When and How Often is Weighted Majority Vote Optimal Under Log Loss?

Steven An[1] **Sanjoy Dasgupta**[1]

[1]Computer Science Dept., University of California San Diego, La Jolla, California, USA

Abstract

Modern machine learning methods often require large, high quality labeled datasets, whose labels are potentially expensive and time consuming to obtain. One solution, seen in Programmatic Weak Supervision, crowdsourcing, and semi-supervised learning, is to cheaply obtain an ensemble of noisy labeling functions (LFs) and combine their predictions. Specifically, weighted majority vote (WMV) is a simple but well studied method to perform such a combination. Weighting strategies for WMV can be derived from a wide ranging set of assumptions (e.g. probabilistic, adversarial, etc.). However, existing analyses often suppose that the LF predictions are fixed, and characterize the conditions when said weighting strategies are optimal (among all weighting strategies). We take a different approach and show that all weighting strategies which only depend on LF accuracies, e.g. majority vote, are optimal (w.r.t. log loss) on a measure zero set of problems. A method to compute the proportion of problems where such aforementioned strategies are ϵ close to being optimal is presented and run. Other contributions include improved analysis of WMV's excess error under log loss.

1 INTRODUCTION

Obtaining a large labeled dataset often incurs high cost monetarily and temporally as a result of the need to hire domain experts to label each datapoint. Multiple areas, including Programmatic Weak Supervision, crowdsourcing, and semi-supervised learning have studied methods which can reduce the cost of acquiring the aforementioned labels. Specifically, one overarching strategy is to cheaply obtain labeling functions (LFs), e.g. crowdsourcing workers, small computer programs, linear classifiers, to noisily label the dataset, and to then combine the LF predictions to obtain a clean labeling. There are two natural questions. First, how should one combine the LF predictions? Second, when should a method of combination be chosen over another?

We focus on using weighted majority vote (WMV) to combine LF predictions and perform our analysis over log loss. Objects of comparison will be weighting strategies, i.e. functions that take information about the problem (LF accuracies, class frequencies, etc.), and assign weights to LFs. Our analysis focuses on two settings: determining when and how often a weighting strategy is optimal (or close to it) when the LF accuracies are known and its performance compared to other weighting strategies when it is not optimal. In particular, we focus on three weighting strategies: majority vote, an intuitive strategy with probabilistic underpinnings, and a consistent one with adversarial underpinnings.

A first pass approach is to give an LF weight proportional to its accuracy. This is intuitively satisfying – we ought to follow accurate LFs more closely and vice versa. As stated above, we consider the question of when and how often this strategy is optimal. The former can be done by determining the assumptions needed to derive this weighting strategy. For example, modeling each LF as a biased coin, i.e. taking the One-Coin Dawid-Skene generative assumption [Li and Yu, 2014], and looking at the posterior label distribution is sufficient. We show that this strategy is minimax optimal if the LF accuracies are fixed, but an adversary is allowed to choose both the LF predictions and ground truth labeling. Thus, satisfying the probabilistic assumptions is not necessary for guaranteeing the performance of this weighting strategy. For the latter question of how often this strategy is optimal, this strategy is optimal on a measure zero set of problems because it does not depend on the LF predictions. This follows as a special case of a more general result we show. We show how to experimentally determine the proportion of problems where this strategy is ϵ close to optimal.

By taking restricting the adversary even more – only allowing it to select the ground truth and not the LF pre-

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:87–132.

dictions, one can obtain a weighting strategy that is optimal for all problems (given the LF accuracies are known). This class of weighting strategies depends on two choices of loss functions. One to measure how far the adversary's choice of ground truth is from our prediction and another to determine the set of labelings the adversary can choose from. This paper will consider log-loss and 0-1 loss respectively, like in [An and Dasgupta, 2024]. A wide variety of losses have been studied, e.g. [Balsubramani and Freund, 2016], [Mazzetto et al., 2021a]. To compare this strategy with the one above, or for any other weighting strategy in the general case where LF accuracies are estimated and not known, we must consider the excess (or generalization) error incurred with respect to the optimal weights.

We show that an existing bound of excess error from [An and Dasgupta, 2024] can be arbitrarily far away from the actual excess loss. Indeed, we conjecture that a similar result holds for bounds derived using similar techniques. This leads to a pessimistic sufficient condition (involving a quantity that we do not know how to compute in practice) for determining when the above adversarial weighting strategy is better than the intuitive one. We instead provide an exact expression of the difference of excess log loss incurred by two arbitrary weighting strategies, allowing us to give a sufficient condition computable in practice for determining when a weighting strategy is better than another. We also provide covariate shift analysis, resulting in a characterization of when one will pick a suboptimal weighting. Our contributions are as follows.

1. An adversarial derivation of the weighting strategy that assigns an LF weight proportional to its accuracy.
2. A characterization of when weights are optimal in terms of LF predictions.
3. A theorem showing that all weighting strategies that only depend on LF accuracies are optimal on a measure zero set of problems under log loss.
4. A method to compute the proportion of problems where a weighting strategy is optimal.
5. Necessary and sufficient conditions for a weighting strategy to outperform another.
6. Experiments to demonstrate the above method and to check our theoretical results.

2 RELATED WORK

While basic, the performance of the majority vote weighting strategy, e.g. weight 1 for every LF, has been well studied. One line of work is in jury theorems, which attempts to characterize the performance of a weighting strategy given some assumptions about the LFs. An early example is from Condorcet [1785], which considers a simple setting where the votes are modeled as a Binomial random variable. I.e. each voter is independent and has the same probability of predicting the correct label. A good survey of jury theorems can be found in [Dietrich and Spiekermann, 2023]. The papers below and our own work can be thought of as in the domain of jury theorems. Narasimhamurthy [2005] perform an adversarial analysis of majority vote under 0-1 loss in the binary labeling case. Li and Yu [2014] show that majority vote is consistent under the general Dawid-Skene generative assumption so long as the average accuracy of LFs is better than random noise. Another line of analysis involves picking out a subset of LFs to perform majority vote on, with the hope of improving over majority vote over all LFs. Ruta and Gabrys [2005] perform an extensive survey of such methods. Mazzetto et al. [2021b] study the same problem in the adversarial setting, with the goal of maximizing the minimum accuracy of the resulting subset of LFs. In another line of work, Littlestone and Warmuth [1994] study weighted majority vote in the online setting.

Now, we give a brief overview of work in different areas, e.g. Programmatic Weak Supervision (PWS), crowdsourcing, and semi-supervised learning, focusing on voting algorithms. For PWS, see [Zhang et al., 2022] for a good survey. The label model of PWS considers the problem of combining LF predictions to denoise their predictions. For example, Ratner et al. [2016] generalize the Dawid-Skene model by modeling dependencies, Fu et al. [2020] propose a fast triplet method, Rühling Cachay et al. [2021] propose using a neural network to combine LF predictions, Wu et al. [2023] propose using a graph neural network. The extant Bayesian PWS models can all be interpreted as weighted majority algorithms: [Kim and Ghahramani, 2012], [Li et al., 2019], [Zhang et al., 2023], [An, 2025].

On the crowdsourcing front, there is the celebrated work of Dawid and Skene [1979], for which we'll study a simplified version of their model. Karger et al. [2011] propose a Belief Propagation inspired algorithm whose output is created via weighted majority vote. Tao et al. [2019] perform their analysis from a domain adaptation point of view and leverage the datapoint features to help compute LF weights. Tian and Zhu [2015] propose a weighted majority vote which maximizes the margin between scores given to the potential true label and other labels.

The performance of the Dawid-Skene (DS) model, which is a generalized version of the weighted majority vote has also been studied extensively. In particular, DS gives each LF k^2 weights, one for each LF prediction/putative label pair, i.e. one for each entry of the confusion matrix. If an LF predicts class ℓ, then the ℓ^{th} row of the confusion matrix is used to create k votes, one for each class. Nitzan and Paroush [1982] show that if the LFs are independent, then the optimal weighting strategy in the binary case is to use the DS weights. Gao and Zhou [2013] provide minimax error convergence rates for the projected EM algorithm in the binary labeling case. Li and Yu [2014] provide expo-

nentially decreasing error bounds for the accuracy of DS while Aeeneh et al. [2025] improve on these bounds. Zhang et al. [2014] propose an initialization method for DS and prove error bounds for the use of EM with their initialization. Berend and Kontorovich [2015] analyze the error of the Naive Bayes model, which is a DS-type model, given supervised estimates of LF confusion matrices. In these works, the author assumes that the DS generative process holds, which allows them to control what the LF predictions look like. In this paper, our analysis will have no assumptions on the LF predictions.

The consistent adversarial weighting strategy discussed above has been studied under the semi-supervised setting. It was originally proposed by Balsubramani and Freund [2015c,b] for 0-1 loss. An and Dasgupta [2024] study the same model but under log loss. Mazuelas et al. [2023] study a similar model, but with more general constraints, which was generalized in [Mazuelas et al., 2022]. These papers use sensitivity analysis of convex programs to derive error bounds. We'll later see an example where that technique yields a weak bound. Other techniques to derive error bounds have appeared in the literature, e.g. PAC-Bayes [Balsubramani and Freund, 2015a], Rademacher averages [Mazzetto et al., 2021a]. For a generalized treatment of the minimax games studied in the above papers, see [Grünwald and Dawid, 2004].

3 PRELIMINARIES

Say we have feature space $\mathcal{X}$ and k labels $\mathcal{Y} = [k] = \{1, \ldots, k\}$. For $X = \{x_1, \ldots, x_n\} \subset \mathcal{X}$, we estimate the conditional label probability $Pr(y_i = \ell \mid x_i)$ using p labeling functions (LFs) $h^{(1)}, \ldots, h^{(p)}$, where $h^{(j)} \colon \mathcal{X} \to \mathcal{Y}$.

Consider weighted majority votes which only depend on the LF predictions. In such a scenario, we can describe X in terms of the LF predictions. That is, observe that there are $\tau \coloneqq k^p$ realizations of $(h^{(1)}(x), \ldots, h^{(p)}(x))$. Call that vector of predictions a *pattern*. Fixing an enumeration of patterns, we will denote the t^{th} pattern by π_t and the prediction of $h^{(j)}$ on that pattern is written as π_{tj}. By definition, the distribution of patterns determines the LF predictions.

Now, if on x_i the LFs predictions match the t^{th} pattern, we estimate $Pr(y_i = \ell \mid x_i)$ by $\widehat{\eta}_{t\ell} = Pr(y = \ell \mid \pi_t)$ or

$$\widehat{\eta}_{t\ell} = Pr(y = \ell \mid h^{(j)}(x) = \pi_{tj}, \forall j \in [p]).$$

For bookkeeping, write $\widehat{\eta}_t = (\widehat{\eta}_{t1}, \ldots, \widehat{\eta}_{tk})$ and $\widehat{\eta} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_\tau)$. In general, we will use carets to denote conditional distributions. Quantities without the carets will be joint distributions, i.e. $\eta \in \Delta_{\tau k}$ where

$$\eta_{t\ell} = Pr(y = \ell, h^{(j)}(x) = \pi_{tj}, \forall j \in [p]).$$

These are considered because we allow the LF predictions to vary, i.e. $Pr(\pi_t)$ varies. We will soon discuss how to construct our estimate for this distribution.

For joint distribution η, the following quantities are important to our analysis. The true LF accuracies are

$$b_j^* = Pr(h^{(j)}(x) = y) = \sum_{t=1}^{\tau} \eta_{t\pi_{tj}}. \tag{1}$$

The true class frequencies are

$$w_\ell^* = Pr(y = \ell) = \sum_{t=1}^{\tau} \eta_{t\ell} \tag{2}$$

and the true pattern frequencies are

$$\alpha_t^* = Pr(\pi_t) = \sum_{\ell=1}^{k} \eta_{t\ell}. \tag{3}$$

When labeling n points, α^* is known if X is known.

3.1 AN EXPONENTIAL FAMILY OF DISTRIBUTIONS

Formally, we seek a function that maps from a pattern to distributions over a class. Specifically, a function of the form $\widehat{g} \colon [\tau] \to \Delta_k$. Our choice for $\widehat{g}$ will lie in an exponential family of distributions (which'll be derived later). From $\widehat{g}$ we'll obtain $g \in \Delta_{k\tau}$, which will be used to estimate η.

Lets say that the LF predictions on datapoint x_i match pattern π_t. We will predict a distribution $\widehat{g}_t \in \Delta_k$ for that datapoint by getting a score for each class and then taking the softmax. For notational convenience, define $\pi_t^\ell \subseteq [p]$ as the set LFs that predict class ℓ on pattern t. Class ℓ's score is the class bias for ℓ plus weights of LFs that predicted ℓ on π_t, i.e. the sum of all θ_j where $j \in \pi_t^\ell$.

So, we assign each LF and class a weight, denoted $\theta \in \mathbb{R}^{p+k}$. Our estimate for $\widehat{\eta}_{t\ell}$ is

$$\widehat{g}_{t\ell}^{(\theta)} = \frac{\exp\left(\theta_{p+\ell} + \sum_{j \in \pi_t^\ell} \theta_j\right)}{\sum_{\ell'=1}^{k} \exp\left(\theta_{p+\ell'} + \sum_{j' \in \pi_t^{\ell'}} \theta_{j'}\right)}. \tag{4}$$

To get $g \in \Delta_{\tau k}$, we need a pattern distribution $\alpha \in \Delta_\tau$. One defines $g_{t\ell} \coloneqq \alpha_t \widehat{g}_{t\ell}^{(\theta)}$ or alternatively uses τ more weights.

Define $\nu \in \mathbb{R}^\tau$ as weights that encode the pattern distribution. Our estimate of $\eta_{t\ell}$ is

$$g_{t\ell}^{(\theta,\nu)} \propto \exp\left(\nu_t + \theta_{p+\ell} + \sum_{j \in \pi_t^\ell} \theta_j\right). \tag{5}$$

If we want g to have pattern distribution α, we choose $\nu_t = \log(\alpha_t / \widehat{Z}_t)$ where $\widehat{Z}_t$ is the normalization factor in Equation 4. We write $g^{(\theta),\alpha} \coloneqq g_{t\ell}^{(\theta,\nu)} = \alpha_t \widehat{g}_{t\ell}$ in that case. If a weighting strategy only specifies θ, then we will assume ν is chosen as above for $\alpha = \alpha^*$. This happens when α^* is known by way of knowing X.

We'll call our family of distributions $\mathcal{G} = \{g^{(\theta,\nu)} \colon \theta \in \mathbb{R}^{p+k}, \nu \in \mathbb{R}^\tau\}$ and will now write it using matrix notation.

3.1.1 A Matrix Representation

To compactly describe $\mathcal{G}$, we introduce two matrices. Let $A \in \{0,1\}^{(p+k)\times\tau k}$ be the matrix that encodes combinations of θ, while $D \in \{0,1\}^{\tau\times\tau k}$ be the matrix that picks out the correct pattern weight ν_t.

Define τ matrices $A^t \in \{0,1\}^{(p+k)\times k}$ so that A is the horizontal concatenation of these matrices. Then,

$$A^t = a^t_{j\ell} = \begin{cases} \mathbf{1}(\pi_{tj} = \ell) & \text{if } j \le p, \\ \mathbf{1}(j - p = \ell) & \text{if } j > p. \end{cases} \tag{6}$$

For D, we'll do something similar with $D^t \in \{0,1\}^{\tau\times k}$.

$$D^t = d^t_{j\ell} = \begin{cases} 1 & \text{if } j = t, \\ 0 & \text{o.w.} \end{cases} \tag{7}$$

Now, it's easy to check that for $\eta \in \Delta_{\tau k}$ that

$$A\eta = (b^*; w^*) \quad \text{and} \quad D\eta = \alpha^*.$$

Now, observe that $\widehat{g}^{(\theta)}_{t\ell} \propto \exp([A^\top\theta]_{t\ell})$. Also, if the pattern distribution of g is α, then $g = D^\top\alpha \circ \widehat{g}$ where $\circ$ is the Hadamard product.

4 EXAMPLES OF WEIGHTING STRATEGIES

We now give three examples of weighting strategies which will be used to instantiate our analysis. Specifically, we discuss majority vote, the intuitive strategy of weighting LFs by their accuracies, and the (adversarial) consistent strategy. When discussing weighting strategies, we will focus on the form of $\theta \in \mathbb{R}^{p+k}$, the weights that determine the estimate of the conditional label distribution $\widehat{g}^{(\theta)}$. This is because we want to study the quality of $g^{(\theta),\alpha}$ as an approximator of η where $\alpha \in \Delta_\tau$ varies. In other words, we want to study how the LF predictions affect how good the conditional label distribution estimates $\widehat{g}^{(\theta)}$ are.

Take for example majority vote (MV). Here, $\theta^{mv} = (\mathbf{1}_p; \mathbf{1}_k)$ where the class bias weights do not affect the prediction. For known X, the MV prediction is

$$g^{(\theta^{mv}),\alpha^*}_{t\ell} = \frac{\alpha^*_t \exp(|\pi^\ell_t|)}{\sum_{\ell'=1}^k \exp(|\pi^{\ell'}_t|)}.$$

It is not hard to construct pattern distributions α^* where majority vote does well/poorly. For example, consider our usual notion of majority vote: we predict the class ℓ on pattern t with the most votes. I.e. $\ell = \arg\max_{\ell'} |\pi^{\ell'}_t|$. Consider an X where $|X| = 3, p = 3, k = 2$ where the labels are all $+1$. If the LF predictions are $(+1,+1,-1)$, $(+1,-1,+1)$, $(-1,+1,+1)$, then majority vote is perfect, while each LF has accuracy of $2/3$.

4.1 AN INTUITIVE WEIGHTING STRATEGY

While majority vote can work well in practice, it is agnostic to LF quality, the class distribution, etc. This makes it susceptible to pathological cases such as the hammer spammer scenario where low numbers of high accuracy LFs (hammers) are overwhelmed by a large number of LFs who are essentially random noise (spammers), or to large numbers of adversarial LFs that consistently predict the wrong label.

A natural shift in strategy is to give an LF weight proportional to its accuracy. I.e. it's intuitively satisfying that LFs that are often right have more "say" in the final prediction and vice versa. Similarly, LFs that are noisy *prima facie* do not contribute anything, so they ought to have little to no weight. The same can be said about class biases. If most of the datapoints have class ℓ, i.e. $w^*_\ell \gg 1/k$, then the scores for class ℓ should be large unless enough highly accurate LFs predict a minority class.

Consider the following weights for $j \in [p], \ell \in [k]$:

$$\theta^{ds}_j = \log\left(\frac{b_j(k-1)}{1-b_j}\right) \quad \text{and} \quad \theta^{ds}_{p+\ell} = \log(w_\ell). \tag{8}$$

The superscript is short for the One-Coin Dawid-Skene model [Li and Yu, 2014] under which this weighting strategy can be derived. One may check that $\theta^{ds}_j = 0$ when $b_j = 1/k$ and $\theta^{ds}_j \uparrow \infty$ (resp. $\downarrow -\infty$) when $b_j \uparrow 1$ (resp. $\downarrow 0$). Note that ν is not defined for this strategy.

The generative process may be stated thusly. Labels and LF predictions are generated independently for each datapoint x. The label is drawn from $y \sim Categorical(w^*)$ and $h^{(j)}(x) = y$ w.p. b^*_j, while it equals an incorrect label $[k]\setminus y$ w.p. $(1-b_j)/(k-1)$. I.e. each LF is modeled as a biased coin. The full derivation is given in the appendix. Since one does not know b^*, w^* in practice, one estimates those quantities, e.g. by EM or something more sophisticated [Zhang et al., 2014], and then uses those weights.

While not depending on the LF predictions makes the characterization of θ^{ds} simple, this will also be the reason that this strategy is "optimal" on a measure zero set of pattern distributions. More formally, we will define the notion of optimality and prove that the measure of $\alpha \in \Delta_\tau$ such that $g^{(\theta^{ds}),\alpha}$ is optimal is zero. To emphasize the fact that the probabilistic assumptions are not the problem, we derive this same strategy using adversarial assumptions.

Say we only know that the LF accuracies b^* and class frequencies w^* and we need to pick a labeling $g \in \Delta_{\tau k}$ to estimate η. Here, we don't know anything about what the LF predictions look like, i.e. we can only say $\alpha^* \in \Delta_\tau$. Now, say an adversary picks z to maximize the loss of our prediction $-z^\top \log g$, with the constraint that η must give the LFs accuracy b^* and must have class frequency w^*, i.e. $Az = (b^*; w^*)$. It turns out the minimax optimal prediction is $g^{(\theta^{ds},\mathbf{0}_\tau)}$ where θ^{ds} uses b^*, w^*.

Formally, define the set of labelings from which the adversary is allowed to choose labelings from as

$$P_{b^*w^*} = \{z \in \Delta_{\tau k} \colon \|Az - (b^*; w^*)\|_\infty \leq 0\}. \quad (9)$$

Our result is written as follows.

Theorem 4.1. *Fix $b^* \in (0,1)^p, w^* \in \Delta_k \cap (0,1)^k$. Then, $g^{(\theta^{ds}),\alpha^{ds}} \in \Delta_{\tau k}$ is the minimax optimal prediction if the adversary chooses distributions in $P_{b^*w^*}$. I.e.*

$$g^{(\theta^{ds}),\alpha^{ds}} = g^{(\theta^{ds},\boldsymbol{\theta}_\tau)} = \arg\min_{g \in \Delta_{\tau k}} \max_{z \in P_{b^*w^*}} -z^\top \log g.$$

We provide the complete proof in the appendix, but note that it suffices to show $g^{(\theta^{ds},\mathbf{0}_\tau)}$ is the maximum entropy distribution in $P_{b^*w^*}$.

For both assumptions, there is an ignorance of what the LF predictions are. In the probabilistic derivation, the predictions on different datapoints are modeled as independent. In the adversarial derivation, one is assumed to be ignorant of what the LF predictions are. Note that we do not mean to say that the only time θ^{ds} is "optimal" is when $\nu = \mathbf{0}_\tau$. We'll soon define the notion of optimality and shall see that there are other choices of ν which make θ^{ds} optimal.

Indeed, Proposition 3 of Mazzetto et al. [2021b], though not presented in the same way, shows a similar result for $0-1$ loss rather than log loss. Specifically, under mild conditions, they show that the minimax optimal prediction one ought to take when the LF predictions are not known is match the predictions of the LF with highest accuracy.

4.2 A CONSISTENT WEIGHTING STRATEGY

We'll now present a weighting strategy that produces weights θ^{bf} which is optimal on all feasible pattern distributions α given b^*, w^*. Suppose we once again knew what b^*, w^*, but we also knew what α^* was. Then, we construct a set of labelings consistent with this knowledge like above with $P_{b^*w^*\alpha^*} = P_{b^*w^*} \cap P_{\alpha^*}$ where

$$P_{\alpha^*} = \{z \in \Delta_{\tau k} \colon \|Dz - \alpha^*\|_\infty \leq 0\}. \quad (10)$$

The resulting prediction is

$$g^{(\theta^{bf}),\alpha^*} = \arg\min_{g \in \Delta_{\tau k}} \max_{z \in P_{b^*w^*\alpha^*}} -z^\top \log g. \quad (11)$$

This is a reformulation of the Balsubramani-Freund (BF) model with log loss [An and Dasgupta, 2024] where the pattern distribution is explicitly stated. In the appendix, we show how θ^{bf} is the result of a convex program. θ^{bf} depends on α^* and it is an open question what the closed form solution of θ^{bf} looks like in terms of b^*, w^*, α^*. In practice, we use $P^\epsilon_{bw} \cap P_{\alpha^*}$ where b, w are estimates of b^*, w^* such that $\|(b; w) - (b^*; w^*)\|_\infty \leq \epsilon$. That is, P^ϵ_{bw} is the set in Equation 9 with ϵ replacing 0. Should one have L labeled points, ϵ is $O(1/\sqrt{L})$.

We take the notion of optimality from [An and Dasgupta, 2024], where a prediction in $g^* \in \mathcal{G}$ is optimal if it minimizes the distance to η with respect to KL divergence. A weighting strategy is consistent if for every tuple of b^*, w^*, α^* where $P_{b^*w^*\alpha^*}$ is nonempty, said weighting strategy returns g^*. In Theorem 6 of [An and Dasgupta, 2024], this was shown to hold for $g^{(\theta^{bf}),\alpha^*}$, while in Lemma 11, the OCDS strategy is shown to be inconsistent.

5 WHEN IS A WEIGHTING STRATEGY OPTIMAL?

Our discussion of what makes a weighting strategy optimal will closely follow [An and Dasgupta, 2024]. The proofs for these results are similar to those in the aforementioned paper and are included in the appendix.

The distance between two distributions $u, v \in \Delta_{\tau k}$ is measured by KL divergence. That is,

$$D_{KL}(u||v) = \sum_{t=1}^{\tau} \sum_{\ell=1}^{k} u_{t\ell} \log\left(\frac{u_{t\ell}}{v_{t\ell}}\right).$$

The optimal prediction $g^* \in \mathcal{G}$ is such that

$$g^* = \arg\min_{g \in \mathcal{G}} D_{KL}(\eta||g).$$

By Theorem 6 of [An and Dasgupta, 2024], we know that g^* is the solution of the BF problem in Equation 10 and is unique. They show that by looking at the KKT conditions, if one can exhibit $g \in \mathcal{G}$ such that $Ag = A\eta = (b^*; w^*)$ and $Dg = D\eta = \alpha^*$, then $g = g^*$.

By choosing KL divergence, one may show a Pythagorean theorem, making it easy to measure the excess loss of an arbitrary prediction $g \in \mathcal{G}$. See appendix for more details.

Lemma 5.1. *Fix $\eta \in \Delta_{\tau k}$. For any $g \in \mathcal{G}$,*

$$D_{KL}(\eta||g) = \underbrace{D_{KL}(\eta||g^*)}_{\textit{unavoidable loss}} + \underbrace{D_{KL}(g^*||g)}_{\textit{avoidable loss}}.$$

Our focus for this paper will be on the second term, which is how much excess loss one incurs by choosing suboptimal weights. In particular, we are interested in when it is equal to 0 as a function of the true pattern distribution α^*. (In [An and Dasgupta, 2024], the two terms on the RHS are called model and approximation uncertainty.)

6 HOW OFTEN IS A WEIGHTING STRATEGY (CLOSE TO) OPTIMAL?

For chosen weights θ, ν, recall that our predicted labeling $g^{(\theta,\nu)} \in \mathcal{G}$ is optimal if $Ag^{(\theta,\nu)} = A\eta = (b^*; w^*)$ and

$Dg^{(\theta,\nu)} = D\eta = \alpha^*$. For the three weighting strategies discussed, they all take b^*, w^*, α^* (or estimates thereof) as input. Moreover, each of the three weighting strategies satisfy $Dg^{(\theta,\nu)} = D\eta = \alpha^*$ either by assumption or construction. Recall that if one knows X, α^* can be computed by looking at the LF predictions on X.

So, it suffices to check when our prediction is such that $Ag^{(\theta),\alpha^*} = A\eta$ where θ is determined by a weighting strategy. We go about this by fixing b^*, w^* and varying α^*. The quantity we're interested in is the probability that a weighting strategy will output g^* for a randomly drawn pattern distribution α^* given that the LF accuracies are b^* and the class frequencies are w^*. I.e. for θ from MV/OCDS/BF etc.,

$$Pr(\theta \text{ is optimal} | b^*, w^*) = Pr(g^{(\theta),\alpha^*} = g^* | b^*, w^*).$$

When θ is from BF, this quantity is 1 by construction (based on BF's consistency), while this same quantity is 0 for MV and OCDS (under mild assumptions).

We'll analyze this probability via Bayes' Theorem. Let $vol(\cdot)$ be the volume operator operating under dimension $\tau - 1$. (I.e. the dimension of Δ_τ.) The joint probability

$$\begin{aligned} &Pr(g^{(\theta),\alpha^*} = g^*, b^*, w^*) \\ &\propto vol(\{\alpha \in \Delta_\tau \colon \|Ag^{(\theta),\alpha} - (b^*; w^*)\|_\infty \leq 0\}). \end{aligned} \quad (12)$$

Call this set being measured $O^{\theta,0}$. Like before, we'll only write the second superscript when it is ϵ (i.e. when the we have $\|\cdot\|_\infty \leq \epsilon$). If $\theta = \theta^{bf}$, we'll write O^{bf} as shorthand for $O^{\theta^{bf}}$. The probability that a pattern distribution can give rise to LF accuracies b^* and class frequencies w^* is

$$\begin{aligned} Pr(b^*, w^*) \propto vol(\{\alpha \in \Delta_\tau \colon \exists z \in \Delta_{\tau k}, \\ \|Az - (b^*; w^*)\|_\infty \leq 0, Dz = \alpha\}). \end{aligned}$$

Call this set T^0. The superscript will be suppressed unless it is nonzero. Note that it can be compactly written as $DP_{b^*w^*}$, the linear transformation of $P_{b^*w^*}$ by D. We will address the question of $Pr(b^*, w^*)$ being positive soon.

We can also restrict the pattern distributions to those realizable over n datapoints. That set is defined as

$$S_n = \{\alpha \in \Delta_\tau \colon n\alpha \in \mathbb{Z}^\tau_{\geq 0}\}. \quad (13)$$

Denote the respective finite versions of the above sets as

$$\tilde{O}^\theta_n = O^\theta \cap S_n \quad \text{and} \quad \tilde{T}_n = T \cap S_n.$$

When referring to the finite case, we'll use $Pr_n(\cdot)$ and use set cardinality rather than volume to measure size.

It turns out that $vol(T) > 0$ can be guaranteed under some mild, soon-to-be stated conditions on p, k, b^*, w^*. Specifically, we'll show that $dim(T) = \tau - 1$ by constructing a simplex of that same dimension with nonzero volume. This is possible because knowing b^*, w^* puts but a mild constraint on the pattern distribution α. In particular, we are able to exhibit τ linearly independent pattern distributions where the t^{th} pattern distribution $\alpha^{(t)}$ is 0 on the t^{th} element, i.e. $\alpha^{(t)}_t = 0$. On the other hand, note that $|\tilde{T}_n| = 0$ for certain choices of n, b^*, w^*. See appendix for more details.

Lemma 6.1 (Informal)**.** $vol(T) > 0$.

Now, we may formally state that BF returns g^*, i.e. is optimal, over all problems (when b^*, w^*, α^* are known). When those quantities are known, one can show that BF minimizes $-\eta^\top \log g$ for $g \in \mathcal{G}$, giving the result below.

Theorem 6.2. *Let θ^{bf} be the weights returned by the BF weighting strategy. Then, $O^{bf} = T$ so that $Pr(g^{(\theta^{bf}),\alpha^*} = g^* | b^*, w^*) = 1$. A fortiori, for all $n \geq 1$, $\tilde{O}^{bf}_n = \tilde{T}_n$ so that $Pr_n(g^{(\theta^{bf}),\alpha^*} = g^* | b^*, w^*) = 1$.*

Recall that θ^{bf} depends on b^*, w^*, α^*, i.e. an LF's weight depends not only on its accuracy, but also how other LFs predict. If this is not the case, i.e. θ only depends on b^*, w^*, the above probability is 0.

Theorem 6.3. *Fix $b^* \in (0,1)^p$, $w^* \in \Delta_k \cap (0,1)^k$. Suppose θ only depends on b^*, w^* and has at least one nonzero element. If $p \geq 3$, $k \geq 3$, and*

$$\max_{j \in [p]} b^*_j \leq (1 + 1/\sqrt{p(p-1)})^{-1},$$

then $Pr(g^{(\theta),\alpha^*} = g^* | b^*, w^*) = 0$.

We show this by proving that O^θ lies in a strictly lower dimensional subspace compared to T. That is, we can write O^θ in the form $\{\alpha \in \Delta_\tau \colon \widehat{A}\alpha = (b^*; w^*)\}$. Then, we show that there are at least two linearly independent equality constraints. As the dimension of a polytope is the dimension of its nullspace, the dimension of O^θ is no larger than $\tau - 2$. We must assume that θ is not the all zeros vector to avoid an edge case where O^θ has but a single equality constraint, giving it dimension $\tau - 1$. Then, we construct a $\tau - 1$ dimension simplex lying inside T by perturbing $Dg^{(\theta^{ds}, \mathbf{0}_\tau)}$ (which itself is in T, c.f. Theorem 4.1) in τ linearly independent directions and taking the convex hull of those perturbations. See Figure 1 for a visualization.

Note that the role of log loss in Theorem 6.3 is only to ensure that $\alpha \in O^\theta$ implies that $g^{(\theta),\alpha}$ is optimal. I.e. for that α, $Ag^{(\theta),\alpha} = (b^*; w^*)$ by construction. The proof for $dim(T)$ did not depend on our choice of log loss.

We may derive the following result by inspecting the definitions of θ for the OCDS and MV strategies. For OCDS, observe that if the class bias weights are the same constant, then those weights are equivalently 0.

Corollary 6.4. *With the assumptions of Theorem 6.3, the MV strategy is optimal w.p. 0 and the same is true for the OCDS strategy when $(b^*; w^*) \neq \mathbf{1}_{p+k}/k$.*

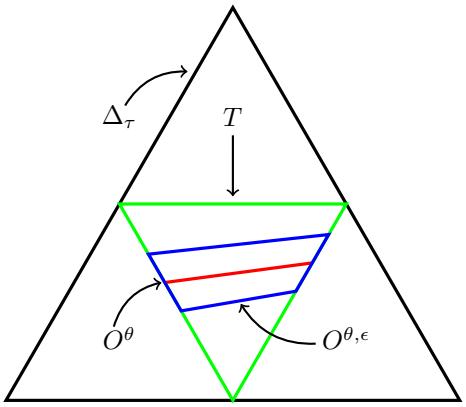


Figure 1: A Visualization of Sets $O^{\theta,\epsilon}$, O^θ, T, and Δ_τ. Observe That O^θ Lies In a Lower Dimension Than T.

Since weighting strategies like MV and OCDS are optimal on a measure zero set, it is natural to ask how often they are close to optimal (e.g. $\epsilon > 0$ close) as a function of the LF accuracies/class frequencies they induce with the resulting labelings. We'll say a prediction $g^{(\theta),\alpha}$ is ϵ-optimal if $\|Ag^{(\theta),\alpha} - (b^*; w^*)\|_\infty \leq \epsilon$. Thus, we may consider the probability $Pr(\theta \text{ is } \epsilon\text{-optimal} | b^*, w^*)$. We have defined this notion of "closeness" to optimality with the sole purpose of estimating the above property algorithmically. A visualization of $O^{\theta,\epsilon}$ can be seen in Figure 1. Once again, see appendix for more details. As we will see in the next section, a prediction g being ϵ-optimal and another g' being ϵ'-optimal with $\epsilon' \leq \epsilon$ does not imply that g' is closer to η than g as such a comparison also depends on their weights.

7 WHEN IS A WEIGHTING STRATEGY BETTER THAN ANOTHER?

In the previous section, we looked at the set of patterns (for fixed b^*, w^*) under which a weighting strategy is optimal and analyzed its size. Because we specified a set, we may characterize when θ' is optimal and θ is not. I.e. we can look at $O^{\theta'} \setminus O^\theta$. As mentioned above, we cannot do the same when we have estimates of b^*, w^*, α^*.

Instead, we will focus on analyzing

$$D_{KL}\left(g^* || g\right) - D_{KL}\left(g^* || g'\right) > 0 \tag{14}$$

to determine when g' is better than g. An and Dasgupta [2024] analyze this expression for $g = g^{(\theta^{ds}),\alpha^*}$ and $g' = g^{(\theta^{bf}),\alpha^*}$ by upper bounding $D_{KL}(g^* || g^{(\theta^{bf}),\alpha^*})$. The idea was to attain an upper bound that does not depend on θ^{bf} as it is hard to control *a priori*. However, their upper bound can be arbitrarily larger than the quantity bounded. This happens because they linearly approximated the entropy function, which can be poor. E.g. approximating $-x^2$ at $x = -2$ via $4x + 4$ yields a poor approximation when $x > -1$.

Theorem 7.1. *Let $Ag^{(\theta^{bf}),\alpha^*} = (b^{bf}; w^{bf})$ be the LF accuracies and class frequencies induced by way of using the BF prediction $g^{(\theta^{bf}),\alpha^*}$ as a putative labeling.*

$$\theta^{*\top}(b^* - b^{bf}; w^* - w^{bf}) \geq D_{KL}\left(g^* || g^{(\theta^{bf}),\alpha^*}\right) + D_{KL}\left(g^{(\theta^{bf}),\alpha^*} || g^*\right)$$

This is shown by analyzing the form of the predictions (Equation 5) and then using the complementary slackness conditions that θ^{bf} satisfies. This immediately implies the upper bound provided by An and Dasgupta [2024], i.e. the LHS bounds the excess loss of $g^{(\theta^{bf}),\alpha^*}$, while having a much simpler proof. We conjecture that bounds derived using sensitivity analysis of convex programs, as found in [Dudík et al., 2004, Mazuelas et al., 2023, An and Dasgupta, 2024], will have similar lower bounds.

Now, we analyze the difference of KL divergences in Equation 14. We will use the superscripts to associate quantities with labelings. E.g. $g' = g^{(\theta'),\alpha'}$, $Ag' = (b'; w')$, $Dg' = \alpha'$. Recall $\widehat{Z}_t$ is the normalizing factor for $\widehat{g}_{t\ell}$. By using the form of distributions in $\mathcal{G}$, Equation 14 becomes the following.

Theorem 7.2. *Let $g, g', g^* \in \mathcal{G}$. Then,*

$$D_{KL}\left(g^* || g\right) - D_{KL}\left(g^* || g'\right) = D_{KL}\left(g' || g\right) + (\theta' - \theta)^\top (b^* - b'; w^* - w') + (\alpha^* - \alpha')^\top \log\left(\frac{\alpha' \widehat{Z}}{\alpha \widehat{Z}'}\right).$$

Observe that on the RHS, there are no references to θ^* or $\widehat{Z}^*$, terms that require knowledge of the optimal weights. All other quantities are known or can be estimated. As such, the remaining results are direct manipulations of the above.

7.1 CASE: PATTERN DISTRIBUTION IS KNOWN

We'll start our analysis in the case where X or equivalently α^* is known. For g' to be better than g, it suffices for

$$(\theta - \theta')^\top (b^* - b'; w^* - w') \leq D_{KL}\left(g' || g\right). \tag{15}$$

I.e. g' is better than g if both $\theta_j \geq \theta'_j$, $b^*_j \leq b'_j$ and vice versa more often than not. In the appendix, we show that the LHS is often negative. This means that if one applies Cauchy-Schwarz as follows, then the test will be very pessimistic. To avoid this situation, we conjecture that one needs lower bounds of $b^* - b'$, $w^* - w'$ in addition to upper bounds.

Lemma 7.3. *g' is better than g (with respect to g^*) if*

$$\|(b^* - b'; w^* - w')\|_2 \leq \frac{D_{KL}\left(g' || g\right)}{\|\theta - \theta'\|_2}.$$

On the other hand, in special cases like the BF weighting strategy, the sign of the weight θ^{bf}_j tells use whether

$b_j^* \geq b_j^{bf} = [Ag^{(\theta^{bf}),\alpha^*}]_j$. I.e. by inspecting the complementary slackness conditions, we can specify when the LHS of Equation 15 is negative.

Lemma 7.4. *Let $\theta^{bf} \in (\mathbb{R} \setminus \{0\})^{p+k}$ be the BF weights and fix $\theta \in \mathbb{R}^{p+k}$. Suppose for $j \in [p+k-1]$ that either $\theta_j^{bf} > 0$ and $\theta_j \leq \theta_j^{bf}$ holds or $\theta_j^{bf} < 0$ and $\theta_j \geq \theta_j^{bf}$ holds. Then, $g^{(\theta^{bf}),\alpha^*}$ is closer to g^* than $g^{(\theta),\alpha^*}$.*

We disallow a BF weight being 0 as then the complementary slackness conditions don't tell us anything. We also do not consider θ^{p+k} because one can always set that to 0 without changing the resulting prediction. I.e. by subtracting θ_{p+k} from $\theta_{p+\ell}$, $\ell \in [k-1]$. In particular, the above result holds for the OCDS weighting strategy and the MV weighting strategy. This gives an *a priori* characterization of a set of problems wherein BF is better than OCDS/MV.

7.2 CASE: PATTERN DISTRIBUTION VARIES

Suppose now that the pattern distribution is not known. That is, while the true pattern distribution is α^*, we receive a pattern distribution of α to determine our conditional label predictions. I.e. α will be used to determine $\widehat{g}, \widehat{g}'$ (with associated weights θ, θ'). α can be thought of as the training pattern distribution. To determine $\widehat{g}, \widehat{g}'$, one approximates $g^{(\theta^*),\alpha}$, but later on (e.g. at test time), we judge how close $g^{(\theta),\alpha^*}, g^{(\theta'),\alpha^*}$ are to $g^{(\theta^*),\alpha^*}$. In other words, a covariate shift occurs. We wish to characterize when $g^{(\theta),\alpha}$ is closer to $g^{(\theta^*),\alpha}$ while $g^{(\theta'),\alpha^*}$ is closer to $g^{(\theta^*),\alpha^*}$ as a function of $\zeta := \alpha^* - \alpha$, the size of the perturbation from α^* to α.

It turns out that the set of ζ satisfying the above can be written as a convex polytope.

Theorem 7.5. *Suppose that*

$$D_{KL}\left(g^{(\theta^*),\alpha}||g^{(\theta),\alpha}\right) \leq D_{KL}\left(g^{(\theta^*),\alpha}||g^{(\theta'),\alpha}\right)$$

and let ζ be a perturbation of α such that $\alpha^* = \alpha + \zeta$. If

$$\zeta \in \Bigg\{\zeta' \in \mathbb{R}^\tau : \alpha + \zeta' \in \Delta_\tau, \sum_{t=1}^{\tau}(\alpha_t + \zeta_t') \times \left[D_{KL}\left(\widehat{g}_t^*||\widehat{g}_t^{(\theta')}\right) - D_{KL}\left(\widehat{g}_t^*||\widehat{g}_t^{(\theta)}\right)\right] \leq 0\Bigg\},$$

then

$$D_{KL}\left(g^{(\theta^*),\alpha*}||g^{(\theta),\alpha*}\right) \geq D_{KL}\left(g^{(\theta^*),\alpha^*}||g^{(\theta'),\alpha^*}\right).$$

The size of this set can be determined algorithmically using LattE [Baldoni et al., 2013] as this is a system of linear inequalities with ζ as our variables.

While this characterization is exact, it only holds for distributions in $\mathcal{G}$. We note that extant methods for covariate shift work on general classifiers. One line of work involves re-weighting examples in the log-likelihood to get a better (max likelihood) estimator under a different covariate distribution [Shimodaira, 2000]. Another line of work uses prediction confidences to estimate performance under covariate shift [Białek et al., 2025].

8 EXPERIMENTS

8.1 COMPUTING THE SIZE OF $O^{\theta,\epsilon}$ AND T^ϵ

Here, we describe how one can determine proportion of pattern distributions wherein a weighting strategy is ϵ-optimal. In particular, we use LattE [Baldoni et al., 2013] to determine the exact volumes of $O^{\theta,\epsilon}, T^\epsilon$ and to count the exact number of pattern distributions in $\tilde{O}_n^{\theta,\epsilon}, \tilde{T}_n^\epsilon$. We perform these computations in some toy settings as these problems are computationally difficult. It is known that exactly counting the number of lattice points in a polytope (e.g. $\tilde{O}_n^{\theta,\epsilon}$) is $\#P$ hard (Table 7.0.1 of [Tóth et al., 2017]). Similarly, computing the exact volume of polytopes is also $\#P$ hard (e.g. [Dyer and Frieze, 1988], [Valiant, 1979]). We do note that there exist parallel algorithms for exact volume computation, e.g. [Avis and Jordan, 2018] and algorithms to estimate the volume, e.g. [Emiris and Fisikopoulos, 2018].

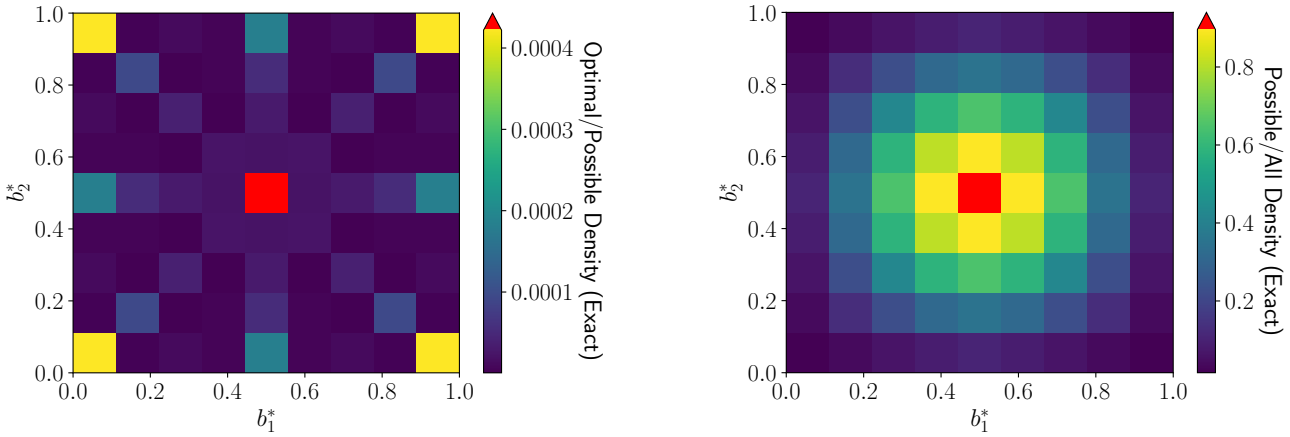
One may find our code and results in the supplementary materials or at `https://github.com/stevenan5/wmv-log-loss-optimality-uai-2026`. Technical details regarding the use of LattE, constructing T, among other considerations are deferred to the appendix.

We evaluate the OCDS strategy on $p = 2, k = 2$ with $w^* = (0.5, 0.5)$ and $b_1^*, b_2^* \in \{i/10\}_{i=1}^9$ for $\epsilon = 0$ and $n = 400$. This n was chosen to meet the various implementation requirements of LattE. See Figure 2. Observe that the probability OCDS is on the order of optimal is ≤ 0.0004 yet $Pr(b^*, w^*)$ can be above 0.8. Also, note the degenerate point where $b_1^* = b_2^* = 0.5$, where all patterns are optimal, colored in red. Here, the OCDS weights is $\mathbf{0}_{p+k}$, a case we specifically avoid in Theorem 6.3. From this, we believe that same theorem can be extended to $p \geq 2, k \geq 2$ while dropping the upper bound requirement on b^*.

We also evaluate the OCDS and MV strategies on $p = 3, k = 2$ with $w^*, n = 1000$ in two cases for $\epsilon = 0.1$. The former is the so-called hammer spammer setup where one LF is highly accurate ($b_1^* = 0.9$) while the others are noisy, but biased toward the truth ($b_2^*, b_3^* = 0.6$). Another is where all LF accuracies are 0.6, i.e. the case where the ensemble is not very accurate. For the first case, the probability OCDS and MV were 0.1-optimal was $0.177\ldots$ and < 0.000581 respectively. For the second case, those probabilities were $0.872\ldots$ versus ≤ 0.198. Perhaps unsurprisingly, we see that OCDS is 0.1-optimal more often.

Table 1: Empirical Evaluation of Tests for Determining When A Weighting Strategy Is Better Than Another. Shown Is the First ϵ for Which A False Negative was Observed.

Dataset	IMDB	Youtube	SMS	CDR	Yelp	Commercial	Tennis	TREC	SemEval	ChemProt	AGNews
Ours (Lem 7.3) BF vs OCDS	0.01	**0.05**	0.01	**0.05**	**0.05**	0.01	**0.2**	0.01	**0.05**	**>0.3**	0.01
Ours (Lem 7.3) BF vs MV	0.01	**0.05**	0.01	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**	**0.1**	**0.1**
[An and Dasgupta, 2024] (Lem 12)	0.169	0.176	0.000123	0.12	0.042	0.114	0.125	0.0112	0.0021	0.064	0.132



(a) $Pr(g^{(\theta^{ds}),\alpha}$ is optimal$|b^*, w^* = (0.5, 0.5))$

(b) $Pr(b^*, w^* = (0.5, 0.5))$

Figure 2: Heatmaps Displaying the Exact Volume of O^{ds}, the Set of Pattern Distributions Wherein the OCDS Weighting Strategy is Optimal, and the Exact Volume of T, the Set of Pattern Distributions Consistent with b^*, w^*. Where $p = k = 2$, $n = 400$ (chosen for technical reasons) and $b_1^*, b_2^* \in \{0.1, 0.2, \ldots, 0.9\}$.

8.2 EVALUATION OF LEMMA 7.3 ON REAL DATASETS

Now, we evaluate Lemma 7.3 on 11 real datasets from WRENCH [Zhang et al., 2021]. In particular, we compare three weighting strategies, BF, OCDS, and MV. The complete details can be found in the appendix.

For BF, we compute weights θ^{bf} by solving the minimax game in Equation 11 with $P^{\epsilon}_{b^* w^* \alpha^*}$ where ϵ is varied between $[0.01, 0.3]$. This is to circumvent problems one encounters when sampling labels for LFs that abstain. For OCDS, we choose the weights in Equation 18 where b^*, w^* are substituted in. For MV, the weights are $\theta = \mathbf{1}_{p+k}$.

In Table 1, we show the first ϵ such that Lemma 7.3 provides a false positive. That is, where the BF prediction is better than OCDS/MV, yet we cannot conclude that using said lemma. If such ϵ is larger than 0.01, we **bold** it. We see that despite using Cauchy-Schwarz, Lemma 7.3 is still usable in practice. We also show the maximum ϵ for which the test provided by [An and Dasgupta, 2024] provides a true positive in determining when BF is better than OCDS. Despite the lower bound of Theorem 7.1 (which affects this test), we see that the ϵ is decently large. However, this test is impractical to use in practice since it requires $\|\theta^*\|_1$.

9 CONCLUSION AND LIMITATIONS

We have shown that even in favorable scenarios, i.e. when the underlying LF accuracies and class distribution are known, an entire class of weighting strategy is optimal on a measure zero set of problems. There are two directions in which our work can be expanded on. First is to generalize Theorem 6.3 to the case of $k = 2$ classes. Another direction is to generalize the representation of the LFs from a single parameter (its accuracy) to more informative measure, e.g. its confusion matrix or even instance specific accuracies. The former may be in reach with our techniques.

By measuring distance via log-loss, checking for g's optimality is reduced to checking whether $Ag = A\eta$ and $Dg = D\eta$. While theoretically convenient, this does not reflect the scenarios where one uses weighted majority vote in non-machine learning contexts. E.g. voting in political elections, voting on a jury, etc. There, 0-1 loss may be more appropriate, though more theoretical machinery would be needed to generalize our results to that scenario.

Acknowledgements

The authors thank the National Science Foundation for their support under grant IIS-2211386.

References

Sina Aeeneh, Nikola Zlatanov, and Jiangshan Yu. New Bounds on the Accuracy of Majority Voting for Multiclass Classification. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):6014–6028, 2025.

Steven An. Statistical Analysis of an Adversarial Bayesian Weak Supervision Method. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 127238–127282. Curran Associates, Inc., 2025.

Steven An and Sanjoy Dasgupta. Convergence Behavior of an Adversarial Weak Supervision Method. In Negar Kiyavash and Joris M. Mooij, editors, *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 1–49. PMLR, 15–19 Jul 2024.

Michael Artin. *Algebra*. Pearson Modern Classics for Advanced Mathematics Series. Pearson, 2018. ISBN 9780134689609.

David Avis and Charles Jordan. mplrs: A Scalable Parallel Vertex/Facet Enumeration Code. *Mathematical Programming Computation*, 10(2):267–302, Jun 2018. ISSN 1867-2957.

Velleda Baldoni, Nicole Berline, Jesús A. De Loera, Brandon E. Dutra, Matthias Köppe, Stanislav Moreinis, Gregory Pinto, Michele Vergne, and Jianqiu Wu. A User's Guide for `LattE integrale` v1.7.2. `LattE` is available at `http://www.math.ucdavis.edu/~latte/`, 2013.

Akshay Balsubramani and Yoav Freund. PAC-Bayes with Minimax for Confidence-Rated Transduction. *CoRR*, abs/1501.03838, 2015a.

Akshay Balsubramani and Yoav Freund. Scalable Semi-Supervised Aggregation of Classifiers. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015b.

Akshay Balsubramani and Yoav Freund. Optimally Combining Classifiers Using Unlabeled Data. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 211–225, Paris, France, 03–06 Jul 2015c. PMLR.

Akshay Balsubramani and Yoav Freund. Optimal Binary Classifier Aggregation for General Losses. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Daniel Berend and Aryeh Kontorovich. A Finite Sample Analysis of the Naive Bayes Classifier. *Journal of Machine Learning Research*, 16(44):1519–1545, 2015.

Jakub Białek, Juhani Kivimäki, Wojciech Kuberski, and Nikolaos Perrakis. Estimating Model Performance Under Covariate Shift Without Labels. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 161084–161115. Curran Associates, Inc., 2025.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. ISBN 9781107394001.

Gary Chartrand and Linda Lesniak. *Graphs & Digraphs*. Chapman & Hall/CRC, Boca Raton, 4th ed. edition, 2005. ISBN 1584883901.

Marie Jean Antoine Nicolas de Caritat, Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. 1785.

John P. D'Angelo and Douglas B. West. *Mathematical Thinking: Problem-Solving and Proofs*. Featured Titles for Transition to Advanced Mathematics. Prentice Hall, 2000. ISBN 9780130144126.

A. Philip Dawid and Allan M. Skene. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.

Franz Dietrich and Kai Spiekermann. Jury Theorems. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2023 edition, 2023.

Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. Performance Guarantees for Regularized Maximum Entropy Density Estimation. In John Shawe-Taylor and Yoram Singer, editors, *Learning Theory*, pages 472–486, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

Martin E. Dyer and Alan M. Frieze. On the Complexity of Computing the Volume of a Polyhedron. *SIAM Journal on Computing*, 17(5):967–974, 1988.

Ioannis Z. Emiris and Vissarion Fisikopoulos. Practical Polytope Volume Approximation. *ACM Trans. Math. Softw.*, 44(4), June 2018. ISSN 0098-3500.

Daniel Y. Fu, Mayee F. Chen, Frederic Sala, Sarah M. Hooper, Kayvon Fatahalian, and Christopher Ré. Fast and Three-rious: Speeding up Weak Supervision with Triplet Methods. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Komei Fukuda. cddlib Reference Manual. https://people.inf.ethz.ch/fukudak/cdd_home/cddlibman2021.pdf, 2021.

Chao Gao and Dengyong Zhou. Minimax Optimal Convergence Rates for Estimating Ground Truth from Crowdsourced Labels. *arXiv preprint arXiv:1310.5764*, 2013.

Peter D. Grünwald and A. Philip Dawid. Game Theory, Maximum Entropy, Minimum Discrepancy and Robust Bayesian Decision Theory. *The Annals of Statistics*, 32 (4):1367–1433, 2004. ISSN 00905364.

David Karger, Sewoong Oh, and Devavrat Shah. Iterative Learning for Reliable Crowdsourcing Systems. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Hyun-Chul Kim and Zoubin Ghahramani. Bayesian Classifier Combination. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 619–627, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

Matthias Köppe. A Primal Barvinok Algorithm Based on Irrational Decompositions. *SIAM Journal on Discrete Mathematics*, 21(1):220–236, 2007. doi: 10.1137/060664768.

Hongwei Li and Bin Yu. Error Rate Bounds and Iterative Weighted Majority Voting for Crowdsourcing, 2014. URL https://arxiv.org/abs/1411.4086.

Yuan Li, Benjamin I. P. Rubinstein, and Trevor Cohn. Exploiting Worker Correlation for Label Aggregation in Crowdsourcing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3886–3895. PMLR, 09–15 Jun 2019.

Nick Littlestone and Manfred K. Warmuth. The Weighted Majority Algorithm. *Information and Computation*, 108 (2):212–261, 1994. ISSN 0890-5401.

Santiago Mazuelas, Yuan Shen, and Aritz Pérez. Generalized Maximum Entropy for Supervised Classification. *IEEE Transactions on Information Theory*, 68(4):2530–2550, 2022.

Santiago Mazuelas, Mauricio Romero, and Peter Grunwald. Minimax Risk Classifiers with 0-1 Loss. *Journal of Machine Learning Research*, 24(208):1–48, 2023.

Alessio Mazzetto, Cyrus Cousins, Dylan Sam, Stephen H Bach, and Eli Upfal. Adversarial Multi Class Learning Under Weak Supervision with Performance Guarantees. In *International Conference on Machine Learning*, pages 7534–7543. PMLR, 2021a.

Alessio Mazzetto, Dylan Sam, Andrew Park, Eli Upfal, and Stephen Bach. Semi-Supervised Aggregation of Dependent Weak Supervision Sources With Performance Guarantees. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3196–3204. PMLR, 13–15 Apr 2021b.

Anand Narasimhamurthy. Theoretical Bounds of Majority Voting Performance for a Binary Classification Problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1988–1995, 2005.

Shmuel Nitzan and Jacob Paroush. Optimal Decision Rules in Uncertain Dichotomous Choice Situations. *International Economic Review*, 23(2):289–297, 1982.

Alexander J. Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data Programming: Creating Large Training Sets, Quickly. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 3567–3575. Curran Associates, Inc., 2016.

Salva Rühling Cachay, Benedikt Boecking, and Artur Dubrawski. End-to-End Weak Supervision. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1845–1857. Curran Associates, Inc., 2021.

Dymitr Ruta and Bogdan Gabrys. Classifier Selection for Majority Voting. *Information Fusion*, 6(1):63–81, 2005. Diversity in Multiple Classifier Systems.

Hidetoshi Shimodaira. Improving Predictive Inference Under Covariate Shift by Weighting the Log-Likelihood Function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. ISSN 0378-3758.

Dapeng Tao, Jun Cheng, Zhengtao Yu, Kun Yue, and Lizhen Wang. Domain-Weighted Majority Voting for Crowdsourcing. *IEEE Transactions on Neural Networks and Learning Systems*, 30(1):163–174, 2019.

Tian Tian and Jun Zhu. Max-Margin Majority Voting for Learning from Crowds. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Csaba D. Tóth, Joseph O'Rourke, and Jacob E. Goodman. *Handbook of Discrete and Computational Geometry*. Chapman and Hall/CRC, Boca Raton, FL, 3rd edition, 2017. ISBN 9781498711395.

Leslie G. Valiant. The Complexity of Computing the Permanent. *Theoretical Computer Science*, 8(2):189–201, 1979. ISSN 0304-3975.

Renzhi Wu, Shen-En Chen, Jieyu Zhang, and Xu Chu. Learning Hyper Label Model for Programmatic Weak Supervision. In *The Eleventh International Conference on Learning Representations*, 2023.

Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. WRENCH: A Comprehensive Benchmark for Weak Supervision. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Jieyu Zhang, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. A Survey on Programmatic Weak Supervision, 2022.

Jieyu Zhang, Linxin Song, and Alex Ratner. Leveraging Instance Features for Label Aggregation in Programmatic Weak Supervision. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 157–171. PMLR, 25–27 Apr 2023.

Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I. Jordan. Spectral Methods Meet EM: A Provably Optimal Algorithm for Crowdsourcing. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 1260–1268, Cambridge, MA, USA, 2014. MIT Press.

When and How Often is Weighted Majority Vote Optimal Under Log Loss? (Appendix)

Steven An[1] **Sanjoy Dasgupta**[1]

[1]Computer Science Dept., University of California San Diego, La Jolla, California, USA

A APPENDIX OVERVIEW

We provided all missing proofs as well as extra experimental results in this appendix. See supplementary material or `https://github.com/stevenan5/wmv-log-loss-optimality-uai-2026` for the code that was run. The appendix is organized as follows.

1. Section B, a recap of the problem setup and notation.
2. Section C, derivation of OCDS weighting strategy from probabilistic assumptions.
3. Section D, proof of Theorem 4.1, the derivation of OCDS weighting strategy via adversarial assumptions.
4. Section E, derivation of the convex program used for the BF weighting strategy.
5. Section F, proof of Lemma 5.1, a Pythagorean type result for KL divergence.
6. Section G, proof of Theorem 7.1, which when used with the arguments of An and Dasgupta [2024] (Theorem 6), provides a new, simpler proof of Theorem 6.2.
7. Section H, proof of Theorem 7.2, Lemma 7.3, and Lemma 7.4, characterizations of when a weighting strategy is better than another when neither is optimal.
8. Section I, proof of Theorem 7.5, our covariate shift analysis.
9. Section J, proof of Lemma 6.1 and Theorem 6.3, regarding the measure zero set of optimal patterns for certain weighting strategies.
10. Section K, the complete experimental results, including our evaluations of Lemma 7.3.

B PROBLEM AND NOTATION RECAP

To start, we briefly recap the problem at hand and review the notation used. Suppose we have a dataset $X = \{x_1, \ldots, x_n\} \subset \mathcal{X}$ where each point has one of k possible labels $\mathcal{Y} = \{1, \ldots, k\} = [k]$. At our disposal is p labeling functions (LFs) $h^{(j)} \colon \mathcal{X} \to \mathcal{Y}$. We will refer to the LF predictions on a datapoint by referring to one of $\tau \coloneqq k^p$ patterns. A pattern is a tuple of k predictions. With respect to patterns, π_{tj} denotes the prediction $h^{(j)}$ on pattern t while π_t^ℓ denotes the subset of $[p]$ for which the LFs predict ℓ on pattern t. That is,

$$\pi_t^\ell = \{j \in [p] \colon \pi_{tj} = \ell\}.$$

We wish to estimate $\widehat{\eta}_{t\ell} = Pr(y = \ell \mid \pi_t)$, the true conditional label probability, via a weighted majority vote. That is, for $p + k$ weights $\theta \in \mathbb{R}^{p+k}$, our estimate for $\widehat{\eta}_{t\ell}$ is $\widehat{g}^{(\theta)}$. See Equation 4 for the functional form. We write $\widehat{\eta} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_\tau)$ as the concatenation of τ distributions in Δ_k to represent all the conditional probabilities. Another quantity of interest is the joint label probability $\eta_{t\ell} = Pr(y = \ell, \pi_t)$. We write all these joint probabilities as $\eta \in \Delta_{\tau k}$. With τ extra weights $\nu \in \mathbb{R}^\tau$, we estimate that quantity with $g^{(\theta,\nu)}$. See Equation 5 for its functional form. In the special case where we know the pattern

distribution α^* and want to force our prediction to adhere to that pattern distribution, we pick $\nu_t = \log(\alpha_t^*/\widehat{Z}_t)$. Recall that $\widehat{Z}_t$ is the normalization factor in Equation 4.

Quantities of interest for us are $b^* \in [0,1]^p$, the accuracy of the p LFs, $w^* \in \Delta_k$, the class frequency distribution, and $\alpha^* \in \Delta_\tau$, the pattern frequency distribution. Estimates of these quantities will be denoted by a lack of $*$ in the superscript. Recall that we can write b^*, w^*, α^* as a matrix multiplication of η. That is, $A\eta = (b^*; w^*)$ and $D\eta = \alpha^*$ where A was defined in Equation 6 and D was defined in Equation 7.

C DERIVING THE OCDS WEIGHTING STRATEGY FROM PROBABILISTIC ASSUMPTIONS

We rehash the findings of Li and Yu [2014] wherein One-Coin Dawid-Skene's (OCDS) posterior label distribution is shown to be a weighted majority vote. Simple results about that posterior label distribution, i.e. prediction, are shown in service of the next section.

We start by laying out the generative process under OCDS. Fix a dataset X and say each of the datapoints has a label $y_i \in \mathcal{Y}$. Define $w^* \in \Delta_k$ to be the underlying class frequency distribution. The OCDS generative process generates label y_i independently by way of $y_i \sim Categorical(w^*)$. LFs are modeled as biased coins so that predictions on different datapoints are modeled as independent coin flips. Moreover, the LF predictions are assumed to be independent to each other on a single datapoint given the label.

Let $b_j^* \in [0,1]$ be the underlying accuracy of LF $h^{(j)}$. Define the distribution $v_{y_i} \in \Delta_k$ where the element in position y_i is b_j^* while every other element is $(1-b_j^*)/(k-1)$. This is the conditional distribution of $h^{(j)}$'s prediction given that the true label is y_i. I.e. $Pr(h^{(j)}(x) = \ell' \mid y = \ell)$ is given by this distribution for $\ell, \ell' \in [k]$. The prediction of $h^{(j)}(x_i) \sim Categorical(v_{y_i})$.

In practice, we are given X (more specifically the LF predictions on X) and wish to infer what its labels are. Specifically, we are interested in the posterior label distribution

$$Pr(y \mid h^{(j)}(x), \forall j \in [p]).$$

By using Bayes' Theorem and the conditional independence of LF predictions, one sees that

$$Pr(y \mid h^{(j)}(x), \forall j \in [p]) = \frac{Pr(y)\prod_{j=1}^p Pr(h^{(j)}(x) \mid y)}{Pr(h^{(j)}(x), \forall j \in [p])}. \tag{16}$$

By definition, the denominator is a pattern distribution and the numerator can be rewritten as follows:

$$Pr(y = \ell, h^{(j)}(x), \forall j \in [p]) = w_\ell^* \prod_{j=1}^p (b_j^*)^{\mathbf{1}(h^{(j)}(x)=\ell)} \prod_{j=1}^p \left(\frac{1-b_j^*}{k-1}\right)^{\mathbf{1}(h^{(j)}(x)\neq\ell)}$$

where $\mathbf{1}(\cdot)$ is the indicator function. Lets suppose that the LF predictions match pattern π_t. We may write the numerator as

$$Pr(y = \ell, \pi_t) = w_\ell^* \prod_{j \in \pi_t^\ell} b_j^* \prod_{j \in [p]\setminus\pi_t^\ell} \frac{1-b_j^*}{k-1}. \tag{17}$$

It's clear the denominator of Equation 16 is $Pr(\pi_t) = \sum_{\ell=1}^k Pr(y = \ell, \pi_t)$.

Now, we wish to show the above can be written as an element of $\mathcal{G}$. Define the OCDS weights as $\theta^{ds} \in \mathbb{R}^{p+k}$ as

$$\theta_j = \log\left(\frac{b_j^*(k-1)}{1-b_j^*}\right) \qquad \text{for } j \in [p] \qquad \text{and} \qquad \theta_{p+\ell} = \log(w_\ell^*) \qquad \text{for } \ell \in [k]. \tag{18}$$

Proposition C.1. *For OCDS weights defined as above (Equation 18), $g^{(\theta^{ds}, \mathbf{0}_\tau)}$ matches the joint probability in Equation 17, which is the OCDS estimate for the joint distribution element $\eta_{t\ell}$. Similarly, $\widehat{g}^{(\theta^{ds})}$ matches the quantity of Equation 16.*

Proof. We start with the first claim. Using the form of prediction for g, i.e. Equation 5, and substituting the weights, we have

$$g_{t\ell}^{(\theta^{ds},\mathbf{0}_\tau)} = \frac{\exp\left(\log(w_\ell^*) + \sum_{j\in\pi_t^\ell} \log\left(\frac{b_j^*(k-1)}{1-b_j^*}\right)\right)}{\sum_{t'=1}^{\tau}\sum_{\ell'=1}^{k} \exp\left(\log(w_{\ell'}^*) + \sum_{j'\in\pi_{t'}^{\ell'}} \log\left(\frac{b_{j'}^*(k-1)}{1-b_{j'}^*}\right)\right)}.$$

Converting the sums into products gives us

$$g_{t\ell}^{(\theta^{ds},\mathbf{0}_\tau)} = \frac{w_\ell^* \prod_{j\in\pi_t^\ell} \frac{b_j^*(k-1)}{1-b_j^*}}{\sum_{t'=1}^{\tau}\sum_{\ell'=1}^{k} w_{\ell'}^* \prod_{j'\in\pi_{t'}^{\ell'}} \frac{b_{j'}^*(k-1)}{1-b_{j'}^*}}.$$

Multiply the numerator and denominator by the same quantity, namely $\prod_{j=1}^{p}[(b_j^*(k-1))/(1-b_j^*)]$. This gives us

$$g_{t\ell}^{(\theta^{ds},\mathbf{0}_\tau)} = \frac{w_\ell^* \prod_{j\in\pi_t^\ell} b_j^* \prod_{j\in[p]\setminus\pi_t^\ell} \frac{1-b_j^*}{k-1}}{\sum_{t'=1}^{\tau}\sum_{\ell'=1}^{k} w_{\ell'}^* \prod_{j'\in\pi_{t'}^{\ell'}} b_{j'}^* \prod_{j'\in[p]\setminus\pi_{t'}^{\ell'}} \frac{1-b_{j'}^*}{k-1}} = \frac{Pr(y=\ell,\pi_t)}{\sum_{t'=1}^{\tau}\sum_{\ell'=1}^{k} Pr(y=\ell,\pi_t)} = Pr(y=\ell,\pi_t).$$

because the denominator sums to 1. To be explicit,

$$g_{t\ell}^{(\theta^{ds},\mathbf{0}_\tau)} = w_\ell^* \prod_{j\in\pi_t^\ell} b_j^* \prod_{j'\in[p]\setminus\pi_t^\ell} \frac{1-b_{j'}^*}{k-1}. \tag{19}$$

□

Now, we show that $Ag^{(\theta^{ds},\mathbf{0}_\tau)} = (b^*; w^*)$ so that $\alpha^{ds} \coloneqq Dg^{(\theta^{ds},\mathbf{0}_\tau)} \in T$. Recall that we have previously defined

$$T \coloneqq \{\alpha \in \Delta_\tau \colon \exists z \in \Delta_{\tau k}, Az = (b^*; w^*), Dz = \alpha\}. \tag{20}$$

To be explicit,

$$\alpha_t^{ds} = \sum_{\ell=1}^{k} w_\ell^* \prod_{j\in\pi_t^\ell} b_j^* \prod_{j'\in[p]\setminus\pi_t^\ell} \frac{1-b_{j'}^*}{k-1}. \tag{21}$$

Proposition C.2. *$Ag^{(\theta^{ds},\mathbf{0}_\tau)} = (b^*; w^*)$ which implies that $\alpha^{ds} = Dg^{(\theta^{ds},\mathbf{0}_\tau)} \in T$. Moreover, if $b^* \in (0,1)^p$ and $w^* \in (0,1)^k \cap \Delta_k$, then every element of $\alpha^{ds} > 0$.*

Proof. This is most easily seen by using the fact that

$$g_{t\ell}^{(\theta^{ds},\mathbf{0}_\tau)} = Pr(y=\ell,\pi_t)$$

with respect to the probability distribution defined using the OCDS generative assumption. Then, take element j of $Ag^{(\theta^{ds},\mathbf{0}_\tau)}$. By the definition of A (Equation 6), it is

$$\sum_{t=1}^{\tau}\sum_{\ell=1}^{k} \mathbf{1}(\pi_{tj}=\ell) Pr(y=\ell,\pi_t) = \sum_{t=1}^{\tau}\sum_{\ell=1}^{k} Pr(y=\ell, h^{(j)}(x)=y, h^{(j')}(x)=\pi_{tj'}, \forall j' \in [p]\setminus\{j\}).$$

Clearly, the class variable y and the other LF predictions are being marginalized out, meaning the above is equal to $Pr(h^{(j)}(x)=y)$. A similar argument works for elements $p+\ell$ of $Ag^{(\theta^{ds},\mathbf{0}_\tau)}$ for $\ell \in [k]$. Our second claim follows by definition of T. To see our final claim, observe that with those assumptions, every element of $g^{(\theta^{ds},\mathbf{0}_\tau)}$ is strictly positive. □

D DERIVATION OF OCDS WEIGHTING STRATEGY FROM ADVERSARIAL ASSUMPTIONS

Now, we derive the same weighting strategy without the use of any probabilistic assumptions on how the data is generated. To do this, we'll analyze the maximum entropy problem as stated in Theorem 4.1 and upper bound the entropy of the solution via sensitivity analysis methods. Specifically, we'll show that the maximum entropy over distributions satisfying the constraints cannot exceed the entropy of $g^{(\theta^{ds},\mathbf{0}_\tau)}$, which is our claimed solution. As $g^{(\theta^{ds},\mathbf{0}_\tau)}$ satisfies the constraints as well (see Proposition C.2), $g^{(\theta^{ds},\mathbf{0}_\tau)}$ must be the maximum entropy solution. In particular, the resulting pattern distribution is α^{ds} (Equation 21).

We note that one may directly solve the maximum entropy problem, doing so is messy.

Theorem D.1 (Theorem 4.1)**.** *Fix some accuracies/class frequencies $b^* \in (0,1)^p, w^* \in \Delta_k \cap (0,1)^k$. Recall we have defined (see Equation 9) the set of labelings satisfying b^*, w^* as*

$$P_{b^*w^*} = \{z \in \Delta_{\tau k} \colon \|Az - (b^*; w^*)\|_\infty \leq 0\}.$$

Define the OCDS weights as $\theta^{ds} \in \mathbb{R}^{p+k}$ as in Equation 18. Then,

$$g^{(\theta^{ds}),\alpha^{ds}} = g^{(\theta^{ds},\mathbf{0}_\tau)} = \operatorname*{arg\,max}_{z \in P_{b^*w^*}} -z^\top \log z = \operatorname*{arg\,min}_{g \in \Delta_{\tau k}} \max_{z \in P_{b^*w^*}} -z^\top \log g.$$

Proof. To show the claim, we will analyze the following.

$$\max_{\alpha \in T} \max_{z \in P_{b^*w^*\alpha}} -z^\top \log z. \tag{22}$$

The third equality follows from a similar argument to that of Theorem 1 in [An and Dasgupta, 2024]. Recall that T (Equation 20) is the set of pattern distributions where it's possible for a labeling $g^{(\theta),\alpha}$ can give rise to LF accuracies b^* and class frequencies w^* (for some $\theta \in \mathbb{R}^{p+k}$) by way of $Ag^{(\theta),\alpha}$. We have seen that this set is nonempty in Proposition C.2.

We will provide an upper bound for the inner maximization problem that does not depend on α. Moreover, we'll demonstrate that the upper bound is tight – it's value is achieved when we look at the entropy of $g^{(\theta^{ds},\mathbf{0}_\tau)}$. Formally, we will show

$$\max_{\alpha \in T} \max_{z \in P^0_{b^*w^*\alpha}} -z^\top \log z \leq -g^{(\theta^{ds},\mathbf{0}_\tau)\top} \log g^{(\theta^{ds},\mathbf{0}_\tau)}.$$

The upper bound is derived by applying sensitivity analysis to the following convex program. To simplify the notation, we'll write $b = (b; w)$.

Call the following convex program the reference problem

$$p(0,0) = \min_{\substack{Az = b^{ref} \\ Dz = \alpha^{ref} \\ z \in \Delta_{\tau k}}} z^\top \log z.$$

Perturbing the equality constraints by the respective values of u, v gives

$$p(u,v) = \min_{\substack{Az - b^{ref} = u \\ Dz - \alpha^{ref} = v \\ z \in \Delta_{\tau k}}} z^\top \log z. \tag{23}$$

For fixed b^{ref}, α^{ref}, suitably chosen u, v can allow one to recreate the set $P^0_{b^*w^*\alpha^*}$, the constraint set of the inner maximization problem from above. This is useful because we will bound $p(u,v)$ in terms of $p(0,0)$ along with some other terms.

Say θ^{ref}, ν^{ref} is the optimal solution to $p(0,0)$. Then, by Section 5.6.2 of [Boyd and Vandenberghe, 2004],

$$-p(u,v) \leq -p(0,0) + \theta^{ref\top} u + \nu^{ref\top} v.$$

If we replace $p(\cdot,\cdot)$ with our objective of negative entropy, we get the following.

$$-g^\top \log g \leq -g^{ref\top} \log g^{ref} + \theta^{ref\top} u + \nu^{ref\top} v.$$

To simplify the notation, observe that by definition, the solution g to the BF problem $p(u, v)$ (Equation 23) is such that $Ag = u + b^{ref}$, $Dg = v + \alpha^{ref}$. So, define $b = u + b^{ref}$, $\alpha = v + \alpha^{ref}$. In a roundabout way, we have $u = b - b^{ref}$ and $v = \alpha - \alpha^{ref}$. Then,

$$\theta^{ref\top} u + \nu^{ref\top} v = \theta^{ref\top} \left(b - b^{ref}\right) + \nu^{ref\top} \left(\alpha - \alpha^{ref}\right).$$

The inequality we have shown so far is

$$-g^\top \log g^\top \leq -g^{ref\top} \log g^{ref} + \theta^{ref\top}(b - b^{ref}) + \nu^{ref\top}(\alpha - \alpha^{ref})$$

in full. To finish the proof, choose $g^{ref} = g^{(\theta^{ds}, \mathbf{0}_\tau)}$. By Proposition C.2, $b^{ref} = Ag^{(\theta^{ds}, \mathbf{0}_\tau)} = b^*$. By construction, $b = b^*$ so that $u = 0$, whereupon we see that $\theta^{ref\top}(b - b^{ref}) = 0$. Also, observe that we have chosen $\nu^{ref} = 0$, which when taken with the above implies that

$$-g^\top \log g \leq -g^{(\theta^{ds}, \mathbf{0}_\tau)\top} \log g^{(\theta^{ds}, \mathbf{0}_\tau)}. \tag{24}$$

Now, we put everything together. By definition, $\alpha^{ds} = Dg^{(\theta^{ds}, \mathbf{0}_\tau)}$. Recall that we have defined g as the solution to the problem $p(u, v)$,

$$g = \underset{\substack{Az - b^* = 0 \\ Dz - \alpha^{ds} = \alpha - \alpha^{ds} \\ z \in \Delta_{\tau k}}}{\arg\min} z^\top \log z$$

where we have substituted $u = 0$, $\alpha^{ref} = \alpha^{ds}$, and $v = \alpha - \alpha^{ds}$. It is easy to check that this is the inner maximization problem of Equation 22. To be explicit, rename g from above to g_α. Equation 24 shows a bound for the entropy of g_α without any dependence on α, thus it must hold when we maximize the entropy of g_α by varying α. Formally,

$$\max_{\alpha \in T} -g_\alpha^\top \log g_\alpha \leq -g^{(\theta^{ds}, \mathbf{0}_\tau)\top} \log g^{(\theta^{ds}, \mathbf{0}_\tau)}.$$

To complete the proof, it suffices to note that the RHS is the entropy of an element of $\mathcal{G}$ and that the weights chosen were θ^{ds}. In other words, the optimal weighting strategy is the OCDS strategy. □

While this result is interesting in the transductive scenario (where the dataset features X is known), one can in theory apply this analysis to the supervised setting. That is, suppose one has X_{train}, X_{test} where the training set is labeled. Suppose further that they have p LFs and wish to label X_{test} via a weighted majority vote of those LFs. Also, say that they wish to fix the predicted conditional label distribution (i.e. their prediction for $Pr(y \mid x)$) by way of fixing θ. The prediction on X_{test} will be $g^{(\theta), \alpha^*}$, where α^* is the pattern distribution for X_{test}.

The question is what θ should look like. On one hand, one may fit a weighted majority vote to X_{train} and use those same weights for X_{test}. In essence, perform Empirical Risk Minimization. Alternatively, they could use X_{train} and its accompanying labels to estimate the LF accuracies and the class frequencies and set θ to be the OCDS weights to get a minimax optimal prediction. (If α^* can be estimated, one may also use the BF weighting strategy gotten by solving the convex program presented in the next section.)

E DERIVING THE BF WEIGHTING STRATEGY

In this section, we show the claims made about the BF weights coming from a convex program. Specifically, we consider a convex program that is a relaxation of the one seen in Equation 11 where the constraint for z is $P^\epsilon_{bw} \cap P^\xi_\alpha$ where b, w are estimates of b^*, w^* and α is an estimate of α^*. The conversion from a minimax problem to a maximum entropy problem follows from Theorem 1 of [An and Dasgupta, 2024].

To reduce the notational burden, we write b in place of $(b; w)$.

Theorem E.1. *The BF game is a maximum entropy problem which has the following dual.*

$$\min_{\substack{b-\epsilon \leq Az \leq b+\epsilon \\ \alpha-\xi \leq Dz \leq \alpha+\xi \\ z \in \Delta_{\tau k}}} z^\top \log z = \max_{\substack{\theta \in \mathbb{R}^{p+k} \\ \nu \in \mathbb{R}^\tau}} \left[b^\top \theta - \epsilon^\top |\theta| + \alpha^\top \nu - \xi^\top |\nu| - \log \left(\sum_{t=1}^{\tau} \sum_{\ell=1}^{k} \left[\exp\left([A^\top \theta]_{t\ell} + \nu_t \right) \right] \right) \right].$$

The form of the BF solution g is

$$g_{t\ell} = \frac{\exp\left([A^\top \theta]_{t\ell} + \nu_t \right)}{\sum_{t'=1}^{\tau} \sum_{\ell'=1}^{k} \left[\exp\left([A^\top \theta]_{t'\ell'} + \nu_{t'} \right) \right]}.$$

Proof. We first write out the Lagrangian.

$$L(z,\sigma,\sigma',\mu,\mu',\lambda) = z^\top \log z + \sigma^\top(Az - b - \epsilon) + \sigma'^\top(b - \epsilon - Az) + \mu^\top(Dz - \alpha - \xi) \\ + \mu'^\top(\alpha - \xi - Dz) + \lambda(z^\top \mathbf{1}_\tau - 1).$$

We do not need to include the constraint ensuring that z is non-negative elementwise because the form of the solution will guarantee that. We now derive the dual function $g(\sigma,\sigma',\mu,\mu',\lambda)$ by minimizing L with respect to z. We have that

$$\nabla_z L = \mathbf{1}_{\tau k} + \log z + A^\top(\sigma - \sigma') + D^\top(\mu - \mu') + \lambda \mathbf{1}_{\tau k}.$$

We now solve for z and ensure that its elements sum to 1. For elementwise exponential,

$$z = \exp\left(-\mathbf{1}_{\tau k} + A^\top(\sigma' - \sigma) + D^\top(\mu' - \mu) - \lambda \mathbf{1}_\tau\right).$$

Plugging this back into the Lagrangian gives

$$-z^\top \mathbf{1}_{\tau k} - \sigma^\top(b + \epsilon) + \sigma'^\top(b - \epsilon) - \mu^\top(\alpha + \xi) + \mu'^\top(\alpha - \xi) - \lambda.$$

To finish, we maximize the above expression with respect to λ, which will give us the dual function. We'll index z the same way as η. $z \in [0,1]^{\tau k}$, so we write as

$$z = (z_1, \ldots, z_\tau)$$

where $z_t \in \Delta_k$. The t there indexes the pattern. The ℓ^{th} element of z_t is just $z_{t\ell}$. Now, notice that every element of z has $\exp(-\lambda)$ in the denominator. Thus, the derivative is

$$\exp(-\lambda - 1) \sum_{t=1}^{\tau} \sum_{\ell=1}^{k} \left[\exp\left([A^\top(\sigma' - \sigma)]_{t\ell} + [D^\top(\mu' - \mu)]_{t\ell}\right)\right] - 1 = 0.$$

This implies that

$$\lambda = 1 + \log\left(\sum_{t=1}^{\tau} \sum_{\ell=1}^{k} \left[\exp\left([A^\top(\sigma' - \sigma)]_{t\ell} + [D^\top(\mu' - \mu)]_{t\ell}\right)\right]\right).$$

The dual function can be written as

$$g(\sigma,\sigma',\mu,\mu') = -\sigma^\top(b + \epsilon) + \sigma'^\top(b - \epsilon) - \mu^\top(\alpha + \xi) + \mu'^\top(\alpha - \xi) \\ - \log\left(\sum_{t=1}^{\tau} \sum_{\ell=1}^{k} \left[\exp\left([A^\top(\sigma' - \sigma)]_{t\ell} + [D^\top(\mu' - \mu)]_{t\ell}\right)\right]\right)$$

We can simplify this more based on the definition of D. $D^\top$ is like an identity matrix of τ dimensions, but each element is now replaced by a *column* vector. This means that $[D^\top \mu]_{tj} = \mu_t$ for each $j \in [k]$. So

$$g(\sigma,\sigma',\mu,\mu') = b^\top(\sigma' - \sigma) - \epsilon(\sigma' + \sigma) + \alpha^\top(\mu' - \mu) - \xi^\top(\mu' + \mu) \\ - \log\left(\sum_{t=1}^{\tau} \sum_{\ell=1}^{k} \left[\exp\left([A^\top(\sigma' - \sigma)]_{t\ell} + (\mu' - \mu)_t\right)\right]\right).$$

With this, we can redefine the Lagrange multipliers so we are dealing with an unconstrained optimization problem. That is, let $\theta := \sigma' - \sigma \in \mathbb{R}^{p+k}$, and $\nu = \mu' - \mu \in \mathbb{R}^\tau$. Now, we get the result of the theorem. The dual for BF can be written as

$$\max_{\substack{\theta \in \mathbb{R}^{p+k} \\ \nu \in \mathbb{R}^\tau}} b^\top \theta - \epsilon^\top |\theta| + \alpha^\top \nu - \xi^\top |\nu| - \log\left(\sum_{t=1}^{\tau} \sum_{\ell=1}^{k} \left[\exp\left([A^\top \theta]_{t\ell} + \nu_t\right)\right]\right).$$

where the absolute value is elementwise. □

Now we derive a result relating the above program to the case where we know α. This is incidentally the case of BF that has traditionally been studied.

Corollary E.2.

$$\min_{\substack{b-\epsilon\leq Az\leq b+\epsilon \\ Dz=\alpha \\ z\in\Delta_{\tau k}}} z^\top \log z = \max_{\theta\in\mathbb{R}^{p+k}} \left[b^\top\theta - \epsilon^\top|\theta| - \sum_{t=1}^{\tau} \alpha_t \log\left(\frac{1}{\alpha_t}\sum_{\ell=1}^{k}\exp\left([A^\top\theta]_{t\ell}\right)\right)\right]$$

The form of the BF solution g is

$$g_{t\ell}^{(\theta),\alpha} = \frac{\alpha_t \exp\left([A^\top\theta]_{t\ell}\right)}{\sum_{\ell'=1}^{k}\left[\exp\left([A^\top\theta]_{t\ell'}\right)\right]}. \tag{25}$$

Proof. We start from Theorem E.1 and maximize the ν's. Note that in this case, $\xi = \mathbf{0}_\tau$. We wish to maximize the following expression over ν_t (for an arbitrary t)

$$g(\theta,\nu) = b^\top\theta - \epsilon^\top|\theta| + \alpha^\top\nu - \log\left(\sum_{t=1}^{\tau}\sum_{\ell=1}^{k}\left[\exp\left([A^\top\theta]_{t\ell}+\nu_t\right)\right]\right).$$

Observe that

$$\frac{\partial g(\theta,\nu)}{\partial\nu_t} = \alpha_t - \frac{\sum_{\ell=1}^{k}\exp\left([A^\top\theta]_{t\ell}+\nu_t\right)}{\sum_{t'=1}^{\tau}\sum_{\ell'=1}^{k}\left[\exp\left([A^\top\theta]_{t'\ell'}+\nu_{t'}\right)\right]}.$$

Setting this equal to 0 and simplifying, we get

$$\alpha_t = \frac{\exp(\nu_t)\sum_{\ell=1}^{k}\exp\left([A^\top\theta]_{t\ell}\right)}{\sum_{t'=1}^{\tau}\exp(\nu_{t'})\sum_{\ell'=1}^{k}\left[\exp\left([A^\top\theta]_{t'\ell'}\right)\right]}.$$

Choose

$$\nu_t = \log\left(\frac{\alpha_t}{\sum_{\ell=1}^{k}\exp\left([A^\top\theta]_{t\ell}\right)}\right),$$

which makes the equality hold (as $\alpha\in\Delta_\tau$). Plugging this into $g(\theta,\nu)$, we have

$$g(\theta,\nu) = b^\top\theta - \epsilon^\top|\theta| - \sum_{t=1}^{\tau}\alpha_t\log\left(\frac{1}{\alpha_t}\sum_{\ell=1}^{k}\exp\left([A^\top\theta]_{t\ell}\right)\right) - \log\left(\sum_{t=1}^{\tau}\alpha_t\right).$$

which matches our claimed dual as the last sum evaluates to 1. To get the claimed form of solution, simply plug in our derived value of ν_t. □

F A PYTHAGOREAN THEOREM TYPE RESULT

Here, we prove Lemma 5.1 of the main paper. This proof closely follows Lemma 19 in the appendix of [An and Dasgupta, 2024]. To recover the stated result, choose $g' = g^*$ where g^* satisfies the Lemma assumption by construction.

Lemma F.1. *Let $g, g' \in \mathcal{G}$ be such that $A\eta = Ag', D\eta = Dg'$.*

$$D_{KL}\left(\eta||g\right) = D_{KL}\left(\eta||g'\right) + D_{KL}\left(g'||g\right).$$

Proof. Observe that

$$D_{KL}\left(\eta||g'\right) - D_{KL}\left(\eta||g\right) = \sum_{t=1}^{\tau}\sum_{\ell=1}^{k}\eta_{t\ell}\log\frac{g_{t\ell}}{g'_{t\ell}}.$$

Now, recall the functional form of g, g'. From Equation E.1, we see that

$$g_{t\ell} = \frac{\exp\left([A^\top\theta]_{t\ell}+\nu_t\right)}{\sum_{t'=1}^{\tau}\sum_{\ell'=1}^{k}\left[\exp\left([A^\top\theta]_{t'\ell'}+\nu_{t'}\right)\right]}.$$

Write the normalizing factor as Z for brevity and observe that by taking the logarithm, we get

$$\log g_{t\ell} = [A^\top \theta]_{t\ell} + \nu_t - \log Z.$$

Something similar can be written for g' but with θ', ν', Z'. It follows that

$$D_{KL}\left(\eta||g'\right) - D_{KL}\left(\eta||g\right) = \sum_{t=1}^{\tau}\sum_{\ell=1}^{k} \eta_{t\ell}\left([A^\top\theta]_{t\ell} + \nu_t - \log Z - [A^\top\theta']_{t\ell} - \nu'_t + \log Z'\right).$$

Regrouping and using vector notation when possible,

$$= \eta^\top (A^\top\theta - A^\top\theta' + D^\top\nu - D^\top\nu') + \sum_{t=1}^{\tau}\sum_{\ell=1}^{k} \eta_{t\ell}\left(\log Z' - \log Z\right).$$

Recall that $D^\top$ takes a vector of length τ and expands it to length τk by repeating each element k times. By assumption, the first term equals

$$g'^\top (A^\top\theta - A^\top\theta' + D^\top\nu - D^\top\nu').$$

We now focus on the second term. Notice that the Z terms do not depend on t, ℓ. Therefore, they can be factored out of the double summation. Since the double summation of $\eta_{t\ell}$ evaluates to 1, we can replace the $\eta_{t\ell}$'s with $g'_{t\ell}$. Going backwards shows that our expression equals $D_{KL}\left(g'||g\right)$ as required. □

G NEW PROOF OF AN ERROR BOUND FOR BF

We now provide a proof of Theorem 7.1. This is a very simple proof that can be used in place of the much longer proof for Theorem 7 of [An and Dasgupta, 2024] to prove Theorem 6 of that same paper. Theorem 6 is equivalent to Theorem 6.2.

We first show a result about the form of the sum of the KL divergence between $g, g' \in \mathcal{G}$ along with the KL divergence with arguments reversed. This result, along with leveraging the complementary slackness conditions once gets by using BF will give us Theorem 7.1.

To start, we'll show a helpful equality which will be applied multiple times.

Lemma G.1. *Let $g, g' \in \mathcal{G}$ and say $Ag = (b; w)$, $Dg = \alpha$, etc.*

$$\theta^\top((b'; w') - (b^*; w^*)) + \nu^\top(\alpha' - \alpha^*) = (g' - \eta)^\top \log g$$

Proof. We start from the LHS and show that it can be transformed into the RHS. To start, we write b, α in terms of the predictions. That is, the LHS is

$$\theta^\top A(g' - \eta) + \nu^\top D(g' - \eta) = (g' - \eta)^\top (A^\top\theta + D^\top\nu).$$

The terms in the second parentheses look like the numerator of the softmax in Equation E.1. If we let Z be the normalization factor for the prediction gotten via weights θ, ν, then observe that

$$(g' - \eta)^\top(-\log(Z)\mathbf{1}_{\tau k}) = -\log Z\left(\sum_{t=1}^{\tau}\sum_{\ell=1}^{k} g' - \sum_{t=1}^{\tau}\sum_{\ell=1}^{k}\eta_{t\ell}\right) = 0$$

where $\mathbf{1}_{\tau k}$ is the vector of τk ones because η, g' are both distributions in $\Delta_{\tau k}$. This means that the LHS is equal to

$$(g' - \eta)^\top[\log\left(\exp(A^\top\theta + D^\top\nu)\right) - \log(Z)]$$

where the log and exponential are elementwise. Equivalently, the above is equal to $(g' - \eta)^\top \log g$, which is what we wanted to show. □

Lemma G.2. *Let $g, g^* \in \mathcal{G}$. Then,*

$$(\theta - \theta^*)^\top((b; w) - (b^*; w^*)) + (\nu - \nu^*)^\top(\alpha - \alpha^*) = D_{KL}\left(g^*||g\right) + D_{KL}\left(g||g^*\right).$$

Alternatively, one may pick any $g' \in \mathcal{G}$ in place of g^*.

Proof. We show this by transforming the RHS into the LHS via application of Lemma G.1. The RHS is equal to

$$g^{*\top} \log g^* - g^\top \log g^* + g^\top \log g - g^{*\top} \log g.$$

Applying the aforementioned Lemma twice gives

$$\theta^{*\top}((b^*; w^*) - (b; w)) + \nu^{*\top}(\alpha^* - \alpha) + \theta^\top((b; w) - (b^*; w^*)) + \nu^\top(\alpha - \alpha^*)$$

which is exactly the LHS. □

It is not hard to construct $g', g \in \mathcal{G}$ where the sum of KL divergences is arbitrarily larger than $D_{KL}(g^*||g)$, the actual quantity we want to bound.

Theorem G.3 (Theorem 7.1)**.** *Suppose the pattern distribution α^* was known and the BF weights were θ^{bf}. Letting $(b^{bf}, w^{bf}) = Ag^{(\theta^{bf}),\alpha^*}$ be the LF accuracies and class frequencies that one would get if the BF labeling were used as the ground truth,*

$$\theta^{*\top}(b^* - b^{bf}; w^* - w^{bf}) \geq D_{KL}\left(g^*||g^{(\theta^{bf}),\alpha^*}\right) + D_{KL}\left(g^{(\theta^{bf}),\alpha^*}||g^*\right).$$

Proof. By construction, $g^{(\theta^{bf}),\alpha^*} \in \mathcal{G}$ so we can apply Lemma G.2 where $g^{(\theta^{bf}),\alpha^*}$ is chosen to be g. It suffices to show that

$$\theta^{bf\top}((b; w) - (b^*; w^*)) \leq 0 \tag{26}$$

as removing those terms would give an upper bound on the sum of KL divergences and give us the claimed result.

Now, recall in the proof of Theorem E.1 that $\theta = \sigma' - \sigma$ where σ' was associated with the constraint $Az \geq b - \epsilon$ (for $z \in \Delta_{\tau k}$) and σ was associated with $Az \leq b + \epsilon$. Specifically, we had abused notation and had b represent both LF accuracies and class frequencies. Moreover, $\sigma, \sigma' \in \mathbb{R}^{p+k}_{\geq 0}$, i.e. are non-negative. The complementary slackness conditions for BF (e.g. see Section 5.5.4 of [Boyd and Vandenberghe, 2004]) imply that for all $j \in [p + k]$, we must have

$$\sigma_j(b_j^{bf} - b_j - \epsilon_j) = 0 \qquad \text{and} \qquad \sigma'_j(b_j - \epsilon_j - b_j^{bf}) = 0.$$

Now, if σ_j and σ'_j are both equal to 0, then the respective term in Equation 26 is 0 and we need not consider it to show the inequality. Moreover, one can show that only one of σ_j, σ'_j is nonzero. (E.g. one can maintain the same θ_j by translating both terms so that one equals 0 and another is non-negative.) The last fact that we need to recall is that by assumption, $b^* \in [b - \epsilon, b + \epsilon]$ where the inclusion is elementwise.

Thus, if $\sigma_j > 0$, then it must be the case that

$$b_j^{bf} - b_j - \epsilon_j = 0 \qquad \text{which implies that} \qquad b_j^{bf} = b_j + \epsilon_j \geq b_j^*.$$

Therefore, $\theta_j \leq 0$ by assumption so that $\theta_j(b_j - b_j^*) \leq 0$. A similar strategy works to show that if $\theta_j \geq 0$, the respective term is also non-positive. □

The same proof strategy as above can be used to shown the following result, where the pattern distribution is unknown and BF outputs ν^{bf}.

Corollary G.4. *Let α^{bf} be the pattern distribution that BF estimates, i.e. $\alpha^{bf} = Dg^{(\theta^{bf},\alpha^{bf})}$.*

$$\theta^{*\top}(b^* - b^{bf}; w^* - w^{bf}) + \nu^{*\top}(\alpha^* - \alpha^{bf}) \geq D_{KL}\left(g^*||g^{(\theta^{bf},\nu^{bf})}\right) + D_{KL}\left(g^{(\theta^{bf},\nu^{bf})}||g^*\right)$$

H NON-OPTIMAL WEIGHT COMPARISON

In this section, we prove Theorem 7.2, Lemma 7.3 and Lemma 7.4.

Before doing that, we first show this simple but useful result, which will be applied twice. It differs from Lemma G.1 in that we expand ν so the pattern distribution is explicit.

Lemma H.1. *Let $g, g', g'' \in \mathcal{G}$ where b', α', etc. is associated with Ag'. Then,*

$$(g' - g'')^\top \log g = \theta^\top((b'; w') - (b''; w'')) + (\log \alpha - \log \widehat{Z})^\top(\alpha' - \alpha'')$$

where $\widehat{Z} = (\widehat{Z}_1, \ldots, \widehat{Z}_t)$ is the normalizing factor of $\widehat{g}$ above.

Proof. We'll transform the LHS into the RHS. First, write

$$g_{t\ell} = \frac{\exp\left([A^\top\theta]_{t\ell} + \log \alpha_t\right)}{\widehat{Z}_t}.$$

Then, the LHS is equal to

$$(g' - g'')^\top(A^\top\theta + \log(D^\top\alpha) - \log(D^\top\widehat{Z})).$$

Simplifying, we get

$$\theta^\top((b'; w') - (b''; w'')) + (\log \alpha - \log \widehat{Z})^\top(\alpha' - \alpha'').$$

□

Theorem H.2 (Theorem 7.2)**.**

$$D_{KL}(g^*||g) - D_{KL}(g^*||g') = D_{KL}(g'||g) + (\theta' - \theta)^\top(b^* - b'; w^* - w') + (\alpha^* - \alpha')^\top \log\left(\frac{\alpha'\widehat{Z}}{\alpha\widehat{Z}'}\right).$$

Proof. Observe that the LHS can be written as

$$g^{*\top}\log g^* - g^{*\top}\log g - g^{*\top}\log g^* + g^{*\top}\log g' = -g^{*\top}\log g + g^{*\top}\log g'.$$

Now, adding 0 twice,

$$= (g^* - g')^\top(\log g' - \log g) + g'^\top \log\left(\frac{g'}{g}\right).$$

The second term is $D_{KL}(g'||g)$ where we may apply Lemma H.1 twice to the first term. The first term is equal to

$$\theta'^\top((b^*; w^*) - (b'; w')) + (\log \alpha' - \log \widehat{Z}')^\top(\alpha^* - \alpha') - \theta^\top((b^*; w^*) - (b'; w')) - (\log \alpha - \log \widehat{Z})^\top(\alpha^* - \alpha').$$

Gathering terms gives the claimed result. □

Now, we assume that $\alpha^* = \alpha'$ and apply Cauchy-Schwarz.

Lemma H.3 (Lemma 7.3)**.** *g' is better than g (with respect to g^*) if*

$$\|(b^* - b', w^* - w')\|_2 \leq \frac{D_{KL}(g'||g)}{\|\theta - \theta'\|_2}.$$

Proof. Observe from Theorem H.2 that in order for the LHS to be non-negative, it suffices for

$$-D_{KL}(g'||g) \leq (\theta' - \theta)^\top(b^* - b'; w^* - w').$$

One may weaken this and require that

$$\left|(\theta' - \theta)^\top(b^* - b'; w^* - w')\right| \leq D_{KL}(g'||g).$$

The LHS can be upper bounded by Cauchy-Schwarz for

$$\|(b^* - b', w^* - w')\|_2\|\theta - \theta'\|_2 \leq D_{KL}(g'||g).$$

Rearranging gives the claimed result. □

In practice, this lemma is weak because it rules out the possibility that the term we applied Cauchy-Schwarz to is negative. In that case, it can be arbitrarily small. See the Section K for those results.

On the other hand, we can use the complementary slackness condition that is present when using the BF weighting strategy to give cases where the BF weighting strategy is better than other strategies.

Lemma H.4 (Lemma 7.4)**.** *Let $\theta^{bf} \in (\mathbb{R} \setminus \{0\})^{p+k}$ be the BF weights and fix $\theta \in \mathbb{R}^{p+k}$. Suppose for $j \in [p+k-1]$ that either $\theta_j^{bf} > 0$ and $\theta_j \leq \theta_j^{bf}$ holds or $\theta_j^{bf} < 0$ and $\theta_j \geq \theta_j^{bf}$ holds. Then, $g^{(\theta^{bf}),\alpha^*}$ is better than $g^{(\theta),\alpha^*}$.*

Proof. We show that given the above assumptions, we can show that

$$(\theta^{bf} - \theta)^\top (b^* - b^{bf}; w^* - w^{bf}) \geq 0.$$

Here, $Ag^{(\theta^{bf}),\alpha^*} = (b^{bf}, w^{bf})$ Because we assumed that the pattern distribution is known, having the above is sufficient for $g^{(\theta^{ds}),\alpha^*}$ to be better than $g^{(\theta),\alpha^*}$ by using Theorem H.2.

To start, observe that we can ignore θ_{p+k}^{bf} and θ_{p+k} because we may set

$$\theta_{p+\ell} \leftarrow \theta_{p+\ell} - \theta_{p+k}$$

(and similarly for θ^{bf}) without changing the prediction.

Now, because we have assumed that $\theta_j^{bf} \neq 0$ (for $j \in [p+k-1]$), we consider the cases where it is positive and negative. Recall from the proof of Theorem G.3 that $\theta_j > 0$ implies that $b_j^{bf} \leq b_j^*$. Similarly, $\theta_j^{bf} < 0$ implies that $b_j^* \leq b_j^{bf}$.

It suffices to observe that $\theta_j^{bf} > 0$ and $\theta_j^{bf} > \theta_j$ (as we had assumed) implies that

$$(\theta_j^{bf} - \theta_j)(b_j^* - b_j^{bf}) \geq 0.$$

Here we've abused notation and written b even though the terms in the second parentheses can be w. Something similar holds for the case where $\theta_j^{bf} < 0$. □

I COVARIATE SHIFT ANALYSIS

Here, we show Theorem 7.5, a result governing how much covariate shift (i.e. pattern distribution change) one can tolerate before another prediction becomes better than the one selected.

Theorem I.1 (Theorem 7.5)**.** *Suppose that*

$$D_{KL}\left(g^{(\theta^*),\alpha} || g^{(\theta),\alpha}\right) \leq D_{KL}\left(g^{(\theta^*),\alpha} || g^{(\theta'),\alpha}\right)$$

and let ζ be a perturbation of α such that $\alpha^* = \alpha + \zeta$. If

$$\zeta \in \left\{ \zeta' \in \mathbb{R}^\tau \colon \alpha + \zeta' \in \Delta_\tau, \sum_{t=1}^{\tau} (\alpha_t + \zeta_t') \left[D_{KL}\left(\widehat{g}_t^* || \widehat{g}_t^{(\theta')}\right) - D_{KL}\left(\widehat{g}_t^* || \widehat{g}_t^{(\theta)}\right) \right] \leq 0 \right\},$$

then

$$D_{KL}\left(g^{(\theta^*),\alpha*} || g^{(\theta),\alpha*}\right) \geq D_{KL}\left(g^{(\theta^*),\alpha^*} || g^{(\theta'),\alpha^*}\right).$$

Proof. Observe that our assumption of $g^{(\theta),\alpha}$ being closer to $g^{(\theta^*),\alpha}$ than $g^{(\theta'),\alpha}$ is equivalent to

$$g^{(\theta^*),\alpha\top} \log\left(\frac{g^{(\theta),\alpha}}{g^{(\theta'),\alpha}}\right) \geq 0$$

Since the two labelings in question both predict the same pattern distribution, we may change the α's in the logarithm to α^*. (See Equation 25 for the form of prediction.) By using the fact that $g^{(\theta^*),\alpha} = D^\top \alpha \circ \widehat{g}^{(\theta^*)}$, we may rewrite the above as

$$\sum_{t=1}^{\tau} \alpha_t \sum_{\ell=1}^{k} \left[\widehat{g}_{t\ell}^{(\theta^*)} \log\left(\frac{g_{t\ell}^{(\theta),\alpha^*}}{g_{t\ell}^{(\theta'),\alpha^*}}\right)\right] = \sum_{t=1}^{\tau} \alpha_t \sum_{\ell=1}^{k} \left[\widehat{g}_{t\ell}^{(\theta^*)} \log\left(\frac{\widehat{g}_{t\ell}^{(\theta)}}{\widehat{g}_{t\ell}^{(\theta')}}\right)\right] \geq 0.$$

Observe that terms in the inner sum have no dependence on α now. Moreover, one may write the inner sum as a difference of KL divergences. Specifically,

$$\sum_{t=1}^{\tau} \alpha_t \left[D_{KL}\left(\widehat{g}_t^{(\theta^*)} || \widehat{g}_t^{(\theta')}\right) - D_{KL}\left(\widehat{g}_t^{(\theta^*)} || \widehat{g}_t^{(\theta)}\right)\right] \geq 0.$$

To show the above result, consider the perturbation ζ needed to make this inequality negative. That is exactly the ζ being specified in the theorem statement. Moreover, once that perturbation has been added on, one may reverse the steps to recover the comparison of KL divergences of predictions with pattern distribution α^*. □

Note that the set in the theorem statement is the intersection of a hyperplane with other constraints ensuring that $\alpha + \zeta$ remains a probability distribution. Thus, if one is inclined, they may determine the size of this set by use of LattE [Baldoni et al., 2013]. We discuss the formalism for this in the Experiments section (Section K) at the end of the appendix.

J OPTIMAL PATTERN DISTRIBUTIONS OF SIZE MEASURE ZERO

Now, we provide the proofs for Lemma 6.1 and Theorem 6.3. To show that theorem, we will need two results. First is that $dim(T)$ is large, specifically that dimension is equal to $\tau - 1$, the dimension of the probability simplex Δ_τ. Second is that $dim(O^\theta)$ is small when θ only depends on the LF accuracies and class frequencies. Concretely, $dim(O^\theta)$ is no larger than $\tau - 2$. The aforementioned lemma is a consequence of the first result as we exhibit τ linearly independent vectors whose convex hull lies in T.

Theorem J.1 (Theorem 6.3)**.** *Fix $b^* \in (0,1)^p$ and $w^* \in \Delta_k \cap (0,1)^k$. Suppose θ only depends on b^*, w^* and has at least one nonzero element. If $p \geq 3$, $k \geq 3$, and*

$$\max_{j \in [p]} b_j^* \leq (1 + 1/\sqrt{p(p-1)})^{-1},$$

then $Pr(g^{(\theta),\alpha^*} = g^* | b^*, w^*) = 0$.

Proof. By Lemma J.2, we know that $dim(O^\theta) \leq \tau - 2$. By Theorem J.3, whose assumptions are satisfied by this theorem's assumptions, we know that $dim(T) = \tau - 1$. Since the volume operator was assumed to operate on dimension $\tau - 1$, $vol(O^\theta) = 0$. □

Lemma J.2. *Suppose that a weighting strategy outputs weights θ based solely on b^*, w^*. Moreover, suppose that at least one element of θ is nonzero. Then, $dim(O^\theta) \leq \tau - 2$.*

Proof. Recall that the dimension of a convex polytope like O^θ is the dimension of the affine hull that contains it. This means that equality constraints in the H-Representation of said polytope reduce the dimension. In our case, we show the existence of 2 linearly independent equality constraints, implying that the dimension of the polytope is at most $\tau - 2$. Recall that $O^\theta \subset \Delta_\tau$ and is defined as

$$O^\theta = \{\alpha \in \Delta_\tau \colon A g^{(\theta),\alpha} = (b^*; w^*)\}.$$

We may rewrite this set as follows. Define a matrix $A^{\widehat{g}} \in [0,1]^{(p+k)\times\tau}$ as follows for some $\widehat{g} = \widehat{g}^{(\theta)} \in \Delta_k^\tau$.

$$A^{\widehat{g}} = a_{jt}^{\widehat{g}} = \begin{cases} \widehat{g}_{t\pi_{tj}} & \text{if } j \leq p \\ \widehat{g}_{t,j-p} & \text{if } j > p. \end{cases} \tag{27}$$

One may check that $A^{\widehat{g}}\alpha = AD^\top \alpha \circ \widehat{g}^{(\theta)}$. Thus, we may write O^θ as

$$O^\theta = \{\alpha \in \Delta_k \colon \|A^{\widehat{g}}\alpha - (b^*; w^*)\|_\infty = 0\}.$$

Since α is a probability distribution, we have the constraint that $\mathbf{1}_\tau^\top \alpha = 1$. It suffices to show that there exists a row of $A^{\widehat{g}}$ that is not a multiple of the all 1's vector. Since we assumed that at least one element of θ is nonzero, $\widehat{g}^{(\theta)}$ cannot equal $1_{\tau k}/k$. Since $\widehat{g}^{(\theta)} \in \Delta_k^\tau$, the concatenation of τ many distributions over k elements, this means that there is at least one pattern t where the distribution is not uniform. Now, observe that one of the two following cases must hold. One possibility

is that for each $\ell \in [k]$, $\widehat{g}_{t\ell}^{(\theta)}$ is the same for all $t \in [\tau]$. By our assumption that θ has at least one nonzero element, there is some pattern t where $\widehat{g}_{t\ell}^{(\theta)} \neq \widehat{g}_{t\ell'}^{(\theta)}$. Otherwise, there is some ℓ wherein there exist $t, t' \in [\tau]$ where $\widehat{g}_{t\ell}^{(\theta)} \neq \widehat{g}_{t'\ell}^{(\theta)}$. In this latter case, we are done as row $p + \ell$ of $A^{\widehat{g}}$ is not a multiple of the all 1's vector. In the former case, observe that an LF never predicts one singular class. (One can see this by observing that over the $\tau = k^p$ patterns, an LF will predict each of the k classes.) Therefore, that LF will predict two different probability masses, meaning the associated row in $A^{\widehat{g}}$ will have two different entries. In other words, it is not a multiple of the all ones vector, finishing our proof.

□

Theorem J.3. *Fix $b^* \in (0,1)^p$ and $w^* \in \Delta_k \cap (0,1)^k$. Suppose $p \geq 3$, $k \geq 3$, and the maximum LF accuracy is not too large, i.e.*

$$\max_{j \in [p]} b_j^* \leq \max_{\ell \in [k]} \frac{1}{1 + w_\ell^* (p(p-1))^{-\frac{1}{2}}} \leq \frac{1}{1 + (p(p-1))^{-\frac{1}{2}}}.$$

Then, $dim(T) = \tau - 1$.

This proof is quite involved, but not very difficult. The idea is to perturb α^{ds} (c.f. Proposition C.2) in $\tau = k^p$ linearly independent directions to get τ points in T. The convex hull of those points will have dimension $\tau - 1$, showing the result. Before presenting the proof, we first give a proof sketch outlining the main steps. In the proof sketch, we will provide extra notation and define certain objects that will be used in the full proof. A table summarizing the new notation introduced will be provided before the full proof is given. Then, we show some helpful results that will be used in the proof. Finally, we give the proof in full.

Proof Sketch. We will exhibit τ linearly independent pattern distributions all in T. The convex combination of these vectors will be shown to have dimension $\tau - 1$. Since $T \subset \Delta_\tau$ and Δ_τ has dimension $\tau - 1$, T has dimension $\tau - 1$.

To exhibit these linearly independent pattern distributions, we will take α^{ds} from Proposition C.2 which is in T. Moreover, our assumptions guarantee that each of its masses are positive. We'll perturb α^{ds} in τ different ways, where the t^{th} perturbation, call it $\alpha^{ds} + r^t \in T$ and is such that $[\alpha^{ds} + r^t]_t = 0$, i.e. the t^{th} element is 0.

It will turn out that showing $r^1, \ldots, r^t$ is an affinely independent set of vectors is enough to guarantee that the perturbed vectors are linearly independent. □

J.1 PROOF PRELIMINARIES AND SOME USEFUL RESULTS

We now review some needed facts and prove some simple results that will aid in our proof of the above theorem.

J.1.1 Patterns and Proto-Patterns

Observe that we may partition the τ patterns by looking at how many classes have predictions. That is, for each $i \in [k]$, one may group all patterns π_t wherein

$$\sum_{\ell=1}^{k} \mathbf{1}(|\pi_t^\ell| > 0) = i.$$

We'll say that π_t has *spread* i. One can go one step further for all patterns with spread i. Consider all partitions of $[p]$ into i nonempty sets. There are $\left\{ {p \atop i} \right\}$ many of these partitions where $\left\{ {p \atop i} \right\}$ is the Stirling number of the second kind. Call each of these $\left\{ {p \atop i} \right\}$ partitions a *proto-pattern* with spread i.

Indeed, observe that from a proto-pattern $\tilde{\pi}$ with spread i that one may create $k!/(k-i)!$ patterns. That is, one has that many ways to assign classes to the i subsets of LFs. Thus, we say that proto-pattern $\tilde{\pi}$ with spread i can spawn $k!/(k-i)!$ patterns. For example, take a proto-pattern $\tilde{\pi}$ with spread 2.

$$\tilde{\pi} = \{\{1\}, \{2,3\}\}.$$

If there are $k = 3$ classes, the patterns it can spawn are

$$(\{1\}, \{2,3\}, \emptyset), \quad (\{1\}, \emptyset, \{2,3\}), \quad \text{etc.}$$

Now, let $V_1, \ldots, V_k$ be a partition of the proto-patterns where V_i contains all proto-patterns with spread i. Also, fix an enumeration of the $\tau = k^p$ patterns. To exhibit the τ linearly independent pattern distributions, we will provide three arguments. One will be for patterns spawned by proto-patterns in V_{k-1}, V_k. Another will be for those spawned by proto-patterns in $V_2, \ldots, V_{k-2}$. Finally, we will provide one more argument spawned by the sole proto-pattern in V_1.

J.1.2 Pattern Mass Redistribution

Recall now the definition of T (Equation 20):

$$T = \{\alpha \in \Delta_\tau \colon \exists z \in \Delta_{\tau k},\ Az = (b^*; w^*),\ Dz = \alpha\}.$$

The way we showed that there is an $\alpha^{ds} \in T$ is to actually demonstrate the existence of $g^{ds} = g^{(\theta^{ds}, \mathbf{0}_\tau)}$ where $Ag^{ds} = (b^*; w^*)$. Then, it's easy to see that $Dg^{ds} \in T$. This means that if we want to perturb α^{ds} by making the t^{th} element 0, then it suffices to exhibit a perturbation of g^{ds} such that $[g^{ds} + r]_{t\ell} = 0$ for each $\ell \in [k]$ and $A(g^{ds} + r) = (b^*; w^*)$. That is to say, we need to redistribute the masses $g^{ds}_{t1}, \ldots, g^{ds}_{tk}$ without changing the LF accuracies and class frequencies.

We'll now demonstrate how to do this on a simple example. The actual proof will require us to address more technical details. Suppose $p = k = 3$ and consider the following pattern:

$$\pi_t = (\{2\}, \{1\}, \{3\}).$$

To make α^{ds}_t equal 0, we must redistribute the masses $g^{ds}_{t1}, g^{ds}_{t2}, g^{ds}_{t3}$, maintaining the LF accuracies and class frequencies. Consider now a perturbation vector $r \in [-1, 1]^{\tau k}$ which is 0 except on r_{t1}, r_{t2}, r_{t3}, where it is

$$r_{t1} = -g^{ds}_{t1}, \quad r_{t2} = -g^{ds}_{t2}, \quad r_{t3} = -g^{ds}_{t3}.$$

Then, $D(g^{ds} + r)$ is 0 on the t^{th} element, but $A(g^{ds} + r) \neq (b^*; w^*)$. In particular,

$$A(g^{ds} + r) = (b^*_1 - g^{ds}_{t2},\ b^*_2 - g^{ds}_{t1},\ b^*_3 - g^{ds}_{t3},\ w^*_1 - g^{ds}_{t1},\ w^*_2 - g^{ds}_{t2},\ w^*_3 - g^{ds}_{t3})^\top.$$

I.e. g^{ds}_{t1} contributes to $h^{(2)}$'s accuracy and w^*_1, the first class' frequency, etc.

Thus, we need to find patterns wherein we can distribute the masses $g^{ds}_{t1}, g^{ds}_{t2}, g^{ds}_{t3}$. The way be do this is to use other patterns spawned by the proto-pattern that spawned π_t. That is, observe the proto-pattern which spawned π_t is

$$\tilde{\pi} = \{\{1\}, \{2\}, \{3\}\}.$$

Consider now the following three patterns:

$$\pi_{t_1} = (\underline{\{2\}}, \{3\}, \{1\}), \quad \pi_{t_2} = (\{3\}, \underline{\{1\}}, \{2\}), \quad \pi_{t_3} = (\{1\}, \{2\}, \underline{\{3\}}).$$

Note that the underlined sets are unchanged from π_t. If we increase $g^{ds}_{t_1,1}$ by g^{ds}_{t1}, then we have increased $h^{(2)}$'s accuracy and the first class frequency by g^{ds}_{t1}. This is exactly the gap in accuracy/class frequency that we got by removing all mass from pattern t. So, change the perturbation vector as follows:

$$r_{t_1,1} = g^{ds}_{t1}, \quad r_{t_2,2} = g^{ds}_{t2}, \quad r_{t_3,3} = g^{ds}_{t3}.$$

Now, $g^{ds} + r \in \Delta_{\tau k}$ and $A(g^{ds} + r) = (b^*; w^*)$ and as before, $[D(g^{ds} + r)]_t = 0$.

Observe that we simply permuted the class labels for the sets of LFs to create the patterns we needed to redistribute the probability mass. We'll generalize this strategy to k classes, though we will need to assume that $k \geq 3$ for this to work! Since writing out the actual permutation of sets of LFs is cumbersome, we will appeal to cycle notation. Consider the permutations over k elements, i.e. functions $\sigma \colon [k] \to [k]$. Let (ij) be a permutation over k elements defined as follows:

$$(ij)(\ell) = \begin{cases} \ell & \text{if } \ell \neq i, j \\ j & \text{if } \ell = i \\ i & \text{if } \ell = j. \end{cases} \tag{28}$$

In words, index i gets sent to j, index j gets sent to i, and all other indices are sent to themselves. Thus, π_{t_1} is gotten by applying (23) to the classes of π_t, π_{t_2} is gotten by applying (13), and π_{t_3} is gotten by applying (12). Written another way, $\pi^\ell_{t_1} = \pi^{(23)(\ell)}_t$.

J.1.3 Linear and Affine Independence

Recall that for a set of vectors $\{u_i\}_{i=1}^n$, they are linearly independent just in case there does not exist an i where one may write

$$u_i = \sum_{\substack{i'=1 \\ i' \neq i}}^{n} a_{i'} u_{i'} \quad \text{for} \quad a_{i'} \in \mathbb{R}.$$

That same set of vectors is affinely independent just in case there is no such i where

$$u_i = \sum_{\substack{i'=1 \\ i' \neq i}}^{n} a_{i'} u_{i'} \quad \text{for} \quad a_{i'} \in \mathbb{R} \quad \text{and} \quad \sum_{\substack{i'=1 \\ i' \neq i}}^{n} a_{i'} = 1.$$

It is easy to see that linear independence implies affine independence. Equivalently, that same set is affinely independent if $\{u_i - u_1\}_{i=2}^n$ is a linearly independent set of vectors.

Proposition J.4. *Let $\alpha \in \Delta_\tau$ and $\{r^t\}_{t=1}^\rho$ be perturbation vectors such that for each $t \in [\rho]$, $\alpha + r^t \in \Delta_\tau$. Then, $\{\alpha + r^t\}_{t=1}^\rho$ is linearly independent if $\{r^t\}_{t=1}^\rho$ is affinely independent.*

Proof. Without loss of generality, suppose that for contradiction $\alpha + r^1$ could be written as a linear combination of the remaining $\rho - 1$ vectors. That is,

$$\alpha + r^1 = \sum_{i=2}^{\rho} a_i(\alpha + r^i)$$

for $a_i \in \mathbb{R}$. This implies that

$$\alpha \left(1 - \sum_{i=2}^{\rho} a_i\right) = \sum_{i=2}^{\rho} a_i r^i - r^1.$$

Since this equality holds, one of two things must be true. Either the set $\{r^i\}_{i=1}^\rho$ is affinely dependent or it's not. However, we had assumed that the set is affinely independent. This means that it is not possible to make $1 - \sum_{i=2}^n a_i = 0$ while also making the RHS equal to 0 at the same time. Thus, we assume that $1 - \sum_{i=2}^n a_i \neq 0$. Our equality can then be rearranged into

$$\alpha = \frac{\sum_{i=2}^{\rho} a_i r^i - r^1}{(1 - \sum_{i=2}^{\rho} a_i)}.$$

Since $\alpha \in \Delta_\tau$, $\mathbf{1}_\tau^\top \alpha = 1$. As $\alpha + r^t \in \Delta_\tau$, $\mathbf{1}_\tau r^t = 0$ for all $t \in [\rho]$. If we take the dot product of each side of our equality with $\mathbf{1}_\tau$, we get $1 = 0$, a contradiction. □

Proposition J.5. *The convex hull of τ linearly independent vectors has dimension $\tau - 1$.*

Proof. Suppose we have τ linearly independent vectors $\{u_1, \ldots, u_\tau\} \subset \mathbb{R}^\tau$. Their convex hull is

$$\left\{ \sum_{t=1}^{\tau} a_t u_t \in \mathbb{R}^\tau : \sum_{t=1}^{\tau} a_t = 1, a_t \geq 0, \forall t \in [\tau] \right\}.$$

Recall that the dimension of a convex hull is defined as the dimension of the affine hull that contains said convex hull. We are thus interested in the dimension of

$$\left\{ \sum_{t=1}^{\tau} a_t u_t \in \mathbb{R}^\tau : \sum_{t=1}^{\tau} a_t = 1 \right\}.$$

This is equivalent to finding the dimension of $\{u_1 - u_\tau, u_2 - u_\tau, \ldots, u_{\tau-1} - u_\tau\}$, a set of $\tau - 1$ vectors. Suppose for contradiction that the set was not linearly independent. Then, there are coefficients a_i, $i \in [\tau - 1]$ which are not all zero, such that

$$\mathbf{0}_\tau = \sum_{i=1}^{\tau-1} a_i(u_i - u_\tau).$$

This implies that

$$\mathbf{0}_\tau = \sum_{i=1}^{\tau-1} a_i u_i - \left(\sum_{i=1}^{\tau-1} a_i\right) u_\tau,$$

which in turn implies that $\{u_1, \ldots, u_\tau\}$ is not linearly independent, a contradiction. Hence, the convex hull of τ linearly independent vectors has dimension $\tau - 1$. □

J.1.4 Weighted Directed Graphs and the Graph Laplacian

Recall that a weighted directed graph G is a tuple (V, E, ω) where V is the set of vertices, E is the set of edges and $\omega \colon E \to \mathbb{R}$ is a function recording the weights of each edge. For $e \in E$, we will write $\omega_e \coloneqq \omega(e)$. If we have $(t, t') \in E$, then we write $\omega_{tt'} \coloneqq \omega((t, t'))$. For our specific case, we assume that there are no self-loops and the following holds:

1. If $(t, t') \in E$, then $(t, t') \in E$,
2. $\omega_{tt'} = \omega_{t't}$.

That is, if node t is connected to node t', then node t' is also connected to t. Moreover, the weights of these edges must be the same. From this, we can construct an undirected weighted graph $G' = (V, E', \omega)$ where edges $(t, t'), (t', t)$ become $\{t, t'\}$ with weight $\omega_{tt'}$.

The weighted Laplacian of G' is defined as $L = D - A \in \mathbb{R}^{|V| \times |V|}$ where D is a diagonal matrix with

$$D = d_{ii} = \sum_{e \in D'} \mathbf{1}(e \text{ connects node } i)\omega_e \tag{29}$$

and A is the weighted adjacency matrix defined as follows.

$$A = a_{ij} = \begin{cases} 0 & \text{if } i = j \\ \omega_{ij} & \text{if } \{i, j\} \in E' \\ 0 & \text{o.w.} \end{cases} \tag{30}$$

Now, we show that if G' has one connected component, then its Laplacian has affinely independent columns.

Proposition J.6. *Suppose that G' has 1 connected component. Then, the columns of its weighted Laplacian matrix L are affinely independent.*

Proof. It is easy to verify that the rank of L is $|V| - 1$ if there is only 1 connected component. I.e. show that $\mathbf{1}_{|V|}$ spans the entire null space of L. Let $u_1, \ldots, u_{|V|}$ be the columns of L. For those columns to be affinely independent, it suffices to show that $\{u_2 - u_1, \ldots, u_{|V|} - u_1\}$ is a linearly independent set.

To do this, observe that

$$u_1 = -\sum_{i=2}^{|V|} u_i.$$

Since L has rank $|V| - 1$, this means that the set $\{u_i\}_{i=2}^{|V|}$ is linearly independent. We wish to show that there does not exist an $i \in [|V|] \setminus \{1\}$ for which

$$u_i - u_1 = \sum_{\substack{i'=2 \\ i' \neq i}}^{n} a_{i'}(u_{i'} - u_1) \quad \text{for} \quad a_{i'} \in \mathbb{R}.$$

If we define $a_i = 1$, then we may write the above as

$$\mathbf{0}_{|V|} = \sum_{i'=2}^{|V|} a_{i'}(u_{i'} - u_1) = \sum_{i'=2}^{|V|} a_{i'} \left(u_{i'} + \sum_{i''=2}^{|V|} u_{i''}\right)$$

where the second equality was by using the fact that u_1 was a linear combination of the other columns of L. Gathering the coefficients for the vectors, we get

$$\mathbf{0}_{|V|} = \sum_{i'=2}^{|V|} a_{i'} u_{i'} + \sum_{i'=2}^{|V|} a_{i'} \left(\sum_{i''=2}^{|V|} u_{i''} \right) = \sum_{i'=2}^{|V|} \left(a_{i'} + \sum_{i''=2}^{|V|} a_{i''} \right) u_{i'}. \tag{31}$$

Since the $u_{i'}$'s being summed are linearly independent, the only way the equality holds is if we can select $a_{i'}$ values so that each expression enclosed by the parentheses is equal to 0. Showing this would complete the proof.

What we'll do is write the coefficients of the $u_{i'}$ as a linear combination of the $a_{i'}$ values. Gather the $a_{i'}$'s into a vector, calling it $a \in \mathbb{R}^{|V|-1}$. The coefficients of $u_{i'}$ can be written in vector form as

$$(I_{|V|-1} + \mathbf{1}_{|V|-1} \mathbf{1}_{|V|-1}^\top) a.$$

If we can show that this matrix is invertible, then the only way for the resulting matrix vector product to be the zeros vector is if the a vector was all zeros.

We demonstrate this by showing its $|V| - 1$ eigenvectors and eigenvalues. The first eigenvector is $\mathbf{1}_{|V|-1}$ with eigenvalue $|V|$. The remaining $|V| - 2$ eigenvectors have eigenvalue 1 and are of the form $e_1 - e_i$ where e_i is the i^{th} canonical basis vector in $|V| - 1$ dimensions for $i \in [|V| - 1] \setminus \{1\}$. Thus, the matrix in question is invertible, and we conclude that the columns of the Laplacian are affinely independent. □

J.1.5 Permutations and the Symmetric Group

To construct the graph(s) for the previous question, we will appeal to group theory. Specifically, we will be dealing with permutations, making it natural to consider the symmetric group.

Recall that a group is a set G and a binary operation, denoted "$\circ$" wherein the following axioms are satisfied. (E.g. see Section 2.2 of [Artin, 2018].)

1. The operation is associative on elements in G
2. There exists an identity element $e \in G$ such that $e \circ g = g$ for any $g \in G$
3. For every $g \in G$, there exists an inverse element $g' \in G$ such that $g \circ g' = e$

We will consider G to be the set of permutations over k elements, which is sometimes written as S_k. Formally, a permutation on k elements is a bijective function $\sigma \colon [k] \to [k]$. It's easy to see that $|S_k| = k!$. The binary operator for this group will be function composition.

We will be interested in a specific set S whereupon one can generate all elements of S_k by operations using elements in S. For a formal definition, see Section 7.10 of [Artin, 2018]. For our purposes, S will be a set of cycles of length 2. (See Equation 28 for the definition of a cycle.)

$$S \coloneqq \{(12), (23), \ldots, (k-1, k)\}$$

Using S and S_k, we are able to construct a graph that has the desired properties for our proof. Specifically, the Cayley color graph of a finite group (say S_k) and a set S which generates said group is a directed graph which represents the relationship between group elements in terms of elements in S. (See Chapter 2.3 of [Chartrand and Lesniak, 2005].) Call this graph $G_C = (V, E)$. There is one node per element of the group, i.e. $|V| = |S_k|$. Each element of the generating set $s \in S$ is associated with a color. If for a node $\sigma \in S_k$ one gets $\sigma' = \sigma \circ s$, then there is an edge (σ, σ') with the color associated with s. If $s \in S$ is its own inverse, i.e. $s \circ s = e$, then $(\sigma, \sigma') \in E$ implies that $(\sigma', \sigma) \in E$. Taken together, this implies that every node has $|S| = k - 1$ incoming and outgoing edges. In this particular case, the edges (σ, σ') and (σ', σ) can be combined to form $\{\sigma, \sigma'\}$, meaning the Cayley graph can be turned into an undirected graph. Moreover, if S generates the entire group S_k, then the Cayley graph is fully connected.

Consider now S_3 where $S = \{(12), (23), (31)\}$. We have added the cycle (31) so that there are 3 incoming and outgoing edges per node. The Cayley color graph can be seen in Figure 3. One can show that the patterns spawned by a proto-pattern of spread k is equivalent to the set of permutations over k elements. Moreover, our strategy of exhibiting a new pattern by permuting the classes predicted by two sets of LFs is equivalent to applying a cycle. We'll leverage these facts in our proof of Lemma J.11.

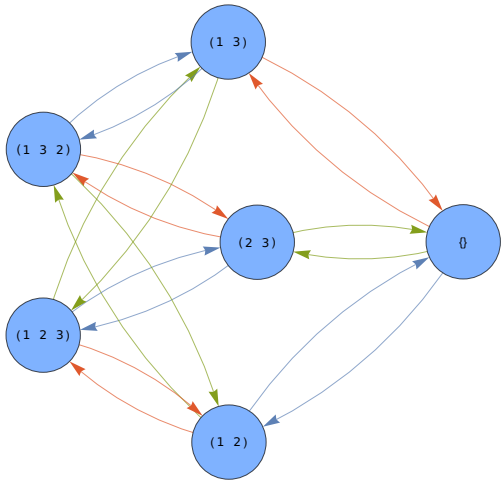


Figure 3: The Cayley Color Graph for S_3 with Generating Set $S = \{(12), (23), (31)\}$. $\{\}$ Denotes the Identity Permutation. Red Edges denote (13), Green Edges (23), and Blue Edges (12).

J.2 NOTATION AND DEFINITION RECAP

See Table 2 for a recap of the notation and definitions.

Table 2: Notation Summary for Theorem J.3's proof

Notation/Word	Explanation
p	number of LFs
k	number of classes
τ	shorthand for k^p, the total number of patterns
g^{ds}	$g^{(\theta^{ds}, \mathbf{0}_\tau)}$ where $g^{ds}_{t\ell}$ is defined in Equation 17
π_t	the t^{th} pattern out of τ, an assignment of LF index to predicted class
π_{tj}	the prediction of the j^{th} LF on pattern t
π_t^ℓ	the subset of LFs in $[p]$ which predict ℓ on pattern t
$\tilde{\pi}$	a proto-pattern, which is a partition of $[p]$ into $i \in [k]$ nonempty subsets
$\left\{ {p \atop i} \right\}$	Stirling number of the second kind, counts the number of proto-patterns for fixed i
$\tilde{\pi}$ spawns π_t	π_t can be created by assigning classes to sets of LFs from the proto-pattern $\tilde{\pi}$
$\tilde{\pi}, \pi$ has spread i	$\tilde{\pi}$ only has $i \in [k]$ sets, π has $i \in [k]$ classes with LF predictions
V_i	a set containing proto-patterns with spread $i \in [k]$
S_k	symmetric group of order k, contains all permutations on k elements
(ij)	cycle notation, represents a permutation which sends element i to position j and vice versa

J.3 A MORE IN DEPTH PROOF STRATEGY FOR THEOREM J.3

Our general strategy is as follows (where b^*, w^* are fixed but arbitrary):

1. Exhibit an $\alpha \in T$ such that all elements of α are positive.
2. Exhibit τ linearly independent vectors all in T by perturbing α.
3. Appeal to Proposition J.5 using those τ vectors to see that $dim(T) \geq \tau - 1$.
4. Use that fact that $T \subset \Delta_\tau$ so that $dim(T) \leq dim(\Delta_\tau) = \tau - 1$ and so $dim(T) = \tau - 1$.

The first point was done in Proposition C.2 and we focus on the second step.

Recall that by Proposition J.4, it suffices for the perturbation vectors (call them r) of α^{ds} to themselves be *affinely* independent for the perturbed vectors ($\alpha^{ds} + r \in \Delta_\tau$) to be *linearly* independent. To get τ perturbation vectors, we will provide τ distributions wherein the t^{th} element of $\alpha^{ds} + r$ is 0. That is to say, we redistribute the mass α^{ds}_t into the other patterns, while maintaining membership in T. (Recall to maintain membership in T, there must be a joint distribution $g' \in \Delta_{\tau k}$ such that $Ag' = (b^*; w^*)$ and $Dg' = \alpha^{ds} + r$.) Of course, we'll need to do this in such a way that the perturbation vectors are affinely independent.

We will be able to write the perturbation vectors in the following matrix, where the matrix is written in block form. The matrix will be a square matrix with τ dimension on each side. On the diagonal, we will have matrices $R_1, \dots R_t$ whose columns are affinely independent. For the last column, we will have matrices $R^1_t, \dots, R^{t-1}_t$.

$$R = \begin{pmatrix} R_1 & \mathbf{0} & \dots & \mathbf{0} & R^1_t \\ \mathbf{0} & R_2 & \dots & \mathbf{0} & R^2_t \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & R_{t-1} & R^{t-1}_t \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & R_t \end{pmatrix} \tag{32}$$

In particular, the first $\tau - k$ columns of R will correspond to patterns with spread > 1 while the last k columns will be associated with the patterns with spread 1. Fix now a proto-pattern $\tilde{\pi}$ with spread $i > 1$. One of the first $t - 1$ (block) columns will be associated with $\tilde{\pi}$. Without loss of generality, let the $k!/(k - i)!$ patterns spawned by $\tilde{\pi}$ have consecutive indices. Now, for a pattern π_t spawned by that proto-pattern, we will only redistribute masses $g^{ds}_{t1}, \dots, g^{ds}_{tk}$ to a pattern also spawned by $\tilde{\pi}$. This allows us to achieve the block diagonal form seen in R.

Proposition J.7. *If the columns of the matrices $R_1, \dots, R_t$ are affinely independent where R_t is a diagonal matrix with nonzero elements on the diagonal, then the columns of R are also affinely independent.*

Proof. For convenience, we'll refer to columns in R by referring to the columns in $R_1, \dots, R_t$, the matrices on the main diagonal. By assumption, if r is a column of $R_{t'}$, $t' \in [t-1]$, then r is affinely independent of other columns in $R_{t'}$. r is also affinely independent of any other column not from R_t as there is no overlap between indices where the vector is nonzero.

We now argue that any column r from $R_1, \dots, R_{t-1}$ is affinely independent from all other columns in R by contradiction. First, observe that coefficients for columns in R_t must be 0 as the last row of r (in block form) is the 0's vector. If r was affinely *dependent*, then it can be written as an affine combination of columns from $R_1, \dots, R_{t-1}$, which we have already argued was not possible.

Finally, we consider a vector r from R_t. Since R_t was assumed to be a diagonal matrix without 0's on the diagonal, r cannot be written as a combination of other column vectors in R. I.e. vectors from $R_1, \dots, R_{t-1}$ are 0 on the last (block) row while R_t being a diagonal matrix implies that the rows are linearly independent. □

Now, all that remains is to prove the affine independence of the columns of each $R_{t'}$ and ensuring that R_t is a diagonal matrix with nonzero elements on its diagonal.

J.4 PROOF STEP: MASS REDISTRIBUTION FOR PATTERNS WITH MAXIMAL SPREAD

Here, we detail how to redistribute the probability masses for patterns that have the maximum spread of k. This is the easiest case and we will build upon the arguments laid out here.

Consider a proto-pattern $\tilde{\pi} \in V_k$, i.e. one with spread k. From $\tilde{\pi}$, we can spawn $k!$ patterns, call them $\pi_1, \dots, \pi_{k!}$. When redistributing the masses $g^{ds}_{11}, \dots, g^{ds}_{1k}$, we will give these masses to the other patterns spawned by $\tilde{\pi}$, i.e. $\pi_2, \dots, \pi_{k!}$. This is necessary to achieve the block diagonal form of R (Equation 32). In words, the perturbation vectors for patterns spawned by $\tilde{\pi}$ will only have nonzero elements on indices in $[k!]$. Indeed, for every fixed proto-pattern with spread > 1 we will only redistribute the pattern mass to patterns spawned by that proto-pattern.

Lets say that the first column of R is associated with this proto-pattern $\tilde{\pi}$. Then, the matrix R_1 is a square matrix of size $k!$. Moreover, columns of R_1 (or actually, the columns of R_1 with $\tau - k!$ 0's appended) record perturbations of the OCDS pattern distribution α^{ds} (see Equation 21) for a total of $k!$ total perturbations. If r is a column of R_1, recall that we require

$$\alpha^{ds} + (r; \mathbf{0}_{\tau - k!}) \in \Delta_\tau \quad \text{and} \quad \alpha^{ds} + (r; \mathbf{0}_{\tau - k!}) \in T.$$

Finally, to meet the conditions of Proposition J.7, the columns of R_1 must be affinely independent. Written out explicitly, these are our requirements for the perturbation vectors r_t:

1. $\alpha + r_t \in T$ for all $t \in [k!]$, i.e. for all patterns spawned by $\tilde{\pi}$,
2. $r_1, \ldots, r_{k!}$ are affinely independent.

We'll address these in turn. For the second point, we will provide some preliminaries.

J.4.1 Constructing Perturbations of α

Recall that for every pattern spawned by $\tilde{\pi}$, we'll redistribute all its mass to other patterns spawned by $\tilde{\pi}$. That is, for π_t spawned by $\tilde{\pi}$, we'll construct a perturbation vector $r^t \in [-1, 1]^\tau$ such that $\alpha + r^t \in T$ and $[\alpha + r^t]_t = 0$. In particular, we will formally generalize the strategy shown in Section J.1.2. To start, we show a useful result about g^{ds}. It is easily generalized to apply to all $g \in \mathcal{G}$.

Proposition J.8. *Let $b^* \in (0,1)^p, w^* \in \Delta_k \cap (0,1)^k$ be fixed. This fixes g^{ds}, which we write as g for short. Also, fix a pattern π_t and class ℓ. For any pattern $\pi_{t'}$ where $\pi_t^\ell = \pi_{t'}^\ell$. I.e. the same set of LFs predict class ℓ on $\pi_t, \pi_{t'}$. Then, if $e_{t\ell}$ is the $(k(t-1)+\ell)^{th}$ canonical basis vector in τk dimensions,*

1. $g_{t\ell} = g_{t'\ell}$
2. $A(g + g_{t\ell}(e_{t'\ell} - e_{t\ell})) = (b^*; w^*)$, *implying that* $D(g + g_{t\ell}(e_{t'\ell} - e_{t\ell})) \in T$.

Proof. We start with the first claim. Reprinting the definition of $g_{t\ell}$ from Equation 19, we have

$$g_{t\ell} = w_\ell^* \prod_{j \in \pi_t^\ell} b_j^* \prod_{j' \in [p] \setminus \pi_t^\ell} \frac{1 - b_{j'}^*}{k-1}.$$

Thus, it only depends on which LFs predict ℓ on pattern π_t. So, if there are two patterns wherein the same set of LFs both predict ℓ, the corresponding masses for class ℓ are equal.

For the second claim, observe that $A(g - g_{t\ell}e_{t\ell})$ reduces the class frequency of ℓ and the accuracies of LFs in π_t^ℓ by $g_{t\ell}$. On the other hand, the addition of the vector $g_{t\ell}e_{t'\ell}$ increases the class frequency of ℓ and the accuracies of LFs in $\pi_t^\ell = \pi_{t'}^\ell$ by $g_{t\ell}$. Thus, there is no net change in the LF accuracies and class frequencies when g is perturbed by $g_{t\ell}(e_{t'\ell} - e_{t\ell})$. Thus, by definition of T, D applied to the perturbed version of g lies in T. □

Remark J.9. Observe that if $k = 2$, then any π_t and $\pi_{t'}$ satisfying the above assumption implies $t = t'$. That is, if $\pi_t^1 = \pi_{t'}^1$, then $\pi_t^2 = \pi_{t'}^2 = [p] \setminus \pi_t^1$. Thus, we will focus on the case where $k \geq 3$.

We'll apply this proposition k times to distribute the mass of α_t in k parts, i.e. by redistributing $g_{t1}^{ds}, \ldots, g_{tk}^{ds}$. What remains is to specify a strategy of getting pattern $\pi_{t'}$ for each ℓ.

The general strategy seen in Subsection J.1.2 is as follows. For fixed $\ell \in [k]$, apply the permutation $(\ell+1, \ell+2)$ to $\pi_t^1, \ldots, \pi_t^k$. Here, the indices are assumed to wrap around insofar as one takes $\ell + 1, \ell + 2$ mod k. E.g. if $\ell + 1 = k + 1$, then the permutation used is $(1, 2)$. Concretely, $\pi_{t'}$ is the pattern such that $\pi_{t'}^{\ell'} = \pi_t^{(\ell+1,\ell+2)(\ell')}$ for each $\ell' \in [k]$ or

$$\pi_{t'}^{\ell+1} = \pi_t^{\ell+2}, \quad \pi_{t'}^{\ell+2} = \pi_t^{\ell+1}, \quad \text{and} \quad \pi_{t'}^{\ell'} = \pi_t^{\ell'} \quad \text{for} \quad \ell' \in [k] \setminus \{\ell+1, \ell+2\}.$$

By definition, $\pi_t^\ell = \pi_{t'}^\ell$, so we can use this pattern with the previous proposition. Moreover, observe that $\pi_{t'}$ is also spawned by $\tilde{\pi}$.

Now, we may formalize the perturbation vector.

Lemma J.10. *Say $k \geq 3$ and $\tilde{\pi}$ is a proto-pattern with spread k. Let $\pi_1, \ldots, \pi_{k!}$ be the patterns spawned by $\tilde{\pi}$. Fix a pattern π_t where $t \in [k!]$ and consider the following k patterns also spawned by $\tilde{\pi}$.*

$$\pi_{t_\ell} \quad \textit{be such that} \quad \pi_{t_\ell}^{\ell'} = \pi_t^{(\ell+1,\ell+2)(\ell')}$$

where any index bigger than k has k subtracted from it so that

$$(\ell+1, \ell+2) \in \{(12), \ldots, (k-1, k), (k1)\}.$$

Now, consider a perturbation vector r^t defined as follows where e_t is the t^{th} canonical basis vector in τ dimensions.

$$r^t = \sum_{\ell=1}^{k} g_{t\ell}^{ds} e_{t_\ell} - \alpha_t^{ds} e_t$$

It has the following properties:

1. $\alpha^{ds} + r^t \in T$
2. $[\alpha^{ds} + r^t]_t = 0$
3. *There are only $k+1$ nonzero elements in r^t*
4. *The last $\tau - k!$ elements of r^t are 0, i.e. the pattern masses are only distributed to patterns spawned by $\tilde{\pi}$.*

Proof. If we apply Proposition J.8 k times, the perturbation of g (where we have redistributed all masses $g_{t1}^{ds}, \ldots, g_{tk}^{ds}$) can be written as

$$\sum_{\ell=1}^{k} g_{t\ell}^{ds} (e_{t_\ell, \ell} - e_{t\ell}).$$

Here, we're abusing notation and $e_{t_\ell,\ell}$ is the $((t_\ell - 1)k + \ell)^{th}$ canonical basis vector in τk dimensions. If D (Equation 7) is applied to that vector, observe that the resulting vector has length τ and is 0 except on $k+1$ elements. Moreover, since $t, t_\ell \leq k!$ (as π_{t_ℓ} are spawned by $\tilde{\pi}$), the last $\tau - k!$ elements of the perturbation vector must be 0, meaning the fourth claim has been shown.

Recall that D computes the pattern distribution from the joint distribution, i.e.

$$D\left(\sum_{t=1}^{\tau}\sum_{\ell=1}^{k} g_{t\ell}^{ds} e_{t\ell}\right) = \sum_{t=1}^{\tau}\left(\sum_{\ell=1}^{k} g_{t\ell}^{ds}\right) e_t.$$

Once again, note that $e_{t\ell}$ is in τk dimensions while e_t is in τ dimensions. So,

$$D\left(\sum_{\ell=1}^{k} g_{t\ell}^{ds} (e_{t_\ell,\ell} - e_{t\ell})\right) = \sum_{\ell=1}^{k} g_{t\ell}^{ds} e_{t_\ell} - \alpha_t^{ds} e_t.$$

In other words, the resulting vector has $-\alpha_t^{ds}$ on index t and $g_{t\ell}^{ds}$ on index t_ℓ. The second and third claims are easy to see by inspection, while our first claim is true by virtue of us using Proposition J.8. □

To get the first block column of R (Equation 32), we horizontally concatenate $r^1, \ldots, r^{k!}$. Since the last $\tau - k!$ elements are guaranteed to be 0, we have maintained the block diagonal form. What remains to be shown is that our perturbation vectors are affinely independent. At a high level, we will appeal to graph theory by showing that R_1 (i.e. the horizontal concatenation of the first $k!$ elements of $r^1, \ldots, r^{k!}$) is a Laplacian matrix for a weighted undirected graph which has one connected component. That will be sufficient. Before we show that, we review some graph theory concepts necessary.

J.4.2 Our Perturbations Form a Laplacian Matrix

Let $r^1, \ldots, r^t \in [-1, 1]^\tau$ be the perturbation vectors exhibited in Lemma J.10. Define $\tilde{r}^t \in [-1, 1]^{k!}$ to be the first $k!$ elements of r^t. Then, the matrix R_1 is defined as follows

$$R_1 \coloneqq \begin{pmatrix} | & | & & | \\ \tilde{r}^1 & \tilde{r}^2 & \ldots & \tilde{r}^{k!} \\ | & | & & | \end{pmatrix}.$$

Lemma J.11. *The matrix $-R_1$ is a Laplacian of a weighted undirected graph that has one connected component and no self loops. Moreover, the graph is isomorphic to an undirected Cayley color graph for S_k, the group of permutations over k elements, generated with the k cycles*

$$\{(12), (23), \ldots, (k-1, k), (k1)\}.$$

Thus, R_1 has affinely independent columns.

Note that the very last implication follows from directly applying Proposition J.6. Thus, we will have shown how to construct $\left\{{p \atop k}\right\} k!$ linearly independent pattern distributions in T (as this is the number of patterns that can be spawned from proto-patterns with maximum spread).

Before proving this result, we review some simple group theory concepts and define the Cayley color graph.

J.4.3 Proof of Lemma J.11

To prove Lemma J.11, it suffices to show the following claims. Let $G_C = (V_C, E_C)$ be the undirected Cayley color graph for S_k with generating set

$$S = \{(12), (23), \ldots, (k-1, k), (k1)\}.$$

Also, let $G = (V, E, \omega)$ denote the weighted undirected graph that R_1 represents. (We'll show this claim.)

1. $-R_1$ is symmetric and can be written in the form $D - A$ for appropriately defined D, A where D is diagonal and A is symmetric. Moreover, each row and column sums to 0. I.e. these matrices result from a weighted undirected graph (see Equations 29, 30).
2. G (without its weights) is isomorphic to G_C. I.e. there is a 1-to-1 mapping of the vertices, call it $f \colon V \to V_C$, where $\{v_1, v_2\} \in E$ if and only if $\{f(v_1), f(v_2)\} \in E_C$.

We start with the first claim where we reference Lemma J.10. Let $D \in [-1, +1]^{k! \times k!}$ be the diagonal elements of $-R_1$:

$$D = d_{ij} = \begin{cases} \alpha_i & \text{if } i = j, \\ 0 & \text{o.w.} \end{cases}$$

To show that A is symmetric, it suffices to show that R_1 is symmetric. Consider column t of R_1. If there is an entry in row t', column t, then that implies there exists $\ell \in [k]$ such that

$$\pi_{t'}^{\ell'} = \pi_t^{(\ell+1, \ell+2)(\ell')} \quad \text{for all } \ell' \in [k].$$

In particular, the entry is $-g_{t\ell}^{ds}$. However, note that we could have swapped t and t' in the above equality as $(\ell + 1, \ell + 2)$ is its own inverse. This implies there is a nonzero entry at row t column t' of R_1. Moreover, by Proposition J.8, the entry is $-g_{t'\ell}^{ds} = g_{t\ell}^{ds}$. Hence, R_1 is symmetric. To see that each row and column of R_1 sums to 0, it suffices to check $\mathbf{1}_\tau^\top r^t$. It's easy to see that it equals 0 because $\alpha_t^{ds} = \sum_{\ell=1}^k g_{t\ell}^{ds}$. Thus, $-R_1$ is a *bona fide* Laplacian matrix for a weighted undirected graph with no self loops.

Now, we show that G is isomorphic to G_C. First, it's easy to see that $|V| = |V_C| = k!$ as there are $k!$ ways to order the k subsets of LFs in the proto-pattern $\tilde{\pi}$. We'll now define the mapping between V and V_C. For convenience, abuse notation and define

$$(\ell + 1, \ell + 2) \circ \pi_t = \pi_{t'} \quad \text{where} \quad \pi_t^{\ell'} = \pi_{t'}^{(\ell+1, \ell+2)(\ell')} \quad \text{for all } \ell' \in [k].$$

Now, fix now an enumeration of the patterns and map the first pattern π_1 to e, the identity permutation. For $t \in [k!] \setminus \{1\}$, π_t will be associated with the permutation $\sigma \in S_k$ such that

$$\pi_t^\ell = \pi_t^{\sigma(\ell)} \quad \text{for all } \ell \in [k].$$

To finish our proof, it suffices to check that the edges are the same after applying the vertex transformation/renaming.

To do this, we analyze how the edges of G, G_C were constructed. For G_C, consider some arbitrary node associated with permutation σ. Since S has k elements, it has k edges where σ is connected to the permutations $(\ell + 1, \ell + 2) \circ \sigma$ (for $(\ell + 1, \ell + 2) \in S$). Thus,

$$E_C = \{\{\sigma, (\ell + 1, \ell + 2) \circ \sigma\} \colon \sigma \in S_k, (\ell + 1, \ell + 2) \in S,\ \ell \in [k]\}$$

Now, let π_t be the pattern associated with σ. Recall how we reassigned the k masses $g^{ds}_{t1}, \ldots, g^{ds}_{tk}$ in Lemma J.10. For each class ℓ, we connected π_t to $\pi_{t'} = (\ell+1, \ell+2) \circ \pi_t$ by reassigning mass $g^{ds}_{t\ell}$ to this second pattern. This can be seen in the Laplacian $-R_1$ as the entry in row t' column t represents an edge between $\pi_t, \pi_{t'}$. (Recall that our choice of g^{ds} has all positive entries, see Equation 19.) Thus,

$$E = \{\{\pi_t, (\ell+1, \ell+2) \circ \pi_t\} : t \in [k!], (\ell+1, \ell+2) \in S,\ \ell \in [k]\}.$$

To go from E to E_C, replace π_t with the permutation σ associated with it. I.e. determine the $\sigma \in S_k$ where

$$\pi_t^\ell = \pi_t^{\sigma(\ell)} \quad \text{for all } \ell \in [k].$$

Similarly, one can go from σ to π_t by finding the pattern π_t where the above equation holds. Thus, G (without its weights) and G_C are isomorphic.

To conclude our proof, observe that we have met the conditions for Proposition J.6, so that R_1 has affinely independent columns as claimed.

J.5 PROOF STEP: MASS REDISTRIBUTION FOR PATTERNS WITH MODERATE SPREAD

In the previous section, our argument relied on the fact that we could exhibit a new pattern by permuting the sets of LFs that predicted on two classes. If a pattern is such that $\pi_t^1 = \pi_t^2 = \emptyset$, then applying the permutation (12) will not give us a new pattern. Thus, we cannot directly apply the argument from the previous section.

Whereas we had assumed the spread was k in the previous section, our arguments there will work when the spread is $k-1$ too. I.e. there's only one class without LFs predicting it, so the above case cannot happen. In other words, the one may distinguish the classes by looking at the sets of LFs (that predict them). So, in this section, we'll consider the case where the spread i is such that $2 \leq i \leq k-2$.

Corollary J.12. *If a proto-pattern $\tilde{\pi}$ has spread $k-1$, then one can use the same construction of perturbations as in Lemma J.10 to get $k!$ linearly independent pattern distributions in T. These pattern distributions will have 0 mass on patterns spawned by $\tilde{\pi}$.*

To handle the case where the spread is between 2 and $k-2$, we will need to build upon the arguments of the previous section. In particular, we will partition the patterns spawned by proto-pattern $\tilde{\pi}$ (which has spread between 2 and $k-2$). We will need more notation to define this partition.

Definition J.13. Suppose a pattern π_t has spread i. The set of all classes which have LF predictions is define as

$$I_t \coloneqq \{\ell \in [k] \colon \pi_t^\ell \neq \emptyset\}.$$

We'll call elements of that set $\ell_1, \ldots, \ell_i$. A *run* (of indices) is a set of classes with contiguous indices where no LFs predict any of those classes. Formally, the set of runs is

$$J_t \coloneqq \{\{\ell_{i'}, \ell_{i'}+1, \ldots, \ell_{i'+1}\} \colon \ell_{i'}, \ell_{i'+1} \in I_t, \ell_{i'+1} - \ell_{i'} > 1\}.$$

When the spread is $k-1$, J has but one element, the set containing class with no LF predictions. Consider the following pattern where $p=3$ and $k=8$.

$$(\emptyset, \emptyset, \{1\}, \emptyset, \{2,3\}, \emptyset, \emptyset, \emptyset).$$

Here, $I = \{3,5\}$ and $J = \{\{1,2\}, \{4\}, \{6,7,8\}\}$ for a total of 3 runs with lengths $2, 1, 3$.

Indeed, it is easy to infer the maximum number of runs possible.

Proposition J.14. *For a pattern with spread $i < k$, the maximum number of runs is*

$$|J| \leq \min\{i+1, k-i\}.$$

Proof. Observe that since runs are separated by indices in I (where $|I| = i$ by definition), one can have a maximum of $i+1$ runs. However, we know that there are only $k-i$ classes without predictions, so the maximum number of runs is as claimed. □

We'll handle the proto-patterns with spread 2 first as that requires a completely different argument. Then, we consider proto-patterns with spread > 2, whereupon we build upon the arguments of the previous section.

J.5.1 Patterns With Spread 2

Like before, we will construct a graph that represents how to distribute the mass from every pattern spawned by a proto-pattern $\tilde{\pi}$ with spread 2. Then, the columns of the Laplacian will form our perturbation vectors (after appending a bunch of 0's).

The graph will be weighted, undirected, and have $k!/(k-2)!$ nodes and each node will have $k-1$ edges. We will have two tasks. For pattern π_t spawned by $\tilde{\pi}$, we have to redistribute the masses $g^{ds}_{t\ell_1}, g^{ds}_{t\ell_2}$ corresponding to the classes that have LF predictions, and the masses for all other classes, which have no LF predictions. Then, we must ensure that this graph is fully connected. When those are done, we can apply Proposition J.6 to conclude that the columns of the Laplacian are affinely independent.

We start with redistributing the masses on classes that have no predictions. Fix now a pattern π_t with spread 2. Observe if we apply the cycle $(\ell_1\ell_2)$ to π_t, we will get a pattern $\pi_{t'}$ where

$$I_t = I_{t'} \quad \text{implying} \quad J_t = J_{t'}.$$

Thus, we can assign the masses $g^{ds}_{t\ell}$, $\ell \in [k] \setminus I_t$ to $\pi_{t'}$ without changing the LF accuracies or class frequencies. I.e. since Proposition J.8 applies. Now, we redistribute the masses $g^{ds}_{t\ell_1}$ and $g^{ds}_{t\ell_2}$. Without loss of generality, consider $g^{ds}_{t\ell_1}$. By assumption, there are $k-2$ classes without a prediction, i.e. the indices in J_t. Thus, applying the permutation $(\ell_2\ell')$ where $\ell' \in [k] \setminus I_t$ gives us patterns where we can assign $g^{ds}_{t\ell_1}$. (Once again, Proposition J.8 is satisfied.) Our strategy will be to place an aliquot portion of $g^{ds}_{t\ell_1}$ in each of these patterns.

Before formalizing this strategy, observe that we only need to consider $k \geq 4$. That is, if $k = 3$, then proto-patterns with spread $2 = k-1$ which we already know how to handle via Corollary J.12.

Lemma J.15. *Suppose $p \geq 2$, $k \geq 4$ and fix a proto-pattern with spread 2. Define a weighted undirected graph $G = (V, E, \omega)$ with no self-loops as follows. Let $|V| = k!/(k-2)!$, where there is a node for every pattern spawned by $\tilde{\pi}$. Now consider the node associated with π_t (spawned by $\tilde{\pi}$). To redistribute the masses on indices in $[k] \setminus I_t$, i.e. the classes with no LF predictions, connect π_t to the pattern $\pi_{t'}$ where*

$$\pi^{\ell}_{t'} = \pi^{(\ell_1\ell_2)(\ell)}_t \quad \textit{for all } \ell \in [k].$$

For the edge connecting those two patterns, assign weight

$$\omega_{tt'} = \sum_{\substack{\ell=1 \\ \ell \notin I_t}}^{k} g^{ds}_{t\ell}.$$

Consider the following $k-2$ patterns for which we assign $g^{ds}_{t\ell_1}$ in aliquot parts.

$$N_t = \{\pi_{t'} \textit{ spawned by } \tilde{\pi} \colon \pi^{\ell}_{t'} = \pi^{(\ell_2\ell')(\ell)}_t \textit{ for } \ell \in [k], \ell' \in [k] \setminus I_t\}.$$

For every $\pi_{t'} \in N_t$, connect it to π_t and assign weight

$$\omega_{tt'} = \frac{g^{ds}_{t\ell_1}}{k-2}.$$

The same can be done with $g^{ds}_{t\ell_2}$, swapping ℓ_2 for ℓ_1.

Now, the following is true.

1. *For every pattern π_t spawned by $\tilde{\pi}$, π_t's masses $g^{ds}_{t1}, \ldots, g^{ds}_{t\ell}$ are only distributed to patterns spawned by $\tilde{\pi}$.*
2. *G has one connected component.*
3. *The columns of G's Laplacian matrix form perturbation vectors r for α where*

$$\alpha + (r; \boldsymbol{0}_{\tau - k!/(k-2)!}) \in T.$$

In particular, the t^{th} element of that vector is 0 if r is the t^{th} column of G's Laplacian (for $t \in [k!/(k-2)!]$). Here, we have enumerated the $k!/(k-2)!$ patterns spawned by $\tilde{\pi}$ as $\pi_1, \ldots \pi_{k!/(k-2)!}$.

Thus, the Laplacian of G has affinely independent columns.

Proof. Observe that the first claim is true by construction.

Now, we show that the graph has one connected component. Suppose we are at the node associated with π_t where $\ell_1, \ell_2 \in I_t$. We'll show how to get to the node associated with an arbitrary (but different) pattern $\pi_{t'}$ by traversing the graph. Say $\ell_1', \ell_2' \in I_{t'}$. First, suppose without loss of generality that $\ell_1 = \ell_1'$, but $\ell_2 \neq \ell_2'$. By construction, N_t contains the pattern $\pi_{t^\star}$ which is such that

$$\pi_{t^\star}^{\ell} = \pi_t^{(\ell_2' \ell_2)(\ell)} \quad \text{for all } \ell \in [k]$$

as we assumed $\ell_2 \neq \ell_2'$. However, this means that

$$\pi_{t^\star}^{\ell_1} = \pi_t^{\ell_1}, \quad \pi_{t^\star}^{\ell_2'} = \pi_t^{\ell_2} \quad \text{and} \quad \pi_{t^\star}^{\ell_2} = \emptyset,$$

wherein we conclude that $\pi_{t^\star} = \pi_{t'}$. In words, we the LFs predicting class ℓ_2 on π_t, i.e. $\pi_t^{\ell_2}$, predict class ℓ_2' on $\pi_{t^\star}$.

Now, consider the case where both $\ell_1 \neq \ell_1'$ and $\ell_2 \neq \ell_2'$. Here, we just apply the previous argument at most twice. That is, we first move $\pi_t^{\ell_2}$ to class ℓ_2'. If $\ell_2' = \ell_1$ and $\ell_1' = \ell_2$, we are done. Otherwise, apply the first argument again, moving $\pi_t^{\ell_1}$ (which might be in class ℓ_2) to ℓ_1'. Thus, we have shown that G is fully connected.

To see the third claim, observe that Proposition J.8 is satisfied for every pattern we assign mass to. Furthermore, observe that we reassign $g_{t1}^{ds}, \ldots, g_{tk}^{ds}$ for each pattern, meaning we have reassigned all the mass associated with one patten.

As before, the final claim follows from Proposition J.6. □

Observe that this strategy for constructing a graph easily generalizes for the case where the spread is k. For a pattern π_t, we may distribute $g_{t\ell}^{ds}$ to any pattern $\pi_{t'}$ where $\pi_t^{\ell} = \pi_{t'}^{\ell}$. If the spread is i, then there are $i-1$ subsets of LFs whose positions we may permute in $k-1$ slots (as we're fixing the position of π_t^{ℓ} in class ℓ). Moreover, one of these permutation is already present in π_t. Thus, there are $(k-1)!/(k-i)! - 1$ *other* patterns which satisfy the assumptions of Proposition J.8.

If the spread is 2, then it is clear that there are $k-2$ positions where we can move $\pi_t^{\ell_2}$. Plugging $i = 2$ into the above formula gives the same result. Thus, the number of outgoing edges from π_t (which has spread 2) is $O(k)$. However, when the spread is k, there are $(k-1)! - 1$ other patterns we could distribute $g_{t\ell}^{ds}$ to. So, if the mass is distributed in equal parts, the number of edges from the node associated with π_t is $O(k!)$.

J.5.2 Patterns With Spread > 2

Like in the case where the spread was 2, we have two main tasks in terms of redistributing the probability mass. I.e. redistributing masses from classes which have LF predictions and masses from classes which have no LF predictions. To redistribute the mass on classes which have LF predictions, we appeal to the argument made above for when the spread was k. That argument will work because we have spread of at least 3, meaning we can apply Lemma J.10 on the classes with LF predictions. For the masses on classes with no predictions, we generalize the strategy given in the previous section. There, we paired up patterns which had LF predictions on the same classes. Here, we'll also pair up patterns, but we'll do it in a more subtle way as by assumption, there are at least 3 classes with LF predictions.

Suppose now that our fixed proto-pattern $\tilde{\pi}$ has spread $i \geq 3$ (and $i \leq k-2$ by assumption). (Note that we can assume $k \geq 5$ because for, $k = 3$ the previous arguments work on proto-patterns with spread $2, 3$ while for $k = 4$, the previous arguments work for spread $2, 3, 4$.) Observe that we may spawn $\binom{k}{i} i!$ patterns from this proto-pattern. $\binom{k}{i}$ ways to pick the indices which have LF predictions (i.e. that many I_t's), $i!$ ways to order the i subsets of LFs on those indices. So, partition the patterns based on which classes have LF predictions, for a total of $\binom{k}{i}$ subsets.

Our strategy is as follows. For each of these $\binom{k}{i}$ subsets, we have $i!$ patterns which we would like to construct a weighted undirected graph (with no self-loops) wherein the edge weights show how to redistribute the probability mass for each pattern (i.e. node). Thus, for the proto-pattern $\tilde{\pi}$ with spread $3 \leq i \leq k-2$, we'll create $\binom{k}{i}$ connected graphs, one for each choice of i classes.

So, consider one of the $\binom{k}{i}$ ways for the LFs to only predict on i classes. Since we assumed the spread $i \geq 3$, we can apply the argument of Lemma J.10. The resulting graph over $i!$ elements is fully connected and the masses of classes which have LF predictions are redistributed. What remains is to distribute the masses on the classes which have no LF predictions. Here,

we adopt a similar strategy wherein we permute the classes which have LF predictions. Recall that we need a notion of reciprocity when redistributing masses. If we distribute mass $g^{ds}_{t\ell}$ to pattern $\pi_{t'}$, then we must distribute the same mass from $\pi_{t'}$ to π_t. When the spread was 2, this was simple to do for the masses on empty classes: we apply the permutation $(\ell_1 \ell_2)$. That permutation is its own inverse, so it's easy to go back and forth from π_t and $\pi_{t'}$. In this case, we have $i \geq 3$ classes with LF predictions, so we need a more subtle argument.

For the $i!$ permutations in S_i, observe that we can assign a parity, even, or odd, to a permutation $\sigma \in S_i$ by seeing whether an even or odd number of swaps is needed to transform the identity permutation to σ. This leads to a natural partition of S_i into two sets of the same size, call it S^e_i, S^o_i for the even and odd permutations respectively (e.g. page 111 of [D'Angelo and West, 2000]). Moreover, since permutations are bijective functions, the permutation (12) is a bijective map between S^e_i and S^o_i. Indeed, since we have fixed the i classes on which the LFs predict, i.e. $\ell_1, \ldots, \ell_i$, these permutations permute the i subsets of classes on $\ell_1, \ldots, \ell_i$. So, we can associate the $i!$ patterns with permutations in S_i a la the proof of Lemma J.11 in Section J.4.3. Thus, by using S^e_i, S^o_i, we have a pairing of patterns wherein we can distribute the mass on classes without LF predictions.

Lemma J.16. *Suppose $p \geq 3$, $k \geq 5$ and fix a proto-pattern $\tilde{\pi}$ with spread $3 \leq i \leq k-2$. Also, fix $\ell_1, \ldots, \ell_i \in [k]$, the set of classes which will have LF predictions. Define a weighted undirected graph $G = (V, E, \omega)$ with no self loops as follows. Let $|V| = i!$, where there's a node for every pattern spawned by $\tilde{\pi}$ where the LF predictions are on classes $\ell_1, \ldots, \ell_i$. Consider the node associated with π_t, an arbitrary pattern spawned by $\tilde{\pi}$ where the LFs predict on classes $\ell_1, \ldots, \ell_i$. Construct i edges and assign weights according to the strategy of Lemma J.10. Here, we apply that argument to the indices $\ell_1, \ldots, \ell_i$. Now, consider the pattern $\pi_{t'}$ (also spawned by $\tilde{\pi}$) where*

$$\pi^{\ell}_{t'} = \pi^{(\ell_1 \ell_2)(\ell)}_t \quad \textit{for all } \ell \in [k].$$

Create an edge between π_t and $\pi_{t'}$ and assign weight

$$\sum_{\substack{\ell=1 \\ \ell \neq \ell_1, \ldots, \ell_i}}^{k} g^{ds}_{t\ell}.$$

The following is true.

1. *We only redistribute mass to patterns spawned by $\tilde{\pi}$ and which have LF predictions on $\ell_1, \ldots, \ell_i$.*
2. *G has one connected component.*
3. *The columns r of G's Laplacian form perturbation vectors r for α where $\alpha + (r; \mathbf{0}_{\tau - i!}) \in T$. In particular, the t^{th} element of that vector is 0 if r is the t^{th} column of G's Laplacian for $t \in [i!]$. Here, we have enumerated the $i!$ patterns spawned by $\tilde{\pi}$ (which have LF predictions on $\ell_1, \ldots, \ell_i$) as $\pi_1, \ldots, \pi_{i!}$.*

Thus, the Laplacian of G has affinely independent columns.

Proof. The first claim should be clear by construction. In applying the argument of Lemma J.10, we only redistribute masses to patterns where there are LF predictions on $\ell_1, \ldots, \ell_i$. Moreover, our construction of $\pi_{t'}$ does not move any LF prediction off of those i indices, meaning the first claim is true.

The second claim is true because the construction in Lemma J.10 guarantees that the graph has one connected component. Moreover, you cannot increase the number of connected components by adding edges to a graph, which is what we do. By our discussion on even and odd permutations, we are justified in adding extra edges.

The third claim is true because we have assigned mass in a manner consistent with Proposition J.8. In particular, we have redistributed all the mass associated with π_t (spawned by $\tilde{\pi}$ and having LF predictions on $\ell_1, \ldots, \ell_i$).

The final claim follows from J.6. □

To summarize, we have now argued how to redistribute the masses from patterns with spread 2 and greater in such a way that the perturbation vectors maintain block diagonal form (see Equation 32). What remains now is to show how to redistribute the mass from a pattern with spread 1.

J.6 PROOF STEP: MASS REDISTRIBUTION FOR PATTERNS WITH MINIMAL SPREAD

Unfortunately, the arguments of the previous section will not, in general, work for the case where the spread is 1. Here, all LFs predict the same class so that there is one proto-pattern in V_1: $\{[p]\}$. It's easy to see that this can spawn k patterns, call them $\pi_1, \ldots, \pi_k$, where $\pi_\ell^\ell = [p]$ and all other $\pi_\ell^{\ell'} = \emptyset$. Observe that the crux of Proposition J.8 was that to reassign the mass of class ℓ for π_t, we need to find another pattern $\pi_{t'}$ where $\pi_t^\ell = \pi_{t'}^\ell$. The LFs that predict class ℓ must match because in general, the class frequency distribution is not uniform. This strategy doesn't work here because there is only one pattern where all classes predict class ℓ! Hence, we cannot distribute g_{11}^{ds} to π_{22} without reducing the class frequency w_1^* by g_{11}^{ds} and increasing the class frequency w_2^* by the same amount. Thus, we must seek an alternate strategy.

Recall that we need to construct a perturbation $r \in [-1, 1]^{\tau k}$ for g^{ds} such that $A(g^{ds} + r) = (b^*; w^*)$. Our perturbation must be such that (without loss of generality) the masses $g_{11}^{ds}, \ldots, g_{1k}^{ds}$ are redistributed to patterns with index greater than k. This is a consequence of our above discussion (and to satisfy Proposition J.7). So, we now discuss how to redistribute the masses of π_1. Of course, this strategy will work for each of the k patterns spawned by the singular proto-pattern with spread 1.

By assumption, $\pi_1^\ell = \emptyset$ for $\ell \geq 2$. We can redistribute masses $g_{t2}^{ds}, \ldots, g_{tk}^{ds}$ to patterns (with spread > 1) which don't have LF predictions on classes $2, \ldots, k$ respectively. I.e. one may just choose $k - 1$ of them. Now, we must redistribute g_{11}^{ds} without changing the LF accuracies or the class frequency w_1^*. We start with the latter.

Consider p patterns $\pi_{t_1}, \ldots, \pi_{t_p}$ where

$$\pi_{t_j}^1 = [p] \setminus \{j\}.$$

If we distribute g_{11}^{ds} to these patterns (on class 1) specifically, notice that we will meet the requirement of not changing the class frequency w_1^*. While we don't need to choose those patterns in specific, that choice is conducive to our proof. While w_1^* remains unchanged, the LF accuracies will have decreased as a result of this.

Proposition J.17. *Suppose we need to redistribute mass g_{11}^{ds}. Split it into p parts so that*

$$g_{11}^{ds} = \sum_{j=1}^{p} m_j \quad \textit{where} \quad m_j \geq 0.$$

If we distribute m_j mass to $g_{t_j 1}^{ds}$, then the first class' frequency won't change. However, the LF accuracies will change. Specifically, the new LF accuracies will be

$$b_j^* - m_j$$

for each $j \in [p]$.

Proof. The first claim should be immediate because the total amount of mass put on class 1 is g_{11}^{ds}, which was the same amount it was decreased by.

For the second claim, fix a LF $h^{(j)}$. Its accuracy is decreased by g_{11}^{ds} because we are redistributing that mass. By definition,

$$j \in \pi_{t_{j'}}^1 \quad \text{for } j' \in [p].$$

This means that the accuracy of $h^{(j)}$ is increased by

$$\sum_{\substack{j'=1 \\ j' \neq j}}^{p} m_{j'} = g_{11}^{ds} - m_j.$$

In other words, the accuracy of $h^{(j)}$ is decreased by g_{11}^{ds} and then increased by $g_{11}^{ds} - m_j$. Thus, after the mass redistribution, $h^{(j)}$'s accuracy is $b_j^* - m_j$. □

This means that if we are able to somehow increase the accuracy of LF $h^{(j)}$ by $m_j \geq 0$ where $\sum_{j=1}^{p} m_j = g_{11}^{ds}$, we will be able to redistribute g_{11}^{ds} *and* maintain the class frequencies and LF accuracies.

Consider a pattern π_t where

$$\pi_t^1 = [p] \setminus \{1, 2\}.$$

If we redistributed g^{ds}_{t1} to $g^{ds}_{t_1,1}$, then $h^{(2)}$'s accuracy increases by g^{ds}_{t1} while no other LF accuracy or class frequency changes. Our strategy is to apply this multiple times. Formally, the strategy can be characterized as follows.

Proposition J.18. *Suppose we have two patterns $\pi_t, \pi_{t'}$ and a class ℓ where*

$$\pi^{\ell}_{t'} \subset \pi^{\ell}_{t}.$$

Then, if we reassign $g^{ds}_{t'\ell}$ to $g^{ds}_{t\ell}$, the only change to the LF accuracies and class frequencies will be an increase of $g^{ds}_{t'\ell}$ to the accuracy of each LF in $\pi^{\ell}_t \setminus \pi^{\ell}_{t'}$.

Proof. Since we are assigning $g^{ds}_{t'\ell}$ to $g^{ds}_{t\ell}$, there is no change to the class frequencies. Similarly, for the LFs in $\pi^{\ell}_t \cap \pi^{\ell}_{t'}$, there is no net change. For the LFs in $\pi^{\ell}_t \setminus \pi^{\ell}_{t'}$, $g^{ds}_{t'\ell}$ do not contribute to its accuracy. However, $g^{ds}_{t\ell}$ *does* contribute to their accuracies. Hence, their accuracies increases by $g^{ds}_{t'\ell}$. □

The main idea of our proof is to use the previous proposition to increase the accuracies of the LFs by a total of g^{ds}_{11}. Then, applying the strategy laid out in Proposition J.17 is sufficient.

For simplicity, we will apply Proposition J.18 to patterns where $p-2$ and $p-1$ LFs predict a single class. Of course, we will require that $p \geq 3$. By repeatedly applying the above proposition, we are able to, under mild conditions, scrounge up enough probability mass to use Proposition J.17 and without changing the LF accuracies. In particular, the increase of accuracies must equal g^{ds}_{11}. To see this is possible, observe that there are $k(k-1)^2$ patterns where LFs $[p] \setminus \{1,2\}$ predict one class and the other LFs predict one of the remaining $k-1$ classes. In particular, we are not limited to increasing the accuracy of the first and second LFs: we may choose any subset of $p-2$ LFs and apply this strategy. We now provide a result showing when this is possible.

Lemma J.19. *Suppose $p \geq 3$ and one of the following conditions hold.*

1. $$\sum_{j=1}^{p} \left(\frac{1-b^*_j}{b^*_j} \right)^2 + \max_{\ell \in [k]} w^*_\ell \leq \left[\sum_{j=1}^{p} \frac{1-b^*_j}{b^*_j} \right]^2.$$
2. $$\max_{j \in [p]} b^*_j \leq \left(1 + \sqrt{\frac{\max_{\ell \in k} w^*_\ell}{p(p-1)}} \right)^{-1}$$

In particular, we may replace the maximum with 1. Fix an arbitrary pattern π_ℓ where all LFs predict class ℓ. Then, there exists a perturbation vector $r \in [-1,1]^\tau$ such that

1. $\alpha + r \in T$
2. $[\alpha + r]_\ell = 0$
3. $r_{\ell'} = 0$ *for all* $\ell' \in [k] \setminus \ell$.

Proof. Here, we will not explicitly construct r, but we'll show that it is possible. To achieve the claims, we must show the existence of a perturbation r to g^{ds} such that $A(g^{ds} + r) = (b^*; w^*)$.

To satisfy the third claim, we will not redistribute the masses of π_ℓ to $\pi_{\ell'}$ for $\ell' \in [k] \setminus \{\ell\}$. In other words, we do not distribute any of π_ℓ's masses to the other proto-patterns generated by $\tilde{\pi} = \{[p]\}$ (in contrast to the previous sections).

The second claim will be done by construction – we'll redistribute $g^{ds}_{\ell 1}, \ldots, g^{ds}_{\ell k}$.

Now, we move on to the first claim. Pick $k-1$ patterns $\pi_{t_1}, \ldots, \pi_{t_{\ell-1}}, \pi_{t_{\ell+1}} \ldots, \pi_{t_k}$ where

$$\pi_{t_{\ell'}} = \emptyset \quad \text{for} \quad \ell' \in [k] \setminus \{\ell\}$$

and $t_1, \ldots, t_k > k$. We'll redistribute $g^{ds}_{\ell\ell'}$ to $\pi_{t_{\ell'}}$, specifically by increasing $g^{ds}_{t_{\ell'}\ell'}$ by $g^{ds}_{\ell\ell'}$. By Proposition J.8, this does not change the LF accuracies or class frequencies. Thus, all that remains is to redistribute $g^{ds}_{\ell\ell}$ without changing the LF accuracies and class frequencies.

As we have discussed, we will repeatedly apply Proposition J.18 so that the accuracy of LF $h^{(j)}$ is increased by m_j wherein

$$\sum_{j=1}^{p} m_j = g_{\ell\ell}^{ds}.$$

Then, we apply the strategy of Proposition J.17 to get a perturbation of α which lies in T.

In words, we want to show that we can increase the accuracy of the LFs by a total of $g_{\ell\ell}^{ds}$. We'll use the aforementioned strategy of redistributing masses from patterns that have $p-2$ LFs predicting on a class to patterns which have $p-1$ LFs predicting the same class. Let $\pi_{t_1}, \ldots, \pi_{t_p}$ be the p patterns defined in Proposition J.17. Here, we're abusing notation and redefining what we mean by these patterns. Specifically, $\pi_{t_j}^{\ell} = [p] \setminus \{j\}$ and $h^{(j)}$ predicts a fixed but arbitrary class in $[k] \setminus \{\ell\}$.

Consider a pattern π_t where LFs $[p] \setminus \{1,2\}$ predict class ℓ. We will reassign $g_{t\ell}^{ds}$ to π_{t_1} to increase $h^{(2)}$'s accuracy by $g_{t\ell}^{ds}$ (i.e. see Proposition J.18). (By Proposition J.8, we could have also assigned this mass to π_{t_2} to increase $h^{(1)}$'s accuracy by the same amount. However, this does not matter because we care about the net accuracy increase over all LFs.) Now, observe that there are $(k-1)^2$ many patterns where $[p] \setminus \{1,2\}$ predict class ℓ. I.e. there are $k-1$ choices of classes to predict for the 2 remaining LFs. By applying this strategy on all those patterns, the total accuracy increase for $h^{(2)}$ is

$$(k-1)^2 w_\ell^* \prod_{j=3}^{p} b_j^* \frac{1-b_1^*}{(k-1)} \frac{1-b_2^*}{(k-1)} = w_\ell^* \prod_{j=1}^{p} b_j^* \frac{1-b_1^*}{b_1^*} \frac{1-b_2^*}{b_2^*}.$$

In particular, we multiplied by 1 twice so the expression is in a more convenient form. Recall that by Proposition J.8, the mass we are reassigning is the same across all $(k-1)^2$ of these patterns because the set of LFs that predict ℓ is the same across all patterns. Now since we could have had $[p] \setminus \{1,2\}$ predict any of the ℓ classes, using all those patterns means the total increase to $h^{(2)}$'s accuracy can be

$$\sum_{\ell=1}^{k} w_\ell^* \prod_{j=1}^{p} b_j^* \frac{1-b_1^*}{b_1^*} \frac{1-b_2^*}{b_2^*} = \prod_{j=1}^{p} b_j^* \frac{1-b_1^*}{b_1^*} \frac{1-b_2^*}{b_2^*}$$

since $w^* \in \Delta_k$.

Observe now that are $\binom{p}{p-2}$ choices of two LFs wherein $\{1,2\}$ was but one. Thus, we can sum the above expression over all choices of two LFs. By doing this, we will increase the accuracies of other LFs, not just $h^{(2)}$. This means that the total increase in LF accuracy will be

$$\prod_{j=1}^{p} b_j^* \left[\sum_{\substack{j,j'=1 \\ j \neq j'}}^{p} \frac{1-b_j^*}{b_j^*} \frac{1-b_{j'}^*}{b_{j'}^*} \right]. \tag{33}$$

To apply Proposition J.17, we require that the above term is no smaller than $g_{\ell\ell}^{ds}$ for each $\ell \in [k]$. If we have such a guarantee, then for π_ℓ, the total accuracy increase we can provide for the p LFs is $g_{\ell\ell}^{ds}$ for each $\ell \in [k]$. In particular, we note that there is no need to redistribute all the mass of the $\binom{p}{p-2}(k-1)^2$ patterns we are considering.

Now, recall that π_ℓ, $\ell \in [k]$ are patterns where all LFs predict class ℓ so that

$$g_{\ell\ell}^{ds} = w_\ell^* \prod_{j=1}^{p} b_j^*.$$

Thus, our above requirement is that we require the expression in Equation 33 to be larger than the above.

$$\prod_{j=1}^{p} b_j^* \left[\sum_{\substack{j,j'=1 \\ j \neq j'}}^{p} \frac{1-b_j^*}{b_j^*} \frac{1-b_{j'}^*}{b_{j'}^*} \right] \geq \max_{\ell \in [k]} w_\ell^* \prod_{j=1}^{p} b_j^*.$$

Equivalently, we just need

$$\left[\sum_{\substack{j,j'=1 \\ j \neq j'}}^{p} \frac{1-b_j^*}{b_j^*} \frac{1-b_{j'}^*}{b_{j'}^*} \right] \geq \max_{\ell \in [k]} w_\ell^*.$$

Observe that the left-hand side can be written as

$$\left[\sum_{j=1}^{p} \frac{1-b_j^*}{b_j^*}\right]^2 - \sum_{j=1}^{p} \left(\frac{1-b_j^*}{b_j^*}\right)^2$$

and that

$$\frac{1-b_j^*}{b_j^*} = \frac{1}{b_j^*} - 1.$$

So, an upper bound on b_j^* gives a lower bound on $\frac{1}{b_j^*}$. Let

$$b = \max_{j \in [p]} b_j^*.$$

Then, the LHS can be bounded as follows:

$$\left[\sum_{\substack{j,j'=1 \\ j \neq j'}}^{p} \frac{1-b_j^*}{b_j^*} \frac{1-b_{j'}^*}{b_{j'}^*}\right] \geq p(p-1)\left(\frac{1}{b} - 1\right)^2.$$

We require that lower bound to be larger or equal to $\max_{\ell \in [k]} w_\ell^*$. One then rearranges.

To complete the proof, note that we can weaken the above conditions by replacing w_ℓ^* with 1 as there's no restriction on how close w_ℓ^* can be to 1. □

With this result, we have shown how to construct perturbation vectors for all τ patterns wherein the perturbed distributions all lie in T and are linearly independent! We now provide the formal proof of Theorem J.3.

J.7 PROOF OF THEOREM J.3

Given our assumptions, we exhibit τ linearly independent pattern distributions in T. Then, by application of Proposition J.5, the convex hull of those τ vectors will have dimension $\tau - 1$. Since $T \subset \Delta_\tau$, $dim(T) \leq dim(\Delta_\tau) = \tau - 1$ so that the dimension of T is lower and upper bounded by $\tau - 1$.

We now demonstrate the existence of those τ pattern distributions. By Lemma C.2 there exists an $\alpha \in T$ with all strictly positive masses. We will construct τ perturbation vectors $r^1 \ldots, r^\tau$ such that for all $t \in [\tau]$,

$$\alpha + r^t \in T.$$

To show that the τ pattern distributions should be linearly independent, it suffices to show that the perturbations r^t are *affinely independent* (Proposition J.4). In particular, the perturbation vectors will have the form exhibited in Equation 32. If that holds and the assumptions in Proposition J.7 are met, then by the τ perturbation vectors are affinely independent.

We will construct the perturbation vectors in parts. Specifically, Lemma J.10, Corollary J.12 show how to construct perturbation vectors for patterns with spread k and $k-1$. Lemma J.11 shows that those perturbation vectors meet our requirements. For patterns with spread 2, Lemma J.15 shows how to construct the perturbation vectors. For pattern vectors with spread between 3 and $k-2$ inclusive, Lemma J.16 shows how to construct satisfactory perturbation vectors. Finally, Lemma J.19 shows how to construct perturbation vectors for patterns with spread 1. As the perturbation vectors were constructed in such a way that the requirements of Proposition J.7 are met, the proof is complete.

K COMPLETE EXPERIMENTAL RESULTS

K.1 EMPIRICAL EVALUATION OF LEMMA 7.3

Here, we present the results of evaluating Lemma 7.3 on 11 real datasets from WRENCH [Zhang et al., 2021]. See Table 3. These datasets were taken from the Supplementary Material of [An, 2025]. The train/validation/test splits of these datasets have been merged and all datapoints without an LF prediction were removed. We note that these datasets contain specialists, i.e. have LFs that abstain, and the domains range from sentiment analysis to spam detection to video frame classification.

Table 3: Dataset Statistics For 11 Real Datasets

Dataset	IMDB	Youtube	SMS	CDR	Yelp	Commercial	Tennis	TREC	SemEval	ChemProt	AG News
Classes (k)	2	2	2	2	2	2	2	6	9	10	4
LFs (p)	5	10	73	33	8	4	4	68	164	26	9
Datapoints (n)	21,914	1,843	2,246	12,562	31,512	81,105	8,803	5,738	2,527	13,768	82,591

We compare three weighting strategies. First is OCDS with the true accuracies and true class frequencies. That is, for $j \in [p]$ and $\ell \in [k]$,

$$\theta_j^{ds} = \log\left(\frac{b_j^*(k-1)}{1-b_j^*}\right) \quad \text{and} \quad \theta_{p+\ell} = \log(w_\ell^*).$$

Note that $b_j^* = Pr(h^{(j)}(x) = y)$. Here an abstention counts as an incorrect prediction! The OCDS prediction is then $g^{(\theta^{ds}),\alpha}$. Second is MV where $\theta = \mathbf{1}_{p+k}$. Third is BF on the polytopes $P^{\epsilon}_{b^*w^*}$ where $\epsilon \in \{0.01, 0.05, 0.1, 0.2, 0.3\}$. This choice of polytopes was done to avoid the problems one gets when using sampled labeled data to estimate LF accuracies. (I.e. LFs with more predictions often get a better estimate of their accuracies.)

Three bounds are evaluated on the comparison of the BF weights versus the OCDS/MV weights. First is the exact difference of excess errors as seen in Theorem 7.2. Second is the aforementioned Lemma 7.3. Third is the bound proposed by An and Dasgupta [2024] in Lemma 12. For this bound, we only evaluate BF against OCDS, since that was how it was presented. Note that the first and third bounds are not computable in practice since we don't know the optimal weights θ^* and we don't know what b^*, w^* are!

We observed that the exact difference of excess errors always correctly predicted when a weighting strategy was better than another. This provides evidence that our result of Theorem 7.2 is correct. Second, we find that Lemma 12 of [An and Dasgupta, 2024] can be okay, despite the lower bound of Theorem 7.1. Finally, we show that Lemma 7.3 can be poor. Specifically, what we mean is that it falsely predicts that g' is worse than g for very small values of ϵ. That is, the result of the test is a false negative as g' is actually better than g. We record the first ϵ for which that happens for every dataset in Table 4. Entries greater than 0.01, i.e. when the bound of Lemma 7.3 is not trivial are **bolded**. We note that this value is quite small most of the time, meaning Lemma 7.3 as stated is probably not very useful as it is conservative. Hence, our suggestion that one attain lower bounds for how well b^*, w^* is estimated. I.e. for $Ag^{(\theta^{bf}),\alpha^*} = (b^{bf}, w^{bf})$, attain lower bounds for $b^* - b^{bf}$ and $w^* - w^{bf}$. For the bound from [An and Dasgupta, 2024], it does not depend on ϵ, so we show the upper bound on ϵ' under which the BF weights gotten using $P^{\epsilon'}_{b^*w^*}$ is guaranteed to better than θ^{ds}.

Table 4: Empirical Evaluation of Tests for Determining When A Weighting Strategy Is Better Than Another. Shown Is the First ϵ for Which A False Negative was Observed.

Dataset	IMDB	Youtube	SMS	CDR	Yelp	Commercial	Tennis	TREC	SemEval	ChemProt	AGNews
Ours (Lemma 7.3) BF vs OCDS	0.01	**0.05**	0.01	**0.05**	**0.05**	0.01	**0.2**	0.01	**0.05**	**>0.3**	0.01
Ours (Lemma 7.3) BF vs MV	0.01	**0.05**	0.01	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**	**0.1**	**0.1**
[An and Dasgupta, 2024] (Lem 12)	0.169	0.176	0.000123	0.12	0.042	0.114	0.125	0.0112	0.0021	0.064	0.132

K.2 DETERMINING THE SIZE OF $O^{\theta,\epsilon}, T^{\epsilon}$

We finish the appendix by providing the complete details for determining the size of $O^{\theta,\epsilon}$ and T^{ϵ}. Not only will we flesh out the details in their entirety, but also address certain issues that arise in using LattE [Baldoni et al., 2013].

To begin, we give some background on the types of objects that cddlib [Fukuda, 2021] and LattE [Baldoni et al., 2013] operate on. Consider a rational polytope defined by a system of linear inequalities. That is, say $A \in \mathbb{Q}^{m\times n}$ and a vector of coefficients $b \in \mathbb{Q}^m$. The polytope in question is defined as

$$P = \{x \in \mathbb{Q}^n \colon Ax \leq b\}.$$

The definition of P above is called the H-Representation, as it is defined as an intersection of half-spaces. We'll suppose that P can be placed into a finitely large hyper-rectangular prism. The P we consider will satisfy this assumption. An alternate representation of P is via its vertices $v_1, \ldots, v_{|V|} \in \mathbb{Q}^n$ where V is the set of all vertices. P can written as the convex combination of these vertices.

$$P = \left\{ \sum_{i=1}^{|V|} a_i v_i \colon a_i \geq 0, \forall i \in [V], \sum_{i=1}^{|V|} a_i = 1 \right\}.$$

cddlib [Fukuda, 2021] implements the Double Description Algorithm, which lets one move from the H-Representation to the V-Representation and vice versa.

If we wish to linearly transform P by a matrix $D \in \mathbb{Q}^{m'\times n}$, then one method is to recover the V-Representation, e.g. using cddlib, then applying D to the vertices. The transformed vertices will be of the transformed polytope DP. One can then acquire the H-Representation of DP via cddlib. This is what we use to construct T^ϵ, i.e. finding the V-Representation of $P^\epsilon_{b^* w^*}$, applying D to the vertices to acquire the vertices of T^ϵ.

Now, we move on to discussing the construction of $O^{\theta,\epsilon}$. Fix $\widehat{g} = \widehat{g}^{(\theta)}$ and recall the definition of $A^{\widehat{g}} \in [0,1]^{(p+k)\times\tau}$ from Equation 27. It is reprinted below.

$$A^{\widehat{g}} = a^{\widehat{g}}_{jt} = \begin{cases} \widehat{g}_{t\pi_{tj}} & \text{if } j \leq p \\ \widehat{g}_{t,j-p} & \text{if } j > p. \end{cases}$$

One may check that $A^{\widehat{g}}\alpha = AD^\top \alpha \circ \widehat{g}^{(\theta)}$. For elementwise inequality, we can write $O^{\theta,\epsilon}$ as

$$O^{\theta,\epsilon} = \{\alpha \in \Delta_\tau \colon -\epsilon \leq A^{\widehat{g}}\alpha - (b^*; w^*) \leq \epsilon\} \tag{34}$$

If the exponential of a weights θ a weighting strategy return is rational, then the definition we wrote above in Equation 34 suffices. However, in the case of majority vote, $\widehat{g}^{(\theta^{mv})} = \widehat{g}^{mv}$ is not rational. To get around this, we truncate elements of $\widehat{g}^{(\theta^{mv})}$ after d decimal places. That is, we approximate $\widehat{g}^{mv}_{t\ell}$ by

$$\frac{\left\lfloor 10^d \widehat{g}^{(\theta^{mv})}_{t\ell} \right\rfloor}{10^d}.$$

The maximum error incurred by this rationalization is $10^{-(d-1)}$. Now as $\widehat{g}^{mv}$ is involved in the construction of $A^{\widehat{g}}$, which in turn is multiplied with $\alpha \in \Delta_\tau$, the total error for this rationalization is $10^{-(d-1)}$. If we call $\widehat{g}^{mv\prime}$ the rationalized version of $\widehat{g}^{mv}$, we will use the following polytope in place of $O^{mv,\epsilon}$.

$$\{\alpha \in \Delta_\tau \colon -\epsilon - 10^{-d} \leq A^{\widehat{g}^{mv\prime}}\alpha - (b^*; w^*) \leq \epsilon + 10^{-d}\}$$

Penultimately, we address the issue of LattE [Baldoni et al., 2013] requiring the polytope P to contain an integral point. This requirement is present regardless of whether we want to determine $|\tilde{O}^{\theta,\epsilon}_n|$ or $vol(O^{\theta,\epsilon})$. In general, the polytopes we consider will not contain an integral point. Therefore, we scale the polytope by a factor of n. Incidentally, this is the number of datapoints. This has a natural interpretation in that we are moving from LF accuracies/class frequencies to counts of correct predictions and counts datapoints which have class ℓ. For the case that we count the number of patterns realizable over n datapoints, we also have to scale the set of possible pattern distributions. That is, we had defined it as

$$S_n = \{\alpha \in \Delta_\tau \colon n\alpha \in \mathbb{Q}^\tau_{\geq 0}\}.$$

Then, as we recall

$$\tilde{O}^{\theta,\epsilon}_n = O^{\theta,\epsilon} \cap S_n.$$

Define the scaled set of S_n, call it S^n as

$$S^n = \{\alpha \in \mathbb{Q}^\tau_{\geq 0} \colon \mathbf{1}_\tau^\top \alpha = n\}.$$

We now count the number of lattice points inside

$$\{\alpha \in S^n \colon \ -n(\epsilon + 10^{-(d-1)}) \leq A^{\widehat{g}^{mv'}}\alpha - n(b^*; w^*) \leq n(\epsilon + 10^{-(d-1)})\}.$$

There are two considerations that arise from this. First is the scaling of $vol(O^{\theta,n\epsilon})$ as a function of n. Since LattE computes the volume of a polytope in the dimension it lies in, e.g. it will return the length of a line that is in three dimensional space, if $d = dim(O^{\theta,\epsilon})$, then $vol(O^{\theta,n\epsilon}) = n^d vol(O^{\theta,\epsilon})$. In particular, if $d < \tau - 1$ ($\tau - 1$ is the maximum possible dimension), then $vol(O^{\theta,n\epsilon})$ will scale slower than $vol(T^{n\epsilon})$. One may observe this behavior if they scale n (as we have).

The second consideration is how the choice of n affects whether O^θ is empty when we are able to exactly represent $A^{\widehat{g}}$. We have abused notation and are referring to the polytope

$$\{\alpha \in S^n \colon A^{\widehat{g}}\alpha = n(b^*; w^*)\}.$$

If elements of the matrix $A^{\widehat{g}}$ are of the form $\frac{i}{5}$, then if $nb_j = i'/11$ for all j, then in general the constraints cannot be satisfied. Thus, when we pick n for the experiments and have $\epsilon = 0$, we ensure that the above polytope is not empty. This can be done by choosing n to be the least common multiple of all entries in $A^{\widehat{g}}$ and b^*, w^*

The final consideration is that LattE requires that the A, b defining a polytope P be integers rather than rational numbers. To meet this requirement, we just scale each inequality (which now only has rational numbers) until the inequality can be written only with integers.

K.2.1 Final Thoughts and Observations

We end by giving a few observations. First is that the polytope T can have a large number of vertices. For example, take $p = 3, k = 2, n = 1000$, $\epsilon = 0$, $w^* = (0.5, 0.5)$, $b^* = (0.8, 0.6, 0.5)$. This polytope has has 188 vertices despite the dimension of the problem being $k^p = 8$. Moreover, if one wants to know $|\tilde{T}_{1000}|$, the computation took > 18 hours, even after playing with the hyperparameters following [Köppe, 2007]. We have included the results of that run, along with the polytope description in the supplementary material. The primary challenge is the cone decomposition required for Barvinok type algorithms. Over 56 million cones were enumerated for that polytope.

When $p = 3, k = 3$, enumerating the vertices of O^θ can take a long time. We attempted a single instance of that, but stopped the code after it had run for 2 hours. For $p, k \geq 3$, it may be faster to reformulate the constraints and do block or Fourier elimination of variables.

In low dimensions, $p = 3, k = 2$, finding the volume is feasible. Below, we show the results of

$$Pr_n(g^{(\theta^{ds}),\alpha} \text{ is optimal} | b^*, w^*) \quad \text{and} \quad Pr_n(b^*; w^*)$$

for the same setup as Figure 2. We see that the values are quite similar and hypothesize that finding the volume (in the underlying dimension, not $\tau - 1$) is a good surrogate for estimating $Pr_n(\cdot)$. I.e. Finding the volume of T from above (where counting the number of patterns took a long time) only took a minute!

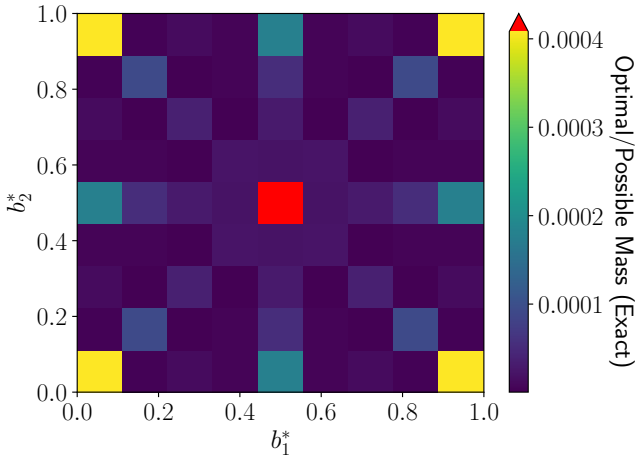


Figure 4: $Pr_n(g^{(\theta^{ds}),\alpha}$ is optimal$|b^*, w^* = (0.5, 0.5))$ When $n = 400$ and b_1^*, b_2^* Vary in $\{0.1, 0.2, \ldots, 0.9\}$.

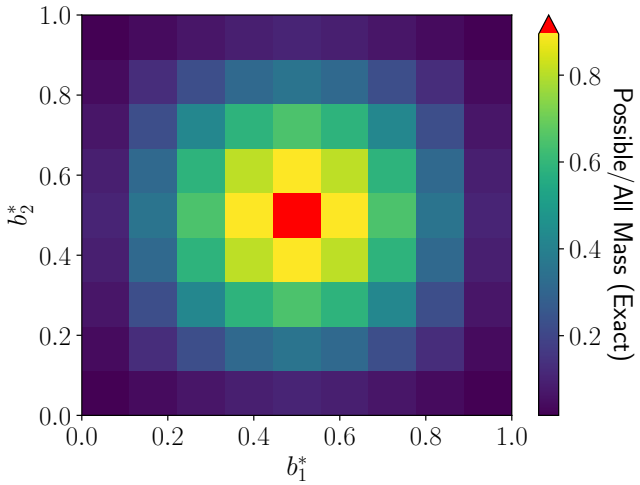


Figure 5: $Pr_n(b^*; w^* = (0.5, 0.5))$ for $n = 400$ and b_1^*, b_2^* Varying in $\{0.1, 0.2, \ldots, 0.9\}$.