On the High Accuracy Limitation of Adaptive Property Estimation

Yanjun Han
Department of Electrical Engineering, Stanford University
yjhan@stanford.edu

Abstract

Recent years have witnessed the success of adaptive (or unified) approaches in estimating symmetric properties of discrete distributions, where the learner first obtains a distribution estimator independent of the target property, and then plugs the estimator into the target property as the final estimator. Several such approaches have been proposed and proved to be adaptively optimal, i.e. they achieve the optimal sample complexity for a large class of properties within a low accuracy, especially for a large estimation error $\varepsilon \gg n^{-1/3}$ where n is the sample size.

In this paper, we characterize the high accuracy limitation, or the penalty for adaptation, for general adaptive approaches. Specifically, we obtain the first known adaptation lower bound that under a mild condition, any adaptive approach cannot achieve the optimal sample complexity for every 1-Lipschitz property within accuracy $\varepsilon \ll n^{-1/3}$. In particular, this result disproves a conjecture in [Acharya et al., 2017a] that the profile maximum likelihood (PML) plug-in approach is optimal in property estimation for all ranges of ε, and confirms a conjecture in [Han and Shiragur, 2021] that their competitive analysis of the PML is tight.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 Introduction and Main Results

Given n i.i.d. samples drawn from a discrete distribution $p = (p_1, \cdots, p_k)$ of support size k, the problem of symmetric (or permutation-invariant) property estimation is to estimate the following quantity

$$F(p) = \sum_{i=1}^{k} f(p_i)$$

or its variants within a small additive error, for a given function $f : [0,1] \to \mathbb{R}$. This is a fundamental problem in computer science and statistics with applications in neuroscience [Rieke et al., 1999], physics [Vinck et al., 2012], ecology [Chao, 1984, Chao and Lee, 1992, Bunge and Fitzpatrick, 1993, Colwell et al., 2012], and others [Plotkin and Wyner, 1996, Porta et al., 2001].

Over the past decade, there are two main lines of research towards the symmetric property estimation. The first line of research aims to work out the minimax estimation rate and construct the minimax rate-optimal estimators for a given property, including entropy [Paninski, 2003, 2004, Valiant and Valiant, 2011a, Jiao et al., 2015, Wu and Yang, 2016], support size [Valiant and Valiant, 2013, Wu and Yang, 2019], support coverage [Orlitsky et al., 2016, Zou et al., 2016], distance to uniformity [Valiant and Valiant, 2011b, Jiao et al., 2018], sorted ℓ_1 distance [Valiant and Valiant, 2011b, Han et al., 2018], Rényi entropy [Acharya et al., 2014, 2017b], nonparametric functionals [Han et al., 2020c,a], and many others. One of the main findings in these work is that, plugging the empirical distribution into the property often leads to a strictly suboptimal estimator, especially when the function f has some non-smooth parts. They also provided general recipes for the construction

of minimax rate-optimal estimators, while the detailed construction crucially depends on the specific property in hand (i.e. classify the smooth and non-smooth parts of f, and apply different procedures).

The other line of research aims to achieve a more ambitious goal: find an adaptive (or unified) estimator that achieves the optimal sample complexity for all (or most of) the above symmetric properties. Specifically, the learner aims to obtain a unified distribution estimator $\widehat{p}$ of the true distribution p independent of the property F in hand, and hopes that the plug-in estimator $F(\widehat{p})$ is minimax rate-optimal in estimating $F(p)$ for a large class of properties F. This goal may sound too good to be true, for at least two reasons:

- as shown above, the plug-in approach of the empirical distribution, possibly the most natural choice of $\widehat{p}$, does not give the rate-optimal estimator;
- the construction of the optimal estimator even for known F is typically quite involved.

However, surprising recent developments show that there does exist such an estimator $\widehat{p}$, and there are even multiple such estimators. One estimator is the *local moment matching* (LMM) estimator in [Han et al., 2018] (and its refinement in [Han and Shiragur, 2021]), which is minimax rate-optimal in estimating the true distribution p up to permutation. Moreover, plugging the LMM estimator into the entropy, power sum function, support size, and all 1-Lipschitz functionals attains the optimal sample complexity for the respective properties within any accuracy $\varepsilon \gg n^{-1/3}$. Another estimator is the *profile maximum likelihood* (PML) estimator proposed in [Orlitsky et al., 2004], whose statistical performance was analyzed in [Acharya et al., 2017a] via a competitive analysis with an amplification factor $\exp(3\sqrt{n})$ of the error probability; this factor was later improved to $\exp(c' n^{1/3+c})$ for any $c > 0$ in [Han and Shiragur, 2021]. Consequently, for a large class of symmetric properties F where there exists a sample-optimal estimator with a sub-Gaussian error probability $\exp(-cn\varepsilon^2)$, the above analyses imply that the PML plug-in approach is also adaptively optimal within any accuracy parameter $\varepsilon \gg n^{-1/3}$.

These adaptive estimators, albeit promising, still leave some questions. Specifically, we notice the following discrepancy: the estimators constructed in the property-specific manner could achieve the optimal sample complexity for the entire accuracy regime $\varepsilon \gg n^{-1/2}$, while both adaptive estimators above are shown to be optimal only when $\varepsilon \gg n^{-1/3}$. This discrepancy leaves alone the following important question:

Is there a fundamental limit for general adaptive approaches of property estimation in the high-accuracy regime where $n^{-1/2} \ll \varepsilon \ll n^{-1/3}$?

Note that there are three possible answers to this question: first, this high-accuracy regime is uncovered simply due to an artifact of the analyses for the above adaptive estimators, and a better theoretical guarantee may be possible. Second, there may exist another fully adaptive estimator which is currently missing. Third, this high-accuracy regime may be a fundamental burden for any adaptive estimator. Specializing this question to the PML, [Acharya et al., 2017a] conjectured that "the PML based approach is indeed optimal for all ranges of ε", while [Han and Shiragur, 2021] conjectured that $\varepsilon \gg n^{-1/3}$ is the best possible range for the PML to be adaptively optimal. However, even for the PML, which is a specific choice of an adaptive estimator, the lower bound analysis is missing.

In this paper, we show that the latter conjecture is true even for general adaptive estimation: there is a phase transition for the performance of adaptive estimators at the accuracy parameter $\varepsilon \asymp n^{-1/3}$, while beyond this point, there is an unavoidable price that *any* adaptive estimator needs to pay on the sample complexity. In other words, for a reasonable family of symmetric properties, although property-specific approaches are optimal for the full accuracy range $\varepsilon \gg n^{-1/2}$, any adaptive approach fails to achieve the optimal sample complexity for at least *one* of the properties if $\varepsilon \ll n^{-1/3}$. Specifically, our main contributions are as follows:

1. We prove the first tight adaptation lower bound for the class of all 1-Lipschitz properties. We show that although the sample complexity for each 1-Lipschitz property is at most $O(k/(\varepsilon^2 \log k))$ for any $\varepsilon \gg n^{-1/2}$, under a mild assumption, any adaptive estimator must incur a sample complexity at least $\Omega(k/\varepsilon^2)$ for every $\varepsilon \ll n^{-1/3}$.
2. As a corollary, we obtain a tight competitive

analysis for the PML plug-in approach. Specifically, we show that the amplification factor of the error probability in the PML competitive analysis is at least $\exp(\Omega(n^{1/3-c}))$ for every $c > 0$, resolving the tightness conjecture of the upper bound $\exp(O(n^{1/3+c}))$ for every $c > 0$ in [Han and Shiragur, 2021].

3. We consider a new class of adaptive estimation problems, where we aim to adapt to a family of loss functions instead of the parameter sets in the traditional setting. We propose a generalized Fano's inequality to establish the adaptation lower bound for the new problem, which could be of independent interest.

1.1 Notations

Throughout the paper we adopt the following notations. Let $\mathbb{N}$ be the set of all positive integers, and for $n \in \mathbb{N}$, let $[n] \triangleq \{1, 2, \ldots, n\}$. For a finite set A, let $|A|$ be the cardinality of A. For $k \in \mathbb{N}$, let $\mathcal{M}_k$ be the set of all discrete distributions supported on $[k]$. For random variables X and Y with joint distribution P_{XY}, let

$$I(X;Y) = \int dP_{XY} \log \frac{dP_{XY}}{dP_X \otimes dP_Y}$$

be the mutual information between X and Y. For $p \in \mathcal{M}_k$, let $\mathbb{P}_p$ and $\mathbb{E}_p$ denote the probability and expectation taken with respect to the i.i.d. samples $X_1, \cdots, X_n \sim p$, respectively. For non-negative sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ (or $a_n = O(b_n)$) to denote that $\limsup_{n\to\infty} a_n/b_n < \infty$, and $a_n \gtrsim b_n$ (or $a_n = \Omega(b_n)$) to denote $b_n \lesssim a_n$, and $a_n \asymp b_n$ (or $a_n = \Theta(b_n)$) to denote both $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We also write $a_n \ll b_n$ to denote that $\limsup_{\varepsilon\to 0} \limsup_{n\to\infty} n^{\varepsilon} a_n/b_n = 0$, and $a_n \gg b_n$ to denote $b_n \ll a_n$.

1.2 Organization

The rest of the paper is organized as follows. Section 2 presents the main adaptation lower bound, together with its implication on PML. In particular, Section 2.3 introduces the new adaptive estimation problem and a generalized Fano's inequality. Section 3 compares our setting and results with an extensive set of prior work. Conclusions and open problems are drawn in Section 4, and the detailed proofs are relegated to the supplementary material.

2 Main Results

This section presents the main results of this paper. In Section 2.1, we introduce the problem setup and state the main adaptation lower bound. In Section 2.2, we review the background and obtain the tight statistical analysis of the PML. In Section 2.3, we introduce the general adaptive estimation problem and present the high-level idea behind the adaptation lower bounds.

2.1 Adaptation Lower Bound

To state the main adaptation lower bound, we first need to define the family of adaptive estimators as well as the family of symmetric property estimation problems in which we aim to achieve adaptation. In this paper, we are interested in characterizing the following *adaptive minimax risk*:

$$R^{\star}_{\text{adaptive}}(n,k) \triangleq \inf_{\widehat{p}} \sup_{F\in\mathcal{F}_{\text{Lip}}} \sup_{p\in\mathcal{M}_k} \mathbb{E}_p |F(\widehat{p}) - F(p)|, \tag{1}$$

where $\mathcal{F}_{\text{Lip}}$ denotes the class of all 1-Lipschitz properties F expressed as $F(p) = \sum_{i=1}^{k} f(p_i)$ with some 1-Lipschitz function $f : [0,1] \to \mathbb{R}$, i.e. $|f(x) - f(y)| \le |x - y|$ for all $x, y \in [0,1]$. Specifically, (1) requires the learner to obtain a single distribution estimator $\widehat{p} = (\widehat{p}_1, \cdots, \widehat{p}_k)$ solely based on the observations X^n, and then use the plug-in estimator $F(\widehat{p})$ to estimate the property $F(p)$. This is exactly the adaptive estimation framework used in [Acharya et al., 2017a, Han et al., 2018]. To measure the performance of the adaptive estimator, we consider the expected estimation error $\mathbb{E}_p|F(\widehat{p}) - F(p)|$ for the worst-case discrete distribution $p \in \mathcal{M}_k$ *and* the worst-case 1-Lipschitz property $F \in \mathcal{F}_{\text{Lip}}$. In other words, we aim to find a single plug-in estimator to perform well on estimating all 1-Lipschitz properties.

Before characterizing the adaptive minimax risk (1), we need the following mild technical assumption on the single distribution estimator $\widehat{p}$.

Assumption 1. *For each $n, k \in \mathbb{N}$, we assume that the distribution estimator $\widehat{p}(X^n) = (\widehat{p}_1, \cdots, \widehat{p}_k)$ satisfies (where $\mathcal{S}_k$ denotes the permutation group over $[k]$)*

$$\sup_{p\in\mathcal{M}_k} \mathbb{E}_p \left[\min_{\sigma\in\mathcal{S}_k} \sum_{i=1}^{k} |\widehat{p}_{\sigma(i)} - p_i| \right] \le A(n) \cdot \sqrt{\frac{k}{n}},$$

with $A(n) \ll n^\delta$ for every $\delta > 0$. We will use $\mathcal{P}$ to denote the class of all such estimators $\widehat{p}$.

Assumption 1 essentially requires that the single distribution estimator $\widehat{p}$ used in the adaptive approach must be a *reasonably* good estimator of the true distribution p up to permutation, where the term *reasonably* means that the estimator cannot be much worse than the empirical estimator. We provide three reasons why we believe this assumption to be mild. First, it is very natural to expect or require that a good distribution estimator used in the adaptive approach should be sound not only after being plugged into various properties, but also *before* the plug-in in terms of the (sorted) distribution estimation. In other words, Assumption 1 could be treated as an additional requirement for any sound adaptive approach. Second, Assumption 1 holds for all known estimators. For example, the empirical distribution satisfies Assumption 1 with $A(n) \equiv 1$ (see, e.g. [Han et al., 2015]), and both known adaptive estimators, i.e. LMM and PML, also belong to $\mathcal{P}$ with $A(n) = \text{polylog}(n)$ (cf. [Han et al., 2018] for the LMM, and the proof of Theorem 2 for the PML). Hence, restricting to the estimator class $\mathcal{P}$ still leads to novel and interesting lower bounds for these known estimators. Third, a larger quantity $A(n)$ in Assumption 1 only shrinks the accuracy regime from $\varepsilon \ll n^{-1/3}$ to $\varepsilon \ll (nA(n))^{-1/3}$, but does not affect the claimed minimax lower bound in the new accuracy regime. In addition, we remark that Assumption 1 is mostly a technical assumption, and conjecture that the following Theorem 1 still holds without it.

Restricting to the estimator class $\mathcal{P}$, the following theorem characterizes the tight adaptive minimax rate for 1-Lipschitz property estimation.

Theorem 1. *For each $n, k \in \mathbb{N}$, it holds that*

$$\inf_{\widehat{p}\in\mathcal{P}} \sup_{F\in\mathcal{F}_{\text{Lip}}} \sup_{p\in\mathcal{M}_k} \mathbb{E}_p|F(\widehat{p}) - F(p)| \asymp \begin{cases} \sqrt{\frac{k}{n\log n}} & \text{if } n^{1/3} \ll k \lesssim n\log n, \\ \sqrt{\frac{k}{n}} & \text{if } 1 \ll k \ll n^{1/3}. \end{cases}$$

Theorem 1 can also be equivalently formulated in terms of the optimal sample complexity.

Corollary 1. *It is sufficient and necessary to have $n = \Theta(k/(\varepsilon^2 \log k))$ samples for the existence of an adaptive estimator in $\mathcal{P}$ to estimate all 1-Lipschitz properties within error ε if $\varepsilon \gg n^{-1/3}$, and it is sufficient and necessary to have $n = \Theta(k/\varepsilon^2)$ samples for the existence of an adaptive estimator in $\mathcal{P}$ to estimate all 1-Lipschitz properties within error ε if $n^{-1/2} \ll \varepsilon \ll n^{-1/3}$.*

Let us appreciate the result of Theorem 1 via some comparisons with following known results. First, there is no phase transition in the high-accuracy regime if we do *not* require an adaptive estimator. Specifically, the following non-adaptive minimax risk was shown in [Hao and Orlitsky, 2019b]:[1]

$$\sup_{F\in\mathcal{F}_{\text{Lip}}} \inf_{\widehat{p}} \sup_{p\in\mathcal{M}_k} \mathbb{E}_p|F(\widehat{p}) - F(p)| \asymp \sqrt{\frac{k}{n\log n}}, \quad 1 \ll k \lesssim n\log n. \tag{2}$$

Comparing Theorem 1 and (2), we observe that after a simple swap of the infimum and supremum, the adaptive minimax risk becomes different from the non-adaptive minimax risk and exhibits an elbow effect at $k \asymp n^{1/3}$ (or equivalently $\varepsilon \asymp n^{-1/3}$). In particular, there is a *strict separation* between the best achievable errors for adaptive and non-adaptive approaches, and the learner need to pay a strict penalty on the estimation error to achieve adaptation below the accuracy level $\varepsilon \ll n^{-1/3}$.

Second, we also compare Theorem 1 with the problem of estimating sorted distribution, where [Han et al., 2018] shows that

$$\inf_{\widehat{p}} \sup_{p\in\mathcal{M}_k} \mathbb{E}_p\left[\sup_{F\in\mathcal{F}_{\text{Lip}}} |F(\widehat{p}) - F(p)|\right] \asymp \begin{cases} \sqrt{\frac{k}{n\log n}} & \text{if } n^{1/3} \ll k \lesssim n\log n, \\ \sqrt{\frac{k}{n}} & \text{if } 1 \ll k \ll n^{1/3}. \end{cases} \tag{3}$$

As $\mathbb{E}[\sup_n X_n] \geq \sup_n \mathbb{E}[X_n]$, the quantity in (3) is no smaller than our adaptive minimax risk in (1), and thus implies the upper bound in Theorem 1. However, the lower bound of Theorem 1 is the most challenging part and stronger than what (3) gives. Comparing the results of Theorem 1 and (3), we remark that the phase transition in (3) characterizes a shift in the estimation difficulty of *a specific problem*, while the phase transition in Theorem 1 characterizes the same shift *only for adaptive approaches*. Therefore, the former transition could be

[1] The original paper did not require to use a plug-in estimator, but any estimator $\widehat{F}$ could be clipped to the range of $F(\cdot)$ and written as $F(\widehat{p})$ for some $\widehat{p} \in \mathcal{M}_k$.

derived by studying different regimes of the problem, while the latter transition crucially requires to also take into account the special nature of the adaptive approach. Technically, we remark that after exchanging the expectation and supremum, the lower bound argument will become fundamentally different, and the traditional approaches fail to give the tight adaptive lower bound (cf. Section 2.3 for more details).

2.2 Lower Bound of PML

The general adaptive lower bound of Theorem 1 also gives tight and non-trivial lower bounds for some known adaptive approaches. For example, for the LMM adaptive approach, Theorem 1 shows that the condition $\varepsilon \gg n^{-1/3}$ required in [Han et al., 2018] for its optimality in property estimation is not superfluous, but in general unavoidable. The implication for the PML adaptive approach [Orlitsky et al., 2004] is even more surprising; to fully describe it we provide a brief review of PML.

Given n i.i.d. observations $X_1, \cdots, X_n$ drawn from a discrete distribution p supported on the domain $[k]$, the *profile* of the observations is defined as a vector $\phi = (\phi_0, \cdots, \phi_n)$ with ϕ_i being the number of domain elements $j \in [k]$ which appear exactly i times in the sample. For example, ϕ_0 is the number of unseen elements, and ϕ_1 is the number of unique elements, i.e. appearing exactly once. Let $\Phi_{n,k}$ be the set of all possible profiles with n observations and support size k. Note that for any $\phi \in \Phi_{n,k}$ and $p \in \mathcal{M}_k$, we could compute the probability that the resulting profile is ϕ under true i.i.d. distribution p, denoted by $\mathbb{P}(p, \phi)$. The *profile maximum likelihood* (PML) distribution estimator is then defined as

$$p^{\mathrm{PML}}(\phi) = \arg\max_{p \in \mathcal{M}_k} \mathbb{P}(p, \phi).$$

In other words, upon observing the profile ϕ, the PML estimator is the discrete distribution which maximizes the probability of observing the given profile ϕ. This estimator is interesting in several aspects. From the optimization side, the probability $\mathbb{P}(p, \phi)$ is a highly non-convex function of p, and it is very challenging to compute the exact or approximate PMLs. From the statistical side, as $\mathbb{P}(p, \phi)$ does not admit an additive form even under i.i.d. models (unlike the traditional log-likelihood), even first-order asymptotic properties are challenging to establish for the PML. After 13 years of its invention, a useful statistical property of the PML was established in [Acharya et al., 2017a] in terms of an interesting *competitive analysis*: for every property F and accuracy parameter ε, it holds that

$$\begin{aligned}&\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|F(p^{\mathrm{PML}}) - F(p)| \geq 2\varepsilon) \\ &\leq \exp(3\sqrt{n}) \cdot \inf_{\widehat{F}} \sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|\widehat{F} - F(p)| \geq \varepsilon), \quad (4)\end{aligned}$$

where the infimum is taken over all possible estimators $\widehat{F}$ depending only on ϕ, which is a natural class as the label information of samples is not needed in symmetric property estimation. Specifically, (4) gives an indirect statistical analysis of the PML plug-in approach which depends on the performance of another estimator. For many properties (such as all 1-Lipschitz properties), the minimax error probability in the RHS of (4) behaves as $\exp(-\Omega(n\varepsilon^2))$ when n exceeds the optimal sample complexity, thus (4) shows that the PML plug-in approach is adaptively optimal for $\varepsilon \gg n^{-1/4}$. The proof of (4) used only the defining property of PML in a delicate way, and the error amplification factor $\exp(3\sqrt{n})$ follows from a simple union bound over the profiles with cardinality $|\Phi_{n,k}| \leq \exp(3\sqrt{n})$.

The paper [Acharya et al., 2017a] asked whether the above error amplification factor $\exp(3\sqrt{n})$ could be improved in general; three years later [Han and Shiragur, 2021] provided an affirmative answer. Specifically, using a chaining property of the PML distributions, [Han and Shiragur, 2021] showed the following improvement

$$\begin{aligned}&\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|F(p^{\mathrm{PML}}) - F(p)| \geq (2 + o(1))\varepsilon) \\ &\leq \exp(c' n^{1/3+c}) \cdot \left(\inf_{\widehat{F}} \sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|\widehat{F} - F(p)| \geq \varepsilon) \right)^{1-c} \quad (5)\end{aligned}$$

for any absolute constant $c > 0$ and some $c' > 0$ depending only on c. Using (5), the accuracy range for the optimality of PML could be improved to $\varepsilon \gg n^{-1/3}$ for the aforementioned properties. It was also conjectured in [Han and Shiragur, 2021] that the new amplification factor in (5) is tight, but little intuition was provided.

Surprisingly, without directly analyzing the PML adaptive approach, Theorem 1 implies the tightness of the error amplification factor in (5), as summarized in our next main theorem.

Theorem 2. *For any given constants $C > 0, c_1 \in$*

$(0, 1/3)$ *and* $c_2 \in (0, 1)$*, it holds that*

$$\liminf_{n\to\infty} n^{-(1/3-c_1)} \cdot \sup_{F\in\mathcal{F}_{\mathrm{Lip}}} \sup_{k,\varepsilon>0} \log \frac{\sup_{p\in\mathcal{M}_k} \mathbb{P}_p(|F(p^{\mathrm{PML}}) - F(p)| \geq C\varepsilon)}{\left(\inf_{\widehat{F}} \sup_{p\in\mathcal{M}_k} \mathbb{P}_p(|\widehat{F} - F(p)| \geq \varepsilon)\right)^{1-c_2}} = +\infty.$$

After some algebra, it is clear that Theorem 2 rules out the possibility that the exponent $O(n^{1/3+c})$ of the amplification factor in (5) could be improved to $O(n^{1/3-c})$ in general. Therefore, Theorem 2 implies that the general competitive analysis of the PML in [Han and Shiragur, 2021] is essentially tight, thereby resolves the conjecture therein.

We provide two additional remarks on Theorem 2. First, the validity of Theorem 2 is irrelevant to Assumption 1, as the PML estimator is an instance which satisfies Assumption 1. Second, the lower bound in Theorem 2 does *not* rule out the possibility that the PML adaptive approach could be fully optimal for *some* property. For example, it was shown in [Charikar et al., 2019] that the PML plug-in approach is fully optimal in estimating the support size. It will be an understanding open question to provide a tight analysis of the PML estimator for specific properties; see also Section 4.

2.3 Generalized Fano's Inequality

The idea to establish the adaptation lower bound in Theorem 1 is useful for a general class of adaptive estimation problems, which we present in this section. We first recall the general decision-theoretic setup [Wald, 1950]. Let $(P_\theta)_{\theta\in\Theta}$ be a generic statistical model with parameter set Θ, and $\mathcal{A}$ be the space of all possible actions the learner could take. In other words, the learner obtains an observation $X \sim P_\theta$ with some unknown θ, and then maps X to a random action $a(X) \in \mathcal{A}$. Let $L : \Theta \times \mathcal{A} \to \mathbb{R}_+$ be any (measurable) loss function, the problem of *minimax estimation* is to characterize the following minimax risk:

$$R^\star(\Theta, \mathcal{A}, L) = \inf_a \sup_{\theta\in\Theta} \mathbb{E}_\theta[L(\theta, a(X))]. \tag{6}$$

Similarly, the problem of *adaptive minimax estimation* with respect to a class of loss functions $\mathcal{L}$ is to characterize the following adaptive minimax risk:

$$R^\star(\Theta, \mathcal{A}, \mathcal{L}) = \inf_a \sup_{\theta\in\Theta} \sup_{L\in\mathcal{L}} \mathbb{E}_\theta[L(\theta, a(X))]. \tag{7}$$

To see how our problem (1) is a special instance of (7), we could set P_θ to be the distribution of n i.i.d. samples from the discrete distribution θ, with $\Theta = \mathcal{M}_k$. Moreover, $\mathcal{A} = \mathcal{M}_k$, $L_F(\theta, a) = |F(\theta) - F(a)|$, and $\mathcal{L} = \{L_F : F \text{ is a 1-Lipschitz property}\}$.

There are several well-known tools to establish the lower bound of (6), where a standard and prominent tool is the reduction to hypothesis testing problems; see, e.g. [Yu, 1997, Tsybakov, 2009]. The main step is to find $\theta_1, \cdots, \theta_M \in \Theta$ such that both the *separation condition* and the *indistinguishability condition* hold: the separation condition typically requires that $\inf_{a\in\mathcal{A}}[L(\theta_i, a) + L(\theta_j, a)] \geq \Delta$ for some separation parameter $\Delta > 0$ and all $i \neq j$, and the indistinguishability condition essentially states that any learner could not determine the true parameter θ_i based on her observations if the truth $i \in [M]$ is chosen uniformly at random. Then it might be tempting to think that one only needs to replace $L(\theta, a)$ by $\sup_{L\in\mathcal{L}} L(\theta, a)$ in the above arguments to lower bound (7). However, this approach will place the supremum in L inside the expectation in (7), and thus provide a lower bound for a larger quantity like (3). An alternative way is to use the trivial inequality $R^\star(\Theta, \mathcal{A}, \mathcal{L}) \geq \sup_{L\in\mathcal{L}} R^\star(\Theta, \mathcal{A}, L)$ and then lower bound the latter quantity. Although this gives a valid lower bound, it is not strong enough in our problem where $R^\star(\Theta, \mathcal{A}, \mathcal{L}) \gg \sup_{L\in\mathcal{L}} R^\star(\Theta, \mathcal{A}, L)$ in view of Theorem 1 and (2).

The main idea to fix the above difficulty is that in addition to choose M points $\theta_1, \cdots, \theta_M \in \Theta$ corresponding to different statistical models, we also find M different loss functions $L_1, \cdots, L_M \in \mathcal{L}$ tailored for the respective models. Specifically, the indistinguishability condition is unchanged as it depends only on $\theta_1, \cdots, \theta_M$, while the separation condition could be replaced by $\inf_{a\in\mathcal{A}}[L_i(\theta_i, a) + L_j(\theta_j, a)] \geq \Delta$ for all $i \neq j$. Motivated by this idea, we propose the following version of the Fano's inequality.

Lemma 1 (Generalized Fano's Inequality)**.** *In the above decision-theoretic setup, suppose that* $\theta_1, \cdots, \theta_M \in \Theta$ *and* $L_1, \cdots, L_M \in \mathcal{L}$ *are chosen. Assume that there exists* $\mathcal{A}_0 \subseteq \mathcal{A}$ *such that*

$$\inf_{a\in\mathcal{A}_0} [L_i(\theta_i, a) + L_j(\theta_j, a)] \geq \Delta > 0, \quad \forall i \neq j \in [M],$$

and an estimator $a(X)$ *satisfies that*

$$P_{\theta_i}(a(X) \in \mathcal{A}_0) \geq p_{\min} > 0$$

for all $i \in [M]$. Then for this estimator we have

$$\sup_{\theta\in\Theta}\sup_{L\in\mathcal{L}} \mathbb{E}_\theta[L(\theta, a(X))] \geq \frac{\Delta}{2}\left(p_{\min} - \frac{I(U;X) + p_{\min}\log 2}{\log M}\right),$$

where $I(U;X)$ denotes the mutual information between $U \sim \mathrm{Uniform}([M])$ and $X \mid U \sim P_{\theta_U}$.

Lemma 1 gives a general lower bound on the adaptive minimax risk under a *soft* separation condition, where the inequality $L_i(\theta_i, a) + L_j(\theta_j, a) \geq \Delta$ holds with a positive probability rather than everywhere. Note that when $L_i \equiv L$ and $p_{\min} = 1$, Lemma 1 reduces to the traditional Fano's inequality [Cover and Thomas, 2006]. Despite the simplicity of the idea underlying Lemma 1, we show in the supplementary material that it gives the desired lower bound for adaptive property estimation.

3 Related Work

3.1 Property Estimation

There has been a rich line of research towards the optimal estimation of properties (or functionals) of high-dimensional parameters, especially in the past decade. Starting from some early work [Lepski et al., 1999, Paninski, 2003, 2004, Cai and Low, 2011, Valiant and Valiant, 2011a,b, 2013], the fully minimax rate-optimal estimators in all accuracy regimes were obtained for the Shannon entropy in [Jiao et al., 2015, Wu and Yang, 2016]. They also provided general recipes for both the estimator construction and tight minimax lower bounds. Specifically, the crux of the optimal estimator construction lies in the classification of smooth and non-smooth regimes and the usage of polynomial approximation to reduce bias in the non-smooth regime, and the minimax lower bound relies on the duality between moment matching and best polynomial approximation. Since then, these general recipes together with their non-trivial extensions have been applied to various other properties, e.g. the Rényi entropy [Acharya et al., 2014, 2017b], support size [Wu and Yang, 2019], support coverage [Orlitsky et al., 2016, Zou et al., 2016, Polyanskiy and Wu, 2019], distance to uniformity [Jiao et al., 2018], general 1-Lipschitz property [Hao and Orlitsky, 2019a,b], L_1 distance [Jiao et al., 2018], KL divergence [Bu et al., 2018, Han et al., 2020b], and nonparametric functionals [Han et al., 2020a,c]. We refer to the survey [Verdú, 2019] for an overview of these results. There is also another line of recent work on estimating a population of parameters or distribution under a Wasserstein distance, a problem closely related to property estimation, via projection-based methods without explicit polynomial approximation [Kong and Valiant, 2017, Tian et al., 2017, Han et al., 2018, Rigollet and Weed, 2019, Vinayak et al., 2019b,a, Wu and Yang, 2020, Jana et al., 2020]. While the above work completely characterized the complexity of many *given problems* in property estimation, the complexity of adaptive estimation in *a family of such problems* is largely missing. For example, the $\Omega(\sqrt{k/(n\log n)})$ lower bound for large k in Theorem 1 simply follows from the complexity of estimating a particular 1-Lipschitz property, but the main $\Omega(\sqrt{k/n})$ lower bound for small k becomes the crucial complexity of adaptive approaches and thus does not follow from the above set of results or tools.

3.2 Adaptive Property Estimation.

More recently the problem of adaptive, or unified, property estimation has drawn several research attention. As reviewed in the introduction, possibly the most well-known adaptive approach is the PML plug-in approach, with early statistical developments in [Orlitsky et al., 2004, 2011, Anevski et al., 2017]. Since [Acharya et al., 2017a] provided the first competitive analysis of the PML plug-in approach, there have been several follow-up papers on the statistical analysis of the PML. Some work focused on the application of the competitive analysis and the construction of the estimator achieving the minimax error probability in (4), e.g. [Hao and Orlitsky, 2019a]. Some work focused on proper modifications of the PML to achieve better adaptation, e.g. [Hao and Orlitsky, 2019a, Charikar et al., 2019]; however, these modified distribution estimators will depend on the target property and are thus not fully unified. Other work aimed to improve the competitive analysis in [Acharya et al., 2017a]; for example, [Hao and Orlitsky, 2020] obtained a distribution-dependent amplification factor without changing the worst-case analysis, and [Han and Shiragur, 2021] improved this factor to $\exp(O(n^{1/3+c}))$ in general. However, none of the above work studied the limitation of the PML plug-in approach, even for concrete examples. There-

fore, the lower bound analysis, especially the possible separation compared with the optimal estimator, of the PML is missing.

Another adaptive approach plugs in the LMM estimator proposed in [Han et al., 2018]. Different from the general competitive analysis of PML, the performance of the LMM approach could be directly analyzed for given properties based on its moment matching performance in each local interval. Built on the LMM performance analysis in estimating entropy, power sum function, and support size, the authors of [Han et al., 2018] commented that the LMM pays some penalty for being a unified approach. However, this comment was only an insight, and there was no lower bound to support it rigorously. The current work fills in this gap and shows that the price observed for the LMM is in fact unavoidable even for general adaptive approaches.

3.3 Adaptation Lower Bound

We also review and compare with some known tools to establish adaptation lower bounds, mainly taken from the statistics literature. Adaptation is an important topic in statistics; for example, in nonparametric estimation one may aim to design a density estimator adapting to different smoothness parameters, or in hypothesis testing one may wish to propose an adaptive test procedure against several different alternatives. However, for some problems the adaptation could be achieved without paying any penalty (e.g. density estimation [Lepskii, 1992, Donoho et al., 1995], L_r norm estimation with non-even r [Han et al., 2020a]), while for others some adaptation penalties are inevitable (e.g. estimating linear [Efromovich and Low, 1994] or quadratic [Efromovich and Low, 1996] functional of densities). The main technical tool to establish tight penalties of adaptation is the *constrained risk inequality* originally developed in [Brown and Low, 1996] and generalized in [Cai and Low, 2011, Duchi and Ruan, 2018]. Roughly speaking, this type of inequality asserts that if an estimator achieves a too small error at one point, it must incur a too large error at another point; therefore, adaptation may incur a penalty as it might be required to adapt to easier problems and achieve a too small error. For testing, there is also another approach to establish adaptation lower bounds, where the key is to use a mixture of different alternative distributions which could be closer to the null than any fixed alternative; see [Spokoiny, 1996] and also [Giné and Nickl, 2016, Chapter 8] for examples.

However, we remark that our adaptive estimation problem in (7) is fundamentally different. In the previous work, the target of adaptive estimation is to adapt to different (usually a nested class of) *parameter sets*, e.g. Hölder balls with different smoothness parameters. Mathematically, the target is to characterize the following optimal penalty for adaptation:

$$\begin{aligned}&\mathrm{Pen}^\star(\{\Theta_m\}_{m=1}^\infty, \mathcal{A}, L)\\ &\triangleq \inf_a \sup_{m\geq 1} \frac{\sup_{\theta\in\Theta_m} \mathbb{E}_\theta[L(\theta, a(X))]}{\inf_{a_m} \sup_{\theta\in\Theta_m} \mathbb{E}_\theta[L(\theta, a_m(X))]},\end{aligned}$$

where $\{\Theta_m\}_{m=1}^\infty$ is the class of parameter sets to which one aims to adapt. In contrast, in (7) we consider a fixed parameter set, but wish to adapt to different *loss functions* for the final estimator. Establishing adaptation lower bounds for different losses is novel to our knowledge, and the previous tools are not applicable in this problem. Consequently, we aim to provide useful tools (e.g. Lemma 1) for this new adaptation problem, and expect them to be a helpful addition to the literature on adaptive estimation.

4 Conclusion and Open Problems

In this paper we showed that there is a high-accuracy limitation for general adaptive approaches of property estimation, which in turn implied tight lower bounds for the known adaptive approaches such as the PML and LMM. A number of directions could be of interest. First, we believe that Assumption 1 is an artifact of our proof and unnecessary for Theorem 1 to hold, and a better choice of the loss functions in Lemma 1 could remove this assumption. Second, the adaptation lower bound for PML does not rule out the possibility that PML could be fully optimal for *certain properties*. However, to show this, one need to go beyond the competitive analysis of the PML and seek for additional properties. Third, our current lower bound for PML only shows the existence of a property requiring $\varepsilon \gg n^{-1/3}$ for the PML to be optimal, and it is interesting to construct such a property explicitly.

Acknowledgement

Yanjun Han is grateful to Kirankumar Shiragur for helpful discussions, and four anonymous reviewers for their valuable comments to improve the presentation of this paper.

References

Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. The complexity of estimating rnyi entropy. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2014.

Jayadev Acharya, Hirakendu Das, Alon Orlitsky, and Ananda Theertha Suresh. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *International Conference on Machine Learning*, pages 11–21, 2017a.

Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. Estimating renyi entropy of discrete distributions. *IEEE Trans. Inf. Theor.*, 63(1):38–56, January 2017b. ISSN 0018-9448. doi: 10.1109/TIT.2016.2620435. URL https://doi.org/10.1109/TIT.2016.2620435.

Dragi Anevski, Richard D Gill, and Stefan Zohren. Estimating a probability mass function with unknown labels. *The Annals of Statistics*, 45(6): 2708–2735, 2017.

Lawrence D Brown and Mark G Low. A constrained risk inequality with applications to nonparametric functional estimation. *The annals of Statistics*, 24(6):2524–2535, 1996.

Yuheng Bu, Shaofeng Zou, Yingbin Liang, and Venugopal V Veeravalli. Estimation of kl divergence: Optimal minimax rate. *IEEE Transactions on Information Theory*, 64(4):2648–2674, 2018.

John Bunge and Michael Fitzpatrick. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88(421): 364–373, 1993.

T Tony Cai and Mark G Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

A Chao. Nonparametric estimation of the number of classes in a population. scandinavianjournal of statistics11, 265-270. *Chao26511Scandinavian Journal of Statistics1984*, 1984.

Anne Chao and Shen-Ming Lee. Estimating the number of classes via sample coverage. *Journal of the American statistical Association*, 87(417): 210–217, 1992.

Moses Charikar, Kirankumar Shiragur, and Aaron Sidford. A general framework for symmetric property estimation. In *Advances in Neural Information Processing Systems*, pages 12426–12436, 2019.

Robert K Colwell, Anne Chao, Nicholas J Gotelli, Shang-Yi Lin, Chang Xuan Mao, Robin L Chazdon, and John T Longino. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of plant ecology*, 5(1):3–21, 2012.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, New York, second edition, 2006.

David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society: Series B (Methodological)*, 57 (2):301–337, 1995.

John C Duchi and Feng Ruan. A constrained risk inequality for general losses. *arXiv preprint arXiv:1804.08116*, 2018.

Sam Efromovich and Mark Low. On optimal adaptive estimation of a quadratic functional. *The Annals of Statistics*, 24(3):1106–1125, 1996.

Sam Efromovich and Mark G Low. Adaptive estimates of linear functionals. *Probability theory and related fields*, 98(2):261–275, 1994.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2016.

Yanjun Han and Kirankumar Shiragur. On the competitive analysis and high accuracy optimality of profile maximum likelihood. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1317–1336. SIAM, 2021.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions under ℓ_1 loss. *IEEE Transactions on Information Theory*, 61(11):6343–6354, 2015.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under wasserstein distance. In *Conference On Learning Theory*, pages 3189–3221, 2018.

Yanjun Han, Jiantao Jiao, and Rajarshi Mukherjee. On estimation of l_r-norms in gaussian white noise models. *Probability Theory and Related Fields*, 177(3):1243–1294, 2020a.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of divergences between discrete distributions. *IEEE Journal on Selected Areas in Information Theory*, 2020b.

Yanjun Han, Jiantao Jiao, Tsachy Weissman, and Yihong Wu. Optimal rates of entropy estimation over lipschitz balls. *Annals of Statistics*, 48(6): 3228–3250, 2020c.

Yi Hao and Alon Orlitsky. The broad optimality of profile maximum likelihood. In *Advances in Neural Information Processing Systems*, pages 10989–11001, 2019a.

Yi Hao and Alon Orlitsky. Unified sample-optimal property estimation in near-linear time. In *Advances in Neural Information Processing Systems*, pages 11104–11114, 2019b.

Yi Hao and Alon Orlitsky. Profile entropy: A fundamental measure for the learnability and compressibility of distributions. In *Advances in Neural Information Processing Systems*, volume 33, pages 6947–6958, 2020.

Soham Jana, Yury Polyanskiy, and Yihong Wu. Extrapolating the profile of a finite population. In *Conference on Learning Theory*, pages 2011–2033. PMLR, 2020.

Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the l_1 distance. *IEEE Transactions on Information Theory*, 64 (10):6672–6706, 2018.

Weihao Kong and Gregory Valiant. Spectrum estimation from samples. *The Annals of Statistics*, 45(5):2218–2247, 2017.

Oleg Lepski, Arkady Nemirovski, and Vladimir Spokoiny. On estimation of the L_r norm of a regression function. *Probability theory and related fields*, 113(2):221–253, 1999.

OV Lepskii. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992.

Alon Orlitsky, Narayana P Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. On modeling profiles instead of values. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 426–435. AUAI Press, 2004.

Alon Orlitsky, NP Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. On estimating the probability multiset. *draft manuscript, June*, 2011.

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

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

Liam Paninski. Estimating entropy on m bins given fewer than m samples. *Information Theory, IEEE Transactions on*, 50(9):2200–2203, 2004.

Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54 (10):4750–4755, 2008.

Nina T. Plotkin and Abraham J. Wyner. *An Entropy Estimator Algorithm and Telecommunications Applications*, pages 351–363. Springer Netherlands, Dordrecht, 1996. ISBN 978-94-015-8729-7. doi: 10.1007/978-94-015-8729-7_29. URL https://doi.org/10.1007/978-94-015-8729-7_29.

Yury Polyanskiy and Yihong Wu. Dualizing le cam's method, with applications to estimating the unseens. *arXiv preprint arXiv:1902.05616*, 2019.

A. Porta, S. Guzzetti, N. Montano, R. Furlan, M. Pagani, A. Malliani, and S. Cerutti. Entropy, entropy rate, and pattern classification as tools to typify complexity in short heart period variability series. *IEEE Transactions on Biomedical Engineering*, 48(11):1282–1291, Nov 2001. ISSN 0018-9294. doi: 10.1109/10.959324.

Fred Rieke, Davd Warland, Rob de Ruyter van Steveninck, and William Bialek. *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-18174-6.

Philippe Rigollet and Jonathan Weed. Uncoupled isotonic regression via minimum wasserstein deconvolution. *Information and Inference: A Journal of the IMA*, 8(4):691–717, 2019.

Vladimir Spokoiny. Adaptive and spatially adaptive testing of a nonparametric hypothesis. 1996.

Kevin Tian, Weihao Kong, and Gregory Valiant. Learning populations of parameters. In *Advances in neural information processing systems*, pages 5778–5787, 2017.

A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer-Verlag, 2009.

Gregory Valiant and Paul Valiant. Estimating the unseen: an $n/\log n$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, pages 685–694. ACM, 2011a.

Gregory Valiant and Paul Valiant. The power of linear estimators. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 403–412. IEEE, 2011b.

Paul Valiant and Gregory Valiant. Estimating the unseen: improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems*, pages 2157–2165, 2013.

Sergio Verdú. Empirical estimation of information measures: A literature guide. *Entropy*, 21(8): 720, 2019.

Ramya Korlakai Vinayak, Weihao Kong, and Sham M Kakade. Optimal estimation of change in a population of parameters. *arXiv preprint arXiv:1911.12568*, 2019a.

Ramya Korlakai Vinayak, Weihao Kong, Gregory Valiant, and Sham M Kakade. Maximum likelihood estimation for learning populations of parameters. *arXiv preprint arXiv:1902.04553*, 2019b.

Martin Vinck, Francesco P. Battaglia, Vladimir B. Balakirsky, A. J. Han Vinck, and Cyriel M. A. Pennartz. Estimation of the entropy based on its polynomial representation. *Phys. Rev. E*, 85: 051139, May 2012. doi: 10.1103/PhysRevE.85.051139. URL https://link.aps.org/doi/10.1103/PhysRevE.85.051139.

Abraham Wald. *Statistical decision functions*. Wiley, 1950.

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 47 (2):857–883, 2019.

Yihong Wu and Pengkun Yang. Optimal estimation of gaussian mixtures via denoised method of moments. *Annals of Statistics*, 48(4):1981–2007, 2020.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

James Zou, Gregory Valiant, Paul Valiant, Konrad Karczewski, Siu On Chan, Kaitlin Samocha, Monkol Lek, Shamil Sunyaev, Mark Daly, and Daniel G. MacArthur. Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nature Communications*, 7:13293 EP –, 10 2016. URL https://doi.org/10.1038/ncomms13293.

On the High Accuracy Limitation of Adaptive Property Estimation: Supplementary Materials

A Proof of Lemma 1

First, as the maximum is no smaller than the average, we have

$$\sup_{\theta\in\Theta}\sup_{L\in\mathcal{L}}\mathbb{E}_\theta[L(\theta,a(X))]\geq\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{\theta_i}[L_i(\theta_i,a(X))]. \tag{8}$$

For each $i\in[M]$, let Q_i be the conditional distribution of $a(X)$ with $X\sim P_{\theta_i}$ conditioning on the event $a(X)\in\mathcal{A}_0$. Then by the non-negativity of each L_i and definition of $p_{\min}$,

$$\mathbb{E}_{\theta_i}[L_i(\theta_i,a(X))]\geq P_{\theta_i}(a(X)\in\mathcal{A}_0)\cdot\mathbb{E}_{a\sim Q_i}[L_i(\theta_i,a)]\geq p_{\min}\cdot\mathbb{E}_{a\sim Q_i}[L_i(\theta_i,a)],$$

and therefore (8) gives

$$\sup_{\theta\in\Theta}\sup_{L\in\mathcal{L}}\mathbb{E}_\theta[L(\theta,a(X))]\geq p_{\min}\cdot\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{a\sim Q_i}[L_i(\theta_i,a)]. \tag{9}$$

The next few steps are similar to the proof of the traditional Fano's inequality. For each $a\in\mathcal{A}_0$, define a test $\Psi(a)=\arg\min_{i\in[M]}L_i(\theta_i,a)$. Then by the separation condition, we have

$$L_i(\theta_i,a)\geq\frac{L_i(\theta_i,a)+L_{\Psi(a)}(\theta_{\Psi(a)},a)}{2}\geq\frac{\Delta}{2}\cdot\mathbb{1}(\Psi(a)\neq i),\quad\forall i\in[M],a\in\mathcal{A}_0,$$

and therefore (9) gives

$$\sup_{\theta\in\Theta}\sup_{L\in\mathcal{L}}\mathbb{E}_\theta[L(\theta,a(X))]\geq\frac{\Delta p_{\min}}{2}\cdot\frac{1}{M}\sum_{i=1}^{M}Q_i(\Psi(a)\neq i)\geq\frac{\Delta p_{\min}}{2}\left(1-\frac{I(U;Y)+\log 2}{\log M}\right), \tag{10}$$

where the second inequality is due to the traditional Fano's inequality [Cover and Thomas, 2006], with $U\sim\mathrm{Uniform}([M])$ and $Y\mid U\sim Q_U$. To proceed, we introduce a few notations: let R_i be the distribution of $a(X)$ with $X\sim P_{\theta_i}$, R be the distribution of $a(X)$ with $X\sim M^{-1}\sum_{i=1}^{M}P_{\theta_i}$, and Q be the restriction of the distribution R to the set $\mathcal{A}_0$. Then

$$\begin{aligned}
I(U;Y)&\overset{\text{(a)}}{\leq}\mathbb{E}_U[D_{\mathrm{KL}}(Q_U\|Q)]\\
&\overset{\text{(b)}}{\leq}\mathbb{E}_U\left[\frac{1}{P_{\theta_U}(a(X)\in\mathcal{A}_0)}\cdot D_{\mathrm{KL}}(R_U\|R)\right]\\
&\overset{\text{(c)}}{\leq}\frac{1}{p_{\min}}\cdot\mathbb{E}_U[D_{\mathrm{KL}}(R_U\|R)]\\
&\overset{\text{(d)}}{=}\frac{I(U;a(X))}{p_{\min}}\overset{\text{(e)}}{\leq}\frac{I(U;X)}{p_{\min}},
\end{aligned}$$

where (a) is due to the variational representation of the mutual information $I(U;Y)=\min_{Q_Y}\mathbb{E}_U[D_{\mathrm{KL}}(P_{Y|U}\|Q_Y)]$, (b) follows from the data-processing property of the KL divergence $D_{\mathrm{KL}}(P\|Q)\geq P(A)\cdot D_{\mathrm{KL}}(P_{\cdot|A}\|Q_{\cdot|A})$, (c) is due to the assumption of Lemma 1, (d) is the definition of the mutual information, and (e) is the data-processing property of the mutual information. Now combining the above inequality with (10) completes the proof of Lemma 1.

B Proof of Theorem 1

Notations: For two probability measures P, Q on the same probability space, let

$$\|P-Q\|_{\mathrm{TV}} = \frac{1}{2}\int |dP - dQ|$$
$$D_{\mathrm{KL}}(P\|Q) = \int dP \log \frac{dP}{dQ}$$
$$\chi^2(P\|Q) = \int \frac{(dP-dQ)^2}{dQ}$$

be the total variation (TV) distance, the Kullback–Leibler (KL) divergence, and the χ^2-divergence between P and Q, respectively.

This section is devoted to the proof of Theorem 1. Note that the upper bound is achieved by the LMM estimator for $k \gg n^{1/3}$ and the empirical distribution for $k \ll n^{1/3}$ [Han et al., 2018][2], and the lower bound for $k \gg n^{1/3}$ follows from the minimax lower bound for estimating a specific 1-Lipschitz property, i.e. the distance to uniformity $F(p) = \sum_{i=1}^k |p_i - 1/k|$ [Jiao et al., 2018]. Therefore, it remains to prove the following adaptation lower bound:

$$\inf_{\widehat{p}\in\mathcal{P}} \sup_{F\in\mathcal{F}_{\mathrm{Lip}}} \sup_{p\in\mathcal{M}_k} \mathbb{E}_p |F(\widehat{p}) - F(p)| \gtrsim \sqrt{\frac{k}{n}}, \quad 1 \ll k \ll n^{1/3}. \tag{11}$$

Recall that to formulate our adaptive property estimation problem in the general framework of (7), we identify $\theta \in \Theta$ and $a \in \mathcal{A}$ with the distributions $p, \widehat{p} \in \mathcal{M}_k$, and $\Theta = \mathcal{A} = \mathcal{M}_k$. Moreover, the loss function is the absolute difference in the property value $L_F(p, \widehat{p}) = |F(p) - F(\widehat{p})|$, and the family of losses is $\mathcal{L} = \{L_F : F \text{ is a 1-Lipschitz property}\}$. In this section, we apply Lemma 1 to a suitable choice of distributions $p_1, \cdots, p_M \in \mathcal{M}_k$ and 1-Lipschitz properties $F_1, \cdots, F_M \in \mathcal{F}_{\mathrm{Lip}}$, and prove the target adaptation lower bound in (11).

Without loss of generality we assume that $k = 2k_0$ is an even integer. Consider the following distribution $p_0 = (p_{0,1}, \cdots, p_{0,k}) \in \mathcal{M}_k$ serving as the "center" of all hypotheses:

$$p_0 = \left(\frac{1}{2k}, \frac{1}{2k} + \frac{1}{k(k-1)}, \frac{1}{2k} + \frac{2}{k(k-1)}, \cdots, \frac{3}{2k}\right).$$

Fix a parameter

$$\delta \in \left(0, \frac{1}{4k(k-1)}\right) \tag{12}$$

to be chosen later, for each $u \in \mathcal{U} \triangleq \{\pm 1\}^{k_0}$ we also associate a distribution $p_u = (p_{u,1}, \cdots, p_{u,k}) \in \mathcal{M}_k$ with

$$p_{u,i} = p_{0,i} + u_i\delta, \quad p_{u,k_0+i} = p_{0,k_0+i} - u_i\delta, \quad \forall i \in [k_0].$$

Clearly each p_u is a valid probability distribution, and this is known as the Paninski's construction [Paninski, 2008]. By the Gilbert–Varshamov bound, there exists $\mathcal{U}_0 \subseteq \mathcal{U}$ such that the minimum pairwise Hamming distance between distinct elements of $\mathcal{U}_0$ is at least $k_0/5$, and $|\mathcal{U}_0| \geq \exp(k_0/8)$. We will set $\{p_u\}_{u\in\mathcal{U}_0}$ as the parameters $\theta_1, \cdots, \theta_M$ in Lemma 1, with $M = |\mathcal{U}_0| \geq \exp(k_0/8)$.

[2]Note that [Han et al., 2018] shows that both the LMM and empirical distributions belong to $\mathcal{P}$.

For each $u \in \mathcal{U}_0$, we also need to specify the associated loss, or equivalently the choice of the 1-Lipschitz property $F_u \in \mathcal{F}_{\text{Lip}}$. The detailed choice of F_u is given by

$$F_u(p) = \sum_{i=1}^{k} f_u(p_i) = \sum_{i=1}^{k} \min_{j\in[k]} |p_i - p_{u,j}|, \quad p = (p_1, \cdots, p_k), u \in \mathcal{U}_0.$$

As the map $x \mapsto |x - x_0|$ is 1-Lipschitz for any $x_0 \in \mathbb{R}$, and the pointwise minimum of 1-Lipschitz functions is still 1-Lipschitz, each F_u is a valid 1-Lipschitz property.

Finally, to apply Lemma 1, it remains to specify the subset $\mathcal{A}_0$. For each $i \in [k]$, let I_i be the open interval $(p_{0,i} - 1/(2k(k-1)), p_{0,i} + 1/(2k(k-1)))$; clearly $I_1, \cdots, I_k$ are disjoint intervals by the definition of p_0. Now we define $\mathcal{A}_0$ as

$$\mathcal{A}_0 \triangleq \left\{ q = (q_1, \cdots, q_k) \in \mathcal{M}_k : \sum_{i=1}^{k} \prod_{j=1}^{k} \mathbb{1}(q_j \notin I_i) \le \frac{k}{10} \right\}.$$

In other words, the subset $\mathcal{A}_0$ consists of all probability vectors which intersect with at least $9/10$ of the intervals $I_1, \cdots, I_k$.

With the above construction and definitions, we are about to use Lemma 1 for the adaptation lower bound. Specifically, we are left with three tasks: to lower bound the separation parameter Δ, to lower bound the minimum probability $p_{\min}$ for all estimators $\widehat{p} \in \mathcal{P}$, and to upper bound the mutual information $I(U; X^n)$.

Lower bound of Δ. First, we aim to find a lower bound of $|F_u(q) - F_u(p_u)| + |F_{u'}(q) - F_{u'}(p_{u'})|$ for all $q \in \mathcal{A}_0$ and $u \neq u' \in \mathcal{U}_0$. By construction of F_u, it is clear that $F_u(p_u) = 0$ for all $u \in \mathcal{U}_0$, and the above quantity can be written as

$$|F_u(q) - F_u(p_u)| + |F_{u'}(q) - F_{u'}(p_{u'})| = \sum_{i=1}^{k} \left(\min_{j\in[k]} |q_i - p_{u,j}| + \min_{j\in[k]} |q_i - p_{u',j}| \right).$$

One could check the following simple fact: if $q_i \in I_{j(i)}$ for some $j(i) \in [k]$, then

$$\min_{j\in[k]} |q_i - p_{u,j}| + \min_{j\in[k]} |q_i - p_{u',j}| \ge |p_{u,j(i)} - p_{u',j(i)}| \in \{0, 2\delta\}.$$

By the definition of $q \in \mathcal{A}_0$, we know that the set $\{j(i)\}_{i\in[k]}$ contains at least $9k/10$ elements of $[k]$. Moreover, by the minimum distance property of $\mathcal{U}_0$, for any $u \neq u' \in \mathcal{U}_0$, there are at least $k/5$ indices $j \in [k]$ such that $|p_{u,j} - p_{u',j}| = 2\delta$. By an inclusion-exclusion principle, there are at least $9k/10 + k/5 - k = k/10$ elements in the set $\{j(i)\}_{i\in[k]}$ such that $|p_{u,j(i)} - p_{u',j(i)}| = 2\delta$, and therefore

$$|F_u(q) - F_u(p_u)| + |F_{u'}(q) - F_{u'}(p_{u'})| \ge \frac{k}{10} \cdot 2\delta = \frac{k\delta}{5}, \quad \forall u \neq u' \in \mathcal{U}_0, q \in \mathcal{A}_0.$$

In other words, $\Delta \ge k\delta/5$ in Lemma 1.

Lower bound of $p_{\min}$. Next, we lower bound the probability $\mathbb{P}_{p_u}(\widehat{p}(X) \in \mathcal{A}_0)$ for all $\widehat{p} \in \mathcal{P}$ and $u \in \mathcal{U}_0$. Here we need to use the definition of $\mathcal{P}$ in Assumption 1. Assume without loss of generality that $\widehat{p}_1 \le \cdots \le \widehat{p}_k$, as any permutation of $\widehat{p}$ does not affect the validity of Assumption 1. Also, by the definition of p_u and the choice of δ in (12), the entries of each p_u are monotonically increasing as well. Consequently, choosing $p = p_u$ in Assumption 1 gives

$$\mathbb{E}_{p_u}\left[\sum_{i=1}^{k} |\widehat{p}_i - p_{u,i}| \right] \le A(n)\sqrt{\frac{k}{n}}.$$

On the other hand, if the event $\widehat{p} \notin \mathcal{A}_0$ occurs, there are at least $k/10$ indices $i \in [k]$ such that $\widehat{p}_j \notin I_i$ for all $j \in [k]$. Consequently, for such an index i, one has $|\widehat{p}_i - p_{u,i}| \geq 1/(2k(k-1)) - \delta \geq 1/(4k(k-1))$ by the choice of δ in (12). Therefore,

$$\sum_{i=1}^{k} |\widehat{p}_i - p_{u,i}| \geq \frac{k}{10} \cdot \frac{1}{4k(k-1)} \cdot \mathbb{1}(\widehat{p} \notin \mathcal{A}_0) \geq \frac{1}{40k} \cdot \mathbb{1}(\widehat{p} \notin \mathcal{A}_0).$$

Combining the above two inequalities, we conclude that

$$\sup_{\widehat{p} \in \mathcal{P}} \max_{u \in \mathcal{U}_0} \mathbb{P}_{p_u}(\widehat{p}(X) \notin \mathcal{A}_0) \leq 40A(n) \cdot \sqrt{\frac{k^3}{n}},$$

which is far smaller than 1 as $k \ll n^{1/3}$ and the assumption $A(n) \ll n^{\delta}$ for all $\delta > 0$. Consequently, we may choose $p_{\min} \geq 1/2$.

Upper bound of $I(U; X^n)$. The upper bound of the mutual information could be established in a similar way as [Han et al., 2018]. Specifically, the following chain of inequalities holds:

$$\begin{aligned} I(U; X^n) &\overset{\text{(a)}}{\leq} \mathbb{E}_U[D_{\mathrm{KL}}(p_U^{\otimes n} \| p_0^{\otimes n})] \\ &\overset{\text{(b)}}{=} n \cdot \mathbb{E}_U[D_{\mathrm{KL}}(p_U \| p_0)] \\ &\overset{\text{(c)}}{\leq} n \cdot \mathbb{E}_U \left[\sum_{i=1}^{k} \frac{(p_{U,i} - p_{0,i})^2}{p_{0,i}} \right] \\ &\overset{\text{(d)}}{\leq} 2nk^2\delta^2, \end{aligned}$$

where (a) is due to the variational representation of the mutual information $I(U; X) = \min_{Q_X} \mathbb{E}_U[D_{\mathrm{KL}}(P_{X|U} \| Q_X)]$ and the fact that $P_{X^n|U} = p_U^{\otimes n}$, (b) follows from the chain rule of the KL divergence, (c) uses the inequality $D_{\mathrm{KL}}(P\|Q) \leq \chi^2(P\|Q)$, and (d) follows from $\min_{i \in [k]} p_{0,i} \geq 1/(2k)$ and simple algebra. Consequently, the mutual information could be upper bounded as $I(U; X^n) \leq 2nk^2\delta^2$.

Combining the above analysis, Lemma 1 gives that

$$\inf_{\widehat{p} \in \mathcal{P}} \sup_{F \in \mathcal{F}_{\mathrm{Lip}}} \sup_{p \in \mathcal{M}_k} \mathbb{E}_p |F(\widehat{p}) - F(p)| \geq \frac{k\delta}{10} \left(\frac{1}{2} - \frac{2nk^2\delta^2 + \log 2}{k/20} \right).$$

Consequently, choosing $\delta = c/\sqrt{nk}$ for a small enough constant $c > 0$ completes the proof of the target lower bound (11) (note that the condition (12) on δ is also fulfilled as $k \ll n^{1/3}$).

C Proof of Theorem 2

This section is devoted to the proof of Theorem 2. The proof consists of two steps: first, we show that the PML distribution belongs to the class $\mathcal{P}$ in Assumption 1, and therefore the adaptation lower bound of Theorem 1 holds for the PML estimator; second, we argue by contradiction that if Theorem 2 is false, then the PML plug-in approach will also achieve the rate-optimal minimax rate for all 1-Lipschitz properties for some $k \ll n^{1/3}$, a contradiction to Theorem 1.

Step I: show that $p^{\mathrm{PML}} \in \mathcal{P}$. First, for the empirical distribution $\widehat{p}$, [Han et al., 2015] shows that

$$\sup_{p \in \mathcal{M}_k} \mathbb{E}_p \left[\sum_{i=1}^{k} |\widehat{p}_i - p_i| \right] \leq \sqrt{\frac{k}{n}}.$$

Moreover, a single perturbation of the observations $X_1, \cdots, X_n$ only changes the quantity $\sum_{i=1}^k |\widehat{p}_i - p_i|$ by at most $2/n$. Hence, by McDiarmid's inequality, we have

$$\sup_{p\in\mathcal{M}_k} \mathbb{P}_p\left[\min_{\sigma\in\mathcal{S}_k}\sum_{i=1}^k |\widehat{p}_{\sigma(i)} - p_i| \geq \sqrt{\frac{k}{n}} + \varepsilon\right] \leq 2\exp\left(-\frac{n\varepsilon^2}{2}\right)$$

for every $\varepsilon > 0$. As for the PML distribution, the competitive analysis of [Acharya et al., 2017a] shows that

$$\sup_{p\in\mathcal{M}_k} \mathbb{P}_p\left[\min_{\sigma\in\mathcal{S}_k}\sum_{i=1}^k |p^{\mathrm{PML}}_{\sigma(i)} - p_i| \geq 2\varepsilon\right] \leq |\Phi_{n,k}| \cdot \sup_{p\in\mathcal{M}_k} \mathbb{P}_p\left[\min_{\sigma\in\mathcal{S}_k}\sum_{i=1}^k |\widehat{p}_{\sigma(i)} - p_i| \geq \varepsilon\right],$$

where $|\Phi_{n,k}|$ is the cardinality of all possible profiles with length n and support size k. Note that trivially $|\Phi_{n,k}| \leq (n+1)^k$ holds, the above two inequalities lead to

$$\sup_{p\in\mathcal{M}_k} \mathbb{P}_p\left[\min_{\sigma\in\mathcal{S}_k}\sum_{i=1}^k |p^{\mathrm{PML}}_{\sigma(i)} - p_i| \geq 2\varepsilon\right] \leq \min\left\{1, 2\exp\left(k\log(n+1) - \frac{n}{2}\left(\varepsilon - \sqrt{\frac{k}{n}}\right)_+^2\right)\right\}. \tag{13}$$

Now integrating the RHS of (13) over $\varepsilon \in (0,\infty)$ gives that $p^{\mathrm{PML}} \in \mathcal{P}$ with $A(n) = O(\sqrt{\log n})$.

Step II: proof by contradiction. Assume by contradiction that Theorem 2 is false, i.e. there exists an absolute constant c_0 such that for some large enough n, it holds that

$$\sup_{p\in\mathcal{M}_k} \mathbb{P}_p(|F(p^{\mathrm{PML}}) - F(p)| \geq C\varepsilon) \leq \exp(c_0 n^{1/3-c_1}) \cdot \left(\inf_{\widehat{F}} \sup_{p\in\mathcal{M}_k} \mathbb{P}_p(|\widehat{F} - F(p)| \geq \varepsilon)\right)^{1-c_2} \tag{14}$$

for all $k \in \mathbb{N}, \varepsilon > 0$ and $F \in \mathcal{F}_{\mathrm{Lip}}$. For any $\varepsilon \gg n^{-1/2}$ and $k \gg 1$, it was shown in [Hao and Orlitsky, 2019b] that the minimax error probability for any 1-Lipschitz property estimation is at most

$$\inf_{\widehat{F}} \sup_{p\in\mathcal{M}_k} \mathbb{P}_p(|\widehat{F} - F(p)| \geq \varepsilon) \leq 2\exp\left(-c_\delta n^{1-\delta}\left(\varepsilon - d_\delta\sqrt{\frac{k}{n\log n}}\right)_+^2\right),$$

for an arbitrary constant $\delta > 0$ and constants $c_\delta, d_\delta > 0$ depending only on δ. Consequently, (14) implies that

$$\sup_{F\in\mathcal{F}_{\mathrm{Lip}}} \sup_{p\in\mathcal{M}_k} \mathbb{P}_p(|F(p^{\mathrm{PML}}) - F(p)| \geq C\varepsilon) \leq 2\exp\left(c_0 n^{1/3-c_1} - (1-c_2)c_\delta n^{1-\delta}\left(\varepsilon - d_\delta\sqrt{\frac{k}{n\log n}}\right)_+^2\right).$$

Choosing $\delta < c_1/4$, $\varepsilon = 2d_\delta\sqrt{k/(n\log n)}$ and $k \asymp n^{1/3-c_1/2}$, the above inequality shows that there exists an absolute constant $c_0' > 0$ depending only on (c_0, c_1, c_2, C) such that

$$\sup_{F\in\mathcal{F}_{\mathrm{Lip}}} \sup_{p\in\mathcal{M}_k} \mathbb{P}_p\left(|F(p^{\mathrm{PML}}) - F(p)| \geq \frac{1}{c_0'}\sqrt{\frac{k}{n\log n}}\right) \leq 2\exp\left(-c_0' n^{1/3-c_1}\right).$$

Hence, using that $\mathbb{E}|X| \leq t + \|X\|_\infty \cdot \mathbb{P}(|X| \geq t)$ for any $t > 0$ implies that for $k \asymp n^{1/3-c_1/2}$ and n tending to infinity (possibly along some subsequence), we arrive at

$$\sup_{F\in\mathcal{F}_{\mathrm{Lip}}} \sup_{p\in\mathcal{M}_k} \mathbb{E}_p|F(p^{\mathrm{PML}}) - F(p)| \lesssim \sqrt{\frac{k}{n\log n}},$$

a contradiction to Theorem 1 as $p^{\mathrm{PML}} \in \mathcal{P}$. Therefore, the inequality (14) does not hold, and the proof of Theorem 2 is completed.