Convergence rate of EM Algorithm on Gaussian Mixture Models

settings : under-specified , exactly-specified , over-specified. ✓

over-specified : the Gaussian Mixture model we want to fit has more mixture components than the true mixture model generating the data.

It's most common to use over-specified mixture models.. So we want to understand how EM Algo. behaves in over-specified mixture model, especially on Gaussian mixture. (Others, like mixture of regressions are discussed in other places).

Problem set-up:

denote $\phi(\cdot; \mu, \Sigma)$ multivariate Gaussian distribution.

To make things easier, we focus on Gaussian Mixture with 2 components.

Say, true model is:

$$f(x; \theta^*, \sigma, \pi) := \pi \cdot \phi(x; \theta^*, \sigma^2 I_d) + (1-\pi) \cdot \phi(x; -\theta^*, \sigma^2 I_d) \quad (1)$$

fixed, known: π, σ, d.

Now we use EM to fit model

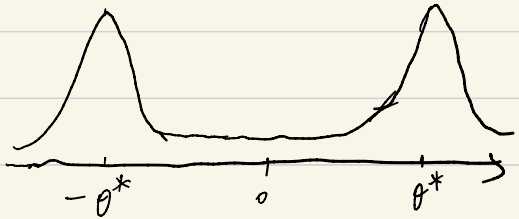
$$f(x, \theta, \sigma, \pi) := \pi \cdot \phi(x; \theta, \sigma^2 I_d) + (1-\pi) \cdot \phi(x; -\theta, \sigma^2 I_d) \quad (2),$$

and use solution $\hat{\theta}_n$ to estimate θ^*

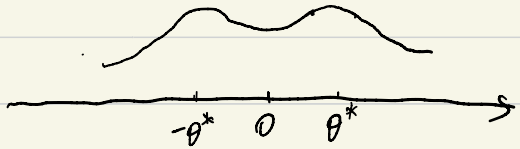
"singal strength" : $\frac{\|\theta^*\|}{\sigma}$, i.e., the separation between the means of mixture components relative to the spread in the components.

e.g. :

strong signal : $\frac{\|\theta^*\|}{\sigma}$ ↑



weak signal : $\frac{\|\theta^*\|}{\sigma}$ ↓.



Particularly interested in the limit of weak signal case when there is no separation, i.e., $\|\theta^*\|_2 = 0$.

Now the true model becomes a single Gaussian, $f(x) = \phi(x; 0, \sigma^2 I_d)$.

we are then in the setting of over-specified.

Why we interest in such limit weak signal, $\theta^* = 0$? See some simulations.

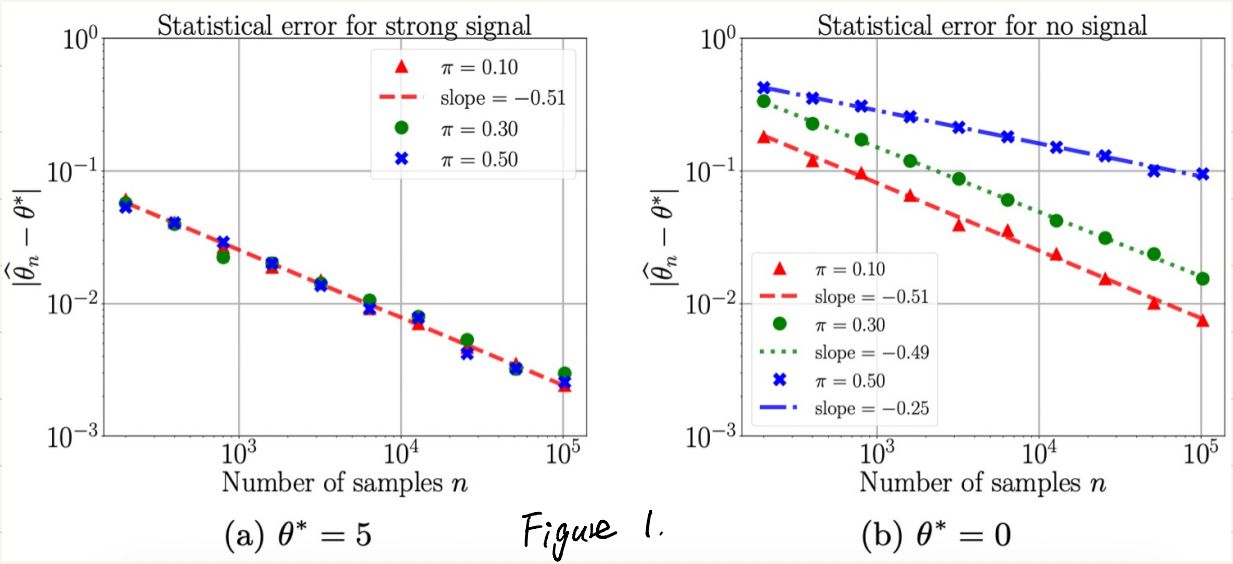


Figure 1.

When true is $f(x) = \phi(x; 0, \sigma^2 I_d)$., but we fit $\pi\,\phi(x; \theta, \sigma^2 I_d) + (1-\pi)\,\phi(x; -\theta, \sigma^2 I_d)$, different weights, error decays in different rate.

And with $\pi \uparrow 0.5$, i.e., more balanced, error decays more slowly.

In the over-specified setting with $\theta^* = 0$,
weights affect EM convergence rate greatly..

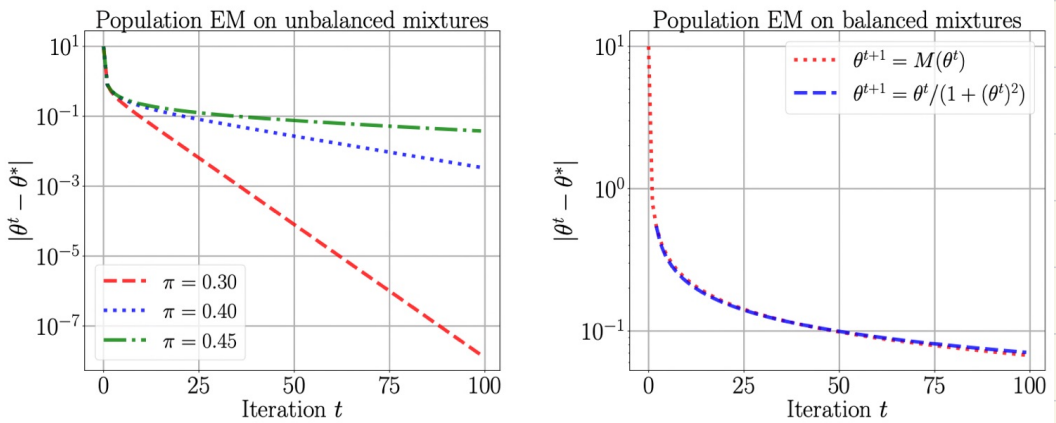


Figure 2.

(a): unbalanced, $\pi \neq \frac{1}{2}$: EM converges geometrically quickly, convergence rate $\downarrow 0$ as $\pi \uparrow \frac{1}{2}$.

(b) balanced $\pi = \frac{1}{2}$: EM convergence rate is sub-geometric, much slower.

Schema:

While true data-generating model is a single Gaussian.

We analyze EM convergence rate of Gaussian Mixture with 2 components.

- unbalanced weights
  - population-based : Thm 1.(a)
  - sample-based : Thm 1.(b).
- balanced weights
  - population-based : Thm 2
  - sample-based : Thm 3 + Thm 4.

Population EM Algo. is the idealized version of EM based on an infinite sample size, the usual sample-based EM Algo is used in practice.

For theoretical purposes, we first analyze the convergence of the population EM updates, and then leverage these findings to understand the behavior of sample EM.

Population EM updates (§ 2.3).

true model. $N(0, \sigma^2 I_d)$. $\theta^* = 0$.

we try to fit $\pi \phi(x; \theta, \sigma^2 I_d) + (1-\pi)\cdot \phi(x; -\theta, \sigma^2 I_d)$. (3).

Latent variable Z: $P(Z=1) = \pi$. $P(Z=0) = 1-\pi$. $Z \sim \text{Bernoulli}(\pi)$.

joint (X, Z). $X|(Z=1) \sim \phi(x; \theta, \sigma^2 I_d)$ $X|(Z=0) \sim \phi(x; -\theta, \sigma^2 I_d)$.

complete data: $(x_1, z_1), (x_2, z_2), \cdots, (x_n, z_n)$., but we only have $x_1, \cdots, x_n$.

population complete log-likelihood:

$$\ell(\theta' | X, Z) = \mathbb{E}_X [\log P(X, Z | \theta')]$$, where the expectation is taken over the true distribution $X \sim N(0, \sigma^2 I_d)$.

$$= \mathbb{E}_X \left\{ \log \left[\left(\pi \cdot \phi(\theta', \sigma^2 I_d) \right)^Z \left((1-\pi) \phi(-\theta', \sigma^2 I_d) \right)^{1-Z} \right] \right\}$$

$$= E_X\left\{ z \cdot \log\left(\pi \cdot \phi\left(\theta', \sigma^2 I_d\right)\right) + (1-z)\cdot \log\left((1-\pi)\cdot \phi\left(-\theta', \sigma^2 I_d\right)\right)\right\}$$

Given θ,

Take expectation of $\mathcal{L}(\theta' | X, z)$ w.r.t. the posterior distribution of z, i.e. $z|X,\theta$, we get $Q(\theta', \theta)$., the lower bound of $\mathcal{L}(\theta'|X,z)$.

$$\begin{aligned} Q(\theta';\theta) &= E_{z|X,\theta}\, \mathcal{L}(\theta|X,z) \\ &= E_{z|X,\theta}\left\{ E_X\left[\log p(X,z|\theta')\right]\right\} \\ &= E_X\left\{ E_{z|X,\theta}\left[\log p(X,z|\theta')\right]\right\} \\ &= E_X\left\{ E_{z|X,\theta}\left[z \log\left(\pi\cdot\phi\left(\theta', \sigma^2 I_d\right)\right) + (1-z)\cdot\log(1-\pi)\cdot\phi\left(-\theta', \sigma^2 I_d\right)\right]\right\}, \end{aligned}$$

denote $w_\theta(x) = E_{z|x,\theta}(z) = p(z=1|x,\theta) = \dfrac{\pi \cdot \exp\left(\frac{-\|\theta-x\|_2^2}{2\sigma^2}\right)}{\pi \cdot \exp\left(\frac{-\|\theta-x\|_2^2}{2\sigma^2}\right) + (1-\pi)\cdot \exp\left(\frac{-\|\theta+x\|_2^2}{2\sigma^2}\right)}$

$$Q(\theta';\theta) = E_x\left[w_\theta(x)\left(\log\pi + \log\phi(\theta', \sigma^2 I_d)\right) + (1-w_\theta(x))\log(1-\pi) + (1-w_\theta(x))\cdot\log\phi(-\theta', \sigma^2 I_d)\right]$$

remove terms unrelated to θ'

$$= -\frac{1}{2} E_x\left[w_\theta(x)\cdot\|x-\theta'\|_2^2 + (1-w_\theta(x))\cdot\|x+\theta'\|_2^2\right]. \quad (7)$$

$$\nabla_{\theta'} Q(\theta';\theta) = -\frac{1}{2} E_x\left[w_\theta(x)(-x-x+2\theta') + (1-w_\theta(x))(x+x+2\theta')\right]$$

$$= E_x\left[(2w_\theta(x)-1)x\right] - \theta' = 0.$$

so population EM operator:

$$M(\theta) = \arg\max_{\theta' \in \mathbb{R}^d} Q(\theta';\theta) = E_x\left[(2w_\theta(x)-1)x\right]. \quad (8)$$

population EM Algo : $\theta^{t+1} = M(\theta^t) = E_X\left[\left(2 w_{\theta^t}(X) - 1\right) X\right]$, $t = 0, 1, 2, \cdots$.

Analogously,

sample-based EM Algo : $\theta^{t+1} = M_n(\theta^t) = \frac{1}{n} \sum_{i=1}^{n} \left[\left(2 w_{\theta^t}(X_i) - 1\right) X_i\right]$, $t = 0, 1, 2, \cdots$. (9).

given the observed data $\{X_i\}_{i=1}^n$

Main results.

§ 3.1. Behavior of EM for unbalance mixtures.

mixture model $\pi \cdot \phi(x; \theta, \sigma^2 I_d) + (1-\pi) \cdot \phi(x; -\theta, \sigma^2 I_d)$ (3).

Thm 1: We fit an unbalanced $(\pi \neq \frac{1}{2})$ of mixture model (3) to $N(0, \sigma^2 I_d)$. $\theta^* = 0$.

Then

(a). The population EM operator (8), $M(\theta) = E_X[(2w_\theta(X) - 1)X]$, is globally strictly contractive, meaning that

population EM. (10a) $\|M(\theta)\|_2 \leq (1 - \frac{\rho^2}{2}) \|\theta\|_2$ for all $\theta \in \mathbb{R}^d$.

with $\rho := |1 - 2\pi| \in (0,1)$.

$1 - \frac{\rho^2}{2} \in (\frac{1}{2}, 1)$

(b) $\exists$ constants c, c' s.t. for $\forall \delta \in (0,1)$, and sample size $n \geq c \cdot \frac{\sigma^2}{\rho^4}\left(d + \log\frac{1}{\delta}\right)$, the sample EM updates $\theta^{t+1} = M_n(\theta^t)$ satisfies the upper bound.

sample EM.

(10b) $$\|\theta^t\|_2 \leq \|\theta^0\|_2 \left(1 - \frac{\rho^2}{2}\right)^t + \frac{c'\left(\|\theta^0\|_2 \sigma^2 + \rho\sigma\right)}{\rho^2}\sqrt{\frac{d + \log\left(\frac{1}{\delta}\right)}{n}}.$$

with probability at least $1-\delta$.

By Thm1 (a).

$$\|\theta^1\|_2 = \|M(\theta^0)\|_2 \leq \left(1 - \frac{\rho^2}{2}\right)\cdot\|\theta^0\|_2$$

$$\|\theta^2\|_2 \leq \left(1 - \frac{\rho^2}{2}\right)\cdot\|\theta^1\|_2 \leq \left(1 - \frac{\rho^2}{2}\right)^2\cdot\|\theta^0\|_2$$

$$\vdots$$

$$\|\theta^T\|_2 \leq \left(1 - \frac{\rho^2}{2}\right)^T\|\theta^0\|_2$$

we want $\|\theta^T\|_2 \le \left(1-\frac{\rho^2}{2}\right)^T \|\theta^0\|_2 \le \varepsilon \quad \forall \varepsilon > 0$. since $\theta^* = 0$.

$$\Rightarrow \left(1-\frac{\rho^2}{2}\right)^T \le \frac{\varepsilon}{\|\theta^0\|_2}$$

$$\Rightarrow \quad T \ge \frac{1}{\log \frac{1}{\left(1-\frac{\rho^2}{2}\right)}} \cdot \log\left(\frac{\|\theta^0\|_2}{\varepsilon}\right)$$

so $\|\theta^T\|_2 \le \varepsilon$ for $T \ge \frac{1}{\log \frac{1}{\left(1-\frac{\rho^2}{2}\right)}} \cdot \log\left(\frac{\|\theta^0\|_2}{\varepsilon}\right)$

It converges in $O\left(\log \frac{1}{\varepsilon}\right)$ steps to an ε-ball around $\theta^* = 0$.

Theoretically, the updates converge at a geometric rate to the true parameter $\theta^* = 0$.

Proof of Thm 1 (a):

$$w_\theta(x) = \frac{\pi \cdot \exp\left(-\frac{\|\theta - x\|_2^2}{2\sigma^2}\right)}{\pi \cdot \exp\left(-\frac{\|\theta - x\|_2^2}{2\sigma^2}\right) + (1-\pi)\cdot \exp\left(-\frac{\|\theta + x\|^2}{2\sigma^2}\right)} = \frac{\pi}{\pi + (1-\pi)\cdot \exp\left(\frac{-2\theta^T x}{\sigma^2}\right)}$$

$$\nabla_\theta w_\theta(x) = \frac{\frac{2\pi(1-\pi)x}{\sigma^2}}{\left[\pi \cdot e^{\frac{\theta^T x}{\sigma^2}} + (1-\pi)\cdot e^{-\frac{\theta^T x}{\sigma^2}}\right]^2}$$

For a scalar $u \in [0,1]$, define function $h(u) = w_{u\theta}(x) = \frac{\pi}{\pi + (1-\pi)\cdot e^{-\frac{2u\theta^T x}{\sigma^2}}}$

$$h'(u) = \frac{2\pi(1-\pi)\cdot\frac{\theta^T x}{\sigma^2}}{\left[\pi \cdot e^{\frac{u\theta^T x}{\sigma^2}} + (1-\pi)\cdot e^{-\frac{u\theta^T x}{\sigma^2}}\right]^2}$$

$$= \nabla w_{u\theta}(x)^T \cdot \theta$$

$\theta_u = u\theta$, $u \in [0,1]$.

By Taylor expansion,

$$\|h(\theta)\|_2 = \|E_x[2X(w_0(X) - w_\theta(X))]\|_2 \qquad w_\theta(X) = w_0(X) + \int_0^1 h'(u)\,du$$

$$= \|E_x[2X \int_0^1 h'(u)\,du]\|_2 \quad \text{(Taylor expansion)}$$

$$= 4\pi(1-\pi) \cdot \left\| \int_0^1 E_x \left[\underbrace{\frac{XX^T}{\sigma^2\left(\pi \cdot \exp\left(\frac{\theta_u^T X}{\sigma^2}\right) + (1-\pi)\exp\left(-\frac{\theta_u^T X}{\sigma^2}\right)\right)^2}}_{\text{matrix } \Gamma_{\theta_u}(X)} \right] \theta\, du \right\|_2$$

$$= 4\pi(1-\pi) \cdot \left\| \theta \cdot \int_0^1 E_x[\Gamma_{\theta_u}(X)]\,du \right\|_2$$

by submultiplicativity property.

$$\leq 4\pi(1-\pi) \cdot \|\theta\|_2 \cdot \left\| \int_0^1 E_x[\Gamma_{\theta_u}(X)]\,du \right\|_2$$

ℓ_2-norm of matrix

$$= 4\pi(1-\pi)\,\|\theta\|_2 \cdot \left\| \int_0^1 E_x[\Gamma_{\theta_u}(X)]\,du \right\|_{op}.$$

$\|\cdot\|_{op}$ is the largest singular value of a matrix.

$$\le 4\pi(1-\pi)\,\|\theta\|_2 \cdot \max_{u\in[0,1]} \left\| E_x[\Gamma_{\theta_u}(X)] \right\|_{op}$$

$$\le 4\pi(1-\pi)\,\|\theta\|_2 \cdot \frac{1-\rho^2/2}{1-\rho^2}, \qquad \pi = \tfrac{1}{2}(1-\rho)$$

$$= \left(1-\frac{\rho^2}{2}\right)\cdot \|\theta\|_2.$$

Also, we need to prove $\max_{u\in[0,1]} \left\| E[\Gamma_{\theta_u}(X)] \right\|_{op} \le \frac{1-\rho^2/2}{1-\rho^2}$.

Thm 1 (b). $\exists$ constants c, c' s.t. for $\forall \delta \in (0,1)$ and a sample size $n \geq c \cdot \frac{\sigma^2}{\rho^4}\left(d + \log\frac{1}{\delta}\right)$,
the sample EM updates $\theta^{t+1} = M_n(\theta^t)$ satisfied the upper bound

sample EM.

(10b) $$\|\theta^t\|_2 \leq \|\theta^0\|_2 \left(1 - \frac{\rho^2}{2}\right)^t + \frac{c'\left(\|\theta^0\|_2 \sigma^2 + \rho\sigma\right)}{\rho^2} \cdot \sqrt{\frac{d + \log(\frac{1}{\delta})}{n}},$$

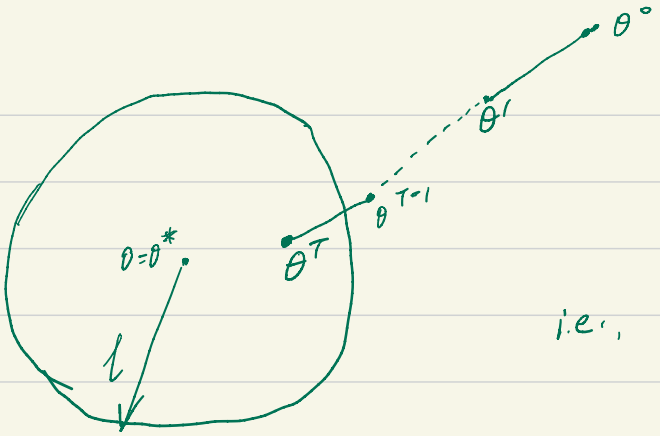
with probability at least $1-\delta$.

Intuition:

Our ultimate interest is in the behavior of the finite-sample EM Also, it is practical, population EM is an idealization.

The finite-sample EM updates are based on $M_n(\theta)$.
population EM updates are based on $M(\theta)$.

so important: the empirical process $\left\{ \|M_n(\theta) - M(\theta)\|_2 : \|\theta - 0\|_2 \leq r \right\}$ ($0 \to \theta^* = 0$).



Let ℓ be an upper bound on the supremum of this empirical process that holds with probability at least $1-\delta$.

i.e., $$\mathbb{P}\left[\sup_{\|\theta-\theta\|\le r} \left\| M_n(\theta) - M(\theta) \right\| \le \ell \right] \ge 1-\delta.$$

So we want to show that the sample-EM iterates converge to a near-optimal solution, i.e., a point whose distance from $\theta^* = 0$ is at most ℓ, or a constant multiple of ℓ.

Proof of Thm 1(b):

By Lemma 1, for $\forall$ radius r, threshold $\delta \in (0,1)$, large enough n, $\exists$ constants c, c'.

we have $P\left[\sup_{\|\theta\|_2 \le r} \|M_n(\theta) - M(\theta)\|_2 \le \frac{c'}{2}\left(6^2 r + \rho\sigma\right)\sqrt{\frac{d + \log(1/\delta)}{n}}\right] \ge 1 - \delta$.

as long as $n \ge c \cdot \frac{\sigma^2}{\rho^4}\left(d + \log\frac{1}{\delta}\right)$.

- $\|\theta' - \theta^*\|_2 = \|M_n(\theta^*) - \theta^*\|_2 = \|M(\theta^*) + M_n(\theta^*) - M(\theta^*)\|_2$

  triangle inequality $\le \|M(\theta^*)\|_2 + \|M_n(\theta^*) - M(\theta^*)\|_2$

  $\le \left(1 - \frac{\rho^2}{2}\right)\|\theta^*\|_2 + \frac{c'}{2}\left(6^2 r + \rho\sigma\right)\sqrt{\frac{d + \log(\frac{1}{\delta})}{n}}$.

  by Thm 1 (a).

By Mathematical induction

$$\|\theta^t - 0\|_2 \le \left(1-\frac{\rho^2}{2}\right)\cdot\|\theta^{t-1}-0\|_2 + \underbrace{\frac{c'}{2}(G^2 r+\rho G)\sqrt{\frac{d+\log(\frac{1}{\delta})}{n}}}_{l}$$

- By iteration, it shows that.

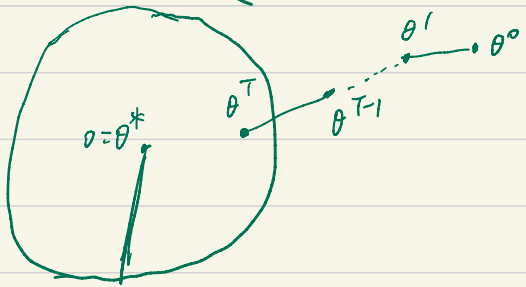
$$\begin{aligned}
\|\theta^t\|_2 &\le \left(1-\frac{\rho^2}{2}\right)\cdot\|\theta^{t-1}\|_2 + l \\
&\le \left(1-\frac{\rho^2}{2}\right)\left[\left(1-\frac{\rho^2}{2}\right)\cdot\|\theta^{t-2}\|_2 + l\right] + l \\
&\le \left(1-\frac{\rho^2}{2}\right)^t\cdot\|\theta^0\|_2 + \left[\sum_{i=0}^{t-1}\left(1-\frac{\rho^2}{2}\right)^i\right]\cdot l \\
&\le \left(1-\frac{\rho^2}{2}\right)^t\cdot\|\theta^0\|_2 + \frac{1}{1-\left(1-\frac{\rho^2}{2}\right)}\cdot l \\
&= \left(1-\frac{\rho^2}{2}\right)^t\cdot\|\theta^0\|_2 + \frac{c'\cdot(G^2 r+\rho G)}{\rho^2}\cdot\sqrt{\frac{d+\log(\frac{1}{\delta})}{n}}
\end{aligned}$$

Since it is for $\forall\, r>0$, we can let $r = \|\theta^0 - 0\|_2$.

So. (10b)

$$\|\theta^t\|_2 \le \|\theta^0\|_2 \left(1-\frac{\rho^2}{2}\right)^t + \frac{c'\left(\|\theta^0\|_2\sigma^2+\rho\sigma\right)}{\rho^2}\sqrt{\frac{d+\log\left(\frac{1}{\delta}\right)}{n}}$$ with prob. at least $1-\delta$.

- Note that bound (10b) has 2 terms., the first decreases geometrically in iteration, the second is independent of t.
  Thus it guarantees that the iterations converge geometrically to a ball of radius $O\left(\frac{c'}{\rho}\left(\|\theta^0\|_2\sigma^2+\rho\sigma\right)\sqrt{\frac{d+\log\left(\frac{1}{\delta}\right)}{n}}\right)$.



- For a fixed sample size n, the bound

(10b) $$\|\theta^t\|_2 \le \|\theta^0\|_2 \left(1-\frac{\rho^2}{2}\right)^t + \underbrace{\frac{c'(\|\theta^0\|_2 \cdot \sigma^2 + \rho\cdot\sigma)}{\rho^2}}_{\alpha} \cdot \sqrt{\frac{d+\log(\frac{t}{\delta})}{n}}.$$

This bound gives a reasonable choice of iteration number T.

For iteration number T s.t. $T \ge \log_{1-\frac{\rho^2}{2}} \frac{\alpha}{\|\theta^0\|_2}$,

$$\left(1-\frac{\rho^2}{2}\right)^T \le \left(1-\frac{\rho^2}{2}\right)^{\log_{1-\frac{\rho^2}{2}} \frac{\alpha}{\|\theta^0\|_2}} = \frac{\alpha}{\|\theta^0\|_2}.$$

so $$\|\theta^T\|_2 \le \|\theta^0\|_2 \left(1-\frac{\rho^2}{2}\right)^T + \alpha = \|\theta^0\|_2 \cdot \frac{\alpha}{\|\theta^0\|_2} + \alpha = 2\alpha.$$

This T-choice guarantees that the first term in (10b) is dominated by the second term, so

$$\|\theta^T\|_2 \le 2\alpha, \text{ with prob. at least } 1-\delta.$$

So sample-EM converges in $O\left(\log\left(\frac{n}{d}\right)\right)$ steps to a ball of radius $\left(\frac{d}{n}\right)^{\frac{1}{2}}$.
It is also a very fast convergence.

— We see fast convergence of EM in unbalanced fit, both sample and population.

Balanced fit. $\pi = \frac{1}{2}$.

EM is worse in balanced case.

Thm 2: convergence rate of population EM update for balance fit.

$\pi = \frac{1}{2}$. we fit model (3). $\frac{1}{2}\phi(x;\theta,\sigma^2 I_d) + \frac{1}{2}\phi(x;-\theta,\sigma^2 I_d)$ to $N(0,\sigma^2 I_d)$.

Then population EM operator

$$M(\theta) = E_X\left[(2w_\theta(X)-1)X\right]$$ satisfies:

(a). for all $\theta \neq 0$, we have

$$(13a)\quad \frac{\|M(\theta)\|}{\|\theta\|_2} \le \gamma_{up}(\theta) := 1-p+\frac{p}{1+\frac{\|\theta\|_2^2}{2\sigma^2}} < 1.$$

$p + \frac{1-p}{1+\frac{\|\theta\|_2^2}{2\sigma^2}}$.

p is a constant.

$p := P(|X|\le 1) + \frac{1}{2}P(|X|>1)$

where $X \sim N(0,1)$.

(b) for all $\theta \neq 0$ s.t. $\|\theta\|_2^2 \leq \frac{56^2}{8}$, we have

(13b) $$\frac{\|M(\theta)\|_2}{\|\theta\|_2} \geq \gamma_{Low}(\theta) := \frac{1}{1 + \frac{2\|\theta\|_2^2}{6^2}}.$$

The contraction parameter is changing with $\|\theta\|$.

(13a): $\|\theta\|_2 \downarrow 0$, $\gamma_{up}(\theta) = p + \frac{1-p}{1+\frac{\|\theta\|_2^2}{26^2}} \uparrow 1$, so convergence rate slows down.

(13b) after $\|\theta\|_2 \leq \frac{56^2}{8}$, contration parameter lower bound $\gamma_{Low}(\theta) \geq \frac{4}{9}$. and $\gamma_{Low}(\theta) \uparrow 1$.

In balanced setting,
population EM still converges to the true parameter $\theta^* = 0$. from arbitrary initialization, but the convergence gets exponentially slower when

the iteration close to the true θ^*.

It is guaranteed to have $\|\theta^T\|_2 \le \sqrt{2}\sigma$ after running $T := \frac{\log\left(\frac{\|\theta^0\|_2^2}{2\sigma^2}\right)}{\log\left(\frac{2}{2-p}\right)}$ iterations.

After $\|\theta^T\|_2 \le \sqrt{2}\sigma$,

the convergence rate becomes sub-geometric,

for $\varepsilon \in (0, \sqrt{2}\sigma)$, $\|\theta^{T_0+t}\|_2 \le \varepsilon$ for $t \ge \frac{c\sigma^2}{\varepsilon^2}\log\left(\frac{\sigma}{\varepsilon}\right)$.