,M_2(t,\hat{b}_{-i}))\right]\right|\leq m\mathcal{L}H\rho. \tag{7}$$

Suppose bidder i has type v_i, how much more utility can she get by misreporting? Since M_2 is IR, she clearly cannot gain by reporting a type t, whose corresponding $\tau_i(t)=\perp$. Next, we argue that she cannot gain much by reporting any other possible types either. If all other bidders report truthfully, bidder i's interim utility for reporting her true type

$$\begin{aligned}
\mathbb{E}_{\hat{b}_{-i}\sim\widehat{\mathcal{F}}_{-i}}\left[u_i(v_i,M_2(v_i,\hat{b}_{-i}))\right] &\geq \mathbb{E}_{b_{-i}\sim\mathcal{F}_{-i}}\left[u_i(v_i,M_1(\tau_i(v_i),b_{-i}))\right]-m\mathcal{L}H\rho\\
&\geq \max_{x\in\mathrm{supp}(\mathcal{F}_i)}\mathbb{E}_{b_{-i}\sim\mathcal{F}_{-i}}\left[u_i(v_i,M_1(x,b_{-i}))\right]-m\mathcal{L}H\rho-\eta\\
&\geq \max_{t:\tau_i(t)\neq\perp}\mathbb{E}_{b_{-i}\sim\mathcal{F}_{-i}}\left[u_i(v_i,M_1(\tau_i(t),b_{-i}))\right]-m\mathcal{L}H\rho-\eta\\
&\geq \max_{t:\tau_i(t)\neq\perp}\mathbb{E}_{\hat{b}_{-i}\sim\widehat{\mathcal{F}}_{-i}}\left[u_i(v_i,M_2(t,\hat{b}_{-i}))\right]-2m\mathcal{L}H\rho-\eta
\end{aligned}$$

The first inequality is due to Inequality (7). The second inequality is true because (a) if $v_i=\tau_i(v_i)$, then

$$\mathbb{E}_{b_{-i}\sim\mathcal{F}_{-i}}\left[u_i(v_i,M_1(\tau_i(v_i),b_{-i}))\right]\geq\max_{x\in\mathrm{supp}(\mathcal{F}_i)}\mathbb{E}_{b_{-i}\sim\mathcal{F}_{-i}}\left[u_i(v_i,M_1(x,b_{-i}))\right]-\eta$$

as M_1 is η-BIC; (b) if $v_i\notin\mathrm{supp}(\mathcal{F}_i)$, then by the definition of $\tau_i(v_i)$,

$$\mathbb{E}_{b_{-i}\sim\mathcal{F}_{-i}}\left[u_i(v_i,M_1(\tau_i(v_i),b_{-i}))\right]\geq\max_{x\in\mathrm{supp}(\mathcal{F}_i)}\mathbb{E}_{b_{-i}\sim\mathcal{F}_{-i}}\left[u_i(v_i,M_1(x,b_{-i}))\right].$$

The third inequality is because when $\tau_i(t) \neq \perp$ it must lie in $\text{supp}(\mathcal{F}_i)$. The last inequality is again due to Inequality (7).

Finally, we show that $\textsc{Rev}_T(M_2, \widehat{\mathcal{F}})$ is not much less than $\textsc{Rev}_T(M_1, \mathcal{F})$. Let $b \sim \mathcal{F}$ and $\hat{b} \sim \widehat{\mathcal{F}}$. There exists a coupling of b and $\hat{b}$ so that they are different w.p. less than ρ. When $b = \hat{b}$, $M_1(b) = M_2(\hat{b})$. When $b \neq \hat{b}$, the revenue in $M_1(b)$ is at most $nm\mathcal{L}H$ more than the revenue in $M_2(\hat{b})$, as both mechanisms are IR. Therefore,

$$\textsc{Rev}_T(M_2, \widehat{\mathcal{F}}) \geq \textsc{Rev}_T(M_1, \mathcal{F}) - nm\mathcal{L}H\rho.$$

□

E.2 Proof of Lemma 2

Proof of Lemma 2: According to Theorem 2, there exists a coupling γ of $\mathcal{F}$ and $\widehat{\mathcal{F}}$ so that

$$\Pr_{(x,y)\sim\gamma}[d(x,y) > \varepsilon] \leq \varepsilon.$$

Now we bound the probability that $r^{(\ell,\delta)}(x) \neq r^{(\ell,\delta)}(y)$, when (x, y) is drawn from γ, and ℓ is drawn from $U[0,\delta]^k$.

$$\begin{aligned}
&\Pr_{\ell\sim U[0,\delta]^k,(x,y)\sim\gamma}\left[r^{(\ell,\delta)}(x) \neq r^{(\ell,\delta)}(y)\right] \\
=&\Pr_{\ell\sim U[0,\delta]^k,(x,y)\sim\gamma}\left[r^{(\ell,\delta)}(x) \neq r^{(\ell,\delta)}(y) \ \wedge\ d(x,y) > \varepsilon\right] \\
&+ \Pr_{\ell\sim U[0,\delta]^k,(x,y)\sim\gamma}\left[r^{(\ell,\delta)}(x) \neq r^{(\ell,\delta)}(y) \ \wedge\ d(x,y) \leq \varepsilon\right] \\
\leq&\Pr_{(x,y)\sim\gamma}[d(x,y) > \varepsilon] + \Pr_{\ell\sim U[0,\delta]^k}\left[r^{(\ell,\delta)}(x) \neq r^{(\ell,\delta)}(y) \mid d(x,y) \leq \varepsilon\right] \cdot \Pr_{(x,y)\sim\gamma}[d(x,y) \leq \varepsilon] \\
\leq&\varepsilon + \Pr_{\ell\sim U[0,\delta]^k}\left[r^{(\ell,\delta)}(x) \neq r^{(\ell,\delta)}(y) \mid d(x,y) \leq \varepsilon\right]
\end{aligned}$$

Now, we bound the probability that $r^{(\ell,\delta)}(\cdot)$ rounds two points x and y to two different points when x and y are within distance ε. For any fixed x and y, we have the following.

$$\begin{aligned}
&\Pr_{\ell\sim U[0,\delta]^k}\left[r^{(\ell,\delta)}(x) \neq r^{(\ell,\delta)}(y)\right] \\
\leq&\sum_{i\in[k]} \Pr_{\ell_i\sim U[0,\delta]}\left[r_i^{(\ell,\delta)}(x) \neq r_i^{(\ell,\delta)}(y)\right] \\
\leq&\sum_{i\in[k]} \frac{|x_i - y_i|}{\delta} \\
=&\frac{d(x,y)}{\delta}
\end{aligned}$$

The first inequality follows from the union bound. Why is the second inequality true? If $|x_i - y_i| \geq \delta$, the inequality clearly holds, so we only need to consider the case where $|x_i - y_i| < \delta$. W.l.o.g. we assume $y_i \geq x_i$ and we consider the following two cases: (i) $\left\lfloor \frac{y_i}{\delta} \right\rfloor = \left\lfloor \frac{x_i}{\delta} \right\rfloor$ and (ii) $\left\lfloor \frac{y_i}{\delta} \right\rfloor = \left\lfloor \frac{x_i}{\delta} \right\rfloor + 1$. In case (i), $r_i^{(\ell,\delta)}(x) \neq r_i^{(\ell,\delta)}(y)$ if and only if $\ell \in \left[x_i - \left\lfloor \frac{x_i}{\delta} \right\rfloor \cdot \delta, y_i - \left\lfloor \frac{y_i}{\delta} \right\rfloor \cdot \delta\right]$. Since ℓ is drawn from the uniform distribution over $[0,\delta]$, this happens with probability exactly $\frac{y_i - x_i}{\delta}$. In case (ii), $r_i^{(\ell,\delta)}(x) \neq r_i^{(\ell,\delta)}(y)$ if and only if $\ell \in \left[x_i - \left\lfloor \frac{x_i}{\delta} \right\rfloor \cdot \delta, \delta\right] \cup \left[0, y_i - \left\lfloor \frac{y_i}{\delta} \right\rfloor \cdot \delta\right]$. This again happens with probability $\frac{y_i - x_i}{\delta}$. Therefore,

$$\Pr_{\ell\sim U[0,\delta]^k}\left[r^{(\ell,\delta)}(x) \neq r^{(\ell,\delta)}(y) \mid d(x,y) \leq \varepsilon\right] \leq \frac{\varepsilon}{\delta}$$

and

$$\Pr_{\ell\sim U[0,\delta]^k,(x,y)\sim\gamma}\left[r^{(\ell,\delta)}(x)\neq r^{(\ell,\delta)}(y)\right]\leq\left(1+\frac{1}{\delta}\right)\varepsilon. \tag{8}$$

Clearly, for any choice of ℓ, $\left\|\lfloor\mathcal{F}\rfloor_{\ell,\delta}-\left\lfloor\widehat{\mathcal{F}}\right\rfloor_{\ell,\delta}\right\|_{TV}\leq\Pr_{(x,y)\sim\gamma}\left[r^{(\ell,\delta)}(x)\neq r^{(\ell,\delta)}(y)\right]$. Combining this inequality with Inequality (8), we have

$$\begin{aligned}
&\mathbb{E}_{\ell\sim U[0,\delta]^k}\left[\left\|\lfloor\mathcal{F}\rfloor_{\ell,\delta}-\left\lfloor\widehat{\mathcal{F}}\right\rfloor_{\ell,\delta}\right\|_{TV}\right]\\
\leq&\mathbb{E}_{\ell\sim U[0,\delta]^k}\left[\Pr_{(x,y)\sim\gamma}\left[r^{(\ell,\delta)}(x)\neq r^{(\ell,\delta)}(y)\right]\right]\\
=&\Pr_{\ell\sim U[0,\delta]^k,(x,y)\sim\gamma}\left[r^{(\ell,\delta)}(x)\neq r^{(\ell,\delta)}(y)\right]\\
\leq&\left(1+\frac{1}{\delta}\right)\varepsilon
\end{aligned}$$

□

E.3 Missing Proofs from Section 3.3

Proof of Lemma 3: We first define $M_1^{(\ell)}$. If the bid profile $w\notin\mathrm{supp}(\underline{\mathcal{D}})$, the mechanism allocates nothing and charges no one. If the bid profile $w\in\mathrm{supp}(\underline{\mathcal{D}})$, for each bidder i sample w_i' independently from the distribution $\mathcal{D}_i\mid\bigtimes_{j=1}^m\beta(w_{ij})$, where $\beta(w_{ij})$ is defined to be $[0,\ell_j)$ if $w_{ij}=0$ and $[w_{ij},w_{ij}+\delta)$ otherwise. Bidder i receives allocation $x_{M,i}(w')$ and pays $(p_{M,i}(w')-m\mathcal{L}\delta)^+=\max\{0,p_{M,i}(w')-m\mathcal{L}\delta\}$. Note that, for any $i\in[n]$, if w_i is drawn from $\lfloor\mathcal{D}_i\rfloor_{\ell,\delta}$ then w_i' is drawn from $\mathcal{D}_i$. If all bidders bid truthfully in $M_1^{(\ell)}$, the revenue is at least $\textsc{Rev}(M,\mathcal{D})-nm\mathcal{L}\delta$. Next, we argue that $M_1^{(\ell)}$ is IR and ξ_1-BIC with $\xi_1=O(m\mathcal{L}\delta)$.

Note that for every bidder i and $w_i\in\mathrm{supp}(\lfloor\mathcal{D}_i\rfloor_{\ell,\delta})$ her interim utility in $M_1^{(\ell)}$ when all other bidders bid truthfully is at least $\mathbb{E}_{w_i'\sim\mathcal{D}_i\mid\bigtimes_{j=1}^m\beta(w_{ij}),w_{-i}'\sim\mathcal{D}_{-i}}\left[u_i(w_i,M(w_i',w_{-i}'))\right]$ due to the definition of $M_1^{(\ell)}$. Now consider every realization of w_i', it must hold that

$$\begin{aligned}
&\mathbb{E}_{w_{-i}'\sim\mathcal{D}_{-i}}\left[u_i(w_i,M(w_i',w_{-i}'))\right]\\
\geq&\mathbb{E}_{w_{-i}'\sim\mathcal{D}_{-i}}\left[u_i(w_i',M(w_i',w_{-i}'))\right]-m\mathcal{L}\delta\\
\geq&\max_{x\in\mathrm{supp}(\mathcal{D}_i)}\mathbb{E}_{w_{-i}'\sim\mathcal{D}_{-i}}\left[u_i(w_i',M(x,w_{-i}'))\right]-m\mathcal{L}\delta\\
\geq&\max_{x\in\mathrm{supp}(\mathcal{D}_i)}\mathbb{E}_{w_{-i}'\sim\mathcal{D}_{-i}}\left[u_i(w_i,M(x,w_{-i}'))\right]-2m\mathcal{L}\delta
\end{aligned}$$

The first and the last inequalities are both due to the fact that the valuation is $\mathcal{L}$-Lipschitz and $\left\|w_i-w_i'\right\|_1\leq m\delta$. The second inequality is because M is BIC w.r.t. $\mathcal{D}$. Hence, bidder i's interim utility in $M_1^{(\ell)}$ is at least $\max_{x\in\mathrm{supp}(\mathcal{D}_i)}\mathbb{E}_{w_{-i}'\sim\mathcal{D}_{-i}}\left[u_i(w_i,M(x,w_{-i}'))\right]-2m\mathcal{L}\delta$.

If bidder i misreports, her utility is no more than

$$\max_{x\in\mathrm{supp}(\mathcal{D}_i)}\mathbb{E}_{w_{-i}'\sim\mathcal{D}_{-i}}\left[u_i(w_i,M(x,w_{-i}'))\right]+m\mathcal{L}\delta,$$

due to the definition of $M_1^{(\ell)}$. Therefore, misreporting can increase bidder i's utility by at most $3m\mathcal{L}\delta$, and $M_1^{(\ell)}$ is $3m\mathcal{L}\delta$-BIC.

Next, we argue that $M_1^{(\ell)}$ is IR. If the $w_{-i} \notin \text{supp}(\underline{\mathcal{D}}_{-i})$, bidder i's utility is 0. So we focus on the case where $w_{-i} \in \text{supp}(\underline{\mathcal{D}}_{-i})$. We will show that for any realization of w_i' and w_{-i}', bidder i's utility is non-negative. If the payment is 0, the claim is trivially true. If the payment is nonzero, bidder i pays $p_{M,i}(w') - m\mathcal{L}\delta$ and has utility $u_i(w_i, M(w_i', w_{-i}'))) + m\mathcal{L}\delta$ which is at least $u_i(w_i', M(w_i', w_{-i}')))$, since the valuation is $\mathcal{L}$-Lipschitz and $\left\|w_i - w_i'\right\|_1 \le m\delta$. As M is IR, $u_i(w_i', M(w_i', w_{-i}'))) \ge 0$. Thus, bidder i's utility is non-negative and $M_1^{(\ell)}$ is IR.□

Proof of Lemma 4:
We first construct $\widehat{M}^{(\ell)}$. For any bid profile w, construct $w' = (r^{(\ell,\delta)}(w_1), \ldots, r^{(\ell,\delta)}(w_n))$, and run $M_2^{(\ell)}$ on w'. Bidder i receives allocation $x_{M_2^{(\ell)},i}(w')$ and pays $\max\{0, p_{M_2^{(\ell)},i}(w') - m\mathcal{L}\delta\}$. Note that if $w_i \sim \widehat{\mathcal{D}}_i$, then $w_i' \sim \left\lfloor \widehat{\mathcal{D}}_i \right\rfloor_{\ell,\delta}$. Assuming all other bidders bid truthfully and bidder i's type is w_i, bidder i's interim utility for bidding truthfully is

$$\begin{aligned}
\mathbb{E}_{b_{-i}\sim\widehat{\mathcal{D}}_{-i}}\left[u_i(w_i, \widehat{M}^{(\ell)}(w_i, b_{-i}))\right] &\ge \mathbb{E}_{b_{-i}'\sim\underline{\widehat{\mathcal{D}}}_{-i}}\left[u_i(w_i, M_2^{(\ell)}(w_i', b_{-i}'))\right] \\
&\ge \mathbb{E}_{b_{-i}\sim\underline{\widehat{\mathcal{D}}}_{-i}}\left[u_i(w_i', M_2^{(\ell)}(w_i', b_{-i}'))\right] - m\mathcal{L}\delta \\
&\ge \max_{x\in\text{supp}(\left\lfloor\widehat{\mathcal{D}}_i\right\rfloor_{\ell,\delta})} \mathbb{E}_{b_{-i}'\sim\underline{\widehat{\mathcal{D}}}_{-i}}\left[u_i(w_i', M_2^{(\ell)}(x, b_{-i}'))\right] - \xi_2 - m\mathcal{L}\delta \\
&\ge \max_{x\in\text{supp}(\left\lfloor\widehat{\mathcal{D}}_i\right\rfloor_{\ell,\delta})} \mathbb{E}_{b_{-i}'\sim\underline{\widehat{\mathcal{D}}}_{-i}}\left[u_i(w_i, M_2^{(\ell)}(x, b_{-i}'))\right] - \xi_2 - 2m\mathcal{L}\delta \\
&\ge \max_{y\in\text{supp}(\widehat{\mathcal{D}}_i)} \mathbb{E}_{b_{-i}\sim\widehat{\mathcal{D}}_{-i}}\left[u_i(w_i, \widehat{M}^{(\ell)}(y, b_{-i}))\right] - \xi_2 - 3m\mathcal{L}\delta
\end{aligned}$$

The first inequality and the last equality are due to the definition of $\widehat{M}^{(\ell)}$. The second and the fourth inequalities are due to the $\mathcal{L}$-Lipschitzness of the valuation function and $\left\|w_i - w_i'\right\|_1 \le m\delta$. The third inequality is because $M_2^{(\ell)}$ is a ξ_2-BIC mechanism w.r.t. $\underline{\widehat{\mathcal{D}}}$. By this chain of inequalities, we know that $\widehat{M}^{(\ell)}$ is a $(\xi_2 + 3m\mathcal{L}\delta)$-BIC mechanism w.r.t. $\widehat{\mathcal{D}}$.

Next, we argue that $\widehat{M}^{(\ell)}$ is also IR. Consider any bidder i and type profile w, $\widehat{M}^{(\ell)}(w)$ has the same allocation as $M_2^{(\ell)}(w')$. When bidder i's payment is 0, her utility is clearly non-negative. When bidder i's payment is $p_{M_2^{(\ell)},i}(w') - m\mathcal{L}\delta$, her utility is at least $u_i(w_i', M_2^{(\ell)}(w'))$ due to the $\mathcal{L}$-Lipschitzness of the valuation function and $\left\|w_i - w_i'\right\|_1 \le m\delta$. Since $M_2^{(\ell)}$ is IR, bidder i's utility in $\widehat{M}^{(\ell)}$ is also non-negative.

Finally, if all bidders bid truthfully in $\widehat{M}^{(\ell)}$ when their types are drawn from $\widehat{\mathcal{D}}$, its revenue under truthful bidding is

$$\textsc{Rev}_T\left(\widehat{M}^{(\ell)}, \widehat{\mathcal{D}}\right) \ge \textsc{Rev}_T\left(M_2^{(\ell)}, \underline{\widehat{\mathcal{D}}}\right) - nm\mathcal{L}\delta.$$

□

Proof of Theorem 3: First, sample ℓ uniformly from $[0,\delta]^m$, and construct $\lfloor\mathcal{D}_i\rfloor_{\ell,\delta}$ for all $i \in [n]$. According to Lemma 3, we can construct a mechanism $M_1^{(\ell)}$ based on M that is $\xi_1 = O(m\mathcal{L}\delta)$-BIC w.r.t. $\bigtimes_{i=1}^n \lfloor\mathcal{D}_i\rfloor_{\ell,\delta}$, IR, and has revenue $\textsc{Rev}_T\left(M_1^{(\ell)}, \bigtimes_{i=1}^n \lfloor\mathcal{D}_i\rfloor_{\ell,\delta}\right) \ge \textsc{Rev}(M, \mathcal{D}) - nm\mathcal{L}\delta$.

Next, we transform $M_1^{(\ell)}$ to $M_2^{(\ell)}$ using Lemma 1. We use $\varepsilon_i^{(\ell)}$ to denote $\left\|\lfloor\mathcal{D}_i\rfloor_{\ell,\delta} - \left\lfloor\widehat{\mathcal{D}}_i\right\rfloor_{\ell,\delta}\right\|_{TV}$ for our sample ℓ and every $i \in [n]$, and $\rho^{(\ell)}$ to denote $\sum_{i\in[n]} \varepsilon_i^{(\ell)}$. For every realization of ℓ, $M_2^{(\ell)}$ is

$\xi_2 = \left(2m\mathcal{L}H\rho^{(\ell)} + \xi_1\right)$-BIC w.r.t. $\bigtimes_{i=1}^n \left\lfloor \widehat{\mathcal{D}}_i \right\rfloor_{\ell,\delta}$ and IR. Its revenue under truthful bidding satisfies

$$\textsc{Rev}_T\left(M_2^{(\ell)}, \bigtimes_{i=1}^n \left\lfloor \widehat{\mathcal{D}}_i \right\rfloor_{\ell,\delta}\right) \geq \textsc{Rev}_T\left(M_1^{(\ell)}, \bigtimes_{i=1}^n \lfloor \mathcal{D}_i \rfloor_{\ell,\delta}\right) - nm\mathcal{L}H\rho^{(\ell)}.$$

Lemma 4 shows that we can construct $\widehat{M}^{(\ell)}$ using $M_2^{(\ell)}$, such that $\widehat{M}^{(\ell)}$ is a $(\xi_2 + 3m\mathcal{L}\delta)$-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR mechanism with revenue

$$\textsc{Rev}_T\left(\widehat{M}^{(\ell)}, \widehat{\mathcal{D}}\right) \geq \textsc{Rev}_T\left(M_2^{(\ell)}, \bigtimes_{i=1}^n \left\lfloor \widehat{\mathcal{D}}_i \right\rfloor_{\ell,\delta}\right) - nm\mathcal{L}\delta.$$

Since $\widehat{M}^{(\ell)}$ is $O(m\mathcal{L}\delta + m\mathcal{L}H\rho^{(\ell)})$-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR for every realization of ℓ, our mechanism $\widehat{M}$ is clearly $O\left(m\mathcal{L}\delta + m\mathcal{L}H \cdot \mathbb{E}_{\ell\sim U[0,\delta]^m}\left[\rho^{(\ell)}\right]\right)$-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR. Moreover, its expected revenue under truthful bidding satisfies

$$\textsc{Rev}_T\left(\widehat{M}, \widehat{\mathcal{D}}\right) \geq \textsc{Rev}\left(M, \mathcal{D}\right) - O\left(nm\mathcal{L}\delta + nm\mathcal{L}H \cdot \mathbb{E}_{\ell\sim U[0,\delta]^m}\left[\rho^{(\ell)}\right]\right).$$

According to Lemma 2,

$$\mathbb{E}_{\ell\sim U[0,\delta]^m}\left[\rho^{(\ell)}\right] \leq n\left(1 + \frac{1}{\delta}\right)\varepsilon.$$

We choose δ to be $\sqrt{nH\varepsilon}$, and $\widehat{M}$ becomes κ-BIC w.r.t. $\widehat{\mathcal{D}}$, where $\kappa = O\left(nm\mathcal{L}H\varepsilon + m\mathcal{L}\sqrt{nH\varepsilon}\right)$, and IR. Furthermore,

$$\textsc{Rev}_T\left(\widehat{M}, \widehat{\mathcal{D}}\right) \geq \textsc{Rev}\left(M, \mathcal{D}\right) - O\left(n\kappa\right).$$

□

E.4 Lipschitz Continuity of the Optimal Revenue in Multi-item Auctions

Using Theorem 3, we can easily prove that the optimal BIC revenue w.r.t. $\mathcal{D}$ and the optimal BIC revenue w.r.t. $\widehat{\mathcal{D}}$ are close as long as $\mathcal{D}_i$ and $\widehat{\mathcal{D}}_i$ are close in either the total variation distance or the Prokhorov distance for all $i \in [n]$.

Theorem 6 (Lipschitz Continuity of the Optimal Revenue). *Consider the general mechanism design setting of Section 2. Recall that $\mathcal{L}$ is the Lipschitz constant of the valuations. For any distributions $\mathcal{D} = \bigtimes_{i=1}^n \mathcal{D}_i$ and $\widehat{\mathcal{D}} = \bigtimes_{i=1}^n \widehat{\mathcal{D}}_i$, where $\mathcal{D}_i$ and $\widehat{\mathcal{D}}_i$ are supported on $[0, H]^m$ for every $i \in [n]$*

- *If $\left\|\mathcal{D}_i - \widehat{\mathcal{D}}_i\right\|_{TV} \leq \varepsilon$ for all $i \in [n]$, then*

$$\left|OPT(\mathcal{D}) - OPT(\widehat{\mathcal{D}})\right| \leq O\left(nm\mathcal{L}H\left(n\varepsilon + \sqrt{n\varepsilon}\right)\right);$$

- *if $\left\|\mathcal{D}_i - \widehat{\mathcal{D}}_i\right\|_{P} \leq \varepsilon$ for all $i \in [n]$, then*

$$\left|OPT(\mathcal{D}) - OPT(\widehat{\mathcal{D}})\right| \leq O\left(n\kappa + n\sqrt{m\mathcal{L}H\kappa}\right),$$

where $\kappa = O\left(nm\mathcal{L}H\varepsilon + m\mathcal{L}\sqrt{nH\varepsilon}\right)$.

Proof of Theorem 6: Let M^* be the optimal BIC mechanism for $\mathcal{D}$. We first prove the Prokorov case. According to Theorem 3, there exists a mechanism $\widehat{M}^*$ such that it is κ-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR. Moreover,

$$\textsc{Rev}_T(\widehat{M}^*, \widehat{\mathcal{D}}) \geq \textsc{Rev}(M^*, \mathcal{D}) - O(n\kappa).$$

By Lemma 1, $\textsc{Rev}_T\left(\widehat{M}^*, \widehat{\mathcal{D}}\right) \le \text{OPT}\left(\widehat{\mathcal{D}}\right) + 2n\sqrt{m\mathcal{L}H\kappa}$. Combining the two inequalities, we have

$$\text{OPT}\left(\widehat{\mathcal{D}}\right) \ge \text{OPT}(\mathcal{D}) - O\left(n\kappa + n\sqrt{m\mathcal{L}H\kappa}\right).$$

By symmetry, we can also argue that

$$\text{OPT}(\mathcal{D}) \ge \text{OPT}\left(\widehat{\mathcal{D}}\right) - O\left(n\kappa + n\sqrt{m\mathcal{L}H\kappa}\right).$$

In the TV case, $\textsc{Rev}_T\left(\widehat{M}^*, \widehat{\mathcal{D}}\right) \ge \textsc{Rev}(M^*, \mathcal{D}) - O(n^2 m\mathcal{L}H\varepsilon)$. Since $\widehat{M}^*$ is $O(mn\mathcal{L}H\varepsilon)$-BIC, $\text{OPT}\left(\widehat{\mathcal{D}}\right) \ge \textsc{Rev}_T\left(\widehat{M}^*, \widehat{\mathcal{D}}\right) - O(nm\mathcal{L}H\sqrt{n\varepsilon})$ due to Lemma 1. By symmetry and the inequalities above, we have $\left|\text{OPT}(\mathcal{D}) - \text{OPT}\left(\widehat{\mathcal{D}}\right)\right| \le O\left(nm\mathcal{L}H(n\varepsilon + \sqrt{n\varepsilon})\right)$. □

E.5 Approximation Preserving Transformation

Theorem 7 (Approximation Preserving Transformation). *Consider the general mechanism design setting of Section 2. Recall that $\mathcal{L}$ is the Lipschitz constant of the valuations. Given $\mathcal{D} = \bigtimes_{i=1}^n \mathcal{D}_i$, where $\mathcal{D}_i$ is a m-dimensional distribution supported on $[0, H]^m$ for all $i \in [n]$, and a BIC w.r.t. $\mathcal{D}$ and IR mechanism M. We use $\widehat{\mathcal{D}} = \bigtimes_{i=1}^n \widehat{\mathcal{D}}_i$ to denote the true but unknown type distribution, and $\widehat{\mathcal{D}}_i$ is supported on $[0, H]^m$ for all $i \in [n]$.*

If $\left\|\mathcal{D}_i - \widehat{\mathcal{D}}_i\right\|_{TV} \le \varepsilon$ for all $i \in [n]$, we can construct a mechanism $\widehat{M}$, in a way that is completely oblivious to the true distribution $\widehat{\mathcal{D}}$, such that

(1) $\widehat{M}$ is η-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR, where $\eta = O(nm\mathcal{L}H\varepsilon)$;
(2) if M is a c-approximation to the optimal BIC revenue for $\mathcal{D}$, then

$$\textit{Rev}_T\left(\widehat{M}, \widehat{\mathcal{D}}\right) \ge c \cdot \textit{OPT}_\eta\left(\widehat{\mathcal{D}}\right) - O\left(nm\mathcal{L}H\left(n\varepsilon + \sqrt{n\varepsilon}\right)\right).$$

If $\left\|\mathcal{D}_i - \widehat{\mathcal{D}}_i\right\|_P \le \varepsilon$ for all $i \in [n]$, we can again construct a mechanism $\widehat{M}$, in a way that is completely oblivious to the true distribution $\widehat{\mathcal{D}}$, such that

(1) $\widehat{M}$ is κ-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR, where $\kappa = O\left(nm\mathcal{L}H\varepsilon + m\mathcal{L}\sqrt{nH\varepsilon}\right)$;
(2) if M is a c-approximation to the optimal BIC revenue for $\mathcal{D}$, then $\widehat{M}$ is almost a c-approximation to the optimal κ-BIC revenue for $\widehat{\mathcal{D}}$, that is,

$$\textit{Rev}_T\left(\widehat{M}, \widehat{\mathcal{D}}\right) \ge c \cdot \textit{OPT}_\kappa\left(\widehat{\mathcal{D}}\right) - O\left(n\kappa + n\sqrt{m\mathcal{L}H\kappa}\right).$$

Proof of Theorem 7: For the TV case, by Theorem 1, we can construct a η-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR mechanism $\widehat{M}$ such that $\textsc{Rev}_T\left(\widehat{M}, \widehat{\mathcal{D}}\right) \ge \textsc{Rev}(M, \mathcal{D}) - O\left(n^2 m\mathcal{L}H\varepsilon\right) \ge c \cdot \text{OPT}(\mathcal{D}) - O\left(n^2 m\mathcal{L}H\varepsilon\right)$. By Theorem 6, $\text{OPT}(\mathcal{D})$ is at least $\text{OPT}(\widehat{\mathcal{D}}) - O\left(nm\mathcal{L}H(n\varepsilon + \sqrt{n\varepsilon})\right)$. Finally, $\text{OPT}(\widehat{\mathcal{D}}) \ge \text{OPT}_\eta(\widehat{\mathcal{D}}) - 2n\sqrt{m\mathcal{L}H\eta}$ due to Lemma 1, so

$$\textsc{Rev}_T\left(\widehat{M}, \widehat{\mathcal{D}}\right) \ge c \cdot \text{OPT}_\eta\left(\widehat{\mathcal{D}}\right) - O\left(nm\mathcal{L}H(n\varepsilon + \sqrt{n\varepsilon})\right).$$

For the Prokhorov case, according to Theorem 3, we can construct a κ-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR mechanism $\widehat{M}$ such that $\textsc{Rev}_T\left(\widehat{M}, \widehat{\mathcal{D}}\right) \ge \textsc{Rev}(M, \mathcal{D}) - O(n\kappa) \ge c \cdot \text{OPT}(\mathcal{D}) - O(n\kappa)$. By Theorem 6 and Lemma 1, $\text{OPT}(\mathcal{D}) \ge \text{OPT}\left(\widehat{\mathcal{D}}\right) - O\left(n\kappa + n\sqrt{m\mathcal{L}H\kappa}\right) \ge \text{OPT}_\kappa(\widehat{\mathcal{D}}) - O\left(n\kappa + n\sqrt{m\mathcal{L}H\kappa}\right)$

Chaining all the inequalities above, we have

$$\textsc{Rev}_T\left(\widehat{M},\widehat{\mathcal{D}}\right) \geq c\cdot \mathrm{OPT}_\kappa(\widehat{\mathcal{D}}) - O\left(n\kappa + n\sqrt{m\mathcal{L}H\kappa}\right).$$

□

If there is a single bidder, we can strengthen Theorem 7 and make constructed mechanism $\widehat{M}$ exactly IC with essentially the same guarantees.

Theorem 8 (Single-Bidder Approximation Preserving Transformation). *Consider the general mechanism design setting of Section 2. Recall that $\mathcal{L}$ is the Lipschitz constant of the valuations. Given a m-dimensional distribution $\mathcal{D}$ supported on $[0,H]^m$, and a IC and IR mechanism M. We use $\widehat{\mathcal{D}}$ to denote the true but unknown type distribution, and $\widehat{\mathcal{D}}$ is also supported on $[0,H]^m$.*

- *If $\left\|\mathcal{D}-\widehat{\mathcal{D}}\right\|_{TV} \leq \varepsilon$, we can construct an IC and IR mechanism $\widehat{M}$, in a way that is completely oblivious to the true distribution $\widehat{\mathcal{D}}$, such that if M is a c-approximation to the optimal BIC revenue for $\mathcal{D}$, then*

$$\textsc{Rev}\left(\widehat{M},\widehat{\mathcal{D}}\right) \geq c\cdot\left(1-O\left(\sqrt{m\mathcal{L}H\varepsilon}\right)\right)\cdot \mathit{OPT}\left(\widehat{\mathcal{D}}\right) - O\left(\left(m\mathcal{L}H+\sqrt{m\mathcal{L}H}\right)\cdot\sqrt{\varepsilon}\right).$$

- *If $\left\|\mathcal{D}-\widehat{\mathcal{D}}\right\|_{P} \leq \varepsilon$, we can again construct an IC and IR mechanism $\widehat{M}$, in a way that is completely oblivious to the true distribution $\widehat{\mathcal{D}}$, such that if M is a c-approximation to the optimal BIC revenue for $\mathcal{D}$,*

$$\textsc{Rev}\left(\widehat{M},\widehat{\mathcal{D}}\right) \geq c\cdot\left(1-\sqrt{\kappa}\right)\cdot \mathit{OPT}\left(\widehat{\mathcal{D}}\right) - O\left(\kappa+\left(\sqrt{m\mathcal{L}H}+1\right)\cdot\sqrt{\kappa}\right),$$

where $\kappa = O\left(m\mathcal{L}H\varepsilon + m\mathcal{L}\sqrt{H\varepsilon}\right)$.

Proof of Theorem 8: We only sketch the proof here. Let M' be the mechanism constructed using Theorem 7, and we construct another mechanism $\widehat{M}$ by modifying M' using Lemma 5. Clearly, $\widehat{M}$ is IC and IR. It is not hard to verify that $\textsc{Rev}\left(\widehat{M},\widehat{\mathcal{D}}\right)$ satisfies the guarantees in the statement by combining the revenue guarantees for $\textsc{Rev}_T\left(M',\widehat{\mathcal{D}}\right)$ as provided by Theorem 7 and the relation between $\textsc{Rev}\left(\widehat{M},\widehat{\mathcal{D}}\right)$ and $\textsc{Rev}_T\left(M',\widehat{\mathcal{D}}\right)$ as stated in Lemma 5. □

F LEARNING MULTI-ITEM AUCTIONS UNDER ITEM INDEPENDENCE

In this section, we show how to derive one of the state-of-the-art learnability results for learning multi-item auctions via our robustness results. We consider the case where every bidder's type distribution is a m-dimensional product distribution. We will show that a generalization of the main result by Gonczarowski and Weinberg [36] follows easily from our robustness result. The main idea is that it suffices to learn the distribution $\mathcal{F}_i$ within small Prokhorov distance for every bidder i, and it only requires polynomial many samples when each $\mathcal{F}_i$ is a product distribution.

Theorem 9. *Consider the general mechanism design setting of Section 2. Recall that $\mathcal{L}$ is the Lipschitz constant of the valuations. For every $\varepsilon,\delta>0$, and for every $\eta \leq \mathrm{poly}(n,m,\mathcal{L},H,\varepsilon)$, we can learn a distribution $\mathcal{D}=\bigtimes_{i\in[n],j\in[m]}\mathcal{D}_{ij}$ with $\mathrm{poly}\left(n,m,\mathcal{L},H,1/\varepsilon,1/\eta,\log(1/\delta)\right)$ samples from $\widehat{\mathcal{D}}=\bigtimes_{i\in[n],j\in[m]}\widehat{\mathcal{D}}_{ij}$, such that, with probability $1-\delta$, we can transform any BIC w.r.t. $\mathcal{D}$, IR, and c-approximation mechanism M to an η-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR mechanism $\widehat{M}$, whose revenue under truthful bidding satisfies*

$$\textsc{Rev}_T\left(\widehat{M},\widehat{\mathcal{D}}\right) \geq c\cdot \mathit{OPT}_\eta\left(\widehat{\mathcal{D}}\right) - \varepsilon.$$

If $n = 1$, the mechanism $\widehat{M}$ will be IR and IC, and

$$\textsc{Rev}\left(\widehat{M}, \widehat{\mathcal{D}}\right) \geq c \cdot (1 - \sqrt{\eta}) \cdot OPT\left(\widehat{\mathcal{D}}\right) - \varepsilon - \sqrt{\eta}.$$

In particular, Gonczarowski and Weinberg [36] proved the $c = 1$ case, and our result applies to any $c \in (0, 1]$. The proof is given in Appendix I. We provide a proof sketch here. We first prove Lemma 12, which shows that polynomially many samples suffice to learn a distribution $\mathcal{D}$ that is close to $\widehat{\mathcal{D}}$ in Prokhorov distance. Now the statement simply follows from Theorem 7.

G MISSING PROOFS FROM SECTION F

We first show that for any product distribution $\mathcal{F}$, we can learn the rounded distribution of $\mathcal{F}$ within small TV distance with polynomially many samples.

Lemma 11. *Let $\mathcal{F} = \bigtimes_{j=1}^{m} \mathcal{F}_j$, where $\mathcal{F}_j$ is an arbitrary distribution supported on $[0, H]$ for every $j \in [m]$. Given $N = O\left(\frac{m^3 H}{\eta^3} \cdot (\log 1/\delta + \log m)\right)$ samples, we can learn a product distribution $\widehat{\mathcal{F}} = \bigtimes_{j=1}^{m} \widehat{\mathcal{F}}_j$ such that*

$$\left\|\mathcal{F} - \widehat{\mathcal{F}}\right\|_P \leq \eta$$

with probability at least $1 - \delta$.

Proof. We denote the samples as $s^1, \ldots, s^N$. Round each sample to multiples of $\eta' = \eta/m$. More specifically, let $\hat{s}^i = \left(\left\lfloor s_1^i/\eta' \right\rfloor \cdot \eta', \ldots, \left\lfloor s_m^i/\eta' \right\rfloor \cdot \eta'\right)$ for every sample $i \in [N]$. Let $\widehat{\mathcal{F}}_j$ be the uniform distribution over $\hat{s}_j^1, \ldots, \hat{s}_j^N$. Let $\overline{\mathcal{F}}_j = \left\lfloor \mathcal{F}_j \right\rfloor_{0,\eta'}$. Note that $\widehat{\mathcal{F}}_j$ is the empirical distribution of N samples from $\overline{\mathcal{F}}_j$. As $\left|\text{supp}(\overline{\mathcal{F}}_j)\right| = \left\lfloor \frac{H}{\eta'} \right\rfloor = \frac{mH}{\eta}$, with $N = O\left(\frac{|\text{supp}(\overline{\mathcal{F}}_j)|}{\eta'^2} \cdot (\log 1/\delta + \log m)\right)$ samples, the empirical distribution $\widehat{\mathcal{F}}_j$ should satisfy $\left\|\widehat{\mathcal{F}}_j - \overline{\mathcal{F}}_j\right\|_{TV} \leq \eta'$ with probability at least $1 - \delta/m$. By the union bound $\left\|\widehat{\mathcal{F}}_j - \overline{\mathcal{F}}_j\right\|_{TV} \leq \eta'$ for all $j \in [m]$ with probability at least $1 - \delta$, which implies $\left\|\widehat{\mathcal{F}} - \overline{\mathcal{F}}\right\|_{TV} \leq \eta$ with probability at least $1 - \delta$. Observe that $\overline{\mathcal{F}}$ and $\mathcal{F}$ can be coupled so that the two samples are always within η in ℓ_1 distance. When $\left\|\widehat{\mathcal{F}} - \overline{\mathcal{F}}\right\|_{TV} \leq \eta$, consider the coupling between $\widehat{\mathcal{F}}$ and $\mathcal{F}$ by composing the optimal coupling between $\widehat{\mathcal{F}}$ and $\overline{\mathcal{F}}$ and the coupling between $\overline{\mathcal{F}}$ and $\mathcal{F}$. Clearly, the two samples from $\widehat{\mathcal{F}}$ and $\mathcal{F}$ are within ℓ_1 distance η with probability at least $1 - \eta$. Due to Theorem 2, the existence of this coupling implies that $\left\|\widehat{\mathcal{F}} - \mathcal{F}\right\|_P \leq \eta$. □

Proof of Theorem 9: We only consider the case, where $\eta \leq \alpha \cdot \min\left\{\frac{\varepsilon}{n}, \frac{\varepsilon^2}{n^2 m \mathcal{L} H}\right\}$. α is an absolute constant and we will specify its choice in the end of the proof.

In light of Lemma 12, we take $N = O\left(\frac{m^3 H}{\sigma^3} \cdot (\log \frac{n}{\delta} + \log m)\right)$ from $\widehat{\mathcal{D}}$ and learn a distribution $\mathcal{D}$ so that, with probability at least $1 - \delta$, $\left\|\mathcal{D}_i - \widehat{\mathcal{D}}_i\right\|_P \leq \sigma$ for all $i \in [n]$. According to Theorem 7, we can transform M into mechanism $\widehat{M}$ that is $O\left(nm\mathcal{L}H\sigma + m\mathcal{L}\sqrt{nH\sigma}\right)$-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR. Choose σ in a way so that $\widehat{M}$ is η-BIC w.r.t. $\widehat{\mathcal{D}}$. Moreover, $\widehat{M}$'s revenue under truthful bidding satisfies

$$\textsc{Rev}_T\left(\widehat{M}, \widehat{\mathcal{D}}\right) \geq c \cdot \text{OPT}_\eta\left(\widehat{\mathcal{D}}\right) - O\left(n\eta + n\sqrt{m\mathcal{L}H\eta}\right).$$

If we choose α to be sufficiently small, then

$$\textsc{Rev}_T\left(\widehat{M}, \widehat{\mathcal{D}}\right) \geq c \cdot \mathrm{OPT}_\eta\left(\widehat{\mathcal{D}}\right) - \varepsilon.$$

When there is only a single-bidder, we can apply Lemma 5 to transform $\widehat{M}$ to an IC and IR mechanism, whose revenue satisfies the guarantee in the statement.□

H OPTIMAL MECHANISM DESIGN UNDER STRUCTURAL ITEM DEPENDENCE

In this section, we go beyond the standard assumption of item-independence, which has been employed in most of prior literature, to consider settings where, as is commonly the case in practice, item values are correlated. Of course, once we embark onto a study of correlated distributions, we should not go all the way to full generality, since exponential sample-size lower bounds are known, even for learning approximately optimal mechanisms in single-bidder unit-demand settings [31]. Besides those sample complexity lower bounds, however, fully general distributions are also not very natural. In practice, high-dimensional distributions are not arbitrary, but have structure, which allows us to perform inference on them and learn them more efficiently. We thus propose the study of optimal mechanism design under the assumption that item values are jointly sampled from a high-dimensional distribution with structure.

There are many probabilistic frameworks that allow modeling structure in a high-dimensional distribution. In this work we consider one of the most prominent ones: *graphical models*, and in particular consider the two most common types of graphical models: *Markov Random Fields* and *Bayesian Networks*.

Definition 10. *A* Markov Random Field (MRF) *is a distribution defined by a hypergraph $G = (V, E)$. Associated with every vertex $v \in V$ is a random variable X_v taking values in some alphabet Σ, as well as a potential function $\psi_v : \Sigma \to [0, 1]$. Associated with every hyperedge $e \subseteq V$ is a potential function $\psi_e : \Sigma^e \to [0, 1]$. In terms of these potentials, we define a probability distribution p associating to each vector $x \in \Sigma^V$ probability $p(x)$ satisfying:*

$$p(x) = \frac{1}{Z} \prod_{v \in V} \psi_v(x_v) \prod_{e \in E} \psi_e(x_e), \tag{9}$$

where for a set of nodes e and a vector x we denote by x_e the restriction of x to the nodes in e, and Z is a normalization constant making sure that p, as defined above, is a distribution. In the degenerate case where the products on the RHS of (9) *always evaluate to* 0*, we assume that p is the uniform distribution over Σ^V. In that case, we get the same distribution by assuming that all potential functions are identically* 1*. Hence, we can in fact assume that the products on the RHS of* (9) *cannot always evaluate to* 0*.*

Definition 11. *A* Bayesian network*, or* Bayesnet*, specifies a probability distribution in terms of a directed acyclic graph G whose nodes V are random variables taking values in some alphabet Σ. To describe the probability distribution, one specifies conditional probabilities $p_{X_v|X_{\Pi_v}}(x_v|x_{\Pi_v})$, for all vertices v in G, and configurations $x_v \in \Sigma$ and $x_{\Pi_v} \in \Sigma^{\Pi_v}$, where Π_v represents the set of parents of v in G, taken to be $\emptyset$ if v has no parents. In terms of these conditional probabilities, a probability distribution over Σ^V is defined as follows:*

$$p(x) = \prod_v p_{X_v|X_{\Pi_v}}(x_v|x_{\Pi_v}), \text{ for all } x \in \Sigma^V.$$

It is important to note that both MRFs and Bayesnets allow the study of distributions in their *full generality*, as long as the graphs on which they are defined are sufficiently dense. In particular,

the graph (hypergraph and DAG respectively) underlying these models captures conditional independence relations, and is sufficiently flexible to capture the structure of intricate dependencies in the data. As such these models have found myriad applications; see e.g. [43, 44, 50, 51] and their references. A common way to control the expressiveness of MRFs and Bayesnets is to vary the maximum size of hyperedges in an MRF and indegree in a Bayesnet. Our sample complexity results presented below will be parametrized according to this measure of complexity in the distributions.

In our results, presented below, we exploit our modular framework to disentangle the identification of good mechanisms for these settings from the intricacies of learning a good model of the underlying distribution from samples. In particular, we are able to pair our mechanism design framework presented in earlier sections with learning results for MRFs and Bayesnets to characterize the sample complexity of learning good mechanisms when the item distributions are MRFs and Bayesnets. Below, we first present our results on the sample complexity of learning good mechanisms in these settings, followed by the learning results for MRFs and Bayesnets that these are modularly dependent on.

H.1 Learning Multi-item Auctions under Structural Item Dependence

In this section, we state our results for learning multi-item auctions when each bidder's values correlated. In particular, we consider two cases: (i) every bidder's type is sampled from an MRF, or (ii) every bidder's type is sampled from a Bayesnet. Our results can accommodate latent variables, that is, some of the variables/nodes of the MRF or Bayesnet are not observable in the samples. We show that the sample complexity for learning an η-BIC and IR mechanism, whose revenue is at most ε less than the optimal revenue achievable by any η-BIC and IR mechanisms, is polynomial in the size of the problem and scales gracefully with the parameters of the graphical models that generate the type distributions. If there is only a single bidder, the mechanism we learn will be exactly IC rather than approximately IC. We derive the sample complexity by combining our robustness result (Theorem 7) with learnability results for MRFs and Bayesnets (Theorem 12 and 13).

Theorem 10 (Optimal Mechanism Design under MRF Item Distributions). *Consider the general mechanism design setting of Section 2. Recall that $\mathcal{L}$ is the Lipschitz constant of the valuations. Let $\widehat{\mathcal{D}} = \bigtimes_{i\in[n]} \widehat{\mathcal{D}}_i$, where each $\widehat{\mathcal{D}}_i$ is a m-dimensional distribution generated by an MRF p_i, as in Definition 10, defined on a graph with $N_i \geq m$ nodes, hyper-edges of size at most d, and $supp(\widehat{\mathcal{D}}_i) \subseteq \Sigma^m \subseteq [0, H]^m$. When $N_i > m$, we say $\widehat{\mathcal{D}}_i$ is generated by an MRF with $N_i - m$ latent variables. We use N to denote $\max_{i\in[n]}\{N_i\}$.*

For every $\varepsilon, \delta > 0$, and $\eta \leq \text{poly}(n, m, \mathcal{L}, H, \varepsilon)$, we can learn, with probability at least $1 - \delta$, an η-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR mechanism $\widehat{M}$, whose revenue under truthful bidding is at most ε smaller than the optimal revenue achievable by any η-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR mechanism, given

- $\frac{poly\left(n, N^d, |\Sigma|^d, \mathcal{L}, H, 1/\eta, \log(1/\delta)\right)}{\varepsilon^4}$ *samples if* ***the alphabet Σ is finite****; when the graph on which p_i is defined is known for each bidder i, then $\frac{poly\left(n, N, \kappa, |\Sigma|^d, \mathcal{L}, H, 1/\eta, \log(1/\delta)\right)}{\varepsilon^4}$-many samples suffice, where κ is an upper bound on the number of edges in all the graphs;*
- $\text{poly}\left(n, N^{d^2}, \left(\frac{H}{\varepsilon}\right)^d, C^d, \mathcal{L}, 1/\eta, \log(1/\delta)\right)$ *samples if* ***the alphabet $\Sigma = [0, H]$****, and the log potentials $\phi_v^{p_i}(\cdot) \equiv \log\left(\psi_v^{p_i}(\cdot)\right)$ and $\phi_e^{p_i}(\cdot) \equiv \log\left(\psi_e^{p_i}(\cdot)\right)$ for every node v and every edge e are C-Lipschitz w.r.t. the ℓ_1-norm, for every bidder i; when the graph on which p_i is defined is known for each bidder i, then poly $\left(n, N, \kappa^d, \left(\frac{H}{\varepsilon}\right)^d, C^d, \mathcal{L}, 1/\eta, \log(1/\delta)\right)$-many samples suffice, where κ is an upper bound on the number of edges in all the graphs.*

If $n = 1$, the mechanism $\widehat{M}$ will be IR and IC, and

$$\textsc{Rev}\left(\widehat{M}, \widehat{\mathcal{D}}\right) \geq \left(1 - \sqrt{\eta}\right) \cdot OPT\left(\widehat{\mathcal{D}}\right) - \varepsilon - \sqrt{\eta}.$$

Theorem 11 (Optimal Mechanism Design under Bayesnet Item Distributions). *Consider the general mechanism design setting of Section 2. Recall that $\mathcal{L}$ is the Lipschitz constant of the valuations. Let $\widehat{\mathcal{D}} = \bigtimes_{i\in[n]} \widehat{\mathcal{D}}_i$, where each $\widehat{\mathcal{D}}_i$ is a m-dimensional distribution generated by a Bayesnet p_i, as in Definition 11, defined on a DAG with $N_i \geq m$ nodes, in-degree at most d, and $supp(\widehat{\mathcal{D}}_i) \subseteq \Sigma^m \subseteq [0, H]^m$. When $N_i > m$, we say $\widehat{\mathcal{D}}_i$ is generated by an MRF with $N_i - m$ latent variables. We use N to denote $\max_{i\in[n]}\{N_i\}$.*

For every $\varepsilon, \delta > 0$, and $\eta \leq \text{poly}(n, m, \mathcal{L}, H, \varepsilon)$, we can learn, with probability at least $1 - \delta$, an η-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR mechanism $\widehat{M}$, whose revenue under truthful bidding is at most ε smaller than the optimal revenue achievable by any η-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR mechanism, with

- $\text{poly}\left(n, d, N, |\Sigma|^{d+1}, \mathcal{L}, H, 1/\eta, 1/\varepsilon, \log(1/\delta)\right)$ *samples if* ***the alphabet Σ is finite****;*
- $\text{poly}\left(n, d^{d+1}, N^{d+1}, (\frac{HC}{\varepsilon})^{d+1}, \mathcal{L}, 1/\eta, \log(1/\delta)\right)$ *samples if* ***the alphabet $\Sigma = [0, H]$****, and for every p_i, the conditional probability of every node v is C-Lipschitz in the ℓ_1-norm (see Theorem 13 for the definition).*

If $n = 1$, the mechanism $\widehat{M}$ will be IR and IC, and

$$\textsc{Rev}\left(\widehat{M}, \widehat{\mathcal{D}}\right) \geq \left(1 - \sqrt{\eta}\right) \cdot OPT\left(\widehat{\mathcal{D}}\right) - \varepsilon - \sqrt{\eta}.$$

H.2 Sample Complexity for Learning MRFs and Bayesnets

In this section, we present the sample complexity of learning an MRF or a Bayesnet. Our sample complexity scales gracefully with the maximum size of hyperedges in an MRF and indegree in a Bayesnet. Furthermore, our results hold even in the presence of latent variables, where we can only observe the values of k variables, out of the total $|V|$ variables, in a sample.

Theorem 12 (Learnability of MRFs in Total Variation and Prokhorov Distance). *Suppose we are given sample access to an MRF p, as in Definition 10, defined on an unknown graph with hyper-edges of size at most d.*

- ***Finite alphabet*** *Σ: Given $\frac{\text{poly}\left(|V|^d, |\Sigma|^d, \log(\frac{1}{\varepsilon})\right)}{\varepsilon^2}$ samples from p we can learn some MRF q whose hyper-edges also have size at most d such that $\|p - q\|_{TV} \leq \varepsilon$. If the graph on which p is defined is known, then $\frac{\text{poly}\left(|V|, |E|, |\Sigma|^d, \log(\frac{1}{\varepsilon})\right)}{\varepsilon^2}$-many samples suffice. Moreover, the polynomial dependence of the sample complexity on $|\Sigma|^d$ cannot be improved, and the dependence on ε is tight up to $\text{poly}(\log \frac{1}{\varepsilon})$ factors.*
- ***Alphabet*** *$\Sigma = [0, H]$: If the log potentials $\phi_v(\cdot) \equiv \log(\psi_v(\cdot))$ and $\phi_e(\cdot) \equiv \log(\psi_e(\cdot))$ for every node v and every edge e are C-Lipschitz w.r.t. the ℓ_1-norm, then given $\text{poly}\left(|V|^{d^2}, \left(\frac{H}{\varepsilon}\right)^d, C^d\right)$ samples from p we can learn some MRF q whose hyper-edges also have size at most d such that $\|p - q\|_P \leq \varepsilon$. If the graph on which p is defined is known, then $\text{poly}\left(|V|, |E|^d, \left(\frac{H}{\varepsilon}\right)^d, C^d\right)$-many samples suffice.*

Our sample complexity bounds can be easily extended to MRFs with latent variables*, i.e. to the case where some subset $V' \subseteq V$ of the variables are observable in each sample we draw from p. Suppose $k = |V'| \leq |V|$ is the number of observable variables. In this case, for all settings discussed above, our sample complexity bound only increases by a $k \cdot \log|V|$ multiplicative factor.*

Theorem 13 (Learnability of Bayesnets in Total Variation and Prokhorov Distance). *Suppose we are given sample access to a Bayesnet p, as in Definition 11, defined on an unknown DAG with in-degree at most d.*

- ***Finite alphabet*** *Σ: Given $O\left(\frac{d|V|\log|V|+|V|\cdot|\Sigma|^{d+1}\log\left(\frac{|V||\Sigma|}{\varepsilon}\right)}{\varepsilon^2}\right)$-many samples from p we can learn some Bayesnet q defined on a DAG whose in-degree is also bounded by d such that $\|p-q\|_{TV}\le\varepsilon$. If the graph on which p is defined is known, then $O\left(\frac{|V|\cdot|\Sigma|^{d+1}\log\left(\frac{|V||\Sigma|}{\varepsilon}\right)}{\varepsilon^2}\right)$-many samples suffice. Moreover, the dependence of the sample complexity on $|\Sigma|^{d+1}$ and $\frac{1}{\varepsilon}$ is tight up to logarithmic factors.*
- ***Alphabet*** *$\Sigma=[0,H]$: Suppose that the conditional probability distribution of every node v is C-Lipschitz in the ℓ_1-norm, that is, $\left\|p_{X_v|X_{\Pi_v}=\sigma}-p_{X_v|X_{\Pi_v}=\sigma'}\right\|_{TV}\le C\cdot\|\sigma-\sigma'\|_1$, $\forall v$ and $\sigma,\sigma'\in\Sigma^{\Pi_v}$. Then, given $O\left(\frac{d|V|\log|V|+|V|\cdot\left(\frac{H|V|dC}{\varepsilon}\right)^{d+1}\log\left(\frac{|V|HdC}{\varepsilon}\right)}{\varepsilon^2}\right)$-many samples from p, we can learn some Bayesnet q defined on a DAG whose in-degree is also bounded by d such that $\|p-q\|_P\le\varepsilon$. If the graph on which p is defined is known, then $O\left(\frac{|V|\cdot\left(\frac{H|V|dC}{\varepsilon}\right)^{d+1}\log\left(\frac{|V|HdC}{\varepsilon}\right)}{\varepsilon^2}\right)$-many samples suffice.*

Our sample complexity bounds can be easily extended to Bayesnets with latent variables*, i.e. to the case where some subset $V'\subseteq V$ of the variables are observable in each sample we draw from p. Suppose $k=|V'|\le|V|$ is the number of observable variables. In this case, for all settings discussed above, our sample complexity bound only increases by a $k\cdot\log|V|$ multiplicative factor.*

In our proof of Theorem 12, we first carefully construct an ε-net over all MRFs with hyperedges of size at most d in either total variation distance or Prokhorov distance, then apply a tournament-style density estimation algorithm [1, 26, 29] to learn a distribution from the ε-net that is at most $O(\varepsilon)$ away from the true distribution using polynomially many samples. Our proof of Theorem 13 follows a similar recipe. The main difference is how we construct the ε-net over all Bayesnets with in-degree at most d. Both proofs are presented in Appendix J.

I MISSING PROOFS FROM SECTION F

We first show that for any product distribution $\mathcal{F}$, we can learn the rounded distribution of $\mathcal{F}$ within small TV distance with polynomially many samples.

Lemma 12. *Let $\mathcal{F}=\bigtimes_{j=1}^m\mathcal{F}_j$, where $\mathcal{F}_j$ is an arbitrary distribution supported on $[0,H]$ for every $j\in[m]$. Given $N=O\left(\frac{m^3H}{\eta^3}\cdot(\log 1/\delta+\log m)\right)$ samples, we can learn a product distribution $\widehat{\mathcal{F}}=\bigtimes_{j=1}^m\widehat{\mathcal{F}}_j$ such that*

$$\left\|\mathcal{F}-\widehat{\mathcal{F}}\right\|_P\le\eta$$

with probability at least $1-\delta$.

Proof. We denote the samples as $s^1,\dots,s^N$. Round each sample to multiples of $\eta'=\eta/m$. More specifically, let $\hat{s}^i=\left(\left\lfloor s_1^i/\eta'\right\rfloor\cdot\eta',\dots,\left\lfloor s_m^i/\eta'\right\rfloor\cdot\eta'\right)$ for every sample $i\in[N]$. Let $\widehat{\mathcal{F}}_j$ be the uniform distribution over $\hat{s}_j^1,\dots,\hat{s}_j^N$. Let $\overline{\mathcal{F}}_j=\left\lfloor\mathcal{F}_j\right\rfloor_{0,\eta'}$. Note that $\widehat{\mathcal{F}}_j$ is the empirical distribution of N samples from $\overline{\mathcal{F}}_j$. As $\left|\mathrm{supp}(\overline{\mathcal{F}}_j)\right|=\left\lfloor\frac{H}{\eta'}\right\rfloor=\frac{mH}{\eta}$, with $N=O\left(\frac{|\mathrm{supp}(\overline{\mathcal{F}}_j)|}{\eta'^2}\cdot(\log 1/\delta+\log m)\right)$ samples,

the empirical distribution $\widehat{\mathcal{F}}_j$ should satisfy $\left\|\widehat{\mathcal{F}}_j - \overline{\mathcal{F}}_j\right\|_{TV} \le \eta'$ with probability at least $1 - \delta/m$. By the union bound $\left\|\widehat{\mathcal{F}}_j - \overline{\mathcal{F}}_j\right\|_{TV} \le \eta'$ for all $j \in [m]$ with probability at least $1 - \delta$, which implies $\left\|\widehat{\mathcal{F}} - \overline{\mathcal{F}}\right\|_{TV} \le \eta$ with probability at least $1 - \delta$. Observe that $\overline{\mathcal{F}}$ and $\mathcal{F}$ can be coupled so that the two samples are always within η in ℓ_1 distance. When $\left\|\widehat{\mathcal{F}} - \overline{\mathcal{F}}\right\|_{TV} \le \eta$, consider the coupling between $\widehat{\mathcal{F}}$ and $\mathcal{F}$ by composing the optimal coupling between $\widehat{\mathcal{F}}$ and $\overline{\mathcal{F}}$ and the coupling between $\overline{\mathcal{F}}$ and $\mathcal{F}$. Clearly, the two samples from $\widehat{\mathcal{F}}$ and $\mathcal{F}$ are within ℓ_1 distance η with probability at least $1 - \eta$. Due to Theorem 2, the existence of this coupling implies that $\left\|\widehat{\mathcal{F}} - \mathcal{F}\right\|_P \le \eta$. □

Proof of Theorem 9: We only consider the case, where $\eta \le \alpha \cdot \min\left\{\frac{\varepsilon}{n}, \frac{\varepsilon^2}{n^2 m \mathcal{L} H}\right\}$. α is an absolute constant and we will specify its choice in the end of the proof.

In light of Lemma 12, we take $N = O\left(\frac{m^3 H}{\sigma^3} \cdot (\log \frac{n}{\delta} + \log m)\right)$ from $\widehat{\mathcal{D}}$ and learn a distribution $\mathcal{D}$ so that, with probability at least $1 - \delta$, $\left\|\mathcal{D}_i - \widehat{\mathcal{D}}_i\right\|_P \le \sigma$ for all $i \in [n]$. According to Theorem 7, we can transform M into mechanism $\widehat{M}$ that is $O\left(nm\mathcal{L}H\sigma + m\mathcal{L}\sqrt{nH\sigma}\right)$-BIC w.r.t. $\widehat{\mathcal{D}}$ and IR. Choose σ in a way so that $\widehat{M}$ is η-BIC w.r.t. $\widehat{\mathcal{D}}$. Moreover, $\widehat{M}$'s revenue under truthful bidding satisfies

$$\textsc{Rev}_T\left(\widehat{M}, \widehat{\mathcal{D}}\right) \ge c \cdot \mathrm{OPT}_\eta\left(\widehat{\mathcal{D}}\right) - O\left(n\eta + n\sqrt{m\mathcal{L}H\eta}\right).$$

If we choose α to be sufficiently small, then

$$\textsc{Rev}_T\left(\widehat{M}, \widehat{\mathcal{D}}\right) \ge c \cdot \mathrm{OPT}_\eta\left(\widehat{\mathcal{D}}\right) - \varepsilon.$$

When there is only a single-bidder, we can apply Lemma 5 to transform $\widehat{M}$ to an IC and IR mechanism, whose revenue satisfies the guarantee in the statement.

□

J MISSING PROOFS FROM SECTION H

J.1 Proof of Theorem 12

Proof of Theorem 12: For the purposes of this proof we take $n = |V|$. We first prove the finite alphabet case, we then extend the result to the infinite alphabet case, and finally we discuss how to accommodate latent variables.

Finite alphabet Σ: We will prove our first sample complexity bound by constructing an ε-cover, in total variation distance, of the set $\mathcal{P}$ of all MRFs with hyperedges of size at most d. We can assume that all $p \in \mathcal{P}$ satisfy the following:

(*A*1) : p is defined on the hypergraph $G = (V, E)$, whose edge set is $E = \binom{V}{d}$, and all its node potential functions are constant and equal 1.

The reason we can assume (*A*1) for all $p \in \mathcal{P}$ is that potentials of nodes and smaller-size hyperedges can always be incorporated into the potentials of some size-d hyperedge that contains them, and the potentials of size-d hyperedges that are not present can always be taken to be constant 1 functions.

Moreover, we can assume the following property for all MRFs $p \in \mathcal{P}$:

(*A*2): $\max_{\sigma \in \Sigma^e} \psi_e(\sigma) = 1, \forall e \in E$.

The reason we can assume $(A2)$ for all $p \in \mathcal{P}$ is that the density of an MRF is invariant to multiplying any single potential function by some scalar.

Now, given some MRF $p \in \mathcal{P}$, satisfying $(A1)$ and $(A2)$, which we can assume without loss of generality, we will make a sequence of transformations to arrive at some MRF $p'' \in \mathcal{P}$ such that $\|p - p''\|_{TV} \leq \varepsilon$ and p'' can be described using $B = \text{poly}\left(|E|, |\Sigma|^d, \log(\frac{1}{\varepsilon})\right)$ bits. This, in turn, will imply that there exists an ε-cover $\mathcal{P}' \subset \mathcal{P}$ that has size 2^B, and the existence of an ε-cover of this size implies that $O(B/\varepsilon^2)$-many samples from any $p \in \mathcal{P}$ suffice to learn some $q \in \mathcal{P}$ such that $\|p - q\|_{TV} \leq O(\varepsilon)$, using a tournament-style density estimation algorithm; see e.g. [1, 26, 29] and their references.

Here are the steps to transform an arbitrary $p \in \mathcal{P}$ into some $p'' \in \mathcal{P}$ of low bit complexity:

- (**Notation**:) From now on we will use $\hat{p}$ to denote unnormalized densities. I.e. if p is defined in terms of potential functions $(\psi_e^p(\cdot))_{e \in E}$, then $\hat{p}(x) = \prod_{e \in E} \psi_e^p(x_e), \forall x \in \Sigma^V$.
- (**Step 1:**) Given some arbitrary $p \in \mathcal{P}$, we construct some $p' \in \mathcal{P}$ such that $\|p - p'\|_{TV} \leq \varepsilon$, p' satisfies $(A1)$ and $(A2)$ and, moreover, the unnormalized density of p' satisfies that, for all $x \in \Sigma^V$, $\hat{p}'(x) = \left(1 + \frac{\varepsilon}{2n^d}\right)^{i_x}$, for some integer i_x. The existence of such p' follows from the invariance of MRFs with respect to multiplying their potential functions by scalars, and the following.

  Claim 3. *Suppose $p, p' \in \mathcal{P}$ satisfy $(A1)$ and are defined in terms of potential functions $(\psi_e^p)_e$ and $(\psi_e^{p'})_e$ respectively. Moreover, suppose that $\forall e, \sigma \in \Sigma^e$:*

$$\psi_e^{p'}(\sigma) \leq \psi_e^p(\sigma) \leq \left(1 + \frac{\varepsilon}{2n^d}\right) \psi_e^{p'}(\sigma).$$

  Then $\|p - p'\|_{TV} \leq \varepsilon$.

  Proof of Claim 3: It follows from the condition in the statement of the claim that, for all $x \in \Sigma^V$:

$$\hat{p}'(x) \leq \hat{p}(x) \leq \left(1 + \frac{\varepsilon}{2n^d}\right)^{\binom{n}{d}} \hat{p}'(x) \leq e^{\varepsilon/2} \hat{p}'(x) \leq (1 + \varepsilon)\hat{p}'(x).$$

  Using the above, let us compare the normalized densities. For all $x \in \Sigma^V$:

$$p(x) = \frac{\hat{p}(x)}{\sum_y \hat{p}(y)} \leq \frac{\hat{p}'(x)(1 + \varepsilon)}{\sum_y \hat{p}'(y)} \leq p'(x)(1 + \varepsilon).$$

  Moreover,

$$p(x) = \frac{\hat{p}(x)}{\sum_y \hat{p}(y)} \geq \frac{\hat{p}'(x)}{\sum_y \hat{p}'(y)(1 + \varepsilon)} \geq p'(x)/(1 + \varepsilon).$$

  Using the above, let us bound the total variation distance between p and p':

$$\begin{aligned}
\|p - p'\|_{TV} &= \frac{1}{2} \sum_x |p(x) - p'(x)| \\
&= \frac{1}{2} \sum_{x : p(x) \geq p'(x)} (p(x) - p'(x)) + \frac{1}{2} \sum_{x : p(x) < p'(x)} (p'(x) - p(x)) \\
&\leq \frac{1}{2} \sum_{x : p(x) \geq p'(x)} \varepsilon p'(x) + \frac{1}{2} \sum_{x : p(x) < p'(x)} \varepsilon p(x) \leq \varepsilon.
\end{aligned}$$

  □

- (**New Notation:**) We introduce some further notation. Let $\left(\psi_e^{p'}\right)_e$ be the potential functions defining distribution $p' \in \mathcal{P}$ from Step 1. We reparametrize these potential functions as follows:

$$\forall e, x \in \Sigma^e : \xi_e^{p'}(x) \equiv \log\left(\psi_e^{p'}(x)\right) / \log\left(1 + \frac{\varepsilon}{2n^d}\right).$$

  Given the definition of p' in Step 1, our new potential functions satisfy the following linear equations:

$$\forall x \in \Sigma^V : \sum_{e \in E} \xi_e^{p'}(x_e) = i_x, \tag{10}$$

  where, because of Assumption (*A2*), satisfied by p', the integers $i_x \leq 0$, for all x.
- (**Step 2:**) We define p'' by setting up a linear program with variables $\xi_e^{p''}(x_e), \forall e, x_e \in \Sigma^E$. In particular, the number of variables of the linear program we are about to write is $L = |E| \cdot |\Sigma|^d$. To define our linear program, we first define $x^* = \text{argmax}_x\, i_x$, and partition Σ^V into two sets $\Sigma^V = \mathcal{G} \sqcup \mathcal{B}$, by taking $\mathcal{G} = \{x \mid i_x \geq i_{x^*} - T\}$, and $\mathcal{B}$ the complement of $\mathcal{G}$, for $T = \frac{4n^d}{\varepsilon}(n \log|\Sigma| + \log(\frac{1}{\varepsilon}))$. In particular, all configurations in $\mathcal{B}$ have probability $p'(x) \leq \varepsilon/|\Sigma|^n$. Our goal is to exhibit that there exists $p'' \in \mathcal{P}$ that (i) satisfies properties (*A1*) and (*A2*); (ii) can be described with poly $\left(|E|, |\Sigma|^d, \log(\frac{1}{\varepsilon})\right)$ bits; and (iii) satisfies $\sum_{x \in \mathcal{B}} p''(x) \leq \varepsilon$ and $p''(x) = p'(x) \cdot (1 + \delta)\ \forall x \in \mathcal{G}$, where $\delta \in \left[-\varepsilon, \frac{\varepsilon}{1-\varepsilon}\right]$. We note that (iii) implies that $\|p' - p''\|_{TV} \leq \varepsilon$, as either $p''(x) \geq p'(x)$ for all $x \in \mathcal{G}$ simultaneously or $p''(x) < p'(x)$ for all $x \in \mathcal{G}$ simultaneously, and the total mass in $\mathcal{B}$ under both p' and p'' are at most ε. Combining (iii) and Claim 3, we have (iv) $\|p - p''\|_{TV} \leq 2\varepsilon$. To exhibit the existence of p'' we write the following linear program:

$$\begin{aligned} \forall x \in \mathcal{G} \setminus \{x^*\} : \sum_{e \in E} \xi_e^{p''}(x_e) - \sum_{e \in E} \xi_e^{p''}(x_e^*) &= i_x - i_{x^*} \\ \forall x \in \mathcal{B} : \sum_{e \in E} \xi_e^{p''}(x_e) - \sum_{e \in E} \xi_e^{p''}(x_e^*) &\leq -T \end{aligned} \tag{11}$$

  Note that, because LP (10) is feasible, it follows that LP (11) is feasible as well. Moreover, the coefficients and constants of LP (11) have absolute value less than T and bit complexity polynomial in d, $\log n$, $\log(\frac{1}{\varepsilon})$ and $\log\log|\Sigma|$, and the number of variables of this LP is $L = |E| \cdot |\Sigma|^d$. From the theory of linear programming it follows that there exists a solution to LP (11) of bit complexity polynomial in $|E|$, $|\Sigma|^d$, $\log n$, and $\log(\frac{1}{\varepsilon})$. Why is (iii) true? It is not hard to see that for any $x \in \mathcal{B}$, $p''(x) \leq \varepsilon/|\Sigma|^n$ due to the second type of constraints in LP (11). For any $x \in \mathcal{G} \setminus \{x^*\}$, $\frac{p''(x)}{p''(x^*)} = \frac{p'(x)}{p'(x^*)}$ due to the first type of constraints in LP (11), so $p''(x) = p'(x) \cdot (1 + \delta)\ \forall x \in \mathcal{G}$ for some constant δ. Since both $\sum_{x \in \mathcal{G}} p'(x)$ and $\sum_{x \in \mathcal{G}} p''(x)$ lie in $[1 - \varepsilon, 1]$, δ lies in $\left[-\varepsilon, \frac{\varepsilon}{1-\varepsilon}\right]$.

To summarize the above (setting $\varepsilon \leftarrow \varepsilon/2$ in the above derivation), given an arbitrary $p \in \mathcal{P}$ we can construct $p'' \in \mathcal{P}$ such that: p'' can be described using $B = \text{poly}\left(|E|, |\Sigma|^d, \log(\frac{1}{\varepsilon})\right)$ bits—by specifying the low complexity solution $\left(\xi_e^{p''}\right)_e$ to LP (11), and p'' satisfies $\|p - p''\|_{TV} \leq \varepsilon$. As we have noted above, the existence of such p'' for every $p \in \mathcal{P}$ implies the existence of an ε-cover, in total variation distance, of $\mathcal{P}$ that has size 2^B, and tournament-style arguments imply then that any $p \in \mathcal{P}$ can be learned to within $O(\varepsilon)$ in total variation distance from $O(\frac{B}{\varepsilon^2})$-many samples, i.e. from $\frac{\text{poly}\left(|E|, |\Sigma|^d, \log(\frac{1}{\varepsilon})\right)}{\varepsilon^2}$-many samples.

We now prove the second part of the statement. If the hypergraph (V, E_p) with respect to which p is defined is known, we redo the above argument, except we take $\mathcal{P}$ to be all MRFs defined on the graph $G = (V, E)$, where E is the union of E_p and all singleton sets corresponding to the nodes V.

For the third part of the statement, we note that an arbitrary distribution p on d variables, each taking values in Σ, can be expressed as a MRF with maximum hyperedge-size d. As such, it is folklore (see e.g. [29]) that $\Omega(|\Sigma|^d/\varepsilon^2)$ samples are necessary to learn p to within ε in total variation distance. This completes the proof for the finite alphabet case.

Next, we show how to extend our sample complexity to the case where the alphabet $\Sigma = [0, H]$.

Alphabet $\Sigma = [0, H]$*:* Let $\delta = \frac{\varepsilon}{8dC(n+1)^d}$, and Σ_δ be the set of all multiples of δ between 0 and H.[10] We first define distribution $\tilde{p}$ to be the rounded version of p using the following coupling. For any sample x drawn from p, create a sample $\tilde{x}$ drawn from $\tilde{p}$ such that $\tilde{x}_v = \left\lfloor \frac{x_v}{\delta} \right\rfloor \cdot \delta$ for every $v \in V$. Note that (i) this coupling makes sure that the two samples from p and $\tilde{p}$ are always within ε of each other in ℓ_1-distance. Our plan is to show that we can (ii) learn an MRF q with polynomially many samples from distribution $\tilde{p}$ such that $\|q - \tilde{p}\|_{TV} = O(\varepsilon)$. Why does this imply our statement? First, we can generate a sample from $\tilde{p}$ using a sample from p due to the coupling between the two distributions. Second, $\|q - \tilde{p}\|_{TV} = O(\varepsilon)$ means that we can couple q and $\tilde{p}$ in a way that the two samples are the same with probability at least $1 - O(\varepsilon)$. Composing this coupling with the coupling between $\tilde{p}$ and p, we have a coupling between p and q so that the two samples are within ε of each other in ℓ_1-distance with probability at least $1 - O(\varepsilon)$. According to Theorem 2, $\|p - q\|_P = O(\varepsilon)$. Now, we focus on proving (ii).

We separate the proof into two steps. In the first step, we show that for any $\tilde{p}$, there is a discretized MRF q' supported on Σ_δ^V with hyperedges of size at most d such that $\|\tilde{p} - q'\|_{TV} \le \varepsilon$ and q' can be described with $B = \text{poly}\left(|E|, |\Sigma_\delta|^d, \log(\frac{1}{\varepsilon})\right)$ bits. In other words, there is a 2^B-sized ε-cover over all possible distributions $\tilde{p}$. In the second step, we show how to learn an MRF q with $O(B/\varepsilon^2)$ samples from $\tilde{p}$ using a tournament-style density estimation algorithm; see e.g. [1, 26, 29] and their references. Before we present the two steps of our proof, and in order to simplify our notation and avoid carrying around node potentials, let us introduce into the edge set E of our hypergraph a singleton edge for every node v, and take the potential of every such edge $e = \{v\}$ to equal the node potential of node v.

- **(Step 1:)** We first define a discrete MRF p' on the same graph $G = (V, E)$ as p with alphabet Σ_δ. Distribution p' is defined by choosing its log-potential $\phi_e^{p'}(x_e)$ to be exactly $\phi_e^p(x_e)$ for every hyperedge $e \in E$ and every possible value $x_e \in \Sigma_\delta^e$. Next, we show that (iii) $\|p' - \tilde{p}\|_{TV} \le \varepsilon/2$. We use A_x to denote the n-dimensional cube $\bigtimes_{v \in V}[x_v, x_v + \delta)$ for any $x \in \Sigma_\delta^V$. Note that

$$\tilde{p}(x) = \frac{\int_{A_x} \exp\left(\sum_e \phi_e^p(y_e)\right) dy}{\int_{[0,H]^n} \exp\left(\sum_e \phi_e^p(y_e)\right) dy} \le \frac{\delta^n \exp\left(\sum_e \phi_e^p(x_e)\right) \cdot \exp(d|E|C\delta)}{\delta^n \sum_{y \in \Sigma_\delta^V} \exp\left(\sum_e \phi_e^p(y_e)\right) \cdot \exp(-d|E|C\delta)} \le p'(x)(1 + \varepsilon/2).$$

  The first inequality is due the C-Lipschitzness of the log potential functions and the second inequality is due to the definition of δ. Similarly,

$$\tilde{p}(x) = \frac{\int_{A_x} \exp\left(\sum_e \phi_e^p(y_e)\right) dy}{\int_{[0,H]^n} \exp\left(\sum_e \phi_e^p(y_e)\right) dy} \ge \frac{\delta^n \exp\left(\sum_e \phi_e^p(x_e)\right) \cdot \exp(-d|E|C\delta)}{\delta^n \sum_{y \in \Sigma_\delta^V} \exp\left(\sum_e \phi_e^p(y_e)\right) \cdot \exp(d|E|C\delta)} \ge \frac{p'(x)}{1 + \varepsilon/2}.$$

[10] We further assume that H is a multiple of δ. If not, let k be the integer such that $\delta \in \left[\frac{H}{2^k}, \frac{H}{2^{k-1}}\right]$, and change δ to be $\frac{H}{2^k}$.

We complete the proof of (iii) by combining the two inequalities.

$$\begin{aligned}\|\tilde{p} - p'\|_{TV} &= \frac{1}{2}\sum_{x\in\Sigma_\delta^V} |\tilde{p}(x) - p'(x)| \\ &= \frac{1}{2}\sum_{x:\tilde{p}(x)\geq p'(x)} (\tilde{p}(x) - p'(x)) + \frac{1}{2}\sum_{x:\tilde{p}(x)<p'(x)} (p'(x) - \tilde{p}(x)) \\ &\leq \frac{1}{2}\sum_{x:\tilde{p}(x)\geq p'(x)} \frac{\varepsilon}{2} p'(x) + \frac{1}{2}\sum_{x:\tilde{p}(x)<p'(x)} \frac{\varepsilon}{2}\tilde{p}(x) \leq \varepsilon/2.\end{aligned}$$

Let $\mathcal{P}$ be the set of all MRFs with hyperedges of size at most d and alphabet Σ_δ. By redoing Step 1 and 2 of the proof for the finite alphabet case, we can show that (iv) for any $\hat{p} \in \mathcal{P}$, there exists another $\hat{p}' \in \mathcal{P}$ describable with $B = \text{poly}\left(|E|, |\Sigma_\delta|^d, \log(\frac{1}{\varepsilon})\right)$ bits such that $\|\hat{p} - \hat{p}'\|_{TV} \leq \varepsilon/2$. Since $p' \in \mathcal{P}$, there exists a $q' \in \mathcal{P}$ describable with B bits such that $\|p' - q'\|_{TV} \leq \varepsilon/2$. Combining this inequality with (iii), we have $\|\tilde{p} - q'\|_{TV} \leq \varepsilon$.

- **(Step 2:)** Let $\mathcal{P}' \subset \mathcal{P}$ be the set of all MRFs in $\mathcal{P}$ with bit complexity at most B from Step 1. Since $\min_{\tilde{q}\in\mathcal{P}'} \|\tilde{q} - \tilde{p}\|_{TV} \leq \varepsilon$, we can learn an MRF $q \in \mathcal{P}'$ such that $\|q - \tilde{p}\|_{TV} \leq O(\varepsilon)$ with $O(B/\varepsilon^2)$ samples from $\tilde{p}$ using a tournament-style density estimation algorithm [1, 26, 29].

To sum up, we can learn an MRF q such that $\|q - p\|_P \leq \varepsilon$ with $\text{poly}\left(|V|^{d^2}, \left(\frac{H}{\varepsilon}\right)^d, C^d\right)$ many samples from p. If the graph G on which p is defined is known, we can choose δ to be $O\left(\frac{\varepsilon}{8dC|E|}\right)$ and improve the sample complexity to $\text{poly}\left(|V|, |E|^d, \left(\frac{H}{\varepsilon}\right)^d, C^d\right)$.

Latent Variable Models: Finally, we consider the case where only k out of the n variables of the MRF are observable. Let S be the set of observable variables, and use p_S to denote the marginal of p on these variables. We will first consider the finite alphabet case. Consider the ε-cover we constructed earlier. We argued that for any MRF p there exists an MRF q in the cover such that $\|p - q\|_{TV} \leq \varepsilon$. For that q we clearly also have $\|p_S - q_S\|_{TV} \leq \varepsilon$. The issue is that we do not know for a given q in the cover which subset of its variables set S might correspond to. But this is not a big deal. We can use our cover to generate an ε-cover of all possible marginals p_S of all possible MRFs p as follows. Indeed, for any q' in the original ε-cover, we include in the new cover the marginal distribution $q'_{S'}$ of every possible subset S' of its variables of size k. This increases the size of our original cover by a multiplicative factor of at most n^k. As a result, the number of samples required for the tournament-style density estimation algorithm to learn a good distribution increases by a multiplicative factor of $k \log n$. For the infinite alphabet case, our statement follows from applying the same modification to the ε-cover of $\tilde{p}$. □

J.2 Proof of Theorem 13

Proof of Theorem 13: We first prove the theorem statement for the finite alphabet case, we then extend it to the infinite alphabet case, and finally show how we can accommodate latent variables as well.

Finite alphabet Σ*:* We prove the claims in the theorem statement in reverse order.

For the third part of the statement, we note that an arbitrary distribution p on $d+1$ variables, each taking values in Σ, can be expressed as a Bayesnet with maximum indegree d. As such, it is folklore (see e.g. [29]) that $\Omega(|\Sigma|^{d+1}/\varepsilon^2)$ samples are necessary to learn p to within ε in total variation distance.

To prove the second part of the statement, we show that there is an ε-cover, in total variation distance, of all Bayesnets $\mathcal{P}$ on a given DAG G of indegree at most d, which has size $B = \left(\frac{n|\Sigma|}{\varepsilon}\right)^{n|\Sigma|^{d+1}}$, where $n = |V|$. The existence of an ε-cover of this size implies that $O(\log(B)/\varepsilon^2)$-many samples from any $p \in \mathcal{P}$ suffice to learn some $q \in \mathcal{P}$ such that $\|p - q\|_{TV} \leq O(\varepsilon)$, using a tournament-style density estimation algorithm; see e.g. [1, 26, 29] and their references. Thus, to prove the second part of the theorem statement it suffices to argue that an ε-cover of size B exists. We prove the existence of this cover by exploiting the following lemma.

Lemma 13. *Suppose p and q are Bayesenets on the same DAG $G = (V, E)$. Suppose that, for all $v \in V$, for all $\sigma \in \Sigma^{\Pi(v)}$, where $\Pi(v)$ are the parents of v in G (using the same notation as in Definition 11), it holds that*

$$\left\| p_{X_v | X_{\Pi_v} = \sigma} - q_{X_v | X_{\Pi_v} = \sigma} \right\|_{TV} \leq \frac{\varepsilon}{|V|}.$$

Then $\|p - q\|_{TV} \leq \varepsilon$.

Proof of Lemma 13: We employ a hybrid argument. First, let us denote $n = |V|$ and label the nodes in V with labels $1, \ldots, n$ according to some topological sorting of G. In particular, the parents (if any) of any node i have indices $< i$. Now, for our hybrid argument we construct the following auxiliary distributions, for $i = 0, \ldots, n$:

$$h^i(x) = \prod_{v=1}^{i} p_{X_v | X_{\Pi_v}}(x_v | x_{\Pi_v}) \prod_{v=i+1}^{n} q_{X_v | X_{\Pi_v}}(x_v | x_{\Pi_v}), \text{for all } x \in \Sigma^V.$$

In particular, $h^0 \equiv q$ and $h^n \equiv p$, and the rest are fictional distributions. By triangle inequality, we have that:

$$\|p - q\|_{TV} \leq \sum_{i=1}^{n} \left\| h^i - h^{i-1} \right\|_{TV}.$$

We will bound each term on the RHS by ε/n to conclude the proof of the lemma. Indeed,

$$
\begin{aligned}
&\left\|h^i - h^{i-1}\right\|_{TV} \\
&= \sum_x \left| \prod_{v=1}^{i} p_{X_v|X_{\Pi_v}}(x_v|x_{\Pi_v}) \cdot \prod_{v=i+1}^{n} q_{X_v|X_{\Pi_v}}(x_v|x_{\Pi_v}) - \prod_{v=1}^{i-1} p_{X_v|X_{\Pi_v}}(x_v|x_{\Pi_v}) \cdot \prod_{v=i}^{n} q_{X_v|X_{\Pi_v}}(x_v|x_{\Pi_v}) \right| \\
&= \sum_x \left| \prod_{v=1}^{i-1} p_{X_v|X_{\Pi_v}}(x_v|x_{\Pi_v}) \cdot \left(p_{X_i|X_{\Pi_i}}(x_i|x_{\Pi_i}) - q_{X_i|X_{\Pi_i}}(x_i|x_{\Pi_i}) \right) \cdot \prod_{v=i+1}^{n} q_{X_v|X_{\Pi_v}}(x_v|x_{\Pi_v}) \right| \\
&= \sum_x \prod_{v=1}^{i-1} p_{X_v|X_{\Pi_v}}(x_v|x_{\Pi_v}) \cdot \left| p_{X_i|X_{\Pi_i}}(x_i|x_{\Pi_i}) - q_{X_i|X_{\Pi_i}}(x_i|x_{\Pi_i}) \right| \cdot \prod_{v=i+1}^{n} q_{X_v|X_{\Pi_v}}(x_v|x_{\Pi_v}) \\
&= \sum_{x_{1\ldots i-1}} \prod_{v=1}^{i-1} p_{X_v|X_{\Pi_v}}(x_v|x_{\Pi_v}) \cdot \\
&\qquad \sum_{x_i} \left(\left| p_{X_i|X_{\Pi_i}}(x_i|x_{\Pi_i}) - q_{X_i|X_{\Pi_i}}(x_i|x_{\Pi_i}) \right| \cdot \sum_{x_{i+1\ldots n}} \left(\prod_{v=i+1}^{n} q_{X_v|X_{\Pi_v}}(x_v|x_{\Pi_v}) \right) \right) \\
&= \sum_{x_{1\ldots i-1}} \left(\prod_{v=1}^{i-1} p_{X_v|X_{\Pi_v}}(x_v|x_{\Pi_v}) \cdot \sum_{x_i} \left(\left| p_{X_i|X_{\Pi_i}}(x_i|x_{\Pi_i}) - q_{X_i|X_{\Pi_i}}(x_i|x_{\Pi_i}) \right| \right) \right) \\
&\leq \sum_{x_{1\ldots i-1}} \left(\prod_{v=1}^{i-1} p_{X_v|X_{\Pi_v}}(x_v|x_{\Pi_v}) \cdot \varepsilon/n \right) \\
&= \varepsilon/n,
\end{aligned}
$$

where for the inequality we used the hypothesis in the statement of the lemma.□

Now suppose $p \in \mathcal{P}$ is an arbitrary Bayesnet defined on G. It follows from Lemma 13 that p lies ε-close in total variation distance to a Bayesnet q such that, for all $v \in V$, and all $\sigma \in \Sigma^{\Pi_v}$, the conditional distribution $q_{X_v|X_{\Pi_v}=\sigma}$ is a discretized version of $p_{X_v|X_{\Pi_v}=\sigma}$ that is $\frac{\varepsilon}{n}$-close in total variation distance. Note that $p_{X_v|X_{\Pi_v}=\sigma}$ is an element of the simplex over $|\Sigma|$ elements, and it is easy to see that this simplex can be $\frac{\varepsilon}{n}$-covered, in total variation distance, using a discrete set of at most $\left(\frac{n|\Sigma|}{\varepsilon}\right)^{|\Sigma|}$-many distributions. As there are at most $n \cdot |\Sigma|^d$ conditional distributions to discretize, a total number of

$$B = \left(\frac{n|\Sigma|}{\varepsilon}\right)^{n|\Sigma|^{d+1}}$$

discretized distributions suffice to cover all $\mathcal{P}$.

To prove the first part of the theorem statement, we proceed in the same way, except that now that we do not know the DAG our cover will be larger. Since there are at most n^{dn} DAGs of indegree at most d on n labeled vertices, and for each DAG there is a cover of all Bayesnets defined on that DAG of size at most B, as above, it follows that there is an ε-cover, in total variation distance, of all Bayesnets of indegree at most d of size:

$$n^{dn} \cdot B.$$

Given the bound on the cover size, the proof concludes by appealing to tournament-style density estimation algorithms, as we did earlier. This completes our proof for the finite alphabet case.

Alphabet $\Sigma = [0, H]$**:** Let $\delta = \frac{\varepsilon}{dCn}$, and Σ_δ be the set of all multiples of δ between 0 and H.[11] For any set of nodes S and $x = (x_v)_{v \in S}$, we use $\lfloor x \rfloor_\delta$ to denote the corresponding rounded vector $\left(\lfloor \frac{x_v}{\delta} \rfloor \cdot \delta\right)_{v \in S}$. We first define distribution $\tilde{p}$ to be the rounded version of p using the following coupling. For any sample x drawn from p, create a sample $\tilde{x} = \lfloor x \rfloor_\delta$ drawn from $\tilde{p}$. Note that (i) this coupling makes sure that the two samples from p and $\tilde{p}$ are always within ε of each other in ℓ_1-distance. Our plan is to show that we can (ii) learn a Bayesnet q with in-degree at most d using polynomially many samples from distribution $\tilde{p}$ such that $\|q - \tilde{p}\|_{TV} = O(\varepsilon)$. Why does this imply our claim? First, we can generate a sample from $\tilde{p}$ using a sample from p due to the coupling between the two distributions. Second, $\|q - \tilde{p}\|_{TV} = O(\varepsilon)$ means that we can couple q and $\tilde{p}$ in a way that the two samples are the same with probability at least $1 - O(\varepsilon)$. Composing this coupling with the coupling between $\tilde{p}$ and p, we have a coupling between p and q such that the two samples are at most ε away from each other in ℓ_1-distance with probability at least $1 - O(\varepsilon)$. This implies, according to Theorem 2, that $\|p - q\|_P = O(\varepsilon)$. Now, we focus on proving (ii) and separate the proof into three steps.

- **(Step 1:)** We first prove that there is a Bayesnet p'' with in-degree at most d and alphabet Σ_δ such that $\|\tilde{p} - p''\|_{TV} \le \varepsilon$. We first construct a Bayesnet p' on the same DAG as p, where the conditional probability distribution for every node v, and $\sigma \in \Sigma^{\Pi_v}$ is defined as

$$p'_{X_v | X_{\Pi(v)=\sigma}} \equiv p_{X_v | X_{\Pi(v)=\lfloor \sigma \rfloor_\delta}}.$$

  Clearly, for any node v, and $\sigma \in \Sigma^{\Pi_v}$,

$$\left\| p_{X_v | X_{\Pi(v)=\sigma}} - p'_{X_v | X_{\Pi(v)=\sigma}} \right\|_{TV} = \left\| p_{X_v | X_{\Pi(v)=\sigma}} - p_{X_v | X_{\Pi(v)=\lfloor \sigma \rfloor_\delta}} \right\|_{TV} \le C \cdot \|\sigma - \lfloor \sigma \rfloor_\delta\|_1 \le Cd\delta \le \frac{\varepsilon}{|V|}.$$

  Hence, Lemma 13 implies that: (iii) $\|p - p'\|_{TV} \le \varepsilon$.[12]
  Next, we construct the rounded distribution p'' of p' via the following coupling. For any sample x' drawn from p', create a sample $x'' = \lfloor x' \rfloor_\delta$ from p''. It is not hard to verify that p'' can also be captured by a Bayesnet defined on the same DAG as p and p'. In particular, for every node v, every $x_v \in \Sigma_\delta$, and $x_{\Pi_v} \in \Sigma_\delta^{\Pi_v}$, the conditional probability is

$$p''_{X_v | X_{\Pi_v}}\left(x_v | x_{\Pi_v}\right) = \int_{x_v}^{x_v + \delta} p'_{X_v | X_{\Pi_v}}\left(z | x_{\Pi_v}\right) dz.$$

  As p'' is the rounded distribution of p', $\tilde{p}$ is the rounded distribution of p, and $\|p - p'\|_{TV} \le \varepsilon$, it must be the case that $\|p'' - \tilde{p}\|_{TV} \le \varepsilon$.
- **(Step 2:)** Let $\mathcal{P}$ be the set of all Bayesnets defined on a DAG with n nodes and in-degree at most d, and which have alphabet Σ_δ. We argue that there is a size $A = n^{dn} \cdot \left(\frac{n|\Sigma_\delta|}{\varepsilon}\right)^{n|\Sigma_\delta|^{d+1}}$ ε-cover $\mathcal{P}'$, in total variation distance, of $\mathcal{P}$, and $\mathcal{P}' \subset \mathcal{P}$. This follows from the same argument we did in the proof for the finite alphabet case. First, there are n^{dn} different DAGs with n nodes and in-degree at most d. Second, for each DAG there are at most $n \cdot |\Sigma_\delta|^d$ conditional distributions. Finally, it suffices to $\frac{\varepsilon}{n}$-cover each conditional distribution, in total variation distance, which can be accomplished by a discrete set of at most $\left(\frac{n|\Sigma_\delta|}{\varepsilon}\right)^{|\Sigma_\delta|}$-many distributions. Since $p'' \in \mathcal{P}$ and $\|p'' - \tilde{p}\|_{TV} \le \varepsilon$, there exists a Bayesnet $\tilde{q}$ from the ε-cover $\mathcal{P}'$ such that $\|\tilde{q} - \tilde{p}\|_{TV} \le 2\varepsilon$.

[11] We further assume that H is a multiple of δ. If not, let k be the integer such that $\delta \in \left[\frac{H}{2^k}, \frac{H}{2^{k-1}}\right]$, and change δ to be $\frac{H}{2^k}$.

[12] Even though Lemma 13 was only proved earlier for a finite alphabet, the same proof extends to when the alphabet is infinite.

- **(Step 3:)** Since $\min_{q' \in \mathcal{P}'} \|q' - \tilde{p}\|_{TV} \leq 2\varepsilon$, we can use a tournament-style density estimation algorithm (see e.g. [1, 26, 29] and their references) to learn a Bayesnet $q \in \mathcal{P}'$ such that $\|q - \tilde{p}\|_{TV} = O(\varepsilon)$ given $O\left(\frac{\log A}{\varepsilon^2}\right)$ samples from $\tilde{p}$.

To sum up, we can learn a Bayesnet q defined on a DAG with in-degree at most d using

$$O\left(\frac{d|V|\log|V| + |V| \cdot \left(\frac{H|V|dC}{\varepsilon}\right)^{d+1} \log\left(\frac{|V|HdC}{\varepsilon}\right)}{\varepsilon^2}\right)$$

samples from p such that $\|q - p\|_P \leq \varepsilon$. If the DAG that p is defined on is known, the sample complexity improves to $O\left(\frac{|V| \cdot \left(\frac{H|V|dC}{\varepsilon}\right)^{d+1} \log\left(\frac{|V|HdC}{\varepsilon}\right)}{\varepsilon^2}\right)$.

Latent Variable Model: Finally, we consider the case where only k out of the n variables of the Bayesnet p are observable. Let S be the set of observable variables, and use p_S to denote the marginal of p on these variables. We will first consider the finite alphabet case. Consider the ε-cover we constructed earlier. We argued that for any Bayesnet p there exists an Bayesnet q in the cover such that $\|p - q\|_{TV} \leq \varepsilon$. For that q we clearly also have $\|p_S - q_S\|_{TV} \leq \varepsilon$. The issue is that we do not know for a given q in the cover which subset of its variables set S might correspond to. But this is not a big deal. We can use our cover to generate an ε-cover of all possible marginals p_S of all possible Bayesnets p as follows. Indeed, for any q' in the original ε-cover, we include in the new cover the marginal distribution $q'_{S'}$ of every possible subset S' of its variables of size k. This increases the size of our original cover by a multiplicative factor of at most n^k. As a result, the number of samples required for the tournament-style density estimation algorithm to learn a good distribution increases by a multiplicative factor of $k \log n$. For the infinite alphabet case, our statement follows from applying the same modification to the ε-cover of $\tilde{p}$. □