# Unsupervised Risk Estimation Using Only Conditional Independence Structure

**Jacob Steinhardt**
Stanford University
jsteinhardt@cs.stanford.edu

**Percy Liang**
Stanford University
pliang@cs.stanford.edu

## Abstract

We show how to estimate a model's test error from unlabeled data, on distributions very different from the training distribution, while assuming only that certain conditional independencies are preserved between train and test. We do not need to assume that the optimal predictor is the same between train and test, or that the true distribution lies in any parametric family. We can also efficiently compute gradients of the estimated error and hence perform unsupervised discriminative learning. Our technical tool is the method of moments, which allows us to exploit conditional independencies in the absence of a fully-specified model. Our framework encompasses a large family of losses including the log and exponential loss, and extends to structured output settings such as conditional random fields.

## 1 Introduction

Can we measure the accuracy of a model at test time without any ground truth labels, and without assuming the test distribution is close to the training distribution? This is the problem of *unsupervised risk estimation* (Donmez et al., 2010): Given a loss function $L(\theta; x, y)$ and a fixed model $\theta$, estimate the risk $R(\theta) \stackrel{\text{def}}{=} \mathbf{E}_{x,y\sim p^*}[L(\theta; x, y)]$ with respect to a test distribution $p^*(x, y)$, given access only to $m$ unlabeled examples $x^{(1:m)} \sim p^*(x)$. Unsupervised risk estimation lets us estimate model accuracy on a novel distribution, and is thus important for building reliable machine learning systems. Beyond evaluating a single model, it also provides a way of harnessing unlabeled data for learning: by minimizing the estimated risk over $\theta$, we can perform unsupervised learning and domain adaptation.

Unsupervised risk estimation is impossible without some assumptions on $p^*$, as otherwise $p^*(y \mid x)$—about which we have no observable information—could be arbitrary. How satisfied we should be with an estimator depends on how strong its underlying assumptions are. In this paper, we present an approach which rests on surprisingly weak assumptions—that $p^*$ satisfies certain conditional independencies, but not that it lies in any parametric family or is close to the training distribution.

To give a flavor for our results, suppose that $y \in \{1, \dots, k\}$ and that the loss decomposes as a sum of three parts: $L(\theta; x, y) = \sum_{v=1}^{3} f_v(\theta; x_v, y)$, where the $x_v$ ($v = 1, 2, 3$) are independent conditioned on $y$. In this case, we show that we can estimate the risk to error $\epsilon$ in $\text{poly}(k)/\epsilon^2$ samples, independently of the dimension of $x$ or $\theta$, with only very mild additional assumptions on $p^*$. In Sections 2 and 3 we generalize to a larger family of losses including the log and exponential losses, and extend beyond the multiclass case to conditional random fields.

Some intuition behind our result is provided in Figure 1. At a fixed value of $x$, we can think of each $f_v$ as "predicting" that $y = j$ if $f_v(x_v, j)$ is low and $f_v(x_v, j')$ is high for $j' \neq j$. Since $f_1$, $f_2$, and $f_3$ all provide independent signals about $y$, their rate of agreement gives information about the model accuracy. If $f_1$, $f_2$, and $f_3$ all predict that $y = 1$, then it is likely that the true $y$ equals 1 and the loss is small. Conversely, if $f_1$, $f_2$, and $f_3$ all predict different values of $y$, then the loss is likely

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

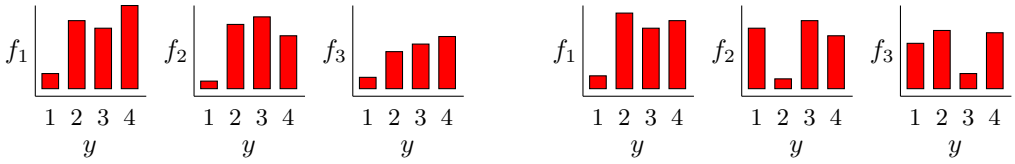



Figure 1: Two possible loss profiles at a given value of $x$. Left: if $f_1$, $f_2$, and $f_3$ are all minimized at the same value of $y$, that is likely to be the correct value and the total loss is likely to be small. Right: conversely, if $f_1$, $f_2$, and $f_3$ are small at differing values of $y$, then the loss is likely to be large.

large. This intuition is formalized by Dawid and Skene (1979) when the $f_v$ measure the $0/1$-loss of independent classifiers; in particular, if $r_v$ is the prediction of a classifier based on $x_v$, then Dawid and Skene model the $r_v$ as independent given $y$: $p(r_1, r_2, r_3) = \sum_{j=1}^{k} p(y = j) \prod_{v=1}^{3} p(r_v \mid y = j)$. They then use the learned parameters of this model to compute the $0/1$-loss.

**Partial specification.** Dawid and Skene's approach relies on the prediction $r_v$ only taking on $k$ values. In this case, the full distribution $p(r_1, r_2, r_3)$ can be parametrized by $k \times k$ conditional probability matrices $p(r_v \,|\, y)$ and marginals $p(y)$. However, as shown in Figure 1, we want to estimate continuous losses such as the log loss. We must therefore work with the prediction vector $f_v \in \mathbf{R}^k$ rather than a single predicted output $r_v \in \{1, \ldots, k\}$. To fully model $p(f_1, f_2, f_3)$ would require nonparametric estimation, resulting in an undesirable sample complexity exponential in $k$—in contrast to the discrete case, conditional independence effectively only *partially specifies* a model for the losses.

To sidestep this issue, we make use of the *method of moments*, which has recently been used to fit non-convex latent variable models (e.g. Anandkumar et al., 2012). In fact, it has a much older history in the econometrics literature, where it is used as a tool for making causal identifications under structural assumptions, even when no explicit form for the likelihood is known (Anderson and Rubin, 1949; 1950; Sargan, 1958; 1959; Hansen, 1982; Powell, 1994; Hansen, 2014). It is this latter perspective that we draw upon. The key insight is that even in the absence of a fully-specified model, certain moment equations–such as $\mathbf{E}[f_1 f_2 \,|\, y] = \mathbf{E}[f_1 \,|\, y]\mathbf{E}[f_2 \,|\, y]$–can be derived solely from the assumed conditional independence. Solving these equations yields estimates of $\mathbf{E}[f_v \,|\, y]$, which can in turn be used to estimate the risk. Importantly, our procedure avoids estimation of the full loss distribution $p(f_1, f_2, f_3)$, on which we make no assumptions other than conditional independence.

Our paper is structured as follows. In Section 2, we present our basic framework, and state and prove our main result on estimating the risk. In Section 3, we extend our framework in several directions, including to conditional random fields. In Section 4, we present a gradient-based learning algorithm and show that the sample complexity needed for learning is $d \cdot \mathrm{poly}(k)/\epsilon^2$, where $d$ is the dimension of the parameters $\theta$. In Section 5, we investigate how our method performs empirically.

**Related Work.** While the formal problem of unsupervised risk estimation was only posed recently by Donmez et al. (2010), several older ideas from domain adaptation and semi-supervised learning are also relevant. The *covariate shift assumption* posits access to labeled samples from a training distribution $p_0(x, y)$ for which $p^*(y \mid x) = p_0(y \mid x)$. If $p^*(x)$ and $p_0(x)$ are close, we can approximate $p^*$ by $p_0$ via importance weighting (Shimodaira, 2000; Quiñonero-Candela et al., 2009). If $p^*$ and $p_0$ are not close, another approach is to assume a well-specified discriminative model family $\Theta$, such that $p_0(y \,|\, x) = p^*(y \,|\, x) = p_{\theta^*}(y \,|\, x)$ for some $\theta^* \in \Theta$; then the only error when moving from $p_0$ to $p^*$ is statistical error in the estimation of $\theta^*$ (Blitzer et al., 2011; Li et al., 2011). Such assumptions are restrictive—importance weighting only allows small perturbations from $p_0$ to $p^*$, and mis-specified models of $p(y \,|\, x)$ are common in practice; many authors report that mis-specification can lead to severe issues in semi-supervised settings (Merialdo, 1994; Nigam et al., 1998; Cozman and Cohen, 2006; Liang and Klein, 2008; Li and Zhou, 2015). More sophisticated approaches based on discrepancy minimization (Mansour et al., 2009) or learning invariant representations (Ben-David et al., 2006; Johansson et al., 2016) typically also make some form of the covariate shift assumption.

Our approach is closest to Dawid and Skene (1979) and some recent extensions (Zhang et al., 2014; Platanios, 2015; Jaffe et al., 2015; Fetaya et al., 2016). Similarly to Zhang et al. (2014) and Jaffe et al. (2015), we use the method of moments for estimating latent-variable models. However, those papers use it for parameter estimation in the face of non-convexity, rather than as a way to avoid full estimation of $p(f_v \,|\, y)$. The insight that the method of moments works under partial specification lets us extend beyond the simple discrete settings they consider to handle more complex continuous and structured losses. The intriguing work of Balasubramanian et al. (2011) provides an alternate approach

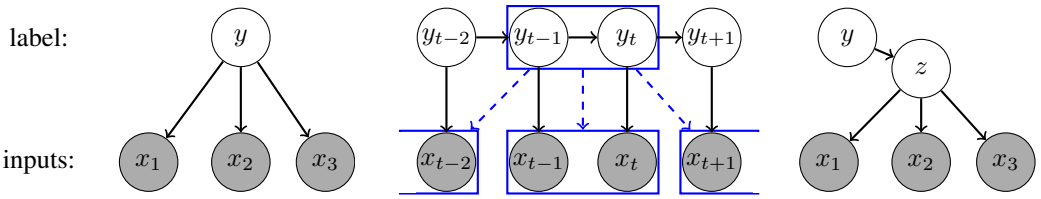



Figure 2: Left: our basic 3-view setup (Assumption 1). Center: Extension 1, to CRFs; the embedding of 3 views into the CRF is indicated in blue. Right: Extension 3, to include a mediating variable $z$.

to continuous losses; they show that the distribution of losses $L \mid y$ is often approximately Gaussian, and use that to estimate the risk. Among all this work, ours is the first to perform gradient-based learning and the first to handle a structured loss (the log loss for conditional random fields).

## 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 $\mathcal{Y} = \{1, \ldots, k\}$, and are given unlabeled samples $x^{(1)}, \ldots, x^{(m)}$ drawn i.i.d. 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 $\theta$ on $p^*$: $R(\theta) \stackrel{\text{def}}{=} \mathbf{E}_{x,y\sim p^*}[L(\theta; x, y)]$. Throughout, we will make the *3-view assumption*:

**Assumption 1** (3-view)**.** *Under $p^*$, $x$ can be split into $x_1, x_2, x_3$, which are conditionally independent given $y$ (see Figure 2). 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 $f_v$.*

Note that each $x_v$ can be large (e.g. they could be vectors in $\mathbf{R}^d$). If we have $V > 3$ views, we can combine views to obtain $V = 3$ without loss of generality. It also suffices for just the $f_v$ to be independent rather than the $x_v$. Given only 2 views, the risk can be shown to be unidentifiable in general, although obtaining upper bounds may be possible.

We give some examples where Assumption 1 holds, then state and prove our main result (see Section 3 for additional examples). We start with logistic regression, which will be our primary focus later on:

**Example 1** (Logistic Regression)**.** Suppose that we have a log-linear model $p_\theta(y \mid x) = \exp\left(\theta^\top (\phi_1(x_1, y) + \phi_2(x_2, y) + \phi_3(x_3, y)) - A(\theta; x)\right)$, where $x_1$, $x_2$, and $x_3$ are independent conditioned on $y$. If our loss function is the log-loss $L(\theta; x, y) = -\log p_\theta(y \mid x)$, then Assumption 1 holds with $f_v(\theta; x_v, y) = \theta^\top \phi_v(x_v, y)$ and $A(\theta; x)$ equal to the partition function of $p_\theta$. □

Assumption 1 does *not* hold for the hinge loss (see Appendix A for details), but it does hold for a modified hinge loss, where we apply the hinge separately to each view:

**Example 2** (Modified Hinge Loss)**.** Suppose that $L(\theta; x, y) = \sum_{v=1}^{3}(1 + \max_{j\neq y} \theta^\top \phi_v(x_v, j) - \theta^\top \phi_v(x_v, y))_+$. In other words, $L$ is the sum of 3 hinge losses, one for each view. Then Assumption 1 holds with $A = 0$, and $-f_v$ equal to the hinge loss for view $v$. □

The model can also be non-linear within each view $x_v$, as long as the views are combined additively:

**Example 3** (Neural Networks)**.** Suppose that for each view $v$ we have a neural network whose output is a score for each of the $k$ classes, $(f_v(\theta; x_v, j))_{j=1}^k$. Sum the scores $f_1 + f_2 + f_3$, apply a soft-max, and evaluate using the log loss; then $L(\theta; x, y) = A(\theta; x) - \sum_{v=1}^{3} f_v(\theta; x_v, y)$, where $A(\theta; x)$ is the log-normalization constant of the softmax, and hence $L$ satisfies Assumption 1. □

We are now ready to present our main result on recovering the risk $R(\theta)$. The key starting point is the *conditional risk matrices* $M_v \in \mathbf{R}^{k\times k}$, defined as (suppressing the dependence on $\theta$)

$$(M_v)_{ij} = \mathbf{E}[f_v(\theta; x_v, i) \mid y = j]. \tag{1}$$

In the case of the $0/1$-loss, the $M_v$ are confusion matrices; in general, $(M_v)_{ij}$ measures how strongly we predict class $i$ when the true class is $j$. If we could recover these matrices along with the marginal class probabilities $\pi_j \stackrel{\text{def}}{=} p^*(y = j)$, then estimating the risk would be straightforward; indeed,

$$R(\theta) = \mathbf{E}\left[A(\theta; x) - \sum_{v=1}^{3} f_v(\theta; x_v, y)\right] = \mathbf{E}[A(\theta; x)] - \sum_{j=1}^{k} \pi_j \sum_{v=1}^{3} (M_v)_{j,j}, \tag{2}$$

where $\mathbf{E}[A(\theta; x)]$ can be estimated from unlabeled data alone.

**Caveat: Class permutation.** Suppose that at training time, we learn to predict whether an image contains the digit 0 or 1. At test time, nothing changes except the definitions of 0 and 1 are reversed. It is clearly impossible to detect this from unlabeled data— mathematically, the risk matrices $M_v$ are only recoverable up to column permutation. We will end up computing the minimum risk over these permutations, which we call the *optimistic risk* and denote $\tilde{R}(\theta) \stackrel{\text{def}}{=} \min_{\sigma \in \text{Sym}(k)} \mathbf{E}_{x,y \sim p^*}[L(\theta; x, \sigma(y))]$. This equals the true risk as long as $\theta$ is at least aligned with the correct classes in the sense that $\mathbf{E}_x[L(\theta; x, j) \mid y = j] \leq \mathbf{E}_x[L(\theta; x, j') \mid y = j]$ for $j' \neq j$. The optimal $\sigma$ can be computed from $M_v$ and $\pi$ in $\mathcal{O}\left(k^3\right)$ time using maximum weight bipartite matching; see Section B for details.

Our main result, Theorem 1, says that we can recover both $M_v$ and $\pi$ up to permutation, with a number of samples that is polynomial in $k$:

**Theorem 1.** *Suppose Assumption 1 holds. Then, for any $\epsilon, \delta \in (0, 1)$, we can estimate $M_v$ and $\pi$ up to column permutation, to error $\epsilon$ (in Frobenius and $\infty$-norm respectively). Our algorithm requires*

$$m = \text{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right) \cdot \frac{\log(2/\delta)}{\epsilon^2} \text{ samples to succeed with probability } 1 - \delta, \text{ where}$$
$$\pi_{\min} \stackrel{\text{def}}{=} \min_{j=1}^{k} p^*(y = j), \quad \tau \stackrel{\text{def}}{=} \mathbf{E}\left[\sum_{v,j} f_v(\theta; x_v, j)^2\right], \quad \text{and} \quad \lambda \stackrel{\text{def}}{=} \min_{v=1}^{3} \sigma_k(M_v), \tag{3}$$

*and $\sigma_k$ denotes the $k$th singular value. Moreover, the algorithm runs in time $m \cdot \text{poly}(k)$.*

Estimates for $M_v$ and $\pi$ imply an estimate for $\tilde{R}$ via (2); see Algorithm 1 below for details. Importantly, the sample complexity in Theorem 1 depends on the number of classes $k$, but not on the dimension $d$ of $\theta$. Moreover, Theorem 1 holds even if $p^*$ lies outside the model family $\theta$, and even if the train and test distributions are very different (in fact, the result is agnostic to how the model $\theta$ was produced). The only requirement is the 3-view assumption for $p^*$ and that $\lambda, \pi_{\min} \neq 0$.

Let us interpret each term in (3). First, $\tau$ tracks the variance of the loss, and we should expect the difficulty of estimating the risk to increase with this variance. The $\frac{\log(2/\delta)}{\epsilon^2}$ term is typical and shows up even when estimating the parameter of a random variable to accuracy $\epsilon$ from $m$ samples. The $\pi_{\min}^{-1}$ term appears because, if one of the classes is very rare, we need to wait a long time to observe even a single sample from that class, and even longer to estimate the risk on that class accurately.

Perhaps least intuitive is the $\lambda^{-1}$ term, which is large e.g. when two classes have similar conditional risk vectors $\mathbf{E}[(f_v(\theta; x_v, i))_{i=1}^k \mid y = j]$. To see why this matters, consider an extreme where $x_1$, $x_2$, and $x_3$ are independent not only of each other but also of $y$. Then $p^*(y)$ is completely unconstrained, and it is impossible to estimate $R$ at all. Why does this not contradict Theorem 1? The answer is that in this case, all rows of $M_v$ are equal and hence $M_v$ has rank 1, $\lambda = 0$, $\lambda^{-1} = \infty$, and we need infinitely many samples for Theorem 1 to hold; $\lambda$ measures how close we are to this degenerate case.

**Proof of Theorem 1.** We now outline a proof of Theorem 1. Recall the goal is to estimate the conditional risk matrices $M_v$, defined as $(M_v)_{ij} = \mathbf{E}[f_v(\theta; x_v, i) \mid y = j]$; from these we can recover the risk itself using (2). The key insight is that certain moments of $p^*(y \mid x)$ can be expressed as polynomial functions of the matrices $M_v$, and therefore we can solve for the $M_v$ even without explicitly estimating $p^*$. Our approach follows the technical machinery behind the spectral method of moments (e.g., Anandkumar et al., 2012), which we explain below for completeness.

Define the loss vector $h_v(x_v) = (f_v(\theta; x_v, i))_{i=1}^k$, which measures the loss that would be incurred under *each* of the $k$ classes. The conditional independence of the $x_v$ means that $\mathbf{E}[h_1(x_1)h_2(x_2)^\top \mid y] = \mathbf{E}[h_1(x_1) \mid y]\mathbf{E}[h_2(x_2) \mid y]^\top$, and similarly for higher-order conditional moments. Marginalizing over $y$, we see that there is low-rank structure in the moments of $h$ that we can exploit; in particular (letting $\otimes$ denote outer product and $A_{\cdot,j}$ denote the $j$th column of $A$):

$$\mathbf{E}[h_v(x_v)] = \sum_{j=1}^{k} \pi_j \cdot (M_v)_{\cdot,j}, \quad \mathbf{E}[h_v(x_v) \otimes h_{v'}(x_{v'})] = \sum_{j=1}^{k} \pi_j \cdot (M_v)_{\cdot,j} \otimes (M_{v'})_{\cdot,j} \text{ for } v \neq v', \text{ and}$$
$$\mathbf{E}[h_1(x_1) \otimes h_2(x_2) \otimes h_3(x_3)] = \sum_{j=1}^{k} \pi_j \cdot (M_1)_{\cdot,j} \otimes (M_2)_{\cdot,j} \otimes (M_3)_{\cdot,j}. \tag{4}$$

The left-hand-side of each equation can be estimated from unlabeled data; using *tensor decomposition* (Lathauwer, 2006; Comon et al., 2009; Anandkumar et al., 2012; 2013; Kuleshov et al., 2015), it is

**Algorithm 1** Algorithm for estimating $\tilde{R}(\theta)$ from unlabeled data.

1: **Input**: unlabeled samples $x^{(1)}, \dots, x^{(m)} \sim p^*(x)$.
2: Estimate the left-hand-side of each term in (4) using $x^{(1:m)}$.
3: Compute approximations $\hat{M}_v$ and $\hat{\pi}$ to $M_v$ and $\pi$ using tensor decomposition.
4: Compute $\sigma$ maximizing $\sum_{j=1}^{k} \hat{\pi}_{\sigma(j)} \sum_{v=1}^{3} (\hat{M}_v)_{j,\sigma(j)}$ using maximum bipartite matching.
5: **Output**: estimated risk, $\frac{1}{m}\sum_{i=1}^{m} A(\theta; x^{(i)}) - \sum_{j=1}^{k} \hat{\pi}_{\sigma(j)} \sum_{v=1}^{3} (\hat{M}_v)_{j,\sigma(j)}$.

then possible to solve for $M_v$ and $\pi$. In particular, we can recover $M$ and $\pi$ up to permutation: that is, we recover $\hat{M}$ and $\hat{\pi}$ such that $M_{i,j} \approx \hat{M}_{i,\sigma(j)}$ and $\pi_j \approx \hat{\pi}_{\sigma(j)}$ for some permutation $\sigma \in \mathrm{Sym}(k)$. This then yields Theorem 1; see Section C for a full proof.

Assumption 1 thus yields a set of moment equations (4) whose solution lets us estimate the risk without any labels $y$. The procedure is summarized in Algorithm 1: we (i) approximate the left-hand-side of each term in (4) by sample averages; (ii) use tensor decomposition to solve for $\pi$ and $M_v$; (iii) use maximum matching to compute the permutation $\sigma$; and (iv) use (2) to obtain $\tilde{R}$ from $\pi$ and $M_v$.

## 3 Extensions

Theorem 1 provides a basic building block which admits several extensions to more complex model structures. We go over several cases below, omitting most proofs to avoid tedium.

**Extension 1** (Conditional Random Field)**.** Most importantly, the variable $y$ need not belong to a small discrete set; we can handle structured outputs such as a CRF as long as $p^*$ has the right structure. This contrasts with previous work on unsupervised risk estimation that was restricted to multiclass classification (though in a different vein, it is close to Proposition 8 of Anandkumar et al. (2012)).

Suppose that $p^*(x_{1:T}, y_{1:T})$ factorizes as a hidden Markov model, and that $p_\theta$ is a CRF respecting the HMM structure: $p_\theta(y_{1:T} \mid x_{1:T}) \propto \prod_{t=2}^{T} f_\theta(y_{t-1}, y_t) \cdot \prod_{t=1}^{T} g_\theta(y_t, x_t)$. For the log-loss $L(\theta; x, y) = -\log p_\theta(y_{1:T} \mid x_{1:T})$, we can exploit the decomposition

$$-\log p_\theta(y_{1:T} \mid x_{1:T}) = \sum_{t=2}^{T} \underbrace{-\log p_\theta(y_{t-1}, y_t \mid x_{1:T})}_{\stackrel{\text{def}}{=} \ell_t} - \sum_{t=1}^{T} \underbrace{-\log p_\theta(y_t \mid x_{1:T})}_{\stackrel{\text{def}}{=} \ell'_t}. \tag{5}$$

Each of the components $\ell_t$ and $\ell'_t$ satisfy Assumption 1 (see Figure 2; for $\ell_t$, the views are $x_{1:t-2}, x_{t-1:t}, x_{t+1:T}$ , and for $\ell'_t$ they are $x_{1:t-1}$, $x_t$, $x_{t+1:T}$). We use Theorem 1 to estimate each $\mathbf{E}[\ell_t]$, $\mathbf{E}[\ell'_t]$ individually, and thus also the full risk $\mathbf{E}[L]$. (We actually estimate the risk for $y_{2:T-1} \mid x_{1:T}$ due to the 3-view assumption failing at the boundaries.)

In general, the idea in (5) applies to any structured output problem that is a sum of local 3-view structures. It would be interesting to extend our results to other structures such as more general graphical models (Chaganty and Liang, 2014) and parse trees (Hsu et al., 2012).

**Extension 2** (Exponential Loss)**.** We can also relax the additivity $L = A - f_1 - f_2 - f_3$ in Assumption 1. For instance, suppose $L(\theta; x, y) = \exp(-\theta^\top \sum_{v=1}^{3} \phi_v(x_v, y))$ is the exponential loss. Theorem 1 lets us estimate the matrices $M_v$ corresponding to $f_v(\theta; x_v, y) = \exp(-\theta^\top \phi_v(x_v, y))$. Then

$$R(\theta) = \mathbf{E}\left[\prod_{v=1}^{3} f_v(\theta; x_v, y)\right] = \sum_j \pi_j \prod_{v=1}^{3} \mathbf{E}\left[f_v(\theta; x_v, j) \mid y = j\right] \tag{6}$$

by conditional independence, so the risk can be computed as $\sum_j \pi_j \prod_{v=1}^{3} (M_v)_{j,j}$. This idea extends to any loss expressible as $L(\theta; x, y) = A(\theta; x) + \sum_{i=1}^{n} \prod_{v=1}^{3} f_i^v(\theta; x_v, y)$ for some functions $f_i^v$.

**Extension 3** (Mediating Variable)**.** Assuming that $x_{1:3}$ are independent conditioned only on $y$ may not be realistic; there might be multiple subclasses of a class (e.g., multiple ways to write the digit 4) which would induce systematic correlations across views. To address this, we show that independence need only hold conditioned on a mediating variable $z$, rather than on the class $y$ itself.

Let $z$ be a refinement of $y$ (in the sense that knowing $z$ determines $y$) which takes on $k'$ values, and suppose that the views $x_1$, $x_2$, $x_3$ are independent conditioned on $z$, as in Figure 2. Then we can

try to estimate the risk by defining $L'(\theta; x, z) = L(\theta; x, y(z))$, which satisfies Assumption 1. The problem is that the corresponding risk matrices $M'_v$ will only have $k$ distinct rows and hence have rank $k < k'$. To fix this, suppose that the loss vector $h_v(x_v) = (f_v(x_v, j))_{j=1}^k$ can be extended to a vector $h'_v(x_v) \in \mathbf{R}^{k'}$, such that (i) the first $k$ coordinates of $h'_v(x_v)$ are $h_v(x_v)$ and (ii) the conditional risk matrix $M'_v$ corresponding to $h'_v$ has full rank. Then, Theorem 1 allows us to recover $M'_v$ and hence also $M_v$ (since it is a sub-matrix of $M'_v$) and thereby estimate the risk.

## 4 From Estimation to Learning

We now turn our attention to unsupervised learning, i.e., minimizing $R(\theta)$ over $\theta \in \mathbf{R}^d$. Unsupervised learning is impossible without some additional information, since even if we could learn the $k$ classes, we wouldn't know which class had which label (this is the same as the class permutation issue from before). Thus we assume that we have a small amount of information to break this symmetry:

**Assumption 2** (Seed Model)**.** *We have access to a "seed model" $\theta_0$ such that $\tilde{R}(\theta_0) = R(\theta_0)$.*

Assumption 2 is very weak — it merely asks for $\theta_0$ to be aligned with the true classes on average. We can obtain $\theta_0$ from a small amount of labeled data (semi-supervised learning) or by training in a nearby domain (domain adaptation). We define $\mathrm{gap}(\theta_0)$ to be the difference between $R(\theta_0)$ and the next smallest permutation of the classes–i.e., $\mathrm{gap}(\theta_0) \stackrel{\text{def}}{=} \min_{\sigma \neq \mathrm{id}} \mathbf{E}[L(\theta_0; x, \sigma(y)) - L(\theta_0; x, y)]$– which will affect the difficulty of learning.

For simplicity we will focus on the case of logistic regression, and show how to learn given only Assumptions 1 and 2. Our algorithm extends to general losses, as we show in Section F.

**Learning from moments.** Note that for logistic regression (Example 1), we have

$$R(\theta) = \mathbf{E}\Big[A(\theta; x) - \theta^\top \sum_{v=1}^{3} \phi_v(x_v, y)\Big] = \mathbf{E}[A(\theta; x)] - \theta^\top \bar{\phi}, \text{ where } \bar{\phi} \stackrel{\text{def}}{=} \sum_{v=1}^{3} \mathbf{E}[\phi_v(x_v, y)]. \quad (7)$$

From (7), we see that it suffices to estimate $\bar{\phi}$, after which all terms on the right-hand-side of (7) are known. Given an approximation $\hat{\phi}$ to $\bar{\phi}$ (we will show how to obtain $\hat{\phi}$ below), we can learn a near-optimal $\theta$ by solving the following convex optimization problem:

$$\hat{\theta} = \operatorname*{arg\,min}_{\|\theta\|_2 \leq \rho} \mathbf{E}[A(\theta; x)] - \theta^\top \hat{\phi}. \quad (8)$$

In practice we would need to approximate $\mathbf{E}[A(\theta; x)]$ by samples, but we ignore this for simplicity (it generally only contributes lower-order terms to the error). The reason for the $\ell^2$-constraint on $\theta$ is that it imparts robustness to the error between $\hat{\phi}$ and $\bar{\phi}$. In particular (see Section D for a proof):

**Lemma 1.** *Suppose $\|\hat{\phi} - \bar{\phi}\|_2 \leq \epsilon$. Then the output $\hat{\theta}$ from (8) satisfies $R(\hat{\theta}) \leq \min_{\|\theta\|_2 \leq \rho} R(\theta) + 2\epsilon\rho$.*

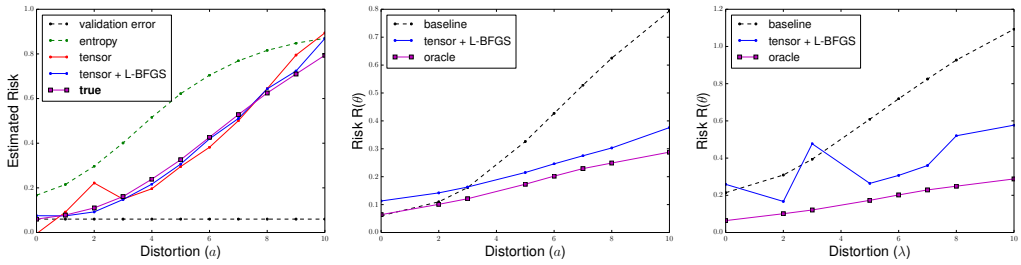
If the optimal $\theta^*$ has $\ell^2$-norm at most $\rho$, Lemma 1 says that $\hat{\theta}$ nearly minimizes the risk: $R(\hat{\theta}) \leq R(\theta^*) + 2\epsilon\rho$. The problem of learning $\theta$ thus reduces to computing a good estimate $\hat{\phi}$ of $\bar{\phi}$.

**Computing $\hat{\phi}$.** Estimating $\bar{\phi}$ can be done in a manner similar to how we estimated $R(\theta)$ in Section 2. In addition to the conditional risk matrix $M_v \in \mathbf{R}^{k \times k}$, we compute the *conditional moment matrix* $G_v \in \mathbf{R}^{dk \times k}$, which tracks the conditional expectation of $\phi_v$: $(G_v)_{i+(r-1)k,j} \stackrel{\text{def}}{=} \mathbf{E}[\phi_v(\theta; x_v, i)_r \mid y = j]$, where $r$ indexes $1, \ldots, d$. We then have $\bar{\phi}_r = \sum_{j=1}^k \pi_j \sum_{v=1}^3 (G_v)_{j+(r-1)k,j}$.

As in Theorem 1, we can solve for $G_1$, $G_2$, and $G_3$ using a tensor factorization similar to (4), though some care is needed to avoid explicitly forming the $(kd) \times (kd) \times (kd)$ tensor that would result (since $\mathcal{O}\left(k^3 d^3\right)$ memory is intractable for even moderate values of $d$). We take a standard approach based on random projections (Halko et al., 2011) and described in Section 6.1.2 of Anandkumar et al. (2013). We refer the reader to the aforementioned references for details, and cite only the resulting sample complexity and runtime, which are both roughly $d$ times larger than in Theorem 1.

**Theorem 2.** *Suppose that Assumptions 1 and 2 hold. Let $\delta < 1$ and $\epsilon < \min(1, \mathrm{gap}(\theta_0))$. Then, given $m = \mathrm{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right) \cdot \frac{\log(2/\delta)}{\epsilon^2}$ samples, where $\lambda$ and $\tau$ are as defined in (3),*

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




(a) (b) (c)

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

*with probability* $1-\delta$ *we can recover* $M_v$ *and* $\pi$ *to error* $\epsilon$*, and* $G_v$ *to error* $(B/\tau)\epsilon$*, where* $B^2 = \mathbf{E}[\sum_{i,v} \|\phi_v(x_v, i)\|_2^2]$ *measures the* $\ell^2$*-norm of the features. The algorithm runs in time* $\mathcal{O}\left(d\left(m + \text{poly}(k)\right)\right)$*, and the errors are in Frobenius norm for* $M$ *and* $G$*, and* $\infty$*-norm for* $\pi$*.*

See Section E for a proof sketch. Whereas before we estimated the risk matrix $M_v$ to error $\epsilon$, now we estimate the gradient matrix $G_v$ (and hence $\bar{\phi}$) to error $(B/\tau)\epsilon$. To achieve error $\epsilon$ in estimating $G_v$ requires $(B/\tau)^2 \cdot \text{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right) \frac{\log(2/\delta)}{\epsilon^2}$ samples, which is $(B/\tau)^2$ times as large as in Theorem 1. The quantity $(B/\tau)^2$ typically grows as $\mathcal{O}(d)$, and so the sample complexity needed to estimate $\bar{\phi}$ is typically $d$ times larger than the sample complexity needed to estimate $R$. This matches the behavior of the supervised case where we need $d$ times as many samples for learning as compared to (supervised) risk estimation of a fixed model.

**Summary.** We have shown how to perform unsupervised logistic regression, given only a seed model $\theta_0$. This enables unsupervised learning under fairly weak assumptions (only the multi-view and seed model assumptions) even for mis-specified models and zero train-test overlap, and without assuming covariate shift. See Section F for learning under more general losses.

## 5 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. To create an image $I$, we sample a class in $\{0, \ldots, 9\}$, then sample 3 images $I_1$, $I_2$, $I_3$ at random from that class, letting every third pixel in $I$ come from the respective image $I_v$. This guarantees there are 3 conditionally independent views. To explore train-test variation, we dim pixel $p$ in the image by $\exp\left(a\left(\|p - p_0\|_2 - 0.4\right)\right)$, where $p_0$ is the image center and distances are normalized to be at most 1. We show example images for $a = 0$ (train) and $a = 5$ (a possible test distribution) in Figure 3.

**Risk estimation.** We use Algorithm 1 to perform unsupervised risk estimation for a model trained on $a = 0$, testing on various values of $a \in [0, 10]$. 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 $\pi$ and $M$ in (4), we first use the tensor power method implemented by Chaganty and Liang (2013) to initialize, and then locally minimize a weighted $\ell^2$-norm of the moment errors in (4) using L-BFGS. We compared with two other methods: (i) validation error from held-out samples (which would be valid if train = test), and (ii) predictive entropy $\sum_j -p_\theta(j \mid x) \log p_\theta(j \mid x)$ on the test set (which would be valid if the predictions were 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.

**Unsupervised domain adaptation.** We next evaluate our learning algorithm in an unsupervised domain adaptation setting, where we receive labeled training data at $a = 0$ and unlabeled test data at a different value of $a$. We use the training data to obtain a seed model $\theta_0$, and then perform

unsupervised learning (Section 4), setting $\rho = 10$ in (8). The results are shown in Figure 4b. For small values of $a$, our algorithm performs worse than the baseline of directly using $\theta_0$, likely due to finite-sample effects. However, our algorithm is far more robust as $a$ increases, and tracks the performance of an oracle that was trained on the same distribution as the test examples.

Because we only need to provide our algorithm with a seed model for disentangling the classes, we do not need much data when training $\theta_0$. To verify this, we tried obtaining $\theta_0$ from only 300 labeled examples. Tensor decomposition sometimes led to bad initializations in this limited data regime, in which case we obtained a different $\theta_0$ by training with a smaller step size. The results are shown in Figure 4c. Our algorithm generally performs well, but has higher variability than before, seemingly due to higher condition number of the matrices $M_v$.

**Summary.** Our experiments show that given 3 views, we can estimate the risk and perform unsupervised domain adaptation, even with limited labeled data from the source domain.

## 6 Discussion

We have presented a method for estimating the risk from unlabeled data, which relies only on conditional independence structure and hence makes no parametric assumptions about the true distribution. Our approach applies to a large family of losses and extends beyond classification tasks to conditional random fields. We can also perform unsupervised learning given only a seed model that can distinguish between classes in expectation; the seed model can be trained on a related domain, on a small amount of labeled data, or any combination of the two, and thus provides a pleasingly general formulation highlighting the similarities between domain adaptation and semi-supervised learning.

Previous approaches to domain adaptation and semi-supervised learning have also exploited multi-view structure. Given two views, Blitzer et al. (2011) perform domain adaptation with zero source/target overlap (covariate shift is still assumed). Two-view approaches (e.g. co-training and CCA) are also used in semi-supervised learning (Blum and Mitchell, 1998; Ando and Zhang, 2007; Kakade and Foster, 2007; Balcan and Blum, 2010). These methods all assume some form of low noise or low regret, as do, e.g., transductive SVMs (Joachims, 1999). By focusing on the central problem of risk estimation, our work connects multi-view learning approaches for domain adaptation and semi-supervised learning, and removes covariate shift and low-noise/low-regret assumptions (though we make stronger independence assumptions, and specialize to discrete prediction tasks).

In addition to reliability and unsupervised learning, our work is motivated by the desire to build *machine learning systems with contracts*, a challenge recently posed by Bottou (2015); the goal is for machine learning systems to satisfy a well-defined input-output contract in analogy with software systems (Sculley et al., 2015). Theorem 1 provides the contract that under the 3-view assumption the test error is close to our estimate of the test error; the typical (weak) contract of ML systems is that if train and test are similar, then the test error is close to the training error. One other interesting contract is to provide prediction *regions* that contain the truth with probability $1 - \epsilon$ (Shafer and Vovk, 2008; Khani et al., 2016), which includes abstaining when uncertain as a special case (Li et al., 2011).

The most restrictive part of our framework is the three-view assumption, which is inappropriate if the views are not completely independent or if the data have structure that is not 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 end with the following questions:

**Open question.** In the 3-view setting, suppose the views are not completely independent. Is it still possible to estimate the risk? How does the degree of dependence affect the number of views needed?

**Open question.** Given only two independent views, can one obtain an upper bound on the risk $R(\theta)$?

The results of this paper have caused us to adopt the following perspective: To leverage 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 truly needed for learning is understanding the *structure* of a domain.

**Acknowledgments.** This research was supported by a Fannie & John Hertz Foundation Fellowship, a NSF Graduate Research Fellowship, and a Future of Life Institute grant.

## References


A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden Markov models. In *COLT*, 2012.

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

T. W. Anderson and H. Rubin. Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, pages 46–63, 1949.

T. W. Anderson and H. Rubin. The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, pages 570–582, 1950.

R. K. Ando and T. Zhang. Two-view feature generation model for semi-supervised learning. In *COLT*, 2007.

K. Balasubramanian, P. Donmez, and G. Lebanon. Unsupervised supervised learning II: Margin-based classification without labels. *JMLR*, 12:3119–3145, 2011.

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

S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *NIPS*, pages 137–144, 2006.

J. Blitzer, S. Kakade, and D. P. Foster. Domain adaptation with coupled subspaces. In *AISTATS*, 2011.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.

L. Bottou. Two high stakes challenges in machine learning. Invited talk at ICML, 2015.

A. Chaganty and P. Liang. Spectral experts for estimating mixtures of linear regressions. In *ICML*, 2013.

A. Chaganty and P. Liang. Estimating latent-variable graphical models using moments and likelihoods. In *ICML*, 2014.

P. Comon, X. Luciani, and A. L. D. Almeida. Tensor decompositions, alternating least squares and other tales. *Journal of Chemometrics*, 23(7):393–405, 2009.

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

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

P. Donmez, G. Lebanon, and K. Balasubramanian. Unsupervised supervised learning I: Estimating classification and regression errors without labels. *JMLR*, 11:1323–1351, 2010.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *COLT*, 2010.

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

E. Fetaya, B. Nadler, A. Jaffe, Y. Kluger, and T. Jiang. Unsupervised ensemble learning with dependent classifiers. In *AISTATS*, pages 351–360, 2016.

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

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 1982.

L. P. Hansen. Uncertainty outside and inside economic models. *Journal of Political Economy*, 122(5), 2014.

D. Hsu, S. M. Kakade, and P. Liang. Identifiability and unmixing of latent parse trees. In *NIPS*, 2012.

A. Jaffe, B. Nadler, and Y. Kluger. Estimating the accuracies of multiple classifiers without labeled data. In *AISTATS*, pages 407–415, 2015.

T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, 1999.

F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *ICML*, 2016.

S. M. Kakade and D. P. Foster. Multi-view regression via canonical correlation analysis. In *COLT*, 2007.

F. Khani, M. Rinard, and P. Liang. Unanimous prediction for 100% precision with application to learning semantic mappings. In *ACL*, 2016.

V. Kuleshov, A. Chaganty, and P. Liang. Tensor factorization via matrix factorization. In *AISTATS*, 2015.

L. D. Lathauwer. A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM Journal of Matrix Analysis and Applications*, 28(3):642–666, 2006.

L. Li, M. L. Littman, T. J. Walsh, and A. L. Strehl. Knows what it knows: a framework for self-aware learning. *Machine learning*, 82(3):399–443, 2011.

Y. Li and Z. Zhou. Towards making unlabeled data never hurt. *IEEE TPAMI*, 37(1):175–188, 2015.

P. Liang and D. Klein. Analyzing the errors of unsupervised learning. In *HLT/ACL*, 2008.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.

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

K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Learning to classify text from labeled and unlabeled documents. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 1998.

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

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

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

J. D. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, 1958.

J. D. Sargan. The estimation of relationships with autocorrelated residuals by the use of instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 91–105, 1959.

D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J. Crespo, and D. Dennison. Hidden technical debt in machine learning systems. In *NIPS*, pages 2494–2502, 2015.

G. Shafer and V. Vovk. A tutorial on conformal prediction. *JMLR*, 9:371–421, 2008.

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

J. Steinhardt, G. Valiant, and S. Wager. Memory, communication, and statistical queries. In *COLT*, 2016.

N. Tomizawa. On some techniques useful for solution of transportation network problems. *Networks*, 1971.

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

## A Hinge Loss and Assumption 1

In Section 2 we stated that the hinge loss does not satisfy Assumption 1. In this section we explain why.

To be explicit, suppose that $x = (x_1, x_2, x_3)$, where the $x_v$ are independent conditioned on $y$, and let $L(\theta; x, y)$ be the multiclass hinge loss:

$$L(\theta; x, y) = \max_{j \neq y} \max \left( 1 + \theta^\top \sum_{v=1}^{3} \left( \phi(x_v, y) - \phi(x_v, j) \right), 0 \right). \tag{9}$$

To satisfy Assumption 1, $L(\theta; x, y)$ should decompose as

$$L(\theta; x, y) = A(\theta; x) - \sum_{v=1}^{3} f_v(\theta; x_v, y). \tag{10}$$

In particular, (10) implies that the dependence on $y$ should be additive over the views $v$. In (9), however, the $\max$ couples all of the views together, so that the decomposition (10) does not hold.

## B Details of Computing $\tilde{R}$ from $M$ and $\pi$

In this section we show how, given $M$, and $\pi$, we can efficiently compute

$$\tilde{R}(\theta) = \mathbf{E}[A(\theta; x)] - \max_{\sigma \in \mathrm{Sym}(k)} \sum_{j=1}^{k} \pi_{\sigma(j)} \sum_{v=1}^{3} (M_v)_{j,\sigma(j)}. \tag{11}$$

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

$$X_{i,j} = \pi_i \sum_{v=1}^{3} (M_v)_{j,i}. \tag{12}$$

Then we are looking for the permutation $\sigma$ 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 $\sigma$ using any maximum-weight matching algorithm such as the Hungarian algorithm, which runs in $\mathcal{O}\left(k^3\right)$ time (Tomizawa, 1971; Edmonds and Karp, 1972).

## C Proof of Theorem 1

**Preliminary reductions.** Our goal is to estimate $M$ and $\pi$ to error $\epsilon$ (with probability of failure $1 - \delta$) in $\mathrm{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right) \cdot \frac{\log(2/\delta)}{\epsilon^2}$ samples. Note that if we can estimate $M$ and $\pi$ to error $\epsilon$ with any fixed probability of success $1 - \delta_0 \geq \frac{3}{4}$, then we can amplify the probability of success to $1 - \delta$ at the cost of $\mathcal{O}\left(\log(2/\delta)\right)$ times as many samples (the idea is to make several independent estimates, then throw out any estimate that is more than $2\epsilon$ away from at least half of the others; all the remaining estimates will then be within distance $3\epsilon$ of the truth with high probability).

**Estimating $M$.** Estimating $\pi$ and $M$ 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. Here $\kappa$ denotes condition number (which is the ratio of $\sigma_1(M)$ to $\sigma_k(M)$, since all matrices in question have $k$ columns).

**Theorem 3** (Anandkumar et al. (2012))**.** *Let $P_{v,v'} \stackrel{\text{def}}{=} \mathbf{E}[h_v(x) \otimes h_{v'}(x)]$, and $P_{1,2,3} \stackrel{\text{def}}{=} \mathbf{E}[h_1(x) \otimes h_2(x) \otimes h_3(x)]$. 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 $\ell^2$-norm of $T$ after unrolling $T$ to a vector (e.g., when $T$ is a matrix $\|T\|_F$ is the Frobenius norm). Suppose that, for some constants $C_{1,2}$, $C_{1,3}$, $C_{1,2,3}$, we have:*

- $\mathbf{P}\left[\|\hat{P}_{v,v'} - P_{v,v'}\|_2 \leq C_{v,v'}\sqrt{\frac{1}{\delta m}}\right] \geq 1-\delta$ *for* $\{v,v'\} \in \{\{1,2\},\{1,3\}\}$, *and*
- $\mathbf{P}\left[\|\hat{P}_{1,2,3} - P_{1,2,3}\|_F \leq C_{1,2,3}\sqrt{\frac{1}{\delta m}}\right] \geq 1-\delta$.

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

$$\sqrt{\frac{k}{\delta m}} \leq C \cdot \frac{\min_{j\neq j'} \|(M_3^\top)_j - (M_3^\top)_{j'}\|_2 \cdot \sigma_k(P_{1,2})}{C_{1,2,3}\cdot k^5 \cdot \kappa(M_1)^4} \cdot \frac{\delta}{\log(k/\delta)} \cdot \epsilon$$
$$\sqrt{\frac{1}{\delta m}} \leq C \cdot \min\left\{\frac{\min_{j\neq j'} \|(M_3^\top)_j - (M_3^\top)_{j'}\|_2 \cdot \sigma_k(P_{1,2})^2}{C_{1,2}\cdot \|P_{1,2,3}\|_F \cdot k^5 \cdot \kappa(M_1)^4} \cdot \frac{\delta}{\log(k/\delta)}, \frac{\sigma_k(P_{1,3})}{C_{1,3}}\right\} \cdot \epsilon$$

*for some* $\epsilon < 1$, *then with probability at least* $1-5\delta$, *we can output* $\hat{M}_3$ *with the following guarantee: there exists a permutation* $\sigma \in \mathrm{Sym}(k)$ *such that for all* $j \in \{1,\ldots,k\}$,

$$\|(M_3^\top)_j - (\hat{M}_3^\top)_{\sigma(j)}\|_2 \leq \max_{j'} \|(M_3^\top)_{j'}\|_2 \cdot \epsilon. \tag{13}$$

By symmetry, we can use Theorem 3 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 a probability of success of $1 - \frac{5}{180}$ for recovering each individual $M_v$, and hence an overall probability of success of $1 - \frac{3\cdot 5}{180} = 11/12$ for this part of the proof.

We now analyze the rate of convergence implied by Theorem 3. Note that by Chebyshev's inequality we can take $C_{1,2,3} = \mathcal{O}\left(\sqrt{\mathbf{E}[\|h_1\|_2^2\|h_2\|_2^2\|h_3\|_2^2]}\right)$, and similarly $C_{v,v'} = \mathcal{O}\left(\sqrt{\mathbf{E}[\|h_v\|_2^2\|h_{v'}\|_2^2]}\right)$. Next, observe that Theorem 3 implies that we can estimate the $M_v$ to error $\epsilon$ given $Z/\epsilon^2$ samples, where $Z$ is polynomial in the following quantities:

1. $k$,
2. $\max_{v=1}^3 \kappa(M_v)$, where $\kappa$ denotes condition number,
3. $\frac{\sqrt{\mathbf{E}[\|h_1\|_2^2\|h_2\|_2^2\|h_3\|_2^2]}}{\left(\min_{j,j'} \|(M_v^\top)_j - (M_v^\top)_{j'}\|_2\right)\cdot \sigma_k(P_{v',v''})}$, where $(v,v',v'')$ is a permutation of $(1,2,3)$,
4. $\frac{\|P_{1,2,3}\|_2}{\left(\min_{j,j'} \|(M_v^\top)_j - (M_v^\top)_{j'}\|_2\right)\cdot \sigma_k(P_{v',v''})}$, where $(v,v',v'')$ is as before, and
5. $\frac{\sqrt{\mathbf{E}[\|h_v\|_2^2\|h_{v'}\|_2^2]}}{\sigma_k\left(P_{v,v'}\right)}$.
6. $\max_{j,v} \|(M_v^\top)_j\|_2$.

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

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

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

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

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

(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

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

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; we end up obtaining a bound of $\frac{\tau^2}{\pi_{\min}\lambda^2}$.

(6) We have $\|(M_v^\top)_j\|_2 = \|\mathbf{E}[h_v \mid y = j]\|_2 \leq \sqrt{\mathbf{E}[\|h_v\|_2^2 \mid y = j]} \leq \tau$.

In sum, we have shown that with probability $\frac{11}{12}$ we can estimate each $M_v$ to column-wise $\ell^2$ error $\epsilon$ using $\mathrm{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right)/\epsilon^2$ samples; since there are only $k$ columns, we can make the total (Frobenius) error be at most $\epsilon$ while still using $\mathrm{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right)/\epsilon^2$ samples. It now remains to estimate $\pi$.

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

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

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 $\|E_1\|_F \leq \sqrt{k}\max_j \|(\hat{M_1}^\top)_j - (M_1^\top)_j\|_2 = \mathrm{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right)/\sqrt{m}$. Therefore:

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

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) $\|\mathbf{E}[h_1]\|_2$: as before, we have $\|\mathbf{E}[h_1]\|_2 \leq \sqrt{\mathbf{E}[\|h_1\|_2^2]} \leq \tau$.

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

(iv) $\|\hat{\mathbf{E}}[h_1] - \mathbf{E}[h_1]\|_2$: with any fixed probability (in this case, $11/12$), this term is $\mathcal{O}\left(\sqrt{\frac{\mathbf{E}[\|h_1\|_2^2]}{m}}\right) = \mathcal{O}\left(\frac{\tau}{\sqrt{m}}\right)$ by Chebyshev's inequality.

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

Since we have shown that we can estimate each of $M$ and $\pi$ individually with probability $\frac{11}{12}$, we can estimate them jointly with probability $\frac{5}{6} > \frac{3}{4}$, thus completing the proof.

## D Proof of Lemma 1

Let $B(\rho) = \{\theta \mid \|\theta\|_2 \leq \rho\}$. First note that $|\theta^\top(\hat{\phi} - \bar{\phi})| \leq \|\theta\|_2\|\hat{\phi} - \bar{\phi}\|_2 \leq \epsilon\rho$ for all $\theta \in B(\rho)$. Letting $\tilde{\theta}$ denote the minimizer of $R(\theta)$ over $B(\rho)$, we obtain

$$
\begin{aligned}
R(\hat{\theta}) &= \mathbf{E}[A(\hat{\theta}; x)] - \hat{\theta}^\top\bar{\phi} && (14)\\
&\leq \mathbf{E}[A(\hat{\theta}; x)] - \hat{\theta}^\top\hat{\phi} + \epsilon\rho && (15)\\
&\overset{(i)}{\leq} \mathbf{E}[A(\tilde{\theta}; x)] - \tilde{\theta}^\top\hat{\phi} + \epsilon\rho && (16)\\
&\leq \mathbf{E}[A(\tilde{\theta}; x)] - \tilde{\theta}^\top\bar{\phi} + 2\epsilon\rho && (17)\\
&= R(\tilde{\theta}) + 2\epsilon\rho, && (18)
\end{aligned}
$$

as claimed, where (i) is because $\hat{\theta}$ is the minimizer of $\mathbf{E}[A(\theta; x)] - \theta^\top\hat{\phi}$, and the remaining inequalities follow from the observation above that $|\theta^\top(\hat{\phi} - \bar{\phi})| \leq \epsilon\rho$.

## E Proof of Theorem 2

We provide here a sketch of the proof of Theorem 2, which essentially follows from the proof of Theorem 1. We note that Theorem 7 of Anandkumar et al. (2012) (and hence Theorem 1 above) does not require that the $M_v$ be $k \times k$, but only that they have $k$ *columns* (the number of rows can be arbitrary). It thus applies for any matrix $M_v'$ whose $j$th columns is obtained as $\mathbf{E}[h_v'(x_v) \mid j]$ for some $h_v' : \mathcal{X}_v \to \mathbf{R}^{d'}$. In our specific case, we will take $h_v' : \mathcal{X}_v \to \mathbf{R}^{k(d+1)}$, where the first $k$ coordinates of $h'(x_v)$ are equal to $h(x_v)$ (i.e., $(f_v(x_v, i))_{i=1}^k$), and the remaining $kd$ coordinates of $h'(x_v)$ are equal to $\frac{\tau}{B}\phi_v(x_v, i)_r$ as in the definition of $G_v$, where the difference is that we have scaled by a factor of $\frac{\tau}{B}$. Note that in this case $M_v' = \begin{bmatrix} M_v \\ \frac{\tau}{B}G_v \end{bmatrix}$. We let $\lambda'$ and $\tau'$ denote the values of $\lambda$ and $\tau$ for $M'$ and $h'$.

Since $M_v$ is a submatrix of $M_v'$, we have $\sigma_k(M_v') \geq \sigma_k(M_v)$, so $\lambda' \geq \lambda$. For $\tau'$, note that

$$
\begin{aligned}
\tau' &= \mathbf{E}\left[\sum_v \|h_v'(x_v)\|_2^2\right] && (19)\\
&= \mathbf{E}\left[\sum_v \|h_v(x_v)\|_2^2 + \frac{\tau^2}{B^2}\sum_{v,i}\|\phi_v(x_v, i)\|_2^2\right] && (20)\\
&= \tau^2 + \frac{\tau^2}{B^2}\mathbf{E}\left[\sum_{v,i}\|\phi_v(x_v, i)\|_2^2\right] && (21)\\
&= 2\tau^2, && (22)
\end{aligned}
$$

so $\tau' \leq \sqrt{2}\tau$. Since $(\lambda')^{-1} = \mathcal{O}(\lambda^{-1})$ and $\tau' = \mathcal{O}(\tau)$, we still obtain a sample complexity of $\text{poly}\left(k, \pi_{\min}^{-1}, \lambda^{-1}, \tau\right) \cdot \frac{\log(2/\delta)}{\epsilon^2}$. Since $\theta_0$ satisfies Assumption 2, we can recover the correct

permutation of the columns of $M_v$ (and hence also of $G_v$, since they are permuted in the same way), which completes the proof.

# F Learning with General Losses

In Section 4, we formed the conditional moment matrix $G_v$, which stored the conditional expectation $\mathbf{E}[\phi_v(x_v, i) \mid y = j]$ for each $j$ and $i$. However, there was nothing special about computing $\phi$ (as opposed to some other moments), and for general losses we can form the conditional gradient matrix $G_v(\theta)$, defined by

$$G_v(\theta)_{i+kr,j} = \mathbf{E}\left[\frac{\partial}{\partial \theta_r} f_v(\theta; x_v, i) \mid y = j\right]. \tag{23}$$

Theorem 2 applies identically to the matrix $G_v(\theta)$ at any fixed $\theta$. We can then compute the gradient $\nabla_\theta R(\theta)$ using the relationship

$$\frac{\partial}{\partial \theta_r} R(\theta) = \mathbf{E}\left[\frac{\partial}{\partial \theta_r} A(\theta; x)\right] - \sum_{j=1}^{k} \pi_j \sum_{v=1}^{3} G_v(\theta)_{j+kr,j}. \tag{24}$$

For clarity, we also use $M_v(\theta)$ to denote the conditional risk matrix at a value $\theta$. To compute the gradient $\nabla_\theta R(\theta)$, we jointly estimate $M_v(\theta_0)$ and $G_v(\theta)$ (note the differing arguments of $\theta_0$ vs. $\theta$). Since the seed model assumption (Assumption 2) allows us to recover the correct column permutation for $M_v(\theta_0)$, estimating $G_v(\theta)$ jointly with $M_v(\theta_0)$ ensures that we recover the correct column permutation for $G_v(\theta)$ as well.

The final ingredient in learning $\theta$ is any gradient descent procedure that is robust to errors in the gradient (so that after $T$ steps with error $\epsilon$ on each step, the total error is $\mathcal{O}(\epsilon)$ and not $\mathcal{O}(\epsilon T)$). Fortunately, this is the case for many gradient descent algorithms, including any algorithm that can be expressed as mirror descent (we omit the details because they are somewhat beyond our scope, but refer the reader to Lemma 21 of Steinhardt et al. (2016) for a proof of this in the case of exponentiated gradient).

The general learning algorithm is given in Algorithm 2:

**Algorithm 2** General algorithm for learning via gradient descent.

Parameters: step size $\eta$
Input: unlabeled samples $x^{(1)}, \ldots, x^{(m)} \sim p^*(x)$, seed model $\theta_0$
$z^{(1)} \leftarrow 0 \in \mathbf{R}^d$
**for** $t = 1$ **to** $T$ **do**
    $\theta^{(t)} \leftarrow \arg\min_\theta \frac{1}{2\eta}\|\theta - \theta_0\|_2^2 - \theta^\top z^{(t)}$
    Compute $(M_v^{(t)}, G_v^{(t)})$ by jointly estimating $M_v(\theta_0), G_v(\theta^{(t)})$ from $x^{(1:m)}$.
    **for** $r = 1$ **to** $d$ **do**
        $g_r \leftarrow \frac{1}{m}\sum_{i=1}^{m} \frac{\partial}{\partial \theta_r} A(\theta^{(t)}; x^{(i)}) - \sum_{j=1}^{k} \pi_j \sum_{v=1}^{3} (G_v^{(t)})_{j+kr,j}$
        $z_r^{(t+1)} \leftarrow z_r^{(t)} + g_r$
    **end for**
**end for**
Output $\frac{1}{T}\left(\theta^{(1)} + \cdots + \theta^{(T)}\right)$.