Spectral M-estimation with Application to Hidden Markov Models: Supplementary Material

Dustin Tran
Harvard University

Minjae Kim
Harvard University

Finale Doshi-Velez
Harvard University

1 Proof of Proposition 1

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

$$\min_{\mathbf{B}_x} \|\mathbf{P}_{3,x,1} - \mathbf{B}_x\mathbf{P}_{2,1}\|_F^2 \tag{1}$$

Equation (1) 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 (1) 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 (1) 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 1 is simply

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

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). □

2 Proof of Proposition 3

Proposition 3. *The gradients are*

$$\nabla_{\mathbf{R}}\mathcal{L} = \mathcal{J}_{\mathbf{R}}^\top\mathbf{W}g(\mathbf{X}, \{\mathbf{R}, \mathbf{S}\}) + \nabla_{\mathbf{R}}P_\alpha(\mathbf{R}, \mathbf{S}) \tag{3}$$

$$\nabla_{\mathbf{S}}\mathcal{L} = \mathcal{J}_{\mathbf{S}}^\top\mathbf{W}g(\mathbf{X}, \{\mathbf{R}, \mathbf{S}\}) + \nabla_{\mathbf{R}}P_\alpha(\mathbf{R}, \mathbf{S}) \tag{4}$$

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

$$[\mathcal{J}_{\mathbf{R}}]_{xij,uvw} = \begin{cases} -[\mathbf{S}_x^\top]_{w\cdot}[\mathbf{P}_{2,1}]_{\cdot j}, & \text{if } x = u,\ i = v\\ 0, & \text{otherwise}\end{cases} \tag{5}$$

and

$$[\mathcal{J}_{\mathbf{S}}]_{xij,uvw} = \begin{cases} -[\mathbf{R}_x]_{iw}[\mathbf{P}_{2,1}]_{vj}, & \text{if } x = u\\ 0, & \text{otherwise}\end{cases} \tag{6}$$

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

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

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

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

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

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

References

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

Byron Boots, Sajid Siddiqi, and Geoffrey J. Gordon. Closing the learning-planning loop with predictive state representations. In *Proceedings of Robotics: Science and Systems*, 2010.

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

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

Furong Huang, Niranjan U. N, Mohammad Umar Hakeem, Prateek Verma, and Animashree Anandkumar. Fast detection of overlapping communities via online tensor methods on gpus. *arXiv preprint arXiv:1309.0787*, 2013.

Nathan Srebro Nati and Tommi Jaakkola. Weighted low-rank approximations. In *International Conference on Machine Learning*, 2003.