Reconstructing weighted voting schemes from partial information about their power indices

Huck Bennett   HDBCO@UMICH.EDU
University of Michigan

Anindya De   ANINDYAD@CIS.UPENN.EDU
University of Pennsylvania

Rocco A. Servedio   ROCCO@CS.COLUMBIA.EDU
Columbia University

Emmanouil-Vasileios Vlatakis-Gkaragkounis   EMVLATAKIS@CS.COLUMBIA.EDU
Columbia University

Abstract

A number of recent works (Goldberg, 2006; O'Donnell and Servedio, 2011; De et al., 2017, 2014) have considered the problem of approximately reconstructing an unknown weighted voting scheme given information about various sorts of "power indices" that characterize the level of control that individual voters have over the final outcome. In the language of theoretical computer science, this is the problem of approximating an unknown linear threshold function (LTF) over $\{-1, 1\}^n$ given some numerical measure (such as the function's n "Chow parameters," a.k.a. its degree-1 Fourier coefficients, or the vector of its n Shapley indices) of how much each of the n individual input variables affects the outcome of the function.

In this paper we consider the problem of reconstructing an LTF given only *partial* information about its Chow parameters or Shapley indices; i.e. we are given only the Chow parameters or the Shapley indices corresponding to a subset $S \subseteq [n]$ of the n input variables. A natural goal in this partial information setting is to find an LTF whose Chow parameters or Shapley indices corresponding to indices in S accurately match the given Chow parameters or Shapley indices of the unknown LTF. We refer to this as the *Partial Inverse Power Index Problem*.

Our main results are a polynomial time algorithm for the (ε-approximate) Chow Parameters Partial Inverse Power Index Problem and a quasi-polynomial time algorithm for the (ε-approximate) Shapley Indices Partial Inverse Power Index Problem.

Keywords: Power Index, Inverse Problem, Chow Parameters, Shapley Indices, Boolean Fourier Analysis Linear Threshold Function

1. Introduction: Power indices and inverse power index problems.

A natural question that arises in voting theory is how to quantify the "power" of an individual voter in a collective decision-making scheme. For simplicity, in this paper we consider only *weighted voting games*; in the language of theoretical computer science, these correspond to *linear threshold functions (LTFs)* $f : \{-1, 1\}^n \to \{-1, 1\}$, $f(x) = \text{sign}(w \cdot x - \theta)$, where each $w_i \geq 0$ is a non-negative *voting weight*. In such a weighted voting game there are n binary voters, each with some amount of non-negative weight, and the collective decision is an affirmative one if and only if the total voting weight of the affirmative voters exceeds the threshold θ.

If all n of the voting weights are the same then it is clear that all n voters have the same amount of "power" over the final outcome, but it is much less clear how to measure the power of a single voter when the voting weights may vary. As a simple example, consider a setting with three voters who have voting weights of $49, 49$ and 2, in which a total of 51 votes are required for the proposition to pass. While the disparity between voting weights may at first suggest that the two voters with 49 votes each have most of the "power," any coalition of two voters is sufficient to pass the proposition and any single voter is insufficient, so the voting power of all three voters is in fact equal. Such examples are not merely hypothetical; in the first voting scheme used by the European Economic Community (the predecessor of the current European Union) in 1957, decisions were accepted if they were supported by at least 12 out 17 votes, and the members' weight distribution was {Germany : 4, France : 4, Italy : 4, The Netherlands : 2, Belgium : 2, Luxembourg : 1} (European-Union, 1957; Leech, 2002). Inspection shows that there is no voting outcome in which Luxembourg could influence the result, and thus its real voting power was null.

A number of different numerical measures, known as "power indices," have been proposed to quantify how much power each voter has in a weighted voting election scheme. These include the Deegan-Packel index (Deegan and Packel, 1978), the Holler index (Holler, 1982; Johnston, 1978), and several others (see the extensive survey of de Keijzer (de Keijzer, 2008)). In the rest of this paper we confine our attention to two particularly well-studied power indices. The first of these are the *Banzhaf indices* (Banzhaf III, 1964); in theoretical computer science these are more commonly known as the *Chow parameters* (Chow, 1961) and we shall henceforth refer to them as such. There are $n + 1$ Chow parameters of an n-variable LTF, and they are simply the constant- and degree-1 Fourier coefficients.[1] The second of these are the *Shapley-Shubik indices* (Shapley and Shubik, 1954), henceforth referred to for brevity as the *Shapley indices*; these are perhaps the best known, and certainly the oldest, power indices studied in the literature. Given an LTF $f : \{-1, 1\}^n \to \{-1, 1\}$ with non-negative weights that satisfy $f((-1)^n) = -1$, $f(1^n) = 1$, the Shapley indices are a vector of n associated probabilities $(\mathring{f}(1), \ldots, \mathring{f}(n))$ that sum to 1. The ith probability is the probability that x_i is the "pivotal index" causing f's value to flip from -1 to 1, starting at the input $(-1)^n$ and flipping indices from -1 to 1 in a random order.

The #P-hardness of counting 0/1 knapsack solutions easily implies that it is #P-hard to exactly compute the Chow parameters of a given LTF, and it turns out that the Shapley indices of LTFs are also #P-hard to compute (Deng and Papadimitriou, 1994). However, simple sampling-based approaches yield efficient algorithms for obtaining highly accurate estimates of the Chow parameters or the Shapley indices (see e.g. (Leech, 2003; Bachrach et al., 2010)). Much more challenging are the *inverse* problems, such as the *(Inverse) Chow Parameters Problem*: given as input the Chow parameters of an unknown LTF (or accurate approximations of the Chow parameters), construct an LTF whose Chow parameters are very close to the input provided. A beautiful result of C.-K. Chow from the 2nd FOCS conference (Chow, 1961) shows that given the exact Chow parameters of an LTF, it is information-theoretically possible to recover the LTF, but the proof is entirely non-constructive. The algorithmic problem of appproximating an unknown LTF from its Chow parameters was studied in a number of recent works (Goldberg, 2006; O'Donnell and Servedio, 2011; De et al., 2014), and more recently the analogous problem for Shapley indices (the *Inverse Shapley Indices Problem)* was studied as well (De et al., 2017). The current state of the art for the Inverse Chow Parameters Problem (De et al., 2014) is an algorithm which, for any constant ε, runs in fixed

---

1. Since every LTF is a unate Boolean function, up to sign the degree-1 Fourier coefficients are the same as the n coordinate influences of the LTF.

poly(n) time and outputs an LTF whose Chow parameters match the given input vector of Chow parameters up to ε-accuracy (in a sense which we make precise later). A similar-in-spirit result (with some technical restrictions and somewhat weaker quantitative bounds; we defer a precise statement until later) was given for the inverse Shapley indices problem in (De et al., 2017). We also remark here that the problem of exactly reconstructing a LTF from its Chow parameters (or its Shapley indices) was recently (Diakonikolas and Pavlou, 2019) shown to be computationally intractable.

2. This work: The Partial Inverse Power Index Problem.

A drawback of the algorithms of (Goldberg, 2006; O'Donnell and Servedio, 2011; De et al., 2017) and (De et al., 2014) for the Inverse Chow Parameters and Inverse Shapley Indices Problems is that they require *full information* about the target vector of power indices: none of these algorithms can be used unless *all* of the target Chow parameters (or Shapley indices) are provided to the algorithm. This is a potentially significant drawback for settings situations in which exhaustive information about the target power indices may not be available.

A natural first idea to deal with partial-information scenarios is to attempt some sort of trivial work-around, for example by simply assigning zero as the target Chow parameter or Shapley index for each coordinate of the vector for which information is not available, and then running a known algorithm such as (Goldberg, 2006; O'Donnell and Servedio, 2011; De et al., 2017, 2014) for the full-information case. It is not difficult to show, though, that a naive approach such as this may lead to the incorrect conclusion that there exists no LTF with the desired vector of power indices, when in fact such an LTF does indeed exist.[2] Thus more sophisticated approaches are required.

This current paper develops such approaches; it studies algorithms for the *Partial Inverse Power Index Problem.* In this partial information version of the problem, the algorithm is only given a subset $S \subset [n]$ of the n "voters" (i.e. coordinates of the unknown LTF f) and the associated power indices (Chow parameters or Shapley indices) corresponding to those coordinates, and the goal is to output a weighted voting game f (i.e. an LTF) such that the power indices of f in coordinates S closely match the input that was provided. We give algorithms for both the Chow Parameters and Shapley Indices version of this problem; to explain our results, we begin by giving a detailed definition of each of these problems below.

2.1. The Partial Chow Parameters Problem

We begin by recalling the definition of the Chow parameters:

Definition 1 *Given $f : \{-1,1\}^n \to \{-1,1\}$, for $0 \leq i \leq n$ the ith* Chow parameter *of f is the value*

$$\hat{f}(i) := \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \{-1,1\}^n}[f(\boldsymbol{x})\boldsymbol{x}_i],$$

where we define $\boldsymbol{x}_0$ to be identically 1 and "$\boldsymbol{x} \sim \{-1,1\}^n$" indicates that $\boldsymbol{x}$ is a uniform random element of $\{-1,1\}^n$.

---

2. As a simple example, suppose that the unknown LTF is the majority function over three variables x_1, x_2, x_3 and only the Chow parameter (degree-1 Fourier coefficient) of x_1 is provided. Each Chow parameter of the three-variable majority function is $1/2$, but there is no 1-variable LTF which has $1/2$ as its Chow parameter for x_1 (the value must be either $-1, 0$, or 1). More sophisticated asymptotic families of similar examples, and analogous examples for Shapley indices, are similarly easy to construct.

Thus the Chow Parameters of a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ are simply its $n+1$ degree-0 and degree-1 Fourier coefficients. The *Chow Parameters Problem* is the problem of (approximately) recovering a weights-based representation of a linear-threshold function (LTF) f given the Chow Parameters of f as input. The (approximate) *Partial Chow Parameters Problem* is the partial information variant of the Chow Parameters Problem where only a *subset* of the Chow Parameters of f corresponding to some subset of indices S are given as input, and the goal is to recover a weights-based representation of an LTF f' such that the "partial Chow distance with respect to S" between f and f', namely $\left(\sum_{i \in S} (\hat{f}(i) - \hat{f}'(i))^2 \right)^{1/2}$, is small:

Definition 2 *The ε-approximate* Partial Chow Parameters Problem *is the promise problem defined as follows. Given $\{(i, \hat{f}(i)) : i \in S\}$ for some LTF $f : \{-1,1\}^n \to \{-1,1\}$ and some $S \subseteq \{0, 1, \ldots, n\}$ as input, output weights $w_1, \ldots, w_n$ and a threshold θ such that $f'(x) := \text{sign}(w \cdot x - \theta)$ satisfies $\left(\sum_{i \in S} (\hat{f}(i) - \hat{f}'(i))^2 \right)^{1/2} \leq \varepsilon$.*

Note that we do not require $\hat{f}(i)$ and $\hat{f}'(i)$ to be close for $i \notin S$, and indeed this would be impossible for any algorithm to achieve (for example, the target LTF f could be any LTF in the extreme case where $S = \emptyset$). We also note that the Partial Chow Parameters Problem generalizes the Chow Parameters Problem, since the latter is simply the special case of the former where $S = \{0, 1, \ldots, n\}$.

2.2. The Partial Shapley Parameters Problem

We begin by defining the notion of the Shapley indices. Given a permutation π mapping $[n]$ to $[n]$, let $x(\pi, i) \in \{-1,1\}^n$ be the string that has a 1 in each coordinate j with $\pi(j) < \pi(i)$ and a -1 in all other coordinates. Define $x^+(\pi, i) \in \{-1,1\}^n$ to be $x(\pi, i)$ but with the ith coordinate flipped from -1 to 1.

Definition 3 *Given a monotone function $f : \{-1,1\}^n \to \{-1,1\}$, the ith (generalized)* Shapley index *of f is the value*

$$\mathring{f}(i) := \frac{1}{2} \mathop{\mathbf{E}}_{\pi \sim \mathbb{S}_n} [f(x^+(\pi, i)) - f(x(\pi, i))].$$

Thus for a non-constant, monotone LTF f, $\mathring{f}(i)$ is the probability that, starting from $x = (-1)^n$ and flipping randomly chosen coordinates of x that are -1 one at a time to 1, i is the unique *pivotal index* for which flipping x_i changes $f(x)$ from -1 to 1. The (approximate) *Partial Shapley Indices Problem* is defined analogously to the Partial Chow Parameters problem:

Definition 4 *The ε-approximate* Partial Shapley Indices Problem *is the promise problem defined as follows. Given $\{(i, \mathring{f}(i)) : i \in S\}$ for some LTF $f : \{-1,1\}^n \to \{-1,1\}$ and some $S \subseteq \{1, \ldots, n\}$ as input, output weights $w_1, \ldots, w_n$ and a threshold θ such that $f'(x) := \text{sign}(w \cdot x - \theta)$ satisfies $\left(\sum_{i \in S} (\mathring{f}(i) - \mathring{f}'(i))^2 \right)^{1/2} \leq \varepsilon$.*

3. Our results

Our first main result is an efficient algorithm for the Chow parameters version of the Partial Inverse Power Index Problem:

4

Theorem 5 (Informal statement) *There is a $\mathrm{poly}(n, 2^{\mathrm{poly}(1/\varepsilon)})$-time algorithm for the ε-approximate Partial Chow Parameters Problem.*

The algorithm of Theorem 5 is an "EPRAS," meaning that its running time is a fixed polynomial in n independent of ε, but depends super-polynomially on ε.

Our second main result is an efficient algorithm for the Shapley parameters version of the Partial Inverse Power Index Problem:

Theorem 6 (Informal statement) *There is a $2^{((\log n)/\varepsilon)^c}$-time algorithm for some absolute constant $c > 0$ for the ε-approximate Partial Shapley Indices Problem.*

Here our algorithm is an "EQPRAS," meaning that its running time is a fixed quasi-polynomial function of n for any value of ε, but depends super-polynomially on ε.

Discussion. We stress that (as is implied by the discussion following Definition 2) there may not be a unique target LTF that is consistent with the given set of Chow parameters or Shapley-Shubik values. For example, suppose the target LTF is a simple unweighted majority of some $n/2$ of the n input variables, and the input set S provided to the algorithm is some subset of $n/4$ of the $n/2$ relevant variables; then clearly any $n/2$-variable majority function including the $n/4$ variables in the input set would be an acceptable solution.

A second note is that our algorithmic results focus on finding *some* LTF whose Chow or Shapley Power Indices (approximately) match a given input set of Chow or Shapley Power Indices, but without giving other guarantees on that LTF. In light of our results, some natural next questions to consider are versions of these Partial Inverse Power Index Problems with additional objectives: can one efficiently find an LTF that completes the given input set of power indices and "spreads" the remaining power that is left to be assigned as evenly (or unevenly) as possible? that minimizes (or maximizes) the total amount of remaining power that needs to be assigned?[3] In fact, our algorithm for the Partial Inverse Chow Parameters Problem is already fairly flexible, and can be adjusted, for example, either to spread or concentrate the remaining power to assign (we leave the details to the interested reader). We expect that the ideas and technical ingredients in this paper will be useful for tackling these next questions, but for now we leave them as intriguing directions for future work.

4. Our techniques for the Chow problem

We begin by giving a high level overview of our algorithm (and associated proof of correctness) for the partial Chow parameters problem. The techniques for the corresponding problem for Shapley indices build on the techniques for the Chow problem.

We begin by recalling the important notion of *regularity* of a linear form (e.g., see Servedio (2007)). A linear form $w \cdot x$ (where $w \neq 0^n$) is said to be τ-*regular* if $\max_j |w_j|/\|w\|_2 \leq \tau$. Regularity plays a crucial role in Boolean function analysis because of the Berry-Esséen theorem, which says that for $x \sim \{-1, 1\}^n$ the random variable $w \cdot x - \theta$ "behaves like a Gaussian with mean $-\theta$ and variance $\|w\|_2$". In fact, the Berry-Esséen theorem can be used to establish analogous statements whenever the n-dimensional random variable x comes from a product distribution with bounded third moments.

---

3. This last question only makes sense for the Chow Parameters variant, since the Shapley indices of any monotone function must always sum to 1.

Moving on to halfspaces, the notion of regularity has played a crucial role in their analysis ever since it was first used in Servedio (2007) (though implicitly it was used in the earlier works of Khot et al. (2007); Mossel et al. (2010)). The reason this notion is useful for us is as follows: Suppose $w \in \mathbb{R}^n$ is a τ-regular vector and $f(x) = \text{sign}(w \cdot x - \theta)$ is a corresponding LTF. Then a result first proven in Matulef et al. (2010) (but which essentially can be derived from Khot et al. (2007)) is that there exists an (explicit) constant c_θ (depending on θ) such that

$$\sum_{i=1}^{n}(\widehat{f}(i) - c_\theta w_i)^2 = O(\tau^{\frac{1}{2}}) \tag{1}$$

(see Proposition 22). In other words, for τ-regular LTFs, the Chow parameters are (close to) a linear rescaling of the LTF's weights.

Now, suppose we were given the promise, in the partial Chow parameters problem, that the target LTF f is in fact τ-regular for $\tau = O(\varepsilon^2)$ (where ε is the desired accuracy parameter for the reconstruction). Then Equation (1) suggests a very simple algorithm for the partial Chow parameters problem in this case:

1. While the threshold parameter θ is not known, we can *guess* it. What this means precisely is the following: since (without loss of generality) we may assume that $\|w\|_2 = 1$, it must be the case that $\theta \in [-\sqrt{n}, \sqrt{n}]$. It is not difficult to show that if instead of having the exact value of θ, we had it up to an additive $\pm\delta$, this adds only a small inaccuracy (roughly, $O(\tau+\delta)$) to the error of the final reconstruction. Thus, what we can do is to try out all possible values of θ in a grid over $[-\sqrt{n}, \sqrt{n}]$ where the width of the grid is some sufficiently small δ. While performing such a guessing step means that we will have a batch of several candidate LTFs, it is straightforward to do *hypothesis testing* at the end, by simply estimating the Chow parameters of each hypothesis LTF and outputting the one which most closely matches the input, to identify a successful candidate from the batch.

   (More generally, several times in this informal description of our algorithms we will employ such a "guessing of parameters." Suppose that the total number of parameters is P and that the grid space for each parameter is L; then iterating over all the possibilities and the subsequent hypothesis testing adds a multiplicative running time overhead of $\approx L^P$. Thus, as long as P is small and L is not too large, the total overhead incurred from guessing parameters is small. In the rest of this informal overview, whenever we "guess a parameter", we will assume that we have its value exactly and not account for (i) either the small inaccuracy due to the granularity of the grid or (ii) multiplicative overhead created by iterating over the possibilities.)

2. Given the parameter θ, we can explicitly compute the constant c_θ. Given the value c_θ, the most obvious approach to the partial Chow parameters problem is to define the quantities $v_1, \ldots, v_n$ as follows: For $i \in S$, we define $v_i := \frac{\widehat{f}(i)}{c_\theta}$. We further define $\text{wt} := \sum_{i \in S} v_i^2$ and define $v_i = \sqrt{(1 - \text{wt})/(n - |S|)}$. Finally, we output the halfspace $g(x) = \text{sign}(v \cdot x - \theta)$. The intuition behind this is that for coordinates $i \in S$, (1) suggests the correct value of w_i (which is what we set v_i to be). For all the other coordinates, we set v_i to be "as regular as

possible".[4] Given (1), it easily follows that $\sum_{i \in S}(\widehat{g}(i) - \widehat{f}(i))^2 = O(\tau^{\frac{1}{2}})$, which is $O(\varepsilon)$ by our choice of parameters.

To handle the case when the unknown LTF f is not τ-regular, we use the "critical index" machinery of Servedio Servedio (2007). To explain how this is done, for ease of exposition let us assume that $f = \text{sign}(w \cdot x - \theta)$ is such that $|w_1| \geq \ldots \geq |w_n|$. The τ-critical index of the vector w (or equivalently, of any linear form $w \cdot x - \theta$) is the smallest index j such that $|w_{j+1}|/\sqrt{\sum_{k>j} w_k^2} \leq \tau$. Thus, a vector w is τ-regular if and only if its τ-critical index is zero. If the τ-critical index of a vector w is not defined, then we say it is ∞.

We now discuss the general algorithmic strategy for the partial Chow parameters problem; as explained below, the strategy depends on the value of the critical index. (While the actual value of the critical index is not known to the algorithm, the algorithm can just "guess" which of the following three cases it is in, followed by hypothesis testing at the very end.)

1. **First case: τ-critical index is large:** This is the case when the τ-critical index $K(\tau)$ is at least $O(\tau^{-2} \log(1/\tau))$. In this case, Servedio Servedio (2007) shows that f is $O(\tau)$-close in Hamming distance to a LTF g which depends only on $K(\tau)$ variables. As f and g are close in Hamming distance, it follows that $\sum_{i \in S}(\widehat{f}(i) - g(i))^2 = O(\tau)$. The algorithm in this case simply enumerates over all LTFs on $K(\tau)$ variables – there are $2^{O(K^2(\tau) \log K(\tau))}$ such LTFs – and for each such LTF g, checks if it is a solution to the partial Chow parameters problem.

2. **Second case: τ-critical index is zero:** This is the case where the linear form $w \cdot x - \theta$ is τ-regular. We have already described the reconstruction algorithm in this case.

3. **Third case: τ-critical index is small:** This is the case when the τ critical index is non-zero but is at most $O(\tau^{-2} \log(1/\tau))$. This case, which is technically the most challenging, combines ingredients from the large and zero critical index cases. Let us assume that the unknown weights are $w_1, \ldots, w_n$ and that the critical index is $K(\tau)$. First of all, the algorithm will guess $K(\tau)$ (note that there are only $O(\tau^{-2} \log(1/\tau))$ possibilities). The algorithm will also guess $w_1, \ldots, w_{K(\tau)}$. Finally, given $\{\widehat{f}(i)\}_{i \in S}$, the algorithm will also guess the subset $T = \{j \in S : j \geq K(\tau)\}$. (Note that the number of choices for T which must be considered can be bounded by $K(\tau)$, since the weights before the critical index (the largest magnitude weights) must correspond to the largest Chow parameters.) Having fixed all these choices, the crucial fact, analogous to Equation (1), is that there exists (an explicitly computable) $c = c(\theta, w_1, \ldots, w_{K(\tau)})$ such that $\sum_{i \geq K}(\widehat{f}(i) - c w_i)^2 = O(\tau^{1/2})$ (see Proposition 24). The algorithm can now compute c and finding a feasible $w_K, \ldots, w_n$ is essentially the same as case (ii), i.e. the zero critical index case.

Finally, we note that the actual algorithm and its analysis is split into two cases, namely, the large versus small critical index cases (and not three cases as described above). In particular, the zero critical index case is subsumed by the small critical index case. However, the small critical index case is both conceptually and technically a combination of the ideas for the zero critical index

---

4. Actually, in a slight deviation from what is described above, our actual algorithm for the regular case performs a slight numerical adjustment to avoid the pathological case in which wt slightly exceeds 1, which would make our algorithm meaningless.

and the large critical index cases. Thus, for expository reasons, in this introduction we have split the analysis into three cases.

While we are glossing over several technical details, the actual algorithm and analysis essentially follows the above description. We now turn to giving a high level overview of the techniques for the partial Shapley value problem.

5. Our techniques for the Shapley problem.

At the highest level, the structure of our algorithm for the Shapley problem is similar to our algorithm for the Chow problem — a case split based on whether the critical index is large, zero, or small — but the analysis and underlying structural results are considerably more involved. (Similar to the Chow problem, the actual algorithm and its analysis has only two cases, the large and the small critical index. However, for the sake of exposition, similar to the Chow problem we describe a three case split here in the introduction.)

Let $g = \text{sign}(v \cdot x - \theta')$ be the target LTF; as an initial pre-processing step, we argue (Theorem 40) that g is close to an LTF $f = \text{sign}(w \cdot x - \theta)$ in which all weights w_i are not-too-large integer multiples of some fixed "granularity" value. We then proceed with a case analysis based on the τ^*-critical index (for a suitable regularity parameter τ^*) of the LTF f. As with the Chow problem, the algorithm proceeds differently depending on whether the target LTF f has large, zero or small τ^*-critical index.

A significant challenge that arises in analyzing these cases for the Shapley problem is the fact that the probabilistic definition of the Shapley indices is much less "clean" than the definition of the Chow parameters. Recall that the i-th Chow parameter $\widehat{f}(i)$ is defined to equal $\mathbf{E}_{\boldsymbol{x} \sim \{-1,1\}^n}[f(\boldsymbol{x}) \cdot \boldsymbol{x}_i]$; the fact that the underlying distribution — uniform over $\{-1,1\}^n$ — is a *product* distribution makes this definition particularly amenable to harmonic analysis and the application of various tools from probability theory. The Shapley indices, on the other hand, do not admit such a clean definition in terms of a product distribution. However, in an attempt to get a syntactically similar definition, De et al. (2017) showed that the i-th Shapley index $\mathring{f}(i)$ is equal to $\alpha \cdot f^*(i) + \beta$, where α and β are fixed values (depending only on n and not on f) and $f^*(i) = \mathbf{E}_{\boldsymbol{x} \sim \mathcal{D}_{\text{Shap}}}[f(\boldsymbol{x}) \cdot \boldsymbol{x}_i]$, where $\mathcal{D}_{\text{Shap}}$ is a certain *symmetric* distribution supported on $\{-1,1\}^n$. Here "symmetric" means that the distribution $\mathcal{D}_{\text{Shap}}$ is invariant under permutation of coordinates (the probability that $\mathcal{D}_{\text{Shap}}$ assigns to a string depends only on the number of 1's in the string). A significant technical complication is that $\mathcal{D}_{\text{Shap}}$ is not a product distribution, and thus several technical tools that are used to analyze the Chow parameters, and that rely on the product distribution structure of the uniform distribution over $\{-1,1\}^n$, are no longer available.

In order to adapt our algorithm for the partial Chow parameters problem to the partial Shapley problem, the main technical statement that is required is that if $f = \text{sign}(w \cdot x - \theta)$ is such that w is τ-regular, then the Shapley indices of f are close to being an affine form of the weights. More precisely, we we want to prove that there are values $\mathring{A} = \mathring{A}(\theta, \|w\|_1)$ and $\mathring{B} = \mathring{B}(\theta, \|w\|_1)$ such that

$$\sum_{i=1}^{n} (\mathring{f}(i) - (\mathring{A} w_i + \mathring{B}))^2 \leq \varepsilon(\tau), \tag{2}$$

where $\varepsilon(\tau) \to 0$ as $\tau \to 0$. To prove this, we first show (Lemma 48) that the Shapley distribution $\mathcal{D}_{\text{Shap}}$ can be approximated by a convex combination of p-biased product distributions u_p^n on the

8

hypercube (here u_p^n is the product distribution in which each marginal $\boldsymbol{x}_i$ has $\mathbf{Pr}[\boldsymbol{x}_i = 1] = p$ and $\mathbf{Pr}[\boldsymbol{x}_i = -1] = 1 - p$). While the distribution $\mathcal{D}_{\text{Shap}}$ cannot be exactly expressed as a convex combination of p-biased distributions on the cube, we show that for any parameter $\delta > 0$, we can express $\mathring{f}(i)$ as a "positive linear combination" of $\big\{\big(f_p^*(i) - \frac{\sum_{j=1}^n f_p^*(j)}{n}\big)\big\}_{p \in [\delta, 1-\delta]}$ up to an error of at most $O(n^2\delta)$. Here $f_p^*(i) = \mathbf{E}[f(\boldsymbol{x}) \cdot \boldsymbol{x}_i]$ where $\boldsymbol{x} \sim u_p^n$, the p-biased distribution on the cube. As $\delta \to 0$, the error of approximating $\{\mathring{f}(i)\}$ goes to zero, but the "positive linear" coefficients of $f_p^*(i)$ (for small values of p) diverge to infinity, thus rendering the expression meaningless. We evade these difficulties by not allowing δ to be too close to 0; more precisely, we choose δ to be a particular $1/\text{poly}(n)$ value, which ensures that $\mathring{f}(i)$ can be expressed as a positive linear combination of $\big\{\big(f_p^*(i) - \frac{\sum_{j=1}^n f_p^*(j)}{n}\big)\big\}_{p \in [\delta, 1-\delta]}$ up to an error of $o(1)$.

Establishing Equation (2) now reduces to showing that there are values $\mathring{A}_p = \mathring{A}_p(\|w\|_1, \theta, p)$ and $\mathring{B}_p = \mathring{B}_p(\|w\|_1, \theta, p)$ such that

$$\sum_{i=1}^n \left(\left(f_p^*(i) - \frac{\sum_{j=1}^n f_p^*(j)}{n}\right) - (\mathring{A}_p w_i + \mathring{B}_p)\right)^2 = \varepsilon_p(\tau), \tag{3}$$

where $\varepsilon_p(\tau) \to 0$ as $\tau \to 0$. We note that when $p = 1/2$, this follows from our analysis for the Chow problem. We carry out a careful adaptation of the machinery developed in the context of LTF analysis for the uniform distribution ($p = 1/2$), including results from Matulef et al. (2010); De et al. (2016), to show that for any $p \in (0, 1)$, we have

$$\sum_{i=1}^n \left(\left(f_p^*(i) - \frac{\sum_{j=1}^n f_p^*(j)}{n}\right) - (\mathring{A}_p w_i + \mathring{B}_p)\right)^2 = O(\sqrt{\tau}).$$

This finishes the sketch of our high level approach for establishing Equation (2).

We now turn to giving an overview of the algorithmic part. As with the partial Chow parameters problem, we choose a suitable value $\tau^* = \tau^*(\varepsilon)$ of the regularity parameter (depending on the desired final accuracy ε), and the algorithmic strategy depends on whether the τ^*-critical index is zero, "large", or "small." As before, the case when the critical index is small is essentially a combination of the first two cases, so in the rest of this intuitive overview, we will just give the high level idea of the algorithmic strategy for the "large" and zero critical index cases.

1. **Case 1: τ^*-critical index is large:** Similar to the partial Chow parameters problem discussed earlier, we would like to argue that that for a suitable threshold $K = K(\tau^*)$, if the τ^*-critical index of a LTF f is larger than K then f is close to a LTF f' on K variables under $\mathcal{D}_{\text{Shap}}$. While the fact that $\mathcal{D}_{\text{Shap}}$ is not a product distribution presents some obstacles, we are able to leverage anti-concentration of certain linear forms under $\mathcal{D}_{\text{Shap}}$ (proved in De et al. (2017)) to argue that if the critical index is larger than essentially $K := O(\log n/(\tau^*)^2)$, then f is close (under $\mathcal{D}_{\text{Shap}}$) to a junta on K variables. Then, as in the partial Chow parameters problem, one can find a suitable LTF by just brute force search over all LTFs on K variables. Note that the threshold K has a $\log n$ dependence on n; this is in contrast with the partial Chow parameters problem, where the corresponding cutoff for "large" critical index is independent of n. This is a bottleneck that results in our algorithm for the partial Shapley problem running in quasipolynomial time (whereas for the partial Chow parameters problem the running time is polynomial in n).

2. **Case 2: τ^*-critical index is zero:** As stated at the beginning of this subsection, we can assume that all the weights in the target LTF are not-too-large integral multiples of some fixed granularity parameter γ^*, and thus we can also assume that the threshold θ is also an not-too-large integer multiple of γ^*. The algorithm guesses two parameters, namely θ and $W = \|w\|_1$; its analysis will exploit the fact that there are only polynomially many possibilities for these parameters. Given these parameters, the algorithm can exactly compute the constants $\mathring{A} = \mathring{A}(\theta, \|w\|_1)$ and $\mathring{B} = \mathring{B}(\theta, \|w\|_1)$ from Equation (2). Now let $S \subseteq [n]$ be the set of indices for which the algorithm is given Shapley indices. The algorithmic problem now reduces to finding a set of weights $w_1, \ldots, w_n$ to

$$\text{Minimize} \sum_{i \in S} (\mathring{\hat{f}}(i) - (\mathring{A}w_i + \mathring{B}))^2 \text{ subject to } \sum_{i=1}^{n} w_i = W; \ \sum_{i=1}^{n} w_i^2 = 1; \ \max_{1 \leq i \leq n} w_i \leq \tau^*;$$

These constraints are non-linear and non-convex, and thus not amenable to techniques from convex programming in any obvious way. However, we show that by exploiting the granularity of the weights w_i (recall that all of them are integral multiples of γ^*), it is possible to use a simple dynamic programming approach to solve this problem.

6. Organization

Due to space constraints, in the rest of the body of the paper we confine ourselves to stating our main structural result for the Partial Chow Parameters Problem, Theorem 7, and describing and analyzing our resulting algorithm. This proves Theorem 5, our main result for the Partial Chow Parameters problem, modulo the proof of Theorem 7. Our Shapley results, which are more technically and conceptually involved than our Chow results, can be found in full in the appendix. More detail on the organization of the paper follows.

In the Appendix, Section A gives basic preliminary definitions and results on LTFs, regularity, various notions of distance between functions that we will use, and various distributions that will arise in our analysis. Section B.1 gives background results from Gaussian analysis and p-biased Fourier analysis of LTFs, and Section B generalizes various technical results on Fourier analysis of regular LTFs under the uniform distribution from (Matulef et al., 2010; De et al., 2016) to the p-biased case. Section B.4 extends some of these results to the case of general LTFs by doing an analysis that works separately with the "head" portion and the (regular) "tail" portion of a general LTF. Section C builds on the $p = 1/2$ case of these structural results to prove Theorem 7. Section D uses the general-p version of these results (with many additional analytic and algorithmic ingredients, as discussed earlier) to prove Theorem 34, our main result for the Partial Shapley Indices problem.

7. Main Structural Result for Partial Chow and Proof of Theorem 5

Our main structural result for the Partial Chow Parameters Problem, proved in Section C, is as follows. The theorem defines a relatively small set of candidate LTFs and asserts that at least one of these must have Chow Parameters that are close to the input Chow Parameters. Our algorithm works by enumerating all of these candidate LTFs, and then checking for each one whether it has Chow Parameters that are close to the input Chow Parameters.

(The notation "$d_{\mathrm{Chow},S}(f,g)$" stands for the partial Chow distance between f and g with respect to S, defined as $d_{\mathrm{Chow},S}(f,g) := \left(\sum_{i \in S} \left(\hat{f}(i) - \hat{g}(i) \right)^2 \right)^{1/2}$.).

Theorem 7 *Let $\varepsilon \in (0, \frac{1}{2})$ and let $\tau = \tau(\varepsilon) = \varepsilon^{1000}$. Let f be an LTF $f(x) = f(x_H, x_T) = \mathrm{sign}(w_H \cdot x_H + w_T \cdot x_T - \theta)$ where H contains all indices $i \in [n]$ with $|\hat{f}(i)| \geq \tau^2$. Let $S \subseteq \{0, 1, \ldots, n\}$. Then one of the following holds:*

1. *$d_{\mathrm{Chow},S}(f,g) \leq O(\varepsilon)$ for some linear threshold function junta g over x_H, or*

2. *$d_{\mathrm{Chow},S}(f,g) \leq O(\varepsilon)$ for some linear threshold function g of the form*

$$g(x) = g(x_H, x_T) = \mathrm{sign}\Big(\sum_{i \in H} v_i \cdot x_i + (\gamma')^{-1} \cdot \Big(\sum_{i \in T \cap S} \hat{f}(i) \cdot x_i + \sum_{i \in T \cap \bar{S}} r \cdot x_i \Big) - \theta' \Big) \quad (4)$$

where $\theta' \in \mathbb{R}, v_H = (v_1, \ldots, v_{|H|}) \in \mathbb{R}^{|H|}$ are such that each v_i is an integer multiple of $\sqrt{\tau}/|H|$ and has magnitude at most $2^{O(|H| \log |H|)} \sqrt{\ln(1/\tau)}$, where γ' satisfies $\gamma' = \gamma$ if $T \subseteq S$ and $\gamma \leq \gamma' \leq \gamma + \tau$ if $T \not\subseteq S$ for $\gamma := \left(\sum_{i \in T} \hat{f}(i)^2 \right)^{1/2}$, and where $r := \left(\frac{(\gamma')^2 - \sum_{i \in T \cap S} \hat{f}(i)^2}{|T \cap \bar{S}|} \right)^{1/2}$.

7.1. Main algorithm for the Partial Chow Parameters Problem

We next present the main algorithm of this section, which is an EPRAS for solving the Partial Inverse Chow Parameters Problem, and which works by leveraging the structural result Theorem 7.

Theorem 8 (Detailed version of Theorem 5) *There exists an algorithm for the Partial Inverse Chow Parameters Problem with the following guarantees. It takes as input four things: (1) a set $\{(i, \hat{f}(i)) : i \in S\}$ for some LTF $f : \{-1, 1\}^n \to \{-1, 1\}$ and some $S \subseteq \{0, 1, \ldots, n\}$, (2) the length n of the input to f, (3) an error parameter $\varepsilon \in (0, \frac{1}{2})$, and (4) a confidence parameter $\delta > 0$. It outputs a weights-based representation of an LTF $g : \{-1, 1\}^n \to \{-1, 1\}$ such that $d_{\mathrm{Chow},S}(f,g) \leq O(\varepsilon)$ with probability $1 - \delta$, and runs in time $n^2 \log n \cdot \log(1/\delta) \cdot 2^{\mathrm{poly}(1/\varepsilon)}$.*

The algorithm consists of three steps: a parameter guessing step, an LTF enumeration step, and an LTF verification step. The second two steps are similar to those in the main algorithm in O'Donnell and Servedio (2011). The two cases in the enumeration step correspond to the two cases in Theorem 7. The algorithm is as follows.

1. Guess the size of the head $|H|$ and a value γ' satisfying the conditions in Theorem 7.

   (a) Compute $H \cap S := \{i \in S : |\hat{f}(i)| \geq \tau^2\}$, $T \cap S := \{i \in S : |\hat{f}(i)| < \tau^2\}$, $|T| := n - |H|$, r from these guesses.

   (b) Set H equal to the union of $H \cap S$ and $|H| - |H \cap S|$ arbitrary indices not in S.

2. For each guess of H and γ' in Step 1, enumerate candidate LTFs using the two cases in Theorem 7:

   (a) Enumerate all junta LTFs g over x_H.

   (b) Enumerate all LTFs g of the form given in Equation (16).

3. For each candidate LTF g generated in Step 2, compute an empirical estimate $\bar{g}(i)$ of each of the Chow Parameters $\hat{g}(i)$ for $i \in S$ so that $|\hat{g}(i) - \bar{g}(i)| \leq \varepsilon/\sqrt{|S|}$ with confidence $1 - \delta/(|S| \cdot M)$, where M is the total number of LTFs enumerated in Step 2. Output (a weights-based representation of) the first g such that $\|(\hat{f}(i))_{i \in S} - (\bar{g}(i))_{i \in S}\| \leq O(\varepsilon)$.

Proof [Proof of Theorem 8] We start by arguing that the above algorithm is correct. By taking a union bound, it holds that all $|S| \cdot M$ estimates $\bar{g}(i)$ of the Chow Parameters $\hat{g}(i)$ of candidate LTFs g with $i \in S$ in Step 3 will be accurate to within a $\varepsilon/\sqrt{|S|}$ additive error factor with probability at least $1 - \delta$. In this case our estimates will all satisfy $\|(\hat{g}(i))_{i \in S} - (\bar{g}(i))_{i \in S}\| \leq \varepsilon$, and hence by the triangle inequality $\|(\hat{f}(i))_{i \in S} - (\bar{g}(i))_{i \in S}\| - \varepsilon \leq d_{\text{Chow},S}(f, g) \leq \|(\hat{f}(i))_{i \in S} - (\bar{g}(i))_{i \in S}\| + \varepsilon$ for every candidate LTF g. So, in this case, we will output a candidate LTF g if and only if it satisfies $d_{\text{Chow},S}(f, g) \leq O(\varepsilon)$. Furthermore, by Theorem 7, one of the candidate LTFs g enumerated in Step 2 will satisfy $d_{\text{Chow},S}(f, g) \leq O(\varepsilon)$, and so with probability at least $1 - \delta$ we will output such a function.

We turn to analyzing the runtime of the algorithm. We start by analyzing the number of guesses that we need for $|H|$ and γ' in Step 1. Each $\hat{f}(i)$ with $i \in H$ satisfies $|\hat{f}(i)| \geq \tau^2$, and because $\sum_{i \in [n]} \hat{f}(i)^2 \leq 1$ this implies that $|H| \leq 1/\tau^4$. Because $0 \leq \gamma \leq 1$, setting γ' to be either 1 or one of the $O(1/\tau)$ integer multiples of τ between 0 and 1 will satisfy the condition $\gamma \leq \gamma' \leq \gamma + \tau$. So, we need $O(1/\tau)$ guesses for γ'. Computing all other quantities given the guesses of $|H|$ and γ' is efficient.

We next upper bound the number M of functions enumerated in Step 3. By Muroga et al. (1961), any junta LTF on $|H|$ variables can be represented using integer weights of magnitude at most $2^{O(|H| \log |H|)}$, meaning that there are at most $2^{O(|H|^2 \log |H|)} = 2^{O(1/\tau^8 \cdot \log(1/\tau))}$ such functions total (where we have used the fact that $|H| \leq 1/\tau^4$).

We next consider functions of the form specified in Equation (16). For fixed H and γ' each such function is uniquely specified by a threshold θ' and head weights v_H, and so the total number of such functions is equal to the total number of possibilities for θ', v_H. Each of the $|H| + 1$ weights θ' and v_i for $i \in H$ is an integer multiple of $\sqrt{\tau}/|H|$ and has magnitude at most $2^{O(|H| \log |H|)} \sqrt{\ln(1/\tau)}$. Therefore, the total number of such functions is upper bounded by $(2^{O(|H| \log^2 |H|)} \sqrt{\ln(1/\tau)/\tau})^{|H|+1} \leq 2^{O(1/\tau^8 \cdot \log^2(1/\tau))}$, where we have again used the fact that $|H| \leq 1/\tau^4$. Combining the upper bounds on the number of juntas on $|H|$ variables and on the number of functions of the form in Equation (16) we get that $M \leq 2^{O(1/\tau^8 \cdot \log^2(1/\tau))}$.

Finally, we upper bound the amount of time necessary to obtain estimates $\bar{g}(i)$ of the Chow Parameters $\hat{g}(i)$ with the desired error and confidence. The following standard Chernoff bound holds for ± 1-valued, i.i.d. Bernoulli random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ each with mean μ:

$$\mathbf{Pr}\left[\left|\left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i\right) - \mu\right| \geq \Delta\right] \leq 2\exp(-\Delta^2 N/2) . \tag{5}$$

Therefore, using $N := O\left(\frac{|S|}{\varepsilon^2} \log\left(\frac{|S|\cdot M}{\delta}\right)\right)$ uniformly random samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \sim \{-1,1\}^n$, the estimator $\bar{g}(i) := \sum_{j=1}^{N} g(\boldsymbol{x}_j) \cdot (\boldsymbol{x}_j)_i$ approximates $\mu := \hat{g}(i)$ to within $\Delta := \varepsilon/\sqrt{|S|}$ additive error with confidence $1 - \delta/(|S| \cdot M)$.

Computing each estimator $\bar{g}(i)$ requires N evaluations of g and uses $O(N)$ additional arithmetic operations. Each function evaluation uses $O(n)$ arithmetic operations, for a total of $O(n \cdot N)$ arithmetic operations. We must compute estimators $\bar{g}(i)$ for $|S| \cdot M$ many Chow Parameters $\hat{g}(i)$, so in total Step 3 requires

$$O(nN\,|S|\,M) = O\left(\frac{n \cdot |S|^2}{\varepsilon^2} \log\left(\frac{|S| \cdot M}{\delta}\right)\right)M\right) \tag{6}$$

time, which also subsumes the amount of time it takes to enumerate all M functions in Step 2.

We conclude by upper bounding the overall runtime of the algorithm by the right-hand side of Equation (6) times the number of guesses we need to make for H and γ' in Step 1:

$$O\left((1/\tau^5) \cdot \frac{n \cdot |S|^2}{\varepsilon^2} \log\left(\frac{|S| \cdot M}{\delta}\right)\right)M\right)$$
$$= \frac{|S|^2}{\varepsilon^2} \cdot \left(\log\frac{|S|}{\delta} + (1/\tau^8 \cdot \log^2(1/\tau))\right)\right) \cdot 2^{O(1/\tau^8 \cdot \log^7(1/\tau))}$$
$$\leq n^2 \log n \cdot \log(1/\delta) \cdot 2^{\mathrm{poly}(1/\varepsilon)} ,$$

where we have used the fact that $\tau = \mathrm{poly}(\varepsilon)$. ∎

Acknowledgments

References

Yoram Bachrach, Evangelos Markakis, Ezra Resnick, Ariel D Procaccia, Jeffrey S Rosenschein, and Amin Saberi. Approximating power indices: theoretical and empirical analysis. *Autonomous Agents and Multi-Agent Systems*, 20(2):105–122, 2010.

John F. Banzhaf III. Weighted voting doesn't work: A mathematical analysis. *Rutgers L. Rev.*, 19: 317, 1964.

R. Bhattacharya and R. Rao. *Normal approximation and asymptotic expansions*. Robert E. Krieger Publishing Company, 1986.

Chao-Kong Chow. On the characterization of threshold functions. In *Switching Circuit Theory and Logical Design, 1961. SWCT 1961. Proceedings of the Second Annual Symposium on*, pages 34–38. IEEE, 1961.

Anindya De, Ilias Diakonikolas, Vitaly Feldman, and Rocco A Servedio. Nearly optimal solutions for the Chow Parameters Problem and low-weight approximation of halfspaces. *Journal of the ACM (JACM)*, 61(2):11, 2014.

Anindya De, Ilias Diakonikolas, and Rocco A. Servedio. A robust Khintchine inequality, and algorithms for computing optimal constants in Fourier analysis and high-dimensional geometry. *SIAM J. Discrete Math.*, 30(2):1058–1094, 2016.

Anindya De, Ilias Diakonikolas, and Rocco A. Servedio. The inverse Shapley value problem. *Games and Economic Behavior*, 105:122–147, 2017. doi: 10.1016/j.geb.2017.06.004. URL https://doi.org/10.1016/j.geb.2017.06.004.

Bart de Keijzer. A survey on the computation of power indices. *Delft University of Technology*, 2008.

John Deegan and Edward W. Packel. A new index of power for simple n-person games. *International Journal of Game Theory*, 7(2):113–123, 1978.

Xiaotie Deng and Christos H. Papadimitriou. On the complexity of cooperative solution concepts. *Mathematics of Operations Research*, 19(2):257–266, 1994.

Ilias Diakonikolas and Chrystalla Pavlou. On the complexity of the inverse semivalue problem for weighted voting games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1869–1876, 2019.

Ilias Diakonikolas and Rocco A. Servedio. Improved approximation of linear threshold functions. *Computational Complexity*, 22(3):623–677, 2013.

Ilias Diakonikolas, Parikshit Gopalan, Rajesh Jaiswal, Rocco A. Servedio, and Emanuele Viola. Bounded independence fools halfspaces. *SIAM Journal on Computing*, 39(8):3441–3462, 2010.

European-Union. Treaty of Rome, 1957.

William Feller. *An Introduction to Probability Theory and its Applications*. John Wiley & Sons, 1968.

Paul W Goldberg. A bound on the precision required to estimate a boolean perceptron from its average satisfying assignment. *SIAM Journal on Discrete Mathematics*, 20(2):328–343, 2006.

Manfred J Holler. Forming coalitions and measuring voting power. *Political studies*, 30(2):262–271, 1982.

Ronald John Johnston. On the measurement of power: Some reactions to Laver. *Environment and Planning A*, 10(8):907–914, 1978.

S. Khot, G. Kindler, E. Mossel, and R. O'Donnell. Optimal inapproximability results for Max-Cut and other 2-variable CSPs? *SIAM Journal on Computing*, 37(1):319–357, 2007.

D. Leech. Computing power indices for large voting games. *Management Science*, 49(6), 2003.

Dennis Leech. Designing the voting system for the Council of the European Union. *Public Choice*, 113(3-4):437–464, 2002.

Kevin Matulef, Ryan O'Donnell, Ronitt Rubinfeld, and Rocco A. Servedio. Testing halfspaces. *SIAM J. on Comput.*, 39(5):2004–2047, 2010.

Elchannan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: Invariance and optimality. *Annals of Mathematics*, 171:295–341, 2010.

Saburo Muroga, Iwao Toda, and Satoru Takasu. Theory of majority switching elements. *J. Franklin Institute*, 271(5):376–418, 1961.

Ryan O'Donnell and Rocco A Servedio. The Chow Parameters Problem. *SIAM Journal on Computing*, 40(1):165–199, 2011.

Rocco A. Servedio. Every linear threshold function has a low-weight approximator. *Comput. Complexity*, 16(2):180–209, 2007.

Lloyd S Shapley and Martin Shubik. A method for evaluating the distribution of power in a committee system. *American Political Science Review*, 48(3):787–792, 1954.

Contents

Appendix A. Background

A.1. Linear threshold functions, regularity, and critical index

Notation	Definition	Description
$\hat{f}(i)$	$\mathbf{E}_{\boldsymbol{x} \sim \{-1,1\}^n}[f(\boldsymbol{x}) \cdot \boldsymbol{x}_i]$	The ith Fourier coefficient/Chow parameter of f.
$\hat{f}(i, p)$	$\mathbf{E}_{\boldsymbol{x} \sim u_p^n}[f(\boldsymbol{x}) \cdot \psi_p(\boldsymbol{x}_i)]$	The ith p-biased Fourier coefficient of f.
$f^*(i, p)$	$\mathbf{E}_{\boldsymbol{x} \sim u_p^n}[f(\boldsymbol{x}) \cdot \boldsymbol{x}_i]$	The ith p-biased coordinate correlation coefficient of f.
$\tilde{f}(i)$	$\mathbf{E}_{\boldsymbol{x} \sim N(0,1)^n}[f(\boldsymbol{x}) \cdot \boldsymbol{x}_i]$	The ith Hermite coefficient of f.
$\mathring{f}(i)$	$\mathbf{E}_{\boldsymbol{\pi} \sim \mathbb{S}_n}[f(x^+(\boldsymbol{\pi}, i)) - f(x(\boldsymbol{\pi}, i))]$	The ith Shapley index (value) of f.
$\overset{\triangle}{f}(i)$	$\mathbf{E}_{\boldsymbol{x} \sim \mathcal{D}_{\mathrm{Shap}}}[f(\boldsymbol{x}) \cdot L_i(\boldsymbol{x})]$	The ith Shapley Fourier coefficient of f.
$f^*(i)$	$\mathbf{E}_{\boldsymbol{x} \sim \mathcal{D}_{\mathrm{Shap}}}[f(\boldsymbol{x}) \cdot \boldsymbol{x}_i]$	The ith Shapley coordinate correlation coefficient of f.

Table 1: Quantities associated with a linear threshold function $f : \{-1, 1\}^n \to \{-1, 1\}$ and index $i \in [n]$.

We recall that a linear threshold function (LTF) is a function $f : \{-1, 1\}^n \to \{-1, 1\}$ defined by $f(x) = \mathrm{sign}(w \cdot x - \theta)$ for some $w \in \mathbb{R}^n, \theta \in \mathbb{R}$ where $\mathrm{sign}(t) = 1$ iff $t \geq 0$. We say that a nonzero vector $w \in \mathbb{R}^n$ is τ-*regular* if $\|w\|_\infty / \|w\|_2 \leq \tau$, and we say that an LTF $f(x) = \mathrm{sign}(w \cdot x - \theta)$ is τ-regular if its weight vector w is τ-regular.

A key ingredient in our proofs is the notion of the *critical index* of an LTF. The critical index was implicitly introduced and used in Servedio (2007) and was explicitly used in Diakonikolas and Servedio (2013); Diakonikolas et al. (2010); O'Donnell and Servedio (2011) and other works. Intuitively, the critical index of w is the first index i such that the sub-vector of w obtained by deleting the i largest-magnitude entries of w is regular. A precise definition follows:

Definition 9 (critical index) *Given a vector* $w \in \mathbb{R}^n$ *such that* $|w_1| \geq \ldots \geq |w_n| > 0$, *for* $k \in [n]$ *we denote by* σ_k *the quantity* $\sqrt{\sum_{i=k}^n w_i^2}$. *We define the* τ-critical index $c(w, \tau)$ *of* w *as the smallest index* $i \in [n]$ *for which* $|w_i| \leq \tau\sigma_i$. *If this inequality does not hold for any* $i \in [n]$, *we define* $c(w, \tau) = \infty$.

Finally, we will use the following lemma, which appears in a number of previous works. The result says that, for weight vectors w with sorted weights, $\sigma_k = \sqrt{\sum_{i=k}^n w_i^2}$, also denoted as $\mathrm{tail}_k(w)$, decreases geometrically for i less than the critical index.

Fact 10 (Fact 25 De et al. (2014))) *Let* $w = (w_1, \ldots, w_n) \in \mathbb{R}^n$ *be such that* $|w_1| \geq \cdots \geq |w_n|$, *and let* $1 \leq a \leq b \leq c(w, \tau)$, *where* $c(w, \tau)$ *is the* τ-critical index of w. *Then* $\mathrm{tail}_b(w) < (1 - \tau^2)^{(b-a)/2} \cdot \mathrm{tail}_a(w)$.

Proof By definition of the critical index, $|w_i| > \tau \cdot \mathrm{tail}_i(w)$ for $i < c(w, \tau)$. Therefore for such an i, $\mathrm{tail}_i(w)^2 = w_i^2 + \mathrm{tail}_{i+1}(w)^2 > \tau^2 \cdot \mathrm{tail}_i(w)^2 + \mathrm{tail}_{i+1}(w)^2$, and so $\mathrm{tail}_{i+1}(w) < (1 - \tau^2)^{1/2} \cdot \mathrm{tail}_i(w)$. The result follows by applying this last inequality repeatedly. ∎

A.2. Boolean functions and distance measures

We assume familiarity with the basics of standard Fourier analysis of Boolean functions with respect to the uniform distribution over $\{-1, 1\}^n$, see Appendix E for a brief overview. (Later in this preliminaries section we will introduce notions of Fourier analysis with respect to other distributions such as product distributions and the "Shapley distribution.")

We will use a range of different notions of distance between Boolean functions $f, g : \{-1, 1\}^n \to \{-1, 1\}$. Let

$$d(f, g) := \Pr_{\boldsymbol{x} \sim \{-1,1\}^n}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})]$$

denote the Hamming distance between f and g. Let

$$d_{\mathrm{Chow}}(f, g) := \left(\sum_{i=1}^n \left(\hat{f}(i) - \hat{g}(i) \right)^2 \right)^{1/2}$$

denote the *Chow distance* between f and g, and let

$$d_{\mathrm{Chow},S}(f, g) := \left(\sum_{i \in S} \left(\hat{f}(i) - \hat{g}(i) \right)^2 \right)^{1/2}$$

denote the *partial Chow distance* between f and g with respect to a subset of indices $S \subseteq \{0, 1, \ldots, n\}$. We similarly define

$$d_{\mathrm{Shapley}}(f, g) := \left(\sum_{i=1}^n \left(\mathring{f}(i) - \mathring{g}(i) \right)^2 \right)^{1/2},$$

the *Shapley distance* between f and g, and

$$d_{\mathrm{Shapley},S}(f, g) := \left(\sum_{i \in S} \left(\mathring{f}(i) - \mathring{g}(i) \right)^2 \right)^{1/2},$$

the *partial Shapley distance* between f and g,

It is clear that $d_{\mathrm{Chow},S}(f, g) \leq d_{\mathrm{Chow}}(f, g)$ and $d_{\mathrm{Shapley},S}(f, g) \leq d_{\mathrm{Shapley}}(f, g)$ for any $S \subseteq \{0, 1, \ldots, n\}$. The following simple result relates Hamming distance and Chow distance:

Proposition 11 ((O'Donnell and Servedio, 2011, Proposition 1.5)) $d_{\mathrm{Chow}}(f, g) \leq 2\sqrt{d(f, g)}$.

Proof For $f, g : \{-1, 1\}^n \to \{-1, 1\}$ we have $d(f, g) = \frac{1}{4} \mathbf{E}[(f(\boldsymbol{x}) - g(\boldsymbol{x}))^2] = \frac{1}{4} \sum_{S \subseteq [n]} (\hat{f}(S) - \hat{g}(S))^2 \geq \frac{1}{4} \sum_{i \in [n]} (\hat{f}(i) - \hat{g}(i))^2 = \frac{1}{4} d_{\mathrm{Chow}}(f, g)^2$, and hence $d_{\mathrm{Chow}}(f, g) \leq 2\sqrt{d(f, g)}$. ∎

A.3. Some useful distributions

A.3.1. THE SHAPLEY DISTRIBUTION $\mathcal{D}_{\mathrm{Shap}}$ AND "FOURIER ANALYSIS" FOR THIS DISTRIBUTION

De et al. (2017) introduced a distribution over $\{-1, 1\}^n$, called the "Shapley distribution" (we write $\mathcal{D}_{\mathrm{Shap}}$ for this distribution though it is denoted by μ in De et al. (2017)), which is very useful for analysis of the Shapley indices. We recall the definition of this distribution: let $Q(n, k) := 1/k + 1/(n-k)$ for $0 < k < n$, and let $\Lambda(n) := \sum_{0<k<n} Q(n, k) = 2H_{n-1}$, where H_n denotes the nth harmonic number. The distribution $\mathcal{D}_{\mathrm{Shap}}$ over $\{-1, 1\}^n$ is defined as follows: it has support $\{-1, 1\}^n \setminus \{(-1)^n, 1^n\}$. To sample a string $\boldsymbol{x} \sim \mathcal{D}_{\mathrm{Shap}}$, first sample $k \in \{1, \ldots, n-1\}$ with probability $Q(n, k)/\Lambda(n)$. Then choose $\boldsymbol{x}$ uniformly from the weight k slice of the hypercube $\{-1, 1\}^n$ (i.e. the set of all $\binom{n}{k}$ many strings in $\{-1, 1\}^n$ with exactly k many 1's).

Following De et al. (2017), we proceed to define a "Fourier basis" under the distribution μ. We define the inner product $\langle f, g \rangle_\mu := \mathbf{E}_{\boldsymbol{x} \sim \mathcal{D}_{\mathrm{Shap}}}[f(\boldsymbol{x})g(\boldsymbol{x})]$, and we define orthonormal functions $L_i : \{-1, 1\}^n \to \mathbb{R}$ for $i = 0, 1, \ldots, n$ so that $\langle L_i, L_j \rangle_\mu = 1$ if $i = j$ and $\langle L_i, L_j \rangle_\mu = 0$ if $i \neq j$. As shown in (De et al., 2017, Lemma 9), we can take $L_0(x) \equiv 1$ and $L_i(x) = a \cdot (\sum_{j=1}^n x_i) + b x_i$ for some values of $a = a(n)$ and $b = b(n)$ satisfying $a = -\Theta(\sqrt{\log n}/n)$ and $b = \Theta(\sqrt{\log n})$. Accordingly, we define *Shapley Fourier coefficients* and *Shapley Fourier distance* with respect to μ as follows. The ith Shapley Fourier coefficient for $i = 0, 1, \ldots, n$ is defined as

$$\overset{\triangle}{f}(i) := \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathrm{Shap}}} [f(\boldsymbol{x}) \cdot L_i(\boldsymbol{x})],$$

and the Shapley Fourier distance between two LTFs f and g is defined as

$$d_{\mathrm{Shapley-Fourier}}(f, g) := \Big(\sum_{i=0}^n (\overset{\triangle}{f}(i) - \overset{\triangle}{g}(i))^2 \Big)^{1/2}. \tag{7}$$

A.3.2. p-BIASED DISTRIBUTIONS AND FOURIER ANALYSIS

We write u_p to denote the p-biased distribution over $\{-1, 1\}$, i.e. a random variable distributed according to u_p takes the value $+1$ with probability p and takes the value -1 with probability $1 - p$. Let

$$\mu_p := 2p - 1 \qquad \text{and} \qquad \sigma_p := 2\sqrt{p(1-p)}$$

denote the mean and standard deviation respectively of such a random variable. We define $\psi_p : \{-1, 1\} \to \mathbb{R}$,

$$\psi_p(x) := \frac{x - \mu_p}{\sigma_p},$$

so if $x \sim u_p$ is a p-biased random variable then $\psi_p(\boldsymbol{x})$ has mean 0 and variance 1. We will overload the above notation, defining $\psi_p^{[w]} : \{-1, 1\} \to \mathbb{R}$,

$$\psi_p^{[w]}(x) := \frac{x - \mu_p \cdot \sum_i w_i}{\sigma_p \|w\|_2} \quad \text{for } w \in \mathbb{R}^n,$$

which gives that

$$\psi_p^{[w]}(x) = \frac{x - \mu_p \cdot \|w\|_1}{\sigma_p \|w\|_2} \quad \text{for } w \in \mathbb{R}^n_{\geq 0}.$$

Fact 12 (Scaling Property) *We have that* $\psi_p^{[w]}(x) = \psi_p^{[\frac{w}{\sigma_p \|w\|_2}]}(\frac{x}{\sigma_p \|w\|_2})$.

The 2^n functions $\{\lambda_{S,p}(x) := \prod_{i \in S} \psi_p(x_i)\}_{S \subseteq [n]}$ are easily seen to constitute an orthonormal basis for the vector space of all real-valued functions on $\{-1, 1\}^n$ under the distribution u_p^n. We write $\hat{f}(S, p)$ to denote the corresponding p-biased Fourier coefficients of a real-valued function f under u_p^n, and we write the p-biased degree-1 coefficient as $\hat{f}(i, p)$ rather than $\hat{f}(\{i\}, p)$. When $p = 1/2$ and we are working with the uniform distribution, we simply write $\hat{f}(S)$ or $\hat{f}(i)$.

A.3.3. GAUSSIAN DISTRIBUTIONS AND HERMITE ANALYSIS

Let $N(\mu, \sigma^2)$ denote the Gaussian distribution with mean μ and variance σ^2. We recall that the n-variable *Hermite polynomials* $\{H_S\}_{S \in \mathbb{N}^n}$ form a complete orthonormal basis for the vector space of all square-integrable functions under the standard n-dimensional Gaussian distribution $N(0, 1)^n$. We write $\tilde{f}(S)$ to denote the S-th Hermite coefficient of a real-valued function f under $N(0, 1)^n$, and we will be particularly interested in f's degree-1 coefficients, i.e., $\tilde{f}(e_i)$, where e_i is the vector which is 1 in the i-th coordinate and 0 elsewhere. See Appendix E for a brief overview of the key notions.

A.4. Miscellaneous notation, terminology, and inequalities

We recall that a function $f : \{-1, 1\}^n \to \{-1, 1\}$ is said to be a *junta* on $J \subseteq [n]$ if f only depends on the coordinates in J. If $|J| = k$ we say that f is a *k-junta*.

Following De et al. (2017), we say that an LTF $f : \{-1, 1\}^n \to \{-1, 1\}$, $f(x) = \text{sign}(w \cdot x - w_0)$ with $w \in \mathbb{R}^n$ is *η-restricted* if $w_0 \in [-(1 - \eta)\|w\|_1, (1 - \eta)\|w\|_1]$. When η is small (as it will be in our Shapley result) this is a mild technical condition on the LTF f (which was also present in De et al. (2017)).

We write "$a \stackrel{k}{\approx} b$" to indicate that $|a - b| \le O(k)$. For $v \in \mathbb{R}^n$ we write "$\|v\|$" to denote the 2-norm $(v_1^2 + \cdots + v_n^2)^{1/2}$.

At various point in our analysis we will need some useful but routine inequalities; we record these in Appendix F.

Appendix B. Useful Fourier analytic results on p-biased Chow parameters of LTFs

B.1. Preliminary results from Gaussian analysis and p-biased Fourier analysis

B.1.1. BACKGROUND ON LTFS AND LINEAR FORMS UNDER THE GAUSSIAN DISTRIBUTION

Let ϕ denote the p.d.f. of a standard normal Gaussian $N(0, 1)$ and let Φ denote the corresponding c.d.f.. We extend the latter notation by writing $\Phi[a, b]$ to denote $\Phi(a) - \Phi(b)$, allowing $b < a$, and we will use the estimate $|\Phi[a, b]| \le |b - a|$ without comment.

Following Matulef et al. (2010), let us define the function $m : [-\infty, \infty] \to [-1, 1]$ by

$$m(\theta) := \left(2 \int_\theta^\infty \phi(x)dx\right) - 1 \tag{8}$$

and the function $W : [-1, 1] \to [0, 2/\pi]$ by

$$W(\nu) = (2\phi(m^{-1}(\nu)))^2 \tag{9}$$

(the latter is well defined since the function m is monotone decreasing with range $[-1, 1]$; we remark that W is a function symmetric about 0, with a peak at $W(0) = 2/\pi$). To motivate these definitions, we observe that $m(\theta)$ corresponds to the expectation $\mathbf{E}_{x \sim N(0,1)}[h_\theta(x)]$ of the univariate function $h_\theta(x) = \text{sign}(x - \theta)$. It is easily verified that

$$\tilde{h}_\theta(1) = \mathbf{E}[h_\theta(x)x] = 2\phi(\theta) \qquad \text{and} \qquad W(\mathbf{E}[h_\theta]) = \tilde{h}_\theta(1)^2. \tag{10}$$

The intuition is that given as input the expected value of some h_θ, the function W outputs the squared degree-1 Hermite coefficient of h_θ. This motivates the following definition, which will be useful later:

Definition 13 *Let* $\alpha(\theta) := \sqrt{W(m(\theta))}$. *View* $[n]$ *as partitioned into* $[n] = H \sqcup T$. *For* $p \in (0, 1)$ *and* $w = (w_H, w_T) \in \mathbb{R}^n$, *let*

$$\alpha(\theta, w_H, w_T, p) := \mathbf{E}_{\rho \sim u_p^{|H|}}[\alpha(\psi_p^{[w_T]}(\theta - w_H \cdot \rho))]. \tag{11}$$

Combining the above observations with the rotational invariance of $N(0, 1)^n$, it is straightforward to establish the following (see Proposition 25 of Matulef et al. (2010) for a proof):

Fact 14 (Hermite Properties of LTFs) *Let* $f : \mathbb{R}^n \to \{-1, 1\}$ *be an LTF* $f(x) = \text{sign}(w \cdot x - \theta)$, *where* $w \in \mathbb{R}^n$ *has* $\|w\| = 1$. *Then the degree-0 and degree-1 Hermite coefficients of* f *satisfy the following properties:*

1. *$\tilde{f}(0) = \mathbf{E}_{x \sim N(0,1)^n}[f(x)] = m(\theta)$;*

2. *$\tilde{f}(e_i) = \sqrt{W(\mathbf{E}_{x \sim N(0,1)^n}[f(x)])}w_i = \sqrt{W(m(\theta))}w_i$;*

3. *$\sum_{i=1}^n \tilde{f}(e_i)^2 = W(\mathbf{E}_{x \sim N(0,1)^n}[f(x)])$.*

We further recall the following useful properties of the functions m and W (see Proposition 24 of Matulef et al. (2010) for the simple proof):

Proposition 15

1. *$\mathbf{E}_{x \sim N(0,1)}[|x - \theta|] = 2\phi(\theta) - \theta m(\theta)$;*

2. *$|m'| \leq \sqrt{2/\pi}$ everywhere and $|W'| < 1$ everywhere;*

3. *If $|\nu| = 1 - \eta$ then $W(\nu) = \Theta(\eta^2 \log(1/\eta))$.*

B.1.2. GAUSSIAN VERSUS p-BIASED LINEAR FORMS

The main reason why the Gaussian distribution is useful for us is because if w is a regular linear form, then the distribution of $w \cdot x$ (when x is uniform random over $\{-1, 1\}^n$ or is drawn from the p-biased distribution u_p^n) can be well approximated in c.d.f. distance by a suitable Gaussian. This is a consequence of the well-known Berry-Esseen theorem, which gives quantitative error bounds on the central limit theorem; in this subsection we state this fundamental result along with a range of consequences and extensions of it which we will use.

Theorem 16 (Berry-Esseen Theorem, Feller (1968)) *Let $X_1, \ldots, X_n$ be independent real-valued random variables with $\mathbf{E}[X_i] = 0$, $\mathbf{E}[X_i^2] = \sigma_i^2 > 0$ and $\rho_i = \mathbf{E}[|X_i|^3] < \infty$ for each $i \in [n]$. Let $\sigma = \left(\sum_{i=1}^{n} \sigma_i^2\right)^{1/2}$ and let $\rho = \sum_{i=1}^{n} \rho_i$. Let $F(x)$ denote the cumulative distribution function of $\sigma^{-1} \cdot \sum_{i=1}^{n} X_i$. Then for all $t \in \mathbb{R}$, it holds that $|F(t) - \Phi(t)| \leq \frac{\rho}{\sigma^3}$, or in more detail,*

$$|F(x) - \Phi(x)| \leq C \cdot \frac{\rho}{\sigma^3} \cdot \frac{1}{1 + |x|^3}$$

for all real x, where C is an absolute constant.

The following is a fairly straightforward consequence of the Berry-Esseen Theorem, and is essentially a p-biased version of (De et al., 2016, Fact 2.6); it says that the value of a regular linear form with input sampled from u_p^n is distributed like a Gaussian up to some small error. For completeness we give the proof in Appendix G.

Fact 17 *Let $0^n \neq w \in \mathbb{R}^n$ be τ-regular, and let $p \in (0, 1)$. The we have the following:*

1. For any interval $[a, b] \subseteq \mathbb{R} \cup \{\pm\infty\}$,

$$\left| \Pr_{x \sim u_p^n} [w \cdot x \in [a, b]] - \left(\Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right) \right) \right| \leq \frac{4\tau}{\sigma_p},$$

where $\mu = \mu_p \cdot \sum_{i=1}^{n} w_i$ and $\sigma = \sigma_p \cdot \|w\|_2$.

2. For any λ and any $\theta \in \mathbb{R}$, we have

$$\Pr_{x \sim u_p^n} [|w \cdot x - \theta| \leq \lambda] \leq 2\frac{\lambda}{\sigma_p \|w\|_2} + 2\frac{\tau}{\sigma_p}.$$

In particular, if $\lambda = O(\tau)$ and $\|w\|_2 = 1$, then we have

$$\Pr[|\sum_i w_i x_i - \theta| \leq \lambda] \leq \frac{O(\tau)}{\sigma_p}.$$

As a p-biased analogue of the (simple) Proposition 31 of Matulef et al. (2010), we note that the Berry-Esseen theorem lets us easily approximate the expected value of a regular LTF under the p-biased distribution:

Lemma 18 *For $f(x) = \text{sign}(w \cdot x - \theta)$ a τ-regular LTF, we have $\mathbf{E}_{x \sim u_p^n}[f(x)] \overset{\frac{\tau}{\sigma_p}}{\approx} m(\psi_p^{[w]}(\theta))$.*

We also have a p-biased analogue of the (more involved) Proposition 32 of Matulef et al. (2010), which gives an approximation for the expected magnitude of the linear form $w \cdot x - \theta$ itself under the p-biased distribution (see Appendix G for the proof):

Lemma 19 *For w a τ-regular LTF, we have*

$$\mathbf{E}_{x \sim u_p^n}[|w \cdot x - \theta|] \overset{\tau \|w\|_2}{\approx} \|w\|_2 \sigma_p \mathbf{E}_{x \sim N(0,1)}\left[|x - \psi_p^{[w]}(\theta)|\right] = \|w\|_2 \sigma_p \left(2\phi(\psi_p^{[w]}(\theta)) - \psi_p^{[w]}(\theta) m(\psi_p^{[w]}(\theta))\right).$$

Bivariate statements. For technical reasons we will also require a two-dimensional analogue of Fact 17. The proof, which uses a multivariate extension of the Berry-Esseen theorem, is sketched in Appendix G and is a p-biased generalization of Theorem 68 of Matulef et al. (2010).

Fact 20 *Let $\boldsymbol{x} \sim u_p^n$ be a p-biased random vector in $\{-1, 1\}^n$, and let $\boldsymbol{y}$ be a random vector in $\{-1, 1\}^n$ that is ρ-correlated with $\boldsymbol{x}$ (meaning that each coordinate $\boldsymbol{y}_i$ is independently set to equal $\boldsymbol{x}_i$ with probability ρ and is set to a random draw from u_p with probability $1 - \rho$) for some ρ that is bounded away from 1. Let $w \in \mathbb{R}^n$ be τ-regular, and let $\ell(x)$ denote the linear form $\sum_{i=1}^n w_i x_i$. Then for any two intervals $[a, b]$ and $[c, d]$ in $\mathbb{R}$, we have*

$$\left| \mathbf{Pr}[(\ell(\boldsymbol{x}), \ell(\boldsymbol{y})) \in [a, b] \times [c, d]] - \Phi_{0,V}\left([\psi_p^{[w]}(a), \psi_p^{[w]}(b)] \times [\psi_p^{[w]}(c), \psi_p^{[w]}(d)]\right) \right| \leq O\left(\frac{\tau}{\sigma_p}\right),$$

where $V = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ and $\Phi_{0,V}$ denotes the distribution of the bivariate Gaussian with zero mean and covariance matrix V.

B.2. A structural theorem on regular LTFs under the p-biased distribution

The following is a p-biased variant of Theorem 48 of Matulef et al. (2010); intuitively, it says that the level-1 Fourier weight of a regular p-biased LTF is captured by the $W(\cdot)$ function that was introduced in Section B.1.1.

Theorem 21 *Let $f_1 : \{-1, 1\}^n \to \{-1, 1\}$ be a τ-regular linear threshold function. Then*

$$\left| \sum_{i=1}^n \hat{f}(i, p)^2 - W\left(\mathop{\mathbf{E}}_{\boldsymbol{x} \sim u_p^n}[f_1(\boldsymbol{x})]\right) \right| \leq O\left(\sqrt{\frac{\tau}{\sigma_p}}\right).$$

Further, suppose that $f_2 : \{-1, 1\}^n \to \{-1, 1\}$ is another τ-regular linear threshold function that can be expressed using the same linear form as f_1, i.e., $f_k = \mathrm{sign}(w \cdot x - \theta_k)$ for some w, θ_1, θ_2 and $k = 1, 2$. Then

$$\left| \left(\sum_{i=1}^n \hat{f}_1(i, p) \hat{f}_2(i, p) \right)^2 - W\left(\mathop{\mathbf{E}}_{\boldsymbol{x} \sim u_p^n}[f_1(\boldsymbol{x})]\right) W\left(\mathop{\mathbf{E}}_{\boldsymbol{x} \sim u_p^n}[f_2(\boldsymbol{x})]\right) \right| \leq O\left(\sqrt{\frac{\tau}{\sigma_p}}\right).$$

Proof We first note that we may assume that $\sqrt{\tau/\sigma_p}$ is bounded below 1, since otherwise the claimed bounds hold for trivial reasons. Using Lemma 18, we have that for $k = 1, 2$,

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim u_p^n}[f_k(\boldsymbol{x})] = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim u_p^n}[\mathrm{sign}(w \cdot \boldsymbol{x} - \theta_k)] \overset{\frac{\tau}{\sigma_p}}{\approx} m(\psi_p^{[w]}(\theta)). \tag{12}$$

Let $\boldsymbol{x} \sim u_p^n$ and let $\boldsymbol{y} \in \{-1, 1\}^n$ be ρ-correlated with $\boldsymbol{x}$ (as described in the statement of Fact 20) where $\rho = \sqrt{\tau/\sigma_p}$ is bounded away from 1. We have that

$$\mathbf{E}[f_1(\boldsymbol{x})f_2(\boldsymbol{y})] = \mathbf{Pr}[f_1(\boldsymbol{x}) = f_2(\boldsymbol{y})] - \mathbf{Pr}[f_1(\boldsymbol{x}) \neq f_2(\boldsymbol{y})] = 2\mathbf{Pr}[(w \cdot \boldsymbol{x}, w \cdot \boldsymbol{y}) \in A \cup B] - 1$$

where $A = [\theta_1, +\infty) \times [\theta_2, +\infty)$ and $B = (-\infty, \theta_1] \times (-\infty, \theta_2]$. Applying Fact 20 and recalling that w is τ-regular, we have that

$$\mathbf{Pr}\left[(w \cdot \boldsymbol{x}, w \cdot \boldsymbol{y}) \in A \cup B\right] \overset{\frac{\tau}{\sigma_p}}{\approx} \mathbf{Pr}\left[(\boldsymbol{X}, \boldsymbol{Y}) \in \widetilde{A} \cup \widetilde{B}\right]$$

where $(\widetilde{\theta}_1, \widetilde{\theta}_2) = (\psi_p^{[w]}(\theta_1), \psi_p^{[w]}(\theta_2))$ and $\widetilde{A} = [\widetilde{\theta}_1, +\infty) \times [\widetilde{\theta}_2, +\infty)$ and $\widetilde{B} = (-\infty, \widetilde{\theta}_2] \times (-\infty, \widetilde{\theta}_2]$ and $(\boldsymbol{X}, \boldsymbol{Y})$ are ρ-correlated $N(0,1)$ Gaussians.

It follows that

$$\mathbf{E}[f_1(\boldsymbol{x})f_2(\boldsymbol{y})] \overset{\frac{\tau}{\sigma_p}}{\approx} \mathbf{E}[h_{\widetilde{\theta}_1}(\boldsymbol{x}), h_{\widetilde{\theta}_2}(\boldsymbol{y})],$$

where $h_{\widetilde{\theta}_1}(\cdot)$ is the function of one Gaussian variable defined as $h_{\widetilde{\theta}_1}(t) := \text{sign}(t - \widetilde{\theta}_1)$. Using the Fourier and Hermite expansions of f_k and $h_{\widetilde{\theta}_k}$ and the fact that $\boldsymbol{x}, \boldsymbol{y}$ are ρ-correlated, we may rewrite the above approximate equality as:

$$\hat{f}_1(\emptyset, p)\hat{f}_2(\emptyset, p) + \rho \cdot \left(\sum_{i=1}^{n} \hat{f}_1(i, p)\hat{f}_2(i, p)\right) + \left(\sum_{|S| \geq 2} \rho^{|S|}\hat{f}_1(S, p)\hat{f}_2(S, p)\right)$$

$$\overset{\frac{\tau}{\sigma_p}}{\approx} \widetilde{h_{\widetilde{\theta}_1}}(0)\widetilde{h_{\widetilde{\theta}_2}}(0) + \rho\widetilde{h_{\widetilde{\theta}_1}}(1)\widetilde{h_{\widetilde{\theta}_2}}(1) + \left(\sum_{k \geq 2} \rho^k \widetilde{h_{\widetilde{\theta}_1}}(k)\widetilde{h_{\widetilde{\theta}_2}}(k)\right).$$

Now by Cauchy-Schwarz (and using the fact that $\rho \geq 0$) we have

$$\left|\sum_{|S| \geq 2} \rho^{|S|}\hat{f}_1(S, p)\hat{f}_2(S, p)\right| \leq \sqrt{\sum_{|S| \geq 2} \rho^{|S|}\hat{f}_1(S, p)^2}\sqrt{\sum_{|S| \geq 2} \rho^{|S|}\hat{f}_2(S, p)^2}$$

$$\leq \rho^2 \sqrt{\sum_{|S| \geq 2} \hat{f}_1(S, p)^2}\sqrt{\sum_{|S| \geq 2} \hat{f}_2(S, p)^2}$$

$$\leq \rho^2 \sqrt{\sum_{|S| \geq 0} \hat{f}_1(S, p)^2}\sqrt{\sum_{|S| \geq 0} \hat{f}_2(S, p)^2} \leq \rho^2.$$

By a similar analysis, we have that

$$\left|\sum_{k \geq 2} \rho^k \widetilde{h_{\widetilde{\theta}_1}}(k)\widetilde{h_{\widetilde{\theta}_2}}(k)\right| \leq \rho^2.$$

We further have by Lemma 18 that

$$\widetilde{h_{\widetilde{\theta}_k}}(0) = \underset{\boldsymbol{x} \sim N(0,1)}{\mathbf{E}}[h_{\widetilde{\theta}_k}(\boldsymbol{x})] = m(\widetilde{\theta}_k) \overset{\frac{\tau}{\sigma_p}}{\approx} \underset{\boldsymbol{x} \sim u_p^n}{\mathbf{E}}[f_k(\boldsymbol{x})] = \hat{f}_k(\emptyset, p),$$

and hence

$$\rho \cdot \left(\sum_{i=1}^{n} \hat{f}_1(i, p)\hat{f}_2(i, p)\right) \overset{\frac{\tau}{\sigma_p} + \rho^2}{\approx} \rho \ \widetilde{h_{\widetilde{\theta}_1}}(1)\widetilde{h_{\widetilde{\theta}_2}}(1) = \rho \cdot 2\phi(\widetilde{\theta}_1) \cdot 2\phi(\widetilde{\theta}_2),$$

where the equality is by Equation (10). Dividing by ρ and using $\frac{\tau}{\rho\sigma_p} + \rho \approx \frac{\tau^{1/2}}{\sigma_p^{1/2}}$ in the error estimate, we get

$$\sum_{i=1}^{n} \hat{f}_1(i,p)\hat{f}_2(i,p) \stackrel{\sqrt{\frac{\tau}{\sigma_p}}}{\approx} 2\phi(\widetilde{\theta}_1) \cdot 2\phi(\widetilde{\theta}_2) = \sqrt{W(m(\widetilde{\theta}_1)) \cdot W(m(\widetilde{\theta}_2))}$$

where the equality is by Equation (10).

Since we may apply this with f_1 and f_2 both equal to f_k, we may also conclude that

$$\sum_{i=1}^{n} \hat{f}(i,p)^2 \stackrel{\sqrt{\frac{\tau}{\sigma_p}}}{\approx} W(m(\widetilde{\theta}_k)).$$

Using the mean value theorem, the fact that $|W'| \leq 1$ on $[-1,1]$, and Equation (12), we can conclude that

$$\sum_{i=1}^{n} \hat{f}_k(i,p)^2 \stackrel{\sqrt{\frac{\tau}{\sigma_p}}}{\approx} W(\mathbf{E}\,[f_k]),$$

giving the first required approximate equality. Similar reasoning yields that

$$\left(\sum_{i=1}^{n} \hat{f}_1(i,p)\hat{f}_2(i,p)\right)^2 \stackrel{\sqrt{\frac{\tau}{\sigma_p}}}{\approx} W(m(\widetilde{\theta}_1)) \cdot W(m(\widetilde{\theta}_2))$$

and the proof is complete. ∎

B.3. p-biased Chow parameters are proportional to weights for regular LTFs

The following is a p-biased analogue of Lemma 6.11 of De et al. (2016); intuitively, it says that for a regular LTF, the vector of weights is close (after a suitable scaling) to the vector of degree-1 Fourier coefficients.

Proposition 22 *Let $f : \{-1,1\}^n \to \{-1,1\}$, $f(x) = \text{sign}(w \cdot x - \theta)$ where w is τ-regular and $\|w\|_2 = 1$. Then*

$$\sum_{i\in[n]} (\hat{f}(i,p) - \alpha(\psi_p^{[w]}(\theta))w_i)^2 \leq O\left(\sqrt{\frac{\tau}{\sigma_p^2}}\right).$$

Proof First we fix some notation: we will write $\mu := \sum_{i=1}^{n} w_i\mu_p$, and we observe that with this notation we have $\psi_p^{[w]}(x) = \frac{x-\mu}{\sigma_p}$. Recalling that $\psi_p(x_i) = \frac{x_i-\mu_p}{\sigma_p}$, we begin by noting that

$$\text{sign}(w \cdot x - \theta) = \text{sign}\left(\left(\sum_{i=1}^{n} w_i\psi_p(x_i)\right) - \psi_p^{[w]}(\theta)\right).$$

26

As we will see, the latter expression is convenient because it contains a linear combination of functions $\psi_p(x_1), \ldots, \psi_p(x_n)$ which are orthonormal under the u_p^n distribution.

We consider two cases depending on the magnitude of $\psi_p^{[w]}(\theta)\sigma_p$.

Case 1: The first case is that $|\psi_p^{[w]}(\theta)\sigma_p| > \sqrt{2\ln(4/\frac{\tau}{\sigma_p})}$, or equivalently, $-\frac{(\psi_p^{[w]}(\theta)\sigma_p)^2}{2} < \ln(\frac{\tau}{\sigma_p}/4)$. In this case since f is ± 1-valued, we have $1 - \mathbf{E}[f(\boldsymbol{x})]^2 = 4\,\mathbf{Pr}[f(\boldsymbol{x}) = 1]\,\mathbf{Pr}[f(\boldsymbol{x}) = -1] \leq 4\,\mathbf{Pr}[f(\boldsymbol{x}) = 1]$, and since

$$1 - \hat{f}(\emptyset, p)^2 = \sum_{|S|>0} \hat{f}(S,p)^2 \geq \sum_{i \in [n]} \hat{f}(i,p)^2,$$

this yields

$$\sum_{i \in [n]} \hat{f}(i,p)^2 \leq 4\,\mathbf{Pr}\left[\sum_{i \in [n]} w_i \psi_p(x_i) \geq \psi_p^{[w]}(\theta)\right].$$

By Hoeffding's inequality and the assumption on $|\psi_p^{[w]}(\theta)\sigma_p|$ that put us in Case 1, we have that

$$\mathbf{Pr}\left[\sum_{i \in [n]} w_i \psi_p(x_i) \geq \psi_p^{[w]}(\theta)\right] \leq \exp\left(\frac{-2(\psi_p^{[w]}(\theta))^2}{4\|w\|_2^2/\sigma_p^2}\right) \leq \frac{1}{4} \cdot \frac{\tau}{\sigma_p},$$

so consequently we have that

$$\sum_{i \in [n]} \hat{f}(i,p)^2 \leq \frac{\tau}{\sigma_p}.$$

On the other hand, by Fact 14 and the definition of $\alpha(\cdot)$ we also have that $\sum_{i \in [n]} \left(\alpha(\psi_p^{[w]}(\theta))w_i\right)^2 = W(m(\psi_p^{[w]}(\theta)))$. Applying Lemma 18, we get that

$$\left|\mathbf{E}_{\boldsymbol{x} \sim u_p^n}[f(\boldsymbol{x})] - m(\psi_p^{[w]}(\theta))\right| \leq O\left(\frac{\tau}{\sigma_p}\right).$$

Observing that that the function $W(\cdot)$ is a contraction (it satisfies $|W'| < 1$), we get that

$$\left|W\left(\mathbf{E}_{\boldsymbol{x} \sim u_p^n}[f(\boldsymbol{x})]\right) - W(m(\psi_p^{[w]}(\theta)))\right| \leq O\left(\frac{\tau}{\sigma_p}\right).$$

Recalling Theorem 21, we further have that:

$$\left|W\left(\mathbf{E}_{\boldsymbol{x} \sim u_p^n}[f(\boldsymbol{x})]\right) - \sum_{i \in [n]} \hat{f}(i,p)^2\right| \leq O\left(\sqrt{\frac{\tau}{\sigma_p}}\right).$$

Putting the pieces together and applying the inequality $(a-b)^2 \leq 2(a^2 + b^2)$, we get that

$$\sum_{i \in [n]} (\hat{f}(i,p) - \alpha(\psi_p^{[w]}(\theta))w_i)^2 \leq 2\sum_{i \in [n]} (\hat{f}(i,p))^2 + 2\sum_{i \in [n]} (\alpha(\psi_p^{[w]}(\theta))w_i)^2 \leq O\left(\sqrt{\frac{\tau}{\sigma_p}}\right)$$

as desired.

Case 2: The remaining case is that $|\psi_p^{[w]}(\theta)\sigma_p| < \sqrt{2\ln(4/\frac{\tau}{\sigma_p})}$.

To show that $\sum_{i\in[n]}(\hat{f}(i,p) - \alpha(\psi_p^{[w]}(\theta))w_i)^2 \leq \varepsilon$, it suffices to show that

$$\sum_{i\in[n]}(\hat{f}(i,p))^2 + \sum_{i\in[n]}(\alpha(\psi_p^{[w]}(\theta))w_i)^2 \overset{\varepsilon}{\approx} 2\sum_{i\in[n]}(\hat{f}(i,p))(\alpha(\psi_p^{[w]}(\theta))w_i) \tag{13}$$

The analysis just given for Case 1 lets us control the LHS of Equation (13) as

$$\sum_{i\in[n]}(\hat{f}(i,p))^2 + \sum_{i\in[n]}(\alpha(\psi_p^{[w]}(\theta))w_i)^2 = \sum_{i\in[n]}(\hat{f}(i,p))^2 + \alpha(\psi_p^{[w]}(\theta))^2 \overset{\sqrt{\frac{\tau}{\sigma_p}}}{\approx} 2W(m(\psi_p^{[w]}(\theta))).$$

Turning to the RHS, we have that

$$2\sum_{i\in[n]}\hat{f}(i,p)\alpha(\psi_p^{[w]}(\theta))w_i = 2\alpha(\psi_p^{[w]}(\theta))\sum_{i\in[n]}\hat{f}(i,p)w_i = 2\alpha(\psi_p^{[w]}(\theta))\underset{\boldsymbol{x}\sim u_p^n}{\mathbf{E}}\left[f(\boldsymbol{x})\sum_{i\in[n]}w_i\psi_p(\boldsymbol{x}_i)\right].$$
$$\tag{14}$$

We can re-express the expectation above as

$$\underset{\boldsymbol{x}\sim u_p^n}{\mathbf{E}}\left[f(\boldsymbol{x})\sum_{i\in[n]}w_i\psi_p(\boldsymbol{x}_i)\right] = \mathbf{E}\left[f(\boldsymbol{x})\left(\sum_{i\in[n]}w_i\psi_p(\boldsymbol{x}_i) - \psi_p^{[w]}(\theta)\right)\right] + \psi_p^{[w]}(\theta)\,\mathbf{E}\left[f(\boldsymbol{x})\right]$$

$$= \mathbf{E}\left[\mathrm{sign}\left(\sum_{i\in[n]}w_i\psi_p(\boldsymbol{x}_i) - \psi_p^{[w]}(\theta)\right)\left(\sum_{i\in[n]}w_i\psi_p(\boldsymbol{x}_i) - \psi_p^{[w]}(\theta)\right)\right]$$

$$+ \psi_p^{[w]}(\theta)\,\mathbf{E}\left[f(\boldsymbol{x})\right]$$

$$= \mathbf{E}\left[|\ell(\boldsymbol{x})|\right] + \psi_p^{[w]}(\theta)\,\mathbf{E}\left[f(\boldsymbol{x})\right],$$

where we write $\ell(x)$ to denote

$$\ell(x) := \left(\sum_{i=1}^n w_i\psi_p(x_i)\right) - \psi_p^{[w]}(\theta).$$

By Lemma 19 we have that

$$\mathbf{E}[|\ell(\boldsymbol{x})|] = \mathbf{E}\left[\left|\sum_{i=1}^n \frac{w_i}{\sigma_p\|w\|_2}\boldsymbol{x}_i - \frac{\theta}{\sigma_p\|w\|_2}\right|\right] \overset{\tau}{\approx} \underset{\boldsymbol{x}\sim N(0,1)}{\mathbf{E}}\left[\left|\boldsymbol{x} - \psi_p^{[\frac{w}{\sigma_p\|w\|_2}]}\left(\frac{\theta}{\sigma_p\|w\|_2}\right)\right|\right] \overset{\text{Fact 12}}{=} \mathbf{E}\left[|\boldsymbol{x} - \psi_p^{[w]}(\theta)|\right].$$

Recalling the last equality of Lemma 19, this gives that

$$\underset{\boldsymbol{x}\sim u_p^n}{\mathbf{E}}[|\ell(\boldsymbol{x})|] \overset{\tau}{\approx} \left(2\phi(\psi_p^{[w]}(\theta)) - \psi_p^{[w]}(\theta)m(\psi_p^{[w]}(\theta))\right).$$

By Lemma 18 we have that

$$\psi_p^{[w]}(\theta) \mathop{\mathbf{E}}_{\boldsymbol{x} \sim u_p^n}[f(\boldsymbol{x})] \overset{\psi_p^{[w]}(\theta)\frac{\tau}{\sigma_p}}{\approx} \psi_p^{[w]}(\theta)m(\psi_p^{[w]}(\theta)).$$

Putting these pieces together, we can re-express Equation (14) as

$$2\alpha(\psi_p^{[w]}(\theta)) \mathop{\mathbf{E}}_{\boldsymbol{x} \sim u_p^n}\left[f(\boldsymbol{x})\sum_{i\in[n]}w_i\psi_p(\boldsymbol{x}_i)\right] = 2\alpha(\psi_p^{[w]}(\theta))\Big(\mathbf{E}\left[|\ell(\boldsymbol{x})|\right] + \psi_p^{[w]}(\theta)\mathbf{E}\left[f(\boldsymbol{x})\right]\Big)$$

$$\overset{(\max_{\theta\in\mathbb{R}}\{\alpha(\theta)\})\cdot(\psi_p^{[w]}+1)\cdot\frac{\tau}{\sigma_p}}{\approx} 2\alpha(\psi_p^{[w]}(\theta))\cdot\Big(2\phi(\psi_p^{[w]}(\theta))\Big)$$

$$\overset{(\psi_p^{[w]}+1)\cdot\frac{\tau}{\sigma_p}}{\approx} 2W(m(\psi_p^{[w]}(\theta))),$$

where for the last line we used the definitions of α and W and the fact that W is uniformly bounded by $\sqrt{2/\pi}$.

Since the above analyses of the LHS and the RHS show that each of these can be approximated by the same quantity $2W(m(\psi_p^{[w]}(\theta)))$, we deduce that

$$\sum_{i\in[n]}(\hat{f}(i,p))^2 + \sum_{i\in[n]}(\alpha(\psi_p^{[w]}(\theta))w_i)^2 \overset{(\psi_p^{[w]}+1)\cdot\frac{\tau}{\sigma_p}+\sqrt{\frac{\tau}{\sigma_p}}}{\approx} 2\sum_{i\in[n]}(\hat{f}(i,p))(\alpha(\psi_p^{[w]}(\theta))w_i).$$

It remains only to verify that

$$(\psi_p^{[w]}+1)\cdot\frac{\tau}{\sigma_p} + \sqrt{\frac{\tau}{\sigma_p}} \leq O\left(\frac{1}{\sigma_p}\sqrt{\log\left(\frac{1}{\tau/\sigma_p}\right)}\cdot\frac{\tau}{\sigma_p} + \sqrt{\frac{\tau}{\sigma_p}}\right) \leq O\left(\sqrt{\frac{\tau}{\sigma_p^2}}\right),$$

and the proof is complete. ∎

The following corollary, which applies to the centralized version of the weight vector w (see the definition immediately before Fact 60) and is an immediate consequence of Proposition 22 and Fact 60, will be useful for our analysis of the Shapley problem:

Corollary 23

$$\sum_{i\in[n]}\left((\sigma_p\hat{f}(i,p) - \frac{1}{n}\sum_{i\in[n]}\sigma_p\hat{f}(i,p)) - \sigma_p\alpha(\psi_p^{[w]}(\theta))(w_i - \frac{1}{n}\sum_{i\in[n]}w_i)\right)^2 \leq O\left(\sqrt{\tau}\right).$$

B.4. Structural results on heads and tails of LTFs (Chow version)

Let $f(x) = \text{sign}(w \cdot x - \theta)$ be an LTF, and to simplify the presentation let us assume that its weights are sorted in magnitude from largest to smallest, i.e., $|w_1| \geq |w_2| \geq \cdots \geq |w_n|$. (Of course this need not hold for the algorithmic problems on LTFs that we consider, but this can be assumed without loss of generality for the structural results we are concerned with in this section.) Let $\tau > 0$.

Although f need not be τ-regular, we can always partition its weights w into "head weights" w_H and "tail weights" w_T such that w_T is τ-regular and any longer suffix of w is not τ-regular. Let $H = \{1, 2, \ldots, \}$ be the set of indices of head weights, and let $T = [n] \setminus H = \{\ldots, n-1, n\}$ be the set of indices of tail weights.

In this section we prove a number of structural results on the p-biased Chow parameters of the head and tail variables in an arbitrary LTF. For the original Chow parameters problem we will only use the $p = 1/2$ case of these results, but the more general case of $p \in (0, 1)$ will be used later in our approach to the Shapley problem.

B.4.1. REGULAR TAIL WEIGHTS ARE PROPORTIONAL TO TAIL CHOW PARAMETERS

We first show that there exists a value $\alpha = \alpha(\theta, w_H, p)$ such that $\hat{f}(i, p) \approx \alpha \cdot w_i$ for all $i \in T$. More precisely, we show that the vector of tail weights w_T is proportional to the vector of tail Fourier coefficients $(\hat{f}(i, p))_{i \in T}$. This characterization will be helpful for recovering the tail weights of an LTF from its tail Fourier coefficients, and as a corollary also gives an approximation of the Fourier weight of f on T.

Proposition 24 *Let* $f((x_H, x_T) = \text{sign}(w_H \cdot x_H + w_T \cdot x_T - \theta)$ *where* w_T *is* τ-*regular and* $\|w_T\|_2 = 1$. *Fix* $p \in (0, 1)$. *Then*

$$\sum_{i \in T} (\hat{f}(i, p) - \alpha(\theta, w_H, w_T, p) \cdot w_i)^2 \leq O\left(\sqrt{\frac{\tau}{\sigma_p^2}}\right),$$

where we recall that $\alpha(\theta, w_H, w_T, p) = \mathbf{E}_{\boldsymbol{\rho} \sim u_p^k}[\alpha(\psi_p^{[w_T]}(\theta - w_H \cdot \boldsymbol{\rho}))]$ *and* $\alpha(\theta) = \sqrt{W(\mu(\theta))}$ *as in Definition 13.*

Proof For $\rho \in \{-1, 1\}^{|H|}$, let $f_\rho(x_T) = f(\rho, x_T)$ denote f with its head variables fixed to ρ. Then

$$\sum_{i \in T} (\hat{f}(i, p) - \alpha(\theta, w_H, w_T, p) \cdot w_i)^2$$

$$= \sum_{i \in T} \left(\mathbf{E}_{\boldsymbol{\rho} \sim u_p^{|H|}}[\hat{f}_{\boldsymbol{\rho}}(i, p) - \alpha(\psi_p^{[w_T]}(\theta - w_H \cdot \boldsymbol{\rho})) \cdot w_i] \right)^2 \quad \text{(By definition)}$$

$$\leq \sum_{i \in T} \mathbf{E}_{\boldsymbol{\rho} \sim u_p^{|H|}}[(\hat{f}_{\boldsymbol{\rho}}(i, p) - \alpha(\psi_p^{[w_T]}(\theta - w_H \cdot \boldsymbol{\rho})) \cdot w_i)^2] \quad \text{(By Jensen's inequality)}$$

$$= \mathbf{E}_{\boldsymbol{\rho} \sim u_p^{|H|}} \left[\sum_{i \in T} (\hat{f}_{\boldsymbol{\rho}}(i, p) - \alpha(\psi_p^{[w_T]}(\theta - w_H \cdot \boldsymbol{\rho})) \cdot w_i)^2 \right] \quad \text{(By linearity of expectation)}$$

$$\leq \max_{\rho \in \{-1,1\}^{|H|}} \sum_{i \in T} (\hat{f}_\rho(i, p) - \alpha(\psi_p^{[w_T]}(\theta - w_H \cdot \rho)) \cdot w_i)^2$$

$$\leq O\left(\sqrt{\frac{\tau}{\sigma_p^2}}\right). \quad \text{(By Proposition 22 applied to } f_\rho)$$

∎

As a corollary, we get that the ℓ_2-weight of the tail of the Chow parameters $\gamma := \|(\hat{f}(i))_{i \in T}\|_2$ and the constant of proportionality $\alpha(\theta, w_H, w_T, p)$ are good approximations of each other for functions f of the form described in Proposition 24.

Corollary 25 *Let $f(x) = f(x_H, x_T) = \text{sign}(w_H \cdot x_H + w_T \cdot x_T - \theta)$ where w_T is τ-regular and $\|w_T\|_2 = 1$. Then $\|(\hat{f}(i))_{i \in T}\|_2 \overset{(\tau/\sigma_p^2)^{1/4}}{\approx} \alpha(\theta, w_H, w_T, p)$.*

Proof Proposition 24 says that the Euclidean distance between the vectors $(\hat{f}(i))_{i \in T}$ and $\alpha(\theta, w_H, w_T, p)w_T$ is at most $O((\tau/\sigma_p^2)^{1/4})$. The corollary follows from Fact 61 since the Euclidean length of the vector $\alpha(\theta, w_H, w_T, p)w_T$ is $\alpha(\theta, w_H, w_T, p)$. ∎

B.4.2. PRESERVING THE HEAD CHOW PARAMETERS

In this section we show that exchanging the tail weights w_T of an LTF f with other weights w_T' of the same ℓ_1 and ℓ_2 norm does not change the head Fourier coefficients $(\hat{f}(i, p))_{i \in H}$ by too much. This will be helpful for handling instances of the Partial Chow Parameters Problem with missing tail Fourier coefficients.

The following lemma is a generalization of (De et al., 2016, Lemma 6.13), and its proof closely follows the proof given there. It essentially says that as far as head Fourier coefficients are concerned, when the tail is regular it does not much matter whether the tail variables are p-biased Boolean random variables or Gaussian random variables with mean and variance matching the p-biased distribution over $\{-1, 1\}$.

Lemma 26 *Let $f(x) = f(x_H, x_T) = \text{sign}(w_H \cdot x_H + w_T \cdot x_T - \theta)$ where $(w_H, w_T) \in (\mathbb{R}^{\geq 0})^n$, and where w_T is τ-regular. Fix $p \in (0, 1)$. Recall that $\hat{f}(i, p) = \mathbf{E}_{\boldsymbol{x} \sim u_p^n}[f(\boldsymbol{x}) \cdot \psi_p(\boldsymbol{x}_i)]$ and let $\check{f}(i, p) = \mathbf{E}_{\boldsymbol{x}_H \sim u_p^{|H|}, \boldsymbol{x}_T \sim N(\mu_p, \sigma_p^2)^{|T|}}[f(\boldsymbol{x}) \cdot \psi_p(\boldsymbol{x}_i)]$. Then*

$$\sum_{i \in H} (\hat{f}(i, p) - \check{f}(i, p))^2 \leq O\left(\frac{\tau^2}{\sigma_p^2}\right) .$$

Proof Define the functions $f', f'' : \{-1, 1\}^{|H|} \to [-1, 1]$ as follows:

$$f'(x_H) = \mathbf{E}_{\boldsymbol{x}_T \sim u_p^{|T|}}[f(x_H, \boldsymbol{x}_T)], \quad f''(x_H) = \mathbf{E}_{\boldsymbol{x}_T \sim N(\mu_p, \sigma_p^2)^{|T|}}[f(x_H, \boldsymbol{x}_T)] .$$

Then

$$\begin{aligned}
\sum_{i \in H} (\hat{f}(i, p) - \check{f}(i, p))^2 &= \sum_{i \in H} (\hat{f}'(i, p) - \hat{f}''(i, p))^2 \\
&\leq \sum_{S \subseteq H} (\hat{f}_p'(S) - \hat{f}_p''(S))^2 \\
&= \mathbf{E}_{\boldsymbol{x} \sim u_p^{|H|}} (f'(\boldsymbol{x}) - f''(\boldsymbol{x}))^2 \\
&\leq \max_{x \in \{-1, 1\}^{|H|}} (f'(x) - f''(x))^2 ,
\end{aligned}$$

where the first equality holds since $\hat{f}'(i, p) = \hat{f}(i, p)$ and $\hat{f}''(i, p) = \check{f}(i, p)$ for all $i \in H$ by definition, and the second equality is by Parseval's identity.

31

To upper bound the last expression, we observe that for every $\rho \in \{-1, 1\}^{|H|}$ it holds that

$$|f'(\rho) - f''(\rho)| = 2 \cdot \left| \Pr_{\boldsymbol{x}_T \sim u_p^{|T|}}[w_T \cdot \boldsymbol{x}_T \geq \theta - w_H \cdot \rho] - \Pr_{\boldsymbol{x}_T \sim N(\mu_p, \sigma_p^2)^{|T|}}[w_T \cdot \boldsymbol{x}_T \geq \theta - w_H \cdot \rho] \right|$$

$$\leq O(\tau/\sigma_p) \, ,$$

where the inequality uses the assumption that w_T is τ-regular to apply Fact 17. ∎

The following theorem is essentially a corollary of Lemma 26. Its proof is similar to (De et al., 2016, Lemma 6.15).

Theorem 27 *Let $f(x) = f(x_H, x_T) = \mathrm{sign}(w_H \cdot x_H + w_T \cdot x_T - \theta)$ and let $g(x) = f'(x_H, x_T) = \mathrm{sign}(w_H \cdot x_H + w_T' \cdot x_T - \theta)$ where $(w_H, w_T) \in (\mathbb{R}^{\geq 0})^n$, and where w_T, w_T' are τ-regular, satisfy $\|w_T\|_1 = \|w_T'\|_1$, and satisfy $\|w_T\|_2 = \|w_T'\|_2$. Fix $p \in (0, 1)$. As in Lemma 26, for $h \in \{f, g\}$ recall that $\hat{h}(i, p) = \mathbf{E}_{\boldsymbol{x} \sim u_p^n}[h(\boldsymbol{x})\psi_p(\boldsymbol{x}_i)]$. Then*

$$\sum_{i \in H}(\hat{f}(i, p) - \hat{g}(i, p))^2 \leq O\left(\frac{\tau^2}{\sigma_p^2}\right) \, .$$

Proof As in Lemma 26, for $h \in \{f, g\}$ let $\check{h}(i, p) = \mathbf{E}_{\boldsymbol{x}_H \sim u_p^{|H|}, \boldsymbol{x}_T \sim N(\mu_p, \sigma_p^2)^{|T|}}[h(\boldsymbol{x})\psi_p(\boldsymbol{x}_i)]$. By the triangle inequality we have that

$$\sqrt{\sum_{i \in H}(\hat{f}(i, p) - \hat{g}(i, p))^2}$$

$$\leq \sqrt{\sum_{i \in H}(\hat{f}(i, p) - \check{f}(i, p))^2} + \sqrt{\sum_{i \in H}(\check{f}(i, p))^2 - \check{g}(i, p))^2} + \sqrt{\sum_{i \in H}(\hat{g}(i, p) - \check{g}(i, p))^2)^2} \, .$$

We prove the theorem by upper bounding each term on the right hand side. By Lemma 26 and the τ-regularity of w_T and w_T', we have that the first and third terms are upper bounded by $O(\tau/\sigma_p)$. Furthermore, we claim that the second term is equal to 0. This follows from the fact that for every $i \in H$,

$$\check{f}(i, p) = \mathbf{E}_{\boldsymbol{x}_H \sim u_p^{|H|}, \boldsymbol{x}_T \sim N(\mu_p, \sigma_p^2)^{|T|}}[\mathrm{sign}(w_H \cdot \boldsymbol{x}_H + w_T \cdot \boldsymbol{x}_T - \theta) \cdot \psi_p(\boldsymbol{x}_i)]$$

$$= \mathbf{E}_{\boldsymbol{x}_H \sim u_p^{|H|}, \boldsymbol{x}_T \sim N(\mu_p, \sigma_p^2)^{|T|}}[\mathrm{sign}(w_H \cdot \boldsymbol{x}_H + w_T' \cdot \boldsymbol{x}_T - \theta) \cdot \psi_p(\boldsymbol{x}_i)]$$

$$= \check{g}(i, p) \, .$$

Here the first and third equalities are by definition. The second equality uses the fact that $w_T \cdot \boldsymbol{x}_T$ and $w_T' \cdot \boldsymbol{x}_T$ are both (exactly) distributed as $N(\mu_p \cdot \|w\|_1, \sigma_p^2 \cdot \|w\|_2^2)$ when $\boldsymbol{x}_T \sim N(\mu_p, \sigma_p^2)^{|T|}$, which holds since $\|w_T\|_1 = \|w_T'\|_1$ and $\|w_T\|_2 = \|w_T'\|_2$. ∎

Corollary 28 *When $p = \frac{1}{2}$, Theorem 27 holds without the assumption that $\|w_T\|_1 = \|w_T'\|_1$.*

Proof When $p = \frac{1}{2}$, $w_T \cdot \boldsymbol{x}_T$ and $w_T' \cdot \boldsymbol{x}_T$ are both (exactly) distributed as $N(0, \|w\|_2^2)$ when $\boldsymbol{x}_T \sim N(\mu_{\frac{1}{2}}, \sigma_{\frac{1}{2}}^2)^{|T|}$ since $\mu_{\frac{1}{2}} = 0$ and $\sigma_{\frac{1}{2}} = 1$. ∎

Appendix C. The Partial Chow Parameters Problem

In this section we prove our main structural result for the Partial Chow Parameters Problem, Theorem 7 (restated as Theorem 30 below). As discussed in Section 6, together with the algorithm and analysis presented in Section 7, this gives an EPRAS for solving the Partial Inverse Chow Parameters Problem.

We recall the following structural theorem of O'Donnell and Servedio (2011):

Theorem 29 ((O'Donnell and Servedio, 2011, Theorem 7.3)) *Let $\varepsilon \in (0, \frac{1}{2})$ and let $\tau = \tau(\varepsilon)$ be a certain $\mathrm{poly}(\varepsilon)$ value. Let f be an LTF $f(x) = f(x_H, x_T) = \mathrm{sign}(w_H \cdot x_H + w_T \cdot x_T - \theta)$ where H contains all indices $i \in [n]$ with $|\hat{f}(i)| \geq \tau^2$. Then at least one of the following holds:*

1. *f is $O(\varepsilon)$-close to a linear threshold function junta f' over x_H, or*

2. *f is $O(\varepsilon)$-close to a linear threshold function f' of the following form:*

$$f'(x) = f'(x_H, x_T) = \mathrm{sign}(\sum_{i \in H} v_i \cdot x_i + \gamma^{-1} \sum_{i \in T} \hat{f}(i) \cdot x_i - \theta') \qquad (15)$$

where $\theta' \in \mathbb{R}, v_H = (v_1, \ldots, v_{|H|}) \in \mathbb{R}^{|H|}$ are such that each v_i is an integer multiple of $\sqrt{\tau}/|H|$ and has magnitude at most $2^{O(|H| \log |H|)}\sqrt{\ln(1/\tau)}$, and where $\gamma := \left(\sum_{i \in T} \hat{f}(i)^2\right)^{1/2}$.

We adapt Theorem 29 to obtain our main structural result for the Partial Chow problem, Theorem 7, which was stated in Section 7 and is restated below. It is syntactically similar, but has the key conceptual difference that it works for the "partial information" case: given only the Chow Parameters of an LTF f corresponding to a *subset* of indices $S \subseteq \{0, 1, \ldots, n\}$, it specifies a relatively small set of LTFs which is guaranteed to include one that is close to f in Partial Chow Distance with respect to S.

Theorem 30 (Same as Theorem 7) *Let $\varepsilon \in (0, \frac{1}{2})$ and let $\tau = \tau(\varepsilon) = \varepsilon^{1000}$. Let f be an LTF $f(x) = f(x_H, x_T) = \mathrm{sign}(w_H \cdot x_H + w_T \cdot x_T - \theta)$ where H contains all indices $i \in [n]$ with $|\hat{f}(i)| \geq \tau^2$. Let $S \subseteq \{0, 1, \ldots, n\}$. Then one of the following holds:*

1. *$d_{\mathrm{Chow}, S}(f, g) \leq O(\varepsilon)$ for some linear threshold function junta g over x_H, or*

2. *$d_{\mathrm{Chow}, S}(f, g) \leq O(\varepsilon)$ for some linear threshold function g of the form*

$$g(x) = g(x_H, x_T) = \mathrm{sign}\Big(\sum_{i \in H} v_i \cdot x_i + (\gamma')^{-1} \cdot \Big(\sum_{i \in T \cap S} \hat{f}(i) \cdot x_i + \sum_{i \in T \cap \bar{S}} r \cdot x_i \Big) - \theta' \Big) \quad (16)$$

where $\theta' \in \mathbb{R}, v_H = (v_1, \ldots, v_{|H|}) \in \mathbb{R}^{|H|}$ are such that each v_i is an integer multiple of $\sqrt{\tau}/|H|$ and has magnitude at most $2^{O(|H| \log |H|)}\sqrt{\ln(1/\tau)}$, where γ' satisfies $\gamma' = \gamma$ if $T \subseteq S$ and $\gamma \leq \gamma' \leq \gamma + \tau$ if $T \not\subseteq S$ for $\gamma := \left(\sum_{i \in T} \hat{f}(i)^2\right)^{1/2}$, and where $r := \left(\frac{(\gamma')^2 - \sum_{i \in T \cap S} \hat{f}(i)^2}{|T \cap \bar{S}|}\right)^{1/2}$.

33

We note the close analogy between Theorem 29 and Theorem 30, and the similarity between Equation (15) and Equation (16): both have tail weight vectors whose ℓ_2 norm is equal to 1, and when $T \subseteq S$ the tail weight vectors are the same. The differences are that in Equation (16), we use the same weight r for all of the variables corresponding to missing tail Chow Parameters,[5] and use γ' as a slight overestimate of γ, the ℓ_2 norm of the tail Chow parameters of f (the slight overestimate is to ensure that r is the square root of a non-negative number).

The main idea behind the proof of Theorem 30 is to upper bound $d_{\text{Chow},S}(f,g)$ using the triangle inequality by

$$d_{\text{Chow},S}(f,g) \leq d_{\text{Chow},H}(f,f') + d_{\text{Chow},H}(f',g) + d_{\text{Chow},T\cap S}(f,g) \qquad (17)$$

for a function f' of the form described in Equation (15) and a function g of the form described in Equation (16), and then to upper bound each term in the right-hand side of Equation (17). Roughly speaking, by Theorem 29 there is a function f' of the form described in Equation (15) so that $d_{\text{Chow},H}(f,f')$ will be small; by the head weight stability result described in Proposition 28 $d_{\text{Chow},H}(f',g)$ will be small; and by the proportionality of the tail weights and Chow Parameters described in Proposition 24, $d_{\text{Chow},T\cap S}(f,g)$ will be small.

We also crucially rely on the regularity of the tails of the weight vectors of the functions f' and g, which is established in Claim 32 and Claim 33 below. Note that we will not explicitly find a function f' of the form described in Equation (15), but merely use its existence to prove Theorem 30.

C.1. Useful facts about tail weights

We will use the following lower bound, from O'Donnell and Servedio (2011), on the tail weight of the LTF in Equation (15).

Fact 31 ((O'Donnell and Servedio, 2011, Equation (8.8))) *Let $f : \{-1,1\}^n \to \{-1,1\}$ be an LTF which has $d(f,g) = \Omega(\varepsilon^2)$ for every Boolean function g which is a junta on the coordinates $H = \{i \in [n] : \hat{f}(i) \geq \tau^2\}$, for $\tau = \Omega(\varepsilon^{288})$. Let $T = [n] \setminus H$. Then $\gamma = \left(\sum_{i \in T} \hat{f}(i)^2 \right)^{1/2} \geq \Omega(\tau^{1/72})$.*

The next two claims establish the regularity of the tails of the weight vectors of the functions f' and g defined in Equation (15) and Equation (16).

Claim 32 *The vector v_T of tail weights of each function f' of the form in Equation (15) is $O(\tau^{143/72})$-regular.*

Proof As stated above, let $\gamma = \left(\sum_{i \in T} \hat{f}(i)^2 \right)^{1/2}$. Since $\|v_T\|_2 = 1$ it suffices to upper bound $\|v_T\|_\infty$ in order to establish regularity. We then have that

$$\|v_T\|_\infty = \max_{i \in T} \frac{|\hat{f}(i)|}{\gamma} \leq \frac{\tau^2}{\Omega(\tau^{1/72})} \leq O(\tau^{143/72}) \,. \qquad (18)$$

---

5. Setting each of the remaining weights equal to r as we do here corresponds to "spreading the remaining power out evenly." However, our techniques are fairly robust to the exact way in which we set the missing tail weights in Equation (16), essentially just requiring that they complete the tail weights to a $\text{poly}(\tau)$-regular vector with unit ℓ_2 norm. For example, if we wanted to "concentrate the remaining power," we could (up to handling minor rounding issues) set $(1 - \sum_{i \in T \cap S} \hat{f}(i)^2)/\tau^2$ missing tail weights equal to τ and the rest equal to zero.

Here the equality uses the definition of v_T, and the first inequality uses the assumption that $|\hat{f}(i)| \leq \tau^2$ for all $i \in T$ to upper bound the numerator and uses Fact 31 to lower bound the denominator. ∎

Claim 33 *The vector v'_T of tail weights of each function g of the form in Equation (16) is $O(\tau^{71/144})$-regular.*

Proof If $T \subseteq S$, then $v'_T = v_T$ and the result follows by Claim 32. So, assume that $T \nsubseteq S$. Let γ, γ', and r be as in Theorem 30. By the definition of r, $\|v'_T\|_2 = 1$ so it suffices to upper bound $\|v'_T\|_\infty$ in order to show regularity. By definition,

$$\|v'_T\|_\infty = (\gamma')^{-1} \cdot \max(\{|\hat{f}(i)| : i \in T \cap S\} \cup \{r\}) .$$

By assumption, $\gamma' \geq \gamma$, and so $(\gamma')^{-1} \cdot \max(\{|\hat{f}(i)| : i \in T \cap S\} \leq \gamma^{-1} \cdot \max(\{|\hat{f}(i)| : i \in T \cap S\} \leq O(\tau^{143/72}) \leq O(\tau^{71/144})$ by Claim 32.

It remains to upper bound r/γ'. Let $n' := |T \cap \bar{S}|$. We will use the fact that

$$\tau^2 \geq \Omega\Big(\frac{1}{\gamma^2} \max_{i \in T} \hat{f}(i)^2\Big) \geq \Omega\Big(\frac{1}{\gamma^2 n'} \sum_{i \in T \cap \bar{S}} \hat{f}(i)^2\Big) = \Omega\Big(\frac{1}{n'} \cdot \Big(1 - \frac{1}{\gamma^2} \sum_{i \in T \cap S} \hat{f}(i)^2\Big)\Big), \quad (19)$$

where the first inequality is from Equation (18) and the equality follows by the definition of γ. We have:

$$
\begin{aligned}
n' \cdot (r/\gamma')^2 &= 1 - \frac{1}{(\gamma')^2} \sum_{i \in T \cap S} \hat{f}(i)^2 && \text{(By definition of } r) \\
&\leq \big(1 - \frac{1}{\gamma^2} \sum_{i \in T \cap S} \hat{f}(i)^2\big) + \big(\big(\frac{\gamma'}{\gamma}\big)^2 - 1\big) && \text{(Multiplying by } (\gamma'/\gamma)^2) \\
&\leq n' \cdot O(\tau^2) + \big(\big(\frac{\gamma'}{\gamma}\big)^2 - 1\big) && \text{(By Equation (19))} \\
&\leq n' \cdot O(\tau^2) + O(\tau/\gamma + (\tau/\gamma)^2) && \text{(Since } \gamma' \leq \gamma + \tau) \\
&\leq n' \cdot O(\tau^2) + O(\tau^{71/72}). && \text{(Since } \gamma \geq \Omega(\tau^{1/72}))
\end{aligned}
$$

Dividing both sides by n' and taking square roots, we get that $r/\gamma' \leq O(\tau^{71/144})$, proving the claim. ∎

C.2. Proof of Theorem 30

We now prove the main structural theorem.

Proof [Proof of Theorem 30] Fix $\tau = \varepsilon^{1000}$. By Theorem 29, there exists a function f' that satisfies $d(f, f') \leq O(\varepsilon^2)$ and is either a junta over x_H or is of the form in Equation (15). By Theorem 11, $d_{\text{Chow},S}(f, f') \leq d_{\text{Chow}}(f, f') \leq 2\sqrt{d(f, f')} = O(\varepsilon)$. If f' is a junta on H or $T \subseteq S$ then the set of functions g defined in Theorem 30 will be the same as the set of functions f' defined in Theorem 29, and therefore we get that $d_{\text{Chow},S}(f, g) \leq O(\varepsilon)$ for some function g defined in Theorem 30 as well.

It therefore remains to show that $d_{\text{Chow},S}(f,g) \leq O(\varepsilon)$ for some function g defined in Theorem 30 in the case where f is $\Omega(\varepsilon^2)$-far from any junta on H, and where $T \not\subseteq S$. Let f' again be a function of the form in Equation (15) that satisfies $d_{\text{Chow}}(f,f') \leq O(\varepsilon)$, and let g be the unique function of the form in Equation (16) with the same threshold θ' and same head weights v_H as f'. By the definition of partial Chow distance and the triangle inequality,

$$
\begin{aligned}
d_{\text{Chow},S}(f,g) &\leq d_{\text{Chow},H\cap S}(f,g) + d_{\text{Chow},T\cap S}(f,g) \\
&\leq d_{\text{Chow},H\cap S}(f,f') + d_{\text{Chow},H\cap S}(f',g) + d_{\text{Chow},T\cap S}(f,g) \\
&\leq d_{\text{Chow},H}(f,f') + d_{\text{Chow},H}(f',g) + d_{\text{Chow},T\cap S}(f,g). \quad (20)
\end{aligned}
$$

We will upper bound each of the terms on the right-hand side of Equation (20). For the first term, using Theorem 29 and Theorem 11, we have that

$$
d_{\text{Chow},H}(f,f') \leq d_{\text{Chow}}(f,f') \leq O(\varepsilon) . \quad (21)
$$

For the second term, let v_T and v'_T denote the tail weight vectors of f' and g, respectively. Using the $O(\tau^{143/72})$-regularity of v_T (Claim 32), the $O(\tau^{71/144})$-regularity of v'_T (Claim 33), and the fact that $\|v_T\| = \|v'_T\| = 1$, we get by Corollary 28 (recalling the setting of τ in terms of ε given in the statement of Theorem 30)

$$
d_{\text{Chow},H}(f',g) \leq O(\tau^{71/144}) \leq O(\varepsilon) . \quad (22)
$$

Finally we upper bound $d_{\text{Chow},T\cap S}(f,g)$. Let $\alpha = \alpha(-v_0, v_T)$ be the constant of proportionality defined in Proposition 24. Then

$$
\begin{aligned}
d_{\text{Chow},T\cap S}(f,g) &\leq \sqrt{\sum_{i\in T\cap S}\left(\hat{g}(i) - \frac{\alpha}{\gamma'}\hat{f}(i)\right)^2} + \left|1 - \frac{\alpha}{\gamma'}\right|\sqrt{\sum_{i\in T\cap S}\hat{f}(i)^2} \\
&= \sqrt{\sum_{i\in T\cap S}\left(\hat{g}(i) - \alpha\cdot v'_i\right)^2} + \left|\frac{\gamma'-\alpha}{\gamma'}\right|\sqrt{\sum_{i\in T\cap S}\hat{f}(i)^2} \\
&\leq O(\tau^{71/1728}) + \left|\frac{\gamma'-\alpha}{\gamma'}\right|\sqrt{\sum_{i\in T\cap S}\hat{f}(i)^2} \\
&\leq O(\tau^{71/1728}) + O\left(\frac{\tau^{1/12}}{\gamma'}\right)\sqrt{\sum_{i\in T\cap S}\hat{f}(i)^2} \\
&\leq O(\tau^{71/1728}) + O(\tau^{5/72}) \\
&\leq O(\varepsilon) .
\end{aligned} \quad (23)
$$

The first inequality is the triangle inequality; the equality follows by the definition of the weights in v'_T for $i \in T \cap S$ as $v'_i = \hat{f}(i)/\gamma'$; the second inequality uses the $O(\tau^{71/144})$-regularity of v'_T to apply Proposition 24; the third inequality holds by the triangle inequality since $\gamma' \overset{\tau}{\approx} \gamma$ and $\gamma \overset{\tau^{1/4}}{\approx} \alpha$ (the former approximation holds by assumption and the latter by Corollary 25); the fourth inequality holds since $\gamma' \geq \gamma \geq \tau^{1/72}$ by Fact 31 and $\sum_{i\in T\cap S}\hat{f}(i)^2 \leq \gamma \leq 1$ (by Parseval's Theorem), and the last inequality uses the setting of τ as a function of ε given in the statement of Theorem 30.

The theorem follows by upper bounding the terms in the right-hand side of Equation (20) using Equation (21), Equation (22), and Equation (23). ∎

Appendix D. The Partial Shapley Indices Problem

In this section we give a quasi-polynomial time algorithm for the Partial Shapley Indices problem by proving the following theorem, which is our second main result. (See Section A.4 for the definition of "η-restricted," and recall from Table 1 that $\mathring{f}(i)$ is the i-th Shapley value of f.)

Theorem 34 *Let* $f(x) = \text{sign}(\ell(x))$ *be an η-restricted LTF where* $\ell(x) = \displaystyle\sum_{i=1}^{n} v_i x_i - t$ *is a linear form with* $v_1, \ldots, v_n \geq 0$ *and η is an absolute constant in* $(1/4, 1]$. *There is an algorithm that, on input* $\{(i, \mathring{f}(i)) : i \in S\}$ *for some* $S \subseteq [n]$ *and a desired accuracy parameter ε satisfying* $\varepsilon \geq 1/n^{1/12}$, *with high probability outputs a weights-based representation of an LTF g that satisfies* $d_{Shapley,S}(f, g) \leq O(\varepsilon)$ *and runs in* $2^{\tilde{O}(\log^{18} n/\varepsilon^{24})}$ *time*

At the highest level the proof is by a case analysis. In Section D.1 we first establish (Theorem 40) a preliminary structural result showing that the target LTF f is closely approximated by an LTF f' with "well-structured" (discretized) weights; having such weights is useful for our algorithm and analysis. We then proceed by a case analysis based on the τ^*-critical index of the approximating LTF f', where

$$\tau^* := \frac{\varepsilon^2}{(\log n)^4} \tag{24}$$

There are two cases: The first case, which corresponds to case 1 of Theorem 55, is that the τ^*-critical index of f' is "large," more precisely at least

$$k^* := \max\left\{\frac{4 \log n}{\tau^2}, \frac{1}{\varepsilon^{12}}\right\}. \tag{25}$$

The algorithm for this case is a relatively straightforward enumeration over candidate junta LTFs. The second and more involved case, which corresponds to case 2 of Theorem 55, is that the τ^*-critical index of f' is between 0 and k^*. This case has an analysis which incorporates ingredients from the aforementioned structural results of both Section B.4.2 and Section D.2 (the latter of which in turn relies on technical results on approximating $\mathcal{D}_{\text{Shap}}$ by a mixture of p-biased product distributions which are given in Section D.1.3), and the algorithm for this case uses dynamic programming. We refer the reader to Section 5 for further high-level description.

D.1. Useful facts for Shapley indices

In this subsection we establish some tools which will be used for the proof of Theorem 34.

D.1.1. BACKGROUND RESULTS

We recall several useful facts about LTFs and Shapley indices, starting with the definition of the Shapley indices:

$$\mathring{f}(i) := \mathop{\mathbf{E}}_{\boldsymbol{\pi} \sim \mathbb{S}_n} [f(x^+(\boldsymbol{\pi}, i)) - f(x(\boldsymbol{\pi}, i))]. \tag{26}$$

We begin with a useful and elementary observation which shows that larger weights in an LTF correspond to larger Shapley indices; the proof is given in Appendix H.

Lemma 35 *Let $f(x) = \text{sign}(\ell(x))$ be an LTF where $\ell(x) = \sum_{i=1}^{n} w_i x_i - \theta$ is a linear form with $w_1, \ldots, w_n \geq 0$. Then for all $i \neq j \in [n]$, it holds that if $w_i \geq w_j$ then $\mathring{f}(i) \geq \mathring{f}(j)$.*

We continue by recalling a theorem about the anti-concentration of measure under the Shapley distribution $\mathcal{D}_{\text{Shap}}$ and some bounds for the Shapley distance between two function f and f'. Recall that an LTF $f : \{-1, 1\}^n \to \{-1, 1\}$, $f(x) = \text{sign}(w \cdot x - \theta)$ with $w \in \mathbb{R}^n$ is said to be η-*restricted* if $\theta \in [-(1-\eta)\|w\|_1, (1-\eta)\|w\|_1]$.

Theorem 36 ((De et al., 2017, Theorem 15)) *Let $\ell(x) = \sum_{i=1}^{n} v_i x_i - \theta'$ be a monotone non-decreasing, η-restricted affine form where η is an absolute constant in $(1/4, 1]$, so $v_i \geq 0$ for $i \in [n]$ and $|\theta'| \leq (1-\eta) \cdot \sum_{i=1}^{n} |v_i|$. Let $12 \leq k \leq n$, and let $r \in \mathbb{R}^+$ be such that $|S| \geq k$ where $S = \{i \in [n] : |v_i| \geq r\}$. Then*

$$\Pr_{\boldsymbol{x} \sim \mathcal{D}_{\text{Shap}}} [|\ell(\boldsymbol{x})| < r] = O\left(\frac{1}{\log n} \cdot \frac{1}{k^{1/6}}\right).$$

We require the following results from De et al. (2017) relating Shapley distance, Shapley Fourier distance, and the Shapley distribution:

Lemma 37 ((De et al., 2017, Lemma 11)) *Let $f, g : \{-1, 1\}^n \to [-1, 1]$ be LTFs. Then*

$$d_{\text{Shapley}}(f, g) \leq \sqrt{2H_{n-1}} \cdot d_{\text{Shapley-Fourier}}(f, g) + \frac{4}{\sqrt{n}},$$

where $H_k = \Theta(\log k)$ is the k-th harmonic number.

Lemma 38 ((De et al., 2017, Special case of Fact 7)) *Let $f, g : \{-1, 1\}^n \to \mathbb{R}$ be LTFs. Then*

$$d_{\text{Shapley-Fourier}}(f, g) \leq 2\sqrt{\Pr_{\boldsymbol{x} \sim \mathcal{D}_{\text{Shap}}} [f(\boldsymbol{x}) \neq g(\boldsymbol{x})]}.$$

By combining Lemma 37 and Lemma 38, we get the following.

Corollary 39 *Let $f, g : \{-1, 1\}^n \to [-1, 1]$ be LTFs. Then*

$$d_{\text{Shapley}}(f, g) \leq O\left(\sqrt{\log n \cdot \Pr_{\boldsymbol{x} \sim \mathcal{D}_{\text{Shap}}} [f(\boldsymbol{x}) \neq g(\boldsymbol{x})]} + \frac{1}{\sqrt{n}}\right).$$

D.1.2. A DISCRETIZATION LEMMA

As described earlier, we will perform our case analysis on an LTF f' which approximates the target LTF f (with respect to Shapley distance) and whose weights (after a suitable rescaling) are integers that are not too large. The following theorem provides the necessary structural result ensuring the existence of such an approximation.

Theorem 40 *Let $\varepsilon \in (\frac{1}{n^{1/12}}, \frac{1}{2})$ and let $f : \{-1,1\}^n \to \{-1,1\}$ be a monotone increasing, η-restricted LTF where η is an absolute constant in $(1/4, 1]$. Then there exists an LTF $f'(x) = \mathrm{sign}(\sum_{i=1}^n w_i \cdot x_i - \theta)$ where $\theta, w_1, \ldots, w_n$ are integer multiples of $1/(n^2 \cdot k^{k/2})$ for $k = 1/\varepsilon^{12}$, with $\theta, w_1, \ldots, w_n \in [0,1]$ and $\max_{i \in [n]} w_i \geq 1/2$, such that $d_{\mathrm{Shapley}}(f, f') \leq O(\varepsilon)$.*

Before giving the proof, to motivate the first structural lemma we will use, consider a linear form $\ell(x) = v \cdot x - \theta'$ and its corresponding LTF $f(x) = \mathrm{sign}(\ell(x))$. We note that given a probability distribution χ over $\{-1,1\}^n$,

- If $\mathbf{Pr}_{x \sim \chi}[|\ell(x)| \text{ is small}]$ is large, then for $\ell'(x) = v' \cdot x - \theta'$ a slight perturbation of $\ell(x)$, the corresponding LTF $f(x) = \mathrm{sign}(\ell'(x))$ could be far from g with respect to χ, i.e. $\mathbf{Pr}_{x \sim \chi}[f(x) \neq f'(x)]$ could possibly be large.

- On the other hand, if $\mathbf{Pr}_{x \sim \chi}[|\ell(x)| \text{ is small}]$ is small, then any slight perturbation $\ell'(x)$ will be such that for the corresponding perturbed LTF f, the probability $\mathbf{Pr}_{x \sim \chi}[f(x) \neq f'(x)]$ must be small.

The following lemma formalizes the above observations; for completeness we give its simple proof.

Lemma 41 *Let $\ell(x), \ell'(x)$ be two linear forms such that $|\ell(x) - \ell'(x)| \leq \varepsilon$ for all $x \in \{-1,1\}^n$. Suppose that ℓ satisfies $\mathbf{Pr}_{x \sim \mathcal{D}_{\mathrm{Shap}}}[|\ell(x)| \leq \varepsilon] \leq \delta$. Then it holds that $\mathbf{Pr}_{x \sim \mathcal{D}_{\mathrm{Shap}}}[f(x) \neq f'(x)] \leq \delta$, where $f(x) = \mathrm{sign}(\ell(x))$ and $f'(x) = \mathrm{sign}(\ell'(x))$.*

Proof

$$
\begin{aligned}
\mathbf{Pr}_{x \sim \mathcal{D}_{\mathrm{Shap}}}[f(x) \neq f'(x)] &= \mathbf{Pr}_{x \sim \mathcal{D}_{\mathrm{Shap}}}[\mathrm{sign}(\ell(x)) \neq \mathrm{sign}(\ell'(x))] \\
&\leq \mathbf{Pr}_{x \sim \mathcal{D}_{\mathrm{Shap}}}[|\ell(x)| < |\ell(x) - \ell'(x)|] \\
&\leq \mathbf{Pr}_{x \sim \mathcal{D}_{\mathrm{Shap}}}[|\ell(x)| < \varepsilon] \leq \delta \,.
\end{aligned}
$$

∎

Theorem 36 provides the desired upper bound on the probability that $x \sim \mathcal{D}_{\mathrm{Shap}}$ has $|\ell(x)|$ being "too small," but to apply it we need to ensure that "many" weights w_i are "not too small." This is ensured by the following lemma:

Theorem 42 ((De et al., 2017, Theorem 3)) *Let $g : \{-1,1\}^n \to \{-1,1\}$ be an η-restricted LTF where $\eta = \Theta(1)$, and let $k \in [2, n]$. There exists a representation of g as $g(x) = \mathrm{sign}(\sum_{i=1}^n v_i x_i - \theta)$ such that (after reordering coordinates so that condition (i) below holds) we have: (i) $|v_i| \geq |v_{i+1}|$, $i \in [n-1]$; (ii) $|\theta| \leq (1 - \eta) \sum_{i=1}^n |v_i|$; and (iii) for all $i \in [0, k-1]$ we have $|v_i| \leq (2/\eta) \cdot \sqrt{n} \cdot k^{\frac{k}{2}} \cdot \sigma_k$, where $\sigma_k := \sqrt{\sum_{j \geq k} v_j^2}$.*

Rescaling the weights so that the largest weight has magnitude 1, Theorem 42 easily yields the following corollary:

Corollary 43 *Let* $g : \{-1,1\}^n \to \{-1,1\}$ *be an* η*-restricted LTF where* $\eta = \Theta(1)$*, and let* $12 \le k \le n$*. Then* g *has a representation as* $g(x) = sign(v \cdot x - \theta)$ *where the largest-magnitude weight has magnitude 1, the* $\Omega(k)$ *many largest-magnitude weights each have magnitude at least* $r := \frac{1}{n \cdot k^{k/2}}$*, and* $|\theta| \le (1 - \eta) \sum_{i=1}^n |v_i|$*.*

Now we can give the proof of Theorem 40.

Proof [Proof of Theorem 40] Recall that $f(x) = \text{sign}(v \cdot x - \theta')$. Let $k = 1/\varepsilon^{12}$ (as in the statement of Theorem 40). Applying Corollary 43, we may express $f(x)$ as $\text{sign}(\ell(x))$, where $\ell(x) = v \cdot x - \theta$ and the weight vector v satisfies the properties stated in that corollary. Since $\ell(x)$ is guaranteed to have "many" weights that are "not too small" we may apply Theorem 36 to it, and we get that

$$\Pr_{\boldsymbol{x} \sim \mathcal{D}_{\text{Shap}}}[|\ell(\boldsymbol{x})| < r] = O\Big(\frac{1}{\log n} \cdot \frac{1}{k^{1/6}}\Big),$$

where $r = \frac{1}{n \cdot k^{k/2}}$.

Now for each $i \in [n]$ we define a rounded version w_i of the weight v_i which is obtained by rounding it to the closest integer multiple of $\frac{1}{n^2 k^{k/2}}$, and we let $\ell'(x)$ be the linear form $w \cdot x - \theta$. It is immediate that for all $x \in \{-1,1\}^n$ we have $|\ell(x) - \ell'(x)| \le \frac{1}{n k^{k/2}} = r$, and that $\max_{i \in [n]} w_i \ge 1/2$. Letting $f'(x) = \text{sign}(\ell'(x))$, by Lemma 41 we have that

$$\Pr_{\boldsymbol{x} \sim \mathcal{D}_{\text{Shap}}}[f(\boldsymbol{x}) \ne f'(\boldsymbol{x})] \le O\Big(\frac{1}{\log n} \cdot \frac{1}{k^{1/6}}\Big).$$

Finally, applying Corollary 39, we get that

$$d_{\text{Shapley}}(f, f') \le O\Big(\sqrt{\log n \cdot \Pr_{\boldsymbol{x} \sim \mathcal{D}_{\text{Shap}}}[f(\boldsymbol{x}) \ne f'(\boldsymbol{x})]} + \frac{1}{\sqrt{n}}\Big) \le O(\varepsilon)$$

as was to be shown. ∎

D.1.3. APPROXIMATING $\mathcal{D}_{\text{Shap}}$ BY A MIXTURE OF p-BIASED PRODUCT DISTRIBUTIONS

We will use the following lemma from De et al. (2017) in order to express the Shapley indices in terms of the coordinate correlation coefficients:

Lemma 44 ((De et al., 2017, Lemma 11)) *For* $f : \{-1,1\}^n \to \{-1,1\}$ *any monotone function, for each* $i = 1, \ldots, n$ *we have*

$$\mathring{f}(i) = \frac{f(1^n) - f((-1)^n)}{n} + \frac{\Lambda(n)}{2} \cdot \left(f^*(i) - \frac{1}{n} \sum_{j=1}^n f^*(j)\right),$$

where

$$f^*(i) = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\text{Shap}}}[f(\boldsymbol{x})\boldsymbol{x}_i].$$

In Appendix B.4 we proved two structural results, Proposition 24 ("tail weights are proportional to tail Chow parameters") and Theorem 27 ("exchanging a regular tail vector for another regular tail vector with the same ℓ_1 and ℓ_2 norm doesn't change the head Chow parameters by much"), for general p-biased input distributions. To analyze our algorithm for the Partial Chow Parameters Problem we only needed the $p = 1/2$ case of these results, but now we will use those structural results in their full generality.

Lemma 44 shows that the Shapley indices are closely related to the distribution $\mathcal{D}_{\text{Shap}}$. Towards the goal of employing the results of Appendix B.4 for the Shapley problem, ideally we would like to define the Shapley distribution $\mathcal{D}_{\text{Shap}}$ as a mixture of p-biased product distributions u_p^n. (As a sanity check of the feasibility of doing this, we note that $\mathcal{D}_{\text{Shap}}$ and each u_p^n are all *exchangeable* distributions: for any one of these distributions, the probability weight assigned to an n-bit string depends only on the number of 1s in the string.) Recall that the Shapley distribution $\mathcal{D}_{\text{Shap}}$ is defined as follows: it puts zero weight on the strings 1^n and $(-1)^n$, and for every other $x \in \{-1, 1\}^n$, it assigns weight

$$\Pr_{\boldsymbol{x} \sim \mathcal{D}_{\text{Shap}}}[\boldsymbol{x} = x] = \frac{\frac{1}{i} + \frac{1}{n-i}}{\Lambda(n)\binom{n}{i}}, \qquad \text{where } i = \text{weight}(x) := |\{k \in [n] : x_k = 1\}|.$$

How can we draw $\boldsymbol{x} \sim \mathcal{D}_{\text{Shap}}$ via a random procedure that uses the p-biased product distributions u_p^n? Towards answering this question, we observe that for $i \in \{1, \cdots, n-1\}$, routine calculus yields that

$$\int_0^1 p^i(1-p)^{n-i} \frac{(\frac{1}{p} + \frac{1}{1-p})}{\Lambda(n)} dp = \frac{\frac{1}{i} + \frac{1}{n-i}}{\Lambda(n)\binom{n}{i}}.$$

Therefore $\mathcal{D}_{\text{Shap}}$ can be alternatively defined as follows:

$$\Pr_{\boldsymbol{x} \sim \mathcal{D}_{\text{Shap}}}[\boldsymbol{x} = x] = \mathbb{1}\{i \notin \{0, n\}\} \cdot \int_0^1 \frac{(\frac{1}{p} + \frac{1}{1-p})}{\Lambda(n)} p^i(1-p)^{n-i} dp, \text{ where } i = \text{weight}(x).$$

This leads to the following natural first attempt to define a new sampling mechanism for making a draw from the Shapley distribution $\mathcal{D}_{\text{Shap}}$ (where we write $\{-1, 1\}^n_{=k}$ to denote $\{x \in \{-1, 1\}^n : \text{weight}(x) = k\}$):

ORIGINAL SHAPLEY DISTRIBUTION SAMPLING MECHANISM:

- Sample layer $\boldsymbol{k} \in \{1, \cdots, n-1\}$ with probability proportional to $\frac{1}{k} + \frac{1}{n-k}$.

- Then sample a uniformly random point from $\{-1, 1\}^n_{=\boldsymbol{k}}$.

$\Rightarrow$

FIRST ATTEMPT AT NEW SHAPLEY DISTRIBUTION SAMPLING MECHANISM:

1. Sample $\boldsymbol{p} \in (0, 1)$ with probability $\mathcal{K}(p)$ proportional to $\frac{(\frac{1}{p} + \frac{1}{1-p})}{\Lambda(n)}$.

2. Then sample layer $\boldsymbol{k} \in \{1, \cdots, n-1\}$ with probability proportional to $\mathcal{K}(\boldsymbol{p})\binom{n}{k}\boldsymbol{p}^k(1-\boldsymbol{p})^{n-k} = \frac{1}{k} + \frac{1}{n-k}$.

3. Finally sample a uniformly random point from $\{-1, 1\}^n_{=\boldsymbol{k}}$.

Unfortunately, there is a crucial flaw in the above new hoped-for sampling mechanism. The flaw is in Step (1): a trivial verification shows that $\int_0^1 \frac{(\frac{1}{p} + \frac{1}{1-p})}{\Lambda(n)} dp = \infty$, and so it is not possible to actually sample $\boldsymbol{p}$ as described in that step.

We get around this challenge by restricting the sampling space in Step (1) above to $[\delta, 1 - \delta]$ instead of $(0, 1)$ where as we will see soon, we take δ to be a very small value (a value which is $1/\mathrm{poly}(n)$ and at most $o(1/n)$). Thus, it is natural for us to consider the continuous probability distribution $\mathcal{K}(\delta)$ supported on $[\delta, 1 - \delta]$, which is defined as follows:

Definition 45 ($\mathcal{K}(\delta)$-distribution) *A random variable $\boldsymbol{p}$ is $\mathcal{K}(\delta)$-distributed if its density is given by $f_{\mathcal{K}(\delta)}(p) = C_\delta \frac{(\frac{1}{p} + \frac{1}{1-p})}{\Lambda(n)}$ for any $p \in [\delta, 1-\delta]$, where $C_\delta := \left(\int_\delta^{1-\delta} \frac{(\frac{1}{p} + \frac{1}{1-p})}{\Lambda(n)} dp \right)^{-1} = \frac{\Lambda(n)}{2\ln(\delta^{-1} - 1)}$. Notice that if δ is inverse polynomial in n then $C_\delta = \Theta(1)$.*

Following the above first attempt at a new Shapley Distribution sampling procedure, we can also define $\mathcal{Q}(\delta)$, a continuous mixture of u_p^n distributions, defined as follows:

Definition 46 ($\mathcal{Q}(\delta)$-distribution) *The $\mathcal{Q}(\delta)$ distribution is supported in $\{-1, 1\}^n$ and is defined as $\mathbf{Pr}_{\boldsymbol{x} \sim \mathcal{Q}(\delta)}[\boldsymbol{x} = x] = \int_\delta^{1-\delta} f_{\mathcal{K}(\delta)}(p) \, \mathbf{Pr}_{\boldsymbol{x} \sim u_p^n}[\boldsymbol{x} = x] dp$.*

Using the above definitions, we establish a useful approximation for the expectation of Boolean functions over $\mathcal{D}_{\mathrm{Shap}}$ in terms of the $\mathcal{Q}(\delta)$ distribution:

Lemma 47 *Let $\delta = 1/n^c$ for some constant $c > 1$ and let $f : \{-1, 1\}^n \to \mathbb{R}$ be such that $\|f\|_\infty \leq O(1)$ and $f(-1)^n = f(1^n) = 0$. Then we have that*

$$\left| \underset{\boldsymbol{x} \sim \mathcal{D}_{\mathrm{Shap}}}{\mathbf{E}} [f(\boldsymbol{x})] - \frac{1}{C_\delta} \underset{\boldsymbol{x} \sim \mathcal{Q}(\delta)}{\mathbf{E}} [f(\boldsymbol{x})] \right| \leq \frac{O(n\delta)}{\Lambda(n)}.$$

Combining Lemma 44 and Lemma 47 and the definition of the $\mathcal{Q}(\delta)$ distribution (which implies that $\mathbf{E}_{\boldsymbol{x} \sim \mathcal{Q}(\delta)}[f(\boldsymbol{x})] = \mathbf{E}_{\boldsymbol{p} \sim \mathcal{K}(\delta)}[\mathbf{E}_{\boldsymbol{x} \sim u_p^n}[f(\boldsymbol{x})]]$), an immediate consequence is the following approximation of Shapley indices which will be the starting point for various structural lemmas in later sections:

Lemma 48 *Let f be any nontrivial monotone LTF (so $f((-1)^n) = -1$ and $f(1^n) = 1$). Then for $\delta = 1/n^c$ for some constant $c > 1$, for each $i \in [n]$, the value $\mathring{f}(i)$ is additively $O(n\delta)$-close to the quantity υ_i defined below:*

$$\mathring{f}(i) \overset{O(n\delta)}{\approx} \frac{2}{n} + \frac{\Lambda(n)}{2} \cdot \frac{1}{C_\delta} \cdot \underset{\boldsymbol{p} \sim \mathcal{K}(\delta)}{\mathbf{E}} \left[f^*(i, \boldsymbol{p}) - \frac{1}{n} \sum_{j=1}^n f^*(j, \boldsymbol{p}) \right] =: \upsilon_i$$

$$= \frac{2}{n} + \frac{\Lambda(n)}{2} \cdot \frac{1}{C_\delta} \cdot \underset{\boldsymbol{p} \sim \mathcal{K}(\delta)}{\mathbf{E}} \left[\sigma_{\boldsymbol{p}} \hat{f}(i, \boldsymbol{p}) - \frac{1}{n} \sum_{j=1}^n \sigma_{\boldsymbol{p}} \hat{f}(j, \boldsymbol{p}) \right].$$

The proofs of Lemma 47 and Lemma 48 are a sequence of routine calculations and are given in Appendix I.

D.1.4. ESTIMATING SHAPLEY INDICES

For completeness we close this subsection with a quick description of a simple sampling-based scheme to approximate the Shapley indices of a given monotone LTF.

Proposition 49 *There is a procedure* ESTIMATESHAPLEY *with the following properties: The procedure is given oracle access to a monotone LTF* $f : \{-1,1\}^n \to \{-1,1\}$, *a desired accuracy parameter* γ, *and a desired failure probability* δ_{fail}. *The procedure makes* $O(n \log(n/\delta_{\text{fail}})/\gamma^2)$ *oracle calls to* f *and runs in time* $O(n^2 \log(n/\delta_{\text{fail}})/\gamma^2)$ *(counting each oracle call to* f *as taking one time step). With probability* $1 - \delta_{\text{fail}}$ *it outputs a list of numbers* $\tilde{a}(1), \ldots, \tilde{a}(n)$ *such that*

$$\sum_{i \in [n]} \left(\tilde{a}(i) - \mathring{f}(i) \right)^2 \leq \gamma.$$

Proof The procedure empirically estimates each $\mathring{f}(j)$, $j = 1, \ldots, n$, to additive accuracy $\gamma/\sqrt{n}$ using the definition of Shapley indices, Equation (26). This is done by generating a uniform random $\boldsymbol{\pi} \sim \mathbb{S}_n$ and then, for each $i = 1, \ldots, n$, constructing the two inputs $x^+(\boldsymbol{\pi}, i)$ and $x(\boldsymbol{\pi}, i)$ and calling the oracle for f twice to compute $f(x^+(\pi, i)) - f(x(\pi, i))$. Since $|f(x^+(\boldsymbol{\pi}, i)) - f(x(\boldsymbol{\pi}, i))| \leq 2$ always, a straightforward application of Hoeffding bounds gives that a sample of $m = O(n \log(n/\delta_{\text{fail}})/\gamma^2)$ permutations suffices to estimate all the $\mathring{f}(i)$ values to additive accuracy $\pm\gamma/\sqrt{n}$ with total failure probability at most δ_{fail}. If each estimate $\tilde{a}(i)$ is additively accurate to within $\pm\gamma/\sqrt{n}$, then $d_{\text{Shapley}}(a, f) \leq \gamma$ as desired. ∎

D.2. Structural results on heads and tails of LTFs (Shapley version)

Working in the same fashion as in Appendix B.4, let $f(x) = \text{sign}(w \cdot x - \theta)$ be an LTF, and to simplify presentation let us assume that its weights are sorted in magnitude from largest to smallest, i.e., $|w_1| \geq |w_2| \geq \cdots \geq |w_n|$. Let $\tau > 0$. Although f need not be τ-regular, we can always partition its weights w into "head weights" w_H and "tail weights" w_T such that w_T is τ-regular and any longer suffix of w is not τ-regular. Let $H = \{1, 2, \ldots, \}$ be the set of indices of head weights, and let $T = [n] \setminus H = \{\ldots, n-1, n\}$ be the set of indices of tail weights.

D.2.1. REGULAR TAIL WEIGHTS ARE APPROXIMATELY AFFINELY RELATED TO TAIL SHAPLEY INDICES

We first show that the vector of tail weights w_T is approximately affinely related to the vector of tail Shapley indices $(\mathring{f}(i))_{i \in T}$; more precisely, there exist real values $\mathring{A} = \mathring{A}(w_H, \|w_T\|_1, \|w_T\|_2, \theta, \delta)$, $\mathring{B} = \mathring{B}(w_H, \|w_T\|_1, \|w_T\|_2, \theta, \delta)$ such that $\mathring{f}(i) \approx \mathring{A} \cdot w_i + \mathring{B}$ for all $i \in T$. This characterization will be helpful for recovering the tail weights of an LTF from the Shapley indices, and as a corollary also gives an approximation of the sum of Shapley indices of f on T.

Theorem 50 *Let* $f(x) = f(x_H, x_T) = \text{sign}(w_H \cdot x_H + w_T \cdot x_T - \theta)$ *be an LTF satisfying* $f(1^n) = 1, f((-1)^n) = -1$ *where* w_T *is* τ-*regular and* $w_1, \ldots, w_n \geq 0$. *There exist real values* $\mathring{A}(w_H, \|w_T\|_1, \|w_T\|_2, \theta, \delta), \mathring{B}(w_H, \|w_T\|_1, \|w_T\|_2, \theta, \delta)$ *such that for* $\delta = 1/n^c$ *for some constant* $c > 1$,

$$\sum_{i \in T} \left(\mathring{f}(i) - (\mathring{A}(w_H, \|w_T\|_1, \|w_T\|_2, \theta, \delta)w_i + \mathring{B}(w_H, \|w_T\|_1, \|w_T\|_2, \theta, \delta) \right)^2 \leq O\left(n^3 \delta^2 + \ln^2(\delta^{-1}) \cdot \left(|H|/n + \sqrt{\tau} \right) \right)$$

43

Initially, we will prove a simplified version of our theorem asserting the extra assumption $\|w_T\|_2 = 1$.

Lemma 51 *Let $f(x) = f(x_H, x_T) = \text{sign}(w_H \cdot x_H + w_T \cdot x_T - \theta)$ be an LTF satisfying $f(1^n) = 1, f((-1)^n) = -1$ where w_T is τ-regular and $\|w_T\|_2 = 1$, and $w_1, \ldots, w_n \geq 0$. There exist real values $\mathring{\Gamma}(w_H, \|w_T\|_1, \theta, \delta), \mathring{\Delta}(w_H, \|w_T\|_1, \theta, \delta)$ such that for $\delta = 1/n^c$ for some constant $c > 1$,*

$$\sum_{i \in T} \left(\mathring{f}(i) - (\mathring{\Gamma}(w_H, \|w_T\|_1, \theta, \delta) w_i + \mathring{\Delta}(w_H, \|w_T\|_1, \theta, \delta) \right)^2 \leq O\left(n^3 \delta^2 + \ln^2(\delta^{-1}) \cdot \left(|H|/n + \sqrt{\tau} \right) \right).$$

Proof

By Lemma 48, we have that

$$\mathring{f}(i) \overset{O(n\delta)}{\approx} \frac{2}{n} + \frac{\Lambda(n)}{2} \cdot \frac{1}{C_\delta} \cdot \underset{\boldsymbol{p} \sim \mathcal{K}(\delta)}{\mathbf{E}} \left[\sigma_{\boldsymbol{p}} \left(\hat{f}(i, \boldsymbol{p}) - \frac{1}{n} \sum_{j=1}^{n} \hat{f}(j, \boldsymbol{p}) \right) \right] = \upsilon_i.$$

Squaring and summing this difference over all $i \in T$, it follows that we have

$$\sum_{i \in T} \left(\mathring{f}(i) - \upsilon_i \right)^2 \leq O(n^3 \delta^2). \tag{27}$$

We now define the quantities $\vartheta_i, \varpi_i, \varrho_i$: The quantity ϑ_i is the same as υ_i, but with the summation over all of $[n]$ inside the expectation operator being instead a sum over all of T. The second is the approximation of ϑ_i that results from using the affine transformation of the weights of the linear form (recall Proposition 24) instead of the actual p-biased Fourier coefficients $\hat{f}(\cdot, p)$. More precisely,

$$\vartheta_i := \frac{2}{n} + \frac{\Lambda(n)}{2} \cdot \frac{1}{C_\delta} \cdot \underset{\boldsymbol{p} \sim \mathcal{K}(\delta)}{\mathbf{E}} \left[\sigma_{\boldsymbol{p}} \left(\hat{f}(i, \boldsymbol{p}) - \frac{1}{n} \sum_{j \in T} \hat{f}(j, \boldsymbol{p}) \right) \right],$$

$$\varpi_i := \frac{2}{n} + \frac{\Lambda(n)}{2} \cdot \frac{1}{C_\delta} \cdot \underset{\boldsymbol{p} \sim \mathcal{K}(\delta)}{\mathbf{E}} \left[\sigma_{\boldsymbol{p}} \alpha(\theta, w_H, w_T, \boldsymbol{p}) \left(w_i - \frac{\sum_{k \in T} w_k}{n} \right) \right].$$

To bound the error incurred by using ϖ_i instead of υ_i, we observe that

$$\sum_{i \in T} (\upsilon_i - \varpi_i)^2 = \frac{\Lambda(n)^2}{4 C_\delta^2} \cdot \sum_{i \in T} \left(\underset{\boldsymbol{p} \sim \mathcal{K}(\delta)}{\mathbf{E}} \left[\sigma_{\boldsymbol{p}} \alpha(\theta, w_H, w_T, \boldsymbol{p}) \left(w_i - \frac{\sum_{k \in T} w_k}{n} \right) - \sigma_{\boldsymbol{p}} \left(\hat{f}(i, \boldsymbol{p}) - \frac{1}{n} \sum_{i \in [n]} \hat{f}(i, \boldsymbol{p}) \right) \right] \right)^2.$$

Since $\delta = 1/n^c$ for some constant $c > 1$, it is easy to see that $\frac{\Lambda(n)^2}{4 C_\delta^2} \leq O\left(\ln^2(\delta^{-1}) \right)$. Using Jensen's inequality and linearity of expectation, we get that

$$\sum_{i \in T} (\upsilon_i - \varpi_i)^2$$

$$\leq O\left(\ln^2(\delta^{-1}) \right) \cdot \underset{\boldsymbol{p} \sim \mathcal{K}(\delta)}{\mathbf{E}} \left[\sigma_{\boldsymbol{p}}^2 \sum_{i \in T} \left(\alpha(\theta, w_H, w_T, \boldsymbol{p}) \left(w_i - \frac{\sum_{k \in T} w_k}{n} \right) - \left(\hat{f}(i, \boldsymbol{p}) - \frac{1}{n} \sum_{i \in [n]} \hat{f}(i, \boldsymbol{p}) \right) \right)^2 \right]. \tag{28}$$

By applying Proposition 24 to f we get that

$$\sum_{i \in T} (\hat{f}(i, p) - \alpha(\theta, w_H, w_T, p) \cdot w_i)^2 \leq O\left(\sqrt{\frac{\tau}{\sigma_p^2}}\right) \tag{a}$$

holds for each $p \in (0, 1)$. Next by applying for any p the first claim of Fact 60 to (a) with the rescaling factor "$c = \frac{|T|}{n}$", we have

$$\sum_{i \in T} \left(\frac{1}{n} \sum_{j \in T} \hat{f}(j, p) - \alpha(\theta, w_H, w_T, p) \cdot \frac{1}{n} \sum_{j \in T} w_j\right)^2 \leq O\left(\sqrt{\frac{\tau}{\sigma_p^2}}\right). \tag{b}$$

Since the sum of all squared p-biased Fourier coefficients is at most 1, by Cauchy-Schwarz we have that

$$\sum_{i \in T} \left(\frac{1}{n} \sum_{j \in H} \hat{f}(j, p)\right)^2 \leq O\left(\frac{|H| \cdot |T|}{n^2}\right) \leq O\left(|H|/n\right). \tag{c}$$

Using Fact 59 to add the last two inequalities, we get:

$$\sum_{i \in T} \left(\frac{1}{n} \sum_{j \in [n]} \hat{f}(j, p) - \alpha(\theta, w_H, w_T, p) \cdot \frac{1}{n} \sum_{j \in T} w_j\right)^2 \leq O\left(\sqrt{\frac{\tau}{\sigma_p^2}}\right) + |H|/n. \tag{d}$$

Multiplying (a), (d) by $\sigma_p^2 = O(1)$ and using Fact 59 to combine them we get:

$$\sum_{i \in T} \left(\sigma_p \alpha(\theta, w_H, w_T, p) \left(w_i - \frac{\sum_{k \in T} w_k}{n}\right) - \sigma_p \left(\hat{f}(i, p) - \frac{1}{n} \sum_{i \in [n]} \hat{f}(i, p)\right)\right)^2 \leq O\left(|H|/n + \sqrt{\tau}\right).$$

Observing that the RHS above has no dependence on p, we can plug this into Equation (28) and we get that

$$\sum_{i \in T} (v_i - \varpi_i)^2 \leq O\left(\ln^2(\delta^{-1})\right) \mathop{\mathbf{E}}_{p \sim \mathcal{K}(\delta)} \left[O\left(|H|/n\right) + O\left(\sqrt{\tau}\right)\right] \leq O\left(\ln^2(\delta^{-1}) \cdot \left(|H|/n + \sqrt{\tau}\right)\right). \tag{29}$$

Finally, combining the bounds from Equation (29) and Equation (27) using Fact 59, we get that

$$\sum_{i \in T} \left(\mathring{f}(i) - \varpi_i\right)^2 \leq O\left(n^3 \delta^2 + \ln^2(\delta^{-1}) \cdot \left(|H|/n + \sqrt{\tau}\right)\right).$$

To finish the proof it remains only to verify that ϖ_i can be written as $\mathring{\Gamma} w_i + \mathring{\Delta}$, where $\mathring{\Gamma} = \mathring{\Gamma}(w_H, \|w_T\|_1, \theta, \delta)$ and $\mathring{\Delta} = \mathring{\Delta}(w_H, \|w_T\|_1, \theta, \delta)$. This holds because

$$
\begin{aligned}
\varpi_i &= \frac{2}{n} + \frac{\Lambda(n)}{2} \cdot \int_\delta^{1-\delta} \frac{1/p + 1/(1-p)}{\Lambda(n)} \left[\sigma_p \alpha(\theta, w_H, w_T, p) \left(w_i - \frac{\sum_{k \in [n]} w_k}{n} \right) \right] dp \\
&= \underbrace{\left(\frac{1}{2} \int_\delta^{1-\delta} \sigma_p \cdot \alpha(\theta, w_H, w_T, p) \cdot \frac{1/p + 1/(1-p)}{n} \, dp \right)}_{\mathring{\Gamma}} w_i \\
&\quad + \underbrace{\left(\frac{2}{n} - \frac{1}{2} \int_\delta^{1-\delta} \sigma_p \cdot \alpha(\theta, w_H, w_T, p) \cdot \frac{1/p + 1/(1-p)}{n} \cdot \|w\|_1 \, dp \right)}_{\mathring{\Delta}}.
\end{aligned}
\tag{30}
$$

It is important to mention that by Definition 13, it holds that

$$
\alpha(\theta, w_H, w_T, p) := \mathbf{E}_{\boldsymbol{\rho} \sim u_p^{|H|}} [\alpha(\psi_p^{[w_T]}(\theta - w_H \cdot \boldsymbol{\rho}))]
$$

where $\psi_p^{[w]}(x) = \frac{x - \mu_p \cdot \|w\|_1}{\sigma_p \|w\|_2}$ for $w \in \mathbb{R}_{\geq 0}^n$. Therefore, the quantity $\alpha(\theta, w_H, w_T, p)$ depends actually only on $(\theta, w_H, \|w_T\|_1, p)$ since $\|w_T\|_2 = 1$. ∎

Proof [Proof of Theorem 50] Finally it is easy to see that we can relax the assumption of $\|w_T\|_1 = 1$ by setting

$$
\begin{cases}
\mathring{A}(w_H, \|w_T\|_1, \|w_T\|_2, \theta, \delta) := \frac{1}{\|w_T\|_2} \mathring{\Gamma}\left(\frac{w_H}{\|w_T\|_2}, \frac{\|w_T\|_1}{\|w_T\|_2}, \frac{\theta}{\|w_T\|_2}, \delta\right) \\
\mathring{B}(w_H, \|w_T\|_1, \|w_T\|_2, \theta, \delta) := \mathring{\Delta}\left(\frac{w_H}{\|w_T\|_2}, \frac{\|w_T\|_1}{\|w_T\|_2}, \frac{\theta}{\|w_T\|_2}, \delta\right)
\end{cases}
\tag{31}
$$

where $\mathring{\Gamma}, \mathring{\Delta}$ are the affine constants of Equation (30). ∎

In the special case in which the entire weight vector w is regular, we get the following:

Corollary 52 *Let* $f(x) = f(x_H, x_T) = \text{sign}(w_H \cdot x_H + w_T \cdot x_T - \theta)$ *be an LTF satisfying* $f(1^n) = 1, f((-1)^n) = -1$ *where w is τ-regular and $w_1, \ldots, w_n \geq 0$. There exist real values* $\mathring{A}(w_H, \|w_T\|_1, \|w_T\|_2, \theta, \delta), \mathring{B}(w_H, \|w_T\|_1, \|w_T\|_2, \theta, \delta)$ *such that for $\delta = 1/n^c$ for some constant $c > 1$,*

$$
\sum_{i \in T} \left(\mathring{f}(i) - (\mathring{A}(w_H, \|w_T\|_1, \|w_T\|_2, \theta, \delta) w_i + \mathring{B}(w_H, \|w_T\|_1, \|w_T\|_2, \theta, \delta) \right)^2 \leq O\left(n^3 \delta^2 + \ln^2(\delta^{-1}) \cdot \left(|H|/n + \sqrt{\tau} \right) \right)
$$

D.2.2. PRESERVING THE HEAD SHAPLEY INDICES

The last structural result we require on Shapley indices is an analogue of Theorem 27 for the head Shapley indices. More precisely, the following theorem shows that exchanging the tail weights w_T of an LTF f with other weights w_T' of the same ℓ_1 and ℓ_2 norm does not change the head Shapley coefficients $(\mathring{f}(i))_{i \in H}$ by too much.

Theorem 53 *Let $f(x) = f(x_H, x_T) = \text{sign}(w_H \cdot x_H + w_T \cdot x_T - \theta)$ and let $g(x) = f'(x_H, x_T) = \text{sign}(w_H \cdot x_H + w'_T \cdot x_T - \theta)$ where $(w_H, w_T) \in (\mathbb{R}^{\geq 0})^n$, and where w_T, w'_T are τ-regular, satisfy $\|w_T\|_1 = \|w'_T\|_1$, and satisfy $\|w_T\|_2 = \|w'_T\|_2$. Suppose that $f(1^n) = g(1^n) = 1, f((-1)^n) = g((-1)^n) = -1$. Then for $\delta = 1/n^c$ for some constant $c > 1$, we have that*

$$\sum_{i \in H} (\mathring{f}(i) - \mathring{g}(i))^2 \leq O\left(|H| n^2 \delta^2 + \Lambda^2(n)(\tau^2 + \frac{\sqrt{\tau}|H|}{n}) \right) .$$

Proof From Lemma 48, we have that each $i \in [n]$ satisfies

$$\mathring{f}(i) \overset{O(n\delta)}{\approx} \frac{2}{n} + \frac{\Lambda(n)}{2} \cdot \frac{1}{C_\delta} \cdot \underset{\boldsymbol{p} \sim \mathcal{K}(\delta)}{\mathbf{E}} \left[\sigma_{\boldsymbol{p}}(\hat{f}(i, \boldsymbol{p}) - \frac{1}{n} \sum_{j=1}^n \hat{f}(j, \boldsymbol{p})) \right] = v_i(f).$$

Applying the above equation with Fact 59, we have

$$\sum_{i \in H} (\mathring{f}(i) - \mathring{g}(i))^2 \leq O(|H| n^2 \delta^2) + \sum_{i \in H} (v_i(f) - v_i(g))^2 \tag{32}$$

Using the above equation with Jensen's inequality, we have

$$\sum_{i \in H} (v_i(f) - v_i(g))^2 \leq \left(\frac{\Lambda(n)}{2 C_\delta} \right)^2 \underset{\boldsymbol{p} \sim \mathcal{K}(\delta)}{\mathbf{E}} \left[\sum_{i \in H} \left(\sigma_p \left(\hat{f}(i, \boldsymbol{p}) - \hat{g}(i, \boldsymbol{p}) - \frac{1}{n} \sum_{j=1}^n \left(\hat{f}(j, \boldsymbol{p}) - \hat{g}(j, \boldsymbol{p}) \right) \right) \right)^2 \right] . \tag{33}$$

To bound the right hand side, we will leverage two facts. First, by applying Proposition 24 to f and Fact 60, for any p, we get that

$$\left| \frac{1}{n} \sum_{i \in T} \hat{f}(i, p) - \alpha(\theta, w_H, w_T, p) \cdot \frac{1}{n} \sum_{i \in T} w_i \right| \leq O\left(\frac{1}{\sqrt{n}} \cdot \left(\frac{\tau}{\sigma_p^2} \right)^{1/4} \right).$$

Similarly, applying Proposition 24 to g and Fact 60, for any p, we get that

$$\left| \frac{1}{n} \sum_{i \in T} \hat{g}(i, p) - \alpha(\theta, w_H, w'_T, p) \cdot \frac{1}{n} \sum_{i \in T} w'_i \right| \leq O\left(\frac{1}{\sqrt{n}} \cdot \left(\frac{\tau}{\sigma_p^2} \right)^{1/4} \right).$$

Recalling that the dependence of $\alpha(\theta, w_H, w_T, p)$ on w_T is only through the quantities $\|w_T\|_1$ and $\|w_T\|_2$, and recalling that $\|w_T\|_1 = \|w'_T\|_1$ and $\|w_T\|_2 = \|w'_T\|_2$, we can combine the last two inequalities to obtain

$$\left| \frac{1}{n} \sum_{i \in T} \hat{g}(i, p) - \frac{1}{n} \sum_{i \in T} \hat{f}(i, p) \right| \leq O\left(\frac{1}{\sqrt{n}} \cdot \left(\frac{\tau}{\sigma_p^2} \right)^{1/4} \right) \tag{34}$$

Next, recalling Theorem 27, we have that

$$\sum_{i \in H} (\hat{f}(i, p) - \hat{g}(i, p))^2 \leq O\left(\frac{\tau^2}{\sigma_p^2} \right).$$

47

Applying the third statement of Fact 60 with its scaling factor "c" set to be $\frac{|H|}{n}$, we get that

$$\sum_{i \in H} \left(\hat{f}(i,p) - \hat{g}(i,p) - \frac{1}{n} \sum_{j \in H} \hat{f}(j,p) - \hat{g}(j,p) \right)^2 \leq O\left(\frac{\tau^2}{\sigma_p^2} \right). \tag{35}$$

By combining Equation (34) and Equation (35) , we get that for all $p \in (0,1)$,

$$\sum_{i \in H} \left(\sigma_p \left(\hat{f}(i,p) - \hat{g}(i,p) - \frac{1}{n} \sum_{j=1}^{n} \hat{f}(j,p) - \hat{g}(j,p) \right) \right)^2 = O\left(\tau^2 + \frac{\sqrt{\tau}|H|}{n} \right).$$

Plugging this back into Equation (32) and Equation (33), we get

$$\sum_{i \in H} (\mathring{\hat{f}}(i) - \mathring{\hat{g}}(i))^2 \leq O\left(|H| n^2 \delta^2 \right) + \left(\frac{\Lambda(n)}{2 C_\delta} \right)^2 \cdot O\left(\tau^2 + \frac{\sqrt{\tau}|H|}{n} \right).$$

■

D.3. Structural Theorem for LTFs under $d_{\textbf{Shapley},S}$

In this section we establish a structural result which is at the heart of our algorithm for the Partial Shapley Indices Problem. This result may be viewed as an analogue of Theorem 7, the structural result that was the core of our algorithm for the Partial Chow Parameters problem.

We will use the following lemma, which appears in a number of previous works (e.g., (De et al., 2014, Fact 25)). Given a vector w of non-negative weights that are sorted by magnitude, so $|w_1| \geq \cdots \geq |w_n| \geq 0$, for $i \in [n]$ let us write $\text{tail}_i(w)$ to denote $(\sum_{j=i}^{n} w_j^2)^{1/2}$. The lemma says that, for weight vectors w with sorted weights as above, this quantity decreases geometrically for i less than the critical index:

Lemma 54 *Let $w = (w_1, \ldots, w_n) \in \mathbb{R}^n$ be such that $|w_1| \geq \cdots \geq |w_n|$, and let $1 \leq a \leq b \leq c(w,\tau)$, where $c(w,\tau)$ is the τ-critical index of w. Then $\text{tail}_b(w) < (1 - \tau^2)^{(b-a)/2} \cdot \text{tail}_a(w)$.*

Proof By definition of the critical index, $|w_i| > \tau \cdot \text{tail}_i(w)$ for $i < c(w,\tau)$. Therefore for such an i, $\text{tail}_i(w)^2 = w_i^2 + \text{tail}_{i+1}(w)^2 > \tau^2 \cdot \text{tail}_i(w)^2 + \text{tail}_{i+1}(w)^2$, and so $\text{tail}_{i+1}(w) < (1 - \tau^2)^{1/2} \cdot \text{tail}_i(w)$. The result follows by applying this last inequality repeatedly. ■

Let us sketch the high level idea for the proof of the large critical index case in the main structural theorem below. (The small critical index case will be an immediate corollary of the previous sections' results.) First, we argue that if the τ^*-critical index $k_{\text{critical}}(\tau^*)$ is sufficiently large, specifically $k_{\text{critical}} > k^*$, then the single weight $w_{k^*/2}$ will have larger magnitude than the ℓ_1 weight of the entire tail $\sum_{i=k^*}^{n} |w_i|$. Then we apply the anti-concentration results Theorem 36 and Lemma 41 to conclude that in this case the tail weights will rarely affect the sign of the affine form $\ell(x) = \sum_{i=1}^{n} w_i x_i - \theta$ and hence that $f(x) = \text{sign}(\ell(x))$ is close in ℓ_1 distance to the junta $f'(x) = \text{sign}(\ell'(x))$ where $\ell'(x) = \sum_{i=1}^{k^*-1} w_i x_i - \theta$. Finally, we show using Corollary 39 that the closeness of two functions in ℓ_1 distance implies closeness in (partial) Shapley distance.

Theorem 55 *Let $\varepsilon \in (\frac{1}{n^{1/14}}, \frac{1}{2})$. Define $\tau^* := (\frac{\varepsilon^2}{\log^4 n})$ and $k^* := \max\{4\frac{\log n}{(\tau^*)^2}, \frac{1}{\varepsilon^{12}}\}$. Let $f(x) = \mathrm{sign}(w \cdot x - \theta)$ be a monotone increasing, η-restricted LTF where η is an absolute constant in $(1/4, 1]$. There is a value $0 \le \hat{k} \le k^*$ such that, taking $H \subset [n]$ to be the indices of the $\hat{k}$ largest-magnitude weights in f and $T := [n] \setminus H$ to be the complementary $n - \hat{k}$ remaining weights, at least one of the following holds:*

1. *$d_{\mathrm{Shapley}}(f, f') = O(\varepsilon)$ for some LTF junta f' over the variables in H; or*

2. *The tail Shapley indices are close to an affine transform of the tail weights in the following sense:*
$$\sum_{i \in T} \left(\mathring{f}(i) - \mathring{A}w_i + \mathring{B} \right)^2 \le O\left(\varepsilon\right) ,$$

   where $\mathring{A}, \mathring{B}$ are the values defined in (31) and w_T vector is τ^*-regular.

Proof

The proof is split into two main cases, according to whether the critical index is large (in which case we show that Item 1 holds) or the critical index is small (in which case we show that Item 2 holds).

Large critical index case $(k_{\mathbf{critical}}(\tau^*) > k^*)$**:**
Without loss of generality, we can assume that the coordinates of the vector w are sorted by magnitude, so $|w_1| \ge \cdots \ge |w_n|$. In this case, we will show that Item 1 holds with $\hat{k} := k^* - 1$ and $H := \{1, \cdots, k^* - 1\}$. Indeed, we have that

$$
\begin{aligned}
\|(w_j)_{j=k^*}^n\|_1^2 &\le n \cdot \|(w_j)_{j=k^*}^n\|_2^2 &&\text{(Cauchy-Schwarz Inequality)}\\
&\le n \cdot (1 - (\tau^*)^2)^{k^*/2} \cdot \|(w_j)_{j=k^*/2}^n\|_2^2 &&\text{(Lemma 54 with } \alpha = k^*/2, \beta = k^*)\\
&\le n^2 \cdot (1 - (\tau^*)^2)^{k^*/2} \cdot w_{k^*/2}^2 \\
&\le w_{k^*/2}^2 &&(k^* \ge 4\tfrac{\log n}{(\tau^*)^2} \ge 4\log_{(1\text{-}(\tau^*)^2)^{\text{-}1}}(n)) .
\end{aligned}
$$

It follows that

$$\|(w_j)_{j=k^*}^n\|_1 \le |w_{k^*/2}| . \tag{36}$$

Having established Equation (36), we are ready to show that if we "zero the tail weights" in $f(x)$ to obtain a $(k^* - 1)$-junta $f'(x)$ then the ℓ_1 distance (with respect to the Shapley distribution $\mathcal{D}_{\mathrm{Shap}}$) between $f(x)$ and $f'(x)$ is not too large.

In more detail, we define the junta $\ell'(x) = \sum_{i=1}^{k^*-1} w_i x_i - \theta$ and $f'(x) = \mathrm{sign}(\ell'(x))$. We can assume without loss of generality that $|\ell(x)| \ne |\ell(x) - \ell'(x)|$ for all $x \in \{-1, 1\}^n$. If not, we can ensure this by perturbing the threshold in one of $\ell(x), \ell'(x)$ slightly without changing the values of

$f(x)$, $f'(x)$ for any x. Then

$$\Pr_{\boldsymbol{x}\sim\mathcal{D}_{\text{Shap}}}[f(\boldsymbol{x}) \neq f'(\boldsymbol{x})] = \Pr_{\boldsymbol{x}\sim\mathcal{D}_{\text{Shap}}}[\text{sign}(\ell(\boldsymbol{x})) \neq \text{sign}(\ell'(\boldsymbol{x}))]$$

$$\leq \Pr_{\boldsymbol{x}\sim\mathcal{D}_{\text{Shap}}}[|\ell(\boldsymbol{x})| < |\ell(\boldsymbol{x}) - \ell'(\boldsymbol{x})|]$$

$$\leq \Pr_{\boldsymbol{x}\sim\mathcal{D}_{\text{Shap}}}[|\ell(\boldsymbol{x})| < \textstyle\sum_{i=k^*}^{n}|w_i|]$$

$$\leq \Pr_{\boldsymbol{x}\sim\mathcal{D}_{\text{Shap}}}[|\ell(\boldsymbol{x})| < |w_{k^*/2}|]$$

$$\leq O((\log n)^{-1} \cdot (k^*/2)^{-1/6}) \,,$$

where the penultimate inequality follows by Equation (36) and the final inequality follows by applying Theorem 36, with its parameters set to $r := |w_{k^*/2}|$, $k := k^*/2$, and $\eta := \eta$. Therefore, by Corollary 39,

$$d_{\text{Shapley}}(f, f') \leq O\left(\sqrt{\log n \cdot \Pr_{\boldsymbol{x}\sim\mathcal{D}_{\text{Shap}}}[f(\boldsymbol{x}) \neq f'(\boldsymbol{x})]} + \frac{1}{\sqrt{n}}\right) \leq O\left((k^*)^{-1/12} + \frac{1}{\sqrt{n}}\right) \leq O(\varepsilon)\,.$$

Small critical index case $(k_{\text{critical}}(\tau^*) \leq k^*)$:
It remains to analyze the case that the (τ^*)-critical index $k_{\text{critical}}(\tau^*)$ is at most k^*. In this case, we will show that Item 2 holds with $\hat{k} := k_{\text{critical}}(\tau^*)$, $H := \{1, \cdots, k_{\text{critical}}(\tau^*)\}$, and $T := \{k_{\text{critical}}(\tau^*) + 1, \cdots, n\}$.

By the definition of the critical index, it is easy to check that w_T is (τ^*)-regular. Thus, as an immediate application of Theorem 50 with $\delta = \frac{1}{n^2}$, $|H| = k_{\text{critical}} \leq k^*$ and $\tau^* = \frac{\varepsilon^2}{\log^4 n}$, we get that there exist real values $\mathring{A} = \mathring{A}(w_H, \|w_T\|_1, \|w_T\|_2, \theta, \delta = \frac{1}{n^2})$, $\mathring{B} = \mathring{B}(w_H, \|w_T\|_1, \|w_T\|_2, \theta, \delta = \frac{1}{n^2})$ such that

$$\sum_{i\in T}\left(\mathring{f}(i) - \mathring{A}w_i + \mathring{B}\right)^2 \leq O\left(\frac{1}{n} + \log^2(n) \cdot \left(\frac{k^*}{n} + \sqrt{\tau^*}\right)\right)$$

$$\leq O(\varepsilon)\,,$$

where the second inequality holds by the definition of τ^*, because $\varepsilon \geq \frac{1}{n^{1/14}} \geq \frac{1}{n}$, and because $k^* \leq \max\{\log^{14} n, n^{12/14}\}$. $\blacksquare$

D.4. An algorithm for the Partial Inverse Shapley Index Problem

In this section, we show how to leverage the structural result Theorem 55 to give an algorithm for recovering the weights of an LTF that are very close to being consistent with a subset S of its Shapley indices.

D.4.1. RECOVERING TAIL WEIGHTS BY DYNAMIC PROGRAMMING

We start by presenting a subroutine, RECOVERWEIGHTS, for recovering weights $w = (w_1, \ldots, w_n)$ corresponding to an LTF that (approximately) minimizes the objective function

$$\sum_{i\in S}(\mathring{f}(i) - \mathring{A}w_i + \mathring{B})^2,$$

subject to certain constraints. Using the characterization in Theorem 55, Item 2 of the tail Shapley values as affine functions of their corresponding input weights, this will allow us to output an LTF g with small $d_{\text{Shapley},S}(f,g)$ for functions f with low critical index.

The algorithm takes the following as input:

- A set $\mathcal{S} = \{(\alpha_i, i) : i \in S\}$ of values α_i (to be thought of as approximations of Shapley indices $\mathring{f}(i)$ of an LTF $f(x) = \text{sgn}(\sum_{i=1}^n v_i x_i - \theta)$) and corresponding indices for some subset $S \subseteq [n]$;

- the number of weights n;

- a granularity parameter γ;

- target ℓ_1 and ℓ_2 norm values W_1 and W_2 for the weight vector w, with W_1 an integer multiple of γ and W_2 an integer multiple of γ^2;

- a regularity parameter τ;

- and constants $\mathring{A}, \mathring{B}$.

The algorithm outputs a vector of non-negative weights $w = (w_1, \ldots, w_n)$ that minimizes the objective function $\sum_{i \in S} (\alpha_i - \mathring{A} w_i + \mathring{B})^2$ subject to the constraints that each w_i is an integer multiple of γ, each $w_i \leq \tau W_2$, $\|w\|_1 = W_1$, and $\|w\|_2 = W_2$.

The algorithm works by dynamic programming. It constructs a table T indexed by three values: an index $i \in \{1, \ldots, n\}$, a target ℓ_1 norm value $W_1' \in \{0, \gamma, 2\gamma, \ldots, W_1\}$, and a target ℓ_2 norm value $W_2' \in \{0, \gamma, \sqrt{2}\gamma, \ldots, W_2\}$, where W_1 is an integer multiple of γ and W_2^2 is an integer multiple of γ^2. Each entry $T(k, W_1', W_2')$ contains a weight vector prefix $w = (w_1, \ldots, w_k)$ of length k that minimizes the objective function $\sum_{i \in S \cap [k]} (\alpha_i - \mathring{A} w_i + \mathring{B})^2$ over all weight vector prefixes $w' = (w_1', \ldots, w_k')$ satisfying the constraints that each w_i' is an non-negative integer multiple of γ, $w_i' \leq \tau W_2$, $\|w'\|_1 = W_1'$, and $\|w'\|_2 = W_2'$. (The entry in $T(k, W_1', W_2')$ contains $\perp$ if no such vector w' exists.)

The algorithm works by constructing "layers" of T indexed by k, starting from $k = 1$, and constructing layer k from layer $k-1$ for $k = 2, \ldots, n$. Its final output is $w := T(n, W_1, W_2)$.

Theorem 56 *The procedure* RECOVERWEIGHTS$(\mathcal{S}, n, \gamma, W_1, W_2, \tau, \mathring{A}, \mathring{B})$ *outputs a weight vector* w *satisfying the conditions that* $w = \gamma z$ *for some* $z \in (\mathbb{Z}^{\geq 0})^n$, $\|w\|_1 = W_1$, $\|w\|_2 = W_2, \|w\|_\infty / \|w\|_2 \leq \tau$ *that minimizes the objective function* $\sum_{i \in S} (\alpha_i - \mathring{A} w_i + \mathring{B})^2$ *over all weight vectors satisfying those conditions. Moreover,* RECOVERWEIGHTS *runs in* $\text{poly}(n, 1/\gamma, W_1)$ *time.*

Proof We prove by induction on k that $T[k, W_1', W_2']$ contains a vector $w = (w_1, \ldots, w_k)$ with $\|w\|_1 = W_1'$ and $\|w\|_2 = W_2'$ that minimizes $\sum_{i \in S \cap [k]} (\alpha_i - \mathring{A} w_i + \mathring{B})^2$ if such a vector exists. The base case where $k = 1$ is clear.

For the inductive case, assume that there exists a weight vector that satisfies all of the required conditions, and let $w = (w_1, \ldots, w_k)$ denote a vector that minimizes the quantity $\sum_{i \in S \cap [k]} (\alpha_i - \mathring{A} w_i + \mathring{B})^2$ among all satisfying vectors. Consider $w^* = (w_1^*, \ldots, w_{k-1}^*) = T[k - 1, W_1' -$

Initialize an $n \times (W_1/\gamma + 1) \times (W_2^2/\gamma^2 + 1)$ table T, by setting each of its entries to $\perp$.

for $w_1 \in \{0, \gamma, 2\gamma, \ldots, \tau W_2\}$ **do**
| Set $T[1, w_1, w_1] \leftarrow w_1$.
end

for $k = 2, \ldots, n$ **do**
| **for** $W_1' \in \{0, \gamma, 2\gamma, \ldots, W_1\}$, $W_2' \in \{0, \gamma, \sqrt{2}\gamma, \ldots, W_2\}$ **do**
| | `// Identify a feasible set of weight vectors of length k`
| | `   achieving the desired ` ℓ_1 ` and ` ℓ_2 ` norm.`
| | $V \leftarrow \{w = (w^*, w_k) : w_k \in \{0, \gamma, 2\gamma, \ldots, \tau W_2\}, w_k \leq W_2',$
| | $\quad w^* = T[k-1, W_1' - w_k, ((W_2')^2 - w_k^2)^{1/2}] \neq \perp\}.$
| | `// Identify a feasible weight vector minimizing the`
| | `   objective function.`
| | $T[k, W_1', W_2'] \leftarrow \arg\min_{w \in V} \sum_{i \in S \cap [k]} (\alpha_i - \mathring{A} w_i + \mathring{B})^2$ (or $\perp$ if $V = \emptyset$).
| **end**
end

return $T[n, W_1, W_2]$.

    Algorithm 1: RECOVERWEIGHTS$(\mathcal{S}, n, \gamma, W_1, W_2, \tau, \mathring{A}, \mathring{B})$

$w_k^*, ((W_2')^2 - (w_k^*)^2)^{1/2}]$, which must exist and satisfy $\sum_{i \in S \cap [k-1]} (\alpha_i - \mathring{A} w_i^* + \mathring{B})^2 \leq \sum_{i \in S \cap [k-1]} (\alpha_i - \mathring{A} w_i + \mathring{B})^2$ by the induction hypothesis. The algorithm will therefore consider the pair $w' = (w^*, w_k)$, which is optimal by the assumption that w is optimal, as needed.

We next turn to analyzing the algorithms's runtime. The table T used in RECOVERWEIGHTS has $O(n \cdot W_1/\gamma \cdot W_2^2/\gamma^2) = O(n \cdot W_1^3/\gamma^3)$ entries. Updating each of these entries (other than those in the first layer) requires computing V, which takes $O(\tau W_2/\gamma) = O(W_1/\gamma)$ time, and checking which $w \in V$ minimizes the objective function, which takes $O(|V| \cdot |S|) = O(n \cdot W_1/\gamma)$ time. The algorithm's runtime is dominated by the total time required to update these entries, which is at most $O(n^2 \cdot W_1^4/\gamma^4)$. ∎

D.4.2. MAIN ALGORITHM FOR THE PARTIAL SHAPLEY VALUES PROBLEM

Now we are ready to present the main algorithm for the Partial Shapley Values Problem. This algorithm takes as input a set of Shapley values and corresponding indices $\mathcal{S} = \{(\mathring{f}(i), i) : i \in S\}$ for some $S \subseteq [n]$ of an LTF $f(x) = f(x_1, \ldots, x_n) = \text{sgn}(\sum_{i=1}^n v_i x_i - \theta)$. The algorithm is analogous to the algorithm in Section 7.1 for the Partial Chow Parameters Problem, and works in three steps.

In the first step, the algorithm sets parameters and guesses the size of the head H and tail T indices of f. As in the first step of the Chow algorithm, the algorithm will only need to know (guess) $|H|$ and $|H \cap S|$; note that fixing a guess for $|H \cap S|$ fixes the corresponding set $H \cap S$ by Lemma 35, and also fixes the sizes and identities of $T \cap S$ and $|T|$. (How the indices not in S are partitioned between H and T is irrelevant since any permutation of indices not in S will result in

candidate LTFs g with the same Partial Shapley Distance with respect to S, $d_{\text{Shapley},S}(f,g)$.) In the second step, the algorithm enumerates all LTFs in a relatively small (quasipolynomial size) set based on the structural result in Theorem 55. Enumerating the LTFs in this set is more nuanced than in the corresponding step in the Partial Chow Parameters Problem, and requires guessing additional values and calling the dynamic programming routine RECOVERWEIGHTS from the previous section. In the final step, the algorithm checks which of the candidate LTFs g generated in the previous step satisfies $d_{\text{Shapley},S}(f,g) \leq O(\varepsilon)$, and outputs one of them. This final step domaintes the algorithm's runtime, which is again quasipolynomial.

The idea behind the algorithm's correctness corresponds to the two cases in Theorem 55. In the first case (the "large critical index" case), we will enumerate all junta LTFs g on H whose weights are discretized to some precision γ. In the second case (the "small critical index" case), we will enumerate all LTFs g of a particular form. We will start by considering a discretized version $f' = f'(x_H, x_T) = \text{sgn}(w'_H \cdot x_H + w'_T \cdot x_T - \theta')$ of f whose weights $w'_1, \ldots, w'_n$ and threshold θ' are integer multiples of γ. The goal of the algorithm will be to (approximately) recover f', which by Theorem 40 is close in Shapley distance to f.

The head weights w_H of g are set to be those of f', which are guessed to some precision γ (in a similar way to the large critical index case). The tail weights w_T of g are set to be (roughly) affine functions of the input Shapley indices $\mathring{f}(i)$ by calling RECOVERWEIGHTS on input values $\alpha_i = \mathring{f}(i)$ for $i \in S \cap T$. (Although our overall goal is to recover an approximation of f, it is useful to think of f' as the "ground truth" function whose tail weights we're trying to recover via the call to the subroutine RECOVERWEIGHTS, and of the input values $\alpha_i = \mathring{f}(i)$ to RECOVERWEIGHTS as noisy versions of $\mathring{f}'(i)$.)

By two applications of the triangle inequality,

$$\begin{aligned} d_{\text{Shapley},S}(f,g) &\leq d_{\text{Shapley},S}(f,f') + d_{\text{Shapley},S}(f',g) \\ &\leq d_{\text{Shapley},S}(f,f') + d_{\text{Shapley},H\cap S}(f',g) + d_{\text{Shapley},T\cap S}(f',g) \,. \end{aligned}$$

We will show that $d_{\text{Shapley},S}(f,g) \leq O(\varepsilon)$ by upper bounding each of the three terms on the right-hand side. Roughly speaking, we will show that $d_{\text{Shapley},S}(f,f') \leq O(\varepsilon)$ by the discretization result in Theorem 40, that $d_{\text{Shapley},H\cap S}(f',g) \leq O(\varepsilon)$ by the head Shapley index stability result in Theorem 53, and that $d_{\text{Shapley},T\cap S}(f',g) \leq O(\varepsilon)$ by the result showing that tail weights are affine functions of their corresponding Shapley indices in Theorem 50.

We next present the full algorithm and analysis for the Partial Shapley Values Problem.

1. (a) Define $\tau := \frac{\varepsilon^2}{\log^4 n}$ and $k := \max\{4\frac{\log^9 n}{\varepsilon^4}, \frac{1}{\varepsilon^{12}}\}$. as in Theorem 55. Fix the granularity parameter $\gamma := 1/(n^2 \cdot k^{k/2})$ as in Theorem 40.

   (b) Guess the size of the head $|H| \in [k]$ and the size of $|H \cap S|$. Identify the $|H \cap S|$ elements of S for which $\mathring{f}(i)$ is largest as the corresponding guess for $H \cap S$. Set H equal to the union of $H \cap S$ and $|H| - |H \cap S|$ arbitrary indices not in S. Set T equal to $[n] \setminus H$.

2. For each setting of H in Step 1, enumerate all LTFs of the following forms (corresponding to the two cases in Theorem 55):

   (a) Enumerate all junta LTFs g on H.

   (b) Enumerate all LTFs g of the form

   $$g(x) = g(x_H, x_T) = \text{sign}(w_H \cdot x_H + w_T \cdot x_T - \theta) ,$$

   obtained by enumerating all combinations of a number of values, and then setting $w_1, \ldots, w_n, \theta$ according to the subsequent procedure.

   Enumerate the following:

   i. Head weights w_H with $w_i \in \{0, \gamma, 2\gamma, \ldots, 1\}$ for $i \in H$,
   ii. The threshold $\theta \in \{0, \gamma, 2\gamma, \ldots, n\}$,
   iii. $W_1 \in \{\gamma, 2\gamma, \ldots, n\}$,
   iv. $W_2 \in \{\gamma, \sqrt{2}\gamma, \ldots, n\}$.

   Set $w_1, \ldots, w_n, \theta$ as follows:

   i. Set the head weights w_H and threshold θ equal to the enumerated values.
   ii. Compute $\mathring{A} = \mathring{A}(w_H, W_1, W_2, \theta, 1/n^2)$, $\mathring{B} = \mathring{B}(w_H, W_1, W_2, \theta, 1/n^2)$ using the formulas in Equation (31).
   iii. Set the tail weights as
   $w_T := \text{RECOVERWEIGHTS}(\{(\mathring{f}(i)), i) : i \in T\}, |T|, \gamma, W_1, W_2, \tau, \mathring{A}, \mathring{B})$.

3. For each candidate LTF g generated in Step 2, compute an empirical estimate $\bar{g}(i)$ of each of the Shapley Values $\mathring{g}(i)$ for $i \in S$ so that $|\mathring{g}(i) - \bar{g}(i)| \le \varepsilon/\sqrt{|S|}$ with confidence $1 - \delta_{\text{fail}}/(|S| \cdot M)$, where M is the total number of LTFs enumerated in Step 2. Output (a weights-based representation of) the first g such that $\|(\mathring{f}(i))_{i \in S} - (\bar{g}(i))_{i \in S}\| \le O(\varepsilon)$.

Theorem 57 *There exists an algorithm for the Partial Inverse Shapley Index Problem with the following guarantees. It takes as input four things: (1) a set $\{(i, \mathring{f}(i)) : i \in S\}$ for some η-restricted, monotone increasing LTF $f : \{-1, 1\}^n \to \{-1, 1\}$ with $\eta \in (1/4, 1]$ and some $S \subseteq \{0, 1, \ldots, n\}$, (2) the length n of the input to f, (3) an error parameter $\varepsilon \in (\frac{1}{n^{1/14}}, \frac{1}{2})$, and (4) a confidence parameter $\delta_{\text{fail}} > 0$. It outputs a weights-based representation of an LTF $g : \{-1, 1\}^n \to \{-1, 1\}$ such that $d_{\text{Shapley},S}(f, g) \le O(\varepsilon)$ with probability $1 - \delta_{\text{fail}}$, and runs in time $2^{\tilde{O}(\log^{18} n/\varepsilon^{24})} \cdot \log(1/\delta_{\text{fail}})$.*

Proof We start by arguing that the above algorithm is correct, beginning with analysis similar to that in the Chow algorithm. By taking a union bound, it holds that all $|S| \cdot M$ estimates $\bar{g}(i)$ of the Shapley Indices $\mathring{g}(i)$ of candidate LTFs g with $i \in S$ in Step 3 will be accurate to within a $\varepsilon/\sqrt{|S|}$ additive error factor with probability at least $1 - \delta_{\text{fail}}$. In this case our estimates will all satisfy $\|(\mathring{g}(i))_{i \in S} - (\bar{g}(i))_{i \in S}\| \leq \varepsilon$, and hence by the triangle inequality $\|(\hat{f}(i))_{i \in S} - (\bar{g}(i))_{i \in S}\| - \varepsilon \leq d_{\text{Shapley},S}(f,g) \leq \|(\hat{f}(i))_{i \in S} - (\bar{g}(i))_{i \in S}\| + \varepsilon$ for every candidate LTF g. So, in this case, we will output a candidate LTF g if and only if it satisfies $d_{\text{Shapley},S}(f,g) \leq O(\varepsilon)$.

In terms of correctness, it remains to show that one of the enumerated LTFs g satisfies $d_{\text{Shapley},S}(f,g) \leq O(\varepsilon)$. By Theorem 40 we have that there exists an LTF f' such that $d_{\text{Shapley}}(f,f') = O(\varepsilon)$ with weights and a threshold which are integer multiples of γ – as defined in Step 1 of the algorithm, $\gamma := 1/(n^2 \cdot k^{k/2})$.

We consider the two cases in Theorem 55 for function $f'(x) = \text{sign}(\sum_{i=1}^{n} w'_i x_i - \theta')$. In the first case, $d_{\text{Shapley}}(f',g) = O(\varepsilon)$ for some junta LTF g on H. All such discretized junta LTFs on H are enumerated in Step 2a, and so the algorithm enumerates a g satisfying $d_{\text{Shapley},S}(f,g) \leq d_{\text{Shapley}}(f,f') + d_{\text{Shapley}}(f',g) \leq O(\varepsilon)$, as needed.

In the second case, we're guaranteed that there exist constants $\mathring{A}', \mathring{B}'$ such that $\sum_{i \in T}(\mathring{f}'(i) - \mathring{A}' w'_i + \mathring{B}')^2 \leq O(\varepsilon)$ and w'_T is $\tau-$regular, where constants $\mathring{A}', \mathring{B}'$ are given by Equation (31) for $\delta = \frac{1}{n^2}$ and w', θ'. We will show that there exists a function g enumerated in Case 2b that satisfies $d_{\text{Shapley},S}(f,g) \leq O(\varepsilon)$. To do this, we observe that by two applications of triangle inequality,

$$\begin{aligned} d_{\text{Shapley},S}(f,g) &\leq d_{\text{Shapley},S}(f,f') + d_{\text{Shapley},S}(f',g) \\ &\leq d_{\text{Shapley},S}(f,f') + d_{\text{Shapley},H \cap S}(f',g) + d_{\text{Shapley},T \cap S}(f',g) \,. \end{aligned}$$

We will show that each of the three terms in the right hand side is upper bounded by $O(\varepsilon)$ in turn. As established earlier, $d_{\text{Shapley},S}(f,f') \leq O(\varepsilon)$.

Next, we argue that $d_{\text{Shapley},H \cap S}(f',g) \leq O(\varepsilon)$. Because f' has weights and a threshold that are integer multiples of γ, the algorithm will enumerate guesses w_H, θ, W_1, and W_2 that are equal to the head weights w'_H, threshold θ', ℓ_1 norm of the tail weights $\|w'_T\|_1$, and ℓ_2 norm of the tail weights $\|w'_T\|_2$ of f', respectively. For such correct guesses, the procedure RECOVERWEIGHTS will output tail weights w_T of g such that $\|w_T\|_1 = \|w'_T\|_1$ and $\|w_T\|_2 = \|w'_T\|_2$. Observe that w_T, w'_T are τ-regular. Thus, by the head Shapley index stability result in Theorem 53 (with $\delta = 1/n^2$), we get $d_{\text{Shapley},H \cap S}(f',g) \leq O(\varepsilon)$.

Finally, we argue that $d_{\text{Shapley},T \cap S}(f',g) \leq O(\varepsilon)$. Recall that

$$\sum_{i \in T}(\mathring{f}'(i) - \mathring{A}' \cdot w'_i + \mathring{B}')^2 \leq O(\varepsilon). \tag{37}$$

Applying $d_{\text{Shapley},S}(f,f') \leq O(\varepsilon)$, we get that

$$\sum_{i \in T}(\mathring{f}(i) - \mathring{A}' \cdot w'_i + \mathring{B}')^2 \leq O(\varepsilon). \tag{38}$$

Consider the case when the algorithm has correctly guessed the head weights w'_H, threshold θ', ℓ_1 norm of the tail weights $\|w'_T\|_1$, and ℓ_2 norm of the tail weights $\|w'_T\|_2$ of f', respectively. By its correctness, RECOVERWEIGHTS will therefore output tail weights w_T of g and threshold θ which

55

satisfy (i) $\|w'_T\|_1 = \|w_T\|_1$ and $\|w'_T\|_2 = \|w_T\|_2$ and (ii) $\theta = \theta'$ and the following holds:

$$\sum_{i \in T} (\mathring{\mathring{f}}(i) - \mathring{A}' \cdot w_i + \mathring{B}')^2 \leq O(\varepsilon). \tag{39}$$

However, since by its correctness, RECOVERWEIGHTS will also guarantee that the weight vector w_T is τ-regular, it will imply that (by Equation (31)),

$$\sum_{i \in T} (\mathring{g}(i) - \mathring{A}' \cdot w_i + \mathring{B}')^2 \leq O(\varepsilon). \tag{40}$$

Thus, applying both Equation (39) and Equation (40),

$$\sum_{i \in T} (\mathring{g}(i) - \mathring{\mathring{f}}(i))^2 \leq O(\varepsilon). \tag{41}$$

Combining with the fact that $d_{\text{Shapley}}(f, f') \leq O(\varepsilon)$ and applying triangle inequality, this implies $d_{\text{Shapley}, T \cap S}(f', g) \leq O(\varepsilon)$.

We now turn to analyzing the runtime of the algorithm. We start by analyzing how many LTFs M' are enumerated in Step 2 for fixed guesses of $|H|$ and $|H \cap S|$ in Step 1. This number is asymptotically dominated by Case 2b, where there are $O((1/\gamma)^k \cdot n/\gamma)$ possible choices for θ and w_i for $i \in H$, $O(n/\gamma)$ choices for W_1, and $O(n^2/\gamma^2)$ choices for W_2. Because $\gamma = 1/(n^2 \cdot k^{k/2})$ and $k \leq 4 \log^9 n/\varepsilon^{12}$, we enumerate a total of

$$M' = (1/\gamma)^{k+4} \cdot \text{poly}(1/\varepsilon) = n^{O(\log^9 n/\varepsilon^{12})} \cdot (\log^9 n/\varepsilon^{12})^{O(\log^{18} n/\varepsilon^{24})} = 2^{\tilde{O}(\log^{18} n/\varepsilon^{24})}.$$

LTFs in Step 2 (for fixed $|H|$).

In Step 1, we make $O(k^2) = O(n^2)$ guesses for $|H|$ and $|H \cap S|$, so we get that the total number of LTFs enumerated by the algorithm is

$$M = M' \cdot O(n^2) = 2^{\tilde{O}(\log^{18} n/\varepsilon^{24})}. \tag{42}$$

Computing $\mathring{A}$, $\mathring{B}$ requires evaluating the formulas in Equation (31), which is efficient. For each guess of w_i for $i \in H$, θ, W_1, W_2, we call RECOVERWEIGHTS, which runs in $\text{poly}(n, 1/\gamma, W_1) = \text{poly}(n, 1/\gamma) = 2^{\tilde{O}(\log^9 n/\varepsilon^{12})}$ time (since $W_1 \leq 1$) which is asymptotically dominated by the upper bound on M in Equation (42). So, the total time needed to enumerate the M LTFs is also $2^{\tilde{O}(\log^{18} n/\varepsilon^{24})}$.

Concluding, the algorithm enumerates $M = 2^{\tilde{O}(\log^{18} n/\varepsilon^{24})}$ LTFs in $\tilde{O}(\log^{18} n/\varepsilon^{24})$ time, so the algorithm's runtime is dominated by the time needed to compute estimates $\bar{g}(i)$ of the Shapley indices in Step 3 for each of the M functions g enumerated in Step 2. The total runtime of the algorithm is the same function of M as in the Chow algorithm (Equation (6)), which is

$$O\left(\frac{n \cdot |S|^2}{\varepsilon^2} \log\left(\frac{|S| \cdot M}{\delta_{\text{fail}}}\right)\right) M\right) = 2^{\tilde{O}(\log^{18} n/\varepsilon^{24})} \cdot \log(1/\delta_{\text{fail}}).$$

■

Appendix E. Fourier and Hermite analysis

E.1. Fourier analysis over $\{-1, 1\}^n$

Viewing $\{-1, 1\}^n$ as endowed with the uniform probability distribution, the set of real-valued functions over $\{-1, 1\}^n$ forms a 2^n-dimensional inner product space with inner product given by $\langle f, g \rangle := \mathbf{E}_{\boldsymbol{x}}[f(\boldsymbol{x})g(\boldsymbol{x})]$. The set of functions $(\chi_S)_{S \subseteq [n]}$ defined by $\chi_S := \prod_{i \in S} x_i$ forms a complete orthonormal basis for this space. Given a function $f : \{-1, 1\}^n \to \{-1, 1\}$ we define its *Fourier coefficients* by $\hat{f}(S) := \mathbf{E}_{\boldsymbol{x}}[f(\boldsymbol{x})\chi_S(\boldsymbol{x})]$, and we have that the *Fourier representation* of f is $f(x) = \sum_{S \subseteq [n]} \hat{f}(S)\chi_S$ (note that this is the unique representation of f as a multilinear real polynomial).

We will be particularly interested in f's degree-1 coefficients, i.e., $\hat{f}(S)$ for $|S| = 1$; we will write these as $\hat{f}(i)$ rather than $\hat{f}(\{i\})$, and we note that these correspond precisely to the Chow Parameters of f. Finally, we recall *Plancherel's identity*, which states that $\langle f, g \rangle = \sum_{S \subseteq [n]} \hat{f}(S)\hat{g}(S)$, and the special case of *Parseval's identity*, which states that $\mathbf{E}_{\boldsymbol{x}}[(f(\boldsymbol{x}))^2] = \sum_{S \subseteq [n]} \hat{f}(S)^2 = 1$.

E.2. Hermite analysis over $\mathbb{R}^n$

Here we consider functions $f : \mathbb{R}^n \to \mathbb{R}$, where we think of the inputs x to f as being distributed according to the standard n-dimensional Gaussian distribution $N(0, 1)^n$. In this context we view the space of all real-valued square-integrable functions as an inner product space with inner product $\langle f, h \rangle = \mathbf{E}_{\boldsymbol{x} \sim N(0,1)^n}[f(\boldsymbol{x})h(\boldsymbol{x})]$. In the case $n = 1$, there is a sequence of Hermite polynomials $h_0(x) \equiv 1, h_1(x) = x, h_2(x) = (x^2 - 1)/\sqrt{2}, \ldots$ that form a complete orthonormal basis for the space. These polynomials can be defined via $\exp(\lambda x - \lambda^2/2) = \sum d = 0^\infty (\lambda^d/\sqrt{d!})h_d(x)$. In the case of general n, we have that the collection of n-variate polynomials $\{H_S(x) := \prod_{i=1}^n h_{S_i}(x_i)\}_{S \in \mathbb{N}^n}$ forms a complete orthonormal basis for the space. Given a square integrable function $f : \mathbb{R}^n \to \mathbb{R}$ we define its Hermite coefficients by $\tilde{f}(S) = \langle f, H_S \rangle$, for $S \in \mathbb{N}^n$ and we have that $f(x) = \sum_S \tilde{f}(S)H_S(x)$ (with the equality holding in $\mathcal{L}^2$). Again, we will be particularly interested in f's degree-1 coefficients, i.e., $\tilde{f}(e_i)$, where e_i is the vector which is 1 in the i-th coordinate and 0 elsewhere; observe that $\tilde{f}(e_i)$ is $\mathbf{E}_{\boldsymbol{x} \sim N(0,1)^n)}[f(\boldsymbol{x})x_i]$. Plancherel's and Parseval's identities are easily seen to hold in this setting.

Appendix F. Useful inequalities

In this section we record some useful elementary inequalities.

Fact 58 *Suppose that A, B are non-negative and $|A - B| \leq \eta$. Then $|\sqrt{A} - \sqrt{B}| \leq \dfrac{\eta}{\sqrt{B}}$.*

Proof $|\sqrt{A} - \sqrt{B}| = \dfrac{|A - B|}{|\sqrt{A} + \sqrt{B}|} \leq \dfrac{\eta}{\sqrt{B}}$. $\blacksquare$

Fact 59 *Let $a, b, c \in \mathbb{R}^n$ with $\|a - b\|_2^2 = O(\varepsilon_1)$ and $\|b - c\|_2^2 = O(\varepsilon_2)$. Then*

$$\|a - c\|_2^2 \leq O(\varepsilon_1 + \varepsilon_2).$$

Proof It is easy to verify that $\|x-y\|_2^2 \leq 2\|x\|_2^2 + 2\|y\|_2^2$, and consequently we have that $\|a-c\|_2^2 \leq 2\|a-b\|_2^2 + 2\|b-c\|_2^2 \leq O(\varepsilon_1 + \varepsilon_2)$. ∎

Given any vector $v \in \mathbb{R}^n$, let us write $v_{\|}$ to denote

$$v_{\|} := \frac{\sum_{i \in n} v_i}{n}(1, \cdots, 1),$$

which we call the *centralized vector* of v.

Fact 60 *Let $a, b \in \mathbb{R}^n$ be such that $\|a - b\|_2 \leq \eta$. Then for any constant $c \in [0, 1]$ it holds that*

$$c\|a_{\|} - b_{\|}\|_2 \leq \eta, \quad \left|\frac{\sum_{i \in n} a_i}{n} - \frac{\sum_{i \in n} b_i}{n}\right| \leq \frac{\eta}{\sqrt{n}}, \quad and \quad \|(a - ca_{\|}) - (b - cb_{\|})\|_2 \leq O(\eta).$$

Proof It suffices to prove only the first claim since the second and the third one can be obtained from the first via the triangle inequality. For the first we have

$$\|a_{\|} - b_{\|}\|_2 = \|(1, \cdots, 1)\|_2 \cdot \left|\sum_{i \in n} \frac{a_i - b_i}{n}\right| \leq \frac{1}{\sqrt{n}} \sum_{i \in n} |a_i - b_i| \leq \frac{1}{\sqrt{n}}\|a - b\|_1 \leq \frac{\sqrt{n}}{\sqrt{n}}\|a - b\|_2 \leq \eta.$$

∎

Fact 61 *Let $a, b \in \mathbb{R}^m$ with $\|a\|_2 \leq 1, \|b\|_2 \leq 1$ such that $\|a - b\|_2^2 \leq \eta$. Then*

$$\left|\|a\|_2 - \|b\|_2\right| \leq \sqrt{\eta} \quad and \quad \left|\|a\|_2^2 - \|b\|_2^2\right| \leq 2\sqrt{\eta}.$$

Proof The first claim holds by the triangle inequality, since $\left|\|a\|_2 - \|b\|_2\right| \leq \|a - b\|_2$. For the second claim we have that $\left|\sum_{i=1}^m (a_i^2 - b_i^2)\right| = \left|\sum_{i=1}^m (a_i - b_i)(a_i + b_i)\right| \leq \sqrt{\sum_{i=1}^m (a_i + b_i)^2}\sqrt{\sum_{i=1}^m (a_i - b_i)^2} \leq$

$\sqrt{2\sum_{i=1}^m (a_i^2 + b_i^2)}\|a - b\|_2 \leq 2\sqrt{\eta}.$ ∎

Fact 62 *Let $a, b, c \in \mathbb{R}^m$ with $\|a - b\|_1 \leq \eta$ and $\|c\|_\infty \leq O(1)$. Then*

$$\left|(a - b) \cdot c\right| \leq O(\eta).$$

Proof We have that

$$|(a - b) \cdot c| = \left|\sum_i (a_i - b_i)c_i\right| \leq \sum_i |a_i - b_i| \cdot |c_i| \leq \|c\|_\infty \cdot \sum_i |a_i - b_i| = \|c\|_\infty \cdot \|a - b\|_1 = O(\eta)$$

as claimed. ∎

Appendix G. Consequences and variants of the Berry-Esseen theorem for p-biased linear forms

Recall Fact 17:

Fact 17. *Let $0^n \neq w \in \mathbb{R}^n$ be τ-regular, and let $p \in (0,1)$. Then we have the following:*

1. *For any interval $[a,b] \subseteq \mathbb{R} \cup \{\pm\infty\}$,*

$$\left| \Pr_{\boldsymbol{x} \sim u_p^n} [w \cdot \boldsymbol{x} \in [a,b]] - \left(\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \right) \right| \leq \frac{4\tau}{\sigma_p},$$

   where $\mu = \mu_p \cdot \sum_{i=1}^n w_i$ and $\sigma = \sigma_p \cdot \|w\|_2$.

2. *For any λ and any $\theta \in \mathbb{R}$, we have*

$$\Pr_{\boldsymbol{x} \sim u_p^n} [|w \cdot \boldsymbol{x} - \theta| \leq \lambda] \leq 2\frac{\lambda}{\sigma_p \|w\|_2} + 2\frac{\tau}{\sigma_p}.$$

   In particular, if $\lambda = O(\tau)$ and $\|w\|_2 = 1$, then we have

$$\Pr[|w \cdot \boldsymbol{x} - \theta| \leq \lambda] \leq \frac{O(\tau)}{\sigma_p}.$$

Proof For part (1), we apply Theorem 16 to the random variables $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ where $\boldsymbol{Y}_i = w_i \boldsymbol{x}_i - \mu_p w_i$ for $i \in [n]$. It is straightforward to check that for each i we have that $\mathbf{E}[\boldsymbol{Y}_i] = 0$, $\mathbf{E}[\boldsymbol{Y}_i^2] = \sigma_i^2 = \sigma_p^2 \cdot w_i^2$, and $\mathbf{E}[|\boldsymbol{Y}_i|^3] = 8p(1-p) \cdot (p^2 + (1-p)^2) \cdot w_i^3 \leq 8p(1-p) \cdot w_i^3$. Therefore $\sigma = \sqrt{\sum_{i=1}^n \sigma_i^2} = \sigma_p \cdot \|w\|_2$ and $\rho = \sum_{i=1}^n \mathbf{E}[|\boldsymbol{Y}_i|^3] \leq 8p(1-p) \cdot \|w\|_3^3 \leq 8p(1-p) \cdot \|w\|_2^2 \cdot \|w\|_\infty$, and hence by Theorem 16 and the τ-regularity of w it holds that for any $\theta \in \mathbb{R}$,

$$\left| \mathbf{Pr}\left[\sigma^{-1} \cdot \sum_{i=1}^n \boldsymbol{Y}_i \leq \theta \right] - \Phi(\theta) \right| \leq \frac{8p(1-p) \cdot \|w\|_2^2 \cdot \|w\|_\infty}{\sigma_p^3 \cdot \|w\|_2^3} \leq \frac{2\tau}{\sigma_p}. \tag{43}$$

We also have

$$\Pr_{\boldsymbol{x} \sim u_p^n} [w \cdot \boldsymbol{x} \leq \theta] = \mathbf{Pr}\left[\sigma^{-1} \cdot \sum_{i=1}^n \boldsymbol{Y}_i \leq \frac{\theta - \mu}{\sigma} \right] \tag{44}$$

where $\mu = \mathbf{E}_{\boldsymbol{x} \sim u_p^n}[w \cdot \boldsymbol{x}]$ (we note for later reference that if all coefficients of w are non-negative, then this value is equal to $\mu_p \cdot \|w\|_1$). We get part (1) of the fact by combining Equations (43) and (44) twice, once setting $\theta = a$ and once setting $\theta = b$.

For part (2), we have

$$\Pr_{\boldsymbol{x} \sim u_p^n} [|w \cdot \boldsymbol{x} - \theta| \leq \lambda] = \Pr[w \cdot \boldsymbol{x} \in [\theta - \lambda, \theta + \lambda]]$$

$$\leq \Phi(\psi_p^{[w]}(\theta - \lambda), \psi_p^{[w]}(\theta + \lambda)) + 2\frac{\tau}{\sigma_p} \leq 2\frac{\lambda}{\sigma_p \|w\|_2} + 2\frac{\tau}{\sigma_p}.$$

■

Recall Lemma 19:

Lemma 19. *For w a τ-regular LTF, we have*

$$\underset{\boldsymbol{x} \sim u_p^n}{\mathbf{E}}[|w \cdot \boldsymbol{x} - \theta|] \overset{\tau\|w\|_2}{\approx} \|w\|_2 \sigma_p \underset{\boldsymbol{x} \sim N(0,1)}{\mathbf{E}}\left[|\boldsymbol{x} - \psi_p^{[w]}(\theta)|\right] = \|w\|_2 \sigma_p\Big(2\phi(\psi_p^{[w]}(\theta)) - \psi_p^{[w]}(\theta)m(\psi_p^{[w]}(\theta))\Big).$$

Proof The proof closely follows the proof of Proposition 32 in Matulef et al. (2010) with minor changes. Using the fact that $\mathbf{E}[r] = \int_0^{+\infty} \mathbf{Pr}[r > s]ds$ for any nonnegative random variable r for which $\mathbf{E}[r] < +\infty$, we have that:

$$
\begin{aligned}
\underset{\boldsymbol{x} \sim u_p^n}{\mathbf{E}}[|w \cdot \boldsymbol{x} - \theta|] &= \int_0^{+\infty} \underset{\boldsymbol{x} \sim u_p^n}{\mathbf{Pr}}[|w \cdot \boldsymbol{x} - \theta| > s]ds \\
&= \int_0^{+\infty} \mathbf{Pr}[w \cdot \boldsymbol{x} > \theta + s]ds + \int_0^{+\infty} \mathbf{Pr}[w \cdot \boldsymbol{x} < \theta - s]ds \\
&= \int_0^{+\infty} 1 - \mathbf{Pr}[w \cdot \boldsymbol{x} < \theta + s]ds + \int_0^{+\infty} \mathbf{Pr}[w \cdot \boldsymbol{x} < \theta - s]ds. \quad (45)
\end{aligned}
$$

It follows from the Berry-Esseen theorem (Theorem 16, the more detailed bound) that $|(45) - (A)| \leq (B)$, where

$$(A) = \int_0^{+\infty} 1 - \Phi(\psi_p^{[w]}(\theta + s)) + \Phi(\psi_p^{[w]}(\theta - s))ds,$$

$$(B) = O\left(\frac{\tau}{\sigma_p}\right)\int_0^{+\infty} \frac{1}{1 + |\psi_p^{[w]}(\theta + s)|^3} + \frac{1}{1 + |\psi_p^{[w]}(\theta - s)|^3}ds.$$

We have that $(B) = O\left(\frac{\tau}{\sigma_p}\right)\int_0^{+\infty} \frac{1}{1 + |\psi_p^{[w]}(\theta + s)|^3} + \frac{1}{1 + |\psi_p^{[w]}(\theta - s)|^3}ds = O\left(\tau\|w\|_2\right).$ Turning to (A), we observe that (A) can be reexpressed as

$$(A) = \int_0^{+\infty} \underset{\boldsymbol{x} \sim N(0,1)}{\mathbf{Pr}}\left[|\mu_p(\sum_i w_i) + \|w\|_2 \sigma_p \boldsymbol{x} - \theta| > s\right]ds = \mathbf{E}\left[|\mu_p(\sum_i w_i) + \|w\|_2 \sigma_p \boldsymbol{x} - \theta|\right].$$

Dividing by $\|w\|_2 \sigma_p$, we have

$$(A) = \mathbf{E}\left[|\mu_p(\sum_i w_i) + \|w\|_2 \sigma_p \boldsymbol{x} - \theta|\right] = \|w\|_2 \sigma_p \mathbf{E}\left[|\boldsymbol{x} - \psi_p^{[w]}(\theta)|\right].$$

Using now part(1) of Proposition 15 we get that

$$(A) = \|w\|_2 \sigma_p\Big(2\phi(\psi_p^{[w]}(\theta)) - \psi_p^{[w]}(\theta)m(\psi_p^{[w]}(\theta))\Big)$$

as desired.
■

G.1. Bivariate bounds.

Recall Fact 20:

Fact 20. *Let $\boldsymbol{x} \sim u_p^n$ be a p-biased random vector in $\{-1, 1\}^n$, and let $\boldsymbol{y}$ be a random vector in $\{-1, 1\}^n$ that is ρ-correlated with $\boldsymbol{x}$ (meaning that each coordinate $\boldsymbol{y}_i$ is independently set to equal $\boldsymbol{x}_i$ with probability ρ and is set to a random draw from u_p with probability $1 - \rho$) for some ρ that is bounded away from 1. Let $w \in \mathbb{R}^n$ be τ-regular, and let $\ell(x)$ denote the linear form $\sum_{i=1}^n w_i x_i$. Then for any two intervals $[a, b]$ and $[c, d]$ in $\mathbb{R}$, we have*

$$\left| \mathbf{Pr}[(\ell(\boldsymbol{x}), \ell(\boldsymbol{y})) \in [a, b] \times [c, d]] - \Phi_{0, V}\left([\psi_p^{[w]}(a), \psi_p^{[w]}(b)] \times [\psi_p^{[w]}(c), \psi_p^{[w]}(d)] \right) \right| \leq O\left(\frac{\tau}{\sigma_p} \right),$$

where $V = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ and $\Phi_{0, V}$ denotes the distribution of the bivariate Gaussian with zero mean and covariance matrix V.

Fact 20 is a p-biased analogue of Theorem 68 of Matulef et al. (2010). The proof uses the following multidimensional analogue of the Berry-Esseen theorem (the statement below can be found as Theorem 16 in Khot et al. (2007) and Corollary 16.3 in Bhattacharya and Rao (1986)):

Theorem 63 (Multi-dimensional Berry Esseen) *Let $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ be independent random vectors in $\mathbb{R}^2$ satisfying:*

- $\mathbf{E}[\boldsymbol{X_j}] = 0$ *for all $j = 1, \ldots, n$, and*

- $\rho_3 = \frac{\sum_{j=1}^n \mathbf{E}[\|\boldsymbol{X}_j\|_2^3]}{n} < \infty.$

Let $V := \frac{\sum_{j=1}^n Cov(\boldsymbol{X}_j)}{n}$, where Cov denotes the covariance matrix, and let λ be the smallest eigenvalue of V and Λ be the largest eigenvalue of V. Let Q_n denote the distribution of $\frac{\sum_{j=1}^n \boldsymbol{X}_j}{\sqrt{n}}$, let $\Phi_{0, V}$ denote the distribution of the bivariate Gaussian with zero-vector mean and covariance matrix V and let $\eta = C \rho_3 \frac{1}{\sqrt{n \lambda^3}}$, where C is a certain universal constant. Then for any Borel set A, it holds that

$$|Q_n(A) - \Phi_{0, V}(A)| \leq \eta + Bound(A),$$

where $Bound(A)$ is the following measure of the boundary of A: $Bound(A) = 2 \sup_{y \in \mathbb{R}^2} \Phi_{0, V}((\partial A)^{\eta'} + y)$, where $\eta' = \sqrt{\Lambda} \eta$ and $(\partial A)^{\eta'}$ denotes the set of points within distance η' of the topological boundary of A.

Proof of Fact 20. We first rewrite $\mathbf{Pr}[(\ell(\boldsymbol{x}), \ell(\boldsymbol{y})) \in [a, b] \times [c, d]]$ as

$$\mathbf{Pr}\left[\left(\frac{\sum_i w_i(\boldsymbol{x}_i - \mu_p)}{\|w\|_2 \sigma_p}, \frac{\sum_i w_i(\boldsymbol{y}_i - \mu_p)}{\|w\|_2 \sigma_p} \right) \in [\psi_p^{[w]}(a), \psi_p^{[w]}(b)] \times [\psi_p^{[w]}(c), \psi_p^{[w]}(d)] \right].$$

We will apply Theorem 63. First we define some new random variables: let

$$\boldsymbol{L}_i := (\boldsymbol{A}_i, \boldsymbol{B}_i) := \left(\frac{\sqrt{n} w_i}{\sigma_p \|w\|_2}(\boldsymbol{x}_i - \mu_p), \frac{\sqrt{n} w_i}{\sigma_p \|w\|_2}(\boldsymbol{y}_i - \mu_p) \right)$$

61

for $i = 1, \ldots, n$. Since each $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ is individually a p-biased random variable over $\{-1, 1\}$, it is easy to see that $\mathbf{E}[\boldsymbol{L}_i] = (0, 0)$, and it is also straightforward to verify that the covariance matrix of $\boldsymbol{L}_i$ is

$$\mathbf{Cov}(\boldsymbol{L}_i) = \begin{bmatrix} \mathbf{Cov}(\boldsymbol{A}_i, \boldsymbol{A}_i) & \mathbf{Cov}(\boldsymbol{A}_i, \boldsymbol{B}_i) \\ \mathbf{Cov}(\boldsymbol{B}_i, \boldsymbol{A}_i) & \mathbf{Cov}(\boldsymbol{B}_i, \boldsymbol{B}_i) \end{bmatrix} = \frac{nw_i^2}{\|w\|_2^2} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

It follows that

$$V = \frac{1}{n} \sum_{i=1}^n \mathbf{Cov}(L_i) = \frac{1}{n} \sum_{i=1}^n \frac{nw_i^2}{\|w\|_2^2} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

and consequently the eigenvalues of V are $\lambda = (1 - \eta)$ and $\Lambda = (1 + \eta)$. We note that $\|\boldsymbol{L}_i\|_2 = \sqrt{\boldsymbol{A}_i^2 + \boldsymbol{B}_i^2} = \frac{\sqrt{n}w_i}{\sigma_p \|w\|_2} \sqrt{(\boldsymbol{x}_i - \mu_p)^2 + (\boldsymbol{y}_i - \mu_p)^2}$, and hence $\mathbf{E}[\|\boldsymbol{L}_i\|_2^2] = \frac{2nw_i^2}{\|w\|_2^2}$. Since $\sqrt{(\boldsymbol{x}_i - \mu_p)^2 + (\boldsymbol{y}_i - \mu_p)^2} \le 2\sqrt{2}$ with probability 1, we have that $\|\boldsymbol{L}_i\|_2 \le \frac{2\sqrt{2} \cdot \sqrt{n}w_i}{\sigma_p \|w\|_2} \le \frac{2\sqrt{2}n\tau}{\sigma_p}$ with probability 1. Consequently we have

$$\rho_3 = \frac{\sum_{i=1}^n \mathbf{E}[\|\boldsymbol{L}_i\|_2^3]}{n} \le \frac{\sum_{i=1}^n \mathbf{E}[\|\boldsymbol{L}_i\|_2^2]}{n} \cdot \max_{i \in [n]} \{\|\boldsymbol{L}_i\|_2\}$$

$$= \frac{2n}{n} \cdot \left(\sum_{i=1}^n \frac{w_i^2}{\|w\|_2^2} \right) \cdot \frac{2\sqrt{2}n\tau}{\sigma_p} = 2^{5/2} n^{1/2} \frac{\tau}{\sigma_p}.$$

Recalling the value of λ and the definition of η, we get that $\eta = O(1 - \rho)^{-3/2} \frac{\tau}{\sigma_p}$ and since ρ is bounded away from 1, this is $O(\tau/\sigma_p)$.

It is easy to check that for any $y \in \mathbb{R}^2$, the measure under $\Phi_{0,V}$ of the y-translate of the set of points within distance η' of the topological boundary of $[\psi_p^{[w]}(a), \psi_p^{[w]}(b)] \times [\psi_p^{[w]}(c), \psi_p^{[w]}(d)]$ is $O(\eta')$. Since $\eta' = (1 + \rho)^{1/2}\eta$, this is also $O(\tau/\sigma_p)$.

Thus it holds that

$$\left| \mathbf{Pr}[(\ell(\boldsymbol{x}), \ell(\boldsymbol{y})) \in [a, b] \times [c, d]] - \Phi_{0,V}\left([\psi_p^{[w]}(a), \psi_p^{[w]}(b)] \times [\psi_p^{[w]}(c), \psi_p^{[w]}(d)] \right) \right| \le O\left(\frac{\tau}{\sigma_p} \right),$$

which is the desired statement.

Appendix H. Proof of Lemma 35: Shapley indices are monotone in LTF weights

Recall Lemma 35:

Lemma 35. *Let $f(x) = \text{sign}(\ell(x))$ be an LTF where $\ell(x) = \sum_{i=1}^n w_i x_i - \theta$ is a linear form with $w_1, \ldots, w_n \ge 0$. Then for all $i \ne j \in [n]$, it holds that if $w_i \ge w_j$ then $\mathring{f}(i) \ge \mathring{f}(j)$.*

Proof

Rephrasing Equation (26), the Shapley value for a voter can be expressed as the fraction of all $n!$ orderings of the n voters in which she casts the pivotal vote. More precisely, for a given ordering (permutation) $\pi \in \mathbb{S}_n$, an index i is the unique pivotal index if starting from $x = (-1)^n$ and flipping coordinates of x from -1 to 1 in the order specified by π, flipping x_i changes $f(x)$ from -1 to 1. We thus have

$$\mathring{f}(i) = \frac{2}{n!} \cdot \sum_{\pi \in \mathbb{S}_n} \mathbb{1}\{i \text{ is the pivotal index in } \pi \text{ order}\} = \frac{2 \cdot |\{\pi \in \mathbb{S}_n : i \text{ is the pivotal index in } \pi \text{ order}\}|}{n!}.$$

Let $h : \mathbb{S}_n \to \mathbb{S}_n$ be the following swapping involution:

$$h(\pi) = \pi' = \begin{cases} \pi'(x) = \pi(x) & x \notin \{i, j\} \\ \pi'(i) = \pi(j) \\ \pi'(j) = \pi(i) \end{cases}.$$

We will show that if j is the pivotal index in permutation π, then i is the pivotal index in permutation $h(\pi)$. For simplicity of notation in the proof, we write $PR(\pi, k)$ to denote the predecessors of k in permutation π, i.e $PR(\pi, k) := \{\ell \in [n] : \pi(\ell) < \pi(k)\}$. Thus equivalently we would like to show that:

$$\text{if } \begin{cases} \sum_{k \in PR(\pi,j)} w_k < \theta \text{ and} \\ \sum_{k \in PR(\pi,j)} w_k + w_j \geq \theta, \end{cases} \quad \text{then} \quad \begin{cases} \sum_{k \in PR(h(\pi),i)} w_k < \theta \text{ and} \\ \sum_{k \in PR(h(\pi),i)} w_k + w_i \geq \theta. \end{cases}$$

To complete the exchange argument, we split the permutations where j is the pivotal index into two cases: whether or not i is the predecessor of j in π.

Case 1: $i \in PR(\pi, j)$. By definition of the swapping involution, $j \in PR(h(\pi), i)$.

Additionally, it easy to check that

$$\begin{cases} PR(\pi, j) \setminus \{i\} = PR(h(\pi), i) \setminus \{j\} & (a) \\ PR(\pi, j) \setminus \{i\} \cup \{j\} = PR(h(\pi), i) & (b) \end{cases}.$$

Since j is pivotal in π, we have that:

$$\{ j \text{ is the pivotal index in permutation } \pi \} \Leftrightarrow \begin{cases} \sum_{k \in PR(\pi,j)} w_k < \theta \\ \sum_{k \in PR(\pi,j)} w_k + w_j \geq \theta \end{cases} \Leftrightarrow$$

$$\begin{cases} \sum_{k \in PR(\pi,j)\setminus\{i\}} w_k + w_i < \theta \\ \sum_{k \in PR(\pi,j)\setminus\{i\}} w_k + w_i + w_j \geq \theta \end{cases} \overset{(a)}{\Leftrightarrow} \begin{cases} \sum_{k \in PR(h(\pi),i)\setminus\{j\}} w_k + w_i < \theta \\ \sum_{k \in PR(h(\pi),i)\setminus\{j\}} w_k + w_i + w_j \geq \theta \end{cases} \Leftrightarrow$$

$$\{ j \text{ is the pivotal index in permutation } \pi \} \overset{w_j < w_i}{\Leftrightarrow} \begin{cases} \sum_{k \in PR(h(\pi),i)\setminus\{j\}} w_k + w_j < \theta \\ \sum_{k \in PR(h(\pi),i)\setminus\{j\}} w_k + w_i + w_j \geq \theta \end{cases} \Leftrightarrow$$

$$\begin{cases} \sum_{k \in PR(h(\pi),i)} w_k < \theta \\ \sum_{k \in PR(h(\pi),i)} w_k + w_i \geq \theta \end{cases} \Leftrightarrow \quad \{ i \text{ is the pivotal index in permutation } h(\pi) \}.$$

Case 2: $i \notin PR(\pi, j)$. By definition of the swapping involution, $j \notin PR(h(\pi), i)$. Additionally, it easy to check that

$$PR(\pi, j) = PR(h(\pi), i) \quad (c)$$

Since j is pivotal in π, we have that:

$$\{ \ j \text{ is the pivotal index in permutation } \pi \ \} \Leftrightarrow \begin{cases} \sum_{k \in PR(\pi,j)} w_k < \theta \\ \sum_{k \in PR(\pi,j)} w_k + w_j \geq \theta \end{cases} \overset{w_j < w_i}{\Leftrightarrow}$$

$$\begin{cases} \sum_{k \in PR(\pi,j)} w_k < \theta \\ \sum_{k \in PR(\pi,j)} w_k + w_i \geq \theta \end{cases} \overset{(c)}{\Leftrightarrow} \begin{cases} \sum_{k \in PR(h(\pi),i)} w_k < \theta \\ \sum_{k \in PR(h(\pi),i)} w_k + w_i \geq \theta \end{cases} \Leftrightarrow$$

$$\{ \ i \text{ is the pivotal index in permutation } h(\pi) \ \}.$$

$\blacksquare$

Appendix I. Proof of Lemma 47 and Lemma 48: The unnormalized $\mathcal{Q}(\delta)$ measure approximates $\mathcal{D}_{\text{Shap}}$ to high accuracy

The following useful result intuitively says that the measure given by $\frac{1}{C_\delta} \mathcal{Q}(\delta)$ can take the place of the Shapley distribution $\mathcal{D}_{\text{Shap}}$ and incur only small error:

Lemma 64 *For $\delta > 0$, we have*

$$d_{\text{TV}}\left(\frac{1}{C_\delta} \mathcal{Q}(\delta), \mathcal{D}_{\text{Shap}} \right) \leq \frac{1}{\Lambda(n)} \cdot O\left(\sum_{k=1}^{n/2} (n\delta)^k \right) + \frac{1}{C_\delta} \Pr_{\boldsymbol{x} \sim \mathcal{Q}(\delta)} [\boldsymbol{x} = -1^n] + \frac{1}{C_\delta} \Pr_{\boldsymbol{x} \sim \mathcal{Q}(\delta)} [\boldsymbol{x} = 1^n],$$

and consequently for $\delta = 1/n^c$ for some constant $c > 1$, it holds that

$$\sum_{k=1}^{n-1} \sum_{x \in \{-1,1\}^n_{=k}} \left| \frac{1}{C_\delta} \Pr_{\boldsymbol{x} \sim \mathcal{Q}(\delta)} [\boldsymbol{x} = x] - \Pr_{\boldsymbol{x} \sim \mathcal{D}_{\text{Shap}}} [\boldsymbol{x} = x] \right| \leq \frac{O(n\delta)}{\Lambda(n)}.$$

Proof Recalling that $\Pr_{\boldsymbol{x} \sim \mathcal{D}_{\text{Shap}}}[\boldsymbol{x} = \pm 1^n] = 0$, let us fix an $x \in \{-1, 1\}^n$ such that $0 < \text{weight}(x) < n$. Let $k = \text{weight}(x)$. Then we have that:

$$\left| \frac{1}{C_\delta} \Pr_{\boldsymbol{x} \sim \mathcal{Q}(\delta)} [\boldsymbol{x} = x] - \Pr_{\boldsymbol{x} \sim \mathcal{D}_{\text{Shap}}} [\boldsymbol{x} = x] \right| = \left| \int_\delta^{1-\delta} \frac{\frac{1}{p} + \frac{1}{1-p}}{\Lambda(n)} p^k (1-p)^{n-k} dp - \int_0^1 \frac{\frac{1}{p} + \frac{1}{1-p}}{\Lambda(n)} p^k (1-p)^{n-k} dp \right|$$

$$= \left| \int_{[0,\delta] \cup [1-\delta,1]} \frac{\frac{1}{p} + \frac{1}{1-p}}{\Lambda(n)} p^k (1-p)^{n-k} dp \right| = \int_{[0,\delta]} \frac{\frac{1}{p} + \frac{1}{1-p}}{\Lambda(n)} \left(p^k (1-p)^{n-k} + p^{n-k} (1-p)^k \right) dp$$

$$= \frac{1}{\Lambda(n)} \int_{[0,\delta]} p^{k-1}(1-p)^{n-k} + p^k(1-p)^{n-k-1} + (1-p)^{k-1} p^{n-k} + (1-p)^k p^{n-k-1} dp$$

$$\leq \frac{1}{\Lambda(n)} \left(\delta^k + \delta^{k+1} + \delta^{n-k} + \delta^{n-k+1} \right).$$

Consequently for any $1 \leq k \leq n-1$ we have that

$$\sum_{x \in \{-1,1\}^n_{=k}} \left| \frac{1}{C_\delta} \Pr_{\boldsymbol{x} \sim \mathcal{Q}(\delta)}[\boldsymbol{x} = 1^n] - \Pr_{\boldsymbol{x} \sim \mathcal{D}_{\mathrm{Shap}}}[\boldsymbol{x} = x] \right| \leq \frac{\binom{n}{k}}{\Lambda(n)} \left(\delta^k + \delta^{k+1} + \delta^{n-k} + \delta^{n-k+1} \right),$$

which yields (assuming without loss of generality for simplicity that n is odd)

$$\sum_{k=1}^{n-1} \sum_{x \in \{-1,1\}^n_{=k}} \left| \frac{1}{C_\delta} \Pr_{\boldsymbol{x} \sim \mathcal{Q}(\delta)}[\boldsymbol{x} = x] - \Pr_{\boldsymbol{x} \sim \mathcal{D}_{\mathrm{Shap}}}[\boldsymbol{x} = x] \right| \leq \sum_{k=1}^{n-1} \frac{\binom{n}{k}}{\Lambda(n)} \left(\delta^k + \delta^{k+1} + \delta^{n-k} + \delta^{n-k+1} \right)$$

$$= 2 \sum_{k=1}^{(n-1)/2} \frac{\binom{n}{k}}{\Lambda(n)} \left(\delta^k + \delta^{k+1} + \delta^{n-k} + \delta^{n-k+1} \right)$$

$$\leq \sum_{k=1}^{n/2} \frac{4n^k}{\Lambda(n)} \left(\delta^k + \delta^{k+1} \right),$$

Thus, it holds that

$$d_{\mathrm{TV}} \left(\frac{1}{C_\delta} \mathcal{Q}(\delta), \mathcal{D}_{\mathrm{Shap}} \right) \leq \frac{1}{\Lambda(n)} \cdot O \left(\sum_{k=1}^{n/2} (n\delta)^k \right) + \frac{1}{C_\delta} \Pr_{\boldsymbol{x} \sim \mathcal{Q}(\delta)}[\boldsymbol{x} = -1^n] + \frac{1}{C_\delta} \Pr_{\boldsymbol{x} \sim \mathcal{Q}(\delta)}[\boldsymbol{x} = 1^n]$$

and the lemma is proved. ∎

Now we are ready to prove Lemma 47 and Lemma 48:

Lemma 47. *Let $\delta = 1/n^c$ for some constant $c > 1$ and let $f : \{-1, 1\}^n \to \mathbb{R}$ be such that $\|f\|_\infty \leq O(1)$ and $f(-1)^n = f(1^n) = 0$. Then it holds that*

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathrm{Shap}}}[f(\boldsymbol{x})] - \frac{1}{C_\delta} \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{Q}(\delta)}[f(\boldsymbol{x})] \right| \leq \frac{O(n\delta)}{\Lambda(n)}.$$

Proof

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathrm{Shap}}}[f(\boldsymbol{x})] - \frac{1}{C_\delta} \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{Q}(\delta)}[f(\boldsymbol{x})] \right| = \left| \sum_{x \in \{-1,1\}^n} f(x) \Pr_{\boldsymbol{x} \sim \mathcal{D}_{\mathrm{Shap}}}[\boldsymbol{x} = x] - f(x) \frac{1}{C_\delta} \Pr_{\boldsymbol{x} \sim \mathcal{Q}(\delta)}[\boldsymbol{x} = x] \right|$$

$$= \left| \sum_{x \in \{-1,1\}^n \setminus \{-1^n, 1^n\}} f(x) \Pr_{\boldsymbol{x} \sim \mathcal{D}_{\mathrm{Shap}}}[\boldsymbol{x} = x] - f(x) \frac{1}{C_\delta} \Pr_{\boldsymbol{x} \sim \mathcal{Q}(\delta)}[\boldsymbol{x} = x] \right|$$

$$\leq \frac{O(n\delta)}{\Lambda(n)} \cdot O(1),$$

where the second equality uses $f((-1)^n) = f(1^n) = 0$ and the last inequality is by Lemma 64 together with a straightforward application of Fact 62. ∎

Lemma 48. *Let f be any nontrivial monotone LTF (so $f((-1)^n) = -1$ and $f(1^n) = 1$). Then for $\delta = 1/n^c$ for some constant $c > 1$, for each $i \in [n]$, the value $\mathring{f}(i)$ is additively $O(n\delta)$-close to the quantity v_i defined below:*

$$\mathring{f}(i) \overset{O(n\delta)}{\approx} \frac{2}{n} + \frac{\Lambda(n)}{2} \cdot \frac{1}{C_\delta} \cdot \underset{\boldsymbol{p} \sim \mathcal{K}(\delta)}{\mathbf{E}} \left[f^*(i, \boldsymbol{p}) - \frac{1}{n} \sum_{j=1}^{n} f^*(j, \boldsymbol{p}) \right] =: v_i$$

$$= \frac{2}{n} + \frac{\Lambda(n)}{2} \cdot \frac{1}{C_\delta} \cdot \underset{\boldsymbol{p} \sim \mathcal{K}(\delta)}{\mathbf{E}} \left[\sigma_{\boldsymbol{p}} \hat{f}(i, \boldsymbol{p}) - \frac{1}{n} \sum_{j=1}^{n} \sigma_{\boldsymbol{p}} \hat{f}(j, \boldsymbol{p}) \right].$$

Proof

$$\mathring{f}(i) = \frac{f(1^n) - f((-1)^n)}{n} + \frac{\Lambda(n)}{2} \cdot \left(f^*(i) - \frac{1}{n} \sum_{j=1}^{n} f^*(j) \right) \qquad \text{(Lemma 44)}$$

$$= \frac{2}{n} + \frac{\Lambda(n)}{2} \cdot \left(\underset{\boldsymbol{x} \sim \mathcal{D}_{\text{Shap}}}{\mathbf{E}} \left[f(\boldsymbol{x}) \left(\boldsymbol{x}_i - \frac{\sum_{k \in [n]} \boldsymbol{x}_k}{n} \right) \right] \right) \qquad \text{(Definition of } f^*(i))$$

$$\overset{O(n\delta)}{\approx} \frac{2}{n} + \frac{\Lambda(n)}{2} \cdot \left(\frac{1}{C_\delta} \cdot \underset{\boldsymbol{x} \sim \mathcal{Q}(\delta)}{\mathbf{E}} \left[f(\boldsymbol{x}) \left(\boldsymbol{x}_i - \frac{\sum_{k \in [n]} \boldsymbol{x}_k}{n} \right) \right] \right) \qquad \text{(Lemma 47)}$$

$$= \frac{2}{n} + \frac{\Lambda(n)}{2} \cdot \frac{1}{C_\delta} \cdot \underset{\boldsymbol{p} \sim \mathcal{K}(\delta)}{\mathbf{E}} \left[f^*(i, \boldsymbol{p}) - \frac{1}{n} \sum_{j=1}^{n} f^*(j, \boldsymbol{p}) \right] = v_i,$$

giving the first claimed approximation (where the last equality holds recalling that $f^*(i, p) = \mathbf{E}_{\boldsymbol{x} \sim u_p^n} [f(\boldsymbol{x}) \boldsymbol{x}_i]$, recall Table 1). For the second statement, observe that as a straightforward consequence of the definition of $f^*(i, p)$ we have that

$$f^*(i, p) = \sigma_p \hat{f}(i, p) + \underset{\boldsymbol{x} \sim u_p^n}{\mathbf{E}} [f(\boldsymbol{x})] \, \mu_p.$$

Using the above equivalent definition we get that:

$$f^*(i, p) - \frac{1}{n} \sum_{j=1}^{n} f^*(i, p) = \sigma_p \hat{f}(i, p) + \underset{\boldsymbol{x} \sim u_p^n}{\mathbf{E}} [f(\boldsymbol{x})] \mu_p - \frac{1}{n} \sum_{j=1}^{n} \left(\sigma_p \hat{f}(j, p) + \underset{\boldsymbol{x} \sim u_p^n}{\mathbf{E}} [f(\boldsymbol{x})] \mu_p \right)$$

$$= \sigma_p \left(\hat{f}(i, p) - \frac{1}{n} \sum_{j=1}^{n} \hat{f}(j, p) \right)$$

which gives the second statement as claimed. $\blacksquare$