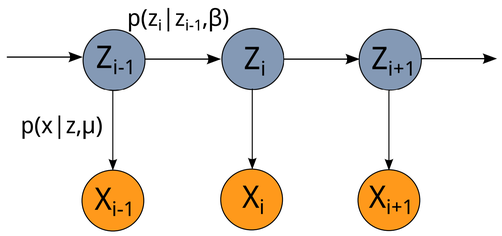
Statistical guarantees for the Baum-Welch algorithm

Fanny Yang[†], Sivaraman Balakrishnan[⋆], Martin Wainwright[†,⋆]
Department of EECS[†], Department of Statistics[⋆], UC Berkeley

Problem statement: Parameter estimation for HMMs

Model: We consider the following Hidden Markov Model (HMM)



with discrete latent variables Z_i (homogeneous Markov Chain with transition parameter β) and observed variables X_i (parameter μ).
The joint probability then reads

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

Goal: Estimate $\theta = (\beta, \mu)$ from a sequence of observations $X_1^n := X_1 \dots X_n$

Provable algorithms: e.g. spectral methods [Hsu et al. '12], parametric-output HMMs [Kontorovich et al. '13] and many more...

$\implies$ Baum-Welch (EM) is easy and works well empirically in two step procedures, but lack of theoretical work

Recap: Baum-Welch/EM algorithm

Motivation: Determine the MLE, i.e.

$$\widehat{\theta}_n := \arg\max_\theta \ell_n(\theta) = \arg\max_\theta \log p(x_1^n; \theta)$$

since it's known to be asymptotically normal (e.g. [Bickel et al. '98])
Catch: MLE not computable directly, but $p(x|z)$ is easy to maximize.

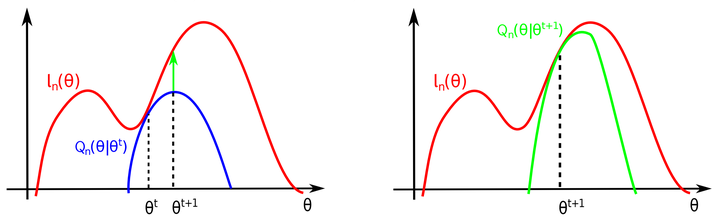
Observations:

- Concave lower bound $Q_n(\cdot \mid \theta')$

$$\ell_n(\theta) \geq \underbrace{\frac{1}{n}\mathbb{E}_{Z_1^n | x_1^n, \theta'}\big[\log p(x_1^n, Z_1^n; \theta)\big] + H_n(\theta')}_{Q_n(\theta \,|\theta')}$$

- Bound matches likelihood for $\theta' = \theta$, i.e. $\ell_n(\theta) = Q_n(\theta \mid \theta)$

$\implies$ **EM algorithm** (called Baum-Welch when applied to HMM):
M-step: Given $\widehat{\theta}^t$ find $\widehat{\theta}^{t+1} = \arg\max_\theta Q_n(\theta \mid \widehat{\theta}^t)$
E-step: Compute $f(\theta) = Q_n(\theta \mid \widehat{\theta}^{t+1})$



Classical EM convergence analysis

Classical convergence guarantees (under regularity assumptions)

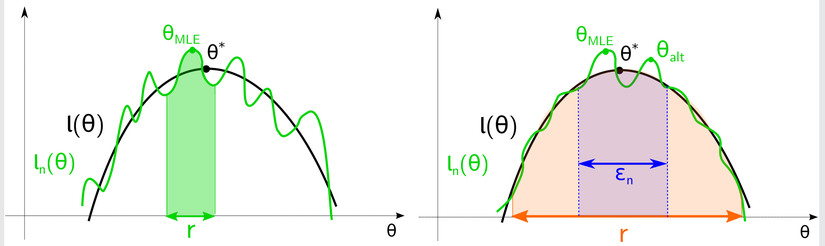
- possibly linear convergence to a stationary point of likelihood
- MLE is a fixed point of Q_n, i.e. $\widehat{\theta}_{MLE} = \arg\max_\theta Q_n(\theta \mid \widehat{\theta}_{MLE})$

Main caveats

- Basin of attraction for $\widehat{\theta}_{MLE}$ could be arbitrarily small
- Growing sample size does not help here, since uniform convergence does not state anything about location of local maxima

$\implies$ Not that useful in practice!
Assuming identifiability, i.e. population optimum is equal true parameter, we do not necessarily need convergence to MLE – any other solution which is close to population optimum is equally valid!

Flavor of our results for the Baum-Welch algorithm



Classical EM analysis

This line of work

Our approach and results:

- Focus on ϵ_n ball around θ^* instead of MLE
- Find a big initialization radius such that it converges to ϵ_n ball, i.e. local minima inside radius r are all within this (much smaller) ball
- Provide a convergence rate given model parameters

Previous works of similar flavor for the i.i.d. case:
Balakrishnan et al. '14, Wang et al '14, Yi, Caramanis '15

References

- S. Balakrishnan, M. Wainwright, B. Yu, "Statistical guarantees for the EM algorithm: From population to sample-based analysis," *arXiv preprint arXiv:1408.2156*, 2014
- Z. Wang, Q. Gu, Y. Ning and H. Liu, "High Dimensional Expectation-Maximization Algorithm: Statistical Optimization and Asymptotic Normality," *arXiv preprint arXiv:1412.8729*, 2014
- X Yi, C. Caramanis, "Regularized EM Algorithms: A Unified Framework and Provable Statistical Guarantees," *arXiv preprint arXiv:1511.08551*, 2014
- D. Hsu, S.M. Kakade, T. Zhang, "A spectral algorithm for learning hidden Markov models," *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1460–1480, 2012.
- A. Kontorovich, B. Nadler and R. Weiss, "On learning parametric-output HMMs," *arXiv preprint arXiv:1302.6009*, 2013
- P. Bickel, Y. Ritov and T. Ryden, "Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models," *The Annals of Statistics*, vol. 26, no. 6, pp. 1614–1635, 1998
- B. Yu, "Rates of convergence for empirical processes of stationary mixing sequences," *The Annals of Probability*, pp. 94-116, 1994

Main result and proof ingredients

Regularity conditions on $Q = \mathbb{E}Q_n$:

- "Lipschitz" condition on gradients:
  $\sup_\theta \|\nabla_\theta Q(\theta \mid \theta') - \nabla_\theta Q(\theta \mid \theta^*)\|_2 \leq L\|\theta' - \theta^*\|_2$
- Strong concavity of Q

Specific to HMM: A mixing Markov Chain with mixing coefficient ρ_{mix} s.t. $\sup_{i,j} |P(Z_k = i | Z_0 = j) - \pi(i)| \leq c\rho_{\text{mix}}^k$ for all k.

Theorem [Y., Balakrishnan and Wainwright '15]
If the regularity conditions hold for $\theta' \in B(r; \theta^*)$ with $\frac{L}{\lambda} < 1$, then

$$\|\widehat{\theta}^t - \theta^*\|_D \leq \underbrace{\kappa^t \|\widehat{\theta}^0 - \theta^*\|_D}_{\text{linear convergence}} + \frac{1}{1-\kappa}\Big(\underbrace{\phi(n, \delta, k)}_{\text{truncation error}} + \underbrace{\epsilon(n, \delta, k)}_{\text{finite sample error}} \Big)$$

with probability $1 - \delta$ and $\kappa = \frac{L}{\lambda}$.

Proof sketch Based on framework in [Balakrishnan et al. '14] for i.i.d. data with the following additional steps:

1. In Q_n replace $\mathbb{E}_{Z_i^{i+1} | X_1^n}$ by $\mathbb{E}_{Z_i^{i+1} | X_{i-k}^{i+k}}$ to obtain truncation error $\phi(n, \delta, k)$

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

2. Treat far away X_{i-k}^{i+k} as independent (see e.g. [Yu '94]), use empirical process theory to get $\epsilon(n, \delta, k)$; bound difference using mixing

Note: Dependence makes regularity conditions much harder to check

Simple example and simulations

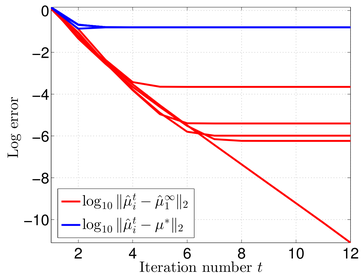
Consider $Z_i \in \{-1, +1\}$ a mixing Markov chain with symmetric transition matrix. The observation densities are $X_i \mid Z_i \sim \mathcal{N}(Z_i\mu^*, \sigma^2 I)$.

Result in terms of $\eta^2 = \frac{\|\mu^*\|_2^2}{\sigma^2}$ and $\kappa := \frac{L}{\lambda} \propto \mathrm{e}^{-c\eta^2}$ reads

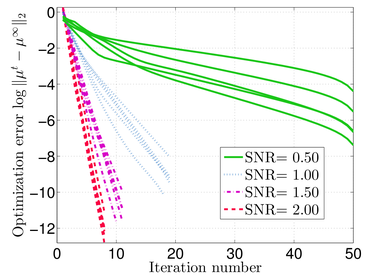
$$\|\widehat{\theta}^t - \theta^*\|_D \leq \kappa^t \|\widehat{\theta}^0 - \theta^*\|_D + \frac{C(\rho_{\text{mix}})}{1-\kappa}\sqrt{\frac{d \log^4 \frac{n}{\delta}}{n}}$$

w.p. $1 - \delta$ if $n \gtrsim d \log^2(d/\delta)$ and the initialization $\hat{\mu}^0 \in \mathbb{B}_2(\frac{\|\mu^*\|_2}{4}, \mu^*)$.

Simulations
Parameters: $d = 10, n = 1000, \sigma = 2, \rho_{\text{mix}} = 0.6$



Convergence to different local minima using different initializations



Rate dependence on SNR