# Spectral M-estimation with Applications to Hidden Markov Models

**Dustin Tran**
Harvard University

**Minjae Kim**
Harvard University

**Finale Doshi-Velez**
Harvard University

## Abstract

Method of moment estimators exhibit appealing statistical properties, such as asymptotic unbiasedness, for nonconvex problems. However, they typically require a large number of samples and are extremely sensitive to model misspecification. In this paper, we apply the framework of M-estimation to develop both a generalized method of moments procedure and a principled method for regularization. Our proposed M-estimator obtains optimal sample efficiency rates (in the class of moment-based estimators) and the same well-known rates on prediction accuracy as other spectral estimators. It also makes it straightforward to incorporate regularization into the sample moment conditions. We demonstrate empirically the gains in sample efficiency from our approach on hidden Markov models.

## 1 Introduction

Developing expressive latent variable models is a fundamental task in statistics and machine learning. However, performing parameter estimation with statistical guarantees remains challenging; in practice, optimization techniques such as the EM algorithm (Dempster et al., 1977) are used to find local solutions to approximate the maximum likelihood estimate (MLE) or maximum a posteriori solution.

Recently, inference techniques based on the method of moments (Pearson, 1894), coined as *spectral learning*, have gained interest because they provide consistent estimators for many classes of models, such as hidden Markov models (Hsu et al., 2012), predictive state representations (Boots et al., 2010), latent tree models (Parikh et al., 2011), weighted automata (Balle and Mohri, 2012), mixture models (Anandkumar et al., 2014b), and mixed membership stochastic blockmodels (Anandkumar et al., 2014a). Spectral methods operate by deriving low-order moment conditions on the model—such as the mean and covariance—and matching these to moments of the observed data. Often this moment-matching process can be solved efficiently with linear algebra routines and can allow for parameter recovery in settings where row-level data is unwieldy to work with (e.g. streaming data) or unavailable (e.g. an institution may only be willing to release summary statistics).

However, current spectral methods are extremely sensitive to poorly-estimated moments and model misspecification. The former problem can be addressed, in part, by robust estimation methods of covariances (Negahban and Wainwright, 2011)—though robust estimation for higher order moments remains an open challenge. When the rank of the model is set too low—a form of model misspecification—Kulesza et al. (2014) demonstrate that naive methods can lead to arbitrarily large prediction error. In practice, there are many occasions where we may wish to learn a low-rank approximation to a complex system.

In contrast, parameters learned from maximum likelihood and other optimization-based estimators are robust (assuming global optimum), as they minimize the Kullback-Leibler divergence from the considered model class to the true data distribution (White, 1982) and can in certain cases achieve consistency (Gourieroux et al., 1984). With finite samples, optimization-based estimators can achieve reasonable variances (Godambe, 1960).

Is such robustness possible for spectral methods? Errors due to both poor moment estimates and model misspecification can be viewed as forms of overfitting. Various heuristics such as early stopping are considered in the literature (Mahoney and Orecchia, 2011), but they fundamentally break assumptions for the statistical guarantees, and are difficult to rigorously characterize; this leads to a disparity between theory and practice.

In this paper, we analyze spectral methods from their traditional–and more general—setting as an *M-estimator*. M-estimation has deep roots in robust statistics (see, e.g., Huber and Ronchetti (2009)). This connection emphasizes the relationship of spectral methods to well-established alternatives such as maximum likelihood. We use this connection to recover the desired properties—sample efficiency and balanced fitting. Specifically, our work makes the following contributions:

Appearing in Proceedings of the $19^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41. Copyright 2016 by the authors.

**Provably optimal sample efficiency with respect to the moments**. With the choice of weighted Frobenius norm as a metric on the moment conditions, the M-estimation procedure corresponds to the *generalized method of moments* (GMM), whose estimator is proven to be *statistically efficient* with respect to the information stored in the moments. Most practically, the GMM is sample efficient and is thus more adaptive to scenarios where the size of the data set is small to moderate or the data collection process results in imbalanced samples for estimation.

**Principled regularization for sparse estimation**. The setting of M-estimation is naturally conducive to penalization in order to regularize parameters, and it is commonly applied to perform robust estimation and variable selection (Owen, 2007; Lambert-Lacroix et al., 2011; Li et al., 2011). From the Bayesian perspective, this can be interpreted as placing priors on the parameters of interest, and where the log-likelihood is replaced by a more general, robust, function of the data and parameters. The proposed M-estimator automatically preserves the same bounds on the predictive accuracy as other spectral algorithms, while also achieving statistical efficiency.

We focus on the application of spectral M-estimation to hidden Markov models in our development of the theory (section 3) and empirical evaluation (section 5); we discuss extensions to other latent variable models in section 6.

## 2 Background

### 2.1 M-estimation

We first review M-estimation (Huber, 1973; Van der Vaart, 2000), which naturally generalizes the moment matching used in spectral methods. Let observations $\mathbf{X}_1, \dots, \mathbf{X}_N \in \mathcal{X}$ be generated from a distribution with unknown parameters $\boldsymbol{\theta}^* \in \Theta$. Consider minimizing the criterion

$$M_N(\boldsymbol{\theta}) = \sum_{n=1}^{N} m(\mathbf{X}_n, \boldsymbol{\theta}),$$

where $m(\cdot,\cdot) : \mathcal{X} \times \Theta \to \mathbb{R}$ are called the estimating functions (Godambe, 1976, 1991). The argument $\widehat{\boldsymbol{\theta}}^{\mathrm{M}}$ which minimizes the criterion is termed the *M-estimator*. Similarly, one may also consider *penalized M-estimation* in which one minimizes the criterion

$$M_N(\boldsymbol{\theta}) = \sum_{n=1}^{N} m(\mathbf{X}_n, \boldsymbol{\theta}) + \lambda P(\boldsymbol{\theta}), \tag{1}$$

where $m(\cdot,\cdot)$ is as before, $\lambda \in \mathbb{R}$ is fixed, and $P(\cdot) : \Theta \to \mathbb{R}$ is a specified penalty function on the parameters.

Let $M(\boldsymbol{\theta}) = \mathbb{E}[m(\mathbf{X}, \boldsymbol{\theta})]$. The M-estimator $\widehat{\boldsymbol{\theta}}^{\mathrm{M}}$ is consistent in that $\frac{1}{N} M_N(\boldsymbol{\theta})$ uniformly converges in probability to $M(\boldsymbol{\theta})$ as $N \to \infty$, and $\widehat{\boldsymbol{\theta}}^{\mathrm{M}}$ converges to $\boldsymbol{\theta}^*$ (or the closest projection, if $\boldsymbol{\theta}^*$ is not among the considered models). In the case of penalization, the intuition is that in the limit, the penalty term $P(\boldsymbol{\theta}^*)$ is dominated by the confidence one has from the data (as the first summation grows with $N$).

### 2.2 Generalized method of moments

A particular case of M-estimation is the *generalized method of moments* (GMM), developed in the econometrics literature (Burguete et al., 1982; Hansen, 1982). Given a vector-valued function $m(\cdot,\cdot) : \mathcal{X} \times \Theta \to \mathbb{R}^k$, the *moment conditions* form

$$M(\boldsymbol{\theta}^*) = \mathbb{E}[m(\mathbf{X}, \boldsymbol{\theta}^*)] = \mathbf{0},$$

where the expectation is taken with respect to the data distribution on $\mathbf{X}$. In practice, we use empirical estimates of the $k$ moment conditions using data, $\sum_{n=1}^{N} m(\mathbf{X}_n, \boldsymbol{\theta})$.[1]

In the setting where $k > |\Theta|$, the problem is overspecified and no root solution exists. One may best hope to find the set of parameters $\boldsymbol{\theta}^*$ which minimizes $\|\mathbb{E}[m(\mathbf{X}, \boldsymbol{\theta})]\|$ for some choice of norm $\|\cdot\|$. The GMM estimator $\widehat{\boldsymbol{\theta}}^{\mathrm{gmm}}$ is given by minimizing a weighted criterion function,

$$M_N(\boldsymbol{\theta}) = \left\| \sum_{n=1}^{N} m(\mathbf{X}_n, \boldsymbol{\theta}) \right\|_{\mathbf{W}}^2, \tag{2}$$

where for a positive definite matrix $\mathbf{W} \in \mathbb{R}^{k\times k}$, the *weighted norm* is $\|\mathbf{v}\|_{\mathbf{W}}^2 = \mathbf{v}^\top \mathbf{W} \mathbf{v}$ for $\mathbf{v} \in \mathbb{R}^k$.

Under standard assumptions, the estimator $\widehat{\boldsymbol{\theta}}^{\mathrm{gmm}}$ is consistent and asymptotically normal. Moreover, if we set $\mathbf{W} \propto \mathbb{E}[m(\mathbf{X}_n, \boldsymbol{\theta}^*) m(\mathbf{X}_n, \boldsymbol{\theta}^*)^\top]^{-1}$, then $\widehat{\boldsymbol{\theta}}^{\mathrm{gmm}}$ is statistically efficient in the class of consistent and asymptotically normal estimators *conditional on the moment conditions*. Therefore, if the moment conditions form a sufficient statistic of the data (as in the MLE), then the GMM estimator is optimal in that its variance asymptotically achieves the optimal Cramér-Rao lower bound. More generally, the GMM estimator achieves the Godambe information.

One can reformulate many, if not all, examples of spectral learning algorithms as special cases of M-estimation, and thus one can recover the set of parameters with maximal sample efficiency using the GMM estimator (Equation 2) and achieve certain robustness properties and regularization by sufficient penalization of the loss (Equation 1).

### 2.3 Hidden Markov models

For the remainder of this paper, we will focus on spectral estimation and associated statistical guarantees for hidden

[1] To simplify presentation, $m(\cdot,\cdot)$ is written as vector-valued to connect to moment estimation. Some simple swapping of symbols can recover the scalar-valued notation in M-estimation.

Markov models (HMMs)—applications to other latent variable models are discussed in Section 6. An HMM is defined by a 5-tuple $\{X, H, \mathbf{T}, \mathbf{O}, \boldsymbol{\pi}\}$ where $X$ is a set of $n$ discrete observations, $H$ is a set of $m$ discrete hidden states, $\boldsymbol{\pi} \in \mathbb{R}^m$ is the initial distribution over hidden states, and the transition $\mathbf{T} \in \mathbb{R}^{m\times m}$ and observation $\mathbf{O} \in \mathbb{R}^{n\times m}$ operators govern the dynamics of the system:

$$\begin{aligned}\mathbf{T}_{ij} &= \Pr(h_{t+1} = i \mid h_t = j),\\ \mathbf{O}_{ij} &= \Pr(x_t = i \mid h_t = j).\end{aligned}$$

Specifically, HMMS assume that given the hidden state $h_t$ at time $t$, the next state $h_{t+1}$ and the current observation $x_t$ is independent of any history before $h_t$.

We are interested in estimating the joint probabilities $\Pr(x_{1:t}) = \Pr(x_1, \dots, x_t)$ and the conditional probabilities $\Pr(x_t \mid x_{1:t-1})$. The model parameters $(\mathbf{T}, \mathbf{O}, \boldsymbol{\pi})$ can also be recovered in our setup, but directly estimating the parameters can be unstable and requires additional assumptions such as coherence (Anandkumar et al., 2014a; Mossel and Roch, 2005).

If $\mathbf{T}$ and $\mathbf{O}$ are full rank, and $\boldsymbol{\pi} > 0$ for all hidden states $h \in [m]$, then Hsu et al. (2012) show that the following statistics are sufficient to consistently estimate the joint probabilities:

$$\begin{aligned}\mathbf{P}_1 &\in \mathbb{R}^n & [\mathbf{P}_1]_i &= \Pr(x_1 = i),\\ \mathbf{P}_{2,1} &\in \mathbb{R}^{n\times n} & [\mathbf{P}_{2,1}]_{ij} &= \Pr(x_2 = i, x_1 = j),\\ \mathbf{P}_{3,x,1} &\in \mathbb{R}^{n\times n} & [\mathbf{P}_{3,x,1}]_{ij} &= \Pr(x_3 = i, x_2 = x, x_1 = j),\end{aligned} \tag{3}$$

where $\mathbf{P}_{3,x,1}$ is written for all $x \in [n]$. We term these statistics *observable*, as they can be estimated directly using triplets of the observations.

Specifically, Hsu et al. (2012) define the spectral model parameters $(\mathbf{b}_1^{\text{spec}}, \mathbf{b}_\infty^{\text{spec}}, \mathbf{B}_x^{\text{spec}})$ as follows. Let $\mathbf{U} \in \mathbb{R}^{n\times m}$ be a matrix such that $\mathbf{U}^\top\mathbf{O}$ is invertible—typically, it is the left singular vectors corresponding to the $m$ largest singular values of $\mathbf{P}_{2,1}$—and set

$$\begin{aligned}\mathbf{b}_1^{\text{spec}} &= \mathbf{U}^\top \mathbf{P}_1,\\ \mathbf{b}_\infty^{\text{spec}} &= (\mathbf{P}_{2,1}^\top \mathbf{U})^\dagger \mathbf{P}_1,\\ \mathbf{B}_x^{\text{spec}} &= \mathbf{U}^\top \mathbf{P}_{3,x,1} (\mathbf{U}^\top \mathbf{P}_{2,1})^\dagger \quad \forall x \in [n].\end{aligned} \tag{4}$$

where $\mathbf{A}^\dagger$ denotes the pseudoinverse of $\mathbf{A}$. Then the joint probability satisfies

$$\Pr(x_{1:t}) = \mathbf{b}_\infty^T \mathbf{B}_{x_t} \cdots \mathbf{B}_{x_1} \mathbf{b}_1. \tag{5}$$

Intuitively, one can think of $\mathbf{b}_1$ as the initial state vector in a projected observable representation space; the matrix $\mathbf{B}_x$ is an observable transition operator which propagates changes in this space; the vector $\mathbf{b}_\infty$ simply acts as a normalizer. From Equation 5, Hsu et al. (2012) demonstrated that the estimator $\widehat{\boldsymbol{\theta}}^{\text{spec}} = (\widehat{\mathbf{b}}_1^{\text{spec}}, \widehat{\mathbf{b}}_\infty^{\text{spec}}, \widehat{\mathbf{B}}_x^{\text{spec}})$, which is constructed from the empirical statistics $\widehat{\mathbf{P}}_1, \widehat{\mathbf{P}}_{2,1}, \widehat{\mathbf{P}}_{3,x,1}$, is asymptotically unbiased as the empirical statistics become exact in the limit. Moreover, the number of observations required to achieve a fixed level of accuracy is only polynomial in the length of the sequence, $t$.

## 3 Spectral M-estimation

Following the results of spectral methods (Hsu et al., 2012; Boots et al., 2010; Balle and Mohri, 2012; Cohen et al., 2012; Arora et al., 2012), it is natural to consider the underlying framework for its methodology, and how it connects to techniques for maximum likelihood estimation. To address this, we start by deriving the usual spectral estimator (4) from the M-estimation setting.

### 3.1 Spectral M-estimator

Denote the parameter triplet $\boldsymbol{\theta} = (\mathbf{b}_1, \mathbf{B}, \mathbf{b}_\infty)$ and define the moment conditions

$$\begin{aligned}m_1(\boldsymbol{\theta}) &= \mathbf{b}_1 - \mathbf{P}_1,\\ m_\infty(\boldsymbol{\theta}) &= \mathbf{P}_{2,1}^\top \mathbf{b}_\infty - \mathbf{P}_1,\\ m_x(\boldsymbol{\theta}) &= \mathbf{P}_{3,x,1} - \mathbf{B}_x \mathbf{P}_{2,1} \quad \forall x \in [n].\end{aligned} \tag{6}$$

Let $\boldsymbol{\theta}^*$ denote the root solution $m_1(\boldsymbol{\theta}^*) = m_\infty(\boldsymbol{\theta}^*) = m_x(\boldsymbol{\theta}^*) = 0$. The vector $\mathbf{b}_1$ is trivially given by $\mathbf{P}_1$, and the solution of $\mathbf{b}_\infty$ to $m_\infty(\cdot)$ is simply the vector of ones, $\mathbf{1}_n$. Thus it suffices to estimate the tensor $\mathbf{B}$.

The standard approach in spectral methods (e.g., Hsu et al. (2012); Boots et al. (2010)) is to first observe that parameter triplets satisfying the joint probability in Equation 5 are equivalent up to a similarity transform: given the triplet $(\mathbf{b}_1, \{\mathbf{B}_x\}, \mathbf{b}_\infty)$ and an invertible matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$, the transformed triplet $(\mathbf{b}_1' = \mathbf{S}\mathbf{b}_1, \{\mathbf{B}_x' = \mathbf{S}\mathbf{B}_x\mathbf{S}^{-1}\}, \mathbf{b}_\infty' = \mathbf{S}^{-T}\mathbf{b}_\infty)$ provide the same quantities. Thus, what we are really interested in is not a unique set of parameters but an equivalence class—governed by the joint probability (5)—and which denote identical parameters up to a similarity transform. The moment conditions (6) are constructed such that the solution $\boldsymbol{\theta}^*$ defines a unique element in this equivalence class (and thus by M-estimation theory, the estimator is identifiable (Van der Vaart, 2000)).

We now formalize the connection to the usual spectral estimator as follows. Let $\mathbf{X} = \{\mathbf{X}_n = (x_{n1}, x_{n2}, x_{n3})\}$ denote the data set of $N$ triplets by which the observable representations $\mathbf{P}_1$, $\mathbf{P}_{2,1}$ and $\mathbf{P}_{3,x,1}$ are estimated. Define

$$M_N(\mathbf{B}) = \sum_{x,i,j\in[n]^3} ([\widehat{\mathbf{P}}_{3,x,1}]_{ij} - [\mathbf{B}_x]_{i\cdot}[\widehat{\mathbf{P}}_{2,1}]_{\cdot j})^2. \tag{7}$$

**Proposition 1** (Equivalence)**.** *Let $\widehat{\boldsymbol{\theta}}^{spec}$ denote the estimator using empirical statistics in Equation 4. Let $\widehat{\boldsymbol{\theta}}^{\text{M}}$ denote the*

*M-estimator given by*

$$\begin{aligned}\widehat{\mathbf{b}}_1^M &= \widehat{\mathbf{P}}_1, \\ \widehat{\mathbf{b}}_\infty^M &= \mathbf{1}_n, \\ \widehat{\mathbf{B}}^M &= \operatorname*{arg\,min}_{\mathbf{B}\in\mathbb{R}^{n\times n\times n}} M_N(\mathbf{B}).\end{aligned}$$

*Then $\widehat{\boldsymbol{\theta}}^{\mathrm{M}}$ is in the same equivalence class as $\widehat{\boldsymbol{\theta}}^{spec}$, so they provide the same probability estimates.*

Proposition 1 allows us to leverage both M-estimation theory and the usual finite sample bounds on accuracy given by Hsu et al. (2012). Specifically, the sample complexity of $\widehat{\boldsymbol{\theta}}^{\mathrm{M}}$ depends polynomially on the singular values $1/\sigma_m(\mathbf{P}_{2,1})$ and $1/\sigma_m(\mathbf{O})$, where $\sigma_m(\cdot)$ denotes the $m^{th}$ largest singular value of its matrix argument.

### 3.2 Regularized Spectral M-estimator: Low Rank Setting

Suppose there is a low rank constraint on the parameters, where $\operatorname{rank}(\mathbf{B}_x) \leq k$ for some $k < m$ and for all matrices $\mathbf{B}_x$. We may impose this constraint for computational tractability, to avoid the $\mathcal{O}(n^3)$ complexity of solving singular value decomposition associated with the dynamical system. It may also occur naturally: the maximal rank of $\mathbf{B}_x$ is $\operatorname{rank}(\mathbf{O}) = \operatorname{rank}(\mathbf{T}) \leq m$, and often the transition operators are low rank. Estimation with this constraint is known as *low rank spectral learning*, Kulesza et al. (2014) show that simply truncating $\mathbf{B}_x$ to a desired rank can lead to poor prediction. Following the M-estimation setting, we now derive a more robust estimator.

To optimize over an unconstrained Euclidean space, we first cast the low rank estimation problem in terms of matrix factorization. Let $\mathbf{B}_x = \mathbf{R}_x\mathbf{S}_x^\top$, where $\mathbf{R}_x, \mathbf{S}_x \in \mathbb{R}^{n\times k}$, and let $\mathbf{R}$ and $\mathbf{S}$ be tensors formed by the collections of matrices $\{\mathbf{R}_x\}$ and $\{\mathbf{S}_x\}$ respectively.

This leads to the criterion function

$$M_N(\mathbf{B}) = \sum_{x,i,j\in[n]^3} ([\widehat{\mathbf{P}}_{3,x,1}]_{ij} - [\mathbf{R}_x]_{i\cdot}\mathbf{S}_x^\top[\widehat{\mathbf{P}}_{2,1}]_{\cdot j})^2, \quad (8)$$

where we use the notation $\mathbf{A}_{i\cdot}$ (and respectively, $\mathbf{A}_{\cdot j}$) to represent the $i^{th}$ row (and $j^{th}$ column) of a matrix.

### 3.3 Regularized Spectral M-estimator: Additional Penalization

Given the M-estimation following Equation 8, we can generalize the procedure further by augmenting the criterion function with a penalty term,

$$M_N(\mathbf{R},\mathbf{S}) + \lambda P_\alpha(\mathbf{R},\mathbf{S}),$$

where $P_\alpha(\mathbf{R},\mathbf{S})$ is a specified penalty function with regularization parameter $\lambda$. However, in general, if $M_N(\mathbf{R},\mathbf{S})$ converges in probability to $M(\mathbf{R},\mathbf{S})$ as in the current setting, we must specify a suitable decaying schedule on the penalty function,

$$M_N(\mathbf{R},\mathbf{S}) + \lambda N^{-p} P_\alpha(\mathbf{R},\mathbf{S})$$

for fixed $p > 0$ (unlike traditional penalized M-estimation, the number of summations remains fixed as $N \to \infty$). Ideally, the penalty function should decay at the slowest possible rate, without affecting the convergence rate of the previous M-estimator (8). We choose $p$ as follows.

**Proposition 2.** *Let $\widehat{\boldsymbol{\theta}}^{\mathrm{M}}$ denote the M-estimator obtained by minimizing the criterion function*

$$M_N(\mathbf{R},\mathbf{S}) + \lambda N^{-p} P_\alpha(\mathbf{R},\mathbf{S}),$$

*where $p > 0$. Then the largest value of $p$ such that the convergence rate of $\widehat{\boldsymbol{\theta}}^{\mathrm{M}}$ does not change is $p = 1/2$.*

Trivially this is the case based on the asymptotic rate of the estimator. In practice, we consider losses of the form

$$\mathcal{L}(\mathbf{R},\mathbf{S}) = M_N(\mathbf{R},\mathbf{S}) + \lambda N^{-1/2}\|\mathbf{R}\|_1. \quad (9)$$

Penalizing only the first factor of $\mathbf{B}$ acts as a proxy for penalizing the observation operator $\mathbf{O}$; that is, by construction one can show that $\mathbf{B}_x = \mathbf{O}\mathbf{A}_x\mathbf{O}^\dagger$, where $\mathbf{A}_x = \mathbf{T}\operatorname{diag}(\mathbf{O}_{x,1},\dots,\mathbf{O}_{x,m})$. We will denote this final criterion function as $\mathcal{L}$ and its M-estimator as $\widehat{\boldsymbol{\theta}}^{\mathrm{M}}$, which also collects the two parameters $\widehat{\mathbf{b}}_1^{\mathrm{M}} = \widehat{\mathbf{P}}_1$ and $\widehat{\mathbf{b}}_\infty^{\mathrm{M}} = \mathbf{1}$.

### 3.4 Sample Efficiency through Generalized Method of Moments

With the low rank and penalization extensions in place, we extend the estimation procedure once more: we define the criterion function $M_N(\mathbf{R},\mathbf{S})$ of Equation 9 in order to obtain optimal sample efficiency.

Let $\mathbf{m}$ be a vector of length $n^3$, which flattens the moment conditions $m_x(\boldsymbol{\theta})$ over $x \in [n]$ and each matrix element $i, j$. More specifically, an index $(x, i, j) \in [n]^3$ into $\mathbf{m}$ is

$$\mathbf{m}_{xij} = [\widehat{\mathbf{P}}_{3,x,1}]_{ij} - [\mathbf{R}_x]_{i\cdot}\mathbf{S}_x^\top[\widehat{\mathbf{P}}_{2,1}]_{\cdot j}.$$

As before, there are $n^3$ moment conditions but now $2n^2k$ parameters due to the low rank structure—corresponding to each element in the $n \times n \times k$ tensors $\mathbf{R}, \mathbf{S}$. The GMM estimator is the minimizer of the criterion function

$$M_N(\mathbf{R},\mathbf{S}) = \sum_{i,j\in[[n]^3]^2} \mathbf{W}_{ij}\mathbf{m}_i\mathbf{m}_j, \quad (10)$$

where $\mathbf{W}$ is a weighting matrix that trades off between errors in the various GMM moment condtions. If $\mathbf{W}$ is the identity $\mathbf{I}$, then each term is $\mathbf{m}_i\mathbf{m}_j$ for all $i, j \in [n]^3$; this recovers the original spectral M-estimation criterion function considered in Equation 8.

To achieve maximum sample efficiency, GMM theory (Hansen, 1982) states that the optimal weighting $\mathbf{W}$ is proportional to the precision matrix,

$$\mathbf{W} \propto \mathbb{E}[m(\mathbf{X}_n, \{\mathbf{R}^*, \mathbf{S}^*\})m(\mathbf{X}_n, \{\mathbf{R}^*, \mathbf{S}^*\})^\top]^{-1}. \quad (11)$$

The optimal $\mathbf{W}$ minimizes the variance of the estimator by calibrating it to the inexactness of the estimated statistics, $\widehat{\mathbf{P}}_1, \widehat{\mathbf{P}}_{2,1}, \widehat{\mathbf{P}}_{3,x,1}$. If the moment conditions form the gradient of the log-likelihood function, $m(\mathbf{X}, \{\mathbf{R}, \mathbf{S}\}) = \nabla\ell(\{\mathbf{R}, \mathbf{S}\}; \mathbf{X})$, the optimal weighting matrix $\mathbf{W}$ becomes the inverse Fisher information evaluated at the true parameters. This recovers a maximum likelihood estimator with minimal asymptotic variance. Analogous to the MLE setting, the choice of the moment conditions $m$ and weighting matrix $\mathbf{W}$ may also be interpreted as minimizing a distance to the true data generating distribution, where the distance between probability distributions is defined by symmetrized KL divergence (Amari and Kawanabe, 1997b,a).

To gain intuition, note that a first-order diagonal approximation to Equation 11 is given by the inverse diagonal entries of the expected outer product. These entries $\mathbf{W}_{ii}$ weight according to the magnitude of error in the sample moments $\mathbf{m}_i$. Large magnitudes for $\mathbf{m}_i$ lead $1/\mathbf{m}_i^2$ to be small; this forces the M-estimator to place less weight on high error moments. With cross-correlation, $\mathbf{W}$ places more weight on other estimates paired with high error moments. For example, a small error moment $\mathbf{m}_j$ leads to a larger weight $1/(\mathbf{m}_i\mathbf{m}_j)^2$. These weights enable more intelligent parameter estimation.

## 4 Algorithm

The criterion function $\mathcal{L}$ of Equation 9 is a quadratic form plus a convex penalty. Moreover, it is strongly convex for $\mathbf{R}$ given $\mathbf{S}$ and $\mathbf{S}$ given $\mathbf{R}$. Hence we proceed with estimation by the procedure of alternating minimization, i.e., apply convex solvers which alternate between estimating each set of parameters.

More specifically, we apply an iterative procedure where we 1. alternate minimizing the loss over $\mathbf{R}$ and $\mathbf{S}$ conditioned on an estimate of $\mathbf{W}$; 2. set $\mathbf{W}$ conditioned on estimates of $\mathbf{R}, \mathbf{S}$; 3. repeat the procedure until convergence. An overview of the procedure is described in Algorithm 1, and we derive gradients in the following proposition.

**Proposition 3.** *The gradients are*

$$\nabla_{\mathbf{R}}\mathcal{L} = \mathcal{J}_{\mathbf{R}}^\top \mathbf{W} m(\mathbf{X}, \{\mathbf{R}, \mathbf{S}\}) + \nabla_{\mathbf{R}} P_\alpha(\mathbf{R}, \mathbf{S}) \quad (12)$$

$$\nabla_{\mathbf{S}}\mathcal{L} = \mathcal{J}_{\mathbf{S}}^\top \mathbf{W} m(\mathbf{X}, \{\mathbf{R}, \mathbf{S}\}) + \nabla_{\mathbf{R}} P_\alpha(\mathbf{R}, \mathbf{S}) \quad (13)$$

*where the matrices* $\mathcal{J}_{\mathbf{R}} \in \mathbb{R}^{n^3 \times n^2 k}$ *and* $\mathcal{J}_{\mathbf{S}} \in \mathbb{R}^{n^3 \times n^2 k}$ *are given by*

$$[\mathcal{J}_{\mathbf{R}}]_{xij,uvw} = \begin{cases} -[\mathbf{S}_x^\top]_{w\cdot}[\mathbf{P}_{2,1}]_{\cdot j}, & \textit{if } x = u,\ i = v \\ 0, & \textit{otherwise} \end{cases} \quad (14)$$

*and*

$$[\mathcal{J}_{\mathbf{S}}]_{xij,uvw} = \begin{cases} -[\mathbf{R}_x]_{iw}[\mathbf{P}_{2,1}]_{vj}, & \textit{if } x = u \\ 0, & \textit{otherwise} \end{cases} \quad (15)$$

Note that we initialize $\widehat{\mathbf{W}} = \mathbf{I}$, so that one loop of Algorithm 1 corresponds to the original spectral estimator of Equation 7. The global optima upon future iterations are refined based on the weighting matrix, and are in fact *guaranteed to perform at least as well* as the minimizer of the original optima. Note also that only one iteration of the loop is necessary for optimal sample efficiency asymptotically, as $\widehat{\mathbf{W}}$ converges in probability to $\mathbb{E}[m(\mathbf{X}_n, \{\mathbf{R}, \mathbf{S}\})m(\mathbf{X}_n, \{\mathbf{R}, \mathbf{S}\})^\top]^{-1}$. However, for finite data we see in experiments that better performance occurs when running the algorithm until convergence.

The matrix factorization view considered here, as well as the introduction of the weighting matrix $\mathbf{W}$, makes the

**Algorithm 1:** Spectral M-estimation for HMMS

```
Input: $N$ observation triplets $\mathbf{X} = \{\mathbf{X}_n : (x_1, x_2, x_3)\}$.
Construct empirical statistics $\widehat{\mathbf{P}}_1, \widehat{\mathbf{P}}_{2,1}, \widehat{\mathbf{P}}_{3,x,1}\ \forall x \in [n]$.
Initialize $\widehat{\mathbf{W}} = I$.
Set iteration counter $s = 1$.
while not converged do
    if $s \geq 2$ then
        $\widehat{\mathbf{W}} = \left(\sum_{n=1}^N m(\mathbf{X}_n, \{\widehat{\mathbf{R}}, \widehat{\mathbf{S}}\})m(\mathbf{X}_n, \{\widehat{\mathbf{R}}, \widehat{\mathbf{S}}\})^\top\right)^{-1}$
    end
    $\widehat{\mathbf{R}}, \widehat{\mathbf{S}} = \arg\min_{\mathbf{R},\mathbf{S}} \mathcal{L}(\mathbf{R}, \mathbf{S})$ (Algorithm 2).
    Increment $s$.
end
$\widehat{\mathbf{b}}_1^{\text{M}} = \widehat{\mathbf{P}}_1$.
$\widehat{\mathbf{B}}^{\text{M}} = \{\widehat{\mathbf{R}}_x \widehat{\mathbf{S}}_x^\top\}$.
$\widehat{\mathbf{b}}_\infty^{\text{M}} = \mathbf{1}_n$.
Return $\widehat{\boldsymbol{\theta}}^{\text{M}} = (\widehat{\mathbf{b}}_1^{\text{M}}, \widehat{\mathbf{B}}^{\text{M}}, \widehat{\mathbf{b}}_\infty^{\text{M}})$.
```

**Algorithm 2:** Alternating minimization, given weights $\mathbf{W}$

```
Input: initial values $\widehat{\mathbf{R}}, \widehat{\mathbf{S}}$.
while not converged do
    $\widehat{\mathbf{R}} = \arg\min_{\mathbf{R}} \mathcal{L}(\mathbf{R}, \mathbf{S})$ (Equation 18)
    $\widehat{\mathbf{S}} = \arg\min_{\mathbf{S}} \mathcal{L}(\mathbf{R}, \mathbf{S})$ (Equation 19)
end
Return $\widehat{\mathbf{R}}, \widehat{\mathbf{S}}$.
```

problem highly nonconvex. However, much recent theory has gone into explaining why simple optimization procedures following alternating minimization typically perform well in practice (Jain et al., 2013; Loh and Wainwright, 2014; Hardt, 2014; Chen and Wainwright, 2015; Bhojanapalli et al., 2015; Garber and Hazan, 2015; Loh, 2015). We also find that in practice the richer information gain from the generalized M-estimation procedure leads to improved estimates. It is an open problem to understand these improvements theoretically. Note that initialization using the original spectral estimator guarantees a global solution to the first iteration without penalization; we can apply it to initialize future iterations of the weighting as well as for nonconvex optimization with a penalty.

For computational efficiency, one can take immediate advantage of the block diagonal structure of the weighting matrix: this comes as a result of the independent sets of parameters in the loss function of Equation 9. That is, the parameter matrices $\mathbf{B}_{x'}$ only appear in the $m_{xij} \in [n]^3$ moments when $x = x'$. Thus it can be embarassingly parallelized into $n$ separate optimizations. We apply individual optimizations on $n$ procedures, each of which have $n^2$ moment conditions and recover a particular $\mathbf{B}_x$. The computational complexity of the algorithm is $\mathcal{O}(n^2)$ per iteration, with a storage complexity of $\mathcal{O}(n^4)$.

## 5 Experiments

We demonstrate the sample efficiency gained by the weighting scheme in the M-estimator and the advantage of sparse estimation due to $L_1$ penalization. We use toy configurations to highlight the M-estimator's robustness to model or rank mismatch, imbalanced observations, low sample size, and overfitting; finally we show results on real data.

For the M-estimator, we initialize using the original spectral estimate and also try several random initializations; we then take the estimates with minimal training loss. As the weighting matrix can become numerically singular, we add $10^{-8}$ to the diagonal. Comparisons are always done on test set evaluations. Note also that evaluations of the loss cannot be compared among algorithms, as the estimators minimize inherently different functions.

| Length | $\widehat{\mathbf{B}}_0^{\text{spec}}$ | $\widehat{\mathbf{B}}_0^{\text{M}}$ |
|---|---|---|
| 10 | **1** | **1** |
| 15 | 0.8889 | **0.8607** |
| 25 | 0.0198 | $\mathbf{3.1521 \cdot 10^{-6}}$ |
| 50 | 0.0008 | $\mathbf{9.9664 \cdot 10^{-9}}$ |

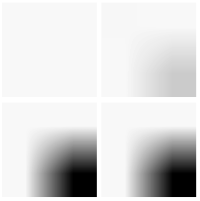
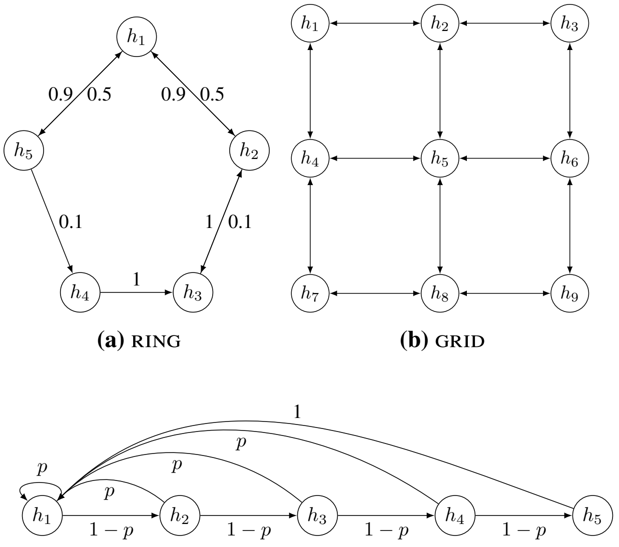


**Figure 1:** Left: Decay of the transition operator $\mathbf{B}_0$ as the length of sequence increases (lower is better); Right: Weighting matrix of $\mathbf{B}_0$ for each length is displayed from top left-right, bottom left-right.



**(a)** RING **(b)** GRID

**(c)** CHAIN

**Figure 2:** HMM configurations. **(a)** RING: The outer loop indicate clockwise transition probabilities, the inner loop indicate counter-clockwise. **(b)** GRID: Each state has equal probability of visiting any neighbor. **(c)** CHAIN: States transition with a probability $p$ of resetting to the first state.

### 5.1 Deterministic sequence

Consider a rank-11 system of two binary states: 0 and 1. The observation sequence deterministically follows the pattern "00000000001... 1", where 0 is always observed for the first 10 steps and 1 is observed for all remaining steps. Suppose that we aim to estimate this with a model of rank 1. Figure 1 displays the original estimator $\widehat{\boldsymbol{\theta}}^{\text{spec}}$ and the M-estimator $\widehat{\boldsymbol{\theta}}^{\text{M}}$. As the length of the sequence increases, we expect $\mathbf{B}_0$, the observable transition operator for the first state, to decay to 0. Our M-estimator achieves this at a much faster rate than $\widehat{\boldsymbol{\theta}}^{\text{spec}}$. It places more weight on the first state, and this weight increases with the length of the sequence.

### 5.2 Ring configuration

| Model rank | $\widehat{\boldsymbol{\theta}}^{\text{spec}}$ | $\widehat{\boldsymbol{\theta}}^{\text{M}}$ | $(\widehat{\boldsymbol{\theta}}^{\text{M}}, \lambda = 0.01)$ |
|---|---|---|---|
| 4 | 1.50 | 1.25 | **1.03** |
| 3 | 1.15 | 1.03 | **0.81** |
| 2 | 0.68 | 0.65 | **0.60** |

**Table 1:** Relative norm difference between estimated and true joint probability, averaged over 100 test examples.

In Section 2a, the hidden states form a ring: $h_1$ has uniform probability of proceeding clockwise to $h_2$ or counter-clockwise to $h_5$; $h_2$ and $h_5$ return back to $h_1$ with probability 0.9 and visit $h_3$ or $h_4$ (respectively) with probability 0.1.

This leads to imbalanced samples where $h_1, h_2, h_5$ are visited most, and one rarely sees $h_3$ and $h_4$. States are correctly observed with 0.6 probability, otherwise we observe any other state uniformly. We train on 100 examples.

Table 1 shows that under difficult settings—with imbalanced states, not enough training examples, and ill-posed rank problems—spectral estimators fit poorly due to the information loss from higher order moments. However, the weighting scheme of $\widehat{\boldsymbol{\theta}}^{\text{M}}$ allows the estimator to compensate for some of these problems, and thus it performs better than $\widehat{\boldsymbol{\theta}}^{\text{spec}}$. Moreover, when used with a $L_1$ penalty of $\lambda = 0.01$, the estimator dominates other algorithms; the value of $\lambda$ was also chosen generally and not optimized over.

### 5.3 Grid configuration

| Grid size | $\widehat{\boldsymbol{\theta}}^{\text{spec}}$ | $\widehat{\boldsymbol{\theta}}^{\text{M}}$ | $(\widehat{\boldsymbol{\theta}}^{\text{M}}, \lambda = 1\text{E-}3)$ |
|---|---|---|---|
| $2 \times 2$ | 0.014 | 0.014 | **0.014** |
| $3 \times 3$ | 0.225 | 0.225 | **0.212** |
| $5 \times 5$ | 0.475 | 0.475 | **0.458** |

**Table 2:** Relative norm difference between estimated and true joint probability, averaged over 100 test examples.

In the grid configuration (Section 2b), each hidden state has an equal probability of transitioning to any one of its neighbors; the observation matrix $\mathbf{O}$ indicates the correct state with 0.9 probability, and any other state otherwise. We use 100,000 training examples and vary the grid size.

Table 2 demonstrates good performance for small grids where the training data is large enough to accurately cover the state space. Note also that the unregularized M-estimator performs the same as the original estimator over all grid sizes. This is because the weighting matrix has no effect due to the the equally likely transitions, which are already well-balanced. However, the role of regularization becomes more important as the grid grows larger; this is because the fixed sample size leads observed states to be more spread out and revisited less often.

### 5.4 Chain configuration

| Reset probability | $\widehat{\boldsymbol{\theta}}^{\text{spec}}$ | $\widehat{\boldsymbol{\theta}}^{\text{M}}$ | $(\widehat{\boldsymbol{\theta}}^{\text{M}}, \lambda = 1\text{E-}3)$ |
|---|---|---|---|
| 0.1 | 0.80 | 0.73 | **0.72** |
| 0.3 | 0.82 | 0.80 | **0.77** |
| 0.5 | 1.24 | 0.96 | **0.69** |

**Table 3:** Relative norm difference between estimated and true joint probability, averaged over 100 test examples.

The chain configuration (Section 2c) mimicks the chain problem in reinforcement learning (Strens, 2000; Poupart et al., 2006). Each hidden state transitions to the next with probability $1-p$ and resets to the first state with probability $p$. We use 50 training examples for each $p$.

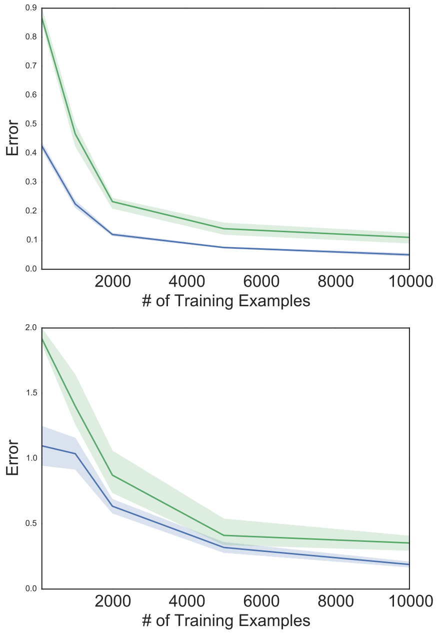


**Figure 3:** Predictive accuracy of original estimator (green) and M-estimator (blue) over # of training examples, with standard error bars taken over 100 simulations. Top: $m = 5$ hidden states with $n = 10$ hidden states. Bottom: $m = 10$ hidden states with $n = 20$ observed states.

As the reset probability increases, the data distribution becomes more heavy tailed. This is reflected in Table 3, as the weighting makes a larger impact over highly skewed distributions. As very few examples are seen with the last few states, the $L_1$ penalty has a growing impact as well.

### 5.5 Synthetic hidden Markov models

We generate two large synthetic data sets following well-behaving HMMS: one system uses $m = 5$ hidden states and $n = 10$ observed states, and the other uses $m = 10$ hidden states and $n = 20$ observed states. We perform both full rank and low rank estimation over $10,000$ training examples and analyze held-out prediction error.

In Figure 3, we see that with few training examples, the M-estimator's optimal weighting scheme is crucial for reasonable performance. Moreover, as explained in theory, the variance of the M-estimator is much lower than the original spectral estimator. The original estimator and the M-estimator converge at the same rate and eventually reach competitive errors. However, the M-estimator achieves this much faster in practice even in these well-behaving dynamical systems.

| Data set | Type | Training set | # Obs. states | $\widehat{\boldsymbol{\theta}}^{\text{spec}}$ | $\widehat{\boldsymbol{\theta}}^{\text{M}}$ | $(\widehat{\boldsymbol{\theta}}^{\text{M}}, \lambda = 1\text{E-}5)$ | $\widehat{\boldsymbol{\theta}}^{\text{EM}}$ |
|---|---|---|---|---|---|---|---|
| *Alice* | Text | 50,000 | 26 | 0.22 | **0.20** | **0.20** | 0.14 |
| *Splice* | DNA | 100,000 | 4 | 0.41 | 0.40 | **0.35** | 0.19 |
| *Bach Chorales* | Music | 4,693 | 20 | 0.31 | 0.28 | **0.25** | 0.24 |
| *Ecoli* | Protein | 1,407 | 20 | 0.14 | **0.13** | 0.15 | 0.12 |
| *Dodgers* | Traffic | 30,000 | 10 | 0.42 | **0.38** | 0.39 | 0.33 |

**Table 4:** Predictive test error for three spectral estimators—Hsu et al. (2012), M-estimator, and regularized M-estimator—and EM. In many cases the M-estimators approach the performance of EM.

### 5.6 Real data sets

We now examine the performance of the estimators for 5 separate data sets: in the *Alice* novel available in Project Gutenberg, the task is to predict characters after having trained over the first 50,000 of them; the *Splice* data set consists of 3,190 examples of DNA sequences which have length 60 and the task is to predict the remaining A,C,T, or G fields; the *Bach Chorales* consists of discrete event sequences in which the task is to predict the correct pitch of melody lines; *Ecoli* describes sequencing information of protein localization sites; *Dodgers* examines link counts over a freeway in Los Angeles. These last four data sets are available from Lichman (2013).

Table 4 indicates the average prediction error on held out data. The results are consistent with that of the toy configurations and synthetic benchmarks. In all data sets, the M-estimator surpasses the original estimator. The benefit of sparse regularization tended to vary, as we did not choose to tune this hyperparameter per data set. We also compared to EM with random initializations as a benchmark to likelihood-based methods. Many local optima performed poorly; the best solutions found after enough random initializations uniformly performed better than the spectral estimators over all data sets.

## 6 Discussion and Related Work

In this work, we focused on the application of M-estimation to estimating parameters of HMMs. Our analysis and algorithms carry over almost identically for predictive state representations (e.g. in Siddiqi et al. (2010); Song et al. (2010); Boots et al. (2010)). Estimating parameters for other latent variable models can also be easily formulated as generalized method of moments problems. For example, following Anandkumar et al. (2012), a mixture model specified by $\Pr(h = j) = \omega_j$ and $\Pr(x = i \mid h = j) = \mathbf{M}_{ij}$ for $i \in [n], j \in [k]$, has moment conditions

$$\begin{aligned} m_1(\mathbf{M}, \omega) &= \mathbf{P}_{2,1} - \mathbf{M}\,\text{diag}(\omega)\mathbf{M}^\top, \\ m_x(\mathbf{M}, \omega) &= \mathbf{P}_{3,x,1} - \mathbf{M}\,\text{diag}(\mathbf{M}^\top e_x)\,\text{diag}(\omega)\mathbf{M}^\top, \end{aligned}$$

for all $x \in [n]$, where $e_x$ is the unit vector equal to one at index $x$. Closest to our approach is that of Kulesza et al. (2015), who propose a weighting scheme to address fundamental issues with low rank spectral learning. Their weighting scheme can be seen as redefining the moment conditions

$$m_x(\boldsymbol{\theta}) = \mathbf{P}_{3,x,1} - \mathbf{B}_x\mathbf{W}\mathbf{P}_{2,1} \quad \forall x \in [n].$$

With this moment condition, solvers using singular value decomposition avoid instabilities as noted in Kulesza et al. (2014). In contrast, our GMM approach takes the direct path of weighting the moment conditions, i.e., the error in the statistics for estimating the moments. Kulesza et al. (2015) also require that a domain expert specify the weighting matrix $\mathbf{W}$; our $\widehat{\boldsymbol{\theta}}^{\text{M}}$ is automatically given by our optimal choice of weighting matrix. That said, in situations where domain experts can connect a choice of $\mathbf{W}$ to a specific task, one can forgo sample efficiency and specify the weighting matrix of the GMM manually.

Also related to our work are methods that use spectral methods to initialize techniques for maximum likelihood estimation (Zhang et al., 2014; Balle et al., 2014). Shaban et al. (2015) follow this approach and propose a two-stage procedure, which corresponds to typical spectral estimation in the first stage and optimization upon the second to ensure feasible solutions (which our method does not). While we also have an iterative procedure that begins with a spectral initialization, each of our steps is still within the spectral framework. Our approach of weighting the moments and considering suitable penalization is orthogonal to the use of the spectral estimates for initializing other estimation techniques. It remains open to explore the benefits of these approaches when merged in practice.

To our knowledge, our work is the first to achieve optimal sample efficiency rates for spectral estimation, and we provide a principled approach to incorporating regularization into the process. However, we now have a highly nonconvex optimization problem, and we also rely on row-level elements of the data. Addressing these concerns, while maintaining sample-efficiency and accuracy bounds, remains an important direction for future work.

### Acknowledgements

We thank Maja Rudolph and Dawen Liang for their helpful comments and support from NSF ACI-1544628.

## References

Amari, S.-I. and Kawanabe, M. (1997a). Estimating functions in semiparametric statistical models. *Lecture Notes-Monograph Series*, pages 65–81.

Amari, S.-I. and Kawanabe, M. (1997b). Information geometry of estimating functions in semi-parametric statistical models. *Bernoulli*, 3(1):29–54.

Anandkumar, A., Ge, R., Hsu, D., and Kakade, S. M. (2014a). A tensor approach to learning mixed membership community models. *Journal of Machine Learning Research*, 15:2239–2312.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014b). Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832.

Anandkumar, A., Hsu, D., and Kakade, S. M. (2012). A method of moments for mixture models and hidden Markov models. *arXiv preprint arXiv:1203.0683*.

Arora, S., Ge, R., and Moitra, A. (2012). Learning topic models–going beyond SVD. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 1–10. IEEE.

Balle, B., Hamilton, W., and Pineau, J. (2014). Methods of moments for learning stochastic languages: Unified presentation and empirical comparison. In *International Conference on Machine Learning*.

Balle, B. and Mohri, M. (2012). Spectral learning of general weighted automata via constrained matrix completion. In *Neural Information Processing Systems 25*.

Bhojanapalli, S., Kyrillidis, A., and Sanghavi, S. (2015). Dropping convexity for faster semi-definite optimization. *arXiv preprint arXiv:1509.03917*.

Boots, B. and Gordon, G. (2011). An online spectral learning algorithm for partially observable nonlinear dynamical systems. In *Proceedings of the 25th National Conference on Artificial Intelligence (AAAI)*.

Boots, B., Siddiqi, S., and Gordon, G. J. (2010). Closing the learning-planning loop with predictive state representations. In *Proceedings of Robotics: Science and Systems*.

Burguete, J. F., Ronald Gallant, A., and Souza, G. (1982). On unification of the asymptotic theory of nonlinear econometric models. *Econometric Reviews*, 1(2):151–190.

Chen, Y. and Wainwright, M. J. (2015). Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*.

Cohen, S. B., Stratos, K., Collins, M., Foster, D. P., and Ungar, L. (2012). Spectral learning of latent-variable PCFGs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 223–231. Association for Computational Linguistics.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Garber, D. and Hazan, E. (2015). Fast and simple PCA via convex optimization. *arXiv preprint arXiv:1509.05647*.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, pages 1208–1211.

Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, 63(2):277–284.

Godambe, V. P. (1991). *Estimating functions*. Clarendon Press Oxford.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica: Journal of the Econometric Society*, pages 681–700.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–54.

Hardt, M. (2014). Understanding alternating minimization for matrix completion. In *FOCS 2014*. IEEE.

Hsu, D., Kakade, S. M., and Zhang, T. (2012). A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480.

Huang, F., N, N. U., Hakeem, M. U., Verma, P., and Anandkumar, A. (2013). Fast detection of overlapping communities via online tensor methods on gpus. *arXiv preprint arXiv:1309.0787*.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, pages 799–821.

Huber, P. J. and Ronchetti, E. (2009). *Robust Statistics*. Wiley Series in Probability and Statistics.

Jain, P., Netrapalli, P., and Sanghavi, S. (2013). Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of Computing*, pages 665–674.

Kulesza, A., Jiang, N., and Singh, S. (2015). Low-rank spectral learning with weighted loss functions. In *Artificial Intelligence and Statistics*.

Kulesza, A., Nadakuditi, R. R., and Singh, S. (2014). Low-rank spectral learning. In *Proceedings of the 17th Conference on Artificial Intelligence and Statistics*.

Lambert-Lacroix, S., Zwald, L., et al. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics*, 5:1015–1053.

Li, G., Peng, H., and Zhu, L. (2011). Nonconcave penalized m-estimation with a diverging number of parameters. *Statistica Sinica*, 21(1):391.

Lichman, M. (2013). UCI machine learning repository.

Loh, P.-L. (2015). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *arXiv preprint arXiv:1501.00312*.

Loh, P.-L. and Wainwright, M. J. (2014). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*.

Mahoney, M. W. and Orecchia, L. (2011). Implementing regularization implicitly via approximate eigenvector computation. In *International Conference on Machine Learning*.

Mossel, E. and Roch, S. (2005). Learning nonsingular phylogenies and hidden Markov models. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 366–375. ACM.

Nati, N. S. and Jaakkola, T. (2003). Weighted low-rank approximations. In *International Conference on Machine Learning*.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097.

Owen, A. B. (2007). A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72.

Parikh, A. P., Song, L., and Xing, E. P. (2011). A spectral algorithm for latent tree graphical models. In *Proceedings of the 28th International Conference on Machine Learning*.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, pages 71–110.

Poupart, P., Vlassis, N., Hoey, J., and Regan, K. (2006). An analytic solution to discrete Bayesian reinforcement learning. In *International Conference on Machine Learning*, pages 697–704.

Shaban, A., Farajtabar, M., Xie, B., Song, L., and Boots, B. (2015). Learning latent variable models by improving spectral solutions with exterior point methods. In *Uncertainty in Artificial Intelligence*.

Siddiqi, S., Boots, B., and Gordon, G. J. (2010). Reduced-rank hidden Markov models. In *Artificial Intelligence and Statistics*.

Song, L., Boots, B., Siddiqi, S. M., Gordon, G. J., and Smola, A. (2010). Hilbert space embeddings of hidden Markov models. In *International Conference on Machine Learning*.

Strens, M. (2000). A Bayesian framework for reinforcement learning. In *International Conference on Machine Learning*, pages 943–950.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge University Press.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. (2014). Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Neural Information Processing Systems*, pages 1260–1268.

## A Proof of Proposition 1

**Proposition 1.** *Let $\widehat{\boldsymbol{\theta}}^{spec}$ denote the estimator using empirical statistics in Equation 4. Let $\widehat{\boldsymbol{\theta}}^{\text{M}}$ denote the M-estimator given by*

$$\begin{aligned}\widehat{\mathbf{b}}_1^M &= \widehat{\mathbf{P}}_1,\\ \widehat{\mathbf{b}}_\infty^M &= \mathbf{1}_n,\\ \widehat{\mathbf{B}}^M &= \operatorname*{arg\,min}_{\mathbf{B}\in\mathbb{R}^{n\times n\times n}} M_N(\mathbf{B}).\end{aligned}$$

*Then $\widehat{\boldsymbol{\theta}}^{\text{M}}$ is in the same equivalence class as $\widehat{\boldsymbol{\theta}}^{spec}$, so they provide the same probability estimates.*

*Proof.* Let $x \in [n]$, and consider a solution to the moment conditions for parameter $\mathbf{B}_x \in \mathbb{R}^{n\times n}$ given by

$$\min_{\mathbf{B}_x} \|\mathbf{P}_{3,x,1} - \mathbf{B}_x\mathbf{P}_{2,1}\|_F^2 \qquad (16)$$

Equation 16 can be solved using any convex program, or, by the Eckart-Young theorem (Eckart and Young, 1936), through singular value decomposition. Thus we recover the original spectral estimator: Equation 16 is equivalent to a singular value decomposition as standard methods in spectral learning do (Hsu et al., 2012; Boots et al., 2010; Boots and Gordon, 2011; Huang et al., 2013). Note further that while this problem is nonconvex, all local optima are also global (Nati and Jaakkola, 2003). Hence the estimates we obtain using optimization routines are consistent.

Hsu et al. (2012) derive Equation 16 from a different standpoint and consider the special case of full rank $k = m$. They proceed to relax the rank constraint by observing that the parameters are learned up to a similarity transform: given the triplet $(\mathbf{b}_1, \{\mathbf{B}_x\}, \mathbf{b}_\infty)$ and an invertible matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$, the transformed triplet $(\mathbf{b}_1' = \mathbf{S}\mathbf{b}_1, \{\mathbf{B}_x' = \mathbf{S}\mathbf{B}_x\mathbf{S}^{-1}\}, \mathbf{b}_\infty' = \mathbf{S}^{-T}\mathbf{b}_\infty)$ provide the same joint probabilities as written in Equation (5).

Instead of choosing an invertible similarity transform, one can find $\mathbf{U} \in \mathbb{R}^{n\times k}$ such that $\mathbf{U}^\top\mathbf{P}_{2,1}$ (equivalently, $\mathbf{U}^\top\mathbf{O}$) is invertible, as any inversions regarding $U$ are only involved through the product $\mathbf{U}^\top\mathbf{P}_{2,1}$. A natural choice is to let $\mathbf{U}$ be the matrix of $k$ left-singular vectors of $\mathbf{P}_{2,1}$ (Hsu et al., 2012, Lemma 2). Then an equivalent optimization procedure to Equation 16 is simply

$$\min_{\mathbf{B}_x'} \|\mathbf{P}_{3,x,1} - \mathbf{B}_x'\mathbf{P}_{2,1}\|_F^2 \qquad (17)$$

where $\mathbf{B}_x' \equiv \mathbf{U}^\top\mathbf{B}_x(\mathbf{U}^T)^\dagger = (\mathbf{U}^\top\mathbf{O})\mathbf{A}_x(\mathbf{U}^\top\mathbf{O})^{-1} \in \mathbb{R}^{k\times k}$. The advantage is that $\mathbf{B}_x'$ is automatically constrained to be of rank $k$ through the similarity transform on $\mathbf{A}_x$ given by $\mathbf{U}^\top\mathbf{O}$. This can be solved trivially with $\mathbf{B}_x' = \mathbf{P}_{3,x,1}\mathbf{P}_{2,1}^\dagger$, and in terms of the original parameter $\mathbf{B}_x = (\mathbf{U}^\top\mathbf{P}_{3,x,1})(\mathbf{U}^\top\mathbf{P}_{2,1})^{-1}$ (Hsu et al., 2012, Proof of Lemma 3). □

## B Proof of Proposition 3

**Proposition 3.** *The gradients are*

$$\nabla_{\mathbf{R}}\mathcal{L} = \mathcal{J}_{\mathbf{R}}^\top\mathbf{W}m(\mathbf{X}, \{\mathbf{R}, \mathbf{S}\}) + \nabla_{\mathbf{R}}P_\alpha(\mathbf{R}, \mathbf{S}) \qquad (18)$$

$$\nabla_{\mathbf{S}}\mathcal{L} = \mathcal{J}_{\mathbf{S}}^\top\mathbf{W}m(\mathbf{X}, \{\mathbf{R}, \mathbf{S}\}) + \nabla_{\mathbf{R}}P_\alpha(\mathbf{R}, \mathbf{S}) \qquad (19)$$

*where the matrices $\mathcal{J}_{\mathbf{R}} \in \mathbb{R}^{n^3\times n^2k}$ and $\mathcal{J}_{\mathbf{S}} \in \mathbb{R}^{n^3\times n^2k}$ are given by*

$$[\mathcal{J}_{\mathbf{R}}]_{xij,uvw} = \begin{cases} -[\mathbf{S}_x^\top]_{w\cdot}[\mathbf{P}_{2,1}]_{\cdot j}, & \textit{if } x = u,\ i = v \\ 0, & \textit{otherwise}\end{cases} \qquad (20)$$

*and*

$$[\mathcal{J}_{\mathbf{S}}]_{xij,uvw} = \begin{cases} -[\mathbf{R}_x]_{iw}[\mathbf{P}_{2,1}]_{vj}, & \textit{if } x = u \\ 0, & \textit{otherwise}\end{cases} \qquad (21)$$

*Proof.* For a general quadratic matrix function $f(\boldsymbol{\theta}) = y(\boldsymbol{\theta})^\top\mathbf{W}y(\boldsymbol{\theta})$ with given matrix $\mathbf{W}$, its gradient is

$$\nabla f(\boldsymbol{\theta}) = [\nabla y(\boldsymbol{\theta})]^\top(\mathbf{W} + \mathbf{W}^\top)y(\boldsymbol{\theta})$$

Hence for our situation where $\mathbf{W}$ is symmetric, it is

$$\begin{aligned}\nabla_R\mathcal{L} = 2&\left[\nabla_R\left[[\widehat{P}_{3,x,1}]_{ij} - [R_x]_{i\cdot}S_x^\top[P_{2,1}]_{\cdot j}\right]_{xij\in[n^3]}\right]^\top \\ &\mathbf{W}\left[\widehat{m}_{xij}(\boldsymbol{\theta})\right]_{xij\in[n^3]} \\ &= 2\mathcal{J}_R^\top\mathbf{W}\left[\widehat{m}_{xij}(\boldsymbol{\theta})\right]_{xij\in[n^3]}\end{aligned}$$

The Jacobian $\mathcal{J}_R$ is a $n^3 \times n^2k$ matrix, with elements $(xij, uvw) \in [n^3] \times [n^2k]$. The $(xij, uvw)^{th}$ entry is the partial derivative of the $xij^{th}$ moment $\widehat{m}_{xij}$ on $[R_u]_{vw}$:

$$\begin{aligned}[\mathcal{J}_R]_{xij,uvw} &= \frac{\partial}{\partial[R_u]_{vw}}\left[-\sum_{r=1}^k [R_x]_{ir}[S_x^\top]_{r\cdot}[P_{2,1}]_{\cdot j}\right] \\ &= \begin{cases} -[S_x^\top]_{w\cdot}[P_{2,1}]_{\cdot j} & \text{if } x = u,\ i = v \\ 0 & \text{otherwise}\end{cases}\end{aligned}$$

Similarly, there is a Jacobian $\mathcal{J}_S$ when taking the gradient with respect to $S$, and by the same logic the Jacobian with respect to $S$ is

$$\begin{aligned}[\mathcal{J}_S]_{xij,uvw} &= \frac{\partial}{\partial[S_u]_{vw}}\left[-\sum_{s=1}^n\sum_{r=1}^k [R_x]_{ir}[S_x]_{sr}[P_{2,1}]_{sj}\right] \\ &= \begin{cases} -[R_x]_{iw}[P_{2,1}]_{vj} & \text{if } x = u \\ 0 & \text{otherwise}\end{cases}\end{aligned}$$

□