Learning Social Welfare Functions

Kanad Shrikar Pardeshi [1] **Itai Shapira** [2] **Ariel D. Procaccia** [2] **Aarti Singh** [2]

Abstract

Is it possible to understand or imitate a policy maker's rationale by looking at past decisions they made? We formalize this question as the problem of learning social welfare functions belonging to the well-studied family of power mean functions. We focus on two learning tasks; in the first, the input is vectors of utilities of an action (decision or policy) for individuals in a group and their associated social welfare as judged by a policy maker, whereas in the second, the input is pairwise comparisons between the welfares associated with a given pair of utility vectors. We show that power mean functions are learnable with polynomial sample complexity in both cases, even if the comparisons are noisy. Finally, we design practical algorithms for these tasks and evaluate their performance on simulated data.

1. Introduction

Consider a standard decision making setting that includes a set of possible actions (decisions or policies), and a set of individuals who assign utilities to the actions. A *social welfare function* aggregates the utilities into a single number, providing a measure for the evaluation of actions with respect to the entire group. Utilitarian social welfare, for example, is the sum of utilities, whereas egalitarian social welfare is the minimum utility. Given two actions that induce the utility vectors $(3, 0)$ and $(1, 1)$ for two individuals, the former is preferred when measured by utilitarian social welfare, whereas the latter is preferred according to egalitarian social welfare.

When competent decision makers adopt policies that affect groups or even entire societies, they may have a social welfare function in mind, but it is typically implicit. Our goal is to *learn* a social welfare function that is consistent with the decision maker's rationale. This learned social welfare function has at least two compelling applications: first, *understanding* the decision maker's priorities and ideas of fairness, and second, potentially *imitating* a successful decision maker's policy choices in future dilemmas or in other domains.

As a motivating example, consider the thousands of decisions made by public health officials in the United States during the Covid pandemic: opening and closing schools, restaurants, and gyms, requirements for masking and social distancing, lockdown recommendations, and so on. Each decision induces utilities for individuals in the population; closing schools, for instance, provides higher utility to medically vulnerable individuals compared to opening them, but arguably has much lower utility for students and parents. Assuming that healthcare officials were acting in the public interest and (approximately) optimizing a social welfare function, which one did they have in mind? Our goal is to answer such questions by learning from example decisions.

In order to formalize this problem, there are two issues we need to address. First, to facilitate sample-efficient learnability, we need to make some structural assumptions on the class of social welfare functions. We focus on the class of *weighted power mean functions*, which includes the most prominent social welfare functions: the aforementioned utilitarian and egalitarian welfare, as well as Nash welfare (the product of utilities). This class is a natural choice, as it is characterized by a set of reasonable social choice axioms such as monotonicity, symmetry, and scale invariance (Roberts, 1980; Cousins, 2023).

Second, we need to specify the input to our learning problem. There are two natural options, and we explore both: utility vectors coupled with their values under a target social welfare function, or pairwise comparisons between utility vectors. We demonstrate sample complexity bounds for both types of inputs, where the social welfare value or comparisons can be noiseless or corrupted by noise. A limitation of our approach is that estimating the utility vector associated with any particular decision or policy may be challenging; see Section 7 for additional discussion of this point.

Our contributions. Learning weighted power mean functions is a non-standard regression or classification problem since the function has a complex, highly nonlinear dependence on the power, which is the parameter of interest. While one can invoke standard hyperparameter selection approaches such as cross-validation to select p from a grid

[1]Carnegie Mellon University [2]Harvard University. Correspondence to: Kanad Pardeshi <kpardesh@andrew.cmu.edu>.

Table 1. A summary of our results regarding the sample complexity of various tasks. Here, $\kappa = \log(u_{\max}/u_{\min})$, with all d individual utilities assumed to be in the range $[u_{\min}, u_{\max}]$. $\rho \in [0, 1/2)$ is the probability of mislabeling for the iid noise model, and $\tau_{\max}$ is the maximum temperature of the logistic noise model

SOCIAL WELFARE INFORMATION	LOSS	KNOWN WEIGHTS	UNKNOWN WEIGHTS
CARDINAL VALUES	ℓ_2 LOSS	$\mathcal{O}(\kappa^4)$	$\mathcal{O}(\kappa^4 d \log d)$
PAIRWISE COMPARISONS	0-1 LOSS	$\mathcal{O}(\log d)$	$\mathcal{O}(d \log d)$
PAIRWISE COMPARISON WITH IID NOISE	0-1 LOSS	$\mathcal{O}\left(\log d/(1-2\rho)^2\right)$	$\mathcal{O}\left(d \log d/(1-2\rho)^2\right)$
PAIRWISE COMPARISONS WITH LOGISTIC NOISE ESTIMATION	LOGISTIC LOSS	$\mathcal{O}(\tau_{\max}^2 \kappa^2)$	$\mathcal{O}(\tau_{\max}^2 \kappa^2 d \log d)$

of values, the infinite domain of p does not allow demonstration of a polynomial sample complexity without deriving an appropriate cover. We derive statistical complexity measures such as pseudo-dimension, covering number, VC dimension and Rademacher complexity for this function class, under both cardinal and ordinal observations of the social welfare function. Our sample complexity bounds are summarized in Table 1. These results may be of interest for other problems where weighted power mean functions are used, such as fairness in federated learning (Li et al., 2020).

Below, we highlight some key contributions of this paper.

- We establish the statistical learnability of popular social welfare functions belonging to the weighted power mean functions family. We derive a polynomial sample complexity of $\mathcal{O}(1)$ for learning using cardinal social welfare values under ℓ_2 loss, and $\mathcal{O}(\log d)$ (where d denotes the number of individuals) for learning using comparisons under $0-1$ loss in the unweighted/known weight setting. The upper bounds are a consequence of the monotonicity of the target functions with p in the cardinal case, and analysis which reveals that the target functions have $\mathcal{O}(d)$ roots in the ordinal case. As expected, the ℓ_2 loss is also sensitive to the range of utility values κ.
- We establish a polynomial sample complexity of $\mathcal{O}(d \log d)$ for both cardinal and ordinal tasks in the setting when the individual weights are unknown. This result is intuitive, as learning an additional d weight parameters incurs a proportional increase in the sample requirement.
- We also analyze the sample complexity for the more practical ordinal task under different noise models (independent and identically distributed aka iid, and logistic noise) and characterize the dependence of sample complexity on the amount of noise. For the iid setting, the sample complexity increases with large noise (large ρ) and reduces to that of noiseless setting when $\rho = 0$. Unlike the iid setting where ρ is known, for the logistic noise, we also consider estimation of the noise level $\tau_{\max}$ and evaluate the likelihood with respect to the noisy distribution. Since the noise is harder to estimate with increasing τ, the sample complexity increases with τ. Also, the likehood is sensitive to the range of utilities κ.
- Despite the non-convexity of the problem, we demonstrate simple, practical algorithms for learning weighted power means functions on the above tasks using simulated data, and observe good performance over a range of d. Additionally, we verify the theoretically derived scaling of sample complexity with d using our simulations.

Related work. Conceptually, our work is related to that of Procaccia et al. (2009). They also study the learnability of decision rules that aggregate individual preferences, but in their case, the individual preferences are represented as rankings over a set of alternatives (rather than cardinal preferences, as in our problem), and the rule to be learned is a voting rule mapping the input rankings to a winning alternative. They provide sample complexity results with respect to two families of voting rules, namely positional scoring rules and voting trees.

Basu & Echenique (2020) derive bounds on VC dimension for additive, Choquet, and max-min expected utility for decision-making under uncertainty, bounding the number of pairwise comparisons needed to falsify a candidate decision rule and estabilishing learnability or non-learnability for these classes. Note that here the decision rule operates on probability distributions instead of utility vectors, and their results are very different from ours on a technical level; for instance, max-min is not learnable in their setting (infinite VC dimension), whereas it is easily learnable in ours.

Kalai (2001) studies the learnability of choice functions, establishing PAC guaranteees for this class of problems. Choice functions are defined with respect to a fixed and finite set of alternatives X, with each sample being a subset from X and the choice over this subset. By contrast, our work involves learning the behavior of a function on an infinite number of actions parameterized by their feature vectors, which requires analysis that is very different from the finite case.

Pellegrini et al. (2021) conduct experiments on learning aggregation functions which are assumed to be a composition of L_p means, observing that they perform favorably in various tasks such as scalar aggregation, set expansion and graph tasks. Our analysis is of a more theoretical nature, establishing the sample-efficient learnability of weighted power mean aggregation functions.

Melnikov & Hüllermeier (2019) consider learning from actions with feature vectors and their global scores, with local scores for each individual not being available for learning. The goal is to learn both local and global score functions, and consider the ordered weighted averaging operator for aggregating local scores. While we assume that each individual's local score is given, the aggregation function belongs to a different function family motivated by social choice theory, and we provide statistical guarantees for learning from this family.

2. Problem Setup

2.1. Basic Terminology

We assume that the decision-making process concerns d individuals. The decision-making setting we consider has each action associated with a positive utility vector $\mathbf{u} \in [u_{\min}, u_{\max}]^d \subset \mathbb{R}_+^d$, which describes the utilities derived by the d individuals.

We encode the impact of each individual $i \in [d]$ on the decision-making process through a weight value $w_i \geq 0$ such that $\sum_{i=1}^d w_i = 1$. These weight values together form a weight vector $\mathbf{w} \in \Delta_{d-1}$. The weight vector might be a known or unknown quantity. A common instance in which the weight vector is known is when all agents are assumed to have an equal say, in which case $\mathbf{w} = \mathbf{1}_d/d$. For all settings considered in this work, we provide PAC guarantees for both known weights and unknown weights.

We assume that the decision-making process provides a cardinal social welfare value to each action. However, this social welfare value can be latent and need not be available to us as data. For the first task concerned with cardinal decision values, the social welfare values are available and can be used for learning. For the second task, both actions in the pair have a latent social welfare which is not available to us; however, the preferred action in the pair is known to us. We consider learning bounds with the empirical risk minimization (ERM) algorithm for all the losses in this work, with $\hat{p}$ being learned when the weights are known, and $(\hat{\mathbf{w}}, \hat{p})$ being learned when the weights are unknown.

2.2. Power Mean

The (weighted) power mean is defined on $p \in \mathbb{R} \cup \{\pm\infty\}$, and for $\mathbf{u} \in \mathbb{R}_+^d, \mathbf{w} \in \Delta_{d-1}$, it is

$$M(\mathbf{u}; \mathbf{w}, p) = \begin{cases} \left(\sum_{i=1}^n w_i u_i^p\right)^{1/p}, & p \neq 0 \\ \prod_{i=1}^d u_i^{w_i}, & p = 0 \end{cases}$$

It is more convenient to use the (natural) *log* power mean than the power mean, which is defined as:

$$\log M(\mathbf{u}; \mathbf{w}, p) = \begin{cases} \frac{\log\left(\sum_{i=1}^d w_i u_i^p\right)}{p}, & p \neq 0 \\ \sum_{i=1}^d w_i \log u_i, & p = 0 \end{cases}$$

Since $\sum_{i=1}^d w_i = 1$, in effect we have d variables, $w_1, \ldots, w_{d-1}$ and p. We refer to the log power mean family with known weight $\mathbf{w}$ as $\mathcal{M}_{\mathbf{w},d} = \{\log M(\cdot; \mathbf{w}, p) | p \in \mathbb{R}\}$. If the weight is unknown, the log power mean family is denoted by $\mathcal{M}_d = \{\log(\cdot; \mathbf{w}, p) | p \in \mathbb{R}, \mathbf{w} \in \Delta_{d-1}\}$.

The power mean family is a natural representation for social welfare functions. Cousins (2023; 2021) puts forward a set of axioms under which the set of possible welfare functions is precisely the weighted power mean family. An unweighted version of these functions results in the family of constant elasticity of substitution (CES) welfare functions (Goel et al., 2019), which are widely studied in econometrics.

The power mean family has some useful properties. An obvious one is that $\log M(\mathbf{u}, \mathbf{w}, p) \in [\log u_{(1)}, \log u_{(d)}]$, where $u_{(1)}$ and $u_{(d)}$ denote the first and d-th order statistics of $\mathbf{u} = (u_1, \ldots, u_n)$. $\log u_{(1)}$ is attained at $p = -\infty$, and $\log u_{(d)}$ is attained at $p = \infty$. A more general observation is the following:

Lemma 2.1. *(a)* $\log M(\mathbf{u}; \mathbf{w}, p)$ *is nondecreasing w.r.t.* p *for all* $\mathbf{u} \in [u_{\min}, u_{\max}]^d$, $\mathbf{w} \in \Delta_{d-1}$.

(b) $\log M(\mathbf{u}; \mathbf{w}, p)$ *is monotonic w.r.t.* $w_i, i \in [d-1]$ *for all* $\mathbf{u} \in [u_{\min}, u_{\max}]^d$, $p \in \mathbb{R}$.

A proof for the above lemma is provided in Appendix A.1. This monotonicity of the power mean in $\mathbf{w}$ and p was also noted by Qi et al. (2000).

To show the generality of this family of functions, we describe a few illustrative cases:

- $M(\mathbf{u}; \mathbf{w}, p = -\infty) = \min_{i \in d} u_i$, which corresponds to egalitarian social welfare.
- $M(\mathbf{u}; \mathbf{w}, p = 0) = \prod_{i=1}^d u_i^{w_i}$, which corresponds to a weighted version of Nash social welfare.
- $M(\mathbf{u}; \mathbf{w}, p = 1) = \sum_{i=1} w_i u_i$, which corresponds to weighted utilitarian welfare.

- $M(\mathbf{u}; \mathbf{w}, p = \infty) = \max_{i \in d} u_i$, which corresponds to egalitarian social *malfare*.

We note that for $p = \pm\infty$, the decision utility is independent of $\mathbf{w}$. With $w_i = 1/d$ for all $i \in [d]$, we get the conventional interpretations of the welfare notions mentioned above.

3. Cardinal Social Welfare

We first consider the case where we know the cardinal value of the social choice associated with each action. Learning in this setting thus corresponds to regression. Formally, we assume an underlying distribution $\mathcal{D} : [u_{\min}, u_{\max}]^d \times [u_{\min}, u_{\max}]$ over the utilities and social welfare values. We receive iid samples $\{(\mathbf{u}_i, y_i)\}_{i=1}^n \sim \mathcal{D}$, $\mathbf{u}_i$ being the utility vector and $y_i \in [u_{i(1)}, u_{i(d)}]$ being the social welfare value associated with action i.

We consider ℓ_2 loss over $\log M(\mathbf{u}_i; \mathbf{w}, p)$ and $\log y_i$. The true risk in this case is

$$R(\mathbf{w}, p) = \mathbb{E}_{(\mathbf{u},y)\sim\mathcal{D}} \left[(\log M(\mathbf{u}; \mathbf{w}, p) - \log y)^2 \right].$$

To analyze the PAC learnability of this setting, we first provide bounds on the pseudo-dimensions of $\mathcal{M}_{\mathbf{w},d}$ and $\mathcal{M}_d$. To achieve this, we use an important property: the log power mean can be expressed using another log power mean with individual utilities in a fixed range

Lemma 3.1. *Let $q = p \log\left(u_{(d)}/u_{(1)}\right)$, and $\mathbf{r} \in [1, e]^d$ such that*

$$r_i = \exp\left(\frac{\log(u_i/u_{(1)})}{\log(u_{(d)}/u_{(1)})}\right)$$

Then,

$$\log M(\mathbf{u}; \mathbf{w}, p) = \log u_{(1)} + \log\left(\frac{u_{(d)}}{u_{(1)}}\right) \log M(\mathbf{r}; \mathbf{w}, q)$$

We prove Lemma 3.1 in Appendix A.2. We note that since $r_i \in [1, e]$ for any $i \in [d]$, $\log M(\mathbf{r}; \mathbf{w}, q) \in [0, 1]$. We now define the function classes $\mathcal{S}_{\mathbf{w},d} = \{f(\mathbf{u}; \mathbf{w}, p) = \log M(\mathbf{r}; \mathbf{w}, q) | (\mathbf{w}, p) \in \Delta_{d-1} \times \mathbb{R}\}$ and $\mathcal{S}_d = \{f(\mathbf{u}; \mathbf{w}, q) = \log M(\mathbf{r}; \mathbf{w}, p) | p \in \mathbb{R}\}$, with $\mathbf{r}$ and q defined according to the above lemma. We have the following bounds on pseudo-dimensions of $\mathcal{S}_{\mathbf{w},d}$ and $\mathcal{S}_d$:

Lemma 3.2. *(a) If $\mathbf{w}$ is known, then Pdim$(\mathcal{S}_{\mathbf{w},d}) = 1$.*

(b) If $\mathbf{w}$ is not known, then Pdim$(\mathcal{S}_d) < 8d(\log_2 d + 1)$.

A detailed proof is provided in Appendix A.4. We highlight the fact that p and $\mathbf{w}$ are the parameters of the log power mean function family, which calls for the novel bounds provided in this work. These bounds on the pseudo-dimensions can now be used to obtain Rademacher complexity bounds.

Lemma 3.3. *Let $\kappa = \log(u_{\max}/u_{\min})$.*

(a) If $\mathbf{w}$ is known, then

$$\hat{\mathfrak{R}}(\mathcal{M}_{\mathbf{w},d}) \leq \kappa \left(\sqrt{\frac{2\log 2 + 2\log n}{n}} + \frac{c}{\sqrt{n}} \right)$$

(b) If $\mathbf{w}$ is unknown, then

$$\hat{\mathfrak{R}}(\mathcal{M}_{\mathbf{w},d}) \leq \kappa \left(\sqrt{\frac{2\log 2 + 16(2\log_2 d + 1)\log n}{n}} + \frac{c}{\sqrt{n}} \right),$$

where $c > 0$ is a constant.

Proof Sketch. $\mathcal{M}_{\mathbf{w},d}$ and $\mathcal{M}_d$ consist of functions which are shifted and scaled versions of functions in $\mathcal{S}_{\mathbf{w},d}$ and $\mathcal{S}_d$. We can make use of the pseudo-dimensions from Lemma 3.2 and scale them appropriately to obtain the above bounds. A more detailed proof is provided in Appendix A.5. □

We observe that the above lemma provides $\mathcal{O}(\sqrt{\log(n)/n})$ and $\mathcal{O}(\sqrt{d\log(d)\log(n)/n})$ bounds on the Rademacher complexity for unknown and known weights respectively. An important aspect of the above bounds is their dependence on $\log\left(u_{\max}/u_{\min}\right)$. Intuitively, this means that the richness of the function class increases as the range over utilities increases. An illustrative case is when $u_{\min} = u_{\max} = u$, in which case all individiuals have identical utilities, and $\log M(\mathbf{u}; \mathbf{w}, p) = u$ for any $\mathbf{w} \in \Delta_{d-1}, p \in \mathbb{R}$. The power mean family is not very expressive in this case, and the bound on the Rademacher complexity reducing to zero also supports this conclusion.

We can now provide PAC bounds for this setting.

Theorem 3.4. *Given a set of samples $\{(\mathbf{u}_i, y_i)\}_{i=1}^n$ drawn from a distribution $\mathcal{D}$, for any $\delta > 0$, the following holds with probability at least $1 - \delta$ with respect to the ℓ_2 loss function:*

(a) If $\mathbf{w}$ is known, then

$$R(\mathbf{w}, \hat{p}) - \inf_{p\in\mathbb{R}} R(\mathbf{w}, p) \leq 16\kappa^2 \left(\sqrt{\frac{2\log 2 + 2\log n}{n}} + \frac{c}{\sqrt{n}} \right) + 6\sqrt{\frac{\log(4/\delta)}{2n}}$$

(b) If $\mathbf{w}$ is unknown, then

$$R(\hat{\mathbf{w}}, \hat{p}) - \inf_{(\mathbf{w},p)\in\Delta_{d-1}\times\mathbb{R}} R(\mathbf{w}, p) \leq 16\kappa^2 \left(\sqrt{\frac{2\log 2 + 16(d\log_2 d + 1)\log n}{n}} + \frac{c}{\sqrt{n}} \right) + 6\sqrt{\frac{\log(4/\delta)}{2n}}$$

where $\kappa = \log(u_{\max}/u_{\min})$.

Proof Sketch. Since $M(\mathbf{u}_i; \mathbf{w}, p) \in [u_{\min}, u_{\max}]$ and $y_i \in [u_{\min}, u_{\max}]$, the ℓ_2 loss function in this case has domain

$[-\log(u_{\max}/u_{\min}), \log(u_{\max}/u_{\min})] = [-\kappa, \kappa]$. It is Lipschitz continuous in this domain with Lipschitz constant 2κ Using lemma 3.3 and Talagrand's contraction lemma, we obtain the bounds $\hat{\mathfrak{R}}(\ell \circ \mathcal{M}_{\mathbf{w},d}) \leq 2\kappa\hat{\mathfrak{R}}(\mathcal{M}_{\mathbf{w},d})$ and $\hat{\mathfrak{R}}(\ell \circ \mathcal{M}_d) \leq 2\kappa\hat{\mathfrak{R}}(\mathcal{M}_d)$. These Rademacher complexity bounds are then used to obtain the uniform convergence bounds stated above. A detailed proof is provided in Appendix A.6. □

These bounds are distribution-free, with the only assumption being that all utilities and social welfare values are in the range $[u_{\min}, u_{\max}]$. They also imply an $\mathcal{O}(1)$ and $\mathcal{O}(d \log d)$ dependence of sample complexity on d for known and unknown weights respectively. Moreover, we observe the extensive dependence of the upper bound on κ for the ℓ_2 loss. Since κ is scale invariant, scaling all utility vectors and social welfare values has no impact on the PAC guarantees achievable in this setting.

Computationally, $\log M(\mathbf{u}; \mathbf{w}, p)$ is non-convex in $\mathbf{w}$ and p, which means that the ℓ_2 loss is also non-convex. This makes ERM challenging in this setting. Conventional methods such as gradient descent might end up stuck in local optima or saddle points. However, as we observe in our experiments, sufficient random sampling before conducting gradient descent can result in reasonable empirical performance up to moderately high d.

Another shortcoming of this setting is that decision-makers are required to provide a social welfare value for each action. A more natural setting might be when decision-makers only provide their preferences between actions — potentially just their *revealed* preferences, i.e., the choices they have made in the past — and we address this case next.

4. Pairwise Preference Between Actions

We assume an underlying distribution

$$\mathcal{D} : [u_{\min}, u_{\max}]^d \times [u_{\min}, u_{\max}]^d \times \{\pm 1\}\,.$$

We obtain i.i.d. samples $\{((\mathbf{u}_i, \mathbf{v}_i), y_i)\}_{i=1}^n \sim \mathcal{D}$, where $(\mathbf{u}_i, \mathbf{v}_i)$ are the utilities for the i-th pair of actions, and y_i is a comparison between their (latent) social choice values. We encode the comparison function as $C : [u_{\min}, u_{\max}]^d \times [u_{\min}, u_{\max}]^d \to \{\pm 1\}$, with

$$C((\mathbf{u}, \mathbf{v}); \mathbf{w}, p) = \operatorname{sign}\left(\log M(\mathbf{u}; \mathbf{w}, p) - \log M(\mathbf{v}; \mathbf{w}, p)\right).$$

We denote the family of above functions by $\mathcal{C}_{\mathbf{w},d} = \{C((\mathbf{u}, \mathbf{v}); \mathbf{w}, p) : p \in \mathbb{R}\}$ when the weights are known, and $\mathcal{C}_d = \{C((\mathbf{u}, \mathbf{v}); \mathbf{w}, p) : p \in \mathbb{R}, \mathbf{w} \in \Delta_{d-1}\}$ when the weights are unknown. We consider learning with $0-1$ loss over $C((\mathbf{u}_i, \mathbf{v}_i); \mathbf{w}, p)$ and y_i. The true risk in this case is

$$R(\mathbf{w}, p) = \mathbb{E}_{((\mathbf{u},\mathbf{v}),y)\sim\mathcal{D}}\left[\frac{(1 + y \cdot C((\mathbf{u}, \mathbf{v}); \mathbf{w}, p))}{2}\right].$$

To provide convergence guarantees for the above setting, we bound the VC dimension of the comparison-based function classes mentioned above

Lemma 4.1. *(a) If $\mathbf{w}$ is known, then VC$(\mathcal{C}_{\mathbf{w},d}) < 2(\log_2 d + 1)$.*

(b) If $\mathbf{w}$ is unknown, then VC$(\mathcal{C}_d) < 8(d \log_2 d + 1)$.

(c) If $\mathbf{w}$ is unknown, then VC$(\mathcal{C}_d) \geq \log_2 d + 1$.

The above lemma is proved in Appendix A.3. Note that the lower bound in Part (c) is relatively weak; we expect the true lower bound to be at least VC$(\mathcal{C}_d) = \Omega(d)$, and improving it is an open problem.

The finiteness of VC dimension guarantees PAC learnability, and we get uniform convergence bounds using the VC theorem.

Theorem 4.2. *Given samples $\{((\mathbf{u}_i, \mathbf{v}_i), y_i)\}_{i=1}^n \sim \mathcal{D}$ and for 0-1 loss and any $\delta > 0$, with probability at least $1 - \delta$,*

(a) If $\mathbf{w}$ is known, then

$$R(\mathbf{w}, \hat{p}) - \inf_{p\in\mathbb{R}} R(\mathbf{w}, p) \leq 16\sqrt{\frac{2(\log_2 d + 1)\log(n + 1) + \log(8/\delta)}{n}}$$

(b) If $\mathbf{w}$ is unknown, then

$$R(\hat{\mathbf{w}}, \hat{p}) - \inf_{(\mathbf{w},p)\in\Delta_{d-1}\times\mathbb{R}} R(\mathbf{w}, p) \leq 16\sqrt{\frac{8(d \log_2 d + 1)\log(n + 1) + \log(8/\delta)}{n}}$$

We note that unlike the bounds on ℓ_2 loss of Theorem 3.4, these bounds on 0/1 loss are independent of the range of utilities κ, and only depend on d. They provide sample complexity bounds which depend on d as $\mathcal{O}(\log d)$ and $\mathcal{O}(d \log d)$ for known and unknown weights respectively. Despite these PAC guarantees, empirical risk minimization can be particularly difficult in this case, since the loss function as well as the function class $\log M(\mathbf{u}; \mathbf{w}, p) - \log M(\mathbf{v}; \mathbf{w}, p)$ can be non-convex. To illustrate this non-convexity, we plot the value of the above function for two pairs of utility vectors with respect to p in Figure 1, with $d = 6$ and $\mathbf{w} = \mathbf{1}_d/d$. We observe that gradient-based methods can potentially get stuck in local optima which do not label all the samples correctly. However, as we show in the experiments, adequate random sampling before using gradient-based methods can lead to good performance for this setting too.

4.1. Convergence Bounds Under IID Noise

Decision making can be especially challenging if two actions are difficult to compare, and the preference data we obtain can potentially be noisy. We first consider each comparison to be mislabeled in an i.i.d. manner with probability $\rho \in [0, 1/2)$. We make use of the framework developed

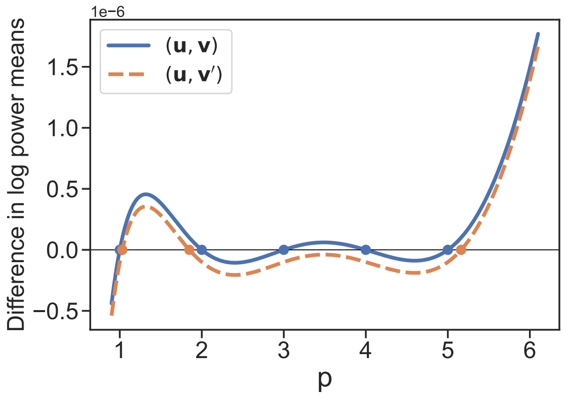


Figure 1. An example showing the non-convexity of $\log M(\mathbf{u}, \mathbf{w}, p) - \log M(\mathbf{v}, \mathbf{w}, p)$. We see that the function has five roots for $(\mathbf{u}, \mathbf{v})$, but is translated downwards for $(\mathbf{u}, \mathbf{v}')$ and has only three roots in this case. If the correct label is 1 for both pairs, then p should be greater than 6; however, gradient-based optimization can stop between 3 and 4, which is a local optimum and does not give correct labels to both points.

by Natarajan et al. (2013), and we consider convergence guarantees under 0-1 loss.

Specifically, the unbiased estimator of ℓ_{0-1} is

$$\tilde{\ell}_{0-1}(t, y) = ((1-\rho)\ell_{0-1}(t, y) - \rho\ell_{0-1}(t, -y))/(1-2\rho).$$

We conduct ERM with respect to $\tilde{\ell}_{0-1}$ to obtain $(\hat{\mathbf{w}}, \hat{p}) \in \Delta_{d-1} \times \mathbb{R}$ (only learning p if weights are known). We observe that $\ell_{0-1}(t, y) = (1+ty)/2$ is 1/2-Lipschitz in t, $\forall t, y \in \{\pm 1\}$. Using Theorem 3 of Natarajan et al. (2013), we get the following convergence bounds:

Theorem 4.3. *Given samples* $\{((\mathbf{u}_i, \mathbf{v}_i), y_i)\}_{i=1}^n \sim \mathcal{D}$, $\forall \delta > 0$, $\forall p \in [0, 1/2)$, *with probability at least* $1-\delta$ *with respect to 0-1 loss,*

(a) If $\mathbf{w}$ *is known, then*

$$R(\hat{\mathbf{w}}, \hat{p}) - \inf_{p \in \mathbb{R}} R(\mathbf{w}, p) \leq \frac{8}{1-2\rho}\sqrt{\frac{(\log_2 d + 1)\log(n+1)}{n}} + 2\sqrt{\frac{\log(1/\delta)}{2n}}$$

(b) If $\mathbf{w}$ *is unknown, then*

$$R(\hat{\mathbf{w}}, \hat{p}) - \inf_{(\mathbf{w}, p) \in \Delta_{d-1} \times \mathbb{R}} R(\mathbf{w}, p) \leq \frac{16}{1-2\rho}\sqrt{\frac{(d\log_2 d + 1)\log(n+1)}{n}} + 2\sqrt{\frac{\log(1/\delta)}{2n}}$$

A detailed proof of the above theorem is provided in Appendix A.7. We note that although ERM is conducted with respect to $\tilde{\ell}_{0-1}$ on the noisy distribution, the risks are defined on the underlying noiseless distribution. This gives $\mathcal{O}(\log d/(1-2\rho)^2)$ and $\mathcal{O}(d \log d/(1-2\rho)^2)$ sample complexities for the known and unknown weights cases respectively. We note that when $\rho = 0$, the above bounds reduce to the bounds in Theorem 4.2. Since the noise level ρ is usually not known to us, it can be estimated using cross-validation as suggested by Natarajan et al. (2013).

However, conducting ERM on $\tilde{\ell}_{0-1}$ might be prohibitively difficult due to the non-convex nature of the function. An iid noise model might also be inappropriate in certain settings; we next consider a more natural noise model.

5. Pairwise Preference With Logistic Noise

Intuitively, we expect that two actions would be harder to compare if their social welfare values are closer to each other. We formalize this intuition in the form of a noise model inspired by the BTL noise model (Bradley & Terry, 1952; Luce, 2005). Let $\mathbf{w}^*$ and p^* be the true power mean parameters, and let $\tau^* \in [0, \tau_{\max}]$ be a temperature parameter. For an action pair $(\mathbf{u}, \mathbf{v})$, we assume that the probability of $\mathbf{u}$ being preferred to $\mathbf{v}$ is

$$\mathbb{P}(y = 1 | (\mathbf{u}, \mathbf{v}); \mathbf{w}^*, p^*, \tau^*) = \frac{1}{1 + \exp\left(-\tau^* \left(\log M(\mathbf{u}; \mathbf{w}^*, p^*) - \log M(\mathbf{v}; \mathbf{w}^*, p^*)\right)\right)} \tag{1}$$

We see that a larger difference between the log power means of $\mathbf{u}$ and $\mathbf{v}$ translates to a higher probability of $\mathbf{u}$ being preferred. If $\mathbf{u}$ and $\mathbf{v}$ lie on the same level set of $\log M(\cdot; \mathbf{w}^*, p^*)$, the probability becomes 0.5, which matches the intuition of both actions being equally preferred. We also note the dependence of the probability on τ^*: a higher τ^* corresponds to more confidence in the preferences, with $\tau^* = 0$ meaning indifference for all pairs of actions. The mislabeling probability is also invariant to scaling $\mathbf{u}$ and $\mathbf{v}$.

Our learning task now becomes estimating $\mathbf{w}$, p and τ given data. We denote the function family in this case by $\mathcal{T}_{\mathbf{w},d} = \{\tau(\log M(\cdot; \mathbf{w}, p) - \log M(\cdot; \mathbf{w}, p)) | \tau, p\}$ when the weights are known, and $\mathcal{T}_d = \{\tau(\log M(\cdot; \mathbf{w}, p) - \log M(\cdot; \mathbf{w}, p)) | \tau, \mathbf{w}, p\}$ when the weights are unknown. A natural loss function to consider in this case is negative log likelihood, and we consider PAC learnability with this loss. Using the framework developed in Section 3, we obtain the following PAC bounds:

Theorem 5.1. *Given samples* $\{((\mathbf{u}_i, \mathbf{v}_i), y_i)\}_{i=1}^n \sim \mathcal{D}$ *and for negative log likelihood loss, for all* $\delta > 0$, *with probability at least* $1-\delta$,

(a) If $\mathbf{w}$ *is known, then*

$$R(\mathbf{w}, \hat{p}) - \inf_{p \in \mathbb{R}} R(\mathbf{w}, p) \leq 16\tau_{\max}\kappa \left(\sqrt{\frac{2\log 2 + 2\log n}{n}} + \frac{c}{\sqrt{n}} \right) + 6\sqrt{\frac{\log(4/\delta)}{2n}}$$

(b) If $\mathbf{w}$ is unknown, then

$$R(\hat{\mathbf{w}}, \hat{p}) - \inf_{(\mathbf{w},p) \in \Delta_{d-1} \times \mathbb{R}} R(\mathbf{w}, p) \leq 16\tau_{\max}\kappa \sqrt{\frac{2\log 2 + 16(d\log_2 d + 1)\log n}{n}} + 16\tau_{\max}\kappa \frac{2c}{\sqrt{n}} + 3\sqrt{\frac{\log(4/\delta)}{2n}}$$

where $\kappa = \log(u_{\max}/u_{\min})$.

Proof Sketch. Since the negative log likelihood function $\ell_{\log}(x) = \log(1 + \exp(-x))$ is 1-Lipschitz, we can use Talagrand's contraction lemma to get $\hat{\mathfrak{R}}(\ell_{\log} \circ \mathcal{T}) \leq \hat{\mathfrak{R}}(\mathcal{T})$. We then observe that $\hat{\mathfrak{R}}(\mathcal{T}) \leq 2\tau_{\max}\hat{\mathfrak{R}}(\mathcal{M})$, and the bound on Rademacher complexity of $\ell_{\log} \circ \mathcal{T}$ can be obtained using Lemma 3.3. The Rademacher bound can then be used to get the uniform bounds stated above. We derive this result in detail in Appendix A.8. □

This gives us sample complexity bounds of $\mathcal{O}(1)$ and $\mathcal{O}(d\log d)$ with respect to d for the known and unknown weights cases respectively, thus establishing PAC learnability. An important distinction between the above theorem and Theorem 4.3 is that the former bounds risk on the noisy distribution with respect to logistic loss, whereas the latter bounds risk on the noiseless distribution with respect to 0-1 loss.

As with the previous cases, non-convexity in this setting also makes global optimization with respect to $\mathbf{w}$ and p (and hence ERM) difficult. Nevertheless, random sampling followed by gradient descent is empirically observed to somewhat overcome this obstacle.

6. Simulations

All of the learning tasks we consider in this work can be computationally hard, since convexity is not guaranteed for any of them. However, we observe through our simulations that weighted power mean functions can still be learned to great accuracy for a modest number of dimensions.

We conduct simulations on cardinal and ordinal data with logistic noise. For each d and n, we construct a dataset in a pre-specified range $[u_{\min}, u_{\max}]^d = [1, 1000]^d$. Each individual i is assumed to have a scaled and translated beta distribution over $[u_{\min}, u_{\max}]$ with the parameters (α_i, β_i) of the beta distribution being different for each i. The utilities for each action are drawn independently for each individual to construct a utility vector. The underlying weight vector is sampled uniformly from Δ_{d-1}.

To learn p (and $\mathbf{w}$ if needed), we first assume p to be in a fixed range, which in this case is $[-10, 10]$. We first conduct a random sampling stage, in which N_{random} instances of p (and $\mathbf{w}$) are uniformly randomly sampled. At the end of this stage, we pick the set of parameters giving the lowest training loss, and then conduct gradient descent for N_{grad} steps. We observe that this simple two-stage method is able to provide good results for the range of values of d we consider. Each setting is run thrice to obtain error bounds on the empirical results.

For the unknown weights case, we observe that we are sampling in d dimensions. As d becomes larger, we increasingly suffer from the curse of dimensionality — N_{random} would have to grow exponentially with d to ensure that we are sampling at the same density across different d. This makes sampling at the same density prohibitively expensive for larger dimensions. As a compromise, we increase N_{random} linearly with d.

6.1. Cardinal Values

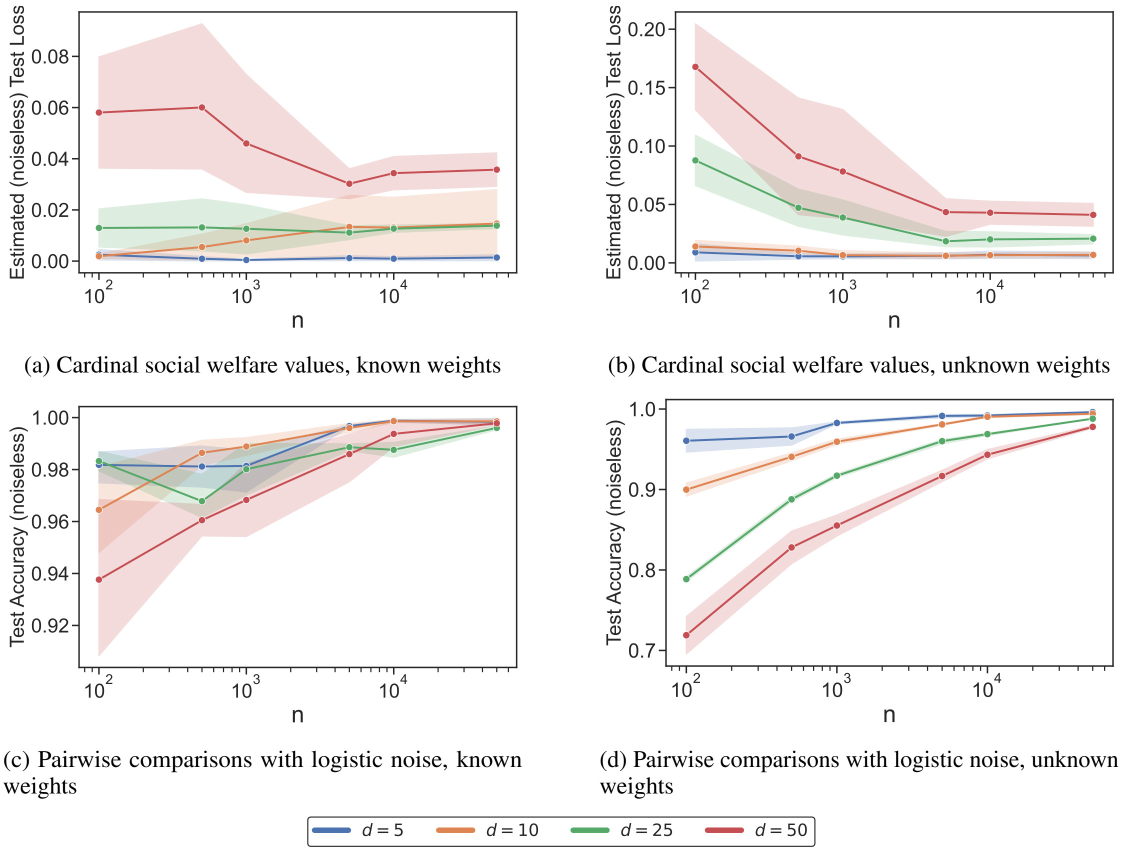
For cardinal values, we further add Gaussian noise to each y_i with standard deviation $(u_{(d)} - u_{(1)})/10$ and clamp the values between $[u_{(1)}, u_{(d)}]$. We conduct experiments for both known and unknown weights by setting $p = -2$. Figure 2a (known weights) and Figure 2b (unknown weights) show the estimated test loss on noiseless test data generated using the true parameters.

We observe that there is relatively little change in the difference of test losses for the case of known weights as n increases. On the other hand, there is greater decrease with increasing n for higher d when the weights are also being learned. The estimated test loss also increases with d, with the trend being stronger for the case of unknown weights.

6.2. Logistic Noise

For logistic noise, we generate pairs of utility vectors with $p = 0.9$ and a $\mathbf{w}$ obtained through random sampling, and then mislabel each instance according to Equation (1) with $\tau^* = 10$. Since we also have to learn τ, we set $\tau_{\max} = 50$, a sufficiently high value, and uniformly randomly sample it along with p (and $\mathbf{w}$). Figure 2a (known weights) and Figure 2b (unknown weights) show the accuracy on noiseless test data of the learned parameters. Across the different settings, the proportion of correctly labeled samples in the training dataset has mean 71.4%, with a maximum value of 86.5%.

For known weights, we observe that accuracy increases with n, and mean accuracy stays high ($> 93\%$) across d. There is



(a) Cardinal social welfare values, known weights

(b) Cardinal social welfare values, unknown weights

(c) Pairwise comparisons with logistic noise, known weights

(d) Pairwise comparisons with logistic noise, unknown weights

Figure 2. Results for synthetic data on cardinal and ordinal logistic tasks with known and unknown weights.

limited distinction between the curves corresponding to different values of d, with all of them approaching near-perfect accuracy as n becomes very high. This suggests that the error bounds in this case should be independent of d. For unknown weights, there is a clear trend of decreasing performance with an increase in d, which is expected because of the $\mathcal{O}(\sqrt{d \log d})$ dependence of logistic loss error bounds. Nevertheless, all settings achieve very high accuracy as n increases. Thus, we observe that up to moderately high d, the logistic noise model finds highly accurate parameters using this simple algorithm, despite the training data having significant mislabeling. We provide empirical verification of our theoretical $\mathcal{O}(d \log d)$ dependence of risk on sample complexity by re-scaling our plots in Appendix B.

7. Discussion

Our work has (at least) several limitations, which can inspire future work. First, as mentioned in Section 1, gaining access to utility vectors — which are needed as input to our problem — can be a challenge. We do not know of existing datasets that include this information. However, utilities are routinely estimated for other economically-motivated algorithms — say, Stackelberg security games (Tambe, 2012) — and there is no fundamental obstacle to estimating them for our purposes. It is an interesting direction of future work to extend our results to the setting where the utility vectors need to be estimated, either by an outside expert, or using input from the individuals themselves (e.g., in the form of pairwise comparisons or rankings of example policies).

Another limitation is that while the performance of our algorithm for small groups of individuals (small d) is very good (more than 90% accuracy for small sample size), scaling up to a national scale — $d = 10^8$, say — is a major challenge for future work. However, this assumes fine-grained utility vectors that assign a utility to each individual. Even at the national scale, a typical scenario plausibly involves "individuals" that are representatives of different sub-populations such as different racial, ethnic, gender or income groups, and are weighted by the size of these groups; this scenario is captured by our framework and leads to relatively small d that is within the range handled by our existing algorithm.

Finally, our work only applies to weighted power mean functions. While we have argued that this family is both expressive and natural, it would be exciting to obtain results for even broader, potentially non-parametric families of social welfare functions.

Impact statement

The ability to learn social welfare functions can enable us to understand a decision maker's priorties and ideas of fairness, based on past decisions they have made. This has direct societal impact as these notions can be used to both understand biases and inform the design of improved fairness metrics. A second potential application is to imitate a successful decision maker's policy choices in future dilemmas or in other domains. This may pose some ethical questions if the learning model is misspecified; however, the restriction of the function class to weighted power means, which is inspired by natural social choice theory axioms, mitigates this risk.

References

Basu, P. and Echenique, F. On the falsifiability and learnability of decision theories. *Theoretical Economics*, 15 (4):1279–1305, 2020.

Bhat, G. S. and Savage, C. D. Balanced Gray codes. *the electronic journal of combinatorics*, 3(1):R25, 1996.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324, 1952.

Cousins, C. An axiomatic theory of provably-fair welfare-centric machine learning. *Advances in Neural Information Processing Systems*, 34:16610–16621, 2021.

Cousins, C. Revisiting fair PAC learning and the axioms of cardinal welfare. In *International Conference on Artificial Intelligence and Statistics*, pp. 6422–6442, 2023.

Goel, A., Hulett, R., and Plaut, B. Markets beyond Nash welfare for Leontief utilities. In *15th International Conference on Web and Internet Economics*, 2019.

Jameson, G. J. Counting zeros of generalised polynomials: Descartes' rule of signs and Laguerre's extensions. *The Mathematical Gazette*, 90(518):223–234, 2006.

Kalai, G. Statistical learnability and rationality of choice. Technical report, Institute of Mathematics and Center for Rationality, The Hebrew University of Jerusalem, 2001.

Li, T., Sanjabi, M., Beirami, A., and Smith, V. Fair resource allocation in federated learning. In *8th International Conference on Learning Representations*, 2020.

Luce, R. *Individual Choice Behavior: A Theoretical Analysis*. Dover Books on Mathematics. Dover Publications, 2005.

Melnikov, V. and Hüllermeier, E. Learning to aggregate: Tackling the aggregation/disaggregation problem for OWA. In *Proceedings of The Eleventh Asian Conference on Machine Learning*, volume 101 of *Proceedings of Machine Learning Research*, pp. 1110–1125, 2019.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

Pellegrini, G., Tibo, A., Frasconi, P., Passerini, A., and Jaeger, M. Learning aggregation functions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2892–2898, 2021.

Procaccia, A. D., Zohar, A., Peleg, Y., and Rosenschein, J. S. The learnability of voting rules. *Artificial Intelligence*, 173(12):1133–1149, 2009.

Qi, F., Mei, J.-Q., Xia, D.-F., and Xu, S.-L. New proofs of weighted power mean inequalities and monotonicity for generalized weighted mean values. *Mathematical Inequalities and Applications*, 3:377–, 09 2000.

Roberts, K. W. S. Interpersonal comparability and social choice theory. *The Review of Economic Studies*, 47(2): 421–439, 1980.

Tambe, M. (ed.). *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned*. Cambridge University Press, 2012.

A. Additional Proofs

A.1. Proof of Lemma 2.1

Proof of Lemma 2.1 Part (a). Let $p > 0$. Define $t_i = \frac{u_i}{u_{\max}}$ for all $i \in [n]$. Since $t_i \leq 1$ for all $i \in [n]$ and given that $\sum_{i=1}^n w_i = 1$, it follows that $\sum_{i=1}^n w_i t_i^p \leq \sum_{i=1}^n w_i = 1$. Therefore, $\log(w_i t_i^p) \leq 0$ for each i. Given $p > q > 0$, we obtain

$$\begin{aligned}
0 &\leq \log\left(\sum_{i=1}^d w_i t_i^p\right)\left(q^{-1} - p^{-1}\right) \\
&= q^{-1}\log\left(\sum_{i=1}^d w_i t_i^p\right) - p^{-1}\log\left(\sum_{i=1}^d w_i t_i^p\right) \\
&\leq q^{-1}\log\left(\sum_{i=1}^d w_i t_i^q\right) - p^{-1}\log\left(\sum_{i=1}^d w_i t_i^p\right) \\
&= \left(u_{\max} + q^{-1}\log\sum_{i=1}^d w_i t_i^q\right) - \left(u_{\max} + p^{-1}\log\sum_{i=1}^d w_i t_i^p\right) \\
&= \log M(\mathbf{u}; \mathbf{w}, p) - \log M(\mathbf{u}; \mathbf{w}, q),
\end{aligned}$$

This derivation similarly holds for the case $0 > p > q$, demonstrating the monotonicity of $\log M(\mathbf{u}; \mathbf{w}, p)$ with respect to p. The continuity of $\log M(\mathbf{u}; \mathbf{w}, p)$ with respect to p at $p = 0$ ensures the monotonicity for all $p \in \mathbb{R}$. □

Proof of Lemma 2.1 Part (a). Since $\sum_{i=1}^d w_i = 1$, we express $w_d = 1 - \sum_{i=1}^{d-1} w_i$ and consider $d-1$ variables. For $p \neq 0$, we have

$$\begin{aligned}
\log M(\mathbf{u}; \mathbf{w}, p) &= p^{-1}\log\left(\sum_{i=1}^{d-1} w_i u_i^p + u_d^p\left(1 - \sum_{i=1}^{d-1} w_i\right)\right) \\
\implies \frac{d\log M(\mathbf{u}; \mathbf{w}, p)}{dw_i} &= p^{-1}\left(\frac{u_i^p - u_d^p}{\sum_{i=1}^{d-1} w_i u_i^p + u_d^p\left(1 - \sum_{i=1}^{d-1} w_i\right)}\right) \\
&= \left(\frac{u_i^p - u_d^p}{\sum_{i=1}^d w_i u_i^p}\right)p^{-1}
\end{aligned}$$

The first term in the derivative is always positive due to the positive sum of weighted utilities. If $u_i > u_d$ and $p > 0$, then $u_i^p - u_d^p > 0$, making the second term positive. Conversely, if $p < 0$, then $u_i^p - u_d^p < 0$, but since p^{-1} is also negative, the product remains positive. Thus, if $u_i > u_d$, the log-norm increases with w_i. A similar argument shows that the log-norm decreases if $u_i < u_d$.

For $p = 0$, we have

$$\begin{aligned}
\frac{d\log M(\mathbf{u}; \mathbf{w}, 0)}{dw_i} &= \frac{d}{dw_i}\left(\sum_{i=1}^{d-1} w_i \log u_i + \left(1 - \sum_{i=1}^{d-1} w_i\right)\log u_d\right) \\
&= \log\left(\frac{u_i}{u_d}\right) \\
&= \lim_{p\to 0}\frac{d\log M(\mathbf{u}; \mathbf{w}, p)}{dw_i}
\end{aligned}$$

This indicates that for $u_i > u_d$, the derivative is positive, implying an increase, and negative for $u_i < u_d$, implying a decrease. Thus, the function is monotonic for all $w_i \in [n-1]$. □

A.2. Proof of Lemma 3.1

$$\begin{aligned}
\log M(\mathbf{u};\mathbf{w},p) &= \frac{\log\left(\sum_{i=1}^d w_i u_i^p\right)}{p} \\
&= \frac{\log\left(u_{(1)}^p\left(\sum_{i=1}^d w_i\left(\frac{u_i}{u_{(1)}}\right)^p\right)\right)}{p} \\
&= \log u_{(1)} + \frac{\log\left(\sum_{i=1}^d w_i \exp\left(p\log\left(\frac{u_i}{u_{(1)}}\right)\right)\right)}{p} \\
&= \log u_{(1)} + \frac{1}{p}\log\left(\sum_{i=1}^d w_i \exp\left(p\log\left(u_{(d)}/u_{(1)}\right)\frac{\log\left(u_i/u_{(1)}\right)}{\log\left(u_{(d)}/u_{(1)}\right)}\right)\right)
\end{aligned}$$

Let $\mathbf{r} \in [0,1]^d$ such that

$$r_i = \exp\left(\frac{\log(u_i/u_{(1)})}{\log(u_{(d)}/u_{(1)})}\right)$$

and $q = p\log(u_{(d)}/u_{(1)})$. We then have

$$\begin{aligned}
\log M(\mathbf{u};\mathbf{w},p) &= \log u_{(1)} + \frac{1}{p}\log\left(\sum_{i=1}^d w_i r_i^{\left(p\log\left(u_{(d)}/u_{(1)}\right)\right)}\right) \\
\log M(\mathbf{u};\mathbf{w},p) &= \log u_{(1)} + \log\left(\frac{u_{(d)}}{u_{(1)}}\right)\frac{1}{q}\log\left(\sum_{i=1}^d w_i r_i^q\right)
\end{aligned}$$

□

A.3. Proof of Lemma 4.1

First, we state a lemma from Jameson (2006):

Lemma A.1 (Jameson (2006), Theorem 4.6)**.** *Let $f:\mathbb{R}\to\mathbb{R}$ be defined as $f(x) = \sum_{i=1}^n a_i \exp(p_i x)$, where $p_1 > p_2 > \ldots > p_n$ and $\sum_{i=1}^n a_i = 0$. Define $A_j := \sum_{i=1}^j a_i$ and denote by $S(A_j)$ the number of sign changes in the sequence $\{A_i\}_{i=1}^j$. Then, the number of unique zeros of f is at most $S(A_j)+1$.*

Consider the function $f(p) = \sum_{i=1}^d w_i u_i^p - \sum_{i=1}^d w_i v_i^p$ for $\mathbf{u},\mathbf{v}\in\mathbb{R}^d$ with disjoint entries:

$$f(p) = \sum_{i=1}^d w_i u_i^p - \sum_{i=1}^d w_i v_i^p = \sum_{i=1}^d w_i \exp(p\log u_i) - \sum_{i=1}^d w_i \exp(p\log v_i)$$

Applying Lemma A.1, if $w_i = w$ for all i, the sequence $\{A_j\}$, consisting of sums of w or $-w$, can have at most $d-1$ sign changes. A sign change at index k implies $A_{k-1}=0$, and the next sign change cannot occur before index $k+2$. Therefore, $f(p)$ has at most d zeros in this case. In the general case, where $w_i \neq w_j$ for some $i\neq j$, a sign change in $\{A_j\}$ can occur at any index except the first and the last. Thus, $f(p)$ can have at most $2d-1$ roots, as sign changes are possible at all intermediate indices. We conclude that $M_p(\mathbf{u};\mathbf{w},p) - M_p(\mathbf{v};\mathbf{w},p)$, defined over $\mathbb{R}\cup\{\pm\infty\}$, can change sign as a function of p at most $d-1$ times if $w_i = \frac{1}{d}$, and up to $2d-1$ times in the general case.

Lemma A.2 (Jameson (2006), Theorem 3.4)**.** *For any $k<d$ and $p_1 < \ldots < p_k \in \mathbb{R}$, there exist $\mathbf{u},\mathbf{v}\in\mathbb{R}_+^d$ and $\mathbf{w}\in\Delta_{d-1}$ such that $M_p(\mathbf{u};\mathbf{w},p_i) = M_p(\mathbf{v};\mathbf{w},p_i)$ for each $i\leq k$. Furthermore, the difference $M_p(\mathbf{u};\mathbf{w},p_i) - M_p(\mathbf{v};\mathbf{w},p_i)$ does not change sign within any interval (p_i,p_{i+1}).*

We now proceed to the proof of Lemma 4.1, which bounds the VC dimensions of the function classes $\mathcal{C}_{\mathbf{w},d}$ and $\mathcal{C}_d$.

Proof of Lemma 4.1 Part (a). As there are at most $2d-1$ roots to $f(p)$, there can be at most $2d-1$ sign changes as p varies from $-\infty$ to ∞. Consequently, the hypothesis class defined by all p (denoted as $\mathcal{M}_{\mathbf{w},d}$) is a subset of the hypothesis class that consists of at most $2d-1$ sign changes on the real line. This larger hypothesis class is denoted by $\mathcal{H}_d$, and we have $\text{VC}(\mathcal{M}_{\mathbf{w},d}) \leq \text{VC}(\mathcal{H}_d)$.

Let us consider m samples $\{\mathbf{u}_i, \mathbf{v}_i\}_{i=1}^m$. For each sample, sign changes occur at most $2d-1$ times, and hence the total number of changes in labeling over the entire real line is bounded by $(2d-1)m$ (as p changes, each change in labeling corresponds to a change in sign for at least one of the samples). This implies that the total number of possible labelings is $(2d-1)m+1$.

If the set of m samples is shattered, the upper bound derived above should be at least as large as the total number of labelings possible. We thus have:

$$(2d-1)m+1 \geq 2^m$$

We can show that $m = 2(\lceil \log_2 d \rceil + 1)$ points cannot be shattered. Consider

$$\begin{aligned}
2^m - (2d-1)m - 1 &= 2^{2(\lceil \log_2 d \rceil + 1)} - 2(2d-1)(\lceil \log_2 d \rceil + 1) - 1 \\
&\geq 2^{2(\log_2 d + 1)} - 4d\lceil \log_2 d \rceil + 2\lceil \log_2 d \rceil - 4d + 1 \\
&= 4d^2 - 4d\lceil \log_2 d \rceil + 2\lceil \log_2 d \rceil - 4d + 1 \\
&= 4d(d - \lceil \log_2 d \rceil) - 4d + 2\lceil \log_2 d \rceil + 1 \\
&\geq 4d - 4d + 2\lceil \log_2 d \rceil + 1 && (d - \lceil \log_2 d \rceil \geq 1 \forall d \in \mathbb{N}) \\
&= \lceil 2 \log_2 d \rceil + 1 > 0
\end{aligned}$$

Thus, $m > 2(\log_2 d + 1)$ points cannot be shattered, meaning that $VC(\mathcal{H}_d) < 2(\log_2 d + 1)$. □

We now bound the VC dimension for the unknown weight case. Consider $p \neq 0$. In this case, a hypothesis $C((\mathbf{u}, \mathbf{v}); \mathbf{w}, p)$ can be expressed as

$$\begin{aligned}
\text{sign}\left(\frac{\log\left(\sum_{i=1}^d w_i u_i^p\right) - \log\left(\sum_{i=1}^d w_i v_i^p\right)}{p}\right) &= \text{sign}(p)\, \text{sign}\left(\log\left(\sum_{i=1}^d w_i u_i^p\right) - \log\left(\sum_{i=1}^d w_i v_i^p\right)\right) \\
&= \text{sign}(p)\, \text{sign}\left(\left(\sum_{i=1}^d w_i u_i^p\right) - \left(\sum_{i=1}^d w_i v_i^p\right)\right) && \text{(log is increasing)} \\
&= \text{sign}(p)\, \text{sign}\left(\langle \mathbf{w}, \mathbf{u}^p - \mathbf{v}^p \rangle\right) = \text{sign}\left(\langle \mathbf{w}, \text{sign}(p)\left(\mathbf{u}^p - \mathbf{v}^p\right)\rangle\right)
\end{aligned}$$

where $\mathbf{u}^p = \begin{pmatrix} u_1^p & \cdots & u_d^p \end{pmatrix}^T$. Thus, for a fixed p, the set of viable $\mathbf{w}$'s spans a halfspace. We note that each component of $\text{sign}(p)\left(\mathbf{u}^p - bfv^p\right)$ is continuous, which means that $\langle \mathbf{w}, \text{sign}(p)\left(\mathbf{u}^p - \mathbf{v}^p\right)\rangle$ is a continuous function in $\mathbf{w}$ and p.

For $n > d$ samples $\{((\mathbf{u}_i, \mathbf{v}_i), y_i\}_{i=1}^n$, we define $\mathbf{h}_i(p) = \text{sign}(p)\left(\mathbf{u}^p - \mathbf{v}^p\right), i \in [n], p \neq 0$. For a fixed p, we note that the set of possible labelings for $\mathbf{w} \in \Delta_{d-1}$ is a subset of the set of possible labelings for $\mathbf{w} \in \mathbb{R}^d$, which in turn is the set of labelings generated by n hyperplanes. Since this problem has VC dimension d, the number of possible labelings for a fixed p is upper bounded by $(n+1)^d$. Let $\mathcal{B}(p)$ denote the set of possible labelings for hyperplanes defined by $\{\mathbf{h}_i(p)\}_{i=1}^n$ for a particular p.

Lemma A.3. *Let p_1 and p_2 have the same sign, with a labeling $\ell \in \{\pm 1\}^n$ such that $\ell \notin \mathcal{B}(p_1)$ but $\ell \in \mathcal{B}(p_2)$. Then, there is a $p \in [p_1, p_2]$ such that there is a set of d linearly dependent vectors $\mathbf{h}_{(1)}(p), \ldots, \mathbf{h}_{(d)}(p)$.*

Proof. Let ℓ be the labeling which is in $\mathcal{B}(p_2)$ but not in $\mathcal{B}(p_1)$. Since this labeling is not in $\mathcal{B}(p_1)$, for each $\mathbf{w}$, there is some hyperplane $\mathbf{h}_i(p_1)$ such that $\ell_i \langle \mathbf{w}, \mathbf{h}_i(p_1)\rangle < 0$. Since this labeling is in $\mathcal{B}(p_2)$, there is some $\mathbf{w}$ such that $\ell_i \langle \mathbf{w}, \mathbf{h}_i(p_2)\rangle \geq 0 \forall i \in [n]$.

Let $\mathfrak{B} \subset \mathbb{R}^d$ denote the unit hypersphere around the origin. Since the labelings are invariant to the scale of $\mathbf{w}$, the set of possible labelings for $\mathbf{w} \in \mathbb{R}^d$ is exactly the set of possible labelings for $\mathbf{w} \in \mathfrak{B}$

Consider the quantity $m(p) = \max_{\mathbf{w}\in\mathfrak{B}} \min_i \ell_i \langle \mathbf{w}, \mathbf{h}_i(p)\rangle$. We observe that if $m(p) < 0$, for each $\mathbf{w}$ there is some $i \in [n]$ such that $\ell_i \langle \mathbf{w}, \mathbf{h}_i(p)\rangle < 0$, i.e., the labeling is not attained at any point. On the other hand, if $m(p) \geq 0$, there is some $\mathbf{w}$ such that the labeling is attained at $\mathbf{w}$. Since $\ell_i \langle \mathbf{w}, \mathbf{h}_i(p)\rangle$ is a continuous function in $\mathbf{w}$ and p for all $i \in [n]$, $\min_i \ell_i \langle \mathbf{w}, \mathbf{h}_i(p)\rangle$ is also a continuous function in $\mathbf{w}$ and p. Thus, $m(p)$ is also a continuous function in p.

Using this fact and the intermediate value theorem, there should be some $p \in [p_1, p_2]$ such that $m(p) = 0$. Let $\mathbf{w}^* \in \mathfrak{B}$ be a vector at which $m(p) = 0$ is attained. We now show that at this p, at least d of the n vectors $\{\mathbf{h}_i(p)\}_{i=1}^n$ are linearly dependent.

Suppose this were not the case, i.e., any set of d vectors in the set is linearly independent. This means that at most $d-1$ of the vectors lie on the hyperplane $\{\mathbf{x} : \langle \mathbf{w}, \mathbf{x}\rangle = 0\}$. Let $\mathbf{h}_{(1)}(p), \ldots, \mathbf{h}_{(k)}(p)$ denote these vectors, with $k \leq d-1$. Since $m(p) = 0$, there should be at least one such vector. Let $\mathbf{H} \in \mathbb{R}^{k\times d}$ be the matrix with these vectors as the rows.

If these k vectors are linearly dependent, we can add any of the remaining $n-k$ vectors to get a set of d linearly dependent vectors. Let us consider the case where they are not linearly dependent. Consider $\{\mathbf{x} : \mathbf{H}\mathbf{x} = \mathbf{1}_k\}$. This is an underdetermined set of linear equations, and the set should be non-empty (because of linear independence of the k vectors). Let $\mathbf{x}_0$ be one of the vectors in this set.

Let $t > \max_{i\in[n]} -\frac{\langle \mathbf{w}, \mathbf{h}_i(p)\rangle}{\langle \mathbf{x}_0, \mathbf{h}_i(p)\rangle}$. We have

$$\begin{aligned}
\langle \mathbf{w} + t\mathbf{x}_0, h_i(p)\rangle &= \langle \mathbf{w}, h_i(p)\rangle + t\,\langle \mathbf{x}_0, h_i(p)\rangle \\
&> \langle \mathbf{w}, h_i(p)\rangle - \max_{j\in[n]} \frac{\langle \mathbf{w}, \mathbf{h}_j(p)\rangle}{\langle \mathbf{x}_0, \mathbf{h}_j(p)\rangle} \langle \mathbf{x}_0, \mathbf{h}_i(p)\rangle \\
&\geq \langle \mathbf{w}, h_i(p)\rangle - \frac{\langle \mathbf{w}, \mathbf{h}_i(p)\rangle}{\langle \mathbf{x}_0, \mathbf{h}_i(p)\rangle} \langle \mathbf{x}_0, \mathbf{h}_i(p)\rangle \\
&= 0
\end{aligned}$$

Thus, $\mathbf{w} + t\mathbf{x}_0$ is a point such that $\langle \mathbf{w} + t\mathbf{x}_0, \mathbf{h}_i(p)\rangle > 0$ for all $i \in [n]$. This means that $m(p) > 0$, which is a contradiction. Intuitively, this means that if any d vectors in $\{h_i(p)\}_{i=1}^n$ are linearly independent, then $m(p) > 0$. Thus, there should be a set of d linearly dependent vectors $h_{(1)}(p), \ldots, h_{(d)}(p)$. □

From the above lemma, we observe that any change in the set of labelings is accompanied by a p which gives d linearly dependent vectors.

Proof of Lemma 4.1 Part (b). Using the lemma above, to bound the number of possible labelings, we first bound the number of p's such that there are d linearly dependent vectors.

Consider a set of d vectors $h_1(p), \ldots, h_d(p)$. As the vectors are linearly dependent, the determinant of the matrix constructed using these vectors should be zero. We should thus have

$$\begin{vmatrix}
\text{sign}\,(p)\,(u_{11}^p - v_{11}^p) & \cdots & \text{sign}\,(p)\,(u_{1n}^p - v_{1d}^p) \\
\vdots & \ddots & \vdots \\
\text{sign}\,(p)\,(u_{d1}^p - v_{11}^p) & \cdots & \text{sign}\,(p)\,(u_{1n}^p - v_{dd}^p)
\end{vmatrix} = 0$$

Upon expanding the determinant, we get an equation of the form $\sum_{i=1}^m a_i u_i^p$ which has $2^d \cdot d!$ terms. From an earlier lemma, we know that this equation should have at most $2^d \cdot d! - 1$ roots. Upon adding the original configuration, we get $2^d \cdot d!$ possible configurations. The choice of the d vectors can be made in $\binom{n}{d}$ ways, and hence we have a bound on the possible changes as $2^d d! \binom{n}{d}$. In the worst case, we assume that all the labelings are changed, and we thus get an upper bound on the changes as

$$(n+1)^d 2^d d! \binom{n}{d}$$

In the beginning of the proof, we had carefully set aside $p = 0$. We now observe that $p = 0$ is a root of the above system of equations. Thus, we are implicitly considering any possible changes at $p = 0$ as well.

We can show that $n = 8(\lceil d \log_2 d \rceil + 1)$ points cannot be shattered.

$$\begin{aligned}(n+1)^d 2^d d! \binom{n}{d} &= (n+1)^d 2^d \cdot n(n-1)\dots(n-d+1) \\ &> (n+1)2^d n^{2d-1} \\ &> 2^{d-1} n^{2d}\end{aligned}$$

We now show that for every $d \in \mathbb{N}$, the inequality $2^{d-1}n^{2d} \leq 2^n$ holds, i.e. $n - d - 2d\log_2 n + 1 > 0$ for $n = 8(\lceil d \log_2 d \rceil + 1)$. For $d \in \{1, 2\}$, this statement can be verified directly. Therefore, it suffices to show that $f(x) := 8(x \log_2 x + 1) - 2x \log_2[8(x \log_2 x + 1)] - x + 1 > 0$ for any $x \geq 3$.

$$\begin{aligned}f(x) &> 8x \log_2 x - 2x \log_2(16x \log_2 x) - x + 9 && (x \log_2 x > 1) \\ &= 6x \log_2 x - 2x \log_2(\log_2 x) - 9x + 9 \\ &> 4x \log_2 x - 9x + 9 = g(x)\end{aligned}$$

We note that $g(3) = 12 \log_2 3 - 18 > 1 > 0$. Moreover, $g'(x) = 4 \log_2 x + 4/\log 2 - 9$, and we note that $g'(2) = 4/\log 2 - 5 > 0.77 > 0$, with $g'(x)$ being an increasing function. Thus, $f(x) > g(x) > 0$ for all $x \geq 3$. This means that $f(x) > 0$, which proves our bound. This implies that the VC dimension is bounded above by $8(d \log_2 d + 1)$ □

Proof of Lemma 4.1 Part (c). Take $m = \log_2(d) + 1$. Let $\sigma_1, \dots, \sigma_{2^m} \in \{\pm 1\}^m$ be a Gray code ordering of the set $\{\pm 1\}^m$, such that two successive values have a Hamming distance of 1, and that the number of changes in different bit positions is at most $\frac{2^m}{2} \leq d$ (for the existence of such an ordering, see (Bhat & Savage, 1996)). For $i < 2^m$, denote by $s_i \in [m]$ the bit position in which σ_i and σ_{i+1} differ. By using Lemma A.2 d times, there exists a sample $\{\mathbf{u}_i, \mathbf{v}_i\}_{i=1}^m$ and $p_1 < \dots < p_{2^m - 1}$, where p_i satisfies $\|\mathbf{u}_j\|_{p_i} = \|\mathbf{v}_j\|_{p_i}$ if $s_i = j$. Furthermore, define $p_0 = -\infty$, and note that each interval (p_i, p_{i+1}) for $0 \leq i < 2^m$ corresponds to a unique combination of labels over $\{\mathbf{u}_i, \mathbf{v}_i\}_{i=1}^m$. □

A.4. Proof of Lemma 3.2

We differentiate between two representations of Rademacher complexity:

$$\begin{aligned}\hat{\mathfrak{R}}(\mathcal{F}) &= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i f(x_i)\right] \\ \hat{\mathfrak{R}}_{abs}(\mathcal{F}) &= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^n \epsilon_i f(x_i)\right|\right]\end{aligned}$$

It is clear that $\hat{\mathfrak{R}}(\mathcal{F}) \leq \hat{\mathfrak{R}}_{abs}(\mathcal{F})$. We now list a few results related to Pollard's pseudo-dimension.

Definition A.4. (Pseudo-shattering) Let $\mathcal{H}$ be a set of real valued functions from input space $\mathcal{X}$. We say $C = (x_1, \dots, x_m)$ is pseudo-shattered if there exists a vector $r = (r_1, \dots, r_m)$ such that for all $b \in \{\pm 1\}^m = (b_1, \dots, b_m)$, there exists $h_b \in \mathcal{H}$ such that $\operatorname{sign}(h_b(x_i) - r_i) = b_i$.

Definition A.5. The pseudo-dimension $\operatorname{Pdim}(\mathcal{H})$ is the cardinality of the largest set pseudo-shattered by $\mathcal{H}$.

The following lemma connects pseudo-dimensions to VC dimensions:

Lemma A.6.

The following lemma bounds the Rademacher complexity using pseudo-dimension and covering numbers.

Lemma A.7. *For $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ with Pdim$(\mathcal{F}) \leq d$,*

$$\mathcal{N}(\epsilon, \mathcal{F}, d_n) \leq \left(\frac{c}{\epsilon}\right)^{2d}$$

where $d_n(f,g) = \left(\frac{1}{n}\sum_{i=1}^n \left(f(x_i) - g(x_i)\right)^2\right)^{1/2}$

We also have the following covering number bound for Rademacher complexity:

Lemma A.8. *For $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$,*

$$\hat{\mathfrak{R}}_{abs}(\mathcal{F}) \leq \inf_{\epsilon>0}\left(\sqrt{\frac{2\log 2\mathcal{N}(\epsilon, \mathcal{F}, d_n)}{n}} + \epsilon\right)$$

We now turn to bounding the complexity for the unknown and known weights cases.:

Proof of Lemma 3.2 Part (a). The function class is:

$$\mathcal{M}_{\mathbf{w},d} = \{\log M(\cdot; \mathbf{1}_d/d, p) | p \in \mathbb{R}\}$$

Moreover, from 2.1, it follows that for a fixed $\mathbf{u} \in \mathbb{R}^d$, $f_p(\mathbf{u})$ is a non-decreasing function with respect to p. Consequently, there exists a $p^* \in \mathbb{R} \cup \{\pm\infty\}$ such that for any $y \in (u_{\min}, u_{\max})$, we have $\log M(\mathbf{u}; \mathbf{w}, p) < y$ for all $p < p^*$, and $\log M(\mathbf{u}; \mathbf{w}, p) \geq y$ for all $p \geq p^*$. This implies that $B_{\log M}(\mathbf{u}, y) = \text{sign}(\log M(\mathbf{u}; \mathbf{w}, p) - y)$ changes its sign exactly once as p increases.

We note that for $B_{\log M}(x, y)$, one point can be shattered (by choosing $p < p^*$ and $p > p^*$). However, for two points $\mathbf{u}$ and $\mathbf{v}$, the number of times a sign change occurs with increasing p for either $\mathbf{u}$ or $\mathbf{v}$ is at most twice, meaning that only 3 labels can be achieved. Thus, 2 points cannot be shattered. □

Proof of Lemma 3.2 Part (b). The function class is:

$$\mathcal{M}_d = \{\log M(\cdot; \mathbf{w}, p) | p \in \mathbb{R}\}$$

We note that

$$\begin{aligned} B_{\log M}(\mathbf{u}, y) &= \text{sign}\left(\log M(\mathbf{u}; \mathbf{w}, p) - y\right) \\ &= \text{sign}\left(\log M(\mathbf{u}; \mathbf{w}, p) - \log M(y \cdot \mathbf{1}_d; \mathbf{w}, p)\right) \end{aligned}$$

which, we observe, is exactly the expression in the noiseless comparison-based setup for the unknown weights case. As we know that the VC dimension for this expression is upper bounded by $8(d\log_2 d + 1)$, our result is proved. □

A.5. Proof of Lemma 3.3

Proof. We prove the result for unknown weights - the result for known weights follows by replacing the pseudo-dimension bound of $\mathcal{S}_d$ by that of $\mathcal{S}_{\mathbf{w},d}$ from Lemma 3.2. Let d_p denote the pseudo-dimension for the unknown weights case.

$$\begin{aligned}
\hat{\mathfrak{R}}(\mathcal{M}_d) &= \mathbb{E}_\epsilon\left[\frac{1}{n}\sup_{(\mathbf{w},p)\in\Delta_{d-1}\times\mathbb{R}}\sum_{i=1}^n \epsilon_i \log M(\mathbf{u}_i;\mathbf{w},p)\right] \\
&= \mathbb{E}_\epsilon\left[\frac{1}{n}\sup_{(\mathbf{w},p)\in\Delta_{d-1}\times\mathbb{R}}\sum_{i=1}^n \epsilon_i \left(u_{i(1)} + \log\left(\frac{u_{i(d)}}{u_{i(1)}}\right)\log M(\mathbf{r}_i;\mathbf{w},q)\right)\right] \\
&= \mathbb{E}_\epsilon\left[\frac{1}{n}\sup_{(\mathbf{w},p)\in\Delta_{d-1}\times\mathbb{R}}\sum_{i=1}^n \epsilon_i \log\left(\frac{u_{i(d)}}{u_{i(1)}}\right)\log M(\mathbf{r}_i;\mathbf{w},q)\right] \\
&\le \mathbb{E}_\epsilon\left[\frac{1}{n}\sup_{(\mathbf{w},p)\in\Delta_{d-1}\times\mathbb{R}}\left|\sum_{i=1}^n \epsilon_i \log\left(\frac{u_{i(d)}}{u_{i(1)}}\right)\log M(\mathbf{r}_i;\mathbf{w},q)\right|\right] \\
&= \log\left(\frac{u_{i(d)}}{u_{i(1)}}\right)\mathbb{E}_\epsilon\left[\frac{1}{n}\sup_{(\mathbf{w},p)\in\Delta_{d-1}\times\mathbb{R}}\left|\sum_{i=1}^n \epsilon_i \log M(\mathbf{r}_i;\mathbf{w},q)\right|\right] \\
&\le \log\left(\frac{u_{\max}}{u_{\min}}\right)\mathbb{E}_\epsilon\left[\frac{1}{n}\sup_{(\mathbf{w},p)\in\Delta_{d-1}\times\mathbb{R}}\left|\sum_{i=1}^n \epsilon_i \log M(\mathbf{r}_i;\mathbf{w},q)\right|\right] \\
&= \kappa\hat{\mathfrak{R}}_{abs}(\mathcal{S}_d)
\end{aligned}$$

From Lemmas A.7 and A.8, and since $\log M(\mathbf{r};\mathbf{w},q)\in[0,1]$, we have

$$\begin{aligned}
\mathcal{N}(\epsilon,\mathcal{S}_d,d_n) &\le \left(\frac{c}{\epsilon}\right)^2 d_p \\
\implies \hat{\mathfrak{R}}_{abs}(\mathcal{S}_d) &\le \inf_{\epsilon>0}\left(\sqrt{\frac{2\log 2 + 4d_p\log(c/\epsilon)}{n}} + \epsilon\right) \\
&\le \sqrt{\frac{2\log 2 + 2d_p\log n}{n}} + \frac{c}{\sqrt{n}} \qquad \text{(setting } \epsilon = c/\sqrt{n}\text{)}
\end{aligned}$$

We thus have

$$\hat{\mathfrak{R}}(\mathcal{M}_d) = \kappa\left(\sqrt{\frac{2\log 2 + 2d_p\log n}{n}} + \frac{c}{\sqrt{n}}\right)$$

Replacing $d_p = 8(d\log_2 d + 1)$ gives us the required bound □

A.6. Proof for Theorem 3.4

Proof. We prove the result for the unknown weights case - the result for known weights follows a similar process. For ℓ_2 loss, our function class is

$$\mathcal{L}_2 = \left\{\ell_2(\log M(\mathbf{u};\mathbf{w},p),y) = (\log y - \log M(\mathbf{u};\mathbf{w},p))^2 | (\mathbf{w},p)\in\Delta_{d-1}\times\mathbb{R}\right\}$$

As $\log M(\mathbf{u};\mathbf{w},p), y_i \in [u_{i(1)}, u_{i(d)}] \subseteq [u_{\min}, u_{\max}]$, we have $\log y - \log M(\mathbf{u};\mathbf{w},p) \in [-\log(u_{\max}/u_{\min}), \log(u_{\max}, u_{\min})]$. Over a bounded range $[-\gamma,\gamma]$, $\ell_2(t,y) = (t-y)^2$ is 2γ Lipschitz continuous w.r.t. t. Thus, using Talagrand's contraction lemma and Lemma 3.3, we have

$$\hat{\mathfrak{R}}(\mathcal{L}_2) = \hat{\mathfrak{R}}(\ell_2\circ\mathcal{M}_d) \le 2\log\left(\frac{u_{\max}}{u_{\min}}\right)\hat{\mathfrak{R}}(\mathcal{M}_d) = 2\kappa\hat{\mathfrak{R}}(\mathcal{M}_d)$$

We then use the uniform convergence bounds for Rademacher complexity to get

$$\sup_{(\mathbf{w},p)\in\Delta_{d-1}\times\mathbb{R}} \left|\hat{R}_n(\mathbf{w},p) - R(\mathbf{w},p)\right| \leq 8\kappa\hat{\mathfrak{R}}(\mathcal{M}_d) + 3\sqrt{\frac{\log(4/\delta)}{2n}} = \epsilon$$

Thus,

$$\begin{aligned} R(\hat{\mathbf{w}},\hat{p}) - R(\mathbf{w},p) &= \left(\hat{R}_n(\hat{\mathbf{w}},\hat{p}) - \hat{R}_n(\mathbf{w},p)\right) + \left(\hat{R}_n(\mathbf{w},p) - R(\mathbf{w},p)\right) + \left(R(\hat{\mathbf{w}},\hat{p}) - \hat{R}_n(\hat{\mathbf{w}},\hat{p})\right) \\ &\leq 0 + \epsilon + \epsilon = 2\epsilon \\ &= 16\kappa\hat{\mathfrak{R}}(\mathcal{M}_d) + 6\sqrt{\frac{\log(4/\delta)}{2n}} \end{aligned}$$

Replacing $\hat{\mathfrak{R}}(\mathcal{M}_d)$ from Lemma 3.3 provides us with the required bounds. □

A.7. Proof for Theorem 4.3

Proof. We prove the result for unknown weights, with the known weights result following similar steps. We consider the function class $\mathcal{C}_d$ as in Section 4, with ℓ_{0-1} loss being $\ell_{0-1}(t,y) = (1+ty)/2, t, y \in \{\pm 1\}$. We observe that ℓ_{0-1} is $1/2$ Lipschitz w.r.t. t. Thus, by applying Theorem 3 of Natarajan et al. (2013), we observe that w.r.t. ℓ_{0-1} on the *noiseless* data distribution,

$$R(\hat{\mathbf{w}},\hat{p}) - R(\mathbf{w},p) \leq 4L_\rho\hat{\mathfrak{R}}(\mathcal{C}_d) + 2\sqrt{\frac{\log(1/\delta)}{2n}} \tag{2}$$

where $L_\rho = (1+|\rho_{+1} - \rho_{-1}|)L/(1-\rho_{+1}-\rho_{-1})$. Here, ρ_{+1} and ρ_{-1} are defined as the probability of mislabeling true positive and true negative examples, which in our case are the same value, ρ. Thus, $L_\rho = 1/(2(1-2\rho))$ in our case. We obtain $\hat{\mathfrak{R}}(\mathcal{C}_d)$ using the VC bound on Rademacher complexity:

$$\hat{\mathfrak{R}}(\mathcal{C}_d) \leq \sqrt{\frac{16(d\log_2 d + 1)\log(n+1)}{n}}$$

Substituting it in 2 concludes our proof. □

A.8. Proof for Theorem 5.1

Proof. We prove the result for unknown weights, with the case for known weights following similar steps. We first establish a bound on the Rademacher complexity of $\mathcal{T}_{\mathbf{w},d} = \{\tau\left(\log M(\cdot;\mathbf{w},p) - \log M(\cdot;\mathbf{w},p)\right)|\tau,p\}$. Let $\log M(\mathbf{u}_i;\mathbf{w},p) =$

$\log u_{i(1)} + \log(u_{i(d)}/u_{i(1)}) \log M(\mathbf{r}_i; \mathbf{w}, q)$, and $\log M(\mathbf{v}_i; \mathbf{w}, p) = \log v_{i(1)} + \log(v_{i(d)}/v_{i(1)}) \log M(\mathbf{r}'_i; \mathbf{w}, q)$.

$$
\begin{aligned}
\hat{\mathfrak{R}}(\mathcal{T}_d) &= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\sum_{i=1}^n \epsilon_i\tau\left(\log M(\mathbf{u}_i;\mathbf{w},p) - \log M(\mathbf{v}_i;\mathbf{w},p)\right)\right] \\
&\leq \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\sum_{i=1}^n \epsilon_i\tau\log M(\mathbf{u}_i;\mathbf{w},p)\right] + \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\sum_{i=1}^n (-\epsilon_i)\tau\log M(\mathbf{v}_i;\mathbf{w},p)\right] \\
&\leq \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\sum_{i=1}^n \epsilon_i\tau\log M(\mathbf{u}_i;\mathbf{w},p)\right] + \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\sum_{i=1}^n \epsilon_i\tau\log M(\mathbf{v}_i;\mathbf{w},p)\right] \\
&\leq \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\sum_{i=1}^n \epsilon_i\tau\left(\log u_{i(1)} + \log(u_{i(d)}/u_{i(1)})\log M(\mathbf{r}_i;\mathbf{w},q)\right)\right] \\
&\quad + \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\sum_{i=1}^n \epsilon_i\tau\left(\log v_{i(1)} + \log(v_{i(d)}/v_{i(1)})\log M(\mathbf{r}'_i;\mathbf{w},q)\right)\right] \\
&= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\sum_{i=1}^n \epsilon_i\tau\log(u_{i(d)}/u_{i(1)})\log M(\mathbf{r}_i;\mathbf{w},q)\right] \\
&\quad + \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\sum_{i=1}^n \epsilon_i\tau\log(v_{i(d)}/v_{i(1)})\log M(\mathbf{r}'_i;\mathbf{w},q)\right] \\
&\leq \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\left|\sum_{i=1}^n \epsilon_i\tau\log(u_{i(d)}/u_{i(1)})\log M(\mathbf{r}_i;\mathbf{w},q)\right|\right] \\
&\quad + \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\left|\sum_{i=1}^n \epsilon_i\tau\log(v_{i(d)}/v_{i(1)})\log M(\mathbf{r}'_i;\mathbf{w},q)\right|\right] \\
&= \frac{1}{n}\log\left(\frac{u_{i(d)}}{u_{i(1)}}\right)\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\left|\sum_{i=1}^n \epsilon_i\tau\log M(\mathbf{r}_i;\mathbf{w},q)\right|\right] \\
&\quad + \frac{1}{n}\log\left(\frac{v_{i(d)}}{v_{i(1)}}\right)\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\left|\sum_{i=1}^n \epsilon_i\tau\log M(\mathbf{r}'_i;\mathbf{w},q)\right|\right] \\
&\leq \frac{\kappa}{n}\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\left|\sum_{i=1}^n \epsilon_i\tau\log M(\mathbf{r}_i;\mathbf{w},q)\right|\right] + \frac{\kappa}{n}\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\left|\sum_{i=1}^n \epsilon_i\tau\log M(\mathbf{r}'_i;\mathbf{w},q)\right|\right] \\
&\leq \frac{\tau_{\max}\kappa}{n}\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\left|\sum_{i=1}^n \epsilon_i\log M(\mathbf{r}_i;\mathbf{w},q)\right|\right] + \frac{\tau_{\max}\kappa}{n}\mathbb{E}_\epsilon\left[\sup_{\tau,\mathbf{w},p}\left|\sum_{i=1}^n \epsilon_i\log M(\mathbf{r}'_i;\mathbf{w},q)\right|\right] \\
&= 2\tau_{\max}\kappa\hat{\mathfrak{R}}_{abs}(\mathcal{S}_d)
\end{aligned}
$$

We now use the bound on $\hat{\mathfrak{R}}_{abs}(\mathcal{S}_d)$ from Appendix A.5 to get the following bound on $\hat{\mathfrak{R}}(\mathcal{T}_d)$

$$
\hat{\mathfrak{R}}(\mathcal{T}_d) \leq 2\tau_{\max}\kappa\left(\sqrt{\frac{2\log 2 + 16(d\log_2 d + 1)\log n}{n}} + \frac{c}{\sqrt{n}}\right)
$$

We then use the uniform convergence bounds obtained using Rademacher complexity to obtain the following PAC bound:

$$
\begin{aligned}
R(\hat{\mathbf{w}},\hat{p}) - R(\mathbf{w},p) &= \left(\hat{R}_n(\hat{\mathbf{w}},\hat{p}) - \hat{R}_n(\mathbf{w},p)\right) + \left(\hat{R}_n(\mathbf{w},p) - R(\mathbf{w},p)\right) + \left(R(\hat{\mathbf{w}},\hat{p}) - \hat{R}_n(\hat{\mathbf{w}},\hat{p})\right) \\
&\leq 0 + \epsilon + \epsilon = 2\epsilon \\
&= 8\kappa\hat{\mathfrak{R}}(\mathcal{T}_d) + 6\sqrt{\frac{\log(4/\delta)}{2n}}
\end{aligned}
$$

□

B. Additional Experiments

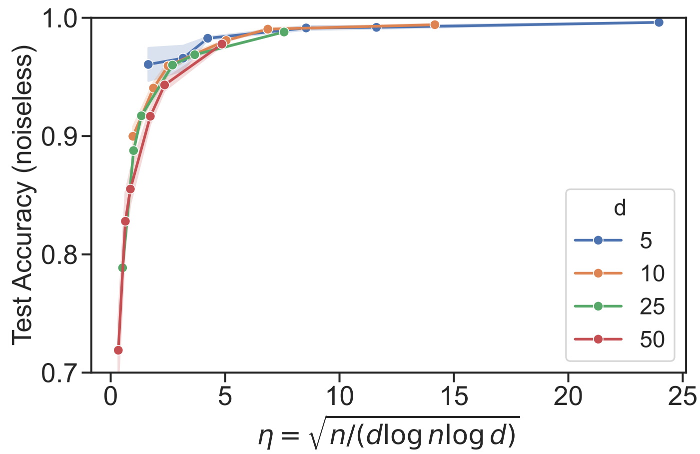


Figure 3. Verification of $\mathcal{O}(d \log d)$ risk bound for ordinal case with logistic noise, unknown weights

In Figure 3 we re-plot the test accuracy α on noiseless data against $\eta = \sqrt{n/(d \log n \log d)}$, a differently scaled version of Figure 2d. Theoretically, α and η are related as $1 - \alpha = \mathcal{O}(1/\eta)$. The alignment of all curves in Figure 3 as compared to the original curves in Figure 2d provides evidence that our risk and sample complexity bounds indeed scale as $d \log n \log d$ for the ordinal case with logistic noise and unknown weights.