Learning Latent Variable Models by Improving Spectral Solutions with Exterior Point Methods

Amirreza Shaban **Mehrdad Farajtabar** **Bo Xie** **Le Song** **Byron Boots**
Georgia Institute of Technology
{amirreza, mehrdad, bo.xie}@gatech.edu {lsong, bboots}@cc.gatech.edu

Abstract

Despite the widespread use of probabilistic latent variable models, identifying the parameters of basic latent variable models continues to be an extremely challenging problem. Traditional maximum likelihood-based learning algorithms find valid parameters, but suffer from high computational cost, slow convergence, and local optima. In contrast, recently developed spectral algorithms are computationally efficient and provide strong statistical guarantees, but are not guaranteed to find valid parameters. In this work, we first use a spectral method of moments algorithm to find a solution that is *close* to the optimal solution but not necessarily in the valid set of model parameters. We then incrementally refine the solution via an exterior point method until a local optima that is arbitrarily near the valid set of parameters is found. Our experiments show that our approach is more accurate than previous work, especially when training data is limited.

1 INTRODUCTION & RELATED WORK

Probabilistic latent variable models are a fundamental tool in statistics and machine learning that have successfully been deployed in a wide range of applied domains including robotics, bioinformatics, speech recognition, document analysis, social network modeling, and economics. Despite their widespread use, identifying the parameters of basic latent variable models like multi-view models and hidden Markov models (HMMs) continues to be an extremely challenging problem. Researchers often resort to local search heuristics such as expectation maximization (EM) [9] that attempt to find parameters that maximize the likelihood of the observed data. Unfortunately, EM has a number of well-documented drawbacks, including high computational cost, slow convergence, and local optima.

In the past 5 years, several techniques based on *method of moments* [15] have been proposed as an alternative to maximum likelihood for learning latent variable models [11, 17, 19, 13, 14, 12, 1, 3, 2, 6, 7, 18]. These algorithms first estimate low-order moments of observations, such as means and covariances, and then apply a sequence of linear algebra to recover the model parameters. Moment estimation is linear in the number of training data samples, and parameter estimation, which relies on techniques like the singular value decomposition (SVD), is typically fast and numerically robust. Spectral algorithms were recently extended to the more difficult problem of estimating the parameters of latent variable models including the stochastic transition and observation matrices of HMMs [3]. The resulting parameters can be used to *initialize* EM in a two-stage learning algorithm [21, 5], resulting in the best of both worlds: parameters found by method of moments provide a good initialization for a maximum likelihood approach.

Unfortunately, spectral method of moments algorithms and two-stage learning algorithms have worked less well in practice. With finite samples, method of moments estimators are *not* guaranteed to find a valid set of parameters. Although error in the estimated parameters are bounded, the parameters themselves may lie outside the class of valid models. To fix these problems, the method of

moments solution is typically projected onto the ℓ_1-ball [10]. In this work, however, we directly initialize our optimization algorithm with the method of moments estimations. We propose a two-stage algorithm for learning the parameters of latent variable models. In the first stage, the parameters are estimated via a spectral method of moments algorithm (similar to [3]). In the second stage, unlike previous work that projects method of moments onto the feasible space of model parameters and then uses the projected parameters to initialize EM [21, 5], we use exterior point methods for non-convex optimization directly initialized with the result of method of moments *without modification*. The exterior point method iteratively refines the solution until a local optima that is arbitrarily close to the valid set of model parameters is found. An important advantage of exterior point methods is that they are likely to achieve a better local minimum than interior point methods when the feasible set is narrow by allowing solutions to exist outside the feasible set during the intermediate steps of the optimization [20].

2 PARAMETER ESTIMATION VIA METHOD OF MOMENTS

Method of moment parameter estimation serves an initial solution to our optimization algorithm. Here we only highlight some important notations of multi-view model. For more details on multi-view model and HMM parameter estimation using method of moment method, the reader may consult [3]. In a multi-view model, observation variables $o_1, o_2, \ldots, o_l$ are conditionally independent given a latent variable h. Assume each observation variable can take one of n_o different values. The observation vector $\boldsymbol{x}_t \in \mathbb{R}^{n_o}$ is defined as follows:

$$\boldsymbol{x}_t = \boldsymbol{e}_j \text{ iff } o_t = j \text{ for } 0 < j \leq n_o, \tag{1}$$

where $\boldsymbol{e}_j$ is the j^{th} canonical basis. In this paper, we consider the case where $l = 3$, however, the techniques can be easily extended to cases where $l > 3$. Let $h \in \{1, \ldots, n_s\}$ be a discrete random variable and $\Pr\{h = j\} = (\boldsymbol{w})_j$, where $\boldsymbol{w} \in \Delta^{n_s - 1}$, then the conditional expectation of observation vector $\boldsymbol{x}_t$ for $t \in \{1, 2, 3\}$ is:

$$\mathbb{E}[\boldsymbol{x}_t \mid h = i] = \boldsymbol{u}_i^t, \tag{2}$$

where $\boldsymbol{u}_i^t \in \Delta^{n_o - 1}$. We define the observation matrix as $\boldsymbol{U}^t = [\boldsymbol{u}_1^t, \ldots, \boldsymbol{u}_{n_s}^t]$ for $t \in \{1, 2, 3\}$, and the diagonal $n_s \times n_s \times n_s$ tensor $\mathcal{H}$, where $\operatorname{diag}(\mathcal{H}) = \boldsymbol{w}$ for ease of notation. The following proposition relates $\boldsymbol{U}^t$s and $\boldsymbol{w}$ to the moments of $\boldsymbol{x}_t$s.

Proposition 1. **[2]** *Assume that columns of $\boldsymbol{U}^t$ are linearly independent for each $t = \{1, 2, 3\}$. Define $\mathcal{M} = \mathbb{E}(\boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3)$, then ($\times_n$ is n-mode product)*

$$\mathcal{M} = \mathcal{H} \times_1 \boldsymbol{U}^1 \times_2 \boldsymbol{U}^2 \times_3 \boldsymbol{U}^3 \tag{3}$$

3 EXTERIOR POINT METHODS

Let $\boldsymbol{v} \in \mathbb{R}^{n_s(3n_o+1)}$ be a vector comprised of the parameters of the multi-view model $\{\boldsymbol{U}^1, \boldsymbol{U}^2, \boldsymbol{U}^3, \operatorname{diag}(\mathcal{H})\}$, and $\mathcal{R}(\boldsymbol{v}) = (\hat{\mathcal{M}} - \mathcal{H} \times_1 \boldsymbol{U}^1 \times_2 \boldsymbol{U}^2 \times_3 \boldsymbol{U}^3)$ be the residual estimation tensor. For ease of notation, we also define function $s(\cdot) : \mathbb{R}^{n_s(3n_o+1)} \to \mathbb{R}^{3n_s+1}$ that computes column sum of $\boldsymbol{U}^1$, $\boldsymbol{U}^2$, $\boldsymbol{U}^3$, and $\operatorname{diag}(\mathcal{H})$. Since in practice method of moments works with estimated empirical moment $\hat{\mathcal{M}}$ rather than the population moment $\mathcal{M}$, the estimated parameters using method of moments do not necessarily minimize the estimation error $||\mathcal{R}(\boldsymbol{v})||_F$ and also may violate the constraints for the model parameters. With these limitations in mind, we rewrite the factorization in Equation (3) in the form of an optimization problem:

$$\begin{aligned} &\text{minimize } \frac{1}{2}||\mathcal{R}(\boldsymbol{v})||_F^2. \\ &\text{s.t. } \boldsymbol{v} \in \mathbb{R}_+^{n_s(3n_o+1)}, \mathrm{s}(\boldsymbol{v}) = \mathbf{1} \end{aligned} \tag{4}$$

Defining optimization problem in this form has the advantage over maximum likelihood optimization schemes that the value of this objective function $||\mathcal{R}(\boldsymbol{v})||_F$ is also defined *outside* of the feasible set. Instead of solving above constrained optimization problem, we solve the following unconstrained optimization problem:

$$\text{minimize } \frac{1}{2}||\mathcal{R}(\boldsymbol{v})||_F^2 + \frac{\lambda_1}{2}||\mathrm{s}(\boldsymbol{v}) - \mathbf{1}||_p^2 + \lambda_2 |\boldsymbol{v}|_-, \tag{5}$$

Algorithm 1 *The exterior point algorithm*

Input: Estimated third order moment $\hat{\mathcal{M}}$, initial point (obtained from method of moments) $\boldsymbol{v}^{(0)}$ is comprised of $\{\hat{\boldsymbol{U}}^1, \hat{\boldsymbol{U}}^2, \hat{\boldsymbol{U}}^3, \text{diag}(\hat{\mathcal{H}})\}$, parameters λ_1, and λ_2, sequence $\{\beta_k\}$, and constant $c > 0$
Output: Convergence point $\boldsymbol{v}^*$

$k \leftarrow 0$
while not converged **do**
  $k \leftarrow k + 1$
  if $|\boldsymbol{v}^{(k-1)}|_- > 0$ **then**
    $\alpha_k \leftarrow max\{c, \beta_k\}$
  else
    $\alpha_k \leftarrow \beta_k$
  end if
  $\tilde{\boldsymbol{v}}^{(k)} \leftarrow \boldsymbol{v}^{(k-1)} - \alpha_k \nabla g(\boldsymbol{v}^{(k-1)})$
  $\boldsymbol{v}^{(k)} \leftarrow \text{prox}(\tilde{\boldsymbol{v}}^{(k)})$ (see Equation (7))
end while

where $|\boldsymbol{v}|_-$ is the absolute sum of all negative elements in the vector $\boldsymbol{v}$, i.e., $|\boldsymbol{v}|_- = \sum_i |(\boldsymbol{v})_i|_-$. We set $p = 2$ in our method. For $p = 1$, there exists a λ_1 and a λ_2 such that the solution to this unconstrained optimization is *also* the solution to the objective in Equation (4). Our approach, however, is different in the sense that for $p = 2$ the solution of our optimization algorithm is not guaranteed to satisfy the constraints in Equation (4), however, we show in Theorem 3 that the solution will be arbitrarily close to the simplex. In return for this relaxation, the above optimization problem can be easily solved by a standard *forward-backward splitting* algorithm [8]. In this method, the function is split into a smooth part:

$$g(\boldsymbol{v}) = \frac{1}{2}||\mathcal{R}(\boldsymbol{v})||_F^2 + \frac{\lambda_1}{2}||\text{s}(\boldsymbol{v}) - \mathbf{1}||_2^2 \tag{6}$$

and a non-smooth part $\lambda_2|\boldsymbol{v}|_-$. We then minimize the objective function by alternating between a gradient step on the smooth part $\nabla g(\boldsymbol{v})$ (forward) and proximal step of the non-smooth part (backward). The overall algorithm is shown in Algorithm 1 where $\text{prox}(\tilde{\boldsymbol{v}}^{(k)})$ function is defined as

$$\boldsymbol{v}^{(k)} = \underset{\boldsymbol{y}}{\text{argmin}}(\alpha_k \lambda_2 |\boldsymbol{y}|_- + \frac{1}{2}||\tilde{\boldsymbol{v}}^{(k)} - \boldsymbol{y}||_F^2). \tag{7}$$

We only need to find the gradient of function $g(\boldsymbol{v})$ for the given model parameter $\boldsymbol{v}$. Under certain assumptions, by using the forward-backward splitting steps in Algorithm 1, we can show that there are lower bounds for λ_1 and λ_2 in which the iterative algorithm converges to a local optimum arbitrarily close to the simplex.

Corollary 2. *For every $\boldsymbol{v}$ which is comprised of the multi-view parameters and is in the compact set $\mathscr{F} = \{\boldsymbol{v} \mid \forall i, 0 < i \leq n_s, 1 < t \leq 3 : ||(\boldsymbol{U}^t)_i||_2 \leq L, ||diag(\mathcal{H})||_2 \leq L\}$, every element of the gradient of the function $r(\boldsymbol{v})$ is bounded as:*

$$|\frac{\partial r(\boldsymbol{v})}{\partial(\boldsymbol{v})_i}| \leq L^3 ||\mathcal{R}(\boldsymbol{v})||_F \tag{8}$$

Theorem 3. *Let $\Upsilon > \sup_k ||\mathcal{R}(\boldsymbol{v}^{(k)})||_F$. For every $\epsilon_1 > 0$, set $\lambda_1 > \frac{L^3\Upsilon}{\epsilon_1}$ and $\lambda_2 > L^3\Upsilon + \lambda_1(\sqrt{n_o}L + 1)$, in Equation* (5)*. For the convergence point of the sequence $\{\boldsymbol{v}^{(k)}\} \subset \mathscr{F}$ which is generated by Algorithm 1 we have: $|\boldsymbol{v}^*|_- = 0$, $||s(\boldsymbol{v}^*) - \mathbf{1}||_2 \leq \epsilon_1$.*

To summarize, we initiate our exterior point method with the result of the method of moments estimator. By choosing large enough λ_1 and λ_2, the convergence point of the exterior point method will be at a local optimum of $g(\boldsymbol{v})$ with the constraint $\boldsymbol{v} \in \mathbb{R}_+^{n_s(3n_o+1)}$ in which the column sum of parameters set is arbitrarily close to 1. It is important to note that the proven lower bounds for λ_1, and λ_2 are sufficient condition for the convergence. In practice, cross-validation can find the best parameter to balance the speed of mapping to the simplex with optimizing the residual function. This method can easily be extended to find HMM parameters. After finding HMM parameters from the method of moments algorithm, Algorithm 1 can be used to further refine the solution. In order to use this algorithm, we just need to define the residual estimation term for HMMs, define the function $g(\cdot)$, and compute its gradient.

4 EXPERIMENTAL RESULTS

We evaluate the performance of our proposed method (EX&SVD) on both synthetic and real world datasets. We compare our approach to several state-of-the-art alternatives including EM initialized with 10 random seeds (EM), EM initialized with the method of moments result described in Section 2 after projecting the estimated parameters into simplex (EM&SVD), and the recently published symmetric tensor decomposition method (STD) [2]. To evaluate the performance gain due to exterior point algorithm, we also included results from method of moments without the additional optimization (SVD) Section 2. Parameters λ_1, and λ_2 controls the speed of mapping parameters into the simplex while estimation error term is optimized simultaneously. We find the best parameters using cross-validation. In our experiments, we sample N training and M testing points from each model. For the evaluation we use $M = 2000$ test samples and calculate normalized ℓ_1 $error = \frac{1}{M}\sum_{i=1}^{M} \frac{|\mathbb{P}(X_i) - \hat{\mathbb{P}}(X_i)|}{\mathbb{P}(X_i)}$.

First, we study the performance of different methods in estimating multi-view model parameters under different parameter set sizes. To this end, we sample $N = 4000$ training points from models with different numbers of hidden states and evaluate the performance of different method in estimating these models parameters. The average error of 10 independent runs is reported in figure 1 for different values of n_s. In each case we set n_o to twice the value of n_s. As n_s increases, the number of model parameters also increases while the number of training points remains fixed. This results in the estimation error increasing as the models get larger for all of the methods. However, the difference between the performance of EX&SVD and other methods becomes more pronounced with n_s, which shows the effectiveness of our method in learning models with large state spaces and relatively smaller datasets. Next, we evaluate the performance of our algorithm by estimating the

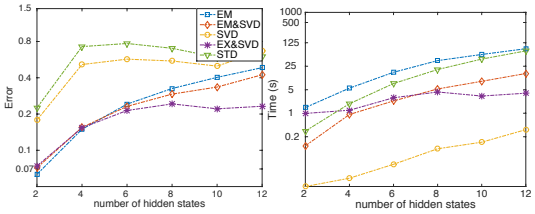

Figure 1: Error vs. n_s (left), time vs. n_s (right) for multi-view model for #training = 4000.

parameters of HMMs on synthetic dataset. Similar to the multi-view case, we randomly sample parameters from two different classes of models to generate synthetic datasets. The first set of models again has 5 hidden states and 10 discrete observations, and the second set of models has 10 hidden states and 20 observations. Figure 2 shows the average error of the implemented algorithms run on 10 different datasets generated by *i.i.d* sampled models. Although, estimated parameters in the SVD method do not have good performance in terms of normalized $l1$ error, using them to initiate an iterative optimization procedure improves performance as demonstrated by both the EM&SVD and the EX&SVD methods. On the other hand, our algorithm can also outperform EM&SVD, especially in the low and medium sample size regions when the error of projection step is relatively high.

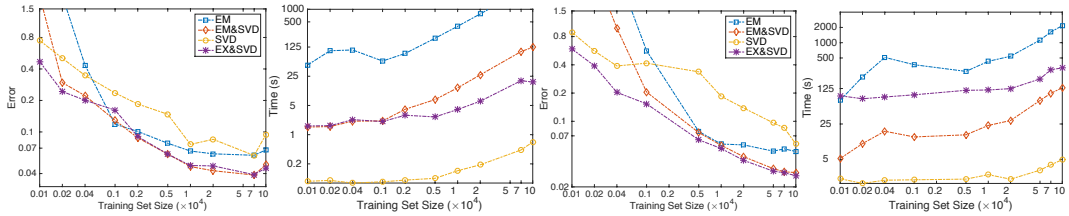

Figure 2: Error vs. #training, and time vs. #training for HMM models. $n_s = 5$, $n_o = 10$ (images #1, and #2), and $n_s = 10$, $n_o = 20$ (images #3, #4).

References

[1] A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y. Liu. Two svds suffice: Spectral decompositions for probabilistic topic modeling and latent dirichlet allocation. *CoRR*, abs/1204.6703, 2012.

[2] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *arXiv preprint arXiv:1210.7559*, 2012.

[3] A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden markov models. In *Proc. Annual Conf. Computational Learning Theory*, pages 33.1–33.34, 2012.

[4] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[5] B. Balle, W. Hamilton, and J. Pineau. Methods of moments for learning stochastic languages: Unified presentation and empirical comparison. In *Proceedings of the International Conference on Machine Learning*, pages 1386–1394, 2014.

[6] B. Balle, A. Quattoni, and X. Carreras. Local loss optimization in operator models: A new insight into spectral learning. In *Proceedings of the International Conference on Machine Learning*, 2012.

[7] S. B. Cohen, K. Stratos, M. Collins, D. P. Foster, and L. Ungar. Experiments with spectral learning of latent-variable PCFGs. In *Proceedings of NAACL*, 2013.

[8] P. L. Combettes and J. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–22, 1977.

[10] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandrae. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the International Conference on Machine Learning*, 2008.

[11] D. Hsu, S. Kakade, and T. Zhang. A spectral algorithm for learning hidden markov models. In *Proc. Annual Conf. Computational Learning Theory*, 2009.

[12] D. Hsu and S.M. Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions, 2012.

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

[14] A. P. Parikh, L. Song, M. Ishteva, G. Teodoru, and E.P. Xing. A spectral algorithm for latent junction trees. In *Conference on Uncertainty in Artificial Intelligence*, 2012.

[15] K. Pearson. Contributions to the mathematical theory of evolution. *Transactions of the Royal Society of London*, 185:71–110, 1894.

[16] S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, pages 1–41, 2013.

[17] S. Siddiqi, B. Boots, and G. J. Gordon. Reduced-rank hidden Markov models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-2010)*, 2010.

[18] L. Song, A. Anamdakumar, B. Dai, and B. Xie. Nonparametric estimation of multi-view latent variable models. In *International Conference on Machine Learning (ICML)*, 2014.

[19] L. Song, B. Boots, S. Siddiqi, G. Gordon, and A. J. Smola. Hilbert space embeddings of hidden markov models. In *International Conference on Machine Learning*, 2010.

[20] H. Yamashita and T. Tanabe. A primal-dual exterior point method for nonlinear optimization. *SIAM Journal on Optimization*, 20(6):3335–3363, 2010.

[21] Y. Zhang, X. Chen, D. Zhou, and M. I Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. In *Advances in Neural Information Processing Systems 14*, pages 1260–1268, 2014.