Estimation of Skill Distributions

Ali Jadbabaie®, Fellow, IEEE, Anuran Makur®, Senior Member, IEEE, and Devavrat Shah®, Fellow, IEEE

Abstract—In this paper, we study the problem of learning the skill distribution of a population of agents from observations of pairwise games in a tournament. These games are played among \( n \) randomly drawn agents from the population. The agents in our model can be individuals, sports teams, or even Wall Street fund managers. Formally, we postulate that the likelihoods of outcomes of games are governed by the parametric Bradley-Terry-Luce (or multinomial logit) model, where the probability of an agent beating another is the ratio between its skill level and the pairwise sum of skill levels, and the skill parameters are drawn from an unknown, non-parametric skill density of interest.

The above problem is, in essence, to learn a distribution from noisy and quantized observations. We propose a surprisingly simple and tractable algorithm that learns the skill density with near-optimal minimax mean squared error scaling as \( n^{-1+\varepsilon} \), for any \( \varepsilon > 0 \), so long as the density is smooth. Our approach brings together prior work on learning skill parameters from pairwise comparisons with kernel density estimation from non-parametric statistics. We then prove information theoretic lower bounds which establish minimax near-optimality of the skill parameter estimation technique used in our algorithm. These bounds utilize a continuum version of Fano’s method along with a careful covering argument. Furthermore, we show that estimation error bounds for the skill density translate to theoretical guarantees on estimating the differential entropy and other bounded statistics of the skill density. Finally, we apply our algorithm to data from soccer world cups and leagues, cricket world cups, and even mutual funds. We find that the differential entropy of a learnt distribution provides a quantitative measure of overall skill in a tournament, which in turn can provide explanations for popular beliefs about perceived qualities of sporting and other tournaments.

Index Terms—Bradley-Terry-Luce model, Parzen-Rosenblatt kernel method, non-parametric density estimation, rank centrality, Fano’s method.

I. INTRODUCTION

The perennially evolving nature of sports and other tournaments has given rise to a multitude of beliefs about the perceived overall “quality” of these tournaments. For example, soccer enthusiasts have observed that recent Soccer World Cups are no longer the goal-fests they once were, and the overall standard of play has greatly improved with matches becoming significantly tighter [2]. Indeed, it is believed that due to the globalization of the sports industry and increased mobility of talented players, recent Soccer World Cups are far more competitive than earlier ones in the 1970s and 1980s. But are such beliefs and assertions backed up by data, or are they just common misconceptions?

In this work (which extends [1]), we partly answer this question by quantifying such observations, beyond mere sports punditry and subjective opinions, using statistical analysis. Our approach can be used to quantify the evolution of the overall “quality” and relative skill levels of agents in any tournament over time. For instance, we later illustrate how it can be used to analyze mutual funds over the years.

To this end, we first delineate the general mathematical setup of our analysis. We posit that the population of agents in a tournament, e.g., EPL teams or mutual fund managers, has an associated distribution of skill levels with a probability density function (PDF) \( P_\alpha \) over \( \mathbb{R}_+ \). Our goal is to learn this \( P_\alpha \). Traditionally, in the non-parametric statistics literature (cf. [3]), one observes samples from the distribution directly to estimate \( P_\alpha \). In our setting, however, we can only observe extremely noisy, quantized values. Specifically, given \( n \geq 2 \) individuals, teams, or players participating in a tournament, indexed by \( n \triangleq \{1, \ldots, n\} \), let their latent skill levels be \( \alpha_i, i \in [n] \), which are sampled independently from \( P_\alpha \). We observe the outcomes of pairwise games or comparisons between them.

More precisely, for each \( i, j \in [n] \) with \( i \neq j \), independently with probability \( p \in (0, 1) \), we observe the outcomes of \( k \geq 1 \) games, and with probability \( 1 - p \), we observe nothing. Let \( G(n, p) \) denote the induced Erdős-Rényi random graph with vertices \([n]\), where an edge \( \{i, j\} \in G(n, p) \) (exists in \( G(n, p) \)) if and only if games between \( i \) and \( j \) are observed. For \( \{i, j\} \in G(n, p) \), let \( Z_m(i, j) \in \{0, 1\} \) denote the outcome of the \( m \)-th game between \( i \) and \( j \) for \( m \in [k] \), with value 1 if \( j \) beats \( i \) and 0 otherwise. By definition, \( Z_m(i, j) + Z_m(j, i) = 1 \).

We assume that the likelihoods of the outcomes of games are determined by the Bradley-Terry-Luce (BTL) [4], [5] or multinomial logit model [6] where:

\[
\forall i \neq j \in [n], \forall m \in [k],\quad P(Z_m(i, j) = 1 | \alpha_1, \ldots, \alpha_n) \triangleq \frac{\alpha_j}{\alpha_i + \alpha_j},
\]

Manuscript received 6 April 2023; revised 15 March 2024; accepted 2 June 2024. Date of publication 10 June 2024; date of current version 20 August 2024. This work was supported in part by the Vannevar Bush Faculty Fellowship, in part by the Army Research Office (ARO) under Grant W911NF-18-S-0001, in part by the National Science Foundation (NSF) under Grant CMMI-1634259 and Grant CNS-1523546, and in part by MIT-KACST Project. An earlier version of this paper was presented in part at the 34th Annual Conference on Neural Information Processing Systems (NeurIPS) 2020 [1]. (Corresponding author: Anuran Makur)

Ali Jadbabaie is with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: jadbabai@mit.edu).

Anuran Makur is with the Department of Computer Science and the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: amakur@purdue.edu).

Devavrat Shah is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: devavrat@mit.edu).

Communicated by Y. Liang, Associate Editor for Machine Learning and Statistics.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2024.34141600.

Digital Object Identifier 10.1109/TIT.2024.34141600

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
and the $Z_m(i, j)$’s (i.e., the outcomes of games) are conditionally independent given $\alpha_1, \ldots, \alpha_n$. We also note that $\mathcal{G}(n, p)$ is independent of $\alpha_1, \ldots, \alpha_n$ and the $Z_m(i, j)$’s. Our choice to model likelihoods of outcomes of games using the BTL model is propelled by a long history of such analyses in the statistics, economics, and psychology literatures (see Section I-C for a brief summary and references).

Our objective is to learn $P_\alpha$ from the observations $\{Z_m(i, j) : \{i, j\} \in \mathcal{G}(n, p), m \in [k]\}$, instead of $\alpha_i, i \in [n]$ as in traditional statistics [3]. For a given, fixed set of $\alpha_i, i \in [n]$, learning them from pairwise comparison data $\{Z_m(i, j) : \{i, j\} \in \mathcal{G}(n, p), m \in [k]\}$ has been extensively studied in the recent literature, cf. [7], [8], [9]. Nevertheless, this line of research does not provide any means to estimate the underlying skill distribution $P_\alpha$.

A. Main Contributions

In this subsection, we delineate the main contributions of this work, which include a new skill density estimation problem setting, an algorithm to solve this problem in Algorithm 1, estimation error analysis in our main technical results (Theorems 1, 2, and 3), several new propositions with useful technical content, and experiments to illustrate our method along with a proposal to measure overall skill score. These contributions are detailed below:

1) Unlike the bulk of prior literature on BTL models (see Section I-C), we assume that skill levels are drawn from an unknown skill distribution $P_\alpha$ and rigorously formulate the problem of non-parametric density estimation for $P_\alpha$, based on pairwise comparison data above and in Section II-A (also see Section I-B below for motivation).

2) We develop a statistically near-optimal and computationally tractable method for estimating the skill distribution $P_\alpha$ from a subset of pairwise comparisons in Algorithm 1. Our estimation method is a two-stage algorithm that first uses the (spectral) rank centrality estimator [7], [8] to estimate skill levels and then applies the Parzen-Rosenblatt kernel density estimator [11], [12] with carefully chosen bandwidth to estimate $P_\alpha$.

In particular, the estimator in the second stage is chosen to be robust to errors in the first stage and to enable tractable analysis. For example, standard analysis of histogram density estimators is inconvenienced by the errors between estimated and true skill levels from the first stage due to the binning element of these estimators.

3) We establish that the minimax mean squared error (MSE) of our method scales as $\tilde{O}(n^{-1+\epsilon})$ for any $P_\alpha$ belonging to an $\eta$-Hölder class in Theorem 2. Our analysis relies on a new MSE decomposition result in Proposition 4 that decouples the overall MSE into the error in density estimation using true skill levels and the error in skill level estimation. Theorem 2 implies that if $P_\alpha$ is smooth ($C^\infty$) with bounded derivatives, then the minimax MSE is $\tilde{O}(n^{-1+\epsilon})$ for any $\epsilon > 0$; see the discussion after Theorem 2 in Section III-B. Somewhat surprisingly, although we do not directly observe $\alpha_i, i \in [n]$, this minimax MSE rate matches the minimax MSE lower bound of $\Omega(n^{-1})$ for smooth $P_\alpha$ even when $\alpha_i, i \in [n]$ are observed [3], [10]. The new technical result in Theorem 2 is presented in Table I.

4) As a key step in our estimation method, we utilize the rank centrality algorithm in [7] and [8] for estimating $\alpha_i, i \in [n]$. While the optimal learning rate of the rank centrality algorithm with respect to relative $\ell^2$-loss is well-understood [7], [8], [9], the optimal learning rates with respect to general relative $\ell^q$-losses with $q \in [1, \infty]$ are not known since we can deduce upper bounds from [9] (see, e.g., Theorem 1 and Proposition 3), but not matching minimax lower bounds. In Theorem 1, we prove minimax lower bounds of $\tilde{\Omega}(n^{-1/2})$ with respect to general relative $\ell^q$-losses. These bounds match the learning rates of the rank centrality algorithm for general relative $\ell^q$-losses that can be deduced from [9], and hence, identify the optimal minimax rates (up to logarithmic factors). We derive these information theoretic lower bounds by employing the generalized Fano’s method—a recently studied variant of Fano’s method [13], [14]. Executing this method in our setting requires a technical covering argument to bound mutual information developed in Proposition 2 (rather than typically used convexity-based bounds), as well as a technical analysis of “small ball probability.” The new technical results in Theorem 1 are presented in Table I.

5) To demonstrate the utility of deriving estimation error bounds on skill distributions, we show that the bound in Theorem 2 carries over to: (a) provide uniform estimation error bounds for bounded statistics using “plug-in” estimators in Proposition 5, and (b) provide an estimation error bound for the differential entropy of the skill distribution in Theorem 3. In particular, our proof of Theorem 3 requires careful rigorous analysis and relies on a key intermediate result about the Lipschitz continuity of differential entropy that we establish in Proposition 6 (also see [15]).

6) To illustrate the utility of our method in Algorithm 1, we perform one synthetic experiment and four experiments on real-world data: Cricket World Cups, Soccer World Cups, English Premier League soccer, and US mutual funds. Intuitively, a concentrated skill

| Estimation problem | Loss | Upper bound | Lower bound |
|--------------------|------|-------------|-------------|
| Smooth $C^\infty$ skill PDF | MSE | $\tilde{O}(n^{-1+\epsilon})$ (Theorem 2) | $\Omega(n^{-1})$ [3], [10] |
| BTL skill parameters | Relative $\ell^q$-norm | $\tilde{O}(n^{-1/2})$ [9] | $\Omega(n^{-1/2})$ (Theorem 1) |
distribution, i.e., one that is close to a Dirac delta measure, corresponds to a tournament with players that are all roughly equally skilled. Hence, the outcomes of games are very random or unpredictable. On the other hand, a skill distribution that is close to uniform suggests a more even representation of different skill levels in the player population. So, the outcomes of games are driven more by skill rather than luck (or random chance). We, therefore, propose to use the negative differential entropy of a learnt skill distribution as a way to measure the “overall skill score.” Notably, in our real-data experiments, for Cricket World Cups, we find that negative entropy decreases from 2003 to 2019. Indeed, this corroborates with fan experience, where in 2003, Australia and India dominated but all other teams were roughly equal, while in 2019, the skill distribution of teams was more uniform. In Soccer World Cups, we observe that the negative entropy has increased over the years, while in English Premier League soccer, we find that the negative entropy has generally decreased over the last few seasons. For US mutual funds, the negative entropy decreases significantly during the Great Recession of 2008, and we generally see more uniform skill distributions post 2008 compared to earlier. Lastly, in our synthetic experiment, we illustrate some intuitively sound trends of the estimation MSE with respect to important problem parameters for a simple example.

B. Motivation

In order to further motivate our problem setting, we provide several reasons to assume the existence of a skill distribution $P_\alpha$ and estimate it rather than just the individual skill levels $\alpha_1, \ldots, \alpha_n$.

1) As mentioned earlier, much of the literature, both classical and current, on estimation of BTL models adopts a frequentist approach where the skill levels $\alpha_1, \ldots, \alpha_n$ are assumed to be unknown, but deterministic, parameters, cf. [4], [8], [9]. However, in many settings such as competitive tournaments among players, it is also reasonable to assume that the skill levels of the set of all possible players in the population form a latent probability distribution. Indeed, in many statistical models, one philosophically adopts this perspective where certain common characteristics of individuals, e.g., human height or weight, are not construed merely as deterministic constants, but as sampled values drawn from an underlying distribution (see, e.g., [16]). Our work follows this line of thought and assumes that skill levels of players also exhibit a distribution of values at the population level, and a tournament is governed by some skill levels that are sampled from this underlying skill distribution. Under this assumption, the underlying skill distribution $P_\alpha$ becomes an important object to learn in addition to individual skill levels $\alpha_1, \ldots, \alpha_n$, because estimating the skill distribution provides a richer population level understanding of skill levels.

2) Furthermore, assuming the existence of a skill distribution $P_\alpha$ in our model introduces a Bayesian flavor to the problem. Initially, the skill distribution is not known to the statistician a priori. However, if the skill distribution is learnt from large amounts of pairwise comparison data for a particular kind of tournament, this learnt distribution could potentially be used in future statistical work that attempts to apply Bayesian methods to perform estimation of skill levels of players in other tournaments of the same kind. Additionally, incorporating $P_\alpha$ in our model also gives it the flavor of a hierarchical or empirical Bayes model (since skill levels are drawn from $P_\alpha$ in the first stage, and pairwise comparisons are performed conditioned on the skill levels in the second stage) [17], [18]. Stemming from simple models of exchangeability, hierarchical models are useful ways to express probability distributions of data that has complex and layered structure, and a suite of techniques exist to perform accurate inference in such models, e.g., transferring information in one dataset to perform inference about another dataset [18]. Thus, framing our model to include and estimate $P_\alpha$ may allow such hierarchical or empirical Bayes methods to be used in future work.

3) In a different vein, estimating $P_\alpha$ is useful even when we seek to estimate aggregate statistics of the skill level parameters. Firstly, the existence of an underlying skill distribution $P_\alpha$ gives probabilistic meaning to such statistics which in turn enables richer analysis. Secondly, although specific functionals of $P_\alpha$, e.g., moments or variance, may be directly estimated from estimates of skill levels, estimating $P_\alpha$ simultaneously recovers information about all such functionals and also provides additional qualitative information. For example, as mentioned above, it is shown in Proposition 5 that MSE guarantees for estimating $P_\alpha$ yield uniform guarantees on estimating bounded statistics of the form $\mathbb{E}[f(\alpha)]$ for functions $f : \mathbb{R} \to \mathbb{R}$, which includes all moments. Since different functionals (or statistics) are pertinent for different applications, a good estimate of the skill distribution $P_\alpha$ can be very useful in general (especially if it is a priori unknown which statistic will be used).

4) Moreover, some functionals of $P_\alpha$, such as differential entropy which we utilize to define overall skill scores, actually require an estimate of $P_\alpha$ to learn the functional. Indeed, standard non-parametric plug-in estimators for differential entropy in the literature, e.g., integral, resubstitution, or splitting data estimators [19], require an estimate of $P_\alpha$ to compute entropy. As mentioned above, we demonstrate how MSE guarantees for estimating $P_\alpha$ yield statistical guarantees on estimating differential entropy in Theorem 3. Therefore, in contexts such as this work, estimating $P_\alpha$ is eminently desirable.

5) Lastly, estimating skill distributions $P_\alpha$ can also be of utility in other applications. For example, tracking how the skill distribution $P_\alpha$ changes over time for a particular tournament can offer new avenues of data analysis and interpretation; we touch upon this in some of our experiments in Section VI. Another possible
application, which we do not explore in this work, is the use of estimates of $P_\alpha$ for bootstrapping [20]. Indeed, certain statistical properties of the estimates of skill levels could potentially be understood by some version of smoothed bootstrapping [20, 21, 22], where one re-samples from the estimate of $P_\alpha$.

C. Related Literature

The problem of estimating distributions of skill levels from tournaments has received increased attention due to the recent advent of fantasy sports platforms, which give rise to new legal and policy making challenges concerned with regulating the accompanying rise of gambling on such platforms (cf. [23] and follow-up work). Indeed, when the distribution of skill levels of players is concentrated around one point, the associated game is essentially one of chance (or luck), and governments may understandably seek to place more betting regulations on such tournaments. While [23] provides an empirical study of an ad hoc measure of skill using fantasy sports data, we consider a rigorous statistical formulation of this problem where the objective is to estimate an unknown PDF of skill levels from partially observed win-loss data of tournaments.

As mentioned earlier, we assume that all players in a tournament have latent “skill” or “merit” parameters that are drawn from an unknown prior skill PDF, and these skill parameters determine the likelihoods of wins and losses in games according to the BTL model. Our algorithm to estimate such skill distributions proceeds by first estimating skill parameters from the observed data, and then estimating the skill distribution based on these parameter estimates. To estimate the skill PDFs from (estimated) skill parameters in the second stage of our algorithm, we exploit kernel density estimation techniques that were originally developed in [11], [12], and [24]. Moreover, as noted in Table I, to evaluate the minimax MSE risk achieved by our algorithm, we compare our MSE risk scaling with well-known minimax lower bounds on density estimation for certain classes of analytic densities (cf. [10] and the references therein). On a separate front, to establish the near-optimality of the skill parameter estimation technique (to be explained in due course) used in our algorithm, we exploit a variant of the generalized Fano’s method. This method was also initially developed in the context of density estimation in [25] and [26]. For the sake of brevity, we do not review the extensive non-parametric density estimation literature any further, and instead refer readers to [3, Chapters 1 and 2], [27, Chapter “Density Estimation”], and the references therein for thorough modern treatments.

Since we assume that the likelihoods of the outcomes of two-player games in a tournament follow the BTL model [4], [5], and estimation of the skill parameters of this model forms the first stage of our proposed algorithm, we outline several relevant aspects of the vast literature concerning the BTL model in the remainder of this section. Indeed, while the BTL model was introduced in statistics to study pairwise comparisons [4], it has a long and diverse history. The model was initially proposed in [28] in the context of chess, and this work also provided an iterative algorithm to compute the maximum likelihood (ML) estimators of the BTL skill parameters. Moreover, the BTL model is a special case of the Plackett-Luce (PL) model [5], [29], which was originally developed in mathematical psychology. The PL model defines a probability distribution over rankings (or permutations) of players that is a natural consequence of Luce’s choice axiom. This axiom can be perceived as a formulation of the “independence of irrelevant alternatives” in social choice theory and econometrics. In fact, the work in [6] on the multinomial logit model in economics is equivalent to the PL model. The earliest known model that is related to the PL model is perhaps the (generalized parametric) Thurstonian model from psychometrics, which provides a probability distribution over rankings using the so called law of comparative judgment [30]. Specifically, Thurstone models a “discriminal process” to rank $n$ items by first associating latent merit parameters $\omega_1, \ldots, \omega_n$ to each of the $n$ items, and then ranking them by ranking the corresponding random variables $\omega_1 + X_1, \ldots, \omega_n + X_n$, where the independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ represent “noise” in the discriminial process. As explained in [31, Section 9D], the resulting distribution over rankings is equivalent to the PL model when the $X_i$’s have Gumbel (or generalized extreme value type-I) distribution, cf. [32]. We refer readers to [31, Sections 9C and 9D] for other models of rankings based on exponential families and further equivalent formulations of the BTL and PL models, and to [33] for a comprehensive discussion on other equivalent models from a modern machine learning perspective. For example, the celebrated Boltzmann-Gibbs distribution in statistical physics and the softmax model in machine learning are also versions of the PL model.

In order to estimate the skill parameters of the BTL model, two main families of algorithms have been developed in the literature. The first of these is a class of minorization-maximization (MM) algorithms that generalize Zermelo’s iterative algorithm in [28]. Much like how Zermelo’s algorithm computes ML estimators of the parameters under a strong connectivity condition [34] (also see [35, Assumption 1] for a graph theoretic interpretation), the more general MM algorithms can be utilized to perform ML estimation for “generalized” BTL models [35]. Moreover, although MM algorithms are typically seen as extending the better known expectation-maximization (EM) algorithms for ML estimation of latent variable models, e.g., Gaussian mixture models [36], the MM algorithms for generalized BTL models can also be construed as special cases of EM algorithms corresponding to certain choices of latent variables [37]. In contrast, in this paper, we utilize the second, more recently discovered, family of spectral algorithms based on the notion of rank centrality introduced in [7] and [8]. The main innovation of such spectral algorithms is to construe (normalized) skill parameters as an invariant distribution of a reversible Markov chain, and armed with this perspective, estimate skill parameters by first estimating the stochastic kernel defining the Markov chain.

Both MM and spectral algorithms have been analyzed extensively in the literature. For instance, [38] proves the consistency and asymptotic normality of ML estimators for skill parameters computed by Zermelo’s algorithm, and [8, Theorems 1 and 2] establish sample complexity bounds for the
relative $\ell^2$-norm estimation error of (normalized) skill parameters. Furthermore, both families of algorithms are shown to be (order-) optimal for recovering the top few ranked players in [9], which presents non-asymptotic analysis for relative $\ell^\infty$ and $\ell^2$-norm losses (although only the ML estimator achieves the optimal constant for $\ell^2$-loss [39]). In particular, [8] and [9] assume that a random Erdős-Rényi graph captures the subset of pairwise games that are observed in a tournament. Our analysis also considers this partial observation model, and exploits the relative $\ell^\infty$ and $\ell^2$-norm loss results of [9]. In a different vein, [40] establishes minimax estimation bounds for squared semi-norm losses defined by graph Laplacian matrices, where the fixed graphs encode the subsets of observed pairwise games (also see follow-up work), [41] demonstrates that the universal singular value thresholding algorithm can be used to estimate (stochastically transitive) "non-parametric" BTL models, and [42] analyzes a weighted least squares method to estimate BTL skill parameters under a chordal (or projection) distance loss. Finally, we refer readers to [37] and [43] and the references therein for other recent research on efficient Bayesian inference for BTL and PL models. As opposed to these works, we analyze minimax estimation of BTL models under a previously unexplored setting where skill parameters are drawn i.i.d. from a prior skill PDF.

Yet another direction of work has concerned the development and analysis of rigorous hypothesis tests for BTL models to offer data-driven approaches that determine when they are representative models [44]. In a complementary vein, several recent works have also studied uncertainty quantification in BTL and related models. For example, [39] develops a unified approach to proving sharp non-asymptotic expansions for both ML and spectral estimators of skill parameters. These expansions are in turn utilized to establish central limit theorems for both estimators and confidence intervals. In a similar setting, [45] proposes a Lagrangian-duality-based de-biasing procedure for ML estimation of skill parameters, performs uncertainty quantification by proving asymptotic normality results for such de-biased estimators under various assumptions, and then uses these results to inform the analysis of specific hypothesis testing problems. On the other hand, [46] considers a new model where BTL skill parameters take the form of affine terms that capture covariate effects. The authors of [46] perform both estimation and uncertainty quantification analysis for the parameters of this new model, deriving estimation error guarantees for ML estimators of parameters in both $\ell^\infty$ and $\ell^2$-norm settings, and then establishing uncertainty quantification results, such as asymptotic normality of ML estimators. Deviating from the pairwise BTL setting, [47] considers a multiway comparison model of data governed by a PL model where the top item is revealed in each comparison. The authors of [47] also analyze both estimation and uncertainty quantification for this model, deriving estimation error guarantees for ML estimators of parameters which generalize the known $\ell^\infty$ and $\ell^2$-norm guarantees in, e.g., [8], [9], and establishing uncertainty quantification results that generalize known results in the BTL setting, e.g., [39]. As noted earlier, our focus on minimax estimation of the underlying skill distributions of BTL models in this work differs from the uncertainty quantification and asymptotic normality results in these works.

D. Notational Preliminaries

We briefly introduce some relevant notation. Let $\mathbb{N} \triangleq \{1, 2, 3, \ldots\}$ denote the set of natural numbers. For any $n \in \mathbb{N}$, let $\mathcal{S}_n$ denote the probability simplex of row probability vectors in $\mathbb{R}^n$, and $\mathcal{S}_{n \times n}$ denote the set of all $n \times n$ row stochastic matrices in $\mathbb{R}^{n \times n}$. For any vector $x \in \mathbb{R}^n$ and any $q \in [1, \infty]$, let $\|x\|_q$ denote the $\ell^q$-norm of $x$, and for any matrix $A \in \mathbb{R}^{n \times n}$, let $\det(A)$ denote the determinant of $A$. For any non-empty interval $I \subseteq \mathbb{R}$ and any (Borel measurable) function $f : I \rightarrow \mathbb{R}$, let $\|f\|_2 \triangleq \left( \int_I f(x)^2 dx \right)^{1/2}$ denote the $L^2$-norm of $f$ (with respect to the Lebesgue measure on $I$), and let $L^2(I)$ denote the Hilbert space of all real-valued (Borel measurable) functions $f : I \rightarrow \mathbb{R}$ with finite $L^2$-norm (with respect to the Lebesgue measure on $I$). Moreover, $\exp(\cdot)$ and $\log(\cdot)$ denote the natural exponential and logarithm functions with base $e$, respectively, $\mathbb{I}(\cdot)$ denotes the indicator function that equals $1$ if its input proposition is true and $0$ otherwise, and $[\cdot]$ and $\lfloor \cdot \rfloor$ denote the ceiling and floor functions, respectively. Finally, we will use standard Bachmann-Landau asymptotic notation, e.g., $O(\cdot), \Omega(\cdot), \Theta(\cdot)$, where it is understood that $n \rightarrow \infty$, and tilde notation, e.g., $\tilde{O}(\cdot), \tilde{\Omega}(\cdot), \tilde{\Theta}(\cdot)$, when we neglect poly$(\log(n))$ factors and problem parameters other than $n$.

Our ensuing analysis will utilize the joint probability law $\mathbb{P}(\cdot)$ of $\alpha_1, \ldots, \alpha_n, G(n, p)$, and $\{Z_m(i, j) : \{i, j\} \in G(n, p), m \in [k]\}$ described earlier, and its associated expectation operator $\mathbb{E}[\cdot]$ (both of which depend on the unknown PDF $P_\alpha$). Moreover, unless stated otherwise, our results will hold for all relevant PDFs $P_\alpha$ that define $\mathbb{P}$.

E. Organization

In closing Section I, we delineate how the remainder of the paper is organized. In Section II, we further clarify the formal setup and describe our estimation algorithm that learns $P_\alpha$ from partial observations of games. In Section III, we present and discuss our main results, i.e., Theorems 1, 2, and 3 mentioned earlier, as well as several auxiliary results. Then, we prove our minimax skill parameter estimation result (Theorem 1) and associated auxiliary results in Section IV. We derive our MSE upper bound on skill PDF estimation (Theorem 2), the implied bound on differential entropy estimation (Theorem 3), and associated auxiliary results in Section V. Finally, we illustrate several simulation results with synthetic data and real-world data pertaining to Cricket World Cups, Soccer World Cups, English Premier League soccer, and US mutual funds in Section VI, and then conclude our discussion in Section VII.

II. ESTIMATION ALGORITHM

Our interest is in estimating the skill PDF $P_\alpha$ from noisy, discrete observations $\{Z_m(i, j) : \{i, j\} \in G(n, p), m \in [k]\}$. Instead, if we had exact knowledge of the samples $\alpha_i, i \in [n]$ from $P_\alpha$, then we could utilize traditional methods from non-parametric statistics such as kernel density estimation.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
However, we do not have access to these samples. So, given pairwise comparisons \( \{ Z_m(i,j) : \{i,j\} \in G(n,p), m \in [k] \} \) generated as per the BTL model with parameters \( \alpha_i, i \in [n] \), we can use some recent developments from the BTL-related literature to estimate these skill parameters. First, a natural two-stage algorithm is to first estimate \( \alpha_i, i \in [n] \) using the observations, and then use these estimated parameters to produce an estimate of \( P_\alpha \). We do precisely this. The key challenge is to ensure that the PDF estimation method is robust to the estimation error in \( \alpha_i, i \in [n] \). As one of our main contributions, we rigorously argue that carefully chosen methods for both steps produces as good an estimation of \( P_\alpha \) robust to the estimation error in \( \alpha_i, i \in [n] \).

### A. Formal Setup

We further formalize the setup here. For any given \( \delta, \epsilon, b \in (0, 1) \) and \( \eta, L_1, B > 0 \), let \( \mathcal{P} = \mathcal{P}(\delta, \epsilon, b, \eta, L_1, B) \) be the (non-parametric) set of all uniformly bounded PDFs with respect to the Lebesgue measure on \( \mathbb{R} \) that have support in \([\delta, 1]\), belong to the \( \eta \)-Hölder class \([3, \text{Definition} 1.2]\), and are lower bounded by \( b \) in an \( \epsilon \)-neighborhood of 1. More precisely, every PDF \( f \in \mathcal{P} \) with support in \([\delta, 1]\) satisfies:

1) \( f \) is bounded (almost everywhere), i.e., \( f(x) \leq B \) for all \( x \in [\delta, 1] \),

2) \( f \) is \( s = \lceil \eta \rceil - 1 \) times differentiable, and its \( s \)th derivative \( f^{(s)} : [\delta, 1] \to \mathbb{R} \) is \( \eta \)-Hölder continuous:

\[
\forall x, y \in [\delta, 1], \quad |f^{(s)}(x) - f^{(s)}(y)| \leq L_1|x - y|^\eta - s, \tag{2}
\]

3) \( f(x) \geq b \) for all \( x \in [1 - \epsilon, 1] \).

As an example, when \( \eta = 1 \), \( \mathcal{P} \) denotes the set of all \( L_1 \)-Lipschitz continuous PDFs on \([\delta, 1]\) that are lower bounded in the neighborhood of unity. We assume in the sequel that \( P_\alpha \in \mathcal{P} \). Furthermore, we define the observation matrix \( Z \in [0, 1]^{n \times n} \), whose \((i,j)\)th element is given by:

\[
Z(i,j) = \begin{cases} 1 \{ \{i,j\} \in G(n,p) \} \sum_{m=1}^{k} Z_m(i,j), & i \neq j \\ 0, & i = j \end{cases} \tag{3}
\]

for all \( i, j \in [n] \).

1) **Minimax Estimation Error Formulation**: It turns out that \( Z \) is a sufficient statistic for the purposes of estimating \( \alpha_i, i \in [n] \) [9, p.2208]. For this reason, we shall restrict our attention to all possible estimators of \( P_\alpha \) using \( Z \). Specifically, let \( \hat{\mathcal{P}} \) be the set of all possible measurable and potentially randomized estimators that map \( Z \) to a Borel measurable function \( \hat{P} : \mathbb{R} \to \mathbb{R} \). Then, the minimax MSE risk is defined as:

\[
RMSE(n) = \inf_{\hat{P} \in \hat{\mathcal{P}}} \sup_{P_\alpha \in \mathcal{P}} \mathbb{E} \left[ \int_{\mathbb{R}} \left( \hat{P}(x) - P_\alpha(x) \right)^2 \right] \tag{4}
\]

where the expectation is with respect to the randomness in \( Z \) as well as within the estimator. Our interest will be in understanding the scaling of \( RMSE(n) \) as a function of \( n \).

In the sequel, we will assume that the parameters \( k, p, \delta, \epsilon, b \) can depend on \( n \), and all other parameters are constant.

2) **More on the Assumptions**: We also briefly explain the motivations behind the assumptions we impose on the non-parametric class of skill densities \( \mathcal{P} \). By restricting \( P_\alpha \) to \( \mathcal{P} \), we are able to perform tractable non-asymptotic analysis in the sequel. Indeed, the Hölder class and boundedness assumptions of \( \mathcal{P} \) are standard in the non-parametric density estimation literature (see, e.g., [3, Section 1.2]). Moreover, since the BTL likelihoods in (1) are invariant to scaling the skill parameters, we may assume without loss of generality that \( \alpha_1, \ldots, \alpha_n \leq 1 \). (The lower bound assumption in a neighborhood of unity ensures that at least one \( \alpha_i \) is close to the upper bound of unity with high probability.) On the other hand, assuming that \( \alpha_1, \ldots, \alpha_n \in [\delta, 1] \).

### B. Description of Algorithm

We now propose an algorithm that constructs a good estimator \( \hat{P} \) of the PDF \( P_\alpha \) based on \( Z \). By analyzing this algorithm, we will eventually obtain an upper bound on (4).

1) **Step 1. Estimate** \( \alpha_i, i \in [n] \): Given the observation matrix \( Z \), let \( S \in \mathbb{R}^{n \times n} \) be the “empirical stochastic matrix” whose \((i,j)\)th element is given by:

\[
\forall i, j \in [n], \quad S(i,j) = \begin{cases} \frac{1}{2np} Z(i,j), & i \neq j \\ 1 - \frac{1}{2np} \sum_{r=1}^{n} Z(i,r), & i = j \end{cases}. \tag{6}
\]

The ensuing proposition shows that \( S \) is indeed row stochastic with high probability when \( p = \Omega(\log(n)/n) \).

**Proposition 1 (Empirical Stochastic Matrix):** If \( n \geq 2 \) and \( p \geq 16(c_1 + 1) \log(n)/(3n) \) for any fixed (universal) constant \( c_1 > 0 \), then we have:

\[
\mathbb{P}(S \in \mathcal{S}_{n \times n}) \geq 1 - \frac{1}{n^{c_1}}.
\]

Proposition 1 is proved in Appendix A-A. Next, inspired by the rank centrality algorithm in [7] and [8], let \( \hat{\pi}_* \in \mathcal{S}_n \) be the invariant distribution of \( S \), given by:

\[
\hat{\pi}_* = \begin{cases} \text{invariant distribution of} \\
\text{any randomly chosen} \\
\text{distribution in} \mathcal{S}_n \end{cases}, \quad \hat{\pi}_* \in \mathcal{S}_{n \times n} \quad \forall S \not\in \mathcal{S}_{n \times n}. \tag{7}
\]

where when \( S \in \mathcal{S}_{n \times n} \), an invariant distribution always exists and we choose one arbitrarily when it is not unique. Then, we can define the following estimates of \( \alpha_1, \ldots, \alpha_n \) based on \( Z \):

\[
\hat{\alpha}_i = \frac{\hat{\pi}_*(i)}{\|\hat{\pi}_*\|_\infty}, \quad \forall i \in [n]. \tag{8}
\]

where \( \hat{\pi}_*(i) \) denotes the \( i \)th entry of \( \hat{\pi}_* \) for \( i \in [n] \).
2) Step 2. Estimate $P_{\alpha}$: Using (8), we construct the Parzen-Rosenblatt (PR) kernel density estimator $\hat{P}^* : \mathbb{R} \to \mathbb{R}$ for $P_{\alpha}$ based on $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$ instead of $\alpha_1, \ldots, \alpha_n$ [11, 12]:

$$\forall x \in \mathbb{R}, \quad \hat{P}^*(x) \triangleq \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{\hat{\alpha}_i - x}{h}\right)$$

(9)

where $h > 0$ is a judiciously chosen bandwidth parameter (see the proof in Section V-B):

$$h = \gamma \max \left\{ \frac{1}{\delta + 1}(pk)^{\frac{1}{\gamma + 2}}, 1 \right\} \left( \log(n) \right)^{\frac{1}{\gamma + 2}}$$

(10)

for any (universal) constant $\gamma > 0$, and $K : [-1,1] \to \mathbb{R}$ is any fixed Lipschitz continuous kernel function of order $s \in \mathbb{N}$, which we define below.

For any $s \in \mathbb{N} \cup \{0\}$, the function $K : [-1,1] \to \mathbb{R}$ is said to be a kernel of order $s$ if it satisfies the following conditions:

1. $K(x) = 0$ for $|x| > 1$,
2. $K$ is (Lebesgue) square-integrable, i.e., $\int_{\mathbb{R}} K(x)^2 dx < \infty$,
3. $\int_{\mathbb{R}} K'(x) dx = 1$,
4. $\int_{\mathbb{R}} x^s K(x) dx = 0$ for all $i \in [s]$ when $s \geq 1$.

Note that $K$ is integrable since it is square-integrable and has support in $[-1,1]$, and any map $\mathbb{R} \ni x \mapsto x^s K(x)$ with $r \geq 0$ is integrable because $K$ is integrable and has support in $[-1,1]$. Such kernels of order $s$ can be constructed using orthogonal polynomials as expounded in [3, Section 1.2.2]. We will additionally assume that:

5. There exists a constant $L_2 > 0$ such that our kernel $K : [-1,1] \to \mathbb{R}$ is $L_2$-Lipschitz continuous:

$$\forall x, y \in \mathbb{R}, \quad |K(x) - K(y)| \leq L_2 |x - y|. \quad (11)$$

This is a mild assumption since several well-known kernels satisfy it. For instance, the (parabolic) Epanechnikov kernel [24]:

$$\forall x \in \mathbb{R}, \quad K_{E}(x) \triangleq \frac{3}{4}(1 - x^2) \mathbb{I}\{|x| \leq 1\}$$

(12)

has order $s = 1$, and is Lipschitz continuous with $L_2 = \frac{3}{2}$. Other examples of valid kernels can be found in [3, p.3 and Section 1.2.2].

3) Summary of Algorithm: We provide the pseudo-code summary of our algorithm in Algorithm 1. With fixed $\delta \in (0,1)$, $\gamma > 0$, and a valid kernel $K : [-1,1] \to \mathbb{R}$, and given knowledge of $k \in \mathbb{N}$ and $p \in (0,1)$ (which can also be easily estimated), Algorithm 1 constructs the estimator (9) for $P_{\alpha}$ based on $Z$. In Algorithm 1, we assume that $S \in \mathcal{S}_{n \times n}$, because this is almost the case in practice. Furthermore, if $k$ varies between players so that $i$ and $j$ play $k_{i,j} = k_{j,i}$ games for $i, j \in [n]$ with $i \neq j$, we can re-define the data $Z(i, j)$ to use $k_{i,j}$ instead of $k$ in (3), and utilize an appropriately altered bandwidth $h$. For example, we can use $k' = \min_{(i,j) \in \mathcal{E}(n,p)} k_{i,j}$ in place of $k$ in (10) to define $h$, which would yield theoretical guarantees akin to Theorem 2 with $k'$. The computational complexity of Algorithm 1 is determined by the running time of rank centrality, e.g., if the spectral gap of $S$ is $\Theta(1)$ and we use power iteration (cf. [48, Section 7.3.1] or [49, Section 4.4.1]) to obtain an

Algorithm 1 Algorithm to Estimate Skill PDF $P_{\alpha}$ Using $Z$

**Input:** Observation matrix $Z \in [0,1]^{n \times n}$ (as defined in (3))

**Output:** Estimator $\hat{P}^* : \mathbb{R} \to \mathbb{R}$ of the unknown PDF $P_{\alpha}$

1. **Step 1: Skill parameter estimation using rank centrality algorithm**
   1. Construct $S \in \mathcal{S}_{n \times n}$ according to (6) using $Z$ (and $p$ and $n$)
   2. Compute leading left eigenvector $\hat{\pi}_\ast \in S_n$ of $S$ in (7)
      $\triangleright$ $\hat{\pi}_\ast$ is the invariant distribution of $S$
   3. Compute estimates $\hat{\alpha}_i = \hat{\pi}_\ast(i)/\|\hat{\pi}_\ast\|_\infty$ for $i = 1, \ldots, n$ via (8)

2. **Step 2: Kernel density estimation using Parzen-Rosenblatt method**
   1. Compute bandwidth $h$ via (10) (using $p$, $k$, $\delta$, $\eta$, and $n$)
   2. Construct $\hat{P}^*$ according to (9) using $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$, $h$, and a valid kernel $K : [-1,1] \to \mathbb{R}$

$O(\text{poly}(n^{-1}))$ $L^2$-approximation of $\hat{\pi}_\ast$, then Algorithm 1 runs in $O(n^2 \log(n))$ time.

C. Intuition for Algorithm

We briefly explain the intuition behind each of the two steps of Algorithm 1. As we mentioned earlier, Step 1 of Algorithm 1 is inspired by the (spectral) rank centrality algorithm in [7] and [8]. To understand this stage, define the row stochastic matrix $D \in \mathcal{S}_{n \times n}$, whose $(i,j)$th element is given by the BTL skill parameters:

$$\forall i, j \in [n], \quad D(i, j) \triangleq \begin{cases} \frac{1}{2n} \left( \frac{\alpha_j}{\alpha_i + \alpha_j}, & i \neq j \\ 1 - \frac{1}{2n} \sum_{r \in [n] \setminus \{i\}} \frac{\alpha_r}{\alpha_i + \alpha_r}, & i = j \end{cases}$$

(13)

where $D(i, j) + D(j, i) = (2n)^{-1}$ for all $i, j \in [n]$ such that $i \neq j$. Note that it is straightforward to verify from (13) that $D(i, i) \geq 0$ for all $i \in [n]$, because:

$$\forall i \in [n], \quad D(i, i) = 1 - \frac{1}{2n} \sum_{r \in [n] \setminus \{i\}} \frac{\alpha_r}{\alpha_i + \alpha_r} \geq 1 - \frac{n - 1}{2n} = \frac{n + 1}{2n} \geq 0. \quad (14)$$

We will construe $D$ as the transition probability matrix of a (time-homogeneous) discrete-time Markov chain on the state space $[n]$ of players. Furthermore, define the “canonically scaled” skill parameters $\pi \in \mathcal{S}_n$ with ith entry given by:

$$\forall i \in [n], \quad \pi(i) \triangleq \frac{\alpha_i}{\alpha_1 + \cdots + \alpha_n}. \quad (15)$$

Clearly, $\pi$ can also be used to define the BTL likelihoods in (1) instead of $\alpha_1, \ldots, \alpha_n$, since $\alpha_j/(\alpha_i + \alpha_j) = \pi(j)/(\pi(i) + \pi(j))$ for all $i, j \in [n]$ with $i \neq j$. Next, following the crucial observation of [8], notice using (13) and (15) that the ensuing detailed balance conditions are satisfied:

$$\forall i, j \in [n], \quad \pi(i)D(i, j) = \pi(j)D(j, i). \quad (16)$$
This implies that \( D \) defines a reversible Markov chain with invariant distribution \( \pi = \pi D \) (see, e.g., [50, Proposition 1.19]). Moreover, this Markov chain is ergodic (i.e., irreducible and aperiodic) because \( D > 0 \) entry-wise, which means that \( \pi \) is the unique invariant distribution of \( D \). This general idea that canonically scaled skill parameters of the BTL model form an invariant distribution of a reversible Markov chain is known as “rank centrality” \([7, 8]\).

Step 1 of Algorithm 1 estimates the BTL skill parameters \( \alpha_1, \ldots, \alpha_n \) using \( \hat{\alpha}_1, \ldots, \hat{\alpha}_n \) in (8) by first estimating the canonically scaled skill parameters \( \pi \in S_n \) in (15). To estimate \( \pi \), it is reasonable to first produce an estimate of \( D \) that is itself a row stochastic matrix (with high probability), and then utilize the corresponding invariant distribution as our estimate of \( \pi \). Notice that:

\[
E[S(\alpha_1, \ldots, \alpha_n)] = D \quad (17)
\]

where \( S \in \mathbb{R}^{n \times n} \) is defined in (6). Hence, \( S \) (which is row stochastic with high probability) can be construed as our estimator of the Markov kernel \( D \). As a consequence, the invariant distribution \( \pi^* \in S_n \) of \( S \) in (7) can be perceived as our estimator of \( \pi \).

Step 2 of Algorithm 1 constructs an estimator for the unknown PDF \( P_\alpha \) of interest using the skill parameter estimates \( \hat{\alpha}_1, \ldots, \hat{\alpha}_n \) obtained from Step 1. Clearly, if we had access to the true i.i.d. samples \( \alpha_1, \ldots, \alpha_n \) from \( P_\alpha \), then we could use the vanilla PR kernel density estimator (see, e.g., (19) in Section III-B) to estimate \( P_\alpha \), because it is known to be minimax optimal for appropriate choices of bandwidth \( h \) and kernel function \( K \) (cf. \([3, 27]\)). However, we do not have access to these true samples. Thus, we utilize the estimates \( \hat{\alpha}_1, \ldots, \hat{\alpha}_n \) to construct an analogous estimator in (9), which is the output of (Step 2 of) Algorithm 1. Intuitively, we expect this estimator to perform well, because \( \hat{\alpha}_1, \ldots, \hat{\alpha}_n \) should be “close” to \( \alpha_1, \ldots, \alpha_n \) when \( n \) is large.

III. MAIN RESULTS AND DISCUSSION

We now present our main results: an achievable minimax MSE for the \( P_\alpha \) estimation method in Algorithm 1, consequences of this result on the estimation of other statistics related to \( P_\alpha \), and minimax lower bounds on estimation of the skill parameters \( \alpha_i, i \in [n] \) from \( Z \) (i.e., Step 1 of Algorithm 1) for any method. This collectively establishes the near-optimality of our proposed method as \( \eta \to \infty \), i.e., as the density becomes smooth (\( C^\infty \)). To this end, we first establish minimax rates for skill parameter estimation, and then derive minimax bounds for PDF estimation.

A. Tight Minimax Bounds on Skill Parameter Estimation

To obtain tight \( P_\alpha \) estimation, it is essential that we have tight skill parameter estimation. Hence, we show that the parameter estimation step performed in (7) has minimax optimal rate (up to logarithmic factors). Specifically, we consider the canonically scaled skill parameters \( \pi \in S_n \) given by (15), which equivalently define the BTL model (1). Building upon the line of work in \([8] \) and \([9] \), the ensuing theorem portrays that for any \( q \in [1, \infty] \), the minimax relative \( \ell^q \)-risk of estimating \( \pi \) based on \( Z \) is \( \Theta(n^{-1/2}) \) (see Table I). For simplicity, we will assume throughout Section III-A on skill parameter estimation that \( \delta, p, \) and \( k \) are \( \Theta(1) \).

**Theorem 1 (Minimax Relative \( \ell^q \)-Risk):** For any \( q \in [1, \infty] \), sufficiently large constants \( c_2, c_3 > 0 \) (which depend on \( \delta, p, \) and \( k \)), and for all sufficiently large \( n \in \mathbb{N} \):

\[
\frac{c_2}{\log(n) \sqrt{n}} \leq \inf_{\pi} \sup_{p_\alpha \in P} \mathbb{E} \left[ \frac{\|\hat{\pi} - \pi\|_q}{\|\pi\|_q} \right] \leq c_3 \frac{\log(n)}{n}
\]

where the infimum is over all estimators \( \hat{\pi} \in S_n \) of \( \pi \) based on \( Z \), and \( \hat{\pi}_* \in S_n \) is defined in (7).

The proof of Theorem 1 can be found in Section IV-D. Theorem 1 states that the rank centrality estimator \( \hat{\pi}_* \) achieves an extremal Bayes relative \( \ell^q \)-risk of \( \tilde{O}(n^{-1/2}) \), and no other estimator can achieve a risk that decays faster than \( \Omega(n^{-1/2}) \). The upper bound in Theorem 1 uses \([9, Theorem 3.1]\) and some calculations to generalize it to arbitrary \( q \in [1, \infty] \), and the lower bound is a new contribution. We prove the lower bound by first lower bounding the minimax risks in terms of Bayes risks in order to circumvent an involved analysis of the infinite-dimensional parameter space \( P \). In particular, we set \( P_\alpha \in P \) to be the uniform PDF over \( [\delta, 1] \), denoted \( \text{unif}([\delta, 1]) \in P \). We then lower bound the Bayes risks using a recent generalization of Fano’s method \([25, 26]\) (also see \([3, 51]\)), which was specifically developed to produce such lower bounds in the setting where the parameter space is a continuum, e.g., \([\delta, 1] \), instead of a finite set \([13, 14, 52], [53]\); see Sections IV-B and IV-C.

The principal analytical difficulty in executing the generalized Fano’s method is in deriving a tight upper bound on the mutual information between \( \pi \) and \( Z \), denoted \( I(\pi; Z) \) (see (21) in Section IV-A for a formal definition), where the probability law of \( \pi \) is defined using \( P_\alpha = \text{unif}([\delta, 1]) \). The ensuing proposition presents our upper bound on \( I(\pi; Z) \).

**Proposition 2 (Covering Number Bound on Mutual Information):** For all \( n \geq 2 \), we have:

\[
I(\pi; Z) \leq \frac{1}{2} \log(n) + \frac{(1 - \delta)^2}{8\delta^2} (2 + \delta + \frac{1}{\delta}) \log(n)
\]

Proposition 2 is proved in Section IV-A. We note that although standard information inequalities (e.g., \([13, Equation (44)] \)) typically suffice to obtain minimax rates for various estimation problems, they only produce a sub-optimal estimate \( I(\pi; Z) = O(n^2) \) in our problem, as explained at the end of Section IV-A, cf. (33). So, to derive the sharper estimate \( I(\pi; Z) = O(n \log(n)) \) in Proposition 2, we execute a careful covering number argument that is inspired by the techniques in \([54]\) (also see the distillation in \([55, Lemma 16.1]\)).

We make three further remarks. Firstly, we note that Theorem 1 holds verbatim if \( P \) is replaced by any set of probability measures with support in \([\delta, 1] \) that contains \( \text{unif}([\delta, 1]) \). Secondly, it is worth juxtaposing our results with \([9, Theorem 5.2]\) and \([8, Theorems 2 and 3]\), which state that the minimax relative \( \ell^q \)-risk of estimating \( \pi \) is \( \Theta(n^{-1/2}) \). This result holds under a worst-case skill parameter value model as opposed to the worst-case prior distribution model of this paper. Finally, while the upper bounds in Theorem 1 utilize a
bound on relative $\ell^\infty$-risk in [9, Theorem 3.1], the bound on relative $\ell^2$-risk in [9, Theorem 5.2] can be used to improve these upper bounds by logarithmic factors when $q \in [1, 2]$. As an illustration of this, in the ensuing proposition, we show that the minimax (relative) $\ell^1$-risk (or total variation distance risk) of estimating $\pi$ based on $Z$ is actually $O(n^{-1/2})$.

**Proposition 3 (Minimax Upper Bound for $q = 1$):** For a sufficiently large constant $c_4 > 0$ (which depends on $\delta$, $p$, and $k$), and for all sufficiently large $n \in \mathbb{N}$:

$$\inf_{\hat{P}_n, P_0 \in \mathcal{P}} \sup_{\pi} \mathbb{E}[[|\hat{\pi} - \pi|_1]] \leq \sup_{\pi} \mathbb{E}[[|\hat{\pi}_* - \pi|_1]] \leq \frac{c_4}{\sqrt{n}},$$

where $\hat{\pi}_* \in \mathcal{S}_n$ is defined in (7).

Proposition 3 is established in Section IV-E.

**B. Minimax Bound on Skill PDF Estimation and Its Consequences**

We now state our main result concerning the estimation error for $P_\alpha$. In particular, we argue that the MSE risk of our estimation algorithm $\hat{P}^*$ (see (9)) scales as $O(n^{-1/(\eta + 1)})$ for any $P_\alpha \in \mathcal{P}$.

**Theorem 2 (MSE Upper Bound):** Fix any sufficiently large constants $c_5, c_6 > 0$ and suppose that $p \geq c_5 \log(n)/(\delta^2 n)$, $b \geq c_6 \sqrt{\log(n)/n}$, $\epsilon \geq 5 \log(n)/(bn)$, and $\lim_{n \to \infty} \delta^{-1}(npk)^{-1/2} \log(n)^{1/2} = 0$. Then, for any $L^2$-Lipschitz continuous kernel $K : [-1, 1] \to \mathbb{R}$ of order $[\eta, 1]$ there exists a sufficiently large constant $c_7 > 0$ (that depends on $\gamma$, $\eta$, $B$, $L_1$, $L_2$, and $K$) such that for all sufficiently large $n \in \mathbb{N}$:

$$R_{\text{MSE}}(n) \leq \sup_{P_\alpha \in \mathcal{P}} \mathbb{E} \left[ \int_\mathbb{R} \left( \hat{P}^*(x) - P_\alpha(x) \right)^2 \, dx \right] \leq c_7 \max \left\{ \left( \frac{1}{\delta^2 pk} \right)^{\frac{1}{\eta + 1}}, 1 \right\} \left( \frac{\log(n)}{n} \right)^{\frac{1}{\eta + 1}}.$$

Theorem 2 is established in Section V-B. We next make several pertinent remarks. Firstly, the condition $p \geq c_5 \log(n)/(\delta^2 n)$ is precisely the critical scaling that ensures that $\mathcal{G}(n, p)$ is connected with high probability, cf. [56, Section 7.1] or [57, Theorem 8.11]. This is essential to estimate $\alpha_1, \ldots, \alpha_n$ in Step 1 of Algorithm 1, since we cannot reasonably compare the skill levels of disconnected players. Secondly, although Theorem 2 holds for $k$ that can vary with $n$ (while satisfying the conditions of the theorem), our main regime of interest is $k = \Theta(1)$, where estimation is intuitively most difficult. Thirdly, while $\hat{P}^*$ can be negative, the non-negative truncated estimator $\hat{P}_+^*(x) = \max\{\hat{P}^*(x), 0\}$ achieves smaller MSE risk than $\hat{P}^*$, cf. [3, p.10]. So it is easy to construct “good” non-negative estimators. Fourthly, we note that similar analyses to Theorem 2 can be carried out for, e.g., Nikol’ski and Sobolev classes of PDFs, cf. [3, Section 1.2.3]. Lastly, it is worth mentioning that BTL models can also be equivalently parametrized using logit parameters $\omega_i = \log(\alpha_i)$, $i \in [n]$, which are drawn independently from the PDF $P_\alpha(x) = e^x P_\alpha(e^x), x \in \mathbb{R}$. When $\delta = \Theta(1)$, it can be shown that the MSE of the estimator $\hat{P}_\text{logit}(x) = e^x \hat{P}^*(e^x), x \in \mathbb{R}$ for $P_\alpha$ is upper bounded by Theorem 2. Therefore, our analysis of Theorem 2 also holds for estimating distributions of logit parameters.

On the other hand, we note that there exists a constant $c_8 > 0$ (depending on $\eta, L_1$) such that for all sufficiently large $n \in \mathbb{N}$, the following minimax lower bound holds [3, Exercise 2.10], [27, Chapter “Density Estimation,” Theorem 6]:

$$R_{\text{MSE}}(n) \geq \inf_{P_\alpha, P_\in \mathcal{P}} \sup_{P_\alpha \in \mathcal{P}} \mathbb{E} \left[ \int_\mathbb{R} \left( \hat{P}_\alpha(x) - P_\alpha(x) \right)^2 \, dx \right] \geq c_8 \left( \frac{1}{n} \right)^{\frac{2}{\eta + 1}},$$

where the infimum is over all estimators $\hat{P}_\alpha : \mathbb{R} \to \mathbb{R}$ of $P_\alpha$ based on $\alpha_1, \ldots, \alpha_n$, and the first inequality holds because the infimum in (4) is over a subset of the class of estimators used in the infimum in (18); indeed, given $\alpha_1, \ldots, \alpha_n$, one can simulate $Z$ via (1) and estimate $P_\alpha$ from $Z$. Thus, when $\eta = 1$, Theorem 2 and (18) show that $R_{\text{MSE}}(n) = O(n^{-1/2})$ and $R_{\text{MSE}}(n) = \Omega(n^{-2/3})$. At the other extreme, if $P_\alpha$ is smooth, i.e., $P_\alpha$ satisfies the boundedness conditions of $\mathcal{P}$ and is infinitely differentiable ($C^\infty$) with all derivatives bounded by $L_1$, then it belongs to $\mathcal{P}(\delta, \epsilon, b, \eta, L_1, B)$ for all $\eta > 0$. Hence, Theorem 2 also holds for all $\eta > 0$ and we may let $\eta \to \infty$. So, for any constant $\epsilon > 0$, letting $\eta = \epsilon^{-1} - 1$ yields an $O(n^{-1+\epsilon})$ MSE upper bound for such smooth PDFs. Likewise, an $O(n^{-1})$ minimax lower bound analogous to (18) also holds for smooth PDFs [3, 10]. Together, these bounds form the first row of Table I.

We next emphasize that the key technical step in the proof of Theorem 2 is the ensuing intermediate result.

**Proposition 4 (MSE Decomposition):** Fix any sufficiently large constants $c_5, c_6, c_7 \geq 0$ and suppose that $p \geq c_5 \log(n)/(\delta^2 n)$, $b \geq c_6 \sqrt{\log(n)/n}$, $\epsilon \geq 5 \log(n)/(bn)$, and $\lim_{n \to \infty} \delta^{-1}(npk)^{-1/2} \log(n)^{1/2} = 0$. Then, for any $P_\alpha \in \mathcal{P}$, any $L_2$-Lipschitz continuous kernel $K : [-1, 1] \to \mathbb{R}$, any bandwidth $h \in (0, 1]$ satisfying $h = \Omega(\max\{1/(\delta \sqrt{pk}), 1\} \sqrt{\log(n)/n})$, and any sufficiently large $n \in \mathbb{N}$:

$$\mathbb{E} \left[ \int_\mathbb{R} \left( \hat{P}_\alpha^*(x) - P_\alpha(x) \right)^2 \, dx \right] \leq 2 \mathbb{E} \left[ \int_\mathbb{R} \left( \hat{P}_\alpha^*(x) - P_\alpha(x) \right)^2 \, dx \right] + \frac{c_7 B^2 L_1^2}{h^2} \mathbb{E} \left[ \max_{i \in [n]} |\alpha_i - \alpha_i|^2 \right] + \frac{c_{10} L_1^2}{n^2 h^4}$$

where $\hat{P}_\alpha^* : \mathbb{R} \to \mathbb{R}$ denotes the classical PR kernel density estimator of $P_\alpha$ [11, 12] based on the true samples $\alpha_1, \ldots, \alpha_n$ (if they were made available by an oracle):

$$\forall x \in \mathbb{R}, \quad \hat{P}_\alpha^*(x) \triangleq \frac{1}{nh} \sum_{i=1}^n K\left( \frac{\alpha_i - x}{h} \right).$$

The proof of Proposition 4 can be found in Section V-A. This result decomposes the MSE between $\hat{P}^*$ (with general $h$) and $P_\alpha$ into two dominant terms: the MSE of estimating $P_\alpha$ using (19), which can be analyzed using a standard bias-variance tradeoff (see Lemma 7 in Section V-B [3, 27]), and the squared $\ell^\infty$-risk of estimating $\alpha_1, \ldots, \alpha_n$ using (8). To analyze the second term, we use a relative $\ell^\infty$-norm bound from [9, Theorem 3.1] (see Lemma 5 in Section IV-D); the same bound is also used to obtain the upper bound in...
Theorem 1. We remark that our proof of Theorem 2 in Section V-B proceeds by balancing the bias in estimating $P_\alpha$ using (19), the variance in estimating $P_\alpha$ using (19), the error in estimating $\alpha_1, \ldots, \alpha_n$ using (8), and lower order terms. As a result, the choice of bandwidth in (10) differs from the usual choice of bandwidth (e.g., in its scaling with $n$) for traditional minimax kernel density estimation from i.i.d. samples, which only balances the bias and variance terms. Hence, solving and tractably analyzing the problem of skill density estimation from pairwise comparison data requires us to carefully choose and tune a density estimation method to combat the error in the skill parameter estimation process (rather than agnostically choosing “good” methods for density and parameter estimation). Indeed, as noted in Section I-A, the density estimator in the second stage of Algorithm 1 is chosen to be robust to errors in the first stage, e.g., standard analysis of histogram density estimators is inconvenient by the errors between estimated and true skill levels from the first stage due to the binning element of these estimators. So, it is more convenient to use kernel density estimators in the second stage. Furthermore, fundamental limits of non-parametric density estimation problems with noisy samples typically depend on the specific structure of the noise distribution (e.g., deconvolution with independent additive noise, cf. [58]). Specifically, in our problem, the error in the estimated skill level parameters is captured by relative additive noise, cf. [58]). Specifically, in our problem, the error in estimating $P_\alpha$ using (19), the variance in estimating $P_\alpha$ using (19), the error in estimating $\alpha_1, \ldots, \alpha_n$ using (8), and lower order terms. As a result, the choice of bandwidth in (10) differs from the usual choice of bandwidth (e.g., in its scaling with $n$) for traditional minimax kernel density estimation from i.i.d. samples, which only balances the bias and variance terms. Hence, solving and tractably analyzing the problem of skill density estimation from pairwise comparison data requires us to carefully choose and tune a density estimation method to combat the error in the skill parameter estimation process (rather than agnostically choosing “good” methods for density and parameter estimation). Indeed, as noted in Section I-A, the density estimator in the second stage of Algorithm 1 is chosen to be robust to errors in the first stage, e.g., standard analysis of histogram density estimators is inconvenient by the errors between estimated and true skill levels from the first stage due to the binning element of these estimators. So, it is more convenient to use kernel density estimators in the second stage. Furthermore, fundamental limits of non-parametric density estimation problems with noisy samples typically depend on the specific structure of the noise distribution (e.g., deconvolution with independent additive noise, cf. [58]). Specifically, in our problem, the error in the estimated skill level parameters is captured by relative $\ell^p$-risks and is correlated across different parameters since the pairwise comparison data is generated by a BTL model. Thus, it is reasonable to analyze our skill density estimation problem by utilizing Proposition 4 to break down the MSE into simpler terms and then balancing these terms by carefully choosing a bandwidth.

Finally, as noted in Section I-A, our analysis of the estimation error of $P_\alpha$ in Theorem 2 also yields theoretical guarantees on the estimation of various other functionals or statistics of $P_\alpha$. As one illustration of this, the ensuing proposition presents an important consequence of Theorem 2.

**Proposition 5 (Estimation of Bounded Statistics):** Fix any sufficiently large constants $c_5, c_6 > 0$ and suppose that $P_\alpha$ is $L_2$-Lipschitz continuous kernel $K : [-1, 1] \rightarrow \mathbb{R}$ of order $|\alpha| - 1$, there exists a sufficiently large constant $c_7 > 0$ (that depends on $\gamma$, $\eta$, $B$, $L_1$, $L_2$, and $K$) such that for all sufficiently large $n \in \mathbb{N}$:

$$\sup_{P_\alpha \in \mathcal{P}} \sup_{f : \mathbb{R} \rightarrow [-1, 1]} \mathbb{E} \left[ \left( \int_{\mathbb{R}} f(x) \hat{P}_\alpha(x) \, dx - \int_{\mathbb{R}} f(x) P_\alpha(x) \, dx \right)^2 \right] \leq 3 c_7 \max \left\{ \frac{1}{(\delta^5 pk)^{\frac{1}{n^2} + 1}}, 1 \right\} \left( \log \frac{n}{\delta} \right)^{\frac{3}{n^2}}$$

where the second supremum is over all (Borel measurable) functions $f : \mathbb{R} \rightarrow [-1, 1]$ that are bounded by 1.

Proposition 5 is derived in Appendix A-B. It conveys that by estimating $P_\alpha$, we obtain uniform guarantees on estimating any bounded statistic of the form $E[f(\alpha)] = \int_{\mathbb{R}} f(x) P_\alpha(x) \, dx$ (for a bounded function $f : \mathbb{R} \rightarrow \mathbb{R}$) using the “plug-in” estimator $\int_{\mathbb{R}} f(x) \hat{P}_\alpha(x) \, dx$. In particular, this implies that we get statistical guarantees on estimating all moments of $P_\alpha$ since $P_\alpha$ has support in $[\delta, 1]$.

As another illustration, the next theorem shows that Theorem 2 also yields a statistical guarantee on the estimation of the differential entropy of $P_\alpha$, denoted $h(P_\alpha)$. Here, and in the setting below, we let $h(f)$ be the differential entropy of a PDF $f$, which is always well-defined and finite in the setting we consider (see (68) in Section V-C for a formal definition).

**Theorem 3 (Estimation of Differential Entropy):** Let $\mathcal{D} = \mathcal{D}(\delta, \alpha, L_3)$ denote the set of all PDFs (with respect to the Lebesgue measure) that have support $[\delta, 1]$, are uniformly lower bounded by a constant $\alpha > 0$, and are $L_3$-Lipschitz continuous with constant $L_3 > 0$. Suppose that the true skill distribution satisfies $P_\alpha \in \mathcal{P} \cap \mathcal{D}$, i.e., $P_\alpha$ is additionally $L_3$-Lipschitz continuous and lower bounded by $a$ on $[\delta, 1]$. Fix any sufficiently large constants $c_5, c_6 > 0$ and suppose that $P \geq c_5 \log(n)/(\delta^5 n)$, $b \geq c_6 \sqrt{\log(n)/n}$, $c \geq 5 \log(n)/(bn)$, and $\lim_{n \rightarrow \infty} \delta^{-1}(npk)^{-1/2} \log(n)^{1/2} = 0$. Moreover, for any $L_2$-Lipschitz continuous kernel $K : [-1, 1] \rightarrow \mathbb{R}$ of order $|\alpha| - 1$, let $\mathcal{P}_{\text{temp}} : \mathcal{P} \rightarrow \mathcal{D}$, $\mathcal{P}_{\text{temp}}(x) = P_{\text{temp}}(x)1[x \in [\delta, 1])$ denote the tempered estimator corresponding to $\mathcal{P}_{\text{proj}}$ in (9), and define the projected PDF estimator $\mathcal{P}_{\text{proj}} : [\delta, 1] \rightarrow \mathbb{R}$ via:

$$\mathcal{P}_{\text{proj}} \triangleq \arg\min_{\mathcal{P} \in \mathcal{D}} \left\| \mathcal{P}_{\text{temp}} - \mathcal{P} \right\|^2_2.$$

Then, there exists a sufficiently large constant $c_{11} > 0$ (that depends on $\gamma$, $\eta$, $B$, $L_1$, $L_2$, and $K$) such that for all sufficiently large $n \in \mathbb{N}$:

$$\sup_{P_\alpha \in \mathcal{P} \cap \mathcal{D}} \mathbb{E} \left( \left( h(\mathcal{P}_{\text{proj}}) - h(P_\alpha) \right)^2 \right) \leq c_{11} \max \left\{ \frac{1}{(\delta^5 pk)^{\frac{1}{n^2} + 1}}, 1 \right\} \left( \log \frac{n}{\delta} \right)^{\frac{3}{n^2}}.$$

Theorem 3 is derived in Section V-C. The proof first develops a result in Proposition 6 that captures the Lipschitz continuity of differential entropy with respect to the $L^2$-norm over all PDFs in $\mathcal{D}$ (cf. [15]). Then, after some careful analysis to ensure that the estimator $h(\mathcal{P}_{\text{proj}})$ is well-defined, the statement of Theorem 3 can be deduced from Theorem 2 and Proposition 6.

We conclude this section with several other remarks about Theorem 3. Firstly, the condition that $P_\alpha \in \mathcal{P} \cap \mathcal{D}$ is not much more stringent than $P_\alpha \in \mathcal{P}$. For example, when $\eta = 1$, $\mathcal{P}$ is the set of all $L_1$-Lipschitz continuous PDFs on $[\delta, 1]$ that are lower bounded in the neighborhood of unity (as noted in Section II-A). On the other hand, $\mathcal{P} \cap \mathcal{D}$ is the set of all $L_1$-Lipschitz continuous PDFs on $[\delta, 1]$ that are lower bounded on $[\delta, 1]$, where we assume that $L_1 \leq L_3$ for simplicity. Hence, we only strengthen the lower bound condition on $P_\alpha$. Secondly, we choose to analyze the estimator $h(\mathcal{P}_{\text{proj}})$ for the true differential entropy $h(P_\alpha)$ in Theorem 3 for reasons of analytical tractability. The goal of Theorem 3 is to illustrate a (minimax) upper bound on entropy estimation that is implied by Theorem 2, and the estimator $h(\mathcal{P}_{\text{proj}})$ serves this purpose. As mentioned in Section I-A, we will use negative differential entropy of a learnt skill distribution as a measure
of “overall skill score” in our experiments in Section VI, but for convenience, we will apply an alternative entropy estimator to generate our plots there. Lastly, it is worth mentioning that the estimators for bounded statistics and differential entropy provided in Proposition 5 and Theorem 3 are not necessarily minimax optimal in our pairwise comparison setting. We have simply shown that non-trivial (minimax) estimation upper bounds on these estimation problems can be obtained from skill distribution estimation bounds. We leave the development of tight minimax estimation guarantees for various statistics and functionals of \( P_\nu \), such as \( h(P_\nu) \) (cf. [19], [59], [60]), as well as the analysis of various non-parametric entropy estimators, such as resubstitution or splitting data estimators [19], in our pairwise comparison setting for future work.

IV. MINIMAX LOWER BOUNDS VIA GENERALIZED FANO’S METHOD

In this section, we establish the minimax bounds in Theorem 1. In order to simplify the exposition, we first establish the upper bound on \( I(\pi; Z) \) in Proposition 2 in Section IV-A, present the generalized Fano’s method in Section IV-B, derive some useful auxiliary lemmata in Section IV-C, and then present the proofs of Theorem 1 and Proposition 3 in Sections IV-D and IV-E, respectively.

A. Covering Numbers and the Proof of Proposition 2

Proposition 2 considers the mutual information between a continuous random variable and a discrete random variable. Since typical definitions of mutual information, e.g., in [61], only consider mutual information between discrete random variables or between continuous random variables, for the benefit of some readers, we commence by presenting the general definition of mutual information. Recall that for any two probability measures \( \mu \) and \( \nu \) over the same measurable space \((\Omega, \mathcal{F})\), the Kullback-Leibler (KL) divergence (or relative entropy) of \( \nu \) from \( \mu \) is defined as [62, Definition 2.1]:

\[
D(\mu||\nu) = \begin{cases} \int_\Omega \log \left( \frac{d\mu}{d\nu} \right) d\mu, & \mu \text{ is absolutely continuous with respect to } \nu \\ +\infty, & \text{otherwise} \end{cases}
\]

(20)

where \( \frac{d\mu}{d\nu} \) is the Radon-Nikodym derivative (or density) of \( \mu \) with respect to \( \nu \). Using (20), for any pair of jointly distributed general random variables \( X \) and \( Y \), we define the mutual information between \( X \) and \( Y \) as [62, Definition 3.1]:

\[
I(X; Y) = D(P_{X,Y} || P_X \otimes P_Y)
\]

(21)

where \( P_{X,Y} \) denotes the joint probability law of \( X \) and \( Y \), and \( P_X \otimes P_Y \) denotes the product measure of the corresponding marginal probability laws of \( X \) and \( Y \), respectively. Note that standard properties of mutual information for discrete random variables, such as the chain rule [61, Theorem 2.5.2] and the data processing inequality [61, Theorem 2.8.1], continue to hold for mutual information of general random variables, cf. [62, Theorem 3.7]. We will utilize some of these properties to prove Proposition 2.

Another set of ideas we will exploit to prove Proposition 2 concerns a powerful and general approach to upper bound mutual information (or more generally, Shannon capacity [62, Sections 5.2-5.3], [61, Chapter 7]) via covering arguments. While there are several variants of such arguments in the literature, in this paper, we will resort to the classical covering argument of [54]. (We refer readers to [13, Section 5] for generalizations of such covering arguments for a class of f-informativities [63].)

Recall the formal setup of our problem in Sections I and II. To present the technique in [54], let us condition on any fixed realization of the underlying Erdős-Rényi random graph \( G(n, p) = G \). Then, the partial observations \( Z \), defined in (3), can be equivalently represented using the random variable:

\[
Z_G \triangleq \{ Z((i, j) : (i, j) \in G \text{ for } i, j \in [n] \text{ with } i < j \}
\]

(22)

where the notation \( \{ i, j \} \in G \) shows that the undirected edge \( \{ i, j \} \) exists in the graph \( G \), and each \( Z((i, j) = \frac{1}{\xi} \sum_{m=1}^M Z(m, (i, j)) \) given \( (i, j) \in G \) (where the likelihoods of the \( Z(m, (i, j)) \)’s are defined via (1)). Moreover, using (22), we have the relation:

\[
I(\alpha_1, \ldots, \alpha_n; Z|G(n, p) = G) = I(\alpha_1, \ldots, \alpha_n; Z_G)
\]

(23)

where the left hand side is well-defined since \( G(n, p) \) is a discrete random variable (cf. [62, Definition 3.6]), and we use the fact that the random variables \( \alpha_1, \ldots, \alpha_n, Z_G \) are independent of \( G(n, p) \). (Note that if \( G \) contains no edges, then \( Z_G \) is a deterministic quantity and \( I(\alpha_1, \ldots, \alpha_n; Z_G) = 0 \).) Now consider the \( n \)-dimensional hypercube \([0, 1]^n\) in which the skill parameter random variables \( \alpha_1, \ldots, \alpha_n \) take values. For any parameter vector (realization) \( \beta = (\beta_1, \ldots, \beta_n) \in [0, 1]^n \), let \( P_{Z_G|\beta} \) denote the conditional probability distribution of \( Z_G \) given \( \alpha_i = \beta_i \) for all \( i \in [n] \) (with abuse of notation); see (1) and (3). Then, for any \( \varepsilon > 0 \), we define an \( \varepsilon \)-covering of \([0, 1]^n\) with finite cardinality \( M \in \mathbb{N} \) to be a subset of parameter vectors \( \{ \beta^{(1)}, \ldots, \beta^{(M)} \} \subset [0, 1]^n \) that satisfies:

\[
\forall \beta \in [0, 1]^n, \exists i \in [M], D( P_{Z_G|\beta} \| P_{Z_G|\beta^{(i)}} ) \leq \varepsilon
\]

(24)

where we use KL divergence as our “distance” measure. Furthermore, for every \( \varepsilon > 0 \), we define the \( \varepsilon \)-covering number as:

\[
M^*(\varepsilon) \triangleq \min \left\{ M \in \mathbb{N} : \exists \text{\varepsilon-covering of } [0, 1]^n \text{ with cardinality } M \right\}
\]

(25)

The next lemma distills the upper bound on mutual information via covering numbers presented in [54, Equation (2), p.1571], and specializes it to our setting of (23).

**Lemma 1 (Covering Number Bound [55, Lemma 16.1]):**

Let \( (\alpha_1, \ldots, \alpha_n) \) be i.i.d. with distribution \( P_\alpha = \text{unif}([0, 1]) \), and recall that the conditional probability distribution of \( Z_G \) given \((\alpha_1, \ldots, \alpha_n)\) is defined by (1) and (3). Then, the mutual information between \((\alpha_1, \ldots, \alpha_n)\) and \( Z_G \) is upper bounded by:

\[
I(\alpha_1, \ldots, \alpha_n; Z_G) \leq \inf_{\varepsilon > 0} \varepsilon + \log(M^*(\varepsilon)).
\]

Using Lemma 1, we can finally prove the upper bound on \( I(\pi; Z) \) in Proposition 2.
Proof of Proposition 2: First, notice that \( \pi \rightarrow (\alpha_1, \ldots, \alpha_n) \rightarrow Z \) forms a Markov chain, because \( \pi \) is a deterministic function of \((\alpha_1, \ldots, \alpha_n)\) (according to (15)). Hence, by the data processing inequality [62, Theorem 3.7], we get:

\[
I(\pi; Z) \leq I(\alpha_1, \ldots, \alpha_n; Z). \tag{26}
\]

Furthermore, notice that \( Z \rightarrow (Z, \mathcal{G}(n, p)) \rightarrow (\alpha_1, \ldots, \alpha_n) \) forms a Markov chain, because \( Z \) is a deterministic (projection) function of \((Z, \mathcal{G}(n, p))\). Thus, by the data processing inequality [62, Theorem 3.7], we also get:

\[
I(\alpha_1, \ldots, \alpha_n; Z) \leq I(\alpha_1, \ldots, \alpha_n; Z, \mathcal{G}(n, p))
\]

\[
= I(\alpha_1, \ldots, \alpha_n; \mathcal{G}(n, p)) + I(\alpha_1, \ldots, \alpha_n; Z | \mathcal{G}(n, p))
\]

\[
= I(\alpha_1, \ldots, \alpha_n; \mathcal{G}(n, p)) \tag{27}
\]

where the conditional mutual information term is well-defined since \( \mathcal{G}(n, p) \) is a discrete random variable (cf. [62, Definition 3.6]), the second equality utilizes the chain rule [62, Theorem 3.7], and the third equality holds because \((\alpha_1, \ldots, \alpha_n) \) and \( \mathcal{G}(n, p) \) are independent, which implies that \( I(\alpha_1, \ldots, \alpha_n; \mathcal{G}(n, p)) = 0 \) (see (21)). Combining (26) and (27), we obtain:

\[
I(\pi; Z) \leq I(\alpha_1, \ldots, \alpha_n; Z | \mathcal{G}(n, p)). \tag{28}
\]

So, it suffices to upper bound the conditional mutual information \( I(\alpha_1, \ldots, \alpha_n; Z | \mathcal{G}(n, p)) \).

To this end, we condition on any fixed realization of the underlying Erdős-Rényi random graph \( \mathcal{G}(n, p) = G \) (as in the discussion preceding this proof), and proceed to establishing an upper bound on \( I(\alpha_1, \ldots, \alpha_n; Z | G) \) by employing Lemma 1. Specifically, we next evaluate the right hand side of the inequality in Lemma 1 above for a judiciously chosen \( \varepsilon \)-covering of \([\delta, 1]^n\). Fix any \( \eta > 0 \) (to be chosen later), and quantize the interval \([\delta, 1]\) using the set of values:

\[
\mathcal{Q} \triangleq \left\{ \delta + \frac{(1-\delta) m}{n^\eta} : m \in \lbrack \lfloor n^\eta \rfloor \rbrack \right\}
\]

which has cardinality \( |\mathcal{Q}| = \lfloor n^\eta \rfloor \), and satisfies the condition:

\[
\forall t \in [\delta, 1], \min_{s \in \mathcal{Q}} |t-s| \leq \frac{1-\delta}{n^\eta} \tag{29}
\]

where the right hand side can be improved to \((1-\delta)/(2n^\eta)\) when \( t \) is not located at the edges of the interval \([\delta, 1]\). The next claim shows that \( Q^n \) is actually an \( \varepsilon \)-covering of \([\delta, 1]^n\) with \( \varepsilon = O(n^{2-2\eta}) \) (neglecting the dependence of \( \varepsilon \) on \( \delta \) and \( k \)).

Claim 1 (\( \varepsilon \)-Covering): \( Q^n \) is an \( \varepsilon \)-covering of \([\delta, 1]^n\) with cardinality \( |Q^n| = \lfloor n^\eta \rfloor^n \leq n^{\eta n} \) and:

\[
\varepsilon = \frac{(1-\delta)^2}{4\delta^2} \left( 2 + \delta + \frac{1}{\delta} \right) \frac{|G|}{n^{2\eta}}
\]

where \( |G| \) denotes the number of edges in the graph \( G \) with abuse of notation. (Since \( |G| \leq \frac{n(n-1)}{2}, \varepsilon = O(n^{2-2\eta}) \).)

To prove this claim, we will need the following coordinate-wise Lipschitz continuity property.

Claim 2 (Coordinate-Wise Lipschitz Continuity): Consider the map \( F : [\delta, \infty)^2 \rightarrow (0, \infty) \):

\[
\forall x, y \geq \delta, \quad F(x, y) \equiv \frac{x}{x + y}
\]

which is used to define the likelihoods of the BTL model in (1). This map is coordinate-wise Lipschitz continuous:

1. For any fixed \( x \in [\delta, \infty) \) and all \( y_1, y_2 \in [\delta, \infty) \):

\[
|F(x, y_1) - F(x, y_2)| \leq \frac{1}{4\delta} |y_1 - y_2|
\]

2. For any fixed \( y \in [\delta, \infty) \) and all \( x_1, x_2 \in [\delta, \infty) \):

\[
|F(x_1, y) - F(x_2, y)| \leq \frac{1}{4\delta} |x_1 - x_2|
\]

Claim 2 is established in Appendix A-D. We next derive Claim 1 using Claim 2.
the sixth inequality follows from a simple upper bound on
\( g_1 \) as before), the largest inequality follows from Claim 2,
and the tenth inequality follows from (30). This establishes
the claim.
Finally, proceeding with the proof of Proposition 2, using
Lemma 1 and Claim 1, we get:
\[
I(\alpha_1, \ldots, \alpha_n; \gamma) \\
\leq (1 - \delta)^2 \left( 2 + \delta + \frac{1}{\delta} \right) \frac{k|G|}{n^{2\delta}} + qn \log(n).
\]
Then, by taking expectations on both sides of this inequality
with respect to the law of \( G(n, p) \), we obtain using (23)
and (28) that:
\[
I(\gamma; Z) \leq (1 - \delta)^2 \left( 2 + \delta + \frac{1}{\delta} \right) \frac{kp(n - 1)}{n^{2\delta}} + qn \log(n) \\
\leq qn \log(n) + (1 - \delta)^2 \left( 2 + \delta + \frac{1}{\delta} \right) kp^{2-2\delta}.
\]
Neglecting \( \delta, k, \) and \( p \), it is clear that setting \( q \geq \frac{1}{2} \) ensures that
this upper bound is \( O(n \log(n)) \). As the proof of Theorem 1
in Section IV-D illustrates, choosing the smallest \( q \) satisfying
\( q \geq \frac{1}{2} \) yields the tightest possible minimax lower bound using
this approach. Thus, we set \( q = \frac{1}{2} \) in the above bound and get:
\[
I(\gamma; Z) \leq \frac{1}{2} n \log(n) + \frac{(1 - \delta)^2}{8}\delta^2 \left( 2 + \delta + \frac{1}{\delta} \right) kp^n
\]
as desired.
In the literature, various information inequalities are
often used to upper bound mutual information terms like
\( I(\alpha_1, \ldots, \alpha_n; Z_G) \), for any realization \( G(n, p) = G \), in a
simpler manner (cf. [13, Equation (44)]). However, these
approaches tend to yield poorer scaling in the upper bound
with \( n \) compared to the covering number argument we utilize
(in Lemma 1). For example, the convexity of KL divergence
immediately yields the bound (see [13, Equation (44)] or [52,
p.1319]):
\[
I(\alpha_1, \ldots, \alpha_n; Z_G) \\
\leq \frac{1}{(1 - \delta)^2} \left( 2 + \delta + \frac{1}{\delta} \right) kp^n D(P_{Z_G}) D(P_{Z_G}) D(\gamma) d\beta d\gamma \\
\leq \sup_{\beta, \gamma} D(P_{Z_G}) D(P_{Z_G}) D(\gamma) \\
= k \sup_{\beta, \gamma} \sum_{i,j} \left( \frac{\beta_i \gamma_j}{\beta_j \gamma_i} \right) \\
\leq k|G| \max_{a,b} D(\gamma) \\
= k|G| D(\gamma) \left( 1 - \delta \right)^2 \left( 1 - \delta \right)^2.
\]
where the third inequality follows from (31) and we use the
notation \( \beta = (\beta_1, \ldots, \beta_n) \) and \( \gamma = (\gamma_1, \ldots, \gamma_n) \) as before,
the maximization in the fourth inequality is over \( a, b \in \left[ \frac{1}{2}, \frac{1}{2} \right] \)
because the map \( 0 < x \mapsto x/(c+x) \) is monotone increasing
for every fixed \( c > 0 \), and the last equality follows from
basic properties of binary KL divergence (see, e.g., [62,
Example 2.11]). As before, by taking expectations on both sides
of (32) with respect to the law of \( G(n, p) \), we obtain using (23)
and (28) that:
\[
I(\gamma; Z) \leq \frac{kpn(n - 1)}{2} D \left( 1 + \delta \right) \left( 1 + \delta \right).
\]
Neglecting \( \delta, k, \) and \( p \), it is clear that \( I(\gamma; Z) = O(n^2) \) in (33),
but the proof of Proposition 2 gives the sharper estimate
\( I(\gamma; Z) = O(n \log(n)) \).
B. Generalized Fano’s Method

We next introduce a canonical approach to obtaining minimax lower bounds in non-parametric estimation problems—the so called Fano’s method, which was introduced in [25], [26] (also see, e.g., [51] and [3, Section 2.7.1] for modern treatments). Fano’s method proceeds by first lower bounding minimax risk by a Bayes risk, where all the prior probability mass is placed over a suitably chosen (and large) finite set of parameters in the (non-parametric or infinite-dimensional) parameter space, then lower bounding this Bayes risk using the probability of error of a multiple hypothesis testing problem, and finally, lower bounding this probability of error using the well-known Fano’s inequality, cf. [61, Theorem 2.10.1]. In the problem of estimating (15) based on $Z$, the parameter space of the minimax risk in Theorem 1 is the infinite-dimensional family of PDFs $P$. For simplicity and analytical tractability, instead of directly applying Fano’s method to this large parameter space $P$, which would involve constructing a prior distribution over some judiciously chosen finite subset of $P$, we first obtain a lower bound on the minimax risk in Theorem 1 in terms of Bayes risk. In particular, as discussed earlier, we set $P_{\alpha} = \text{unif}([\delta, 1]) \in P$ throughout Section IV so that $\alpha_1, \ldots, \alpha_n$ are i.i.d. $P_{\alpha} = \text{unif}([\delta, 1])$. Hence, $P(\cdot)$ denotes the joint probability law of $\alpha_1, \ldots, \alpha_n$, $G(n, p)$, and $\{Z_{m}(i, j) : \{i, j\} \in G(n, p), m \in [k]\}$ with $P_{\alpha} = \text{unif}([\delta, 1])$ in the sequel, and $E[\cdot]$ denotes the corresponding expectation operator. This yields the following lower bound on the minimax relative $\ell^1$-norm risk in Theorem 1 for any $q \in [1, \infty)$:

$$\inf \sup P_{\alpha_1} \left[ \frac{\|\hat{\pi} - \pi\|_q}{\|\pi\|_q} \right] \geq \inf \mathbb{E}_\pi \left[ \frac{\|\hat{\pi} - \pi\|_q}{\|\pi\|_q} \right]$$

(34)

where the infimum is over all (measurable) randomized estimators $\hat{\pi} \in S_n$ of the canonically scaled skill parameters $\pi$ based on the observation matrix $Z$, and $E_{\pi}[\cdot]$ denotes the expectation operator with respect to general (not necessarily uniform) $P_{\alpha}$. Therefore, we can focus on the simpler problem of lower bounding the Bayes risk on the right hand side of (34) for any $q \in [1, \infty]$.

Unfortunately, while Fano’s method is very effective at lower bounding non-parametric minimax risks, it cannot lower bound Bayes risks where the parameter space is not a discrete and finite set, because the classical Fano’s inequality only holds for discrete and finite parameter sets [61, Theorem 2.10.1]. To remedy this dearth of Fano-based techniques to lower bound Bayes risks where the parameter space is a continuum, the so called generalized Fano’s method has been developed in the recent literature [13], [14], [52]. One of the first results in this line of work was a generalization of Fano’s inequality to the continuum Fano inequality in [53, Proposition 2], which had useful consequences for minimax estimation with a specific zero-one valued loss function [53, Section 3]. The techniques in [53] have been vastly generalized by [13] and [14] to obtain lower bounds on Bayes risks in terms of $f$-informativity [63] and conditional mutual information (with auxiliary random variables), respectively. In this paper, we will utilize the key result in [14, Theorem 1, Equation (6)]. The lemma below presents the result in [14, Theorem 1, Equation (6)] specialized to our relative $\ell^q$-loss setting.

Lemma 2 (Generalized Fano’s Method [14, Theorem 1]):

For any $q \in [1, \infty]$, the Bayes risk on the right hand side of (34) is lower bounded by:

$$\inf P_{\alpha} \mathbb{E} \left[ \frac{\|\hat{\pi} - \pi\|_q}{\|\pi\|_q} \right] \geq \inf P_{\alpha} \mathbb{E} \left[ \frac{\|\hat{\pi} - \pi\|_q}{\|\pi\|_q} \right] \geq \sup_{t > 0} \left( 1 - \frac{I(\pi; Z) + \log(2)}{\log(1/L_q(t))} \right)$$

(35)

where we define the small ball probability $L_q(t)$ as [14, Equation (2)]:

$$\forall \epsilon > 0, \ L_q(t) \triangleq \sup_{\pi \in S_n} \mathbb{P}\left( \frac{\|\pi - \nu\|_q}{\|\pi\|_q} \leq t \right)$$

and $I(\pi; Z)$ denotes the mutual information (defined in (21)) between the canonically scaled skill parameters $\pi$ (defined in (15)) and the observation matrix $Z$ (defined in (3)).

We remark that several variants of Lemma 2 exist in the literature, such as [52, Theorem 6.1] and [13, Remark 10, Corollary 12(i)]. As expounded in [13], in order to compute lower bounds such as that in Lemma 2, we need to establish two things:

1) Tight upper bounds on the mutual information $I(\pi; Z)$.

2) Tight upper bounds on the small ball probability $L_q(t)$.

We have already derived an upper bound on $I(\pi; Z)$ in Proposition 2 using the covering number argument presented in Section IV-A. Next, we prove an upper bound on the small ball probability $L_q(t)$ for $q \in [1, \infty]$.

C. Upper Bound on Small Ball Probability

As noted by both [13] and [14], there is no general recipe for obtaining upper bounds on $L_q(t)$. So, we develop our bound via direct computation. To this end, the ensuing lemma presents an upper bound on the mode of the joint PDF of $\pi$, or more precisely, the joint PDF of:

$$\hat{\pi} \triangleq (\pi(1), \ldots, \pi(n - 1))$$

(36)

with respect to the Lebesgue measure on $\mathbb{R}^{n-1}$, which excludes $\pi(n)$, because $\pi(n) = 1 - \pi(1) - \cdots - \pi(n - 1)$.

Lemma 3 (Bound on Mode of Joint PDF of $\hat{\pi}$): Let the joint PDF of $\hat{\pi}$ with respect to the Lebesgue measure on $\mathbb{R}^{n-1}$ be denoted $P_{\hat{\pi}}$. The mode of $P_{\hat{\pi}}$ is upper bounded by:

$$\text{ess sup}_{\tau \in \mathbb{R}^{n-1}} P_{\hat{\pi}}(\tau) \leq \frac{n^{n-1}}{(1 - \delta)^n}$$

where ess sup denotes the essential supremum.

Proof: First, consider the map $h : [\delta, 1]^n \rightarrow \text{im}(h)$:

$$h(\beta) \triangleq \left( \frac{\beta_1}{\sum_{i=1}^n \beta_i}, \ldots, \frac{\beta_{n-1}}{\sum_{i=1}^n \beta_i}, \frac{n}{\sum_{i=1}^n \beta_i} \right)$$

where $\text{im}(h) \triangleq \left\{ (\tau_1, \ldots, \tau_{n-1}, 1) : \exists \beta_1, \ldots, \beta_n \in [\delta, 1] \text{ such that } \tau_i = \beta_i/\sigma \text{ for all } i \in [n-1] \text{ and } \sigma = \sum_{i=1}^n \beta_i \right\}$ denotes the range (or image) of $h$. Clearly, we have $h(1, 1, \ldots, 1) = (\pi, \alpha_1 + \cdots + \alpha_n)$.
\( \alpha_n \) using (15) and (36). Furthermore, \( h \) is a bijection with inverse function \( h^{-1} : \text{im}(h) \to [\delta, 1]^n \):

\[
\forall (\tau_1, \ldots, \tau_{n-1}, \sigma) \in \text{im}(h), \quad h^{-1}(\tau_1, \ldots, \tau_{n-1}, \sigma) = \left( \sigma \tau_1, \ldots, \sigma \tau_{n-1}, \sigma \left( 1 - \sum_{i=1}^{n-1} \tau_i \right) \right).
\]

By direct evaluation, the Jacobian matrix of \( h^{-1} \), denoted \( \nabla h^{-1} : \text{im}(h) \to \mathbb{R}^{n \times n} \), is:

\[
[\nabla h^{-1}]_{i,j} = \begin{cases} 
\sigma \mathbb{1}(i = j), & i, j \in [n - 1] \\
-\sigma, & i = n, j \in [n - 1] \\
\tau_i, & i \in [n - 1], j = n \\
1 - \sum_{i=1}^{n-1} \tau_i, & i = j = n
\end{cases}
\]

for all \((\tau_1, \ldots, \tau_{n-1}, \sigma) \in \text{im}(h)\), where \([\nabla h^{-1}]_{i,j}\) denotes the \((i, j)\)th entry of the matrix \(\nabla h^{-1}\) for \(i, j \in [n]\). (Note that \(\nabla h^{-1}\) is also well-defined on the boundary of \(\text{im}(h)\), because there exists an open set containing \(\text{im}(h)\) such that the first partial derivatives of \(h^{-1}\) exist on this open set.) Now, for \(r \in \{0, 1, \ldots, n - 2\} \), define the successive sub-matrices:

\[
M_{n-r} \triangleq \begin{bmatrix}
1 & 0 & \cdots & 0 & \tau_{r+1} \\
0 & 1 & \cdots & 0 & \tau_{r+2} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 1 & \tau_{n-1} \\
-1 & -1 & \cdots & -1 & 1 - \sum_{i=1}^{n-1} \tau_i
\end{bmatrix} \in \mathbb{R}^{(n-r) \times (n-r)}
\]

where \(M_n\) is closely related to \(\nabla h^{-1}\) (as shown below in (37)), and let the transpose of the Frobenius companion matrix of the monic polynomial \(q_n(t) = 1 + t + t^2 + \cdots + t^n\) be [68, Definition 3.3.13]:

\[
C_n \triangleq \begin{bmatrix}
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1 \\
-1 & -1 & -1 & \cdots & -1
\end{bmatrix} \in \mathbb{R}^{n \times n}.
\]

Then, the corresponding Jacobian determinant satisfies the recurrence relation:

\[
\det(\nabla h^{-1}) = \det \left( \begin{bmatrix}
\sigma & 0 & \cdots & 0 & \tau_1 \\
0 & \sigma & \cdots & 0 & \tau_2 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & \sigma & \tau_{n-1} \\
-\sigma & -\sigma & \cdots & -\sigma & 1 - \sum_{i=1}^{n-1} \tau_i
\end{bmatrix} \right)
\]

\[
= \sigma^{n-1} \det(M_n) 
\]

\[
= \sigma^{n-1} (\det(M_{n-1}) + (-1)^{n-1} \tau_1 \det(C_{n-1}))
\]

for all \((\tau_1, \ldots, \tau_{n-1}, \sigma) \in \text{im}(h)\), where (37) follows from the multilinearity of the determinant, and (38) uses the Laplace (cofactor) expansion of determinants by minors along the first row [68, Section 0.3.1]. We next compute this Jacobian determinant.

It is easy to calculate \(\det(C_{n-1})\) in (38), because \(q_{n-1}(t)\) is also the characteristic polynomial of its (adjoint) companion matrix \(C_{n-1}\) [68, Theorem 3.3.14]. The \(n - 1\) distinct roots of \(q_{n-1}(t)\) are the following \(n\)th roots of unity:

\[
\forall r \in [n - 1], \quad q_{n-1}(\omega^r) = 0
\]

where \(\omega = \exp\left(\frac{2\pi i}{n}\right)\). (Note that unlike the rest of this paper, in the definition of \(\omega\), we use \(i \) and \(\pi\) to represent the imaginary unit \(i = \sqrt{-1}\) and the mathematical constant \(\pi = 3.14159\ldots\), respectively.) Hence, \(\{\omega^r : r \in [n - 1]\}\) are the eigenvalues of \(C_{n-1}\), and we have:

\[
\det(C_{n-1}) = \prod_{r=1}^{n-1} \omega^r = \omega^{n(n-1)/2} = (-1)^{n-1}
\]

since the determinant is the product of the eigenvalues [68, Section 1.2]. Combining this with (38), for all \((\tau_1, \ldots, \tau_{n-1}, \sigma) \in \text{im}(h)\), we get:

\[
\det(\nabla h^{-1}) = \sigma^{n-1} (\det(M_{n-1}) + (-1)^{n+1} \tau_1 (-1)^{n-1})
\]

\[
= \sigma^{n-1} (\det(M_{n-1}) + \tau_1)
\]

\[
= \sigma^{n-1} \left( \det(M_2) + \sum_{j=1}^{n-2} \tau_j \right)
\]

\[
= \sigma^{n-1} \left( 1 - \sum_{j=1}^{n-1} \tau_j + \tau_{n-1} + \sum_{j=1}^{n-2} \tau_j \right)
\]

\[
= \sigma^{n-1} \left( 1 - \sum_{j=1}^{n-1} \tau_j \right)
\]

where the third equality follows from unwinding the recursion in the second line.

Now observe that the joint PDF of \(\langle \alpha_1, \ldots, \alpha_n \rangle\) (with respect to the Lebesgue measure on \(\mathbb{R}^n\)) is given by:

\[
\forall \beta \in \mathbb{R}^n, \quad P_{\alpha_1, \ldots, \alpha_n}(\beta) = \frac{1}{(1 - \delta)^n} I\{\beta \in [\delta, 1]^n\}
\]

since \(\alpha_1, \ldots, \alpha_n\) are i.i.d. \(P_\alpha = \text{unif}([\delta, 1])\). As a consequence, the joint PDF of \(h(\alpha_1, \ldots, \alpha_n) = (\tilde{p}, \alpha_1 + \cdots + \alpha_n)\) (with respect to the Lebesgue measure on \(\mathbb{R}^n\)) is given by the change-of-variables formula:

\[
P_{\tilde{p}, \alpha_1 + \cdots + \alpha_n}(\tau_1, \ldots, \tau_{n-1}, \sigma)
\]

\[
= P_{\alpha_1, \ldots, \alpha_n}(h^{-1}(\tau_1, \ldots, \tau_{n-1}, \sigma)) \det(\nabla h^{-1})
\]

\[
= \frac{\sigma^{n-1}}{(1 - \delta)^n} I\{\tau_1, \ldots, \tau_{n-1}, \sigma \in \text{im}(h)\}
\]

(40)

for all \((\tau_1, \ldots, \tau_{n-1}, \sigma) \in \mathbb{R}^n\), where we utilize our earlier computation of the Jacobian determinant in (39). Although we only seek to bound the joint PDF of \(\tilde{p}\), the joint PDF in (40) includes an additional random variable \(\alpha_1 + \cdots + \alpha_n\) as an artifact of our calculation approach (which requires an invertible map \(h\) with a well-defined and invertible Jacobian matrix \(\nabla h\)).

So, in the final step of this proof, we marginalize the joint PDF in (40) and then bound the desired joint PDF of \(\tilde{p}\) (with
respect to the Lebesgue measure on $\mathbb{R}^{n-1}$: 
\[
P_\varepsilon(\tau) = \mathbb{I}\{\tau \in \tilde{S}_n\} \int_{[n,\delta,n]} P_{\pi,\alpha_1,\ldots,\alpha_n}(\tau, \sigma) \, d\sigma 
\]
\[
= \frac{1}{(1-\delta)^n} \int_{[n,\delta,n]} \sigma^{n-1} \mathbb{I}\{\tau(\sigma) \in \text{im}(h)\} \, d\sigma 
\]
\[
\leq \frac{1}{(1-\delta)^n} \int_{[n,\delta,n]} \sigma^{n-1} \, d\sigma 
\]
\[
= n^{n-1}(1-\delta)^n \mathbb{I}\{\tau \in \tilde{S}_n\} 
\]
\[
\leq n^{n-1}(1-\delta)^n \mathbb{I}\{\tau \in \tilde{S}_n\}, 
\]
where $\tilde{S}_n \triangleq \{(\tau_1, \ldots, \tau_{n-1}) \in \left[n-\frac{\delta}{1-\delta}, 1+\frac{1}{\delta(n-1)}\right]^{n-1} : 
\exists \beta_1, \ldots, \beta_n \in [\delta, 1] \text{ such that } \tau_i = \beta_i / \left(\sum_{j=1}^{n} \beta_j\right) \text{ for all } i \in [n-1]\}$. Taking the (essential) supremum over all $\tau \in \mathbb{R}^{n-1}$ in the above bound completes the proof. 

We now use Lemma 3 to upper bound the small ball probability $L_q(t)$ for $q \in [1, \infty]$ in the lemma below.

**Lemma 4 (Upper Bound on Small Ball Probability):** For any $q \in [1, \infty]$ and every $t > 0$, we have:

\[
L_q(t) \leq (\frac{8}{\delta(1-\delta)})^n t^{n-1}. 
\]

**Proof:** Starting with (35), observe that for all $t > 0$:

\[
L_q(t) = \mathbb{P}_{\nu \in S_n}\left(\|\nu - \hat{\nu}\|_q \leq t \|\pi\|_q\right) 
\]
\[
\leq \mathbb{P}_{\nu \in S_n}\left(\|\nu - \hat{\nu}\|_q \leq \frac{t}{\delta(n(q-1)/q)\|\pi\|_q}\right) 
\]
\[
\leq \mathbb{P}_{\nu \in S_n}\left(\|\hat{\pi} - \hat{\nu}\|_q \leq \frac{t}{\delta(n(q-1)/q)\|\pi\|_q}\right) 
\]
\[
= \mathbb{P}_{\nu \in S_n} \left(\int_{\mathbb{R}^{n-1}} P_\varepsilon(\tau) \left\{\|\tau - \hat{\nu}\|_q \leq \frac{t}{\delta(n(q-1)/q)\|\pi\|_q}\right\} d\tau \right) 
\]
\[
\leq \frac{n^{n-1}}{(1-\delta)^n} \sup_{\nu \in S_n} \left(\int_{\mathbb{R}^{n-1}} \mathbb{I}\{\|\tau - \hat{\nu}\|_q \leq \frac{t}{\delta(n(q-1)/q)\|\pi\|_q}\} d\tau \right) 
\]
\[
= \frac{n^{n-1}}{(1-\delta)^n} \int_{\mathbb{R}^{n-1}} \mathbb{I}\{\|\tau - \hat{\nu}\|_q \leq \frac{t}{\delta(n(q-1)/q)\|\pi\|_q}\} d\tau 
\]
\[
= \frac{n^{n-1}}{(1-\delta)^n} \Gamma\left(\frac{1}{q} + 1\right)^{n-1} \frac{(2t)^n}{\delta^{n-1}n(n(q-1)/q)} 
\]
\[
\leq \left(\frac{2}{\delta(1-\delta)}\right)^n t^{n-1} \frac{\Gamma\left(\frac{1}{q} + 1\right)^{n-1} n(n(q-1)/q)}{\Gamma\left(\frac{n-1}{q} + 1\right)} 
\]
\[
\leq \frac{e}{\sqrt{2\pi}} \left(\frac{2}{\delta(1-\delta)}\right)^n \left(\frac{2}{\delta(1-\delta)}\right)^{n-1} t^{n-1} 
\]
\[
\leq \frac{8}{\delta(1-\delta)} t^{n-1}, 
\]
where the second inequality uses the bound:

\[
\|\pi\|_q \leq n^{1/q} \|\pi\|_\infty = \frac{n^{1/q} \max_{i \in [n]} \alpha_i}{\sum_{i=1}^{n} \alpha_i} \leq \frac{1}{\delta(n(q-1)/q)} 
\]

which follows from (15) and the fact that $\alpha_1, \ldots, \alpha_n \in [\delta, 1]$, the third inequality uses (36) and the fact that $\|\pi - \nu\|_q \geq \|\pi - \hat{\nu}\|_q$ where $\hat{\nu} = (\nu_1, \ldots, \nu_{n-1})$, the fifth inequality follows from Lemma 3, the sixth equality uses the well-known volume of the $\ell^q$-ball with radius $t/(\delta n(q-1)/q)$, cf. [69], where $\Gamma : (0, \infty) \rightarrow \mathbb{R}$ denotes the gamma function, the seventh inequality follows from the fact that $2/\delta \geq 1$, the ninth inequality holds because $2e^{1/12} \sqrt{\pi} < 4$ and $e/(4\sqrt{2\pi}) < 1$, and the eighth inequality holds because:

\[
\Gamma\left(\frac{1}{q} + 1\right)^{n-1} n(n(q-1)/q) 
\]
\[
\leq \frac{e}{\sqrt{2\pi}} \left(\frac{2}{\delta(1-\delta)}\right)^n \left(\frac{1}{q} + 1\right)^{n-1} n(n(q-1)/q)(12(1+1/q)) 
\]
\[
\leq \frac{e}{\sqrt{2\pi}} \left(\frac{2}{\delta(1-\delta)}\right)^n \left(1 + \frac{1}{q}\right)^{n-1} 
\]

which follows from the Stirling’s formula bounds [70, Theorem 5] (also see, e.g., [71, Chapter II, Section 9, Equation (9.15)]):

\[
\forall x > 0, \sqrt{2\pi} x^{x-1/2} e^{-x} \leq \Gamma(x) \leq \sqrt{2\pi} x^{x-1/2} e^{-x} e^{1/(12x)}, 
\]

and the facts that $q \in [1, \infty]$, $n \geq 2$, $(1 + q^{-1})(n(q-1)/2)/[1 + (n-1)q^{-1}] \leq 2(n^{-1}/2, n^{-1}/q + 1/n) \geq 1/q$, and $(1 + q^{-1})/q \leq e$ (which is equivalent to log$(1 + q^{-1}) \leq q$). This completes the proof.

We note that the constant $8/(\delta(1-\delta))$ in Lemma 4 can obviously be improved for various special cases of $q \in [1, \infty]$. For example, when $q = \infty$, we can use the simplified formula for the volume of a hypercube (or $\ell^\infty$-norm ball) to obtain the bound (cf. (41)):

\[
\forall t > 0, L_\infty(t) \leq \left(\frac{2}{\delta(1-\delta)}\right)^n t^{n-1}. 
\]

Similarly, when $q = 1$, we can use the simplified formula for the volume of a cross-polytope (or $\ell^1$-norm ball) and Stirling’s formula (cf. [71, Chapter II, Section 9, Equation (9.15)]) to obtain the bound:

\[
\forall t > 0, L_1(t) \leq \left(\frac{2}{1-\delta}\right)^n t^{n-1}. 
\]

Next, we provide proofs of Theorem 1 and Proposition 3 using Proposition 2, Lemmata 2 and 4, and the results in [9, Theorems 3.1 and 5.2].

**D. Proof of Theorem 1**

In order to prove Theorem 1, we require the following known result from the literature [9], which upper bounds the relative $\ell^\infty$-norm loss between $\hat{\pi}_s$ defined in (7) and the canonically scaled skill parameters in (15). (As explained earlier in Section II-C, $\hat{\pi}_s$ is intuitively a good estimator of $\pi$.)
Lemma 5 (Relative $\ell^\infty$-Loss Bound [9, Theorem 3.1]): Suppose that $p \geq c_5 \log(n)/(\delta^5 n)$ for some sufficiently large constant $c_5 > 0$. Then, there exist (universal) constants $c_{12}, c_{13} > 0$ such that for all sufficiently large $n \in \mathbb{N}$, we have:

\[
\mathbb{P}\left(\frac{\|\hat{\pi}_* - \pi\|_\infty}{\|\pi\|_\infty} \leq \frac{c_{12}}{\delta} \sqrt{\frac{\log(n)}{npk}} \right) \geq 1 - \frac{c_{13}}{n^5}
\]

where the probability is computed with respect to the conditional distribution of the observation matrix $Z$ and the random graph $\mathcal{G}(n, p)$ given any realizations of the skill parameters $\alpha_1, \ldots, \alpha_n$.

We remark that the proof of Lemma 5 in [9] crucially uses the assumption that $\alpha_1, \ldots, \alpha_n \in [\delta, 1]$. We also note that the conditioning on $\alpha_1, \ldots, \alpha_n$ in Lemma 5 reflects the fact that [9] considers a non-Bayesian scenario where $\alpha_1, \ldots, \alpha_n$ are deterministic (and unknown). In contrast, this work assumes that $\alpha_1, \ldots, \alpha_n$ are drawn i.i.d. from a prior PDF $P_\alpha$.

We now proceed to establishing Theorem 1.

Proof of Theorem 1: Fix any $q \in [1, \infty]$. We first prove the minimax upper bound. The inequality:

\[
\inf_{\hat{\pi}} \sup_{P_\alpha \in \mathcal{P}} \mathbb{E}_{P_\alpha} \left[ \frac{\|\hat{\pi}_* - \pi\|_q}{\|\pi\|_q} \right] \leq \sup_{P_\alpha \in \mathcal{P}} \mathbb{E}_{P_\alpha} \left[ \frac{\|\hat{\pi}_* - \pi\|_q}{\|\pi\|_q} \right]
\]

holds trivially, because $\hat{\pi}_* \in S_\alpha$ in (7) is an estimator for $\pi$ based on $Z$. To prove an upper bound on the extremal Bayes risk on the right hand side of this inequality, we define the event in Lemma 5 as:

\[
A \triangleq \left\{ \frac{\|\hat{\pi}_* - \pi\|_\infty}{\|\pi\|_\infty} \leq \frac{c_{12}}{\delta} \sqrt{\frac{\log(n)}{npk}} \right\}
\]

where $c_{12} > 0$ is the universal constant from Lemma 5. Then, (57) states that for any PDF $P_\alpha \in \mathcal{P}$ and for all sufficiently large $n \in \mathbb{N}$:

\[
\mathbb{P}_{P_\alpha}(A) \geq 1 - \frac{c_{13}}{n^5}
\]

(45)

where $c_{13} > 0$ is another universal constant from Lemma 5, and $\mathbb{P}_{P_\alpha}(\cdot)$ denotes the probability measure with respect to general (not necessarily uniform) $P_\alpha$. If we condition on the event $A$, then it follows that:

\[
\frac{\|\hat{\pi}_* - \pi\|_q}{\|\pi\|_q} \leq \frac{\|\hat{\pi}_* - \pi\|_\infty}{\|\pi\|_\infty} \frac{n^{(q-1)/q}}{n^{3q}} \leq \frac{\|\hat{\pi}_* - \pi\|_\infty}{\|\pi\|_\infty} \frac{1}{\delta^2} \leq \frac{c_{12}}{\delta^2} \sqrt{\frac{\log(n)}{npk}}
\]

(46)

where the first inequality follows from the bound:

\[
\|\pi\|_q \geq n^{1/q} \min_{i \in [n]} \alpha_i \geq \frac{\delta}{n^{(q-1)/q}}
\]

which follows from (15) and the fact that $\alpha_1, \ldots, \alpha_n \in [\delta, 1]$, the second inequality follows from the bound $\|\hat{\pi}_* - \pi\|_q \leq n^{1/q} \|\hat{\pi}_* - \pi\|_\infty$. The third inequality follows from the bound $\|\pi\|_\infty \leq 1/(n\delta)$ (cf. (42)), and the last inequality follows from the event $A$. Hence, for every PDF $P_\alpha \in \mathcal{P}$ and for all sufficiently large $n$, we have:

\[
\mathbb{E}_{P_\alpha} \left[ \frac{\|\hat{\pi}_* - \pi\|_q}{\|\pi\|_q} \right] = \mathbb{E}_{P_\alpha} \left[ \frac{\|\hat{\pi}_* - \pi\|_q}{\|\pi\|_q} | A \right] \mathbb{P}_{P_\alpha}(A)
\]

\[
+ \mathbb{E}_{P_\alpha} \left[ \frac{\|\hat{\pi}_* - \pi\|_q}{\|\pi\|_q} | A^c \right] \mathbb{P}_{P_\alpha}(A^c)
\]

\[
\leq \frac{c_{12}}{\delta^2} \sqrt{\frac{\log(n)}{npk}} + \frac{c_{13}}{n^5}
\]

(48)

where the first equality uses the law of total expectation, the second inequality follows from (45), (46), and the fact that:

\[
\frac{\|\hat{\pi}_* - \pi\|_q}{\|\pi\|_q} \leq \frac{n}{\delta}
\]

which follows from (47) and $\|\hat{\pi}_* - \pi\|_\infty \leq 1$, and the third inequality (48) holds for all sufficiently large $n$ because $k = \Theta(1)$. Letting $c_3 = 2c_{12}/(\delta^3 \sqrt{npk})$ and substituting it into (48), and then taking the supremum in (48) over all PDFs $P_\alpha \in \mathcal{P}$ yields the desired upper bound in the theorem statement. (We remark that when $q = \infty$, the dependence of $c_3$ on $\delta$ can be improved to $c_3 = 2c_{12}/(\delta \sqrt{pk})$ by following the same proof strategy as above.)

We next prove the information theoretic lower bound. Fix any $\varepsilon > 0$, and consider any sufficiently large $n \geq 2$ such that:

\[
n \geq \max \left\{ 2 + \frac{1}{\varepsilon} \left( \frac{8}{\delta^2} \right)^{4/\varepsilon} \right\},
\]

\[
\exp \left( \frac{(1 - \delta)^2 \left( 2 + \frac{1}{\delta^2} \right) \kappa p + 4 \log(2) \delta^2 \varepsilon}{\delta^2 \varepsilon} \right)
\]

(49)

Then, observe that:

\[
\inf_{\hat{\pi}} \sup_{P_\alpha \in \mathcal{P}} \mathbb{E}_{P_\alpha} \left[ \frac{\|\hat{\pi} - \pi\|_q}{\|\pi\|_q} \right]
\]

\[
\geq \sup_{t > 0} \left( 1 - \frac{I(\pi; Z) + \log(2)}{\log(1/L_q(t))} \right)
\]

\[
\geq \sup_{t > 0} \left( 1 - \frac{1}{2} n \log(n) + c(\delta, p, k)n + \log(2) \right)
\]

\[
\geq \sup_{t > 0} \left( 1 - \frac{1}{2} n \log(n) + c(\delta, p, k)n + \log(2) \right)
\]

\[
\geq \sup_{t > 0} \left( 1 - \frac{1}{n} \log(1/t) - \log(8/(\delta(1 - \delta)))n \right)
\]

\[
\geq \sup_{t > 0} \left( 1 - \frac{1 + c(\delta, p, k)n + \log(4)}{n \log(n)} \right)
\]

\[
\geq \frac{1}{n^{1/\varepsilon}} \left( 1 - \frac{1 + c(\delta, p, k)n + \log(4)}{n \log(n)} \right)
\]

\[
\geq \frac{1}{n^{1/\varepsilon}} \left( 1 - \frac{1 + \frac{\varepsilon}{2}}{1 + \frac{\varepsilon}{2}} \right)
\]

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
where the first inequality follows from (34) and Lemma 2, the second inequality follows from Proposition 2 and we let:

\[
c(\delta, p, k) = \frac{(1 - \delta)^2}{8 \delta^2} \left( 2 + \delta + \frac{1}{\delta} \right) kp
\]

for clarity, the third inequality holds due to Lemma 4, the fifth inequality follows from setting \( t = n^{-1/2} - \varepsilon \), and the sixth inequality follows from (49), which implies the following bounds:

\[
n \geq \exp \left( \frac{(1 - \delta)^2 (2 + \delta + \frac{1}{\delta}) kp + 4 \log(2) \delta^2}{\delta^2} \right)
\]

\[
\Rightarrow \quad 2c(\delta, p, k) \frac{\log(4)}{n \log(n)} \leq \frac{\varepsilon}{4},
\]

\[
n \geq 2 + \frac{1}{\varepsilon} \quad \Rightarrow \quad 1 - \frac{1}{n} \geq \frac{1 + \varepsilon}{1 + 2 \varepsilon},
\]

\[
n \geq \left( \frac{8}{\delta(1 - \delta)} \right)^{4/\varepsilon} \quad \Rightarrow \quad 2 \log \left( \frac{8}{\delta(1 - \delta)} \right) \leq \frac{\varepsilon}{2}.
\]

Now, let us define the constant:

\[
c_{14} = \max \left\{ 4 \log \left( \frac{8}{\delta(1 - \delta)} \right), \frac{(1 - \delta)^2 (2 + \delta + \frac{1}{\delta}) kp + 4 \log(2) \delta^2}{\delta^2} \right\}
\]

and set \( \varepsilon = c_{14}/\log(n) \). It is straightforward to verify that (49) is satisfied for this choice of \( \varepsilon \) for all sufficiently large \( n \). Moreover, since \( \varepsilon \leq 1 \) for all sufficiently large \( n \), we see that (50) can be recast as:

\[
\inf \sup P_\alpha \left[ \left\| \hat{\pi} - \pi \right\|_q \right] \geq \frac{c_{14}}{6 \log(n) n^{1/2 + \varepsilon} / \log(n)}
\]

\[
= \frac{c_{14}}{6 \log(n) n^{1/2 + \varepsilon} / \log(n)} \frac{1}{\log(n)/n}
\]

for all sufficiently large \( n \). Finally, letting \( c_2 = c_{14}/(6 \exp(c_{14})) \) yields the minimax lower bound in the theorem statement. This completes the proof. \( \blacksquare \)

**E. Proof of Proposition 3**

As noted in Section III-A, although the upper bounds in Theorem 1 are established using Lemma 5, when \( q \in [1, 2] \), we can remove an extra factor of \( \sqrt{\log(n)} \) from these upper bounds by utilizing [9, Theorem 5.2] (also see [8, Theorem 2]). So, we present this result in the lemma below.

**Lemma 6 (Relative \( \ell^2 \)-Loss Bound [9, Theorem 5.2]):** Suppose that \( \delta = \Theta(1) \) and \( p \geq c_{15} \log(n)/n \) for some sufficiently large constant \( c_{15} > 0 \) (which may depend on \( \delta \)). Then, there exists a constant \( c_{16} > 0 \) (which may depend on \( \delta \)) and a (universal) constant \( c_{17} > 0 \) such that for all sufficiently large \( n \in \mathbb{N} \), we have:

\[
P\left( \left\| \hat{\pi} - \pi \right\|_2 \geq \frac{c_{16}}{\sqrt{npk}} \right| \alpha_1, \ldots, \alpha_n \right) \geq 1 - \frac{c_{17}}{n^5}
\]

where the probability is computed with respect to the conditional distribution of the observation matrix \( Z \) and the random graph \( \mathcal{G}(n, p) \) given any realizations of the skill parameters \( \alpha_1, \ldots, \alpha_n \).

This lemma is an analog of Lemma 5, but for \( \ell^2 \)-norm instead of \( \ell^\infty \)-norm. As remarked after Lemma 5, in contrast to this work, the conditioning on \( \alpha_1, \ldots, \alpha_n \) in Lemma 6 reflects the non-Bayesian scenario considered in [9] (where \( \alpha_1, \ldots, \alpha_n \) are deterministic). We next derive Proposition 3.

**Proof of Proposition 3:** The proof strategy is similar to the argument for the minimax upper bound in Theorem 1, but we present the details here since they differ. Before, the inequality:

\[
\inf \sup P_\alpha \left[ \left\| \hat{\pi} - \pi \right\|_1 \leq \sup P_\alpha \left[ \left\| \hat{\pi} - \pi \right\|_1 \right] \right]
\]

holds trivially. To prove an upper bound on the extremal Bayes risk on the right hand side of this inequality, we define the event in Lemma 6 as:

\[
A \triangleq \left\{ \left\| \hat{\pi} - \pi \right\|_2 \leq \frac{c_{16}}{\sqrt{n pk}} \right\}.
\]

Then, after taking expectations in Lemma 6 with respect to the law of \( \alpha_1, \ldots, \alpha_n \), we get that for any PDF \( P_\alpha \in \mathcal{P} \) and for all sufficiently large \( n \in \mathbb{N} \):

\[
P_{P_\alpha}(A) \geq 1 - \frac{c_{17}}{n^5}.
\]

Hence, for every PDF \( P_\alpha \in \mathcal{P} \) and for all sufficiently large \( n \), we have:

\[
E_{P_\alpha}[\| \hat{\pi} - \pi \|_1] = E_{P_\alpha}[\| \hat{\pi} - \pi \|_1 | A] P_{P_\alpha}(A)
\]

\[
+ E_{P_\alpha}[\| \hat{\pi} - \pi \|_1 | A^c] P_{P_\alpha}(A^c)
\]

\[
\leq \frac{c_{16}}{\sqrt{n pk}} + \frac{2c_{17}}{n^5}
\]

where the first inequality uses the law of total expectation, the second inequality follows from (51) and the facts that \( \| \hat{\pi} - \pi \|_1 \leq \| \hat{\pi} \|_1 + \| \pi \|_1 = 2 \) (via the triangle inequality), and conditioned on \( A \), we get:

\[
\| \hat{\pi} - \pi \|_1 \leq \sqrt{n} \| \hat{\pi} - \pi \|_2 \leq \frac{c_{16}}{\sqrt{npk}} \| \pi \|_2 \leq \frac{c_{16}}{\delta \sqrt{npk}}
\]

which, in turn, uses the equivalence of \( \ell^1 \) and \( \ell^2 \)-norms (via the Cauchy-Schwarz-Bunyakovsky inequality) and the simple bound \( \| \pi \|_2 \leq 1/(\delta \sqrt{n}) \) (due to (15) and \( \alpha_1, \ldots, \alpha_n \in [\delta, 1] \), and the third inequality (52) holds for all sufficiently large \( n \) because \( k = \Theta(1) \). Letting \( c_2 = \frac{2 c_{16}}{\delta \sqrt{npk}} \) and substituting it into (52), and then taking the supremum in (52) over all PDFs \( P_\alpha \in \mathcal{P} \) yields the desired upper bound in the proposition statement. \( \blacksquare \)

**V. MSE UPPER BOUND ON SKILL DENSITY ESTIMATION**

In this section, we will prove Proposition 4 and Theorems 2 and 3 in Sections V-A, V-B, and V-C, respectively. We note that, unless stated otherwise, \( \mathbb{E}[\cdot] \) denotes the joint probability law of \( \alpha_1, \ldots, \alpha_n, \mathcal{G}(n, p) \), and \( Z_{m(i,j)} : \{i,j\} \in \mathcal{G}(n, p), m \in [k] \) with general \( P_\alpha \) throughout this section, and \( \mathbb{E}[\cdot] \) denotes the corresponding expectation operator.
A. Proof of Proposition 4

We first establish Proposition 4 using Lemma 5.

Proof of Proposition 4: We commence this proof with the following bound on the MSE between \( \hat{\mathcal{P}}^* \) and \( P_\alpha \):

\[
E \left[ \int_{R} \left( \hat{\mathcal{P}}^*(x) - P_\alpha(x) \right)^2 \, dx \right] \\
\leq 2 E \left[ \int_{R} \left( \hat{\mathcal{P}}^*_\alpha(x) - \hat{\mathcal{P}}^*_\alpha(x) + \hat{\mathcal{P}}^*_\alpha(x) - P_\alpha(x) \right)^2 \, dx \right] \\
+ 2 E \left[ \int_{R} \left( \hat{\mathcal{P}}^*(x) - \hat{\mathcal{P}}^*_\alpha(x) \right)^2 \, dx \right] \\
= 2 E \left[ \int_{R} \left( \hat{\mathcal{P}}^*_\alpha(x) - P_\alpha(x) \right)^2 \, dx \right] \\
+ 2 E \left[ \int_{R} \left( \hat{\mathcal{P}}^*(x) - \hat{\mathcal{P}}^*_\alpha(x) \right)^2 \, dx \right] \\
\leq \frac{L^2}{h^2} E \left[ \max_{i \in [n]} \left| \hat{\alpha}_i - \alpha_i \right|^2 \right] \\
\leq \frac{L^2}{h^2} E \left[ \sum_{i \in S_x} \left| \hat{\alpha}_i - \alpha_i \right|^2 \right] \\
(54)
\]

where the first inequality holds because the map \( R \ni t \mapsto K((t-x)/h) \) has support contained inside the interval \([x-h, x+h] \), the second inequality follows from the triangle inequality, and the third inequality follows from the fact that the kernel \( K \) is \( L_2 \)-Lipschitz continuous. To further upper bound (54), we now present three claims.

The first claim is a rudimentary auxiliary result that is frequently used in the high-dimensional and non-parametric statistics and theoretical machine learning literature. It says that the intersection of high probability events is itself a high probability event.

Claim 3 (Intersection of High Probability Events): Consider any two events \( A_1 \) and \( A_2 \) with probabilities satisfying \( P(A_1) \geq 1 - \varepsilon_1 \) and \( P(A_2) \geq 1 - \varepsilon_2 \) for any constants \( \varepsilon_1, \varepsilon_2 \in [0, 1] \). Then, we have:

\[
P(A_1 \cap A_2) \geq 1 - \varepsilon_1 - \varepsilon_2.
\]

Proof: This is an immediate corollary of the inclusion-exclusion principle.

The second claim utilizes Lemma 5 to show that \( |\hat{\alpha}_i - \alpha_i| = O \left( \max\{1/(\delta \sqrt{pk}), 1\} \sqrt{\log(n)/n} \right) \) for every \( i \in [n] \) with high probability for all sufficiently large \( n \).

Claim 4 (\( E^n \)-Norm Bound on Skill Parameter Estimation): There exists a (universal) constant \( c_{18} > 0 \) such that for all sufficiently large \( n \in \mathbb{N} \):

\[
P \left( \max_{i \in [n]} |\hat{\alpha}_i - \alpha_i| \leq c_{18} \max \{1/(\delta \sqrt{pk}), 1\} \sqrt{\log(n)/n} \right) \geq 1 - \frac{c_{13} + 1}{n^\beta}
\]

where \( c_{13} > 0 \) is the fixed constant from Lemma 5.

Proof: We prove this claim in four steps. Firstly, we establish that for all sufficiently large \( n \in \mathbb{N} \):

\[
P \left( \max_{i \in [n]} \alpha_i \geq 1 - \frac{5 \log(n)}{bn} \right) \geq 1 - \frac{1}{n^\beta}.
\]  

To establish (55), note that:

\[
P \left( \forall i \in [n], \alpha_i < 1 - \frac{5 \log(n)}{bn} \right) = P \left( \alpha_1 < 1 - \frac{5 \log(n)}{bn} \right)^n = \left( 1 - \left( 1 - \frac{5 \log(n)}{n} \right) \right)^n \leq \left( 1 - \frac{5 \log(n)}{n} \right)^n = \exp \left( n \log \left( 1 - \frac{5 \log(n)}{n} \right) \right) \leq \frac{1}{n^\beta}
\]
where the first equality holds because $\alpha_1, \ldots, \alpha_n$ are i.i.d.,
the third inequality holds because we have assumed that $\epsilon \geq 5 \log(n)/(bn)$ and $P_\alpha$ satisfies the lower bound $P_\alpha(t) \geq b$
for all $t \in [1 - \epsilon, 1]$, and for every large enough $n$, the fifth
inequality follows from the well-known bound $\log(1+x) \leq x$
for all $x > -1$. This produces the desired bound (55).

Secondly, we define the normalized skill parameter random variables:

$$\forall i \in [n], \quad \hat{\alpha}_i \triangleq \frac{\alpha_i}{\max_j 1/|\pi_j|} = \frac{\pi(i)}{\|\pi\|_\infty}$$

where $\pi$ denotes the probability vector of canonically scaled skill parameters in (15). It turns out that for all sufficiently
large $n \in \mathbb{N}$:

$$\mathbb{P} \left( \forall i \in [n], \ 0 \leq \hat{\alpha}_i - \alpha_i \leq \frac{10}{c_6} \sqrt{\frac{\log(n)}{n}} \right) \geq 1 - \frac{1}{n^5} \quad (56)$$
i.e., the normalized skill parameters are close to the true skill parameters with high probability. To derive (56), suppose that
the event in (55) occurs. Then, we have:

$$1 - \frac{5 \log(n)}{bn} \leq \max_{i \in [n]} \alpha_i \leq 1$$

which implies that for every $i \in [n]$:

$$\alpha_i \leq \hat{\alpha}_i \leq \alpha_i \left( 1 - \frac{5 \log(n)}{bn} \right) \leq \alpha_i \left( 1 - \frac{5}{c_6} \sqrt{\frac{\log(n)}{n}} \right) \leq \alpha_i \left( 1 + \frac{10}{c_6} \sqrt{\frac{\log(n)}{n}} \right)$$

where the third inequality follows from the lower bound we have assumed on $b$ in the proposition statement, and the fourth
inequality holds for all sufficiently large $n \in \mathbb{N}$, because the bound $(1-x)^{-1} = 1+x+x^2(1-x)^{-1} \leq 1+x+x^2 \leq 1+2x$ holds for any $x \in (0, \frac{1}{2}]$. Thus, since $\alpha_i \leq 1$ for all $i \in [n]$, we get (56).

Thirdly, we approximate the normalized skill parameters $\hat{\alpha}_i$ using the estimates $\hat{\alpha}_i$. Recall from Lemma 5 that for all sufficiently
large $n \in \mathbb{N}$:

$$\mathbb{P} \left( \frac{\|\hat{\pi}_* - \pi\|_\infty}{\|\pi\|_\infty} \leq \frac{c_{12}}{\delta} \sqrt{\frac{\log(n)}{npk}} \right) \geq 1 - \frac{c_{13}}{n^5} \quad (57)$$

where we take expectations with respect to the law of $\alpha_1, \ldots, \alpha_n$. Suppose that the event in (57) happens. Then, for any
$i \in [n]$, notice that for all sufficiently large $n \in \mathbb{N}$:

$$\hat{\alpha}_i = \frac{\pi(i)}{\|\pi\|_\infty} \leq \frac{\pi(i) + (\hat{\pi}_*(i) - \pi(i))}{\|\hat{\pi}_* - \pi\|_\infty} \leq \frac{\pi(i) + \|\hat{\pi}_* - \pi\|_\infty}{\|\pi\|_\infty - \|\pi\|_\infty} \leq \frac{\pi(i) + \|\pi\|_\infty}{\|\pi\|_\infty - \|\pi\|_\infty}$$

where the first equality follows from (8), the fourth inequality follows from the (reverse) Minkowski inequality, the sixth
inequality holds for all sufficiently large $n \in \mathbb{N}$ because: 1) the event in (57) occurs, 2) we have assumed that $\lim_{n \to \infty} \delta^{-1}(npk)^{-1/2} \log(n)^{1/2} = 0$ in the proposition statement, and 3) we use the bound $(1-x)^{-1} \leq 1+2x$ for $x \in (0, \frac{1}{2}]$, and the seventh inequality holds because $\hat{\alpha}_i \leq 1$ for all $i \in [n]$. Likewise, using analogous reasoning, observe that for all sufficiently large $n \in \mathbb{N}$:

$$\hat{\alpha}_i \geq \left( \frac{\pi(i) - \|\pi\|_\infty}{\|\pi\|_\infty - \|\pi\|_\infty} \right) \geq \frac{\pi(i) - \|\pi\|_\infty}{\|\pi\|_\infty - \|\pi\|_\infty}$$

Therefore, using Lemma 5, we obtain that for all sufficiently
large $n \in \mathbb{N}$:

$$\mathbb{P} \left( \max_{i \in [n]} |\hat{\alpha}_i - \alpha_i| \leq 4 \frac{\|\hat{\pi}_* - \pi\|_\infty}{\|\pi\|_\infty} \right) \geq 1 - \frac{c_{13}}{n^5} \quad (58)$$

Finally, we combine (56) and (58) together. Suppose that the events in both (56) and (57) occur. Then, we get:

$$\max_{i \in [n]} |\hat{\alpha}_i - \alpha_i| \leq \max_{i \in [n]} |\hat{\alpha}_i - \alpha_i| + \max_{i \in [n]} |\hat{\alpha}_i - \alpha_i| \leq 4 \frac{\|\hat{\pi}_* - \pi\|_\infty}{\|\pi\|_\infty} + \frac{10}{c_6} \sqrt{\frac{\log(n)}{n}} \leq 4c_{12} \delta \frac{1}{\sqrt{npk}} + \frac{10}{c_6} \sqrt{\frac{\log(n)}{n}}$$

Finally, our third claim uses Claim 4 to argue that with high probability, the cardinality of $S_\alpha$ is not too large.

**Claim 5 (Cardinality Bound for $S_\alpha$):** There exists a sufficiently large (universal) constant $c_{19} > 0$ such that for every
sufficiently large \( n \in \mathbb{N} \), we have:
\[
P(|S_n| \leq c_{19} Bhn) \geq 1 - \frac{c_{13} + 2}{n^5}.
\]

**Proof:** First, for any constant \( \tau > 1 \) (to be chosen later), we define \( N_h \in [n] \cup \{0\} \) to be the discrete random variable representing the number of players \( i \in [n] \) for whom \( \alpha_i \) belongs to the interval \([x - \tau h, x + \tau h]\):
\[
N_h \triangleq \sum_{i=1}^{n} \mathbb{I}\{\alpha_i \in [x - \tau h, x + \tau h]\}.
\]
Then, fix any \( \varepsilon \geq 0 \), and observe using Lemma 9 in Appendix B that:
\[
P\left(\frac{1}{n} N_h - P(\alpha_i \in [x - \tau h, x + \tau h]) > \varepsilon\right) \leq \exp\left(-2n\varepsilon^2\right)
\]
which uses the fact that \( \{\alpha_i \in [x - \tau h, x + \tau h]\} : i \in [n] \) are i.i.d. Bernoulli random variables with mean \( P(\alpha_i \in [x - \tau h, x + \tau h]) \), since \( \alpha_1, \ldots, \alpha_n \) are drawn i.i.d. from \( P_a \). At this point, letting \( \varepsilon = \sqrt{5 \log(n) / (2n)} \) yields:
\[
P\left(\frac{1}{n} N_h - P(\alpha_i \in [x - \tau h, x + \tau h]) > \sqrt{5 \log(n) / 2n}\right) \leq \exp\left(-\frac{5n \log(n)}{n}\right) = \frac{1}{n^5}.
\]
(59)

Next, recall that \( P_a \in \mathcal{P} \) is uniformly bounded (almost everywhere) by \( B > 0 \), i.e., \( P_a(t) \leq B \) for all \( t \in \mathbb{R} \). Using this bound, we obtain:
\[
P(\alpha_i \in [x - \tau h, x + \tau h]) = \int_{x - \tau h}^{x + \tau h} P_a(t) dt \leq 2B\tau h.
\]
(60)
Hence, combining (59) and (60), we have that with probability at least \( 1 - n^{-5} \):
\[
\frac{1}{n} N_h \leq P(\alpha_i \in [x - \tau h, x + \tau h]) + \sqrt{\frac{5 \log(n)}{2n}} \leq 2B\tau h + \sqrt{\frac{5 \log(n)}{2n}}.
\]
Equivalently, we have derived the following bound:
\[
P\left(N_h \leq 2B\tau h + \sqrt{\frac{5 \log(n)}{2n}}\right) \geq 1 - \frac{1}{n^5}.
\]
(61)
Now, recall from the proposition statement that \( h = \Omega(\max\{1/((\delta/\sqrt{pk}), 1)\sqrt{\log(n)/n}\}) \). Hence, we may choose \( \tau > 1 \) large enough so that for all sufficiently large \( n \), we have:
\[
(\tau - 1)h \geq c_{18} \max\left\{\frac{1}{\delta/\sqrt{pk}}, 1\right\}\sqrt{\log(n) / n}.
\]
(62)
where \( c_{18} > 0 \) is the constant from Claim 4. Assume that both the events in (61) and Claim 4 occur. Then, for any \( i \in [n] \), if \( \alpha_i \in [x - h, x + h] \), then we trivially have \( \alpha_i \in [x - \tau h, x + \tau h] \) since \( \tau > 1 \). On the other hand, if \( \alpha_i \in [x - h, x + h] \), then we have \( \alpha_i \in [x - \tau h, x + \tau h] \), because \( |\alpha_i - \alpha_1| \leq c_{18} \max\{1/((\delta/\sqrt{pk}), 1)\sqrt{\log(n)/n}\} \leq (\tau - 1)h \) by Claim 4 and (62). Thus, we get:
\[
|S_n| \leq N_h
\]
with \( \tau \) chosen according to (62). Applying Claim 3, Claim 4, and (61) together yields:
\[
P\left(|S_n| \leq 2B\tau h + \sqrt{\frac{5 \log(n)}{2n}}\right) \geq 1 - \frac{c_{13} + 2}{n^5}
\]
for every sufficiently large \( n \). Lastly, let \( c_{19} > 2\tau \) be a sufficiently large constant so that for all sufficiently large \( n \):
\[
(c_{19} - 2\tau)Bh \geq \sqrt{\frac{5 \log(n)}{2n}}
\]
which is consistent with our assumption that \( h = \Omega(\max\{1/((\delta/\sqrt{pk}, 1)\sqrt{\log(n)/n}\}) \). Therefore, we may write:
\[
P(|S_n| \leq c_{19} Bhn) \geq 1 - \frac{c_{13} + 2}{n^5}
\]
for every sufficiently large \( n \), which completes the proof. ■

Having developed Claim 5, we are now in a position to upper bound (54). For any sufficiently large \( n \), define the event in Claim 5 as:
\[
A_n \triangleq \{|S_n| \leq c_{19} Bhn\}.
\]
Then, observe that:
\[
E\left[|S_n|^2 \max_{i \in [n]} |\hat{\alpha}_i - \alpha_i|^2\right]
\]
\[
= E\left[|S_n|^2 \max_{i \in [n]} |\hat{\alpha}_i - \alpha_i|^2 \mathbb{I}_{A_n}\right] P(A_n)
\]
\[
+ E\left[|S_n|^2 \max_{i \in [n]} |\hat{\alpha}_i - \alpha_i|^2 \mathbb{I}_{A_n^c}\right] P(A_n^c)
\]
\[
\leq c_{19} B^2 h^2 n^2 E\left[\max_{i \in [n]} |\hat{\alpha}_i - \alpha_i|^2 \mathbb{I}_{A_n}\right] P(A_n) + \frac{c_{13} + 2}{n^3}
\]
\[
\leq c_{19} B^2 h^2 n^2 \left( E\left[\max_{i \in [n]} |\hat{\alpha}_i - \alpha_i|^2 \mathbb{I}_{A_n}\right] P(A_n)\right)
\]
\[
+ E\left[\max_{i \in [n]} |\hat{\alpha}_i - \alpha_i|^2 \mathbb{I}_{A_n^c}\right] P(A_n^c)\right) + \frac{c_{13} + 2}{n^3}
\]
\[
= c_{19} B^2 h^2 n^2 \left( E\left[\max_{i \in [n]} |\hat{\alpha}_i - \alpha_i|^2 \right] P(A_n)\right)
\]
\[
+ \frac{c_{13} + 2}{n^3}
\]
(63)
where the first and fourth equalities follow from the tower property, and the second inequality uses Claim 5 and the facts that \( |S_n| \leq n \) and \( |\hat{\alpha}_i - \alpha_i| \leq 1 \) for all \( i \in [n] \) (see (8)). Plugging (63) into (54) produces:
\[
E\left[\sum_{i=1}^{n} K\left(\frac{\hat{\alpha}_i - x}{h}\right) - K\left(\frac{\alpha_i - x}{h}\right)\right]^2\right]
\]
\[
\leq \frac{L^2}{h^2} \left( c_{19} B^2 h^2 n^2 E\left[\max_{i \in [n]} |\hat{\alpha}_i - \alpha_i|^2 \right] + \frac{c_{13} + 2}{n^3}\right).
\]
We can then substitute this bound into (53) to obtain:
\[
E\left(\int_{\mathbb{R}} (\hat{P}_n(x) - P_n(x))^2 dx\right)
\]
\[
\leq 2E\left(\int_{\mathbb{R}} (\hat{P}_n(x) - P_n(x))^2 dx\right)
\]
\[
+ \frac{2L^2}{n^2 h^4} \int_{-1}^{2} c_{19} B^2 h^2 n^2 \left( E\left[\max_{i \in [n]} |\hat{\alpha}_i - \alpha_i|^2 \right] + \frac{c_{13} + 2}{n^3}\right) dx
\]
\[
\begin{align*}
&= 2 \mathbb{E} \left[ \int_{\mathbb{R}} (\hat{P}_{\alpha_n}(x) - P_\alpha(x))^2 \, dx \right] \\
&+ 6L_2^2 \left( \frac{c_1^2 B^2 h^2 n^2}{n^2 h^4} \mathbb{E} \left[ \max_{i \in [n]} |\hat{\alpha}_i - \alpha_i|^2 \right] + \frac{c_{13} + 2}{n^3} \right) \\
&= 2 \mathbb{E} \left[ \int_{\mathbb{R}} (\hat{P}_{\alpha_n}(x) - P_\alpha(x))^2 \, dx \right] \\
&+ \frac{6c_1^2 B^2 L_2^2}{h^2} \mathbb{E} \left[ \max_{i \in [n]} |\hat{\alpha}_i - \alpha_i|^2 \right] + \frac{6(c_{13} + 2)L_2^2}{n^3 h^4}
\end{align*}
\]
for all sufficiently large \( n \). This completes the proof after letting \( c_9 = 6c_1^2 \) and \( c_{10} = 6(c_{13} + 2) \).

\[\text{Lemma 7 (Bias-Variance Tradeoff [3], [27]): For any } P_\alpha \in \mathcal{P}, \text{ any kernel } K : [-1,1] \rightarrow \mathbb{R}, \text{ and any bandwidth } h \in (0,1), \text{ we have:}
\]
\[
\mathbb{E} \left[ \int_{\mathbb{R}} (\hat{P}_{\alpha_n}(x) - P_\alpha(x))^2 \, dx \right] \\
\leq \frac{1}{nh} \int_{-1}^{1} K(x)^2 \, dx + 3h^{2\eta} \left( \frac{L_1}{8!} \int_{-1}^{1} |x|^{\eta} |K(x)| \, dx \right)^2
\]
where \( \alpha_1, \ldots, \alpha_n \) are i.i.d. with distribution \( P_\alpha \).

In Lemma 7, it is well-known that the first term captures the variance of \( \hat{P}_{\alpha_n} \), and the second term bounds the squared bias of \( \hat{P}_{\alpha_n} \). Specifically, as shown in [3] and [27], the bound on the variance term uses the property that the kernel is square-integrable, and the bound on the bias term uses the other properties in the definition of a kernel as well as the Hölder class assumption on \( P_\alpha \) (outlined earlier). Furthermore, we remark that the bound on the bias term in Lemma 7 follows from [3, Proposition 1.2] by noting that \( P_\alpha \) has its support in the interval \([-1,1]\] and \( \hat{P}_{\alpha_n} \) has its support in the interval \([-1,2]\] (because the kernel \( K \) has its support in \([-1,1]\) and the bandwidth \( h \leq 1 \)). In fact, the length of the interval \([-1,1]\) and \([-1,2]\) is what gives rise to the constant 3 in Lemma 7.

We next prove Theorem 2 using Lemmata 5 and 7 and Proposition 4.

\[\text{Proof of Theorem 2: We begin by recalling the result of Proposition 4. There exist sufficiently large constants } c_9, c_{10} > 0 \text{ such that for all sufficiently large } n \in \mathbb{N}: \]
\[
\mathbb{E} \left[ \int_{\mathbb{R}} (\hat{P}^*(x) - P_\alpha(x))^2 \, dx \right] \\
\leq 2 \mathbb{E} \left[ \int_{\mathbb{R}} (\hat{P}_{\alpha_n}(x) - P_\alpha(x))^2 \, dx \right] \\
+ \frac{c_9 B^2 L_2^2}{h^2} \mathbb{E} \left[ \max_{i \in [n]} |\hat{\alpha}_i - \alpha_i|^2 \right] + \frac{c_{10} L_2^2}{n^3 h^4}
\]
Since we must have \( \lim_{n \to \infty} \frac{c_{20}}{nh} = c_2 h^{2n} \), which implies that the right hand side of (67) is \( \Theta(n^{-(2n-1)/(2n+1)}) \). On the other hand, if we balance the second and third terms, we get:

\[
\frac{c_{21} h^{2n}}{nh} \Rightarrow h = \Theta\left(\frac{1}{\delta \pi^2 (pk)^{n+1}}\right)
\]

where \( c_7 > 0 \) is a sufficiently large constant that depends on \( \gamma, \eta, B, L_1, L_2, \) and the kernel \( K \). This completes the proof.

\section*{C. Proof of Theorem 3}

To prove Theorem 3, we first need to establish a version of the main result in [15], which demonstrates the Lipschitz continuity of differential entropy with respect to the 2-Wasserstein distance (and was developed in the context of network information theory). In particular, our result will demonstrate the Lipschitz continuity of differential entropy with respect to the \( \ell^2 \)-distance. To this end, similar to the statement of Theorem 3, for any non-empty compact interval \( I \subseteq \mathbb{R} \), let \( D = D(I, a, L_3) \) be the set of all PDFs (with respect to the Lebesgue measure) that have support \( I \), are uniformly lower bounded by a constant \( a > 0 \), and are Lipschitz continuous with Lipschitz constant \( L_3 > 0 \). (Note that \( I = [\delta, 1] \) in the statement of Theorem 3, but we proceed with the slightly more general development here.) Hence, every PDF \( f \in D \) with support \( I \) satisfies:

1) \( f \) is lower bounded, i.e., \( f(x) \geq a \) for all \( x \in I \),
2) \( f \) is \( L_3 \)-Lipschitz continuous, i.e., \( |f(x) - f(y)| \leq L_3 |x - y| \) for all \( x, y \in I \).

Since \( I \) is compact and each \( f \in D \) is continuous, and hence, bounded (by the boundedness theorem), each \( f \in D \) has finite \( \ell^2 \)-norm, and \( D \subseteq \ell^2(I) \). So, \( D \) is a subset of the Hilbert space \( \ell^2(I) \). The next lemma shows that \( D \) is closed and convex, where we define closure with respect to the \( \ell^2 \)-distance.

\begin{lemma} (Closed and Convex): \( D \) is a closed and convex set in \( \ell^2(I) \).
\end{lemma}

Lemma 8 is proved in Appendix A-C. Since we identify all functions in \( \ell^2(I) \) that differ on null sets of Lebesgue measure 0, in the context of Hilbert spaces, \( D \) is technically a set of equivalence classes of functions that differ on null sets from \( L_3 \)-Lipschitz continuous PDFs on \( \mathbb{R} \) that are lower bounded by \( a \). As shown in Appendix A-C, we can represent any such equivalence class with a single function in \( D \) that is lower bounded by \( a \) and \( L_3 \)-Lipschitz continuous everywhere. For this reason, we will construe \( D \) as the set of functions defined above rather than equivalence classes.

Next, recall that for any PDF \( f \in D \), its differential entropy is given by [61, Section 8.1]:

\[
h(f) \triangleq - \int f(x) \log(f(x)) \, dx
\]

which is always well-defined and finite, because \( I \) is compact and each \( f \) is continuous, and hence, bounded (by the boundedness theorem) [72, Section 8.3]. The ensuing proposition illustrates that the differential entropy functional is Lipschitz continuous over the PDFs in \( D \).

\begin{proposition} (Lipschitz Continuity of Differential Entropy): For every \( f, g \in D \), we have:

\[
|h(f) - h(g)| \leq \frac{L_3 |I|^{3/2}}{2a} \|f - g\|_2
\]

where \( |I| \) denotes the length of the interval \( I \).
\end{proposition}

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Proof: Fix any \( f, g \in \mathcal{D} \), and consider a random variable \( X \in I \) with PDF \( f \) and a random variable \( Y \in I \) with PDF \( g \). Let \( P_{X,Y} \) be any coupling of \( X \) and \( Y \), i.e., any joint probability distribution of \( X \) and \( Y \) such that its marginal distributions are \( P_X = f \) and \( P_Y = g \) (with abuse of notation). Then, observe that:

\[
|\log(f(X)) - \log(f(Y))| \leq \frac{1}{a} |f(X) - f(Y)|
\]

where the first inequality holds because \( \sup_{t \geq a} \frac{d}{dt} \log(t) = \frac{1}{a} \) and \( f, g \) are lower bounded by \( a \), and the second inequality holds because \( f \in \mathcal{D} \). Taking expectations of both sides with respect to the coupling \( P_{X,Y} \), we obtain:

\[
\mathbb{E}\left[ |\log\left(\frac{f(X)}{f(Y)}\right)| \right] \leq \frac{L_3}{a} \mathbb{P}(X \neq Y)
\]

where \( \mathbb{E}[\cdot] \) and \( \mathbb{P}(\cdot) \) denote the expectation and probability operators with respect to \( P_{X,Y} \) here. Then, minimizing over all couplings \( P_{X,Y} \) on both sides yields:

\[
\mathbb{E}\left[ |\log\left(\frac{f(X)}{f(Y)}\right)| \right] \leq \frac{L_3}{2a} \int_I |f(t) - g(t)| \, dt
\]

where \( \mathbb{E}[\cdot] \) now denotes the expectation with respect to the optimal coupling of \( X \) and \( Y \) on the left hand side (and it will continue to denote this from hereon), and the right hand side follows from the optimal coupling representation of total variation distance (cf. [50], [73]):

\[
\frac{1}{2} \int_I |f(t) - g(t)| \, dt = \min_{P_{X,Y}} \mathbb{P}(X \neq Y)
\]

where the minimum exists and is taken over all couplings \( P_{X,Y} \). Next, applying the Cauchy-Schwarz-Bunyakovsky inequality, we get:

\[
\mathbb{E}\left[ |\log\left(\frac{f(X)}{f(Y)}\right)| \right] \leq \frac{L_3}{2a} \| f - g \|_2^2.
\] (69)

Finally, observe that:

\[
|h(g) - h(f)| \leq h(g) - h(f) + D(g||f)
\]

\[
= - \int_I g(x) \log(g(x)) \, dx + \int_I f(x) \log(f(x)) \, dx + \int_I g(x) \log\left(\frac{g(x)}{f(x)}\right) \, dx - \int_I f(x) \log\left(\frac{f(x)}{g(x)}\right) \, dx
\]

\[
= \mathbb{E}[\log(f(X))] - \mathbb{E}[\log(f(Y))]
\]

\[
= \mathbb{E}\left[ |\log\left(\frac{f(X)}{f(Y)}\right)| \right]
\]

\[
\leq \frac{L_3}{2a} \| f - g \|_2^2
\]

where the first inequality uses the non-negativity of Kullback-Leibler divergence \( D(g||f) \) (see (20) for a general definition), and the last inequality follows from (69). Then, by symmetry, we obtain:

\[
|h(g) - h(f)| \leq \frac{L_3}{3/2} \left\| f - g \right\|_2^2\frac{3}{2a}
\]

which completes the proof.

Since Proposition 6 is a variant of a result in [15], we briefly discuss [15, Proposition 1] and its proof technique in comparison to our result. Specifically, [15, Proposition 1] states that for continuously differentiable density functions \( p, q \) with support \( \mathbb{R}^m \) that have second moments bounded by \( C_0 > 0 \) and satisfy \( \max\{\|\nabla \log(p(x))\|_2, \|\nabla \log(q(x))\|_2\} \leq C_1 \|x\|_2 + C_2 \) for some \( C_1 > 0 \) and \( C_2 \geq 0 \), we have:

\[
|h(p) - h(q)| \leq \left( C_1 \sqrt{C_0} + C_2 \right) W_2(p, q)
\] (70)

where \( W_2(p, q) \) is the 2-Wasserstein distance. Here, the regularity condition, \( \max\{\|\nabla \log(p(x))\|_2, \|\nabla \log(q(x))\|_2\} \leq C_1 \|x\|_2 + C_2 \), is imposed in [15] for the purposes of their proof. Indeed, this condition can be used to write the absolute difference \( |\log(p(X)) - \log(p(Y))| \) in terms of an integral of \( x \mapsto \nabla \log(p(x)) \), where \( X \) and \( Y \) are distributed according to \( p \) and \( q \), respectively, and then apply the Cauchy-Schwarz inequality to the integrand to obtain a bound in terms of \( \|X - Y\|_2 \); these are the first few steps in their proof. In contrast, as stated earlier, our result in Proposition 6 applies to densities with conditions, such as Lipschitz continuity, that are similar to what we require for skill distribution estimation. Our conditions are also simple enough so that we can prove properties like the closure of \( D \) in Lemma 8, which are needed to ensure that \( \mathbb{P}_{proj} \) is well-defined in Theorem 3. Furthermore, our result in Proposition 6 illustrates the Lipschitz continuity of differential entropy in \( \mathcal{L}_2 \)-distance instead of 2-Wasserstein distance, because this enables us to translate our MSE bound on skill distribution estimation in Theorem 2 (which uses \( \mathcal{L}_2 \)-distance between densities) into an MSE bound on differential entropy estimation in Theorem 3 (as shown in the next proof). Moving onto the proofs, as alluded to above, the high-level idea of the proof in [15] is to bound \( |\log(p(X)) - \log(p(Y))| \) in terms of \( \|X - Y\|_2 \) (using an integral representation and the Cauchy-Schwarz inequality), so that one can then take expectations with respect to the optimal 2-Wasserstein coupling of \( X \) and \( Y \), where \( X \) and \( Y \) are distributed according to \( p \) and \( q \), respectively. Taking these expectations produces an upper bound in terms of the 2-Wasserstein distance \( W_2(p, q) \) after some manipulation. On the other hand, in our proof of Proposition 6 above, for any PDFs \( f, g \in \mathcal{D} \) and random variables \( X \) and \( Y \) distributed according to \( f \) and \( g \), respectively, we also upper bound \( |\log(f(X)) - \log(f(Y))| \) in terms of \( \|X - Y\|_2 \), albeit using our assumptions rather than the integral representation and Cauchy-Schwarz argument of [15]. But after this, since we seek a bound in terms of the \( \mathcal{L}_2 \)-distance \( \|f - g\|_2 \), we bound \( |X - Y| \) with \( |f| \mathbb{I}\{X \neq Y\} \) using the compactness of the support \( f \). We can then take expectations with respect to the optimal total variation coupling of \( X \) and \( Y \), and after performing some further steps, obtain an upper bound in...
terms of the $L^2$-distance $\|f - g\|_2$. In this way, we loosely follow the proof strategy in [15]. Although we follow the original strategy in [15], we note that results like Proposition 6 can alternatively be succinctly derived using the Lipschitz continuity of $t \mapsto t \log(t)$ induced by the assumptions on $\mathcal{D}$.

Equipped with this key intermediate result, we next prove Theorem 3.

Proof of Theorem 3: Let the interval $I = [\delta, 1]$ in the above discussion with $|I| \leq 1$. Then, the set $\mathcal{D} = \mathcal{D}(I, a, L_3)$ above reduces to the set $\mathcal{D} = \mathcal{D}(\delta, a, L_3)$ in the theorem statement. We first argue that the estimator $h(\hat{P}_{\text{proj}})$ in the theorem statement for the true differential entropy $h(P_\alpha)$ is well-defined.

Notice that the tempered estimator $\hat{P}_{\text{temp}} \in L^2(I)$. Indeed, using the definition of $\hat{P}_{\text{temp}}$, (9), and the triangle inequality, we have:

$$\|\hat{P}_{\text{temp}}\|_2 \leq \|\hat{P}^\ast\|_2$$

$$\leq \frac{1}{n\hbar} \sum_{i=1}^{n} \sqrt{h} \|K\|_2$$

$$\leq \frac{1}{\sqrt{h}} \|K\|_2 < +\infty$$

where the bandwidth $h$ is given by (10) and $\|K\|_2$ is finite by assumption (see Section II-B). Since $\hat{P}_{\text{temp}} \in L^2(I)$ and $\mathcal{D}$ is a closed and convex subset of $L^2(I)$ by Lemma 8, employing the Hilbert projection theorem (cf. [74, Chapter 6, Theorem 2], [75, Theorem 3.14]), we see that the projected PDF estimator:

$$\hat{P}_{\text{proj}} = \arg\min_{\hat{P} \in \mathcal{D}} \|\hat{P}_{\text{temp}} - \hat{P}\|_2$$

is well-defined and unique. (Furthermore, as noted earlier, we may assume that the PDF $\hat{P}_{\text{proj}}$ is $L_3$-Lipschitz continuous and lower bounded by a everywhere on $I$, by choosing an appropriate representative in $\mathcal{D}$ as in Appendix A-C.) Therefore, the differential entropy $h(\hat{P}_{\text{proj}})$ is also well-defined and finite as explained after (68).

Next, observe using Proposition 6 that:

$$\left(h(\hat{P}_{\text{proj}}) - h(P_\alpha)\right)^2 \leq \frac{L^2}{4a^2} \|\hat{P}_{\text{proj}} - P_\alpha\|_2^2$$

$$\leq \frac{L^2}{4a^2} \|\hat{P}_{\text{temp}} - P_\alpha\|_2^2$$

$$\leq \frac{L^2}{4a^2} \|\hat{P}^\ast - P_\alpha\|_2^2$$

where the first inequality uses the facts that $\hat{P}_{\text{proj}} \in \mathcal{D}$ (by construction), $P_\alpha \in D$ (by assumption), and $|I| \leq 1$, the second inequality holds because the projection map on $\mathcal{D}$ is non-expansive (cf. [75, Theorem 3.14, Proposition 4.8], [76, Theorem 3]), and the third inequality holds by definition of $\hat{P}_{\text{temp}}$:

$$\|\hat{P}^\ast - P_\alpha\|_2^2 = \int_{\mathbb{R}} \left(\hat{P}^\ast(x) - P_\alpha(x)\right)^2 dx$$

$$\geq \int_{I} \left(\hat{P}^\ast(x) - P_\alpha(x)\right)^2 dx$$

$$= \int_{I} \left(\hat{P}_{\text{temp}}(x) - P_\alpha(x)\right)^2 dx$$

$$= \left\|\hat{P}_{\text{temp}} - P_\alpha\right\|_2^2.$$

Hence, taking expectations on both sides, we have for all $P_\alpha \in \mathcal{P} \cap \mathcal{D}$:

$$\mathbb{E} \left[\left(h(\hat{P}_{\text{proj}}) - h(P_\alpha)\right)^2\right]$$

$$\leq \frac{L^2}{4a^2} \mathbb{E} \left[\left(\hat{P}^\ast(x) - P_\alpha(x)\right)^2 dx\right]$$

$$\leq c_{11} \max\left\{\frac{1}{\delta^2 pk}, 1\right\} \left(\frac{\log(n)}{n}\right)^{\frac{n}{n+1}}$$

where the second inequality holds for all sufficiently large $n \in \mathbb{N}$ due to Theorem 2 and we let $c_{11} = L^2_{\hbar} v_\gamma/(4a^2)$. Taking the supremum over all $P_\alpha \in \mathcal{P} \cap \mathcal{D}$ on the left hand side completes the proof.

VI. EXPERIMENTS

In this section, we apply our method to several real-world datasets and a synthetic example to demonstrate its utility. Specifically, Algorithm 1 produces estimates of skill distributions. In order to compare skill distributions across different scenarios as well as capture their essence, it is desirable to compute a single score that holistically measures the variation of levels of skill in a tournament (to be explained). We propose such a score and discuss it in Sections VI-A and VI-B. Our algorithmic choices in Section VI-C, portray our numerical experiments on Cricket World Cups, Soccer World Cups, English Premier League soccer, and US mutual funds in Sections VI-D, VI-E, VI-F, and VI-G, respectively, and present our synthetic experiment in Section VI-H.

A. Overall Skill Score

Intuitively, a Dirac delta measure (i.e., all skill levels are equal) represents a setting where all game outcomes are completely random; there is no role of skill. On the other hand, the uniform PDF $\text{unif}([0, 1])$ (assuming $\delta$ is very small) typifies a setting where players are endowed with a broad variety of skill levels. We refer readers to [23] for a related discussion. Propelled by this intuition, any “distance” between $P_\alpha$ and $\text{unif}([0, 1])$ that is maximized when $P_\alpha$ is a Dirac delta measure serves as a reasonable score, which is larger when luck plays a greater role in determining the outcomes of games. Therefore, we propose to use the negative differential entropy of $P_\alpha$ as an overall score to measure skill in a tournament [61], [62]:

$$-h(P_\alpha) \triangleq \int_{\mathbb{R}} P_\alpha(t) \log(P_\alpha(t)) dt \in [0, +\infty].$$

This is a well-defined quantity (for all PDFs of interest $P_\alpha$) that is equal to the KL divergence between $P_\alpha$ and $\text{unif}([0, 1])$: $-h(P_\alpha) = D(P_\alpha || \text{unif}([0, 1]))$ (cf. (20) in Section IV-A). Moreover, $-h(P_\alpha) = 0$ when $P_\alpha = \text{unif}([0, 1])$, and $-h(P_\alpha) = +\infty$ when $P_\alpha$ is a Dirac delta measure. In the sequel, we will use the phrase “variation of skill levels” to mean variation in the sense of entropy of the skill distribution (rather than other notions of variation like the variance; we
compare entropy with variance in Section VI-B). For simplicity, to estimate \(-h(P_\alpha)\) from data, we will use a version of the resubstitution estimator based on \(\hat{P}^*\) and \(\hat{\alpha}_1, \ldots, \hat{\alpha}_n\) [19], [59]:

\[
-\hat{H}_Z \triangleq \frac{1}{n} \sum_{i=1}^{n} \log \left( \hat{P}^*(\hat{\alpha}_i) \right)
\]

where \(\hat{P}^*\) and \(\hat{\alpha}_1, \ldots, \hat{\alpha}_n\) are the estimators for \(P_\alpha\) and \(\alpha_1, \ldots, \alpha_n\) based on the data \(Z\) in (9) and (8), respectively. Note that using other entropy estimation methods, such those akin to the approaches delineated in [19] or Theorem 3, would demonstrate the same trends in the sequel.

### B. Discussion of Overall Skill Score

We further discuss some properties and considerations of our choice of “overall skill score” in this subsection. As noted above, we will use the negative differential entropy of a pdf as the overall skill score in our experiments. Neglecting \(\delta > 0\) since it is usually very small, we consider skill distributions on the interval \([0, 1]\). As also noted above, the negative differential entropy of a pdf \(P\) on \([0, 1]\) is given by the KL divergence:

\[
-h(P) = D(P||\text{unif}([0, 1]))
\]

Hence, the extreme values of negative entropy can be determined using properties of KL divergence.

Indeed, the negative entropy is minimized with value \(-h(P) = 0\) (where KL divergence is 0) if and only if \(P = \text{unif}([0, 1])\) is the uniform pdf on \([0, 1]\) (cf. [62]). On the other hand, the negative entropy is maximized with value \(-h(P) = +\infty\) under a variety of scenarios. For example, since the KL divergence \(D(P||\text{unif}([0, 1]))\) is known to be \(+\infty\) whenever \(P\) is not absolutely continuous with respect to \(\text{unif}([0, 1])\) as probability measures (cf. (20) in Section IV-A), \(-h(P) = +\infty\) whenever \(P\) is actually a probability mass function (i.e., a discrete distribution), e.g., \(P\) is a single Dirac delta measure corresponding to a constant random variable. But negative entropy can also be \(+\infty\) for non-degenerate pdfs. For example, consider a collection of disjoint intervals \(\{I_m : m = 2, 3, 4, \ldots\}\) in \([0, 1]\) with lengths \(|I_m| = (m \log(m))^{-2}\) (where \(\sum_{m=2}^{\infty} |I_m| = \sum_{m=2}^{\infty} (m \log(m))^{-2} < 1\)), and a pdf \(P\) such that:

\[
P(t) = \begin{cases} \frac{1}{C}, & t \in I_m \text{ for } m = 2, 3, 4, \ldots \\ 0, & \text{otherwise} \end{cases}
\]

where \(C = \sum_{m=2}^{\infty} m^{-1} (\log(m))^{-2}\) [72, Section 8.3]. Then, a simple calculation in [72, Section 8.3] shows that \(-h(P) = +\infty\). As explained in [62], these two cases of infinite KL divergence are consistent, because in both cases, there exists a sequence of finite partitions of \([0, 1]\) such that the KL divergence between the probability mass functions induced by these partitions from \(P\) and \(\text{unif}([0, 1])\) diverges. We also note that if a valid pdf \(P\) on \([0, 1]\) is bounded, then \(-h(P)\) must be finite [72, Section 8.3] (and the example in (74) is indeed unbounded). This discussion conveys that while the minimum value of our overall skill score scale is uniquely achieved by the uniform pdf, the maximum value of the scale can be achieved by many distributions, e.g., discrete distributions, certain unbounded pdfs, etc. So, one must interpret large values of overall skill score with care, ensuring that the skill pdf is indeed approaching a Dirac delta measure (where all skill levels are equal) rather than other possibilities.

Intuitively, we often interpret differential entropy as a measure of “variation” of a pdf. For nice “Gaussian-like” densities, this notion of variation is similar to that provided by the variance, because the differential entropy of a Gaussian pdf with variance \(\sigma^2 > 0\) is \(\frac{1}{2} \log(2\pi e \sigma^2)\). However, the notions of variance and differential entropy are of course different in general. In particular, the extreme values of differential entropy and variance are not achieved by the same classes of distributions. Thus, in our experiments, we will use “variation of skill levels” to mean variation in the entropic sense (as mentioned earlier). In the sequel, we reiterate the differences between differential entropy and variance using some examples.

Firstly, we portray that the extreme values of variance are different to those of differential entropy. As before, let us consider any bound random variable \(X \in [0, 1]\) (almost surely). On the one hand, the variance \(\text{VAR}(X)\) is minimized with value \(\text{VAR}(X) = 0\) if and only if \(X\) is a constant almost surely, i.e., the probability distribution of \(X\) is a single Dirac delta measure. On the other hand, the variance \(\text{VAR}(X)\) is maximized at the value \(\text{VAR}(X) = \frac{1}{4}\) if and only if \(X\) is a uniform Bernoulli random variable with \(F(X = 1) = F(X = 0) = \frac{1}{2}\). This latter fact requires an argument. To this end, recall that Popoviciu’s inequality states that for any random variable \(X \in [0, 1]\), its variance is bounded by [77]:

\[
\text{VAR}(X) \leq \frac{1}{4}.
\]

It turns out that this inequality is met with equality if and only if \(X\) is a uniform Bernoulli random variable. Since this equality case is seldom proved when considering general random variables, we provide a proof for it here. Clearly, a uniform Bernoulli random variable has variance \(\frac{1}{4}\), so it suffices to prove that \(X \in [0, 1]\) and \(\text{VAR}(X) = \frac{1}{4}\) implies that \(X\) is a uniform Bernoulli random variable. To show this, recall that the Bhattacharyya inequality states that for any random variable \(X \in [0, 1]\), its variance is bounded by [78]:

\[
\text{VAR}(X) \leq \text{E}[X](1 - \text{E}[X]) \leq \frac{1}{4}.
\]

Hence, the assumption \(\text{VAR}(X) = \frac{1}{4}\) implies that \(\text{E}[X] = \frac{1}{2}\), which in turn implies that \(\text{E}[X^2] = \text{VAR}(X) + \text{E}[X]^2 = \frac{1}{4}\). So, we get \(\text{E}[X(1 - X)] = \text{E}[X] - \text{E}[X^2] = 0\) for the almost surely non-negative random variable \(X(1 - X) \geq 0\) (since \(X \in [0, 1]\)). This means that \(X(1 - X) = 0\) almost surely, which yields that \(X \in \{0, 1\}\) almost surely, i.e., \(X\) is a Bernoulli random variable. The variance of a Bernoulli random variable is \(\frac{1}{4}\), if and only if it is uniform. Thus, \(X\) is a uniform Bernoulli random variable; this establishes the equality case of (75).

Secondly, we delineate two examples which demonstrate that variance can vary arbitrarily for fixed differential entropy, and differential entropy can vary arbitrarily for fixed variance.
respectively. In the first example, consider a random variable \( X \in [0,1] \) distributed according to the PDF \( P_{\Delta,\tau} \) with \( 0 < \Delta < \tau \leq \frac{1}{2} \) such that:

\[
P_{\Delta,\tau}(t) = \begin{cases} \frac{1}{2\Delta}, & t \in [\tau - \Delta, \tau) \cup [1 - \tau, 1 - \tau + \Delta], \\ 0, & \text{otherwise}. \end{cases}
\] (77)

This PDF has differential entropy \( h(P_{\Delta,\tau}) = \log(2\Delta) \) and variance \( \text{VAR}(X) = \tau(\tau - \Delta - 1) + \frac{2\Delta^2 + 3\Delta}{6} + \frac{1}{4} \) by direct calculation. Then, for any very small fixed \( \Delta \), setting \( \tau = \frac{1}{2} \) produces a minimal variance of \( \text{VAR}(X) = \frac{\Delta^2}{4} = O(\Delta^2) \), and as \( \tau \rightarrow \Delta \), we obtain a maximal variance of \( \text{VAR}(X) = \frac{1}{4} - \frac{\Delta^2}{4} + \frac{\Delta^2}{4} = \frac{1}{4} - \Theta(\Delta) \). Therefore, for very small fixed \( \Delta \), the PDF \( P_{\Delta,\tau} \) has constant differential entropy but can achieve almost the entire range of variance values by varying \( \tau \).

In the second example, consider a random variable \( X \in [0,1] \) distributed according to the PDF \( P_{\Delta,k} \) with \( k \in \mathbb{N} \setminus \{1,2\} \) and \( 0 < \Delta < \frac{1}{k} \) such that:

\[
P_{\Delta,k}(t) = \begin{cases} \frac{1}{2\Delta}, & t \in \bigcup_{m=0}^{k-1} \left[ \frac{m}{k}, \frac{m}{k} + \Delta \right], \\ 0, & \text{otherwise}. \end{cases}
\] (78)

This PDF has differential entropy \( h(P_{\Delta,k}) = \log(k\Delta) \) and variance \( \text{VAR}(X) = \frac{1}{k} - \frac{3}{k^2} \left( \frac{1}{k^2} - \Delta^2 \right) \) by direct calculation. Then, for any very large fixed \( k \), \( \text{VAR}(X) \approx \frac{1}{k^2} \) (or more precisely, \( \frac{1}{k^2} - \Delta^2 \) \( \leq \) \( \text{VAR}(X) \leq \frac{1}{k^2} \)) regardless of the value of \( \Delta \). However, as \( \Delta \rightarrow \frac{1}{k} \), we obtain a maximal differential entropy value of 0, and as \( \Delta \rightarrow 0 \), we obtain a minimal differential entropy value of \(-\infty\). Therefore, for very large fixed \( k \), the PDF \( P_{\Delta,k} \) has roughly constant variance but can achieve the entire range of differential entropy values by varying \( \Delta \).

In light of the above discussion, we use negative differential entropy as an overall skill score in our experiments with the understanding that it measures variation of skill levels in a specific sense, and not necessarily in the sense of variance. We leave the investigation of various other notions of dispersion of skill distributions, e.g., variance, interquartile range, etc., as future work.

C. Algorithmic Choices and Data Preprocessing

In all our simulations, we assume that \( \eta = 1 \) and use the Epanechnikov kernel \( K_{\text{E}} \) defined in (12). In our real-data simulations in Sections VI-D, VI-E, and VI-G, we set the bandwidth to \( h = 0.3 \ n^{-1/4} \), and in Section VI-F, we set the bandwidth to \( h = 0.43 \ n^{-1/4} \); indeed, \( h \) is typically chosen using ad hoc data-driven techniques in practice [3, Section 1.4]. In our synthetic experiment in Section VI-H, we can directly use the formula for bandwidth given in (10) with \( \gamma = 0.1 \).

The real-world data in Sections VI-D, VI-E, VI-F, and VI-G is available in the form of wins, losses, and draws in tournaments (after additional processing in the case of Section VI-G). For simplicity, we ignore draws and only utilize wins and losses (as in the assumed BTL model). To allow for regularization in the small data regime for our real-data simulations, we apply Laplace smoothing so that between any pair of players, each observed game is counted as 20 games, and

1 additional win is added for each player; this effectively means that \( p = 1 \). In contrast, such smoothing is not needed in our synthetic experiment.

We remark that the constants used to define \( h \), e.g., 0.3 or \( \gamma = 0.1 \), and the level of smoothing mentioned above are chosen to generate “smooth” PDFs in Figures 1, 2, 3, 4, and 5. Moreover, the qualitative results and trends in the sequel remain the same for a range of values around these constants and the chosen level of smoothing.

D. Cricket World Cups

We utilize publicly available data from Wikipedia for international (ICC) Cricket World Cups held in 2003, 2007, 2011, 2015, and 2019. Each World Cup has between \( n = 10 \) to \( n = 16 \) teams, with each pair of teams playing 0, 1, or (rarely) 2 matches against one another. We learn the skill distributions for each World Cup separately as portrayed in Figure 1a. The corresponding negative entropies are reported in Figure 1b. As can be seen, there is a clear decrease in negative entropy from 2003 to 2019, reaching close to 0 in 2019. Figure 1a elegantly quantifies sports intuition that the skill levels of Australia and India dominated those of other teams in 2003, cf. [79], with other teams having roughly similar and lower skill levels. Moreover, Figure 1b shows that the variation of skill levels increased over the years, and the 2019 World Cup had a far greater variation of skill levels than previous years.
E. Soccer World Cups

Again, we use publicly available data from Wikipedia for FIFA Soccer World Cups in 1970, 1974, 1978, ..., 2010, 2014, and 2018. Each World Cup has between \( n = 16 \) to \( n = 32 \) teams, with each pair of teams playing 0, 1, or (rarely) 2 matches. Figures 2a, 2b, and 2c depict the skill distributions and associated negative entropies of Soccer World Cups over the years. It is evident that the negative entropies have increased from 1978 to 2002, and then plateaued (i.e., remained roughly constant and away from 0) from 2002 onwards. Indeed, the skill densities post 1998 in Figure 2b appear to be more concentrated (with less variation of skill levels) than those prior to 1990 in Figure 2a. Since more concentrated skill densities suggests that different teams are more likely to have similar skill levels, this trend in overall skill scores seems consistent with fan experience that Soccer World Cups have evolved over the last half-century to have tighter matches [2].

F. English Premier League Soccer

Yet again, we use publicly available data from Wikipedia for English Premier League (EPL) soccer championships in the 2018-2019, 2019-2020, 2020-2021, 2021-2022, and 2022-2023 seasons. Each league has \( n = 20 \) teams, with every pair of teams playing 0, 1, or 2 times against each other (excluding ties). The structure of EPL championships, where each pair of teams plays 2 matches, differs from Soccer World Cups. So, this experiment demonstrates the use of our algorithm when data is generated under different conditions to the previous experiment. Figure 3a illustrates the skill PDFs of EPL championships, and Figure 3b portrays the negative entropies of these skill PDFs. (Note that for convenience, Figure 3b reports the negative entropy of a season using its second year, e.g., the 2018-2019 season is reported as 2019.) The negative entropies in Figure 3b appear to be on a decreasing trend in general, with a particularly low negative entropy in the 2020-2021 season. This is corroborated by the estimated skill PDFs, where in particular, the 2020-2021 skill PDF is “more uniform” than the skill PDFs of earlier seasons in Figure 3a.

G. US Mutual Funds

Our final real-data experiments are calculated based on the CRSP US Survivor-Bias-Free Mutual Funds Database that is made available by the Center for Research in Security Prices (CRSP), The University of Chicago Booth School of Business, and can be obtained through the Wharton Research Data Services (WRDS) [80]. We consider \( n = 3260 \) mutual funds in this dataset that have monthly net asset values recorded from January 2005 to December 2018. These values are preprocessed by computing monthly returns (i.e., change in net asset value normalized by the previous month’s value) for all funds, which provide a fair measure of monthly performance. Then, we perceive each year as a tournament where each fund plays \( k = 12 \) monthly games against every other fund, and one fund beats another in a month if it has a larger monthly return. Figures 4a and 4b depict the skill PDFs obtained by applying our algorithm to the win-loss data (produced by the method above) every year, and Figure 4c presents the associated negative entropies. Clearly, 2017 and the Great Recession in 2008 were the times where negative entropy was maximized and minimized, respectively, in Figure 4c. Figures 4a and 4b unveil that the skill PDF has much greater variation of skill levels in 2008 compared to 2017, which contains a large peak near 0. Furthermore, Figures 4a, 4b, and 4c convey that we generally have “more uniform” skill distributions after 2008, e.g., in 2009-2016, compared to before 2008, e.g., in 2005-2007. This increased variation of skill levels could be due to a bulk of lower-skilled mutual funds becoming more competent due to the financial crisis. These observations elucidate the potential utility of our algorithm in identifying...
Fig. 3. Plots 3a and 3b illustrate the estimated skill PDFs and corresponding estimated negative differential entropies, respectively, of EPL championships from the 2018-2019 to 2022-2023 seasons. (Note that Figure 3b represents a season using its second year, e.g., the 2018-2019 season is reported as 2019.) and explaining trends in other kinds of data, such as financial data.

H. Synthetic Experiment

We also perform a synthetic experiment to further illustrate the skill PDF estimation procedure in Algorithm 1 on an example where the ground truth is known, and to display qualitative trends of the estimation MSE with respect to important problem parameters. In this experiment, we assume that the true skill PDF \( P_\alpha = \text{unif}([\delta, 1]) \) is a uniform distribution on \([\delta, 1]\) with \( \delta = 0.1 \). (The choice \( \delta = 0.1 \) is arbitrary and the general trends do not change when \( \delta \) is altered slightly.)

For any fixed values of number of agents \( n \), number of comparisons per pair \( k \), and Erdős-Rényi parameter \( p \), we can generate an estimated skill PDF as follows. We generate pairwise comparison data by first sampling \( P_\alpha = \text{unif}([\delta, 1]) \) to obtain \( n \) i.i.d. skill levels, then generating outcomes of \( k \) independent games per pair of agents for all pairs using the BTL model (see (1)), and then applying an “Erdős-Rényi mask” that only keeps the game outcomes for each pair of agents with probability \( p \); see Section I. We next apply Algorithm 1 on this data (with the choices outlined in Section VI-C) and obtain an estimated skill PDF \( \hat{P}_\alpha \).

To illustrate how “close” an estimated skill PDF might look to the true skill PDF, Figure 5a displays the uniform true PDF and an estimated PDF produced by the above procedure in the setting where \( n = 500 \), \( k = 1 \), and \( p = 1 \).

As mentioned earlier, this synthetic setting can be used to portray qualitative trends for the estimation MSE. To this end, for any fixed \( n \), \( k \), and \( p \), we can estimate the estimation MSE (or risk), \( \mathbb{E}[\int_{\mathbb{R}} (\hat{P}_\alpha(x) - P_\alpha(x))^2 \, dx] \), where \( P_\alpha = \text{unif}([\delta, 1]) \), in simulations as follows. Observe that the above procedure produces \( \hat{P}_\alpha \) for one realization of skill levels, outcomes of games, and the Erdős-Rényi graph. So, we can compute one realization of the integrated squared error \( \int_{\mathbb{R}} (\hat{P}_\alpha(x) - P_\alpha(x))^2 \, dx \) using numerical integration via the trapezoidal rule. Furthermore, by repeating the above procedure (with fixed \( n, k, \) and \( p \)) several times, we can obtain
Fig. 5. Plot 5a illustrates an example of skill PDF estimation for \( n = 500 \), \( k = 1 \), and \( p = 1 \). Plots 5b, 5c, and 5d illustrate the decay MSE for skill PDF estimation with \( n \) (where \( k = 1 \) and \( p = 1 \)), \( k \) (where \( n = 500 \) and \( p = 1 \)), and \( p \) (where \( n = 500 \) and \( k = 1 \)), respectively.

In this paper, we presented and analyzed a statistical model that enabled us to rigorously quantify the overall “quality” of relative skill levels in a tournament. Specifically, we assumed that the outcomes of pairwise games between agents are determined by the BTL model, and the skill parameters of this model are drawn from an unknown PDF belonging to a non-parametric class. We then proposed an efficient two-step algorithm to learn these skill distributions from win-loss data of tournaments. Furthermore, we established tight minimax bounds on the skill parameter estimation step of our method (up to logarithmic factors), and then proved that our entire algorithm is minimax near-optimal in the MSE sense when the skill PDF is smooth. We also showed that our estimation error guarantee for the skill PDF implied corresponding statistical guarantees for the estimation of various statistics and functionals of the skill PDF, such as its differential entropy. Finally, using the negative differential entropy of a learnt distribution as an overall skill score, we demonstrated the utility of our algorithm in discerning various trends in sports and financial data.

In closing, we suggest some broad future directions. Firstly, it would be useful to develop minimax optimal algorithms that directly estimate meaningful overall skill scores from tournament data, e.g., KL divergence or Wasserstein distance between \( P_\alpha \) and some fixed distribution. Secondly, the skill distribution estimation problem could be re-formulated and analyzed using different approaches to contend with the inher-
ent scale invariance of BTL skill levels. Thirdly, the BTL model for pairwise comparisons can be too simplistic in certain real-world scenarios. In these cases, it may be suitable to analyze the estimation of skill distributions for other models of pairwise comparisons from the literature, such as the (generalized) Thurstonian model [30] or more general stochastically transitive models, cf. [41]. Lastly, the aforementioned issue could also be partly addressed by deriving “good” hypothesis tests that verify whether observed tournament data truly obeys BTL or related models, cf. [44].

Appendix A

Proofs of Auxiliary Results

In this appendix, we provide proofs of several auxiliary results.

A. Proof of Proposition 1

Proof: First, define the random variables:

\[ \forall m \in [n], \quad M_m \triangleq \sum_{r \in [n] \setminus [m]} 1 \{ \{ m, r \} \in \mathcal{G}(n, p) \} \]

which count the numbers of outcomes of games observed for each player. Furthermore, define the event:

\[ A \triangleq \{ \forall m \in [n], \quad M_m \leq (n-1)p \} \]

Then, we can verify that \( A \subseteq \{ S \in S_{n \times n} \} \). Indeed, if \( A \) occurs, then we have:

\[ \forall m \in [n], \quad \frac{1}{2np} \sum_{r \in [n] \setminus [m]} 1 \{ \{ m, r \} \in \mathcal{G}(n, p) \} \leq \frac{M_m}{2(n-1)p} \leq 1 \]

\[ \Rightarrow \quad \forall m \in [n], \quad \frac{1}{2np} \sum_{r \in [n] \setminus [m]} Z(m, r) \leq 1 \]

\[ \Leftrightarrow \quad \forall m \in [n], \quad S(m, m) \geq 0 \]

\[ \Leftrightarrow \quad S \in S_{n \times n} \]

where the implication follows from (3), and the equivalences follow from (6). Hence, it suffices to prove that \( \mathbb{P}(A^c) \leq n^{-c_1} \).

To this end, notice that:

\[ \mathbb{P}(A^c) = \mathbb{P}(\exists m \in [n], \quad M_m > 2(n-1)p) \]

\[ \leq n \mathbb{P} \left( \frac{1}{n-1} \sum_{r = 2}^{n} 1 \{ \{ 1, r \} \in \mathcal{G}(n, p) \} - p \geq p \right) \]

\[ \leq n \exp \left( -\frac{3(n-1)p}{8} \right) \]

\[ \leq n \exp \left( -\frac{2(c_1 + 1)(n-1) \log(n)}{n} \right) \]

\[ \leq \frac{1}{n^{c_1}} \]

where the second inequality follows from the union bound and the fact that \( \{ M_m : m \in [n] \} \) are identically distributed random variables, the third inequality follows from Lemma 10 in Appendix B since each indicator random variable \( 1 \{ \{ 1, r \} \in \mathcal{G}(n, p) \} \) has mean \( p \), variance \( p(1-p) \leq p \), and satisfies the bound \( 1 \{ \{ 1, r \} \in \mathcal{G}(n, p) \} - p \leq 1 \), the fourth inequality follows from our assumption that \( p \geq 16(c_1 + 1) \log(n)/(3n) \), and the fifth inequality holds because \( n \geq 2 \). This completes the proof. \( \blacksquare \)

B. Proof of Proposition 5

Proof: This is a corollary of Theorem 2. Observe that for any \( P_n \in \mathbb{P} \) and any (Borel measurable) function \( f : \mathbb{R} \to [-1,1] \) bounded by 1, we have:

\[ \mathbb{E} \left[ \left( \int_{\mathbb{R}} f(x) \tilde{P}^*(x) \, dx - \int_{\mathbb{R}} f(x) P_n(x) \, dx \right)^2 \right] \]

\[ = \mathbb{E} \left[ \left( \int_{\mathbb{R}} f(x) \left( \tilde{P}^*(x) - P_n(x) \right) \, dx \right)^2 \right] \]

\[ \leq \mathbb{E} \left[ \left( \int_{\mathbb{R}} \left| \tilde{P}^*(x) - P_n(x) \right| \, dx \right)^2 \right] \]

\[ \leq 3 \mathbb{E} \left[ \int_{-1}^{1} \left( \tilde{P}^*(x) - P_n(x) \right)^2 \, dx \right] \]

where the second inequality follows from Hölder’s inequality, the third equality holds because the bandwidth parameter in (10) satisfies \( h \in (0,1] \) for all sufficiently large \( n \in \mathbb{N} \) (and hence, \( \tilde{P}^* \) has support in \([−2,2]\)), and the fourth inequality follows from the Cauchy-Schwarz-Bunyakovsky inequality. Then, taking suprema over \( P_n \in \mathbb{P} \) and \( f : \mathbb{R} \to [-1,1] \) on both sides yields:

\[ \sup_{P_n \in \mathbb{P}} \sup_{f : \mathbb{R} \to [-1,1]} \mathbb{E} \left[ \left( \int_{\mathbb{R}} f(x) \tilde{P}^*(x) \, dx - \int_{\mathbb{R}} f(x) P_n(x) \, dx \right)^2 \right] \]

\[ \leq 3 \sup_{P_n \in \mathbb{P}} \mathbb{E} \left[ \int_{-1}^{1} \left( \tilde{P}^*(x) - P_n(x) \right)^2 \, dx \right] \]

\[ \leq 3 c_7 \max \left\{ \left( \frac{1}{\delta^p pk} \right)^{\frac{1}{\pi+1}}, 1 \right\} \left( \frac{\log(n)}{n} \right)^{\frac{1}{\pi+1}} \]

where the second inequality follows from Theorem 2 and holds for all sufficiently large \( n \in \mathbb{N} \). This completes the proof. \( \blacksquare \)

C. Proof of Lemma 8

Proof: We first show that \( \mathcal{D} \) is a convex set. For any PDFs \( f, g \in \mathcal{D} \) and every \( \lambda \in [0,1] \), consider the function \( \lambda f + (1-\lambda)g : I \to \mathbb{R} \). It is straightforward to verify that \( \lambda f + (1-\lambda)g \) is a valid PDF that is lower bounded by \( a \). To see that \( \lambda f + (1-\lambda)g \) is \( L_3 \)-Lipschitz continuous, observe that for any \( x, y \in I \):

\[ |\lambda f(x) + (1-\lambda)g(x) - \lambda f(y) - (1-\lambda)g(y)| \]

\[ \leq \lambda |f(x) - f(y)| + (1-\lambda)|g(x) - g(y)| \]

\[ \leq L_3 |x - y| \]

using the triangle inequality and the fact that \( f, g \in \mathcal{D} \). Thus, \( \lambda f + (1-\lambda)g \in \mathcal{D} \), and \( \mathcal{D} \) is convex.

We next show that \( \mathcal{D} \) is closed in \( L^2(I) \). Let \( f \in L^2(I) \) be any limit point of \( \mathcal{D} \). Then, there exists a sequence of PDFs \( f_m \in \mathcal{D} \) with \( m \in \mathbb{N} \) such that \( \lim_{m \to \infty} ||f_m - f||_2 = 0 \). Hence, using the first step of the classical proof of the Riesz-Fischer theorem (cf. [81], Chapter 2, Theorem 2.2 and...
Corollary 2.3, and Chapter 4, Theorem 1.2]), there exists a subsequence of PDFs \( f_n = f_{m_n} \in D \) with \( n \in \mathbb{N} \) such that \( f_n \) converges pointwise to \( f \) almost everywhere, i.e., \( \lim_{n \to \infty} f_n(x) = f(x) \) for all \( x \in \mathbb{I} \cap \mathbb{N} \), where \( \mathbb{I} \subseteq \mathbb{I} \) is a null set with Lebesgue measure 0. It is straightforward to verify that \( f \) is bounded away from almost everywhere (on \( \mathbb{I} \cap \mathbb{N} \)). Moreover, since \( f \in D \), we have that:

\[
\forall x \in I, \ f_n(x) - f_n(x^*) \leq L_3|x - x^*| \leq L_3|I|
\]

where \( x^* \in I \) is such that \( f_n(x^*) = \min_{x \in I} f_n(x) \) (which is well-defined by the extreme value theorem), and \( |I| \) denotes the length of the interval \( I \). Since \( f_n \in D \), we have \( f_n(x^*) \leq |I|^{-1} \), which implies the universal upper bound:

\[
\forall x \in I, \ f_n(x) \leq \frac{1}{|I|} + L_3|I|.
\]

Therefore, by the bounded convergence theorem, we also have that \( \int f_n(x) \, dx = \lim_{n \to \infty} \int f_n(x) \, dx = 1 \). This means that \( f \) is a valid PDF that is bounded by \( a \) almost everywhere. It remains to show that \( f \) is \( L_3 \)-Lipschitz continuous.

To this end, notice that for any \( x, y \in \mathbb{I} \cap \mathbb{N} \):

\[
|f(x) - f(y)| \leq \lim_{n \to \infty} |f(x) - f_n(x)| + |f_n(x) - f_n(y)| + |f_n(y) - f(y)| \leq L_3|x - y| + \lim_{n \to \infty} |f_n(x) - f_n(y)| + |f_n(y) - f(y)| \leq L_3|x - y| \quad (79)
\]

using the triangle inequality and the facts that \( f_n \in D \) and \( f_n \) converges pointwise to \( f \) almost everywhere (on \( \mathbb{I} \cap \mathbb{N} \)). Now, since functions that differ on null sets of Lebesgue measure 0 are considered equivalent in \( L^2(I) \), let us specify the values of \( f \) on \( \mathbb{N} \) as follows. For every \( x \in \mathbb{N} \), consider a sequence \( x_k(x) \in \mathbb{I} \cap \mathbb{N} \) with \( k \in \mathbb{N} \) such that \( \lim_{k \to \infty} x_k(x) = x \); it is straightforward to argue that such a sequence in \( \mathbb{I} \cap \mathbb{N} \) exists because \( \mathbb{N} \) is a null set. Then, the sequence \( \{f(x_k(x)) : k \in \mathbb{N}\} \) is Cauchy due to (79), and must converge since \( \mathbb{R} \) is complete. So, for every \( x \in \mathbb{N} \), let \( f(x) = \lim_{k \to \infty} f(x_k(x)) \). Thus, we have defined a representative \( f : \mathbb{I} \to \mathbb{R} \) (with abuse of notation) of the equivalence class of all functions that differ from \( f \) on null sets:

\[
f(x) = \begin{cases} \lim_{n \to \infty} f_n(x), & x \in \mathbb{I} \cap \mathbb{N}, \\ \lim_{k \to \infty} f(x_k(x)), & x \in \mathbb{N}.
\end{cases}
\]

It is easy to see that this representative \( f \) is lower bounded by \( a \) everywhere (i.e., \( f(x) \geq a \) for all \( x \in I \)). Moreover, for any \( x \in \mathbb{N} \) and \( y \in \mathbb{I} \cap \mathbb{N} \), the representative \( f \) satisfies:

\[
|f(x) - f(y)| \leq \lim_{k \to \infty} |f(x) - f_k(x)| + |f_k(x) - f(y)| \leq \lim_{k \to \infty} |f(x) - f_k(x)| + L_3 \lim_{k \to \infty} |x - y| \leq L_3|x - y|
\]

using the triangle inequality and (79). Likewise, for any \( x, y \in \mathbb{N} \), the representative \( f \) satisfies \( |f(x) - f(y)| \leq L_3|x - y| \). Hence, the representative \( f \) is \( L_3 \)-Lipschitz continuous everywhere (i.e., \( |f(x) - f(y)| \leq L_3|x - y| \) for all \( x, y \in I \)). Overall, we have established that the carefully defined representative above for any limit point of \( \mathbb{D} \) lives in \( D \). Thus, \( D \) is a closed set in \( L^2(I) \). This completes the proof.

\[\blacksquare\]

D. Proof of Claim 2

Proof: This is a straightforward exercise in calculus. Indeed, we have for all \( x, y \geq \delta \):

\[
\frac{\partial F}{\partial y}(x, y) = \frac{x}{(x+y)^2} \quad \text{and} \quad \frac{\partial F}{\partial x}(x, y) = \frac{y}{(x+y)^2},
\]

which implies that:

\[
\max_{x, y \geq \delta} \left| \frac{\partial F}{\partial y}(x, y) \right| = \max_{x, y \geq \delta} \left| \frac{\partial F}{\partial x}(x, y) \right| = \max_{t \geq \delta} \frac{t}{(t+\delta)^2} = \frac{1}{4\delta},
\]

where the final equality holds because it is easy to verify that the map \( \delta \leq t \mapsto t/(t+\delta)^2 \) is globally maximized at \( t = \delta \). This establishes the Lipschitz constants in parts 1 and 2 of Claim 2.

\[\blacksquare\]

APPENDIX B

CONCENTRATION OF MEASURE INEQUALITIES

In this appendix, we present two well-known exponential concentration of measure inequalities that are used in this paper. The first of these results bounds the tail probability of the empirical average of a collection of independent bounded random variables.

Lemma 9 (Hoeffding’s Inequality [82, Theorems 1 and 2]): Given independent random variables \( X_1, \ldots, X_n \in [a, b] \), for some constants \( a < b \), we have for every \( \varepsilon > 0 \):

\[
\Pr\left( \frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}[X_i] \geq \varepsilon \right) \leq \exp\left( -\frac{2n\varepsilon^2}{(b-a)^2} \right).
\]

The second of these results provides a tighter bound on the tail probability of the empirical average of a collection of independent bounded random variables using information about the variances of the random variables.

Lemma 10 (Bernstein’s Inequality [83]): Given independent random variables \( X_1, \ldots, X_n \) such that for some constants \( a, b > 0 \), \( |X_i - \mathbb{E}[X_i]| \leq a \) and \( \forall \bar{X} \mathbb{E}(X_i) \leq b \) for all \( i \in [n] \), we have for every \( \varepsilon > 0 \):

\[
\Pr\left( \frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}[X_i] \geq \varepsilon \right) \leq \exp\left( -\frac{n\varepsilon^2}{2b + \frac{4}{3}a^2} \right).
\]

ACKNOWLEDGMENT

Note: The author ordering is alphabetical.

REFERENCES

[1] A. Jadbabaie, A. Makur, and D. Shah, “Estimation of skill distribution from a tournament,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Vancouver, BC, Canada, Dec. 2020, pp. 8418–8429.

[2] P. Scott and A. Kirk. (Jun. 13, 2018). Is the World Cup as Exciting as It Used to be?. The Telegraph. Accessed: May 29, 2020. [Online]. Available: https://www.telegraph.co.uk/world-cup/how-has-the-wc-changed/

[3] A. B. Tsybakov, Introduction To Nonparametric Estimation (Springer Series in Statistics). New York, NY, USA: Springer, 2009.

[4] R. A. Bradley and M. E. Terry, “Rank analysis of incomplete block designs. I. The method of paired comparisons,” Biometrika, vol. 39, nos. 3–4, pp. 324–345, Dec. 1952.
I. Ahmad and P.-E. Lin, “A nonparametric estimation of the entropy for absolutely continuous distributions (corresp.),” IEEE Trans. Inf. Theory, vol. IT-22, no. 3, pp. 372–375, May 1976.

Y. Han, J. Jiao, T. Weissman, and Y. Wu, “Optimal rates of entropy estimation over Lipschitz balls,” Ann. Statist., vol. 48, no. 6, pp. 3228–3250, Dec. 2020.

D. M. Covington and J. A. Thomas, Elements of Information Theory, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

Y. Polyanskiy and Y. Wu, Information Theory: From Coding to Learning. New York, NY, USA: Cambridge Univ. Press, 2023.

I. Csiszár, “A class of measures of informativity of observation channels,” Periodica Mathematica Hungarica, vol. 2, nos. 1–4, pp. 191–213, Mar. 1972.

F. E. Su, “Methods for quantifying rates of convergence for random walks on groups,” Ph.D. thesis, Dept. Math., Harvard Univ., Cambridge, MA, USA, 1995.

A. Makur and L. Zheng, “Bounds between contraction coefficients,” in Proc. 53rd Annu. Allerton Conf. Commun., Control, Comput. (Allerton), Monticello, IL, USA, Sep. 2015, pp. 1422–1429.

A. Makur and L. Zheng, “Comparison of contraction coefficients for $f$-divergences,” Problems Inf. Transmiss., vol. 56, no. 2, pp. 103–156, Apr. 2020.

A. Makur, “Information contraction and decomposition,” Sc.D. thesis, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, May 2019.

R. A. Horn and C. R. Johnson, Matrix Analysis. Cambridge, MA, USA, 1995.

X. Wang, “Volumes of generalized unit balls,” Math. Mag., vol. 78, no. 5, pp. 390–397, Oct. 2005.

L. Gordon, “A stochastic approach to the gamma function,” Amer. Math. Monthly, vol. 101, no. 9, pp. 858–865, Nov. 1994.

W. Feller, An Introduction to Probability Theory and Its Applications, vol. 1, 3rd ed. New York, NY, USA: Wiley, 1968.

R. B. Ash, Information Theory (Interscience Tracts in Pure and Applied Mathematics), no. 19. New York, NY, USA: Wiley, 1965.

C. Villani, Optimal Transport: Old and New (Grundlehren der mathematischen Wissenschaften), vol. 338. Berlin, Germany: Springer, 2009.

P. D. Lax, Functional Analysis. New York, NY, USA: Wiley, 2002.

H. H. Bauschke and P. L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces (CMS Books in Mathematics). New York, NY, USA: Springer, 2011.

W. Cheney and A. A. Goldstein, “Proximity maps for convex sets,” Proc. Amer. Math. Soc., vol. 10, no. 3, pp. 448–450, Jun. 1959.

T. Popoviciu, “Sur les équations algébriques ayant toutes leurs racines réelles,” Mathematica, vol. 9, pp. 129–145, Aug. 1935.

R. Bhatia and C. Davis, “A better bound on the variance,” Amer. Math. Monthly, vol. 107, no. 4, pp. 353–357, Apr. 2000.

A. Peter. (Jun. 18, 2020), The Greatest ODI teams in history: Ricky Ponting and His All-Conquering Australia Squad of 2002-04. Sport 360. Accessed: Sep. 3, 2023. [Online]. Available: https://sport360.com/article/cricket/international-cricket/5343492/the-greatest-odi-teams-in-history-ricky-ponting-and-his-all-conquering-australia-squad-of-2002-04

CRSP US Survivor-Bias-Free Mutual Funds Database, Center for Research in Security Prices (CRSP), Booth School of Business, The University of Chicago, Chicago, IL, USA. Wharton Research Data Services (WRDS), The Wharton School, Univ. Pennsylvania, Philadelphia, PA, USA. Accessed: 2020. [Online]. Available: https://wrds-www.wharton.upenn.edu/pages/about/data-vendors/center-for-research-in-security-prices-crsp/

E. M. Stein and R. Shakarchi, Real Analysis: Measure Theory, Integration, and Hilbert Spaces (Princeton Lectures in Analysis), vol. 3. Princeton, NJ, USA: Princeton Univ. Press, 2005.

W. Hoeffding, “Probability inequalities for sums of bounded random variables,” J. Amer. Stat. Assoc., vol. 58, no. 301, pp. 13–30, 1963.

S. N. Bernstein, The Theory of Probabilities. Moscow, Russia: Gasteizdat Publishing House, 1946.

Ali Jadbabaie (Fellow, IEEE) received the B.S. degree from the Sharif University of Technology, Tehran, Iran, in 1995, the M.S. degree in electrical and computer engineering from The University of New Mexico, Albuquerque, NM, USA, in 1997, and the Ph.D. degree in control and dynamical systems from California Institute of Technology, Pasadena, CA, USA, in 2000. He was a Post-Doctoral Scholar with Yale University, New Haven, CT, USA, before joining as a Faculty Member with the University of Pennsylvania, Philadelphia, PA, USA, where he was subsequently promoted through the ranks and was the Alfred Fitler Moore Professor of network science with the Electrical and Systems Engineering Department with secondary appointments in computer and information science and operations, information and decisions with the Wharton School. As a member of the General Robotics, Automation, Sensing and Perception Laboratory, University of Pennsylvania, he was also the Co-Founder and the Director of the Raj and Neera Singh Program in networked and social systems engineering (NETS), a new undergraduate interdisciplinary degree program. He is currently the JR East Professor and the Head of the Department of Civil and Environmental Engineering, a Core Faculty Member with the Institute for Data, Systems, and Society (IDSS), and a PI with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. Previously, he was the Director of the Sociotechnical Systems Research Center, MIT, and the Associate Director of IDSS, MIT, which he helped founded in 2015. His current research interests include the interplay of dynamic systems and networks with specific emphasis on multi-agent coordination and control, distributed optimization, network science, and network economics. He was a recipient of the National Science Foundation CAREER Award, the Office of Naval Research Young Investigator Award, the O. Hugo Schuck Best Paper Award from the American Automatic Control Council, the George S. Axelby Best Paper Award from the IEEE Control Systems Society, and the Vannevar Bush Fellowship from the Office of the Secretary of Defense. He is a senior author of several student best paper awards. He was the Inaugural Editor-in-Chief of IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, a new interdisciplinary journal sponsored by several IEEE societies.

Anuran Makur (Senior Member, IEEE) received the B.S. degree (summa cum laude) from the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley (UC Berkeley), CA, USA, in 2013, and the S.M. and Sc.D. degrees from the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2015 and 2019, respectively. From 2019 to 2021, he was a Post-Doctoral Researcher with the Laboratory for Information and Decision Systems and the Institute for Data, Systems, and Society, MIT. He is currently an Assistant Professor with the Department of Computer Science and the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. His research interests include theoretical statistics and machine learning, information theory, and other areas in applied probability. He was a recipient of the Arthur M. Hopkin Award from UC Berkeley in 2013, the Jacobs Presidential Fellowship from MIT in 2013, the Ernst A. Guillemin Master’s Thesis Award from MIT in 2015, the Jin Au Kong Doctoral Thesis Award from MIT in 2020, the Thomas M. Cover Dissertation Award from the IEEE Information Theory Society in 2021, and the CAREER Award from the National Science Foundation in 2023.

Devavrat Shah (Fellow, IEEE) is currently the Andrew (1956) and Erna Viterbi Professor with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT). He is also a member of the Laboratory for Information and Decision Systems and the Statistics and Data Science Center, Institute for Data, Systems, and Society, MIT. His current research interests include graphical models, causal inference, stochastic networks, and social data processing. His work has been recognized through prize paper awards in machine learning, operations research, electrical engineering, and computer science, as well as career prizes, including the 2010 Erlang Prize from the INFORMS Applied Probability Society and the 2008 ACM Sigmetrics Rising Star Award. He is a Distinguished Alumni of his alma mater HIT Bombay and a Kavli Fellow of the National Academy of Sciences.