Unsupervised Risk Estimation with only Structural Assumptions

Jacob Steinhardt JSTEINHARDT@CS.STANFORD.EDU
Percy Liang PLIANG@CS.STANFORD.EDU
Stanford University, 353 Serra Mall, Stanford, CA 94305

Abstract

Given a model θ and *unlabeled* samples from a distribution p^*, we show how to estimate the *labeled* risk of θ while only making *structural* (i.e., conditional independence) assumptions about p^*. This lets us estimate a model's test error on distributions very different than its training distribution, thus performing unsupervised domain adaptation even without assuming the true predictor remains constant (covariate shift). Furthermore, we can perform discriminative semi-supervised learning, even under model mis-specification. Our technical tool is the method of moments, which allows us to exploit conditional independencies without relying on a specific parametric model. Finally, we introduce a new theoretical framework for grappling with the non-identifiability of the class identities fundamental to unsupervised learning.

1. Introduction

We study the problem of *unsupervised risk estimation* — that is, given a loss function $L(\theta; x, y)$, and a test distribution $p^*(x, y)$, estimate the risk $R(\theta) \overset{\text{def}}{=} \mathbb{E}_{x,y\sim p^*}[L(\theta; x, y)]$ given access only to m unlabeled examples $x^{(1:m)} \overset{i.i.d.}{\sim} p^*(x)$. Can we do this without making strong parametric assumptions about p^*? Although perhaps daunting at first glance, previous work has successfully made progress by only requiring parametric assumptions on the *losses*; for instance, Balasubramanian et al. (2011) assume Gaussianity of model scores, while a line of work starting with Dawid & Skene (1979) assume multiple classifiers with known dependency structure, which specifies a complete generative distribution for the 0/1-loss (Zhang et al., 2014; Platanios, 2015; Jaffe et al., 2015).

In this work, we show that modeling the losses is unnecessary, and that we can in fact recover the risk for multiclass classification while only making *conditional independence* assumptions about p^* — in particular, that the input x is split into three views which are independent conditioned on the true label, and that the loss decomposes over these views (see Figure 1). As an example, these conditions hold for many groups of diseases/symptoms in the QMR knowledge base (Halpern & Sontag, 2013); more generally, the 3-view assumption is the workhorse of the method of moments for estimating latent-variable models (e.g., Anandkumar et al., 2013). Indeed, the method of moments is one of the main technical tools in this paper.

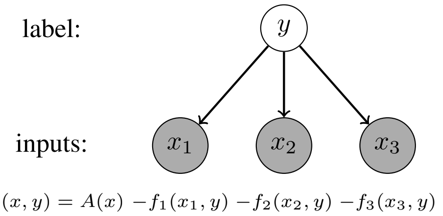


Figure 1: In this paper, we make the "3-view" assumption illustrated above, but do not make any additional generative or discriminative assumptions about the true distribution p^*. In particular, we do not estimate $p^*(x \mid y)$ or $p^*(y \mid x)$, and instead solve for the risk directly using the method of moments. We assume that the loss decomposes across the views, which is the case for many models under both the log and squared loss (but not the hinge loss).

We view unsupervised risk estimation as a prerequisite of any attempt to harness unlabeled data. Two use cases of interest are:

1. **Domain adaptation:** given an initial model trained on a source domain, adapt it to a target domain given only unlabeled target data. If θ_0 is the initial model, we can perform the adaptation as $\min_{\|\theta-\theta_0\|\le r} R(\theta)$.
2. **Semi-supervised learning:** given a small number of labeled examples and many unlabeled examples, fit an accurate model. If we let R_{labeled} denote the risk on the labeled examples, we can do this by solving $\min_\theta R_{\text{labeled}}(\theta) + R(\theta)$.

Traditional approaches to domain adaptation typically assume covariate shift and overlap between the source and target distributions (Shimodaira, 2000; Quiñonero-Candela et al., 2009). Blitzer et al. (2011) show that under a two-

This is a preliminary manuscript, last updated February 2016. Please send any errors or comments to the first author.

view assumption, source-target overlap is unnecessary. Our results show that with three independent views, even the covariate shift assumption can be done away with.

One approach to semi-supervised learning is to build a generative model over x and y, and include the marginal likelihood on the unlabeled examples as part of the cost function during learning. However, a wide body of empirical evidence shows that, when the generative model is misspecified, the unlabeled examples can actually degrade performance (Merialdo, 1994; Cozman & Cohen, 2006; Liang & Klein, 2008; Li & Zhou, 2015). Because of this, two-view assumptions have been used as an alternative to the generative approaches (Blum & Mitchell, 1998; Ando & Zhang, 2007; Kakade & Foster, 2007; Balcan & Blum, 2010). These methods all assume some form of low noise or low regret, as do other methods such as transductive SVMs (Joachims, 1999). Our results imply that, with three independent views, such assumptions are unnecessary.

By focusing on the central problem of risk estimation, our work connects multi-view learning approaches for domain adaptation and semi-supervised learning, and extends them to remove covariate shift and low-noise assumptions. Our work does *not* strictly generalize the work above, as for instance Kakade & Foster (2007) and Blitzer et al. (2011) assume low regret but not independence, and consider regression rather than classification.

Finally, we treat a fundamental identifiability issue — that the class identities are only recoverable up to permutation in any unsupervised setting. Previous work required strong assumptions to circumvent this issue (e.g. that the class probabilities are distinct and known (Balasubramanian et al., 2011) or that the classifiers are correct on average (Jaffe et al., 2015)). We instead show that, as long as the model slightly outperforms random guessing, the class identities are correct with high probability. We do this by formulating a robust Bayesian hypothesis test whose performance is justified via a novel notion of *identifiability index*, using classical tools such as metric entropy (Kolmogorov & Tikhomirov, 1959; Lorentz, 1966), fractional covering numbers (Lovász, 1975), and Lipschitz concentration of Gaussians (Tsirelson et al., 1976).

While previous work has used multi-view assumptions to estimate the $0/1$-risk, our setting is quite different, as the assumptions in e.g. Zhang et al. (2014) or Jaffe et al. (2015) yield a fully-specified family for the distribution of $0/1$-losses, and they proceed by estimating this distribution. In contrast, we directly estimate the risk without needing to estimate the underlying distribution of losses.

Our main technical results are:

- **Risk Estimation:** we estimate the risk $R(\theta)$ to error ϵ given a number of unlabeled samples that depends on ϵ and the number of classes k, but not on the dimension d of θ (Theorem 2.2).
- **Learning:** given $\mathrm{poly}(k) \cdot \frac{d \log(d)}{\epsilon^2}$ unlabeled samples, we learn the optimal parameters θ up to error ϵ (Corollary 3.2).
- **Identifiability:** We develop a robust Bayesian hypothesis test (Algorithm 2), and show that if the estimated risk is at least slightly better than random guessing, then we have very likely identified the true classes.

2. Framework and Estimation Algorithm

We will focus on multiclass classification; we assume an unknown true distribution $p^*(x, y)$ over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{1, \dots, k\}$, and are given unlabeled samples $x^{(1)}, \dots, x^{(m)}$ drawn independently from $p^*(x)$. Given parameters $\theta \in \mathbb{R}^d$ and a loss function $L(\theta; x, y)$, our goal is to estimate the risk of θ on p^*:

$$R(\theta) \overset{\text{def}}{=} \mathbb{E}_{x,y\sim p^*}[L(\theta; x, y)]. \tag{1}$$

Throughout, we will make the *3-view assumption*:

Assumption 2.1 (3-view)**.** *Under p^*, x can be split into x_1, x_2, x_3, which are conditionally independent given y (see Figure 1). Moreover, the loss decomposes additively across views: $L(\theta; x, y) = A(\theta; x) - \sum_{v=1}^{3} f_v(\theta; x_v, y)$, for some functions A and f_v.*

Often L will be the log loss for a model $p_\theta(y \mid x) \propto \exp(\theta^\top \sum_{v=1}^{3} \phi_v(x_v, y))$, in which case $f_v(\theta; x_v, y) = \theta^\top \phi_v(x_v, y)$ and $A(\theta; x)$ is the log partition function.

Note that Assumption 2.1 is not enough to recover R, because permuting the classes $\{1, \dots, k\}$ will preserve $p^*(x)$ (as well as the conditional independence) but will change the risk R. However, it turns out that this is the *only* thing that is unknown; in particular, define the optimistic risk $\tilde{R}$ as the minimum risk over all permutations of the classes:

$$\tilde{R}(\theta) \overset{\text{def}}{=} \min_{\sigma \in \mathrm{Sym}(k)} \mathbb{E}_{x,y\sim p^*}[L(\theta; x, \sigma(y))], \tag{2}$$

where $\mathrm{Sym}(k)$ is the group of permutations on $\{1, \dots, k\}$. We will show that $\tilde{R}$ *can* be recovered, as Theorem 2.2 indicates below. The key insight is that, even without estimating $p^*(y \mid x)$, we can express $\tilde{R}(\theta)$ in terms of certain moments of p^*, which are obtained as the solution to a system of cubic equations (corresponding to 3rd-order moments) derived from Assumption 2.1.

We start by expanding the definition of R:

$$\begin{aligned} R(\theta) &= \bar{A}(\theta) - R_{\text{linear}}(\theta), \text{ where} \\ \bar{A}(\theta) &\overset{\text{def}}{=} \mathbb{E}_{x\sim p^*}[A(\theta; x)], \\ R_{\text{linear}}(\theta) &\overset{\text{def}}{=} \sum_{j=1}^{k} p^*(y = j) \sum_{v=1}^{3} \mathbb{E}[f_v(\theta; x_v, j) \mid y = j]. \end{aligned}$$

The first term $\bar{A}$ can be estimated from only unlabeled data, while the second term R_{linear} can be expressed in terms of the conditional expectations $\mu_{j,v} = \mathbb{E}[f_v \mid y = j]$, more formally defined as (suppressing the dependence on θ):

$$\mu_{j,v} \overset{\text{def}}{=} \mathbb{E}[h_v(x) \mid y = j], \text{ where}$$
$$h_v(x) \overset{\text{def}}{=} \left[f_v(\theta; x_v, 1) \ \cdots \ f_v(\theta; x_v, k)\right]^\top. \tag{3}$$

Thus $h_v(x)$ is the vector of model scores, while $\mu_{j,v}$ is the mean score vector across class j. If we let $\pi_j = p^*(y = j)$ for convenience, then $R_{\text{linear}}(\theta) = \sum_{j=1}^{k} \pi_j \sum_{v=1}^{3} (\mu_{j,v})_j$. This is useful as it implies that, to estimate R, we need only estimate π and μ.

From here, we follow the technical machinery behind the spectral method of moments (e.g., Anandkumar et al., 2012), which we explain for completeness. The conditional independence of the x_v means that conditioned on $y = j$, the second- and third-order moments of h are products of the first-order moments $\mu_{j,v}$. By marginalizing over y, we obtain the following equations, where $\otimes$ is the Kronecker product (also called outer product or tensor product):

$$\mathbb{E}[h_v(x)] = \sum_{j=1}^{k} \pi_j \mu_{j,v}$$
$$\mathbb{E}[h_v(x) \otimes h_{v'}(x)] = \sum_{j=1}^{k} \pi_j \mu_{j,v} \otimes \mu_{j,v'} \text{ for } v \neq v'$$
$$\mathbb{E}[h_1(x) \otimes h_2(x) \otimes h_3(x)] = \sum_{j=1}^{k} \pi_j \mu_{j,1} \otimes \mu_{j,2} \otimes \mu_{j,3} \tag{4}$$

Note that the left-hand-side of each equation can be estimated from unlabeled data. There are more independent equations than unknowns in (4) for any $k \geq 2$; in particular, Anandkumar et al. (2012) show (see Theorem 7 therein) that we can recover π and μ up to permutation: that is, there is some permutation $\sigma \in \mathrm{Sym}(k)$ such that $p^*(y = j) \approx \hat{\pi}_{\sigma(j)}$ and $\mathbb{E}[h_v(x) \mid y = j] \approx \hat{\mu}_{\sigma(j),v}$.

Implications. Once we have π and μ, we can plug back into R. If we had σ, we could plug in exactly:

$$R(\theta) = \bar{A}(\theta) - \sum_{j=1}^{k} p^*(y = j) \sum_{v=1}^{3} \mathbb{E}[f_v(\theta; x_v, j) \mid y = j]$$
$$\approx \bar{A}(\theta) - \sum_{j=1}^{k} \hat{\pi}_{\sigma(j)} \sum_{v=1}^{3} (\hat{\mu}_{\sigma(j),v})_j$$

Since we don't know σ, we can instead take the minimum over all σ, which yields $\tilde{R}$. This minimization is an instance of maximum weight bipartite matching, and can be solved in $\mathcal{O}\left(k^3\right)$ time; see Section A for details.

Putting all of the above ideas together, we obtain Theorem 2.2, which gives a sample complexity bound for estimating $\tilde{R}$ that depends on the number of classes k, the minimum class probability $\pi_{\min}$, the second moment of the loss τ, and the minimum singular value $\sigma_k(M_v)$, where $M_v \overset{\text{def}}{=} [\mu_{1,v} \ \cdots \ \mu_{k,v}]$. More formally:

Theorem 2.2. *Suppose Assumption 2.1 holds. Then, we can estimate $\tilde{R}(\theta)$ to accuracy ϵ with probability $1 - \delta$ for any $0 < \epsilon, \delta < 1$ using*

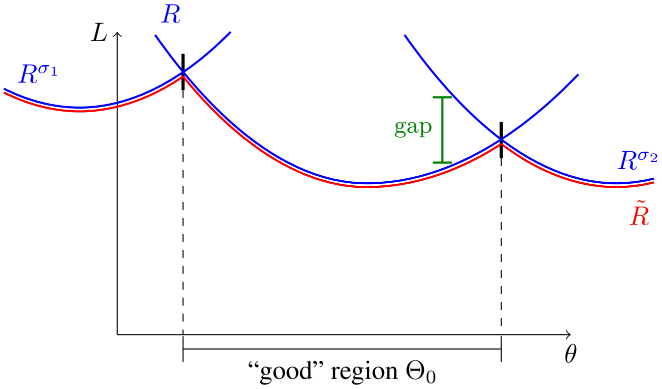
$$m = \mathrm{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right) \cdot \frac{\log(2/\delta)}{\epsilon^2} \textit{ samples, where}$$
$$\pi_{\min} \overset{\text{def}}{=} \min_{j=1}^{k} \pi_j,$$
$$\tau \overset{\text{def}}{=} \max\left(\max_{j,v} \sqrt{\mathbb{E}[\|h_v\|_2^2 \mid y = j]}, \sqrt{\mathbb{E}[A^2]}\right), \textit{ and}$$
$$\lambda \overset{\text{def}}{=} \min_{v=1}^{3} \sigma_k(M_v). \tag{5}$$

For a full proof, see Section B. In summary, Assumption 2.1 yields a set of moment equations (4) that, when solved, allow us to estimate the optimistic risk $\tilde{R}(\theta)$.

3. From Estimation to Learning

In the previous section, we saw how to estimate the risk $R(\theta)$ up to permutation of the labels. We now turn to the problem of learning, i.e., minimizing over $\theta \in \mathbb{R}^d$. We will first show how to compute the gradient $\nabla_\theta \tilde{R}$, and next attend to the difference between minimizing $\tilde{R}(\theta)$ and minimizing $R(\theta)$. For the latter, we assume that we have at least an initial point θ_0 in a certain "good" region Θ_0, often obtainable by training on a related task or from a small amount of supervised data. To elaborate on this, we provide some geometric intuition about R and $\tilde{R}$.

Let $R^\sigma(\theta)$ be the risk when the labels are permuted by σ: $R^\sigma(\theta) \overset{\text{def}}{=} \mathbb{E}_{x,y\sim p^*}[L(\theta; x, \sigma(y))]$. Then $R = R^{\text{id}}$, where id is the identity permutation, while $\tilde{R}$ is the minimum of the functions R^σ. This is illustrated below:



We define Θ_0 to be the region where $R(\theta) = \tilde{R}(\theta)$, or equivalently where $\min_{\sigma \neq \text{id}} R^\sigma(\theta) - R^{\text{id}}(\theta) \geq 0$. Define $\mathrm{gap}(\theta)$ to be the difference between the smallest and second-smallest values (over σ) of $R^\sigma(\theta)$; then $\mathrm{gap}(\theta) = 0$

at the boundary of Θ_0. The gap measures how easy it is to mix up two of the curves R^σ in the presence of noise.

Computing the gradient. To compute $\nabla_\theta R$, the straightforward approach is (recalling $R(\theta) = \bar{A}(\theta) - \sum_{v=1}^3 \mathbb{E}[f_v(\theta; x, y)]$) to take the vectors $\nabla_\theta f_v$ and compute their expectations similarly to Theorem 2.2. However, equation (4) would then involve a $kd \times kd \times kd$ tensor, which is prohibitive even for e.g. $k = 2$, $d = 10^4$. Instead, we take an approach inspired by Nesterov & Spokoiny (2011), which is to compute the directional derivative $u^\top \nabla_\theta f_v$ (requiring only a $k \times k \times k$ tensor) and from several such u approximate the full gradient.

In particular, we take $u \in \mathbb{R}^d$ and replace f_v in equation (4) with $u^\top \nabla_\theta f_v$, thus estimating $u^\top \nabla_\theta R_{\text{linear}} = \sum_{v=1}^3 \mathbb{E}_{x,y}[u^\top \nabla_\theta f_v]$ from only $\text{poly}(k)$ samples. By sampling many such u, we can estimate the full gradient as $\sum_{v=1}^3 \mathbb{E}_{u,x,y}[uu^\top \nabla_\theta f_v(x,y)]$, assuming $\mathbb{E}[uu^\top] = I_{d\times d}$.

This still does not quite work, because we only obtain each $u^\top \nabla_\theta f_v$ up to a permutation σ_u, which will be different for different u. We remedy this by *simultaneously* estimating f_v and $u^\top \nabla_\theta f_v$; since f_v is the same each time, we can use it to undo the permutations σ_u, allowing us to correctly average across u. The full procedure is given as Algorithm 1, and requires $\log(d/\epsilon)$ times as many samples as in Theorem 2.2. In particular (see Section C for proof):

Theorem 3.1. *Suppose that Algorithm 1 is run with* $m = \text{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right) \cdot \frac{\log(d/\epsilon\delta)}{\epsilon^2}$ *samples as input at a point* θ_0, *and that* $\epsilon < \min(1, \text{gap}(\theta_0))$. *Then with probability* $1 - \delta$, *the output of Algorithm 1 is a vector* $\hat{g}$ *satisfying*

$$\left\| \hat{g} - \nabla \tilde{R}_{\text{linear}}(\theta_0) \right\|_2 \le \epsilon B, \tag{6}$$

where $B^2 = \max_{v=1}^3 \mathbb{E}[\max_{j=1}^k \|\nabla_\theta f_v(\theta_0; x, j)\|_2^2]$ *measures the squared* ℓ^2*-norm of the gradient.*

Note that Algorithm 1 can likely be improved substantially by replacing the random projections with fast low-rank projections (Halko et al., 2011), which would allow efficient direct computation of the full gradient.

Linear models. When the f_v are *linear* in θ — e.g. for multiclass logistic regression — evaluating the gradient once is sufficient for all iterations of learning. The reason is that $R(\theta)$ is then equal to $\bar{A}(\theta) - \theta^\top \nabla_\theta \tilde{R}_{\text{linear}}(\theta_0)$, for any vector $\theta_0 \in \Theta_0$. As long as we regularize θ by constraining $\|\theta\|_2 \le \rho$, approximating $\nabla \tilde{R}_{\text{linear}}$ with $\hat{g}$ will still yield a near-optimal θ. In particular (again see Section C):

Corollary 3.2. *Suppose that* $f_v(\theta; x_v, y) = \theta^\top \phi_v(x_v, y)$, *where* $\|\phi_v(x_v, j)\|_2 \le B$ *for all* v, x_v, j *and that* $A(\theta; x)$ *is* B*-Lipschitz in* θ. *Also suppose that we run Algorithm 1 at a point* $\theta_0 \in \Theta_0$ *to obtain* $\hat{g}$. *Then, if* ϵ *and* m *satisfy the conditions in Theorem 3.1, and we let*

$$\hat{\theta} = \operatorname*{arg\,min}_{\|\theta\|_2 \le \rho} \frac{1}{m} \sum_{i=1}^m A(\theta; x^{(i)}) - \theta^\top \hat{g}, \tag{7}$$

then $R(\hat{\theta}) \le \min_{\|\theta\|_2 \le \rho} R(\theta) + 2\epsilon B \rho$ *with probability* $1 - 2\delta$.

Algorithm 1 Algorithm for computing the gradient of $\tilde{R}$.

```
Input: initial parameters $\theta_0 \in \mathbb{R}^d$, samples $x^{(1:m)} \sim p^*$.
Define $h_v(x,u) \in \mathbb{R}^{2k}$ as
    $h_v(x,u) = [f_v(\theta_0; x_v, j) \;\; u^\top \nabla_\theta f_v(\theta_0; x_v, j)/B]_{j=1}^k$
for $t = 1$ to $T = \frac{75d\log(2d/\delta)}{\epsilon^2}$ do
  Sample $u_t$ uniformly from $\{\pm 1\}^d$.
  Using $x^{(1:m)}$, estimate the moments
      $\mathbb{E}[h_v(x,u_t)], \mathbb{E}[h_{v_1}(x,u_t) \otimes h_{v_2}(x,u_t)]$, and
      $\mathbb{E}[h_1(x,u_t) \otimes h_2(x,u_t) \otimes h_3(x,u_t)]$.
  Use tensor decomposition to compute $\hat{\pi}_{j,t} \approx \pi_j$, $\hat{\mu}_{j,t} \approx \sum_v \mu_{j,v}$, and $\hat{g}_{j,t} \approx \sum_v \mathbb{E}[u_t^\top \nabla_\theta f_v(\theta_0; x_v, j) \mid j]$.
  Permute the rows of $\hat{\mu}_{j,t}$ (and simultaneously $\hat{g}_{j,t}$) such that $\sum_j \hat{\pi}_{j,t}(\hat{\mu}_{j,t})_j$ is maximized.
end for
Output $\hat{g} = \frac{1}{T}\sum_{t=1}^T u_t \sum_{j=1}^k \hat{\pi}_{j,t}\hat{g}_{j,t}$.
```

If e.g. $\|\theta\|_2 = \|\phi\|_\infty = 1$, then we will have $\rho = 1$ and $B \le \sqrt{d}$, whence we need $\epsilon \ll 1/\sqrt{d}$ to obtain good bounds on $R(\hat{\theta})$; Corollary 3.2 then implies a sample complexity of $\text{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right) \cdot d\log(d/\epsilon\delta)$.

A general criterion. If the f_v are non-linear, we likely need to evaluate $\tilde{R}$ at more than one point, and there is no guarantee that $\tilde{R}(\theta) = R(\theta)$ once we move away from θ_0. To help address this, we provide a criterion which certifies that, given an initial point $\theta_0 \in \Theta_0$, a new point θ will also lie in Θ_0:

Lemma 3.3. *Suppose* $\theta_0 \in \Theta_0$, *and* θ *satisfies*

$$\mathbb{E}_{x\sim p^*}\left[\max_{j=1}^k |f(\theta; x, j) - f(\theta_0; x, j)|\right] < \frac{1}{2}\left(\text{gap}(\theta_0) + \text{gap}(\theta)\right), \tag{8}$$

where $f = \sum_{v=1}^3 f_v$. *Then,* $\theta \in \Theta_0$ *as well.*

See Section D for a proof. The idea is that the left hand side of (8) bounds the amount that any of the R^σ (or more precisely, $R^\sigma - \bar{A}$) can move. Note that the left-hand-side of (8) can be estimated from unlabeled data.

We can use Lemma 3.3 to modify any optimization procedure: given any proposed next point θ_{t+1}, backtrack in the direction of the current point θ_t until (8) holds, and then continue the optimization. Note however that if (8) is overly conservative then we may not reach the optimum.

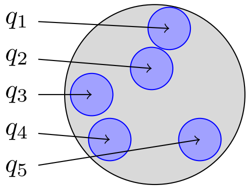


Figure 2: Illustration of Definition 4.1. Each ball is a set $B_r(q) = \{p \mid \mathrm{D}(p \parallel q) \le r\}$. If no single ball has large probability mass under ν, then no small number of distributions $q_1, \dots, q_M$ can fit a random distribution from ν well.

4. Identifiability Index

In Section 2, we showed how to identify the risk R up to permutation of the classes (thus instead obtaining the optimistic risk $\tilde{R}$). We now study more carefully how this optimism can affect our estimate of the risk. This will help us avoid false negatives where $\tilde{R}$ is small but R is large; at a high level, we will see that if $\tilde{R}$ is better than random guessing, then $\tilde{R} = R$ with high probability.

Intuitively, a generic "bad" distribution is simply not very related to the model, and so none of its permutation will be, either. To formalize what is meant by generic, we adopt a robust Bayesian perspective: we define a prior ν over possible input distributions p^*, and design a test with low false negative rate under the prior, where the test is (nearly) independent of the choice of ν. In the following, we restrict to the log loss, as the proofs are already hard for this case.

Identifiability index. If we think of the $k!$ permutations of the classes as simply $k!$ candidate models, then the question becomes: how high can we make the probability that at least one of the models looks good by chance? We adopt a Bayesian approach and suppose that the true distribution $p^*(y \mid x)$ is drawn from a prior ν. For a given $q(y \mid x)$, we can consider the ball of possible p^* that are fit well by q; if all such balls have low mass under ν, then it is impossible to fit p^* well by chance (this is illustrated in Figure 2).

Formally, we take the ball $B_r(q) \overset{\text{def}}{=} \{p \mid \mathrm{D}(p \parallel q) \le r\}$; here the divergence $\mathrm{D}(p \parallel q)$ is the average KL divergence

$$\mathrm{D}(p \parallel q) \overset{\text{def}}{=} \mathbb{E}_{x \sim \pi}[\mathrm{KL}(p(\cdot \mid x) \parallel q(\cdot \mid x))], \tag{9}$$

where $\pi(x)$ is the distribution over x. We then define the identifiability index of ν to be the most mass that can be covered by a single such ball:

Definition 4.1 (Identifiability index)**.** For a given r, the *identifiability index* α_r for a prior ν is defined as $\sup_q \nu(B_r(q))$.

In particular, if $\alpha_r \ll \frac{1}{k!}$, then it is unlikely that *any* of the $k!$ permutations of a model p_θ would have low risk by chance. We will see below (Theorem 4.4) that α_r is often $\exp(-\Omega(d))$ for d-dimensional exponential families.

Algorithm 2 Algorithm for checking that $\tilde{R}(\theta) = R(\theta)$.

```
Input: model $p_\theta$, failure probability $\delta$, samples $x^{(1:m)}$.
Let $r_0$ satisfy $\alpha_{r_0} \le \frac{\delta}{k!}$.
Return true if $\tilde{R}(\theta) \le r_0$, else false.
```

In addition to the identifiability index, we also care about how well-separated the classes are — if they are not very well-separated, then it is more likely that two classes have been mixed up. We can formalize this by looking at the minimum gap in risk from getting a single class wrong:

Definition 4.2 (Class separation)**.** The *class separation* $\gamma(\theta)$ of p_θ is defined as $\gamma(\theta) \overset{\text{def}}{=} \min_{j' \ne j} (\mu_j)_j - (\mu_j)_{j'}$, where $\mu_j = \sum_{v=1}^{3} \mu_{j,v}$ is the mean score across class j.

Note that we cannot estimate the class separation (since it depends on the true permutation of the μ_j), so we assume that *either* $\gamma(\theta) > 0$, *or* p^* is drawn at random from ν. We can then check that $R(\theta) = \tilde{R}(\theta)$ (with failure probability δ) by checking that $\tilde{R}(\theta) \le r_0$, where $\alpha_{r_0} \le \delta / k!$. This test is depicted in Algorithm 2. The validity of this test is certified by Proposition 4.3 (proved in Section E):

Proposition 4.3. *Suppose that ν has identifiability index α_r, and that ν' induces positive class separation ($\gamma(\theta) > 0$) with probability 1. Then, if $p^* \sim \hat{\nu}$, with $\hat{\nu} \in \{\nu, \nu'\}$, we have the false negative bound*

$$\max_{\hat{\nu} \in \{\nu, \nu'\}} \mathbb{P}_{\hat{\nu}}[\tilde{R}(\theta) \le r \wedge \tilde{R}(\theta) \ne R(\theta)] \le k!\, \alpha_r. \tag{10}$$

Thus, assuming that p^* either has positive separation or is drawn at random from ν, Algorithm 2 has a false negative rate of at most δ. The test depends on the prior ν only through the identifiability index, which we show below is small for many priors, even when $r = \log(k) - \epsilon$ (note that $\log(k)$ is the risk of uniform guessing).

Identifiability and learning. The intuition in Figure 2 also applies in the learning setting, if instead of the $k!$ permutations of a fixed model we consider all models in some family Θ. We can bound the false negative rate under $p^* \sim \nu$ in terms of the identifiability index as well as the covering number of Θ (i.e., the minimum number of balls $B_r(q)$ needed to cover Θ); details are in Section F.

Computing α_r. We referred to Algorithm 2 as a robust Bayesian hypothesis test because the test depends on the prior ν only through the identifiability index α_r. We strengthen the robustness argument by showing that α_r is small — typically exponentially small in the dimension.

To start, for exponential families, the identifiability index α_r is large under a Gaussian prior:

Theorem 4.4. *Consider an exponential family defined by ϕ, i.e. $p_\beta(y \mid x) \propto \exp(\beta_y^\top \phi(x))$. Assume that $\phi(x) \ne 0$ almost surely, and let γ be the maximum singular value of*

$\mathbb{E}_{x\sim\pi}[\phi(x)\phi(x)^\top/\|\phi(x)\|_2^2]$. *Then, there exists a multivariate normal distribution ν over β with identifiability index $\alpha_{\log(k)-\epsilon} = \exp((\epsilon/3k)^4\,\gamma^{-1})$.*

Note that $\gamma^{-1} \approx d$ assuming the features are well-conditioned. For large d Theorem 4.4 implies that $\alpha_r \ll 1/k!$ when r is even slightly less than $\log(k)$. The proof of Theorem 4.4 is based on Lipschitz concentration bounds for Gaussian distributions and is given in Section 5.

More generally, we can consider any family of distributions $\mathcal{P}$, and ask whether there is a prior over $\mathcal{P}$ with small identifiability index. As might be expected from Figure 2, the existence of such a prior depends only on the covering number $N_{\mathcal{P}}(\cdot)$. Some care is needed because the divergence $\mathrm{D}\left(\cdot \,\|\, \cdot\right)$ is not a metric, but we can nevertheless show:

Theorem 4.5. *Given a space $\mathcal{P}$, let α_r^* be the minimum identifiability index of any prior ν on $\mathcal{P}$. Also assume that α_r^* is continuous in r and that $N_{\mathcal{P}}(r) < \infty$ for all $r > 0$. Then for all $0 < \epsilon < 1$,*

$$\frac{N_{\mathcal{P}}\left(r-\epsilon\right)}{1+\log N_{\mathcal{P}}(k^{-\frac{1}{2}}\left(\epsilon/26\right)^3)} \le (\alpha_r^*)^{-1} \le N_{\mathcal{P}}\left(r\right). \qquad (11)$$

Therefore, for any sufficiently large space $\mathcal{P}$ we can obtain a prior with small identifiability index. Theorem 4.5 formalizes a simple intuition: if it takes a large number of distributions to cover $\mathcal{P}$, then a small number of distributions can only cover a small fraction of $\mathcal{P}$. Our proof, given in Section 5, exploits a duality between the identifiability index and fractional covering number, as well as certain approximate triangle inequalities for the KL divergence.

5. Identifiability Index: Proofs

We now provide proofs of Theorems 4.4 and 4.5. This section may be skipped if desired, but we include it for the inclined reader because the ideas are novel. To preserve the flow, we sometimes gloss over parts of the argument, in which case we will include a reference to the part of the supplement where the argument is fleshed out.

Proof of Theorem 4.4

Recall that we have $p_\beta(Y = j \mid x) \propto \exp(\beta_j^\top\phi(x))$, and want a prior ν over β such that $\mathbb{P}_\beta[\mathrm{D}\left(p_\beta \,\|\, q\right) \le \log k - \epsilon]$ is small for all q. We will take each $\beta_j \sim \mathcal{N}(0, I_{d\times d})$, and then scale by a large enough constant τ that $p_\beta(y \mid x)$ can be treated (see G.1) as a point mass:

$$p_\beta(y \mid x) = \delta_{y_\beta(x)}, \quad y_\beta(x) \overset{\text{def}}{=} \arg\max_j \beta_j^\top\phi(x). \qquad (12)$$

At a high level, we prove a concentration inequality: for any fixed q, $\mathbb{E}_{x\sim\pi}[q(y_\beta(x) \mid x)] \approx \frac{1}{k}$ with high probability (over β). We do this by constructing a Lipschitz (in β) approximation to q and then applying known results on Lipschitz functions of Gaussians (Tsirelson et al., 1976).

To start, note that $\mathrm{D}\left(p_\beta \,\|\, q\right) = \mathbb{E}_{x\sim\pi}[-\log q(y_\beta(x) \mid x)]$. Since $-\log(q) \ge \log(k) + 1 - kq$, we can show that

$$\mathrm{D}\left(p_\beta \,\|\, q\right) \le \log(k) - \epsilon \implies \mathbb{E}_{x\sim\pi}[q(y_\beta(x) \mid x)] \ge \frac{1+\epsilon}{k}. \qquad (13)$$

To avoid technical issues, we approximate the distribution over x by m samples and take $m \to \infty$, yielding (see G.2):

$$\lim_{m\to\infty} \mathbb{P}_{\beta, x_{1:m}}\left[\frac{1}{m}\sum_{i=1}^m q(y_\beta(x_i) \mid x_i) \ge \frac{1+\epsilon}{k}\right]. \qquad (14)$$

To bound the sample average in (14), we will exploit the Lipschitz concentration of multivariate Gaussians:

Theorem 5.1 (Boucheron et al. (2013), Theorem 5.6)**.** *If f is L-Lipschitz in ℓ^2-norm and $\beta \sim \mathcal{N}(0, I)$, then*

$$\mathbb{P}[f(\beta) - \mathbb{E}[f(\beta)] \ge t] \le \exp(-t^2/2L^2). \qquad (15)$$

We interpret $\beta \mapsto \frac{1}{m}\left(q(y_\beta(x_1) \mid x_1) + \cdots + q(y_\beta(x_m) \mid x_m)\right)$ as a composition of simpler functions, each of which is Lipschitz. Let $t_j^i = \phi(x_i)^\top\beta_j/\|\phi(x_i)\|_2$ (so that $y_\beta(x_i) = \arg\max_j t_j^i$), and note that $[t_j^1 \;\cdots\; t_j^k] = \beta_j^\top\Phi$, where $\Phi = [\phi(x_i)/\|\phi(x_i)\|_2]_{i=1}^m$. We have (see G.3) the following diagram, with Lipschitz constants in blue:

$$\begin{tikzcd}
& \beta_1,\dots,\beta_k \arrow[dl] \arrow[dr, "\|\Phi\|_{\mathrm{op}}"] & \\
t_1^1,\dots,t_k^1 \arrow[d] & \cdots & t_1^m,\dots,t_k^m \arrow[d, "???"] \\
q(y_\beta(x_1) \mid x_1) \arrow[dr] & \cdots & q(y_\beta(x_m) \mid x_m) \arrow[dl, "\frac{1}{\sqrt{m}}"] \\
& \frac{1}{m}\sum_{i=1}^m q(y_\beta(x_i) \mid x_i) &
\end{tikzcd}$$

The first step $\beta_{1:k} \mapsto t_{1:k}^{1:m}$ has Lipschitz constant $\|\Phi\|_{\mathrm{op}}$ (where $\|\cdot\|_{\mathrm{op}}$ indicates operator norm), and the last has Lipschitz constant $1/\sqrt{m}$. However, the argmax in the middle stage is not even continuous, so has Lipschitz constant ∞!

We fix this by giving q credit for being close: i.e. if $y' \neq y$ is distance $s < \delta$ from being the argmax, add in $\left(1 - \frac{s}{\delta}\right)q(y')$. This increases the expectation of q by at most δ (see G.4) while making the function $t \mapsto q$ be $\frac{\sqrt{2}}{\delta}$-Lipschitz. Taking $\delta = \frac{\epsilon}{2k}$, we obtain a $\frac{2\sqrt{2}k\|\Phi\|_{\mathrm{op}}}{\epsilon\sqrt{m}}$-Lipschitz composition, and want it to exceed its expectation (of at most $\frac{1+\epsilon/2}{k}$) by less than $\frac{\epsilon}{2k}$. Applying Theorem 5.1, the probability that this does not happen is at most (see G.5) $\exp\left(-\frac{\epsilon^4}{64k^4}\frac{m}{\|\Phi\|_{\mathrm{op}}^2}\right)$. Since $\frac{1}{m}\Phi\Phi^\top$ converges almost surely to $\mathbb{E}[\phi(x)\phi(x)^\top/\|\phi(x)\|_2^2]$, (14) is bounded by (see G.6) $\exp(-(\epsilon^4/64k^4)\gamma^{-1}) < \exp(-(\epsilon/3k)^4\gamma^{-1})$, from which Theorem 4.4 follows.

Proof of Theorem 4.5

To upper bound $(\alpha_r^*)^{-1}$, take a minimal covering $\{q_m\}_{m=1}^{N_{\mathcal{P}}(r)}$; note that $\nu\left(\cup_m B_r(q_m)\right) = 1$ for any ν, hence at least some $B_r(q_m)$ must have mass at least $\frac{1}{N_{\mathcal{P}}(r)}$.

For the lower bound, we want to show that if $\mathcal{P}$ needs many balls $B_r(q)$ to be covered, then a single ball can cover only a small fraction of $\mathcal{P}$. Let $\mathcal{U}$ be the space of all distributions q, and let $f : \mathcal{U} \to \{0, 1\}$ select a single $q \in \mathcal{U}$. We can then write down a min-max integer linear program between ν and f in order to find the distribution ν over $\mathcal{P}$ that is hardest to cover. In the equations below, we interpret f as a member of the functions from $\mathcal{U}$ to $\{0, 1\}$ with finite support, and write $\|f\|_1 = \sum_{q:f(q)\neq 0}|f(q)|$. Then $\|f\|_1 = 1$, and applying the minimax theorem (see H.2), we have

$$\alpha_r^* = \min_{\nu} \sup_{\substack{f:\mathcal{U}\to\{0,1\}\\ \|f\|_1=1}} \mathbb{E}_{p\sim\nu}\left[\overbrace{\sum_{q:B_r(q)\ni p} f(q)}^{1 \text{ if } B_r(q) \text{ covers } p}\right]$$

$$= \sup_{\substack{f:\mathcal{U}\to[0,1]\\ \|f\|_1=1}} \min_{p\in\mathcal{P}} \sum_{q:B_r(q)\ni p} f(q) = \frac{1}{N_{\text{frac}}}, \qquad (16)$$

where $N_{\text{frac}} = \inf \|f\|_1$ s.t. $\sum_{q:B_r(q)\ni p} f(q) \geq 1 \; \forall p \in \mathcal{P}$. The quantity N_{frac} is the *fractional covering number* and is well-studied. For instance (Lovász, 1975):

Lemma 5.2. *For a collection of subsets $\mathcal{C}$ of a space $\mathcal{P}$, let $N(\mathcal{P}, \mathcal{C})$ denote the covering number and $N_{\text{frac}}(\mathcal{P}, \mathcal{C})$ the fractional covering number. Then $N_{\text{frac}}(\mathcal{P}, \mathcal{C}) \geq \frac{N(\mathcal{P},\mathcal{C})}{1+\log|\mathcal{P}|}$.*

The proof is a simple randomization argument; see H.1. In our case, $\mathcal{C} = \{B_r(q) \mid q \in \mathcal{U}\}$. Lemma 5.2 does not directly apply because $|\mathcal{P}|$ is infinite, but the following approximate triangle inequality (proved in H.3) lets us discretize, showing that we can replace $\mathcal{P}$ with a sufficiently fine covering $\hat{\mathcal{P}}$ of $\mathcal{P}$ while only changing KL divergences by a small amount:

Lemma 5.3. *Suppose that $p, \hat{p}$ satisfy $\mathrm{D}\left(p \parallel \hat{p}\right) \leq \epsilon$ for some $\epsilon < 1/4$. Then for any q, the mixture distribution $\overline{q} = (1 - \sqrt{\epsilon})q + \sqrt{\epsilon}u$ (where $u(y \mid x)$ is uniform) satisfies:*

1. $\mathrm{D}\left(p \parallel \overline{q}\right) \leq \mathrm{D}\left(\hat{p} \parallel q\right) + 5\sqrt{\epsilon}\log(k/\epsilon)$
2. $\mathrm{D}\left(\hat{p} \parallel \overline{q}\right) \leq \mathrm{D}\left(p \parallel q\right) + 5\sqrt{\epsilon}\log(k/\epsilon)$

If $\hat{\mathcal{P}}$ is a ϵ-covering of $\mathcal{P}$, we can thus transform any r-covering covering of $\hat{\mathcal{P}}$ into a $(r + 5\sqrt{\epsilon}\log(k/\epsilon))$-covering of $\mathcal{P}$ and vice versa. Applying Lemma 5.2 to $\hat{\mathcal{P}}$ then yields Theorem 4.5. (See H.4 for a more detailed justification.)

6. Experiments

To better understand the behavior of our algorithms, we perform experiments on a version of the MNIST data set that is modified to ensure that the 3-view assumption holds. Specifically, to create an image, we sample a class in $\{0, \ldots, 9\}$, then sample 3 images at random from that class, letting every third pixel come from the respective image. This guarantees that there will be 3 conditionally independent views. To explore train-test variation, we dim pixel p in the image by $\exp\left(\lambda\left(\|p - p_0\|_2 - 0.4\right)\right)$, where p_0 is the image center and the distance is normalized to have maximum value 1. We show example images for $\lambda = 0$ (train) and $\lambda = 5$ (a possible test distribution) in Figure 3.

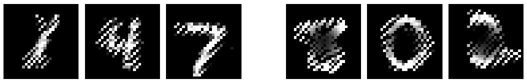


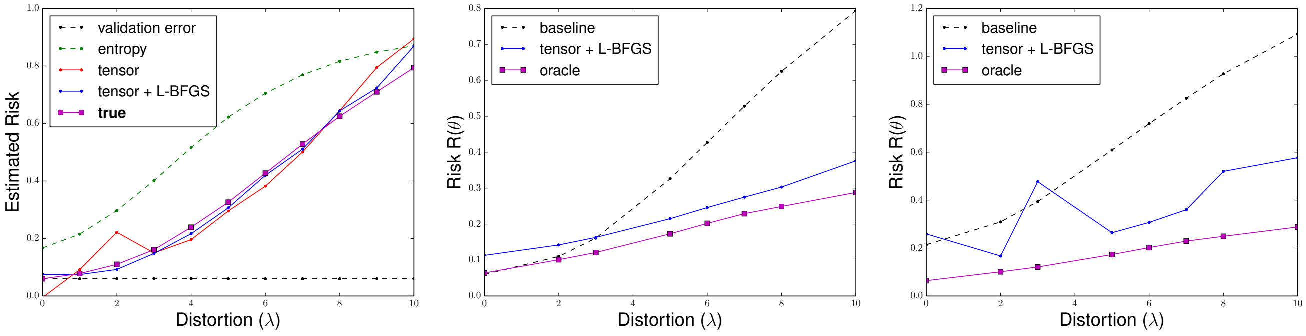
Figure 3: Sample train images (left) and test images (right) from the modified MNIST data set.

Risk estimation. We use unsupervised risk estimation (Theorem 2.2) to estimate the risk of a model trained on $\lambda = 0$ and tested on various values of $\lambda \in [0, 10]$. We trained the model with AdaGrad (Duchi et al., 2010) on $10,000$ training examples, and used $10,000$ test examples to estimate the risk. To solve for π and μ in (4), we first use the tensor power method from Anandkumar et al. (2013) to initialize, and then locally minimize a weighted ℓ^2-norm of the moment errors with L-BFGS. As a baseline, we take the validation error for $\lambda = 0$ (i.e., assume train = test), as well as the predictive entropy $-\sum_j p_\theta(j \mid x)\log p_\theta(j \mid x)$ on the test set (i.e., assume the predictions are well-calibrated). The results are shown in Figure 4a; both the tensor method in isolation and tensor + L-BFGS estimate the risk accurately, with the latter performing slightly better.

Domain adaptation. We next test the efficacy of our learning algorithm. For θ_0 we used the trained model at $\lambda = 0$, using $T = 400,000$ random projections to estimate ∇R_{linear}, and then minimizing over $\|\theta\|_2 \leq 10$. The results are shown in Figure 4b. For small values of λ, Algorithm 1 is less data-efficient than the baseline of directly using θ_0. However, our algorithm is far more robust as λ increases, and tracks the performance of an oracle that minimizes the (supervised) test error subject to $\|\theta\|_2 \leq 10$.

Semi-supervised learning. Finally, to test the semi-supervised aspect of our method, we also considered a θ_0 that was obtained from only 300 training examples, again at $\lambda = 0$. The tensor method sometimes led to bad initializations, in which case we obtained a different θ_0 by training with a smaller step size. The results are shown in Figure 4c. We generally outperform the baseline of θ_0, but our learned parameters are higher-variance than before, seemingly due to higher condition number of the matrices $P_{v,v'}$.

Summary. Our experiments show that given 3 views, we can estimate the risk and perform domain adaptation, even from a small amount of supervised data.



(a) (b) (c)

Figure 4: Results on the modified MNIST data set. (a) Risk estimation for varying degrees of distortion λ. (b) Transfer learning with $10,000$ training and $10,000$ test examples. (c) Transfer learning with 300 training and $10,000$ test examples.

7. Discussion

We have presented a method for estimating the risk from unlabeled data, which relies only on assumptions about the conditional independence structure and hence makes no parametric assumptions about the true distribution. Moreover, using the identifiability index, we have established criteria under which the optimistic risk $\tilde{R}$ equals the true risk R. We require only a "seed model" θ_0 that does sufficiently better than random guessing (Proposition 4.3), from which we can then learn from only unlabeled data (Corollary 3.2). This seed model could have been trained on a related domain, on a small amount of supervised data, or any combination of the two, and thus provides a pleasingly general contract that highlights the similarities between domain adaptation and semi-supervised learning.

Both our work and previous work on unsupervised risk estimation focuses on the model predictions, rather than the full distribution of data, which allows efficient estimation of the risk even though the data distribution itself is treated nonparametrically. This "semi-parametric approach" has been well-studied in the econometrics literature (Powell, 1994), which to us presents an exciting opportunity to import a new set of tools into the sphere of machine learning. Econometrics has also used the generalized method of moments as a tool for handling model mis-specification (Hansen, 1982; Newey & McFadden, 1994), further suggesting a convergence of tools and goals.

Unsupervised risk estimation touches upon the recently-posed problem of what Bottou (2015) calls *machine learning with contracts*, which asks that a machine learning system satisfy a well-defined input-output contract in analogy with software systems (Sculley et al., 2015). Theorem 2.2 provides the contract that under the 3-view assumption the test error is close to our estimate of the test error. This contrasts with the typical weak contract that if train and test are similar, then the test error is close to the training error.

The most restrictive part of our framework is the 3-view assumption, which may be inappropriate if the views are not completely independent or if the data have other structure that is not well-captured in terms of multiple views. Since Balasubramanian et al. (2011) obtain results under Gaussianity (which would be implied by many somewhat dependent views), we are optimistic that unsupervised risk estimation is possible for a wider family of structures. Along these lines, we pose the following two questions:

Open question. In the multi-view setting, suppose that the views are not completely independent. Is it still possible to estimate the risk without assuming a distribution over the losses? How does the degree of independence affect the number of views needed?

Open question. Given a general Bayes net structure on x and y, when is unsupervised risk estimation possible?

On a more technical level, while Algorithm 1 accurately computes the gradient, it is slow, and more sophisticated approaches would likely work better. In addition, though Lemma 3.3 provides some guidance, we do not have a clear algorithm for optimizing $R(\theta)$ in the case that the f_v are non-linear in θ (e.g. for neural nets); a general algorithm for non-linear, non-convex losses would be desirable. Finally, we believe the correct dependence in Theorem 4.4 should be $\exp(-C \cdot (\epsilon/k)^2 d)$, which likely requires a more precise argument than the Lipschitz bound presented here.

The results of this paper have caused us to adopt the following perspective: To handle unlabeled data, we should make *generative* structural assumptions, but still optimize *discriminative* model performance. This hybrid approach allows us to satisfy the traditional machine learning goal of predictive accuracy, while handling lack of supervision and under-specification in a principled way. Perhaps, then, what is needed for learning is to understand the *structure* of a domain.

References

Anandkumar, A., Hsu, D., and Kakade, S. M. A method of moments for mixture models and hidden Markov models. In *Conference on Learning Theory (COLT)*, 2012.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. Tensor decompositions for learning latent variable models. *arXiv*, 2013.

Ando, R. K. and Zhang, T. Two-view feature generation model for semi-supervised learning. In *Conference on Learning Theory (COLT)*, pp. 25–32, 2007.

Balasubramanian, K., Donmez, P., and Lebanon, G. Unsupervised supervised learning II: Margin-based classification without labels. *Journal of Machine Learning Research (JMLR)*, 12:3119–3145, 2011.

Balcan, M. and Blum, A. A discriminative model for semi-supervised learning. *Journal of the ACM (JACM)*, 57(3), 2010.

Blitzer, J., Kakade, S., and Foster, D. P. Domain adaptation with coupled subspaces. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 173–181, 2011.

Blum, A. and Mitchell, T. Combining labeled and unlabeled data with co-training. In *Conference on Learning Theory (COLT)*, 1998.

Bottou, L. Two high stakes challenges in machine learning. Invited talk at the 32nd International Conference on Machine Learning, 2015.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Cozman, F. and Cohen, I. Risks of semi-supervised learning: How unlabeled data can degrade performance of generative classifiers. In *Semi-Supervised Learning*. 2006.

Dawid, A. P. and Skene, A. M. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 1:20–28, 1979.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. In *Conference on Learning Theory (COLT)*, 2010.

Edmonds, J. and Karp, R. M. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM (JACM)*, 19(2):248–264, 1972.

Halko, N., Martinsson, P., and Tropp, J. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53:217–288, 2011.

Halpern, Y. and Sontag, D. Unsupervised learning of noisy-or Bayesian networks. In *Uncertainty in Artificial Intelligence (UAI)*, 2013.

Hansen, L. P. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 50:1029–1054, 1982.

Jaffe, A., Nadler, B., and Kluger, Y. Estimating the accuracies of multiple classifiers without labeled data. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 407–415, 2015.

Joachims, T. Transductive inference for text classification using support vector machines. In *International Conference on Machine Learning (ICML)*, 1999.

Kakade, S. M. and Foster, D. P. Multi-view regression via canonical correlation analysis. In *Conference on Learning Theory (COLT)*, pp. 82–96, 2007.

Kolmogorov, A. N. and Tikhomirov, V. M. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

Li, Y. and Zhou, Z. Towards making unlabeled data never hurt. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):175–188, 2015.

Liang, P. and Klein, D. Analyzing the errors of unsupervised learning. In *Human Language Technology and Association for Computational Linguistics (HLT/ACL)*, 2008.

Lorentz, G. G. Metric entropy and approximation. *Bulletin of the American Mathematical Society*, 72(6):903–937, 1966.

Lovász, L. On the ratio of optimal integral and fractional covers. *Discrete Mathematics*, 13(4):383–390, 1975.

Merialdo, B. Tagging English text with a probabilistic model. *Computational Linguistics*, 20:155–171, 1994.

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, pp. 1–40, 2011.

Newey, W. K. and McFadden, D. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pp. 2111–2245. 1994.

Platanios, E. A. Estimating accuracy from unlabeled data. Master's thesis, Carnegie Mellon University, 2015.

Powell, J. L. Estimation of semiparametric models. In *Handbook of Econometrics*, volume 4, pp. 2443–2521. 1994.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. The MIT Press, 2009.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J., and Dennison, D. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2494–2502, 2015.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244, 2000.

Sion, M. On general minimax theorems. *Pacific journal of mathematics*, 8(1):171–176, 1958.

Tomizawa, N. On some techniques useful for solution of transportation network problems. *Networks*, 1(2):173–194, 1971.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

Tsirelson, B. S., Ibragimov, I. A., and Sudakov, V. N. Norms of Gaussian sample functions. In *Proceedings of the Third Japan-USSR Symposium on Probability Theory*, pp. 20–41, 1976.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *arXiv*, 2014.

A. Details of Computing $\tilde{R}$ from μ

In this section we show how, given $\bar{A}$, μ, and π, we can efficiently compute

$$\tilde{R}(\theta) = \bar{A}(\theta) - \max_{\sigma \in \mathrm{Sym}(k)} \sum_{j=1}^{k} \pi_{\sigma(j)} \sum_{v=1}^{3} (\mu_{\sigma(j),v})_j. \tag{17}$$

The only bottleneck is the maximum over $\sigma \in \mathrm{Sym}(k)$, which would naïvely require considering $k!$ possibilities. However, we can instead cast this as a form of maximum matching. In particular, form the $k \times k$ matrix

$$X_{i,j} = \pi_i \sum_{v=1}^{3} (\mu_{i,v})_j. \tag{18}$$

Then we are looking for the permutation σ such that $\sum_{j=1}^{k} X_{\sigma(j),j}$ is maximized. If we consider each $X_{i,j}$ to be the weight of edge (i,j) in a complete bipartite graph, then this is equivalent to asking for a matching of i to j with maximum weight, hence we can maximize over σ using any maximum-weight matching algorithm such as the Hungarian algorithm, which runs in $\mathcal{O}\left(k^3\right)$ time (Tomizawa, 1971; Edmonds & Karp, 1972).

B. Proof of Theorem 2.2

Preliminary reductions. Our goal is to estimate $\tilde{R}$ to error ϵ (with probability of failure $1-\delta$) in $\mathrm{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right) \cdot \frac{\log(1/\delta)}{\epsilon^2}$ samples. Since $\tilde{R}$ is a scalar parameter, if we can estimate $\tilde{R}$ to error ϵ with any fixed probability of success $1 - \delta_0 > 1/2$, then by taking the median of $\mathcal{O}\left(\log(1/\delta)/\delta_0\right)$ independent estimates, we can amplify the probability of success to $1-\delta$. In this argument we will focus on achieving a probability of success of $3/4$.

Letting $R_{\mathrm{linear}}^{\sigma}$ be the linear part of the risk if the labels are permuted by σ, note that

$$\tilde{R} = \bar{A} - \min_{\sigma} R_{\mathrm{linear}}^{\sigma} \tag{19}$$

$$= \bar{A} - \min_{\sigma} \sum_{j=1}^{k} \pi_{\sigma(j)} \sum_{v=1}^{3} (\mu_{\sigma(j),v})_j \tag{20}$$

Note that $0 \le \pi_{\sigma(j)} \le 1$ and $-\tau \le (\mu_{\sigma(j),v})_j \le \tau$. Therefore, as long as each of the quantities ($\bar{A}$, π, μ) can be estimated to error ϵ in $\mathrm{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right)/\epsilon^2$ samples, by taking an appropriately scaled down ϵ (roughly $k\tau$ times smaller to account for the product terms of π with μ), we can estimate the entire expression for $\tilde{R}$ to within error ϵ as well. We therefore focus the remainder of the argument on estimating $\bar{A}$, π, and μ individually.

Estimatimg $\bar{A}$. Remember that $\bar{A} = \mathbb{E}_x[A(\theta; x)]$, and that $\mathbb{E}_x[A^2] \le \tau^2$ by assumption. Therefore, $\bar{A}$ can be estimated to error ϵ (with probability, say, $11/12$) using $\mathrm{poly}(\tau)/\epsilon^2$ samples.

Estimating μ. Estimating π and μ is mostly an exercise in interpreting Theorem 7 of Anandkumar et al. (2012), which we recall below, modifying the statement slightly to fit our language.

Theorem B.1 (Anandkumar et al. (2012)). *Let $P_{v,v'} \overset{\mathrm{def}}{=} \mathbb{E}[h_v(x) \otimes h_{v'}(x)]$, and $P_{1,2,3} \overset{\mathrm{def}}{=} \mathbb{E}[h_1(x) \otimes h_2(x) \otimes h_3(x)]$. Also let $\hat{P}_{v,v'}$ and $\hat{P}_{1,2,3}$ be sample estimates of $P_{v,v'}$, $P_{1,2,3}$ that are (for technical convenience) estimated from independent samples of size m. Let $\|T\|_F$ denote the ℓ^2-norm of T after unrolling T to a vector. Suppose that:*

- $\mathbb{P}\left[\|\hat{P}_{v,v'} - P_{v,v'}\|_2 \le C_{v,v'}\sqrt{\frac{\log(1/\delta)}{m}}\right] \ge 1-\delta$ *for* $\{v,v'\} \in \{\{1,2\},\{1,3\}\}$, *and*
- $\mathbb{P}\left[\|\hat{P}_{1,2,3} - P_{1,2,3}\|_F \le C_{1,2,3}\sqrt{\frac{\log(1/\delta)}{m}}\right] \ge 1-\delta.$

Then, there exists constants C, m_0, δ_0 such that the following holds: if $m \geq m_0$ and $\delta \leq \delta_0$ and

$$\sqrt{\frac{\log(k/\delta)}{m}} \leq C \cdot \frac{\min_{j \neq j'} \|\mu_{j,3} - \mu_{j',3}\|_2 \cdot \sigma_k(P_{1,2})}{C_{1,2,3} \cdot k^5 \cdot \kappa(M_1)^4} \cdot \frac{\delta}{\log(k/\delta)} \cdot \epsilon,$$
$$\sqrt{\frac{\log(1/\delta)}{m}} \leq C \cdot \min\left\{ \frac{\min_{j \neq j'} \|\mu_{j,3} - \mu_{j',3}\|_2 \cdot \sigma_k(P_{1,2})^2}{C_{1,2} \cdot \|P_{1,2,3}\|_F \cdot k^5 \cdot \kappa(M_1)^4} \cdot \frac{\delta}{\log(k/\delta)}, \frac{\sigma_k(P_{1,3})}{C_{1,3}} \right\} \cdot \epsilon,$$

then with probability at least $1 - 5\delta$, we can output $\hat{M}_3 = [\hat{\mu}_{1,3} \ \cdots \ \hat{\mu}_{k,3}]$ with the following guarantee: there exists a permutation $\sigma \in \mathrm{Sym}(k)$ such that for all $j \in \{1, \dots, k\}$,

$$\|\hat{\mu}_{j,3} - \mu_{\sigma(j),3}\|_2 \leq \max_{j'} \|\mu_{j',3}\|_2 \cdot \epsilon. \tag{21}$$

By symmetry, we can use Theorem B.1 to recover each of the matrices M_v, $v = 1, 2, 3$, up to permutation of the columns. Furthermore, Anandkumar et al. (2012) show in Appendix B.4 of their paper how to match up the columns of the different M_v, so that only a single unknown permutation is applied to each of the M_v simultaneously. We will set $\delta = 1/180$, which yields an overall probability of success of $11/12$ for this part of the proof.

We now analyze the rate of convergence implied by Theorem B.1. Note that we can take $C_{1,2,3} = \mathcal{O}\left(\sqrt{\mathbb{E}[\|h_1\|_2^2 \|h_2\|_2^2 \|h_3\|_2^2]}\right)$, and similarly $C_{v,v'} = \mathcal{O}\left(\sqrt{\mathbb{E}[\|h_v\|_2^2 \|h_{v'}\|_2^2]}\right)$. Then, since we only care about polynomial factors, it is enough to note that we can estimate the M_v to error ϵ given Z/ϵ^2 samples, where Z is polynomial in the following quantities:

1. k,

2. $\max_{v=1}^3 \kappa(M_v)$, where κ denotes condition number,

3. $\frac{\sqrt{\mathbb{E}[\|h_1\|_2^2 \|h_2\|_2^2 \|h_3\|_2^2]}}{\left(\min_{j,j'} \|\mu_{j,v} - \mu_{j',v}\|_2\right) \cdot \sigma_k(P_{v',v''})}$, where (v, v', v'') is a permutation of $(1, 2, 3)$,

4. $\frac{\|P_{1,2,3}\|_2}{\left(\min_{j,j'} \|\mu_{j,v} - \mu_{j',v}\|_2\right) \cdot \sigma_k(P_{v',v''})}$, where (v, v', v'') is as before, and

5. $\frac{\sqrt{\mathbb{E}[\|h_v\|_2^2 \|h_{v'}\|_2^2]}}{\sigma_k\left(P_{v,v'}\right)}$.

6. $\max_{j,v} \|\mu_{j,v}\|_2$.

It suffices to show that each of these quantities are polynomial in k, $\pi_{\min}^{-1}$, τ, and λ^{-1}.

(1) k is trivially polynomial in itself.

(2) Note that $\kappa(M_v) \leq \sigma_1(M_v)/\lambda \leq \|M_v\|_F/\lambda$. Furthermore, $\|M_v\|_F^2 = \sum_j \|\mathbb{E}[h_v \mid j]\|_2^2 \leq \sum_j \mathbb{E}[\|h_v\|_2^2 \mid j] \leq k\tau^2$. In all, $\kappa(M_v) \leq \sqrt{k}\tau/\lambda$, which is polynomial in k and τ/λ.

(3) We first note that $\min_{j \neq j'} \|\mu_{j,v} - \mu_{j',v}\|_2 = \sqrt{2} \min_{j \neq j'} \|M_v(e_j - e_{j'})\|_2 / \|e_j - e_{j'}\|_2 \geq \sigma_k(M_v)$. Also, $\sigma_k(P_{v',v''}) = \sigma_k(M_{v'} \operatorname{diag}(\pi) M_{v''}) \geq \sigma_k(M_{v'}) \pi_{\min} \sigma_k(M_{v''})$. We can thus upper-bound the quantity in (3.) by

$$\frac{\sqrt{\mathbb{E}[\|h_1\|_2^2 \|h_2\|_2^2 \|h_3\|_2^2]}}{\sqrt{2}\pi_{\min} \sigma_k(M_1) \sigma_k(M_2) \sigma_k(M_3)} \leq \frac{\tau^3}{\sqrt{2}\pi_{\min} \lambda^3},$$

which is polynomial in $\pi_{\min}^{-1}$, τ/λ.

(4) We can perform the same calculations as in (3), but now we have to bound $\|P_{1,2,3}\|_2$. However, it is easy to see that

$$\begin{aligned}\|P_{1,2,3}\|_2 &= \sqrt{\|\mathbb{E}[h_1 \otimes h_2 \otimes h_3]\|_2^2} \\ &\leq \sqrt{\mathbb{E}[\|h_1 \otimes h_2 \otimes h_3\|_2^2]} \\ &= \sqrt{\mathbb{E}[\|h_1\|_2^2 \|h_2\|_2^2 \|h_3\|_2^2]} \\ &= \sqrt{\sum_{j=1}^{k} \pi_j \prod_{v=1}^{3} \mathbb{E}[\|h_v\|_2^2 \mid y = j]} \\ &\leq \tau^3,\end{aligned}$$

which yields the same upper bound as in (3).

(5) We can again perform the same calculations as in (3), where we now only have to deal with a subset of the variables, thus obtaining a bound of $\frac{\tau^2}{\pi_{\min}\lambda^2}$.

(6) We have $\|\mu_{j,v}\|_2 \leq \sqrt{\mathbb{E}[\|h_v\|_2^2 \mid y = j]} \leq \tau$.

In sum, we have shown that with probability $\frac{11}{12}$ we can estimate each $\mu_{j,v}$ to ℓ^2 error ϵ using $\mathrm{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right)/\epsilon^2$ samples. It now remains to estimate π.

Estimating π. This part of the argument follows Appendix B.5 of Anandkumar et al. (2012). Noting that $\pi = M_1^{-1}\mathbb{E}[h_1]$, we can estimate π as $\hat{\pi} = \hat{M_1}^{-1}\hat{\mathbb{E}}[h_1]$, where $\hat{\mathbb{E}}$ denotes the empirical expectation. Hence, we have

$$\begin{aligned}\|\pi - \hat{\pi}\|_\infty &\leq \left\|(\hat{M_1}^{-1} - M_1^{-1})\mathbb{E}[h_1] + M_1^{-1}(\hat{\mathbb{E}}[h_1] - \mathbb{E}[h_1]) + (\hat{M_1}^{-1} - M_1^{-1})(\hat{\mathbb{E}}[h_1] - \mathbb{E}[h_1])\right\|_\infty \\ &\leq \underbrace{\|\hat{M_1}^{-1} - M_1^{-1}\|_F}_{(i)} \underbrace{\|\mathbb{E}[h_1]\|_2}_{(ii)} + \underbrace{\|M_1^{-1}\|_F}_{(iii)} \underbrace{\|\hat{\mathbb{E}}[h_1] - \mathbb{E}[h_1]\|_2}_{(iv)} + \underbrace{\|\hat{M_1}^{-1} - M_1^{-1}\|_F}_{(i)} \underbrace{\|\hat{\mathbb{E}}[h_1] - \mathbb{E}[h_1]\|_2}_{(iv)}.\end{aligned}$$

We will bound each of these factors in turn:

(i) $\|\hat{M_1}^{-1} - M_1^{-1}\|_F$: let $E_1 = \hat{M_1} - M_1$, which by the previous part satisfies $\|E_1\|_F \leq \sqrt{k}\max_j \|\hat{\mu}_{j,1} - \mu_{j,1}\|_2 = \mathrm{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right)/\sqrt{m}$. Therefore:

$$\begin{aligned}\|\hat{M_1}^{-1} - M_1^{-1}\|_F &\leq \|(M_1 + E_1)^{-1} - M_1^{-1}\|_F \\ &= \|M_1^{-1}(I + E_1 M_1^{-1})^{-1} - M_1^{-1}\|_F \\ &\leq \|M_1^{-1}\|_F \|(I + E_1 M_1^{-1})^{-1} - I\|_F \\ &\leq k\lambda^{-1}\sigma_1\left(I + E_1 M_1^{-1})^{-1} - I\right) \\ &\leq k\lambda^{-1}\frac{\sigma_1(E_1 M_1^{-1})}{1 - \sigma_1(E_1 M_1^{-1})} \\ &\leq k\lambda^{-2}\frac{\|E_1\|_F}{1 - \lambda^{-1}\|E_1\|_F} \\ &\leq \frac{\mathrm{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right)}{1 - \mathrm{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right)/\sqrt{m}} \cdot \frac{1}{\sqrt{m}}.\end{aligned}$$

We can assume that $m \geq \mathrm{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right)$ without loss of generality (since otherwise we can trivially obtain the desired bound on $\|\pi - \hat{\pi}\|_\infty$ by simply guessing the uniform distribution), in which case the above quantity is $\mathrm{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right) \cdot \frac{1}{\sqrt{m}}$.

(ii) $\|\mathbb{E}[h_1]\|_2$: we have $\|\mathbb{E}[h_1]\|_2 \leq \sqrt{\mathbb{E}[\|h_1\|_2^2]} \leq \tau$.

(iii) $\|M_1^{-1}\|_F$: since $\|X\|_F \leq \sqrt{k}\sigma_1(F)$, we have $\|M_1^{-1}\|_F \leq \sqrt{k}\lambda^{-1}$.

(iv) $\|\hat{\mathbb{E}}[h_1] - \mathbb{E}[h_1]\|_2$: with any fixed probability (say $11/12$), this term is $\mathcal{O}\left(\sqrt{\frac{\mathbb{E}[\|h_1\|_2^2]}{m}}\right) = \mathcal{O}\left(\frac{\tau}{\sqrt{m}}\right)$.

In sum, with probability at least $\frac{11}{12}$ all of the terms are poly $\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right)$, and at least one factor in each term has a $\frac{1}{\sqrt{m}}$ decay. Therefore, we have $\|\pi - \hat{\pi}\|_\infty \leq$ poly $\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right) \cdot \sqrt{\frac{1}{m}}$.

Since we have shown that we can estimate each of $\bar{A}$, π, and μ individually with probability $\frac{11}{12}$, we can also estimate $\tilde{R}$ with probability $\frac{3}{4}$, thus completing the proof.

C. Proofs of Theorem 3.1 and Corollary 3.2

Proof of Theorem 3.1. First, we note that Theorem 2.2 can be extended to the case where instead of just estimating μ, we also estimate $g_t = u_t^\top \nabla_\theta R_{\text{linear}}$. Since re-proving Theorem 3.1 in this case would be tedious, we simply highlight the main differences.

First, M is now a $2k \times k$ matrix rather than a $k \times k$ matrix; moreover, since all we do to M is add k additional rows, the minimum singular value λ can only increase (which would decrease the sample complexity). On the other hand, τ could increase to $\tau^2 + B^2$ once we add in $u^\top \nabla_\theta h_v$. To deal with this, we first scale $u^\top \nabla_\theta f_v$ down by a factor of B (and then scale it up afterwards), which is why our final error bound will depend on $B\epsilon$ rather than just ϵ. Finally, $\pi_{\min}$ and k remain unchanged. We note that we need the slightly stronger guarantee that $R^\sigma(\theta)$ and $u_t^\top \nabla_\theta R_{\text{linear}}^\sigma$ are uniformly well-estimated for all σ, which is a straightforward consequence of Theorem B.1.

Now, we can assume that for all $t = 1, \dots, T$ (where $T = \mathcal{O}\left(d \log(d/\delta)/\epsilon^2\right)$), we successfully estimate each R^σ to within error $\epsilon/5$ with probability $1 - \delta/2$ (where we union bound over t, using the fact that we the number of samples m is required to grow with $\log(T/\delta)$). Since $\epsilon < \text{gap}(\theta_0)$, we thus necessarily recover the correct permutation σ, and can undo it to obtain $\hat{g}_t$ such that $|\hat{g}_t - u_t^\top \nabla_\theta \tilde{R}_{\text{linear}}| \leq \epsilon'$, where $\epsilon' = B\epsilon/5$. Let $\epsilon_t = \hat{g}_t - u_t^\top \nabla_\theta \tilde{R}_{\text{linear}}$. We then have

$$\begin{aligned}
\left\|\frac{1}{T}\sum_{t=1}^T u_t \hat{g}_t - \nabla_\theta \tilde{R}_{\text{linear}}\right\|_2 &\leq \left\|\frac{1}{T}\sum_{t=1}^T u_t u_t^\top \nabla_\theta \tilde{R}_{\text{linear}} - \nabla_\theta \tilde{R}_{\text{linear}}\right\|_2 + \left\|\frac{1}{T}\sum_{t=1}^T u_t (u_t^\top \nabla_\theta \tilde{R}_{\text{linear}} - \hat{g}_t)\right\|_2 \\
&= \left\|\frac{1}{T}\sum_{t=1}^T u_t u_t^\top - I\right\|_{\text{op}} \left\|\nabla_\theta \tilde{R}_{\text{linear}}\right\|_2 + \frac{1}{T}\left\|\sum_{t=1}^T \epsilon_t u_t\right\|_2 \\
&\leq \left\|\frac{1}{T}\sum_{t=1}^T u_t u_t^\top - I\right\|_{\text{op}} \left\|\nabla_\theta \tilde{R}_{\text{linear}}\right\|_2 + \frac{1}{T}\|[u_1 \ \cdots \ u_T]\|_{\text{op}} \|\epsilon_{1:T}\|_2 \\
&\leq 3B\left\|\frac{1}{T}\sum_{t=1}^T u_t u_t^\top - I\right\|_{\text{op}} + \epsilon' \left\|\frac{1}{T}\sum_{t=1}^T u_t u_t^\top\right\|_{\text{op}}^{\frac{1}{2}}.
\end{aligned}$$

Now, by Theorem 1.2 of Tropp (2012), we have for $\epsilon < 1$ that

$$\mathbb{P}\left[\sigma_{\max}\left(\frac{1}{T}\sum_{t=1}^T u_t u_t^\top\right) \geq 1 + \epsilon/5\right] \leq d \exp\left(-\frac{\epsilon^2 T}{75d}\right). \tag{22}$$

Thus, setting $T = \frac{75d}{\epsilon^2}\log(2d/\delta)$, we get that, with probability $1 - \delta/2$, we have

$$\left\|\frac{1}{T}\sum_{t=1}^T u_t \hat{g}_t - \nabla_\theta \tilde{R}_{\text{linear}}\right\|_2 \leq 3B\epsilon/5 + 2\epsilon' = B\epsilon, \tag{23}$$

as was to be shown. □

Proof of Corollary 3.2. Since $\theta_0 \in \Theta_0$, we know that $\nabla_{\theta_0} \tilde{R}_{\text{linear}} = \nabla_{\theta_0} R_{\text{linear}}$. Now note that $\theta^\top \left(\hat{g} - \nabla_{\theta_0} R_{\text{linear}}\right) \leq \|\theta\|_2 \|\hat{g} - \nabla_{\theta_0} R_{\text{linear}}\|_2 \leq \epsilon B \rho$ for all θ with $\|\theta\|_2 \leq \rho$. In addition, by standard Rademacher complexity bounds we have with probability $1 - \delta$ that $\|\bar{A}(\theta) - \frac{1}{m}\sum_{i=1}^m A(\theta; x^{(i)})\|_2 \leq \epsilon B \rho$ for all $\|\theta\|_2 \leq \rho$, provided that $m = \Omega\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, since $A(\theta; x^{(i)})$ is B-Lipschitz. Putting these together, the result follows by a standard uniform convergence argument. □

D. Proof of Lemma 3.3

Recall that we are given $\theta_0 \in \Theta_0$, and θ such that $\mathbb{E}_{x \sim p^*}\left[\max_{j=1}^{k}|f(\theta; x, j) - f(\theta_0; x, j)|\right] \leq \frac{1}{2}\left(\mathrm{gap}(\theta) + \mathrm{gap}(\theta_0)\right)$. Our goal is to show that $\theta \in \Theta_0$ as well.

By the definition of $\mathrm{gap}(\theta_0)$, we have $R^\sigma(\theta_0) \geq R(\theta_0) + \mathrm{gap}(\theta_0)$ for all $\sigma \neq \mathrm{id}$. Assume for the sake of contradiction that $\tilde{R}(\theta) \neq R(\theta)$; then again by the definition of gap, there exists a $\sigma \neq \mathrm{id}$ with $R^\sigma(\theta) \leq R(\theta) - \mathrm{gap}(\theta)$.

To derive a contradiction, we will study R_{linear} and $R^\sigma_{\mathrm{linear}}$. We have that

$$\begin{aligned}
R_{\mathrm{linear}}(\theta_0) &\leq R^\sigma_{\mathrm{linear}}(\theta_0) - \mathrm{gap}(\theta_0) && (24)\\
&\leq |R^\sigma_{\mathrm{linear}}(\theta_0) - R^\sigma_{\mathrm{linear}}(\theta)| + R^\sigma_{\mathrm{linear}}(\theta) - \mathrm{gap}(\theta_0) && (25)\\
&\leq |R^\sigma_{\mathrm{linear}}(\theta_0) - R^\sigma_{\mathrm{linear}}(\theta)| + R_{\mathrm{linear}}(\theta) - (\mathrm{gap}(\theta_0) + \mathrm{gap}(\theta)) && (26)\\
&\leq |R^\sigma_{\mathrm{linear}}(\theta_0) - R^\sigma_{\mathrm{linear}}(\theta)| + |R_{\mathrm{linear}}(\theta_0) - R_{\mathrm{linear}}(\theta)| + R_{\mathrm{linear}}(\theta_0) - (\mathrm{gap}(\theta_0) + \mathrm{gap}(\theta)). && (27)
\end{aligned}$$

On the other hand, for *all* permutations σ we have:

$$\begin{aligned}
|R^\sigma_{\mathrm{linear}}(\theta_0) - R^\sigma_{\mathrm{linear}}(\theta)| &= \left|\sum_{j=1}^{k} \pi_j \mathbb{E}[f(\theta_0; x, \sigma(j)) - f(\theta; x, \sigma(j)) \mid y = j]\right| && (28)\\
&\leq \sum_{j=1}^{k} \pi_j \mathbb{E}\left[\max_{j'=1}^{k} |f(\theta_0; x, j') - f(\theta; x, j')| \;\middle|\; y = j\right] && (29)\\
&= \mathbb{E}\left[\max_{j'=1}^{k} |f(\theta_0; x, j') - f(\theta; x, j')|\right] && (30)\\
&< \frac{1}{2}\left(\mathrm{gap}(\theta) + \mathrm{gap}(\theta_0)\right). && (31)
\end{aligned}$$

Substituting this in for both $R^\sigma_{\mathrm{linear}}$ and $R_{\mathrm{linear}} = R^{\mathrm{id}}_{\mathrm{linear}}$ in (27) above yields the desired contradiction and hence completes the proof.

E. Proof of Proposition 4.3

First, suppose that $p^* \sim \nu'$. Note that $(\mu_j)_j - (\mu_j)_{j'}$ measures how much larger $\mathbb{E}[-\log p_\theta(j \mid x) \mid y = j]$ would become if we assign class j to label j'. Therefore, the condition $\gamma(\theta) > 0$ implies that $\mathbb{E}[-\log p_\theta(j' \mid x) \mid y = j] > \mathbb{E}[-\log p_\theta(j \mid x) \mid y = j]$ for all $j \neq j'$. Therefore, any $\sigma \neq \mathrm{id}$ will have $R^\sigma(\theta) > R(\theta)$, and so $\tilde{R}(\theta) = R(\theta)$.

Next, suppose that $p^* \sim \nu$. Then, we have

$$\begin{aligned}
\mathbb{P}[\tilde{R}(\theta) \leq r_0] &= \mathbb{P}[\min_\sigma R^\sigma(\theta) \leq r_0] && (32)\\
&= \mathbb{P}[\min_\sigma \mathbb{E}[-\log p_\theta(\sigma(y) \mid x)] \leq r_0] && (33)\\
&\leq \mathbb{P}[\min_\sigma \mathrm{D}\left(p_\theta(\sigma(y) \mid x) \parallel p^*(y \mid x)\right) \leq r_0] && (34)\\
&\leq \sum_\sigma \mathbb{P}[\mathrm{D}\left(p_\theta(\sigma(y) \mid x) \parallel p^*(y \mid x)\right) \leq r_0] && (35)\\
&\leq k!\, \alpha_{r_0}. && (36)
\end{aligned}$$

Putting these together, we have

$$\max_{\hat{\nu} \in \{\nu, \nu'\}} \mathbb{P}_{\hat{\nu}}[\tilde{R}(\theta) \leq r_0 \wedge \tilde{R}(\theta) \neq R(\theta)] \leq k!\, \alpha_{r_0}, \tag{37}$$

as was to be shown.

F. Extension of Proposition F.1

In this section, we establish the following extension of Proposition F.1:

Proposition F.1. *Suppose that $p^* \sim \nu$. Then for any $0 < \epsilon < 2$, we have the uniform false negative bound*

$$\mathbb{P}[\exists \theta, \tilde{R}(\theta) < r_0] \leq k! \, N_\Theta((\epsilon/4\sqrt{k})^4)\alpha_{r_0+\epsilon}, \tag{38}$$

where N_Θ is the covering number of Θ under $\mathrm{D}\left(\cdot \parallel \cdot\right)$.

Note that N_Θ is typically exponential in the dimension of Θ; comparing to e.g. Theorem 4.4 below, we thus need Θ to have smaller dimension than the model family of p^* for (38) to be non-trivial (which could be true in e.g. transfer learning settings where we tune a subset of the parameters to a new domain).

Proof of Proposition F.1. We want to bound

$$\mathbb{P}[\exists \theta \in \Theta, \tilde{R}(\theta) \leq r_0]. \tag{39}$$

To start, let $\mathcal{Q}$ be a minimal covering of Θ under $\mathrm{D}\left(\cdot \parallel \cdot\right)$ of radius ϵ_0, where ϵ_0 will be chosen later. We will replace each $q \in \mathcal{Q}$ with the distribution $\overline{q} = (1-\delta)q + \delta u$, where u is the uniform distribution and $\delta > 0$ will also be chosen later. Let $\overline{\mathcal{Q}}$ denote this new set of distributions.

By the same logic as Proposition 4.3, the probability that $\min_{q \in \overline{\mathcal{Q}}} \tilde{R}(q) \leq r_0 + \epsilon$ is at most $k! \, |\mathcal{Q}| \alpha_{r_0+\epsilon}$. We would now like to use this to say something about Θ. We will be able to do so with the following lemma:

Lemma F.2. *Suppose that $\mathrm{D}\left(p \parallel q\right) \leq \epsilon_0$. Then, for all distributions p^*,*

$$\mathrm{D}\left(p^* \parallel \overline{q}\right) \leq \mathrm{D}\left(p^* \parallel p\right) + \log\left(\frac{1}{1-\delta}\right) + (k/\delta)\sqrt{\epsilon_0/2}. \tag{40}$$

Roughly, this says: "if q covers p and p covers p^*, then $\overline{q}$ covers p^*". For our purposes, we will set $\delta = \sqrt{k}(\epsilon_0/8)^{1/4}$. Assuming that $\delta \leq 1/2$, we have the bound $\log \frac{1}{1-\delta} \leq 2\delta$, so that Lemma F.2 would imply the bound $\mathrm{D}\left(p^* \parallel \overline{q}\right) \leq \mathrm{D}\left(p^* \parallel p\right) + 2\sqrt{k}(2\epsilon_0)^{1/4}$.

To apply Lemma F.2, consider the event E that $\min_{\theta \in \Theta} \tilde{R}(\theta) \leq r_0$. In the following argument, we will let p^σ denote the distribution with $p^\sigma(y \mid x) = p(\sigma(y) \mid x)$.

The event E implies that there exists p_θ, σ with $\mathrm{D}\left(p^* \parallel p_\theta^\sigma\right) \leq r_0$, and because $\mathcal{Q}$ covers Θ, there is also a q with $\mathrm{D}\left(p_\theta \parallel q\right) \leq \epsilon_0$ (and hence also $\mathrm{D}\left(p_\theta^\sigma \parallel q^\sigma\right) \leq \epsilon_0$). Invoking Lemma F.2, there is thus a $\overline{q} \in \overline{\mathcal{Q}}$ such that $\mathrm{D}\left(p^* \parallel \overline{q}^\sigma\right) \leq r_0 + 2\sqrt{k}(2\epsilon_0)^{1/4}$. We will now choose $\epsilon_0 = (\epsilon/4\sqrt{k})^4$, which is enough to ensure that $\mathrm{D}\left(p^* \parallel \overline{q}^\sigma\right) \leq r_0 + \epsilon$; hence $\mathbb{P}[E] \leq \mathbb{P}[\min_{q \in \overline{\mathcal{Q}}} \tilde{R}(q) \leq r_0 + \epsilon] \leq k! \, |\mathcal{Q}| \alpha_{r_0+\epsilon}$, as was to be shown. □

Two remaining details are to check that $\delta < 1/2$, and to prove Lemma F.2. For the first, since $\delta = \sqrt{k}(\epsilon_0/8)^{1/4} < \sqrt{k}(\epsilon/4\sqrt{k}) = \epsilon/4$, it suffices to constrain $\epsilon < 2$, which we assumed. We now turn to Lemma F.2.

Proof of Lemma F.2. We will first prove the equivalent result for KL divergence $\mathrm{KL}\left(\cdot \parallel \cdot\right)$ (instead of the averaged KL divergence $\mathrm{D}\left(\cdot \parallel \cdot\right)$). Note that

$$\begin{aligned}
\mathrm{KL}\left(p^* \parallel \overline{q}\right) - \mathrm{KL}\left(p^* \parallel \overline{p}\right) &= \mathbb{E}_{p^*}\left[\log \frac{\overline{p}(x)}{\overline{q}(x)}\right] \\
&\leq \mathbb{E}_{p^*}\left[\max\left(\log \frac{\overline{p}(x)}{\overline{q}(x)}, 0\right)\right] \\
&\overset{(i)}{\leq} \frac{k}{\delta}\mathbb{E}_{\overline{q}}\left[\max\left(\log \frac{\overline{p}(x)}{\overline{q}(x)}, 0\right)\right] \\
&\overset{(ii)}{\leq} \frac{k}{\delta}\mathbb{E}_{\overline{q}}\left[\max\left(\frac{\overline{p}(x) - \overline{q}(x)}{\overline{q}(x)}, 0\right)\right] \\
&= \frac{k}{\delta}\|\overline{p} - \overline{q}\|_{TV} \\
&\leq \frac{k}{\delta}\|p - q\|_{TV},
\end{aligned}$$

where (i) uses the fact that $p^* \leq (k/\delta)\overline{q}$ and (ii) uses $\log(a/b) \leq (a-b)/b$ for $a \geq b$. We then note that $\|p-q\|_{TV}\leq \sqrt{\mathrm{KL}\left(p \parallel q\right)/2}$ by Pinsker's inequality, and also that $\mathrm{KL}\left(p^* \parallel \overline{p}\right) \leq \int p^*(x)\log\frac{p^*(x)}{(1-\delta)p(x)}dx \leq \mathrm{KL}\left(p^* \parallel p\right) + \log\left(\frac{1}{1-\delta}\right)$. Averaging over $x \sim \pi$, we have

$$\begin{aligned} \mathrm{D}\left(p^* \parallel \overline{q}\right) &\leq \mathrm{D}\left(p^* \parallel p\right) + \log\left(\frac{1}{1-\delta}\right) + (k/\delta)\mathbb{E}_{x\sim\pi}\left[\sqrt{\mathrm{KL}\left(p(Y \mid x) \parallel q(Y \mid x)\right)}\right] && (41) \\ &\leq \mathrm{D}\left(p^* \parallel p\right) + \log\left(\frac{1}{1-\delta}\right) + (k/\delta)\sqrt{\mathrm{D}\left(p \parallel q\right)/2}, && (42) \end{aligned}$$

which completes the proof. □

G. Justification of Some Details for Theorem 4.4

G.1. Treating p_β as a point mass

Here we justify why, for sufficiently large τ, if each $\beta_j \sim \mathcal{N}(0, \cdot I_{d\times d})$, then we can treat $p_{\beta,\tau}(Y = j \mid x) \propto \exp\left(\tau\beta_j^\top\phi(x)\right)$ as a point mass. Precisely, we have the following result:

Lemma G.1. *Let $\beta_j \sim \mathcal{N}(0, I_{d\times d})$. For any τ, let*

$$p_{\beta,\tau}(Y = j \mid x) \propto \exp\left(\tau \cdot \beta_j^\top\phi(x)\right), \text{ and let } p_{\beta,\infty}(Y = j \mid x) \stackrel{\text{def}}{=} \mathbb{I}\left[j = \arg\max_{j'} \beta_{j'}^\top\phi(x)\right]. \tag{43}$$

Then, if $\mathcal{F}$ is the family of functions $f : \mathcal{X} \times \mathcal{Y} \to [-1, 1]$, we have, for every $\delta > 0$,

$$\lim_{\tau\to\infty} \mathbb{P}_{\beta_{1:k}\sim\mathcal{N}(0,I)}\left[\sup_{f\in\mathcal{F}} \mathbb{E}_{p_{\beta,\tau}}[f(x,y)] - \mathbb{E}_{p_{\beta,\infty}}[f(x,y)] > \delta\right] = 0 \tag{44}$$

In particular, since in light of (13) we only need to worry about the expectation of $q(y \mid x)$, which itself lies in $[0, 1]$, for any desired error threshold δ the probability that the difference between $\mathbb{E}_{p_{\beta,\tau}}[q(y \mid x)]$ and $\mathbb{E}_{p_{\beta,\infty}}[q(y \mid x)]$ exceeds δ goes to 0 uniformly (in q) as $\tau \to \infty$. Since the constants we provide in Theorem 4.4 are slightly larger than those actually implied by our proofs, there is some non-infinite value of τ that implies Theorem 4.4 with the stated constants, even though we otherwise only establish the result for $p_{\beta,\infty}$.

Proof of Lemma G.1. Note that the supremum in (44) is simply the expected total variational distance $\mathbb{E}_{x\sim\pi}\left[\|p_{\beta,\tau}(Y \mid x) - p_{\beta,\infty}(Y \mid x)\|_{TV}\right]$. We then have

$$\begin{aligned} &\lim_{\tau\to\infty} \mathbb{P}_{\beta_{1:k}\sim\mathcal{N}(0,I)}\left[\mathbb{E}_{x\sim\pi}\left[\|p_{\beta,\tau}(Y \mid x) - p_{\beta,\infty}(Y \mid x)\|_{TV}\right] > \delta\right] && (45) \\ &\quad\leq \delta^{-1} \lim_{\tau\to\infty} \mathbb{E}_{\beta_{1:k},x}\left[\|p_{\beta,\tau}(Y \mid x) - p_{\beta,\infty}(Y \mid x)\|_{TV}\right] && (46) \\ &\quad\stackrel{(i)}{=} 0. && (47) \end{aligned}$$

To justify (i), note that since $\phi(x) \neq 0$ almost surely, $p_{\beta,\tau}(y \mid x) \to p_{\beta,\infty}(y \mid x)$ almost surely as $\tau \to \infty$. Since total variational distance is bounded by 1, by Lebesgue's dominated convergence theorem, the limit of the expectation is indeed zero. □

G.2. Approximating π by samples

Applying Lebesgue's dominated convergence theorem at (i) below, we have

$$\lim_{m\to\infty} \mathbb{P}_{\beta,x_{1:m}}\left[\frac{1}{m}\sum_{i=1}^{m} q(y_\beta(x_i)\mid x_i) \geq \frac{1+\epsilon}{k}\right] = \lim_{m\to\infty} \mathbb{E}_{\beta,x_1,x_2,\dots}\left[\mathbb{I}\left[\frac{1}{m}\sum_{i=1}^{m} q(y_\beta(x_i)\mid x_i) \geq \frac{1+\epsilon}{k}\right]\right] \tag{48}$$

$$\overset{(i)}{=} \mathbb{E}_{\beta,x_1,x_2,\dots}\left[\lim_{m\to\infty} \mathbb{I}\left[\frac{1}{m}\sum_{i=1}^{m} q(y_\beta(x_i)\mid x_i) \geq \frac{1+\epsilon}{k}\right]\right] \tag{49}$$

$$= \mathbb{E}_{\beta,x_1,x_2,\dots}\left[\mathbb{I}\left[\mathbb{E}_{x\sim\pi}[q(y_\beta(x)\mid x)] \geq \frac{1+\epsilon}{k}\right]\right] \tag{50}$$

$$= \mathbb{P}_{\beta}\left[\mathbb{E}_{x\sim\pi}[q(y_\beta(x)\mid x)] \geq \frac{1+\epsilon}{k}\right]. \tag{51}$$

G.3. Computing Lipschitz constants

Recall we want to show that the map $\beta_{1:k} \mapsto t_{1:k}^{1:m}$ is $\|\Phi\|_{\mathrm{op}}$-Lipschitz, and that the map $(q(y_\beta(x_i) \mid x_i))_{i=1}^m \mapsto \frac{1}{m}\sum_{i=1}^m q(y_\beta(x_i) \mid x_i)$ is $(1/\sqrt{m})$-Lipschitz.

In the first case, remember that t_j^i is defined to be $\beta_j^\top \phi(x_i)/\|\phi(x_i)\|_2$. Therefore, the map $\beta \mapsto t$ is in fact linear, with corresponding matrix $\Phi^\top$. The Lipschitz constant of a linear map is simply its operator norm, i.e. $\|\Phi\|_{\mathrm{op}}$, as claimed.

In the second case, our map is again linear, and is equivalent to the map $v \mapsto (1/m)\mathbb{1}^\top v$, where $v \in \mathbb{R}^m$. This map has Lipschitz constant $(1/m)\|\mathbb{1}\|_2 = 1/\sqrt{m}$, again as claimed.

G.4. Modifying q to be Lipschitz

Let us be a bit more formal about how we modify q. We will define the function

$$r_\beta(x) \overset{\text{def}}{=} \sum_{j=1}^{k} \max\left(0, 1 - \frac{1}{\delta}\max_{i=1}^{k}(\beta_i - \beta_j)^\top \frac{\phi(x)}{\|\phi(x)\|_2}\right) q(Y = j \mid x). \tag{52}$$

Therefore, $r_\beta(x) \geq q(y_\beta(x) \mid x)$ always, and will be bigger if there are other values β_i such that $\beta_i^\top \phi(x)$ is close to the maximum.

Note that $(\beta_i - \beta_j)^\top \frac{\phi(x)}{\|\phi(x)\|_2}$ is $\sqrt{2}$-Lipschitz, hence each term is $\frac{\sqrt{2}}{\delta}q(Y = j \mid x)$-Lipschitz; since $\sum_j q(Y = j \mid x) = 1$, the entire expression is $\frac{\sqrt{2}}{\delta}$-Lipschitz, as claimed. We also have the following bound on its expectation:

Lemma G.2. *For any x, if $\beta_{1:k} \sim \mathcal{N}(0, I)$, then*

$$\mathbb{E}_\beta[r_\beta(x)] \leq \frac{1}{k} + \delta. \tag{53}$$

Proof. Since $\sum_{j=1}^k q(Y = j \mid x) = 1$, it suffices to show that for each j,

$$\mathbb{E}\left[\max\left(0, 1 - \delta^{-1}\max_{i}(\beta_i - \beta_j)^\top \frac{\phi(x)}{\|\phi(x)\|_2}\right)\right] \leq \frac{1}{k} + \delta. \tag{54}$$

Let $t_i = \beta_i^\top \frac{\phi(x)}{\|\phi(x)\|_2}$; note that the t_i are independent Gaussians with mean zero and variance 1. Note also that the term in the expectation is always at most 1, and is only non-zero if either (i) t_j is the largest of the t_i (which occurs with probability $1/k$) or if (ii) $M < t_j < M + \delta$, where $M = \max_{i\neq j} t_i$. Conditioned on M, this latter event always has probability at most $\frac{\delta}{\sqrt{2\pi}} < \delta$ (since the density of a Gaussian is bounded everywhere by $\frac{1}{\sqrt{2\pi}}$). Marginalizing over M, the overall probability is also at most δ, and so the overall expectation in (54) is at most $\frac{1}{k} + \delta$, as was to be shown. □

G.5. Obtaining the final bound

Take $\delta = \frac{\epsilon}{2k}$, so that $\mathbb{E}[r_\beta(x)] \leq \frac{1+\epsilon/2}{k}$. Also define $\overline{r} = \frac{1}{m}\sum_{i=1}^m r_\beta(x_i)$. Then, since $q(y_\beta(x) \mid x) \leq r_\beta(x)$ (see the remarks in G.4), we have

$$\begin{aligned}
\mathbb{P}_\beta\left[\frac{1}{m}\sum_{i=1}^m q(y_\beta(x_i) \mid x_i) \geq \frac{1+\epsilon}{k}\right] &\leq \mathbb{P}_\beta\left[\frac{1}{m}\sum_{i=1}^m r_\beta(x_i) \geq \frac{1+\epsilon}{k}\right] && (55)\\
&= \mathbb{P}_\beta\left[\overline{r} \geq \frac{1+\epsilon}{k}\right] && (56)\\
&\leq \mathbb{P}_\beta[\overline{r} - \mathbb{E}[\overline{r}] \geq \frac{\epsilon}{2k}] && (57)\\
&\leq \exp\left(-\frac{\epsilon^4}{64k^4}\frac{m}{\|\Phi\|_{\mathrm{op}}^2}\right), && (58)
\end{aligned}$$

where the final line invokes Theorem 5.1, using the fact that $\beta \mapsto \overline{r}$ is $(\|\Phi\|_{\mathrm{op}}\cdot(2\sqrt{2}k/\epsilon)\cdot(1/\sqrt{m}))$-Lipschitz.

G.6. Almost sure convergence

We have

$$\begin{aligned}
&\lim_{m\to\infty}\mathbb{P}_{\beta,x_{1:m}}\left[\frac{1}{m}\sum_{i=1}^m q(y_\beta(x_i) \mid x_i) \geq \frac{1+\epsilon}{k}\right] && (59)\\
&\overset{(58)}{\leq} \lim_{m\to\infty}\mathbb{E}_{x_{1:m}}\left[\exp\left(-\frac{\epsilon^4}{64k^4}\frac{m}{\|\Phi(x_{1:m})\|_{\mathrm{op}}^2}\right)\right] && (60)\\
&\overset{(i)}{=} \mathbb{E}_{x_1,x_2,\dots}\left[\lim_{m\to\infty}\exp\left(-\frac{\epsilon^4}{64k^4}\frac{m}{\|\Phi(x_{1:m})\|_{\mathrm{op}}^2}\right)\right] && (61)\\
&\overset{(ii)}{=} \exp\left(-\frac{\epsilon^4}{64k^4}\gamma^{-1}\right), && (62)
\end{aligned}$$

where (i) is Lebesgue's dominated convergence theorem and (ii) is because $\|\Phi\|_{\mathrm{op}}^2/m \to \gamma^{-1}$ almost surely as $m \to \infty$.

H. Justification of Some Details for Theorem 4.5

H.1. Proof of Lemma 5.2

Let $f : \mathcal{C} \to [0,1]$ be an optimal fractional covering. For $t = 1, \dots, T$, sample $S_t \in \mathcal{C}$ with probability $\frac{f(S_t)}{N_{\mathrm{frac}}}$. Then for any fixed t and $p \in \mathcal{P}$, we have $\mathbb{P}[p \text{ is covered by } S_t] \geq \frac{1}{N_{\mathrm{frac}}}$. After $T = \lceil N_{\mathrm{frac}} \log|\mathcal{P}|\rceil$ samples, we thus have $\mathbb{P}[p \text{ is not covered by any } S_t] \leq (1 - \frac{1}{N_{\mathrm{frac}}})^{N_{\mathrm{frac}}\log|\mathcal{P}|} < e^{-\log|\mathcal{P}|} = \frac{1}{|\mathcal{P}|}$. So with non-zero probability, we cover all $p \in \mathcal{P}$ after $\lceil N_{\mathrm{frac}} \log|\mathcal{P}|\rceil$ samples, and so $N \leq \lceil N_{\mathrm{frac}} \log|\mathcal{P}|\rceil \leq N_{\mathrm{frac}}(\log|\mathcal{P}|+1)$.

H.2. Applying the minimax theorem

To expand on (16), we have

$$
\begin{aligned}
\alpha_r^* &= \min_{\nu} \sup_{\substack{f:\mathcal{U}\to\{0,1\} \\ \|f\|_1=1}} \mathbb{E}_{p\sim\nu}\left[\overbrace{\sum_{q:B_r(q)\ni p} f(q)}^{1 \text{ if } B_r(q) \text{ covers } p}\right] \\
&\overset{(i)}{\leq} \min_{\nu} \sup_{\substack{f:\mathcal{U}\to[0,1] \\ \|f\|_1=1}} \mathbb{E}_{p\sim\nu}\left[\sum_{q:B_r(q)\ni p} f(q)\right] \\
&\overset{(ii)}{=} \sup_{\substack{f:\mathcal{U}\to[0,1] \\ \|f\|_1=1}} \min_{\nu} \mathbb{E}_{p\sim\nu}\left[\sum_{q:B_r(q)\ni p} f(q)\right] \\
&= \sup_{\substack{f:\mathcal{U}\to[0,1] \\ \|f\|_1=1}} \min_{p\in\mathcal{P}} \sum_{q:B_r(q)\ni p} f(q) = \frac{1}{N_{\text{frac}}},
\end{aligned}
$$

Here (i) just relaxes the integer program to a linear program, and (ii) applies the following minimax theorem, which is Theorem 4.1 of Sion (1958):

Theorem H.1. *Let X and Y be any spaces, h a function on $X\times Y$ that is concave-convexlike. If for any $c < \inf_y \sup_x h(x,y)$ there exists a finite set $X_0 \subset X$ such that $\inf_y \max_{x\in X_0} h(x,y) > c$, then $\inf_y \sup_x h(x,y) = \sup_x \inf_y h(x,y)$.*

Here concave-convexlike is defined as follows: (i) for every $x_1, x_2 \in X$ and $t \in [0,1]$, there is a $x_0 \in X$ such that $t\cdot h(x_1,y) + (1-t)\cdot h(x_2,y) \leq h(x_0,y)$ for all $y \in Y$; (ii) for every $y_1, y_2 \in Y$ and $t\in[0,1]$, there is a $y_0 \in Y$ such that $t\cdot h(x,y_1) + (1-t)\cdot h(x,y_2) \geq h(x,y_0)$ for all $x \in X$.

We intend to apply Theorem H.1 with $X = \mathcal{F}_{\text{finite}}(\mathcal{U};1)$ (the space of finitely supported functions from $\mathcal{U}$ to $[0,1]$ satisfying $\|f\|_1 = 1$), $Y = \Delta(\mathcal{P})$ (the space of probability distributions on $\mathcal{P}$), and $h(f,\nu) = \mathbb{E}_{p\sim\nu}\left[\sum_{q:B_r(q)\ni p} f(q)\right]$.

First, we check the convexity and concavity properties. Given $\nu_1,\nu_2 \in \Delta(\mathcal{P})$ and $t \in [0,1]$, we can clearly take $\nu_0 = t\nu_1 + (1-t)\nu_2$. Similarly, given $f_1, f_2 \in \mathcal{F}_{\text{finite}}(\mathcal{U};1)$ and $t\in[0,1]$, we can clearly take $f_0 = tf_1 + (1-t)f_2$, which is finitely supported if f_1 and f_2 are.

Next, take some $c < \inf_{\nu\in\Delta(\mathcal{P})} \sup_{f\in\mathcal{F}_{\text{finite}}(\mathcal{U};1)} h(f,\nu)$. This means that in particular, $c < \inf_{\nu\in\Delta(\mathcal{P})} \sup_q \nu(B_r(q)) = \alpha_r^*$. Since α_r^* is continuous by assumption, we thus also have that $c < \inf_{\nu\in\Delta(\mathcal{P})} \sup_q \nu(B_q(r-\delta))$ for some $\delta > 0$. Now, for some ϵ (to be determined later), take a finite ϵ^4-covering $q_1,\ldots,q_N$ of $\mathcal{P}$, and also set $\overline{q}_n = (1-\epsilon)q_n + \epsilon u$, where u is the uniform distribution. We claim that $\min_{n=1}^N \nu(B_r(\overline{q}_n)) > c$ for all $\nu \in \Delta(\mathcal{P})$, whence we can take the functions $f_n = \mathbb{I}[q = \overline{q}_n]$ and $X_0 = \{f_n\}_{n=1}^N$ to satisfy the conditions of Theorem H.1.

To show this, we will show that for each q, $B_{r-\delta}(q) \subseteq B_r(\overline{q}_n)$ for some n, so that $\sup_q \nu(B_{r-\delta}(q)) \leq \max_{n=1}^n \nu(B_r(\overline{q}_n))$. Indeed, take n such that $\mathrm{D}\left(q \,\|\, q_n\right) \leq \epsilon^4$. By Lemma F.2, we then have that for any $p \in B_{r-\delta}(q)$, $\mathrm{D}\left(p \,\|\, \overline{q}_n\right) \leq \mathrm{D}\left(p \,\|\, q\right) + \log\left(\frac{1}{1-\epsilon}\right) + (k/\epsilon)\sqrt{\epsilon^4/2}$. Since $\mathrm{D}\left(p \,\|\, q\right) \leq r-\delta$, we can ensure that $\mathrm{D}\left(p \,\|\, \overline{q}_n\right) \leq r$ for sufficiently small ϵ, as claimed.

H.3. Proof of Lemma 5.3

As before, let $\overline{p} = (1-\delta)p + \delta u$, where u is the uniform distribution; we will take $\delta = \sqrt{\epsilon}$. We will make extensive use of the following helper lemma regarding the behavior of $\overline{p}$, proved later in this section:

Lemma H.2. *Let $H(\delta) = \delta\log\left(\frac{1}{\delta}\right) + (1-\delta)\log\left(\frac{1}{1-\delta}\right)$. Also suppose that $\delta < \frac{1}{2}$. Then, for any p and q, we have:*

$$\mathrm{KL}\left(p \,\|\, \overline{q}\right) \leq \mathrm{KL}\left(p \,\|\, q\right) + \log\left(\frac{1}{1-\delta}\right) \tag{63}$$

$$\mathrm{KL}\left(\overline{q} \,\|\, p\right) \geq (1-\delta)\,\mathrm{KL}\left(p \,\|\, q\right) - H(\delta) \tag{64}$$

$$\mathrm{KL}\left(\overline{p} \,\|\, \overline{q}\right) \leq (1-\delta)\,\mathrm{KL}\left(p \,\|\, q\right) + H(\delta) \tag{65}$$

$$\mathrm{KL}\left(\overline{p} \,\|\, \overline{q}\right) \leq 2(1-\delta)\|p-q\|_{TV}\log(k/\delta). \tag{66}$$

Also under the assumption $\delta < 1/2$, we have $\log(1/(1-\delta)) \leq 2\delta$, which we will often use to simplify some of the above expressions. Now, starting with the actual proof, we have

$$\begin{aligned}
\mathrm{KL}\left(p \parallel \overline{q}\right) &= \mathrm{KL}\left(\overline{\hat{p}} \parallel \overline{q}\right) + \left(\mathrm{KL}\left(p \parallel \overline{q}\right) - \mathrm{KL}\left(\overline{\hat{p}} \parallel \overline{q}\right)\right) \\
&\overset{(65)}{\leq} (1-\delta)\,\mathrm{KL}\left(\hat{p} \parallel q\right) + H(\delta) + \left(\mathrm{KL}\left(p \parallel \overline{q}\right) - \mathrm{KL}\left(\overline{\hat{p}} \parallel \overline{q}\right)\right).
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
\mathrm{KL}\left(p \parallel \overline{q}\right) - \mathrm{KL}\left(\overline{\hat{p}} \parallel \overline{q}\right) &= \int p(x) \log \frac{p(x)}{\overline{q}(x)} dx - \int \overline{\hat{p}}(x) \log \frac{\overline{\hat{p}}(x)}{\overline{q}(x)} dx \\
&= \mathrm{KL}\left(p \parallel \overline{\hat{p}}\right) + \int (p(x) - \overline{\hat{p}}(x)) \log \frac{\overline{\hat{p}}(x)}{\overline{q}(x)} dx \\
&\leq \mathrm{KL}\left(p \parallel \overline{\hat{p}}\right) + \left| \int (p(x) - \overline{\hat{p}}(x)) dx \right| \log(k/\delta) \\
&\leq \mathrm{KL}\left(p \parallel \hat{p}\right) + 2\delta + 2\|p - \overline{\hat{p}}\|_{TV} \log(k/\delta) \\
&\leq \mathrm{KL}\left(p \parallel \hat{p}\right) + 2\delta + 2(\|p - \hat{p}\|_{TV} + \|\hat{p} - \overline{\hat{p}}\|_{TV}) \log(k/\delta) \\
&\leq \epsilon + 2\delta + 2(\sqrt{\epsilon/2} + \delta) \log(k/\delta),
\end{aligned}$$

in which case

$$\begin{aligned}
\mathrm{KL}\left(p \parallel \overline{q}\right) &\leq (1-\delta)\,\mathrm{KL}\left(\hat{p} \parallel q\right) + \epsilon + \sqrt{2\epsilon} \log(k/\delta) + H(\delta) + 2\delta\left(\log(k/\delta) + 1\right) \\
&\leq \mathrm{KL}\left(\hat{p} \parallel q\right) + \sqrt{\epsilon}\left((\sqrt{2} + 2)\log(k) + (\sqrt{2} + 3)\log(1/\sqrt{\epsilon}) + 4 + \sqrt{\epsilon}\right) \\
&\leq \mathrm{KL}\left(\hat{p} \parallel q\right) + 5\sqrt{\epsilon} \log(2k/\epsilon).
\end{aligned}$$

For the second part, by symmetry we will suppose instead that $\mathrm{KL}\left(\hat{p} \parallel p\right) \leq \epsilon$ and prove the same inequality as before. We have

$$\begin{aligned}
\mathrm{KL}\left(p \parallel \overline{q}\right) &= \mathrm{KL}\left(\hat{p} \parallel q\right) + \left(\mathrm{KL}\left(\hat{p} \parallel \overline{q}\right) - \mathrm{KL}\left(\hat{p} \parallel q\right)\right) + \left(\mathrm{KL}\left(\overline{p} \parallel \overline{q}\right) - \mathrm{KL}\left(\hat{p} \parallel \overline{q}\right)\right) + \left(\mathrm{KL}\left(p \parallel \overline{q}\right) - \mathrm{KL}\left(\overline{p} \parallel \overline{q}\right)\right) \\
&\leq \mathrm{KL}\left(\hat{p} \parallel q\right) + 2\delta + H(\delta) + \delta\,\mathrm{KL}\left(p \parallel \overline{q}\right) + \left(\mathrm{KL}\left(\overline{p} \parallel \overline{q}\right) - \mathrm{KL}\left(\hat{p} \parallel \overline{q}\right)\right) \\
&\leq \mathrm{KL}\left(\hat{p} \parallel q\right) + 2\delta + H(\delta) + \delta \log(k/\delta) + \left(\mathrm{KL}\left(\overline{p} \parallel \overline{q}\right) - \mathrm{KL}\left(\hat{p} \parallel \overline{q}\right)\right).
\end{aligned}$$

Using essentially the same argument as before, we have

$$\begin{aligned}
\mathrm{KL}\left(\overline{p} \parallel \overline{q}\right) - \mathrm{KL}\left(\hat{p} \parallel \overline{q}\right) &= \int \overline{p}(x) \log \frac{\overline{p}(x)}{\overline{q}(x)} dx - \int \hat{p}(x) \log \frac{\hat{p}(x)}{\overline{q}(x)} dx \\
&= -\,\mathrm{KL}\left(\hat{p} \parallel \overline{p}\right) + \int (\overline{p}(x) - \hat{p}(x)) \log \frac{\overline{p}(x)}{\overline{q}(x)} dx \\
&\leq \int |\overline{p}(x) - \hat{p}(x)| \log(k/\delta) dx \\
&= 2\|\overline{p} - \hat{p}\|_{TV} \log(k/\delta) \\
&\leq 2\left(\sqrt{\epsilon/2} + \delta\right) \log(k/\delta),
\end{aligned}$$

in which case

$$\begin{aligned}
\mathrm{KL}\left(p \parallel \overline{q}\right) &\leq \mathrm{KL}\left(\hat{p} \parallel q\right) + 2\delta + H(\delta) + \delta \log(k/\delta) + 2\left(\sqrt{\epsilon/2} + \delta\right) \log(k/\delta) \\
&= \mathrm{KL}\left(\hat{p} \parallel q\right) + \sqrt{\epsilon}\left((\sqrt{2} + 3)\log(k) + (\sqrt{2} + 4)\log(1/\sqrt{\epsilon}) + 4\right) \\
&\leq \mathrm{KL}\left(\hat{p} \parallel q\right) + 5\sqrt{\epsilon} \log(2k/\epsilon).
\end{aligned}$$

It only remains to prove Lemma H.2.

Proof of Lemma H.2. First,

$$\begin{aligned}\mathrm{KL}\left(p \,\|\, \overline{q}\right) &= \int p(x) \log \frac{p(x)}{\overline{q}(x)} dx \\ &\leq \int p(x) \log \frac{p(x)}{(1-\delta)q(x)} dx \\ &= \mathrm{KL}\left(p \,\|\, q\right) + \log \frac{1}{1-\delta}.\end{aligned}$$

Next,

$$\begin{aligned}\mathrm{KL}\left(\overline{p} \,\|\, q\right) &= \int \left[(1-\delta)p(x) + \delta u(x)\right] \log \frac{(1-\delta)p(x) + \delta u(x)}{q(x)} dx \\ &\geq (1-\delta) \int p(x) \log \frac{(1-\delta)p(x)}{q(x)} dx + \delta \int u(x) \log \frac{\delta u(x)}{q(x)} dx \\ &= (1-\delta)\,\mathrm{KL}\left(p \,\|\, q\right) + \delta\,\mathrm{KL}\left(u \,\|\, q\right) - H(\delta) \\ &\geq (1-\delta)\,\mathrm{KL}\left(p \,\|\, q\right) - H(\delta).\end{aligned}$$

In addition,

$$\begin{aligned}\mathrm{KL}\left(\overline{p} \,\|\, \overline{q}\right) &\leq (1-\delta)\,\mathrm{KL}\left(p \,\|\, \overline{q}\right) + \delta\,\mathrm{KL}\left(u \,\|\, \overline{q}\right) \\ &\leq (1-\delta)\left(\mathrm{KL}\left(p \,\|\, q\right) + \log \frac{1}{1-\delta}\right) + \delta \int u(x) \log \frac{u(x)}{(1-\delta)q(x) + \delta u(x)} dx \\ &\leq (1-\delta)\,\mathrm{KL}\left(p \,\|\, q\right) + (1-\delta) \log \frac{1}{1-\delta} + \delta \log \frac{1}{\delta}.\end{aligned}$$

Finally,

$$\begin{aligned}\mathrm{KL}\left(\overline{p} \,\|\, \overline{q}\right) &\leq \mathrm{KL}\left(\overline{p} \,\|\, \overline{q}\right) + \mathrm{KL}\left(\overline{q} \,\|\, \overline{p}\right) \\ &= (1-\delta) \int \left(p(x) - q(x)\right)\left(\log \overline{p}(x) - \log \overline{q}(x)\right) dx \\ &\leq (1-\delta) \int |p(x) - q(x)| \log(k/\delta) dx \\ &= 2(1-\delta)\|p - q\|_{TV} \log(k/\delta).\end{aligned}$$

□

H.4. Replacing $\mathcal{P}$ by a covering

Let f be a fractional covering of $\mathcal{P}$ of radius $r_0 - \epsilon$, and let $\hat{\mathcal{P}}$ be a covering of $\mathcal{P}$ of radius ϵ_0, where ϵ_0 will be determined later. Then by Lemma 5.3, f is a fractional covering of $\hat{\mathcal{P}}$ of radius $r_0 - \epsilon + 5\sqrt{\epsilon_0}\log(2k/\epsilon_0)$. However, again by Lemma 5.3, note that any (non-fractional) covering of $\hat{\mathcal{P}}$ of radius $r_0 - \epsilon + 5\sqrt{\epsilon_0}\log(2k/\epsilon_0)$ is also a (non-fractional) covering of $\mathcal{P}$ of radius $r_0 - \epsilon + 10\sqrt{\epsilon_0}\log(2k/\epsilon_0)$. Combining with Lemma 5.2, we have

$$\begin{aligned}N_{\mathrm{frac}}(\mathcal{P}, r_0 - \epsilon) &\geq N_{\mathrm{frac}}(\hat{\mathcal{P}}, r_0 - \epsilon + 5\sqrt{\epsilon_0}\log(2k/\epsilon_0)) && (67) \\ &\geq N(\hat{\mathcal{P}}, r_0 - \epsilon + 5\sqrt{\epsilon_0}\log(2k/\epsilon_0))/(1 + \log|\hat{\mathcal{P}}|) && (68) \\ &\geq N(\mathcal{P}, r_0 - \epsilon + 10\sqrt{\epsilon_0}\log(2k/\epsilon_0))/(1 + \log|\hat{\mathcal{P}}|). && (69)\end{aligned}$$

We want to take ϵ_0 such that $10\sqrt{\epsilon_0}\log(2k/\epsilon_0) < \epsilon$. Here we will use the fact that $\sqrt{t}\log(1/t) \leq 2.3t^{1/3}$, applied at $t = \epsilon_0/2k$, to obtain $10\sqrt{\epsilon_0}\log(2k/\epsilon_0) < 23\sqrt{2k}(\epsilon_0/2k)^{1/3} = 23(\sqrt{2k}\epsilon_0)^{1/3} < 26(\sqrt{k}\epsilon_0)^{1/3}$, and hence we can take $\epsilon_0 = (\epsilon/26)^3/\sqrt{k}$, as was to be shown. We also observe that this is less than $1/4$ for any $\epsilon \leq 1$, so that the precondition of Lemma 5.3 is always satisfied.