Statistical and computational guarantees for the Baum-Welch algorithm

Fanny Yang*, Sivaraman Balakrishnan†, Martin Wainwright*,†

EECS Department*, Statistics Department†
UC Berkeley

53rd Annual Allerton Conference
Allerton, September 30th, 2015
Outline

1. Background
   - Hidden Markov Model and algorithms
   - Classical convergence analysis of the EM algorithm

2. Main results
   - Convergence guarantees for the Baum-Welch algorithm
   - Special case - Gaussian output HMM
   - Discussion
Hidden Markov Models

Parametric Hidden Markov Model (HMM) with discrete latent variables $Z_i$ and observed variables $X_i$

$$p(z_i|z_{i-1}, \beta)$$

Model assumptions

- Markov chain $\{Z_i\}_{i \in \mathbb{Z}}$ has a unique stationary distribution, is sufficiently mixing
Hidden Markov Models

Parametric Hidden Markov Model (HMM) with discrete latent variables $Z_i$ and observed variables $X_i$

Model assumptions

- Markov chain $\{Z_i\}_{i \in \mathbb{Z}}$ has a unique stationary distribution, is sufficiently mixing
- HMM parameterized by $\theta = (\beta, \mu)$
Hidden Markov Models

Parametric Hidden Markov Model (HMM) with discrete latent variables $Z_i$ and observed variables $X_i$

Model assumptions
- Markov chain $\{Z_i\}_{i \in \mathbb{Z}}$ has a unique stationary distribution, is sufficiently mixing
- HMM parameterized by $\theta = (\beta, \mu)$

Goal: Reconstruct $\theta$ from a sequence of observations $X_1^n := X_1, \ldots, X_n$ drawn from the HMM
Hidden Markov Models - Reconstruction algorithms

Algorithms for general HMM reconstruction, e.g.

- Classical approach: Baum-Welch algorithm (Baum et al. 1970) and EM variants (with k-means) (Dasgupta and Schulman 2007)
- Spectral methods (Hsu, Kakade and Zhang 2012)
- Convex programs for parametric-output HMMs (Kontorovich et al. 2013)
Hidden Markov Models - Reconstruction algorithms

Algorithms for general HMM reconstruction, e.g.
- Classical approach: Baum-Welch algorithm (Baum et al. 1970) and EM variants (with k-means) (Dasgupta and Schulman 2007)
- Spectral methods (Hsu, Kakade and Zhang 2012)
- Convex programs for parametric-output HMMs (Kontorovich et al. 2013)

Motivation to analyze EM
- Widely used and easy to implement for many models
- Empirically high precision when initialized correctly
- But lack of theoretical work which explains this behavior (for i.i.d. data Balakrishnan et al. 2014)
Hidden Markov Models - EM algorithm

Recall the complete log likelihood

$$
\log p(z_1^n, x_1^n; \theta) = \log \left[ \pi_1(z_1) \prod_{i=2}^{n} p(z_i | z_{i-1}; \beta) \prod_{i=1}^{n} p(x_i | z_i, \mu) \right]
$$

$$
= \sum_{i=1}^{n} \log f(z_{i-1}^i; \beta, \mu)
$$

EM maximizes the log likelihood

$$
\hat{\theta}_{\text{MLE}} := \arg \max_{\theta} \ell_n(\theta) := \arg \max_{\theta} \left[ \log p(x_1^n; \theta) \right]
$$

$$
= \arg \max_{\theta} \left[ \log \sum_{z_1^n} p(z_1^n, x_1^n; \theta) \right]
$$
Hidden Markov Models - EM algorithm

Recall the complete log likelihood

\[
\log p(z_1^n, x_1^n; \theta) = \log \left[ \pi_1(z_1) \prod_{i=2}^{n} p(z_i \mid z_{i-1}; \beta) \prod_{i=1}^{n} p(x_i \mid z_i, \mu) \right]
\]

\[= \sum_{i=1}^{n} \log f(z_{i-1}; \beta, \mu) \]

EM maximizes the log likelihood

\[\hat{\theta}_{MLE} := \arg \max_{\theta} \ell_n(\theta) := \arg \max_{\theta} \left[ \log p(x_1^n; \theta) \right] \]

\[= \arg \max_{\theta} \left[ \log \sum_{z_1^n} p(z_1^n, x_1^n; \theta) \right] \]

**Key of EM**: Optimize over the *expected* complete log likelihood
Observe the lower bound for all $\theta'$

$$\ell_n(\theta) \geq \frac{1}{n} \mathbb{E}_{Z_1^n|X_1^n,\theta'} [\log p(X_1^n, Z_1^n; \theta)] + H(\theta')$$

$$=: Q_n(\theta | \theta').$$

with $\ell_n(\theta) = Q_n(\theta | \theta)$. 
EM algorithm for latent variable models

Using \( \ell_n(\theta) \geq Q_n(\theta | \theta') \) and \( \ell_n(\theta') = Q_n(\theta' | \theta') \)

Algorithm (EM algorithm)

*Initialize with \( \hat{\theta}^0 \) and then iterate until convergence*

\[
\hat{\theta}^{t+1} = \arg \max_{\theta \in \Theta} Q_n(\theta | \hat{\theta}^t)
\]
EM algorithm for latent variable models

Using $\ell_n(\theta) \geq Q_n(\theta \mid \theta')$ and $\ell_n(\theta') = Q_n(\theta' \mid \theta')$

Algorithm (EM algorithm)

Initialize with $\hat{\theta}^0$ and then iterate until convergence

$\hat{\theta}^{t+1} = \arg \max_{\theta \in \Theta} Q_n(\theta \mid \hat{\theta}^t)$
EM algorithm for latent variable models

Using $\ell_n(\theta) \geq Q_n(\theta \mid \theta')$ and $\ell_n(\theta') = Q_n(\theta' \mid \theta')$

Algorithm (EM algorithm)

Initialize with $\hat{\theta}^0$ and then iterate until convergence

$$\hat{\theta}^{t+1} = \arg\max_{\theta \in \Theta} Q_n(\theta \mid \hat{\theta}^t)$$
EM algorithm for latent variable models

Using $\ell_n(\theta) \geq Q_n(\theta \mid \theta')$ and $\ell_n(\theta') = Q_n(\theta' \mid \theta')$

Algorithm (EM algorithm)

Initialize with $\hat{\theta}^0$ and then iterate until convergence

$$\hat{\theta}^{t+1} = \arg\max_{\theta \in \Theta} Q_n(\theta \mid \hat{\theta}^t)$$
EM algorithm - classical convergence analysis


- MLE is a fixed point of $Q_n$, i.e. $\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} Q_n(\theta | \hat{\theta}_{\text{MLE}})$
- Under regularity assumptions $\hat{\theta}^t$ converges to a stationary point of the sample likelihood
- Linear convergence close to $\hat{\theta}_{\text{MLE}}$ via Taylor’s theorem

- MLE is a fixed point of $Q_n$, i.e. $\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} Q_n(\theta | \hat{\theta}_{\text{MLE}})$
- Under regularity assumptions $\hat{\theta}^t$ converges to a stationary point of the sample likelihood
- Linear convergence close to $\hat{\theta}_{\text{MLE}}$ via Taylor’s theorem

Main caveat: necessary initialization radius or linear convergence rates only given in an **arbitrarily small ball** around the MLE
Classical analysis - caveats

Only guarantees convergence to some stationary point of \( \ell_n \) for convergence to MLE very close initialization necessary. In fact arbitrarily small if the local maxima are close.
Classical analysis - caveats

- Only guarantees convergence to some stationary point of $\ell_n$
Classical analysis - caveats

- Only guarantees convergence to some stationary point of $\ell_n$

$\rightarrow$ for convergence to MLE very close initialization necessary
Classical analysis - caveats

- Only guarantees convergence to some stationary point of $\ell_n$
  $\rightarrow$ for convergence to MLE very close initialization necessary
- In fact arbitrarily small if the local maxima are close
Only guarantees convergence to some stationary point of $\ell_n$

$\rightarrow$ for convergence to MLE very close initialization necessary

In fact arbitrarily small if the local maxima are close
Characterize initialization radius $r$ around $\theta^*$ such that

- all stationary points in $r$ are in an $\epsilon_n$ ball around $\theta^*$
- Baum-Welch converges linearly to the $\epsilon_n$ ball
Our contributions

Characterize initialization radius $r$ around $\theta^*$ such that

- all stationary points in $r$ are in an $\epsilon_n$ ball around $\theta^*$
- Baum-Welch converges linearly to the $\epsilon_n$ ball
Characterize initialization radius $r$ around $\theta^*$ such that

- all stationary points in $r$ are in an $\epsilon_n$ ball around $\theta^*$
- Baum-Welch converges linearly to the $\epsilon_n$ ball
Baum-Welch algorithm - empirical observations

\[ \log_{10} \| \hat{\theta}_i^t - \hat{\theta}_1^\infty \|_2 \]

\[ \log_{10} \| \hat{\theta}_i^t - \theta^* \|_2 \]
Regularity conditions:

- "Lipschitz condition" on $Q = \mathbb{E} Q_n$, i.e.
  $$\sup_{\theta} \| \nabla_{\theta} Q(\theta \mid \theta') - \nabla_{\theta} Q(\theta \mid \theta^*) \|_2 \leq L \| \theta' - \theta^* \|_2$$

- Strong concavity of $Q$, i.e.
  $$Q(\theta_1) - Q(\theta_2) \leq \nabla Q(\theta_2)^T (\theta_1 - \theta_2) - \frac{\lambda}{2} \| \theta_1 - \theta_2 \|_2^2$$
Regularity conditions:

- “Lipschitz condition” on $Q = \mathbb{E} Q_n$, i.e.
  \[ \sup_{\theta} \| \nabla_\theta Q(\theta | \theta') - \nabla_\theta Q(\theta | \theta^*) \|_2 \leq L \| \theta' - \theta^* \|_2 \]

- Strong concavity of $Q$, i.e.
  \[ Q(\theta_1) - Q(\theta_2) \leq \nabla Q(\theta_2)^T (\theta_1 - \theta_2) - \frac{\lambda}{2} \| \theta_1 - \theta_2 \|_2^2 \]

**Theorem (Y., Balakrishnan, and Wainwright '15)**

If the regularity conditions hold for $\theta' \in B(r; \theta^*)$ with $L, \lambda$ then

\[ \| \hat{\theta}^t - \theta^* \|_2 \leq \kappa^t \| \hat{\theta}^0 - \theta^* \|_2 + \frac{1}{1 - \kappa} \left( \phi(n, \delta, k) + \epsilon(n, \delta, k) \right) \]

with linear convergence, truncation error, and finite sample error with probability $1 - \delta$ and $\kappa = \frac{L}{\lambda}$. 
Baum-Welch algorithm - theoretical guarantees

Regularity conditions:

- “Lipschitz condition” on \( Q = \mathbb{E} Q_n \), i.e.
  \[
  \sup_{\theta} \| \nabla_{\theta} Q(\theta \mid \theta') - \nabla_{\theta} Q(\theta \mid \theta^*) \|_2 \leq L \| \theta' - \theta^* \|_2
  \]

- Strong concavity of \( Q \), i.e.
  \[
  Q(\theta_1) - Q(\theta_2) \leq \nabla Q(\theta_2)^T (\theta_1 - \theta_2) - \frac{\lambda}{2} \| \theta_1 - \theta_2 \|_2^2
  \]

Theorem (Y., Balakrishnan and Wainwright '15)

If the regularity conditions hold for \( \theta' \in B(r; \theta^*) \) with \( L, \lambda \) then

\[
\| \hat{\theta}^t - \theta^* \|_2 \leq \kappa^t \| \hat{\theta}^0 - \theta^* \|_2 + \frac{1}{1 - \kappa} \left( \phi(n, \delta, k) + \epsilon(n, \delta, k) \right)
\]

with linear convergence \( \kappa^t \)

with probability \( 1 - \delta \) and \( \kappa = \frac{L}{\lambda} \).
Regularity conditions:

- “Lipschitz condition” on $Q = \mathbb{E} Q_n$, i.e.
  $$\sup_{\theta} \| \nabla_{\theta} Q(\theta | \theta') - \nabla_{\theta} Q(\theta | \theta^*) \|_2 \leq L \| \theta' - \theta^* \|_2$$

- Strong concavity of $Q$, i.e.
  $$Q(\theta_1) - Q(\theta_2) \leq \nabla Q(\theta_2)^T (\theta_1 - \theta_2) - \frac{\lambda}{2} \| \theta_1 - \theta_2 \|_2^2$$

---

Theorem (Y., Balakrishnan and Wainwright ’15)

If the regularity conditions hold for $\theta' \in B(r; \theta^*)$ with $L, \lambda$ then

$$\| \hat{\theta}^t - \theta^* \|_2 \leq \kappa^t \| \hat{\theta}^0 - \theta^* \|_2 + \frac{1}{1 - \kappa} \left( \phi(n, \delta, k) + \epsilon(n, \delta, k) \right)$$

with linear convergence, truncation error, and finite sample error.

with probability $1 - \delta$ and $\kappa = \frac{L}{\lambda}$.
Baum-Welch algorithm - theoretical guarantees

Regularity conditions:

- "Lipschitz condition" on $Q = \mathbb{E} Q_n$, i.e.
  \[
  \sup_{\theta} \| \nabla_{\theta} Q(\theta | \theta') - \nabla_{\theta} Q(\theta | \theta^*) \|_2 \leq L \| \theta' - \theta^* \|_2
  \]

- Strong concavity of $Q$, i.e.
  \[
  Q(\theta_1) - Q(\theta_2) \leq \nabla Q(\theta_2)^T (\theta_1 - \theta_2) - \frac{\lambda}{2} \| \theta_1 - \theta_2 \|_2^2
  \]

Theorem (Y., Balakrishnan and Wainwright '15)

If the regularity conditions hold for $\theta' \in B(r; \theta^*)$ with $L, \lambda$ then

\[
\| \hat{\theta}^t - \theta^* \|_2 \leq \kappa^t \| \hat{\theta}^0 - \theta^* \|_2 + \frac{1}{1 - \kappa} \left( \phi(n, \delta, k) + \epsilon(n, \delta, k) \right)
\]

with linear convergence, truncation error, finite sample error, with probability $1 - \delta$ and $\kappa = \frac{L}{\lambda}$. 
Proof idea - Additional steps for dependent data

Based on a framework by Balakrishnan et al. 2014 for i.i.d. data

For example, $\mathbb{E} Q_n$ depends on $n$ and it is a sum of dependent random variables:

$$Q_n(\theta | \theta') \sim \frac{1}{n} \mathbb{E}_{Z^1_n|X^1_n,\theta'} \left[ \log p(X^n, Z^n; \theta) \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{Z^{i+1}_i|X^n,\theta'} f_i(Z^{i+1}_i, X_i, \theta)$$
Proof idea - Additional steps for dependent data

Based on a framework by Balakrishnan et al. 2014 for i.i.d. data

For example \( \mathbb{E} Q_n \) depends on \( n \) and it is a sum of dependent random variables:

\[
Q_n(\theta \mid \theta') \sim \frac{1}{n} \mathbb{E} Z_1^n | X_1^n, \theta' \left[ \log p(X_1^n, Z_1^n; \theta) \right]
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} Z_{i+1}^n | X_1^n, \theta' f_i(Z_{i+1}^n, X_i, \theta)
\]

Key feature is mixing, i.e. \( \sup_{i,j} |P(Z_k = i \mid Z_0 = j) - \pi(i)| \leq c \rho^k \)
Based on a framework by Balakrishnan et al. 2014 for i.i.d. data

For example $\mathbb{E} Q_n$ depends on $n$ and it is a sum of dependent random variables:

$$Q_n(\theta | \theta') \sim \frac{1}{n} \mathbb{E}_{Z_1^n | X_1^n, \theta'} \left[ \log p(X_1^n, Z_1^n; \theta) \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{Z_{i+1}^i | X_1^n, \theta'} f_i(Z_{i+1}, X_i, \theta)$$

Key feature is mixing, i.e. $\sup_{i,j} |P(Z_k = i | Z_0 = j) - \pi(i)| \leq c \rho^k$

- Replace $\mathbb{E}_{Z_{i+1}^i | X_1^n}$ by $\mathbb{E}_{Z_{i+1}^i | X_{i-k}^{i+k}} \rightarrow$ truncation error $\phi(n, \delta, k)$
  - decaying exponentially with $k$
Proof idea - Additional steps for dependent data

Based on a framework by Balakrishnan et al. 2014 for i.i.d. data

For example $\mathbb{E} Q_n$ depends on $n$ and it is a sum of dependent random variables:

$$Q_n(\theta \mid \theta') \sim \frac{1}{n} \mathbb{E}_{Z_1^n \mid X_1^n, \theta'} \left[ \log p(X_1^n, Z_1^n; \theta) \right]$$

$$\sim \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{Z_{i+1}^{i+k} \mid X_{i-k}^i, \theta'} f_i(Z_{i+1}^{i+k}, X_i, \theta)$$

Key feature is mixing, i.e. $\sup_{i,j} |P(Z_k = i \mid Z_0 = j) - \pi(i)| \leq c \rho^k$

- Replace $\mathbb{E}_{Z_{i+1}^n \mid X_1^n}$ by $\mathbb{E}_{Z_{i+1}^{i+k} \mid X_{i-k}^i}$ → truncation error $\phi(n, \delta, k)$ decaying exponentially with $k$
Proof idea - Additional steps for dependent data

Based on a framework by Balakrishnan et al. 2014 for i.i.d. data

For example $\mathbb{E}Q_n$ depends on $n$ and it is a sum of dependent random variables:

$$Q_n(\theta \mid \theta') \sim \frac{1}{n} \mathbb{E}_{Z^n|X^{n}_1,\theta'} \left[ \log p(X^n_1, Z^n_1; \theta) \right]$$

$$\sim \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{Z^{i+1}_i|X^{i+k}_{i-k},\theta'} f_i(Z^{i+1}_i, X_i, \theta)$$

Key feature is mixing, i.e. $\sup_{i,j} |P(Z_k = i \mid Z_0 = j) - \pi(i)| \leq c \rho^k$

- Replace $\mathbb{E}_{Z^{i+1}_i|X^n_1}$ by $\mathbb{E}_{Z^{i+1}_i|X^{i+k}_{i-k}} \rightarrow$ truncation error $\phi(n, \delta, k)$ decaying exponentially with $k$
- Use that far away blocks $\{X^{i+k}_{i-k}\}_{i=1,...,n}$ are almost independent
Guarantees for Gaussian output HMM

Model assumptions
- $Z_i \in \{-1, +1\}$ is a mixing Markov chain with symmetric transition matrix
- Normal observation densities $X_i \mid Z_i \sim \mathcal{N}(Z_i \mu^*, \sigma^2 I)$.

Defining the SNR $\eta^2 = \frac{\|\mu^*\|^2}{\sigma^2}$ we have
Guarantees for Gaussian output HMM

Model assumptions

- \( Z_i \in \{-1, +1\} \) is a mixing Markov chain with symmetric transition matrix
- Normal observation densities \( X_i \mid Z_i \sim \mathcal{N}(Z_i \mu^*, \sigma^2 I) \).

Defining the SNR \( \eta^2 = \frac{||\mu^*||_2^2}{\sigma^2} \) we have

**Corollary (Y., Balakrishnan and Wainwright ’15)**

*Given \( n \gtrsim d \log^2(d/\delta) \) and the initialization \( \hat{\mu}^0 \in B_2\left(\frac{||\mu^*||_2}{4}; \mu^*\right) \), we obtain*

\[
||\hat{\theta}^t - \theta^*||_2 \leq \kappa^t ||\hat{\theta}^0 - \theta^*||_2 + \frac{C}{1 - \kappa} \left[ \sigma \sqrt{\frac{d \log^2(n/\delta)}{n}} + ||\mu^*||_2 \sqrt{\frac{\log^2(n/\delta)}{n}} \right]
\]

*with probability at least \( 1 - \delta \) and \( \kappa \propto e^{-c\eta^2 \log d} \).*
 guarantees for Gaussian output HMM

Model assumptions

- $Z_i \in \{-1, +1\}$ is a mixing Markov chain with symmetric transition matrix
- Normal observation densities $X_i | Z_i \sim \mathcal{N}(Z_i \mu^*, \sigma^2 I)$.

Defining the SNR $\eta^2 = \frac{\|\mu^*\|^2}{\sigma^2}$ we have

**Corollary (Y., Balakrishnan and Wainwright ’15)**

*Given $n \gtrsim d \log^2 (d/\delta)$ and the initialization $\hat{\mu}^0 \in \mathbb{B}_2 \left( \frac{\|\mu^*\|_2}{4}; \mu^* \right)$, we obtain*

$$
\| \hat{\theta}^t - \theta^* \|_2 \leq \kappa^t \| \hat{\theta}^0 - \theta^* \|_2 + \frac{C}{1 - \kappa} \left[ \sigma \sqrt{\frac{d \log^2 (n/\delta)}{n}} + \| \mu^* \|_2 \sqrt{\frac{\log^2 (n/\delta)}{n}} \right]
$$

*with probability at least $1 - \delta$ and $\kappa \propto e^{-c\eta^2 \log d}$. 

Guarantees for Gaussian output HMM

Model assumptions

- $Z_i \in \{-1, +1\}$ is a mixing Markov chain with symmetric transition matrix

- Normal observation densities $X_i \mid Z_i \sim \mathcal{N}(Z_i \mu^*, \sigma^2 I)$.

Defining the SNR $\eta^2 = \frac{\|\mu^*\|^2}{\sigma^2}$ we have

Corollary (Y., Balakrishnan and Wainwright '15)

Given $n \gtrsim d \log^2 (d/\delta)$ and the initialization $\hat{\mu}^0 \in B_2(\frac{\|\mu^*\|_2}{4}; \mu^*)$, we obtain

$$\|\hat{\theta}^t - \theta^*\|_2 \leq \kappa^t \|\hat{\theta}^0 - \theta^*\|_2 + \frac{C}{1 - \kappa} \left[ \sigma \sqrt{\frac{d \log^2 (n/\delta)}{n}} + \|\mu^*\|_2 \sqrt{\frac{\log^2 (n/\delta)}{n}} \right]$$

with probability at least $1 - \delta$ and $\kappa \propto e^{-c\eta^2 \log d}$. 
Guarantees for Gaussian output HMM

Model assumptions

- \( Z_i \in \{-1, +1\} \) is a mixing Markov chain with symmetric transition matrix

- Normal observation densities \( X_i \mid Z_i \sim \mathcal{N}(Z_i \mu^*, \sigma^2 I) \).

Defining the SNR \( \eta^2 = \frac{\|\mu^*\|_2^2}{\sigma^2} \) we have

**Corollary (Y., Balakrishnan and Wainwright ’15)**

\[
\text{Given } n \gtrsim d \log^2(d/\delta) \text{ and the initialization } \hat{\mu}^0 \in \mathbb{B}_2\left(\frac{\|\mu^*\|_2}{4}; \mu^*\right), \text{ we obtain}
\]

\[
\|\hat{\theta}^t - \theta^*\|_2 \leq \kappa^t \|\hat{\theta}^0 - \theta^*\|_2 + \frac{C}{1 - \kappa} \left[ \sigma \sqrt{\frac{d \log^2(n/\delta)}{n}} + \|\mu^*\|_2 \sqrt{\frac{\log^2(n/\delta)}{n}} \right]
\]

with probability at least \( 1 - \delta \) and \( \kappa \propto e^{-c\eta^2 \log d} \).
Convergence rate dependence on SNR

Parameter settings: $d = 10$, $n = 1000$, $\sigma = 2$
Contributions

In this work we

- guarantee that the Baum-Welch estimate gets close to $\theta^*$
- characterize the basin of attraction depending on the model
- give a linear convergence rate of the BW algorithm in this basin
- provide explicit values for the Gaussian output HMM example
Some open questions

- How hard is it to compute Lipschitz constants for models where the output distribution is not an exponential family?
- Can we extend the framework to more general graphical models like lattices etc.?
Some open questions

- How hard is it to compute Lipschitz constants for models where the output distribution is not an exponential family?
- Can we extend the framework to more general graphical models like lattices etc.?

THANK YOU!