A reunification of optimization and spectral methods: optimal learning in partially observable settings

Dustin Tran

(draft) April 30, 2015

Harvard University, School of Engineering and Applied Sciences

How can we best learn parameters in a latent variable model?

Motivation

In statistical estimation theory, the ideal estimator satisfies four requirements:

- consistency
- statistical efficiency
- computational efficiency
- numerical stability

Motivation

The EM algorithm satisfies:

- consistency ✗
- statistical efficiency ✓ (assuming global)
- computational efficiency ✗
- numerical stability ✓

Motivation

Variational inference satisfies:

- consistency ✗
- statistical efficiency ✓ (assuming global; SVI with averaging)
- computational efficiency ✓ (SVI)
- numerical stability ✓ (implicit/proximal SVI)

Motivation

Markov chain Monte carlo satisfies:

- consistency ✓
- statistical efficiency ✓
- computational efficiency ✗
- numerical stability ✗(??)

What are spectral methods?

They are matrix (tensor) decompositions which generate sample moments, i.e., observable representations that are functions of the parameters.

- Can solve for them and asymptotically recover the ground truth in expectation, i.e., lead to consistent estimators
- Are consistent estimators for (certain) nonconvex problems!

Motivation

Spectral methods satisfy:

- consistency ✓
- statistical efficiency ✗
- computational efficiency ✓
- numerical stability ✗

Can we leverage the optimization viewpoint to obtain statistical efficiency and numerical stability?

Motivation

Given a cost function $\ell(\theta; Y)$, can:

- Add regularization/priors $f(\theta)$
- Add weighted distance (generalized method of moments)
- Use efficient and numerically stable optimization routines (stochastic gradient methods)

Background: Generalized method of moments (GMMs)

Let $Y_1, \ldots, Y_N$ be $(d+1)$-dimensional observations generated from some model with unknown parameters $\theta \in \Theta$.

The k *moment conditions* for a vector-valued function $g(Y, \cdot) : \Theta \to \mathbb{R}^k$ is

$$m(\theta^*) \stackrel{\text{def}}{=} \mathbb{E}[g(Y, \theta^*)] = 0_{k \times 1}$$

The *observable representations*, or sample moments, are

$$\widehat{m}(\theta) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} g(Y_n, \theta)$$

We aim to minimize the normed distance $\|\widehat{m}(\theta)\|$ for some choice of $\|\cdot\|$. Define the *weighted Frobenius norm* as

$$\|\widehat{m}(\theta)\|_W^2 \stackrel{\text{def}}{=} \widehat{m}(\theta)^T W \widehat{m}(\theta),$$

where W is a positive definite matrix. The *generalized method of moments* (GMM) estimator is

$$\theta^{gmm} = \arg\min_{\theta \in \Theta} \|\widehat{m}(\theta)\|_W^2$$

Under standard assumptions*, the GMM estimator θ^{gmm} is consistent and asymptotically normal. Moreover, if

$$W \stackrel{\text{def}}{=} \mathbb{E}[g(Y_n, \theta^*)g(Y_n, \theta^*)^T]^{-1}$$

then θ^{gmm} is statistically efficient in the class of asymptotically normal estimators (and conditioned only on the information in the moment!).

Hidden markov model

Time $t \in \{1, 2, \ldots\}$

- Hidden states $h_t \in \{1, \ldots, m\}$
- Observations $x_t \in \{1, \ldots, n\}$ $(m \leq n)$

Given full rank matrices $T \in \mathbb{R}^{m \times m}$, $O \in \mathbb{R}^{n \times m}$, the dynamical system for HMMs is given by

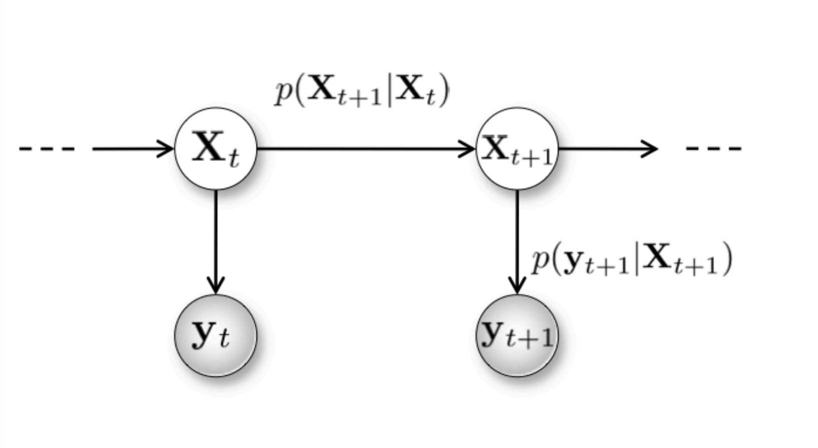
$$T_{ij} \stackrel{\text{def}}{=} P[h_{t+1} = i \mid h_t = j] \iff P[h_{t+1}] = TP[h_t]$$

$$O_{ij} \stackrel{\text{def}}{=} P[x_t = i \mid h_t = j] \iff P[x_t] = OP[h_t]$$

and $\pi \stackrel{\text{def}}{=} P[h_1] \in \mathbb{R}^m$ the initial state distribution.

Goal: Estimate the joint distribution $P[x_{1:t}]$, which also allows one to make predictions: $P[x_{t+1} \mid x_{1:t}] = [x_{1:t+1}]/P[x_{1:t}]$.

Hidden markov model



Notation: $P[x, y, \cdots]_{ij\cdots} \stackrel{\text{def}}{=} P(x = i, y = j, \cdots)$

Theorem

The joint probability matrix $P[x_{t+1}, x_t]$ satisfies the following for all columns $j \in \{1, \ldots, n\}$:

$$P[x_{t+1}, x_t]_{\cdot j} = \Phi_j P[x_t] \qquad \Phi_j \stackrel{\text{def}}{=} OT \operatorname{diag}(O_{j\cdot}) O^\dagger$$

Furthermore,

$$P[x_{t+1}, x_t, x_{1:t-1}]_{\cdot j\vec{k}} = \Phi_j P[x_t, x_{1:t-1}]_{\cdot\vec{k}}$$

where $\vec{k}$ represents a sequence of states for $x_{1:t-1}$, i.e., $x_1 = k_1, \ldots, x_{t-1} = k_{t-1}$.

We aim to solve

$$\min_{\Phi_j} \|\widehat{P}[x_{t+1}, x_t, x_{1:t-1}]_{\cdot j\vec{k}} - \Phi_j \widehat{P}[x_t, x_{1:t-1}]_{\cdot \vec{k}}\|_F$$

such that $\text{rank}(\Phi_j) \leq m$ for all $j \in \{1, \ldots, n\}$. By the Eckart-Young-Mirsky theorem, this is equivalent to solving

$$\min_{\Phi_j} \|U_j^T \widehat{P}[x_{t+1}, x_t, x_{1:t-1}]_{\cdot j\vec{k}} - \Phi_j U_j^T \widehat{P}[x_t, x_{1:t-1}]_{\cdot \vec{k}}\|_F$$

where U_j is the matrix of m left-singular vectors of

$$\mathbb{E}[\widehat{P}[x_{t+1}, x_t, x_{1:t-1}]_{\cdot j\vec{k}} \widehat{P}[x_t, x_{1:t-1}]_{\cdot \vec{k}}^T]$$

- It's convex! (optimization leads to consistent estimator)
- This recovers the same **(???)** parameters as the method of moments estimation derived in Hsu et al. (2009).

Use weighted Frobenius norm instead for GMM estimation:

$$\min_{\Phi_j} \|\widehat{P}[x_{t+1}, x_t, x_{1:t-1}]_{\cdot j\vec{k}} - \Phi_j \widehat{P}[x_t, x_{1:t-1}]_{\cdot\vec{k}}\|_W^2$$

such that $\text{rank}(\Phi_j) \leq m$.

Hidden markov model

And add your favorite regularizers/priors!

$$\min_{\Phi_j} \|\widehat{P}[x_{t+1}, x_t, x_{1:t-1}]_{\cdot j\vec{k}} - \Phi_j \widehat{P}[x_t, x_{1:t-1}]_{\cdot \vec{k}}\|_W^2 + \alpha\|\Phi_j\|_1 + (1-\alpha)\|\Phi_j\|_F^2$$

such that $\mathrm{rank}(\Phi_j) \leq m$.

Hidden markov model

$$\min_{\Phi_j} \|\widehat{P}[x_{t+1}, x_t, x_{1:t-1}]_{\cdot j\vec{k}} - \Phi_j \widehat{P}[x_t, x_{1:t-1}]_{\cdot \vec{k}}\|_W^2 \qquad \text{rank}(\Phi_j) \leq m$$

1. This extends naturally to predictive state representations (Boots and Gordon, 2010) (**???** work in progress).
2. Remark on local optima in the optimization

- In the unweighted case (MoM), any local optima are necessarily global
- In the weighted case (GMM), this no longer holds (Nati and Jaakola, 2003)

$$\min_{\Phi_j} \|\widehat{P}[x_{t+1}, x_t, x_{1:t-1}]_{\cdot j\vec{k}} - \Phi_j \widehat{P}[x_t, x_{1:t-1}]_{\cdot \vec{k}}\|_W^2 \qquad \operatorname{rank}(\Phi_j) \leq m$$

If the weighting introduces local optima, then what's the point?

Conclusion

Reframing moment estimation from spectral methods as optimization leads to an alternative viewpoint on improving estimation:

- consistency ✓ **(???)**
- statistical efficiency ✓ (GMM estimator, using SGD with averaging)
- computational efficiency ✓ (SGD)
- numerical stability ✓ (implicit/proximal SGD)

Questions?