Homeomorphic-Invariance of EM: Non-Asymptotic Convergence in KL Divergence for Exponential Families via Mirror Descent

Frederik Kunstner
University of British Columbia

Raunak Kumar
Cornell University

Mark Schmidt
University of British Columbia
Canada CIFAR AI Chair (Amii)

Recently awarded AI/Stats 2021 Best Paper Prize

Learning with Missing Values

Missing values are very common in real datasets.

Models/data often have unobserved/hidden/latent values.

For example, we may want to fit a Gaussian to a dataset like this:

$$X = \begin{bmatrix} 0.3 & ? & 5 & -1 \\ -0.2 & 10 & 1 & +1 \\ 0.1 & ? & 2 & -1 \\ 0.1 & 22 & 0 & ? \end{bmatrix}.$$

One of the most common algorithms for this setting is EM.

"Expectation maximization".

Applies when problem is "easy" to solve with no missing values.

Uses probabilistic "soft" assignments to missing variables.

For many problems it leads to simple closed-form updates.

Expectation Maximization: Optimization with MAR Variables

EM was independently invented for a variety of different problems.

Paper giving general form is among most-cited across all fields:

Maximum Likelihood from Incomplete Data Via the *EM* Algorithm
AP Dempster, NM Laird, DB Rubin - Journal of the Royal Statistical Society, 1977
Cited by 63685 Related articles All 72 versions

Some common applications:

- Mixture of Gaussians.
- Multivariate student t.
- Hidden Markov models.
- Factor analysis.
- Semi-supervised learning.
- Graphical models with missing data.

In many problems, we introduce missing variables to use EM.

MLE from Incomplete Data via the EM Algorithm (Exponential Families)

Maximum likelihood with observed data x and missing z:

$$\mathcal{L}(\theta) = -\log p(x \mid \theta) = -\log \overbrace{\int p(x, z \mid \theta)\, \mathrm{d}z}^{\text{average over missing data}}$$

Most classic EM applications have complete data in **exponential family**,

$$p(x, z \mid \theta) \propto \exp(\langle T(x, z), \theta \rangle - A(\theta))$$

E-step: Compute the expected sufficient statistics

$$\bar{\mu}_t = \mathbb{E}_{z \sim p(z \mid x, \theta_t)}[T(x, z)]$$

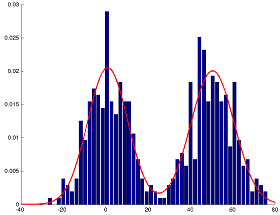
M-step: Maximum likelihood/Moment matching

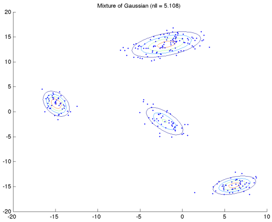
$$\text{Find } \theta_{t+1} \text{ such that } \mathbb{E}_{x,z \sim p(x,z \mid \theta_{t+1})}[T(x, z)] = \bar{\mu}_t$$

Increases likelihood, parameterization invariant, converges to stationary*.

Example: Mixture of Gaussians

Application: modeling multi-modal data with mixture of Gaussians





We introduce missing variable for each sample ("which Gaussian?").

Yields an intuitive EM update:

E-step: compute pr("example comes from each Gaussian").

M-step: update cluster parameters using examples "in" cluster.

Convergence Rate of EM

Is EM a good optimization algorithm?

How fast does it converge?

Previous results:

Asymptotically, EM has linear convergence rate.

No dependence on parameterization.

Instead depends on "amount of missing information".

But we may never reach asymptotic regime.

Non-asymptotically, "at least as a fast as gradient descent".

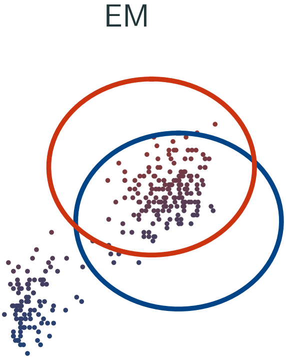
Dependent on parameterization.

Misses dependence on "amount of missing information".

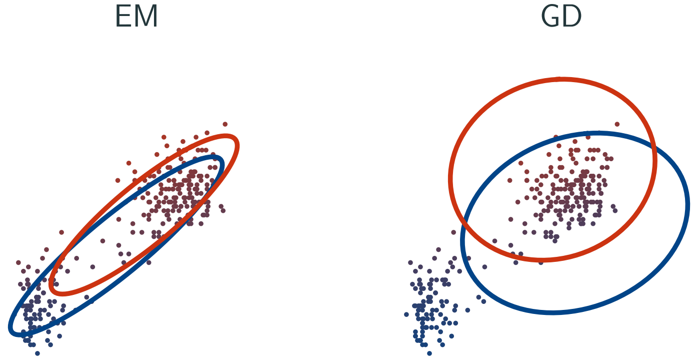
In practice EM is faster than gradient descent.

This work: parameterization-invariant non-asymptotic EM analysis.

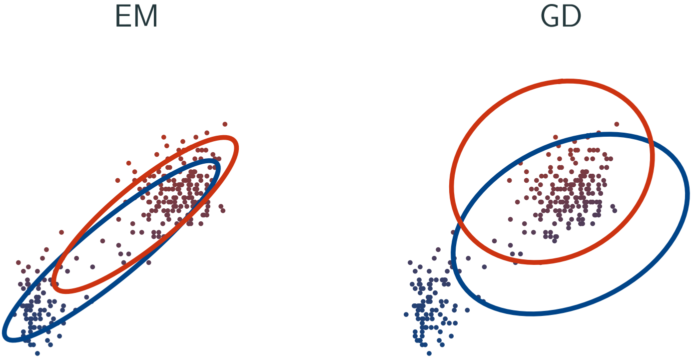
“EM is at least as fast as GD”



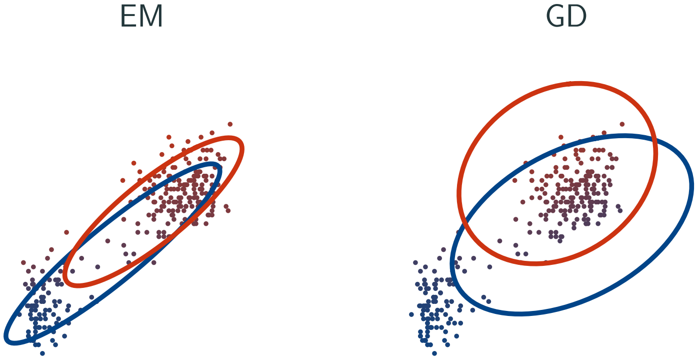
“EM is at least as fast as GD”



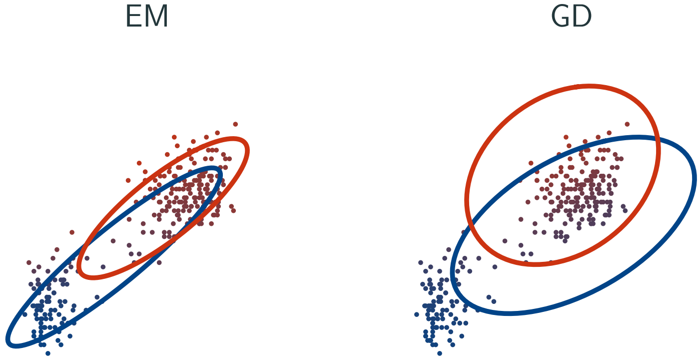
“EM is at least as fast as GD”



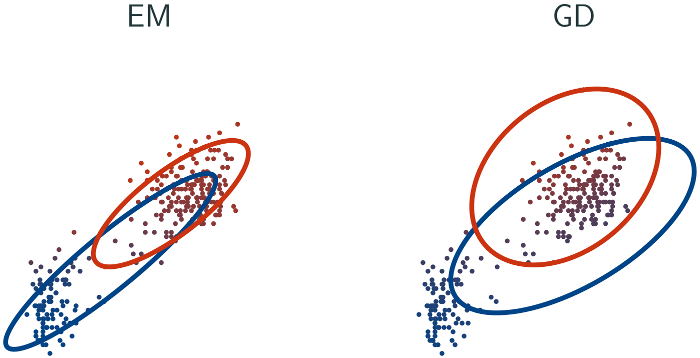
“EM is at least as fast as GD”



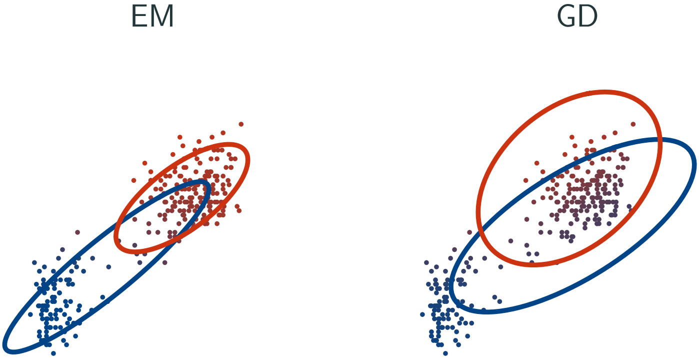
“EM is at least as fast as GD”



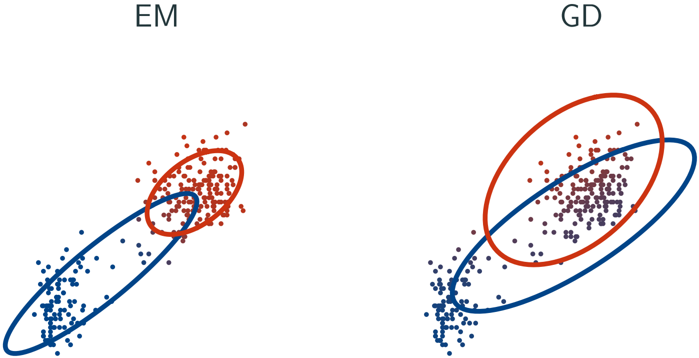
“EM is at least as fast as GD”



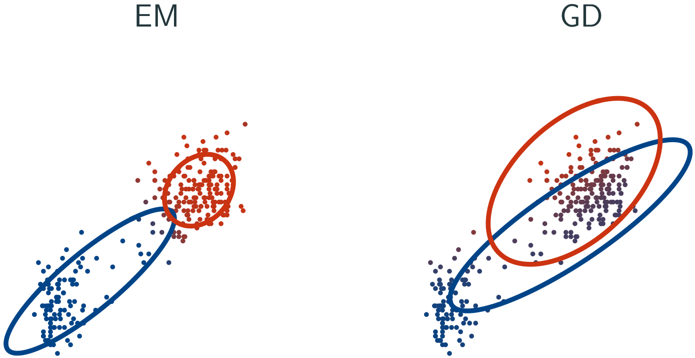
“EM is at least as fast as GD”



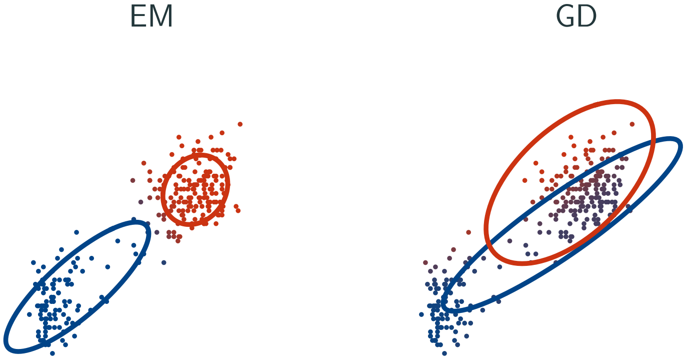
“EM is at least as fast as GD”



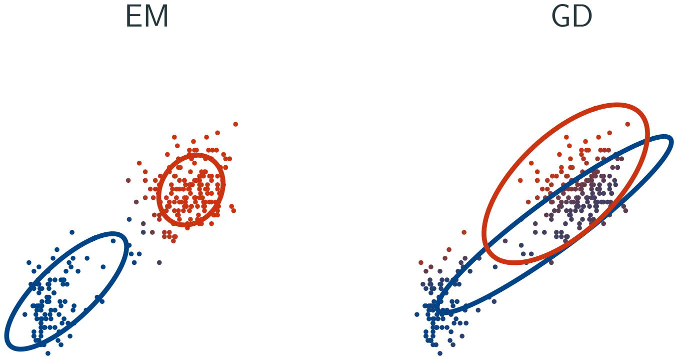
“EM is at least as fast as GD”



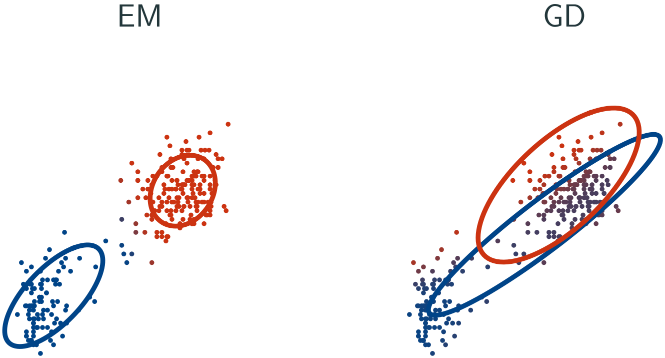
"EM is at least as fast as GD"



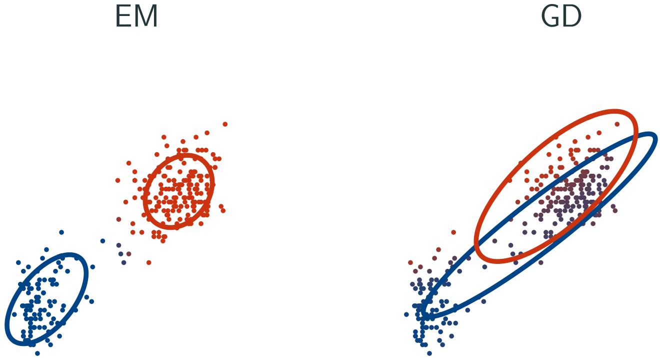
“EM is at least as fast as GD”



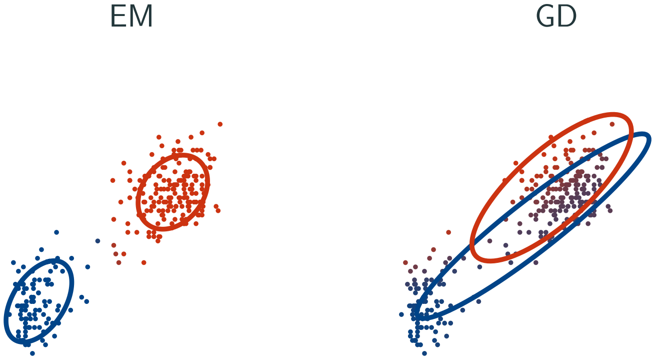
“EM is at least as fast as GD”



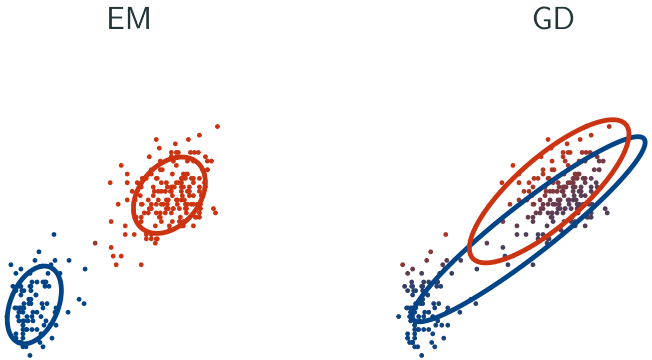
“EM is at least as fast as GD”



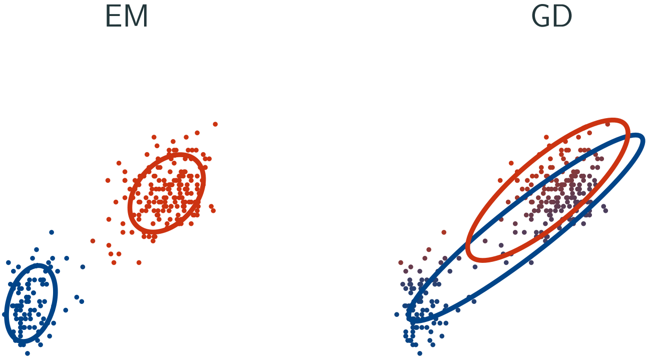
“EM is at least as fast as GD”



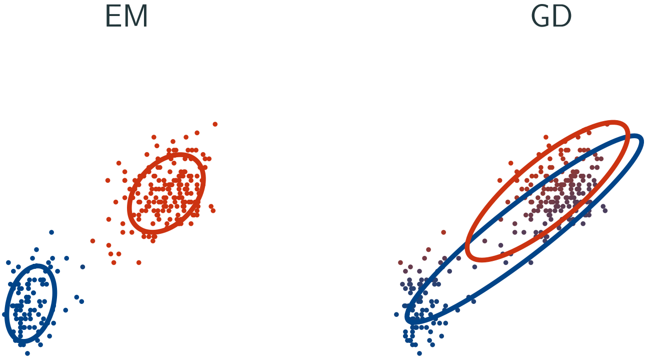
“EM is at least as fast as GD”



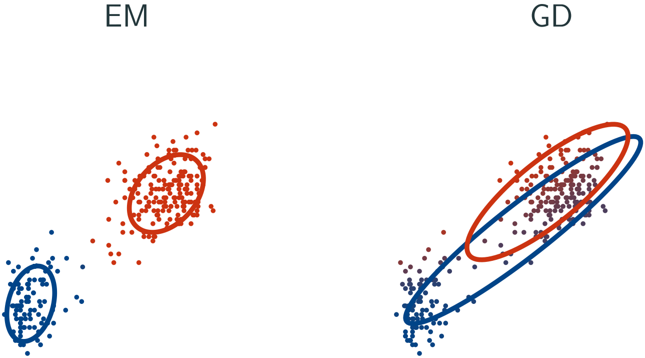
“EM is at least as fast as GD”



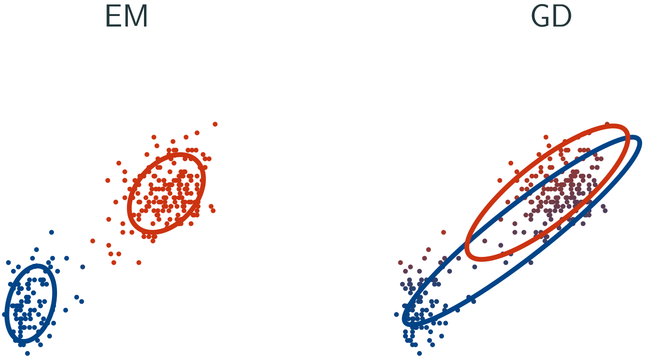
“EM is at least as fast as GD”



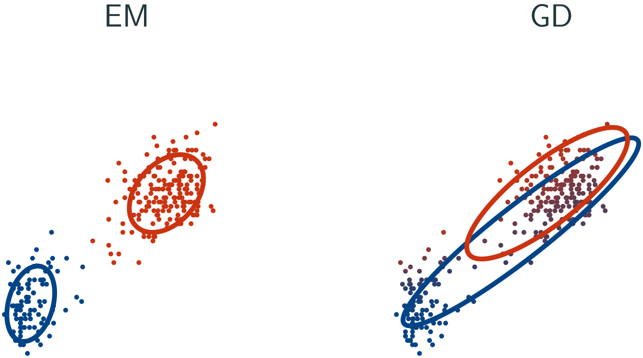
“EM is at least as fast as GD”



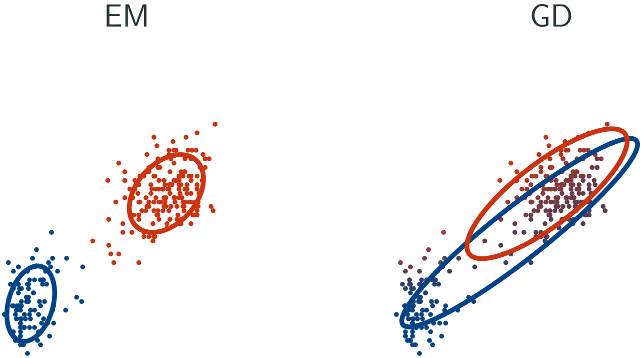
"EM is at least as fast as GD"



"EM is at least as fast as GD"

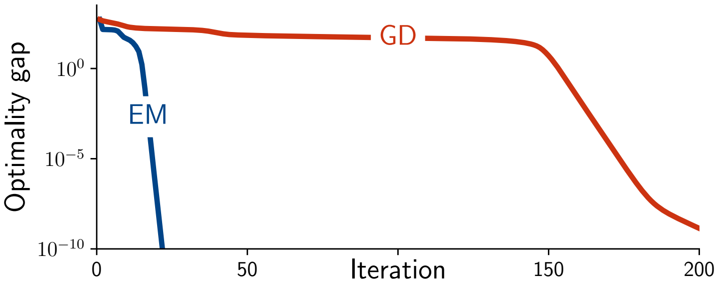


“EM is at least as fast as GD”



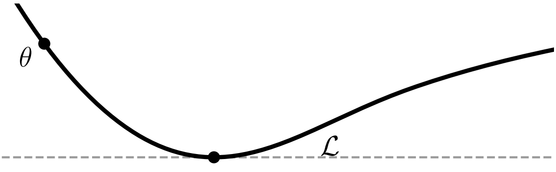
"EM is at least as fast as GD"

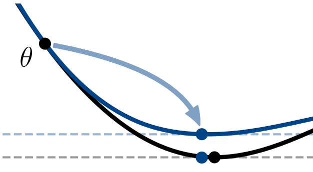
EM GD



EM is much faster. What are we missing?

Previous work





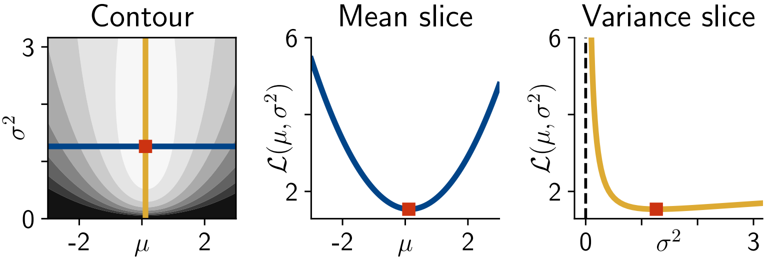
$$\mathcal{L}(\theta) \leq \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta - \theta_t \rangle + \frac{L}{2}\|\theta - \theta_t\|^2$$

$$\implies \quad \min_{t \leq T} \frac{1}{2}\|\nabla \mathcal{L}(\theta_t)\|^2 \leq \frac{L}{T}(\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*))$$

- ✗ depends on the parametrization
- ✗ unknown constant, $L = \infty$?

Most models are not smooth

fitting $\mathcal{N}(\mu, \sigma^2)$



Variance can not be upper-bounded by a quadratic

EM for EFs is Mirror Descent

Gradient descent with step-size α:

$$\begin{aligned}\theta_{t+1} &= \theta_t - \alpha \nabla \mathcal{L}(\theta_t) \\ &\in \arg\min_\theta \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta - \theta_t \rangle + \frac{1}{\alpha} \frac{1}{2} \|\theta - \theta_t\|^2\end{aligned}$$

Converges if $\mathcal{L}$ is $(1/\alpha)$-smooth.

Mirror descent is a generalization allowing a Bregman divergences:

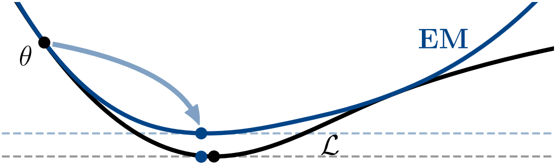
$$\theta_{t+1} \in \arg\min_\theta \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta - \theta_t \rangle + \frac{1}{\alpha} D_h(\theta, \theta_t)$$

Converges if $\mathcal{L}$ is $(1/\alpha)$-smooth **relative to a reference function** h.

We show EM for EFs is mirror descent (1-smooth relative to A):

$$\theta_{t+1} = \arg\min_\theta \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta - \theta_t \rangle + \underbrace{D_A(\theta, \theta_t)}_{\mathrm{KL}[p_{\theta_t} \,\|\, p_\theta]}$$

Our approach



$$\mathcal{L}(\theta) \leq \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta - \theta_t \rangle + \mathrm{KL}[p_{\theta_t} \parallel p_\theta]$$

$$\min_{t \leq T} \mathrm{KL}[p_{\theta_{t+1}} \parallel p_{\theta_t}] \leq \frac{1}{T}(\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*))$$

- ✓ parametrization invariant
- ✓ no unknown/infinite constant

Stationary points in KL divergence

GD $$\min_{t \le T} \frac{1}{2}\|\nabla \mathcal{L}(\theta_t)\|^2 \le \boldsymbol{L}\frac{\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*)}{T}$$

EM $$\min_{t \le T} D_A(\theta_t, \theta_{t+1}) \le \frac{\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*)}{T}$$

How does $D_A(\theta_t, \theta_t)$ relate to stationarity?

GD $$\|\nabla \mathcal{L}(\theta_t)\| = \|\bar{\mu}_t - \mu_t\|.$$

EM $$D_A(\theta_t, \theta_{t+1}) = D_{A^*}(\bar{\mu}_t, \mu_t).$$

GD tries to shrink gradient, EM tries to shrink natural gradient.

$$D_A(\theta_t, \theta_{t+1}) \approx \frac{1}{2}\|\nabla \mathcal{L}(\theta_t)\|^2_{I(\theta_t)^{-1}}$$

Convergence near a strict local optimum

Known asymptotically: EM has linear convergence rate,

$$\text{as } t \to \infty \qquad \mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_*) \leq r(\theta_*)[\mathcal{L}(\theta_t) - \mathcal{L}(\theta_*)]$$

$$r(\theta) = \lambda\big(I_{z\,|\,x}(\theta) I_{x,z}(\theta)^{-1}\big) \qquad \text{“How much information is missing”}$$

Non-asymptotic: Strong-convexity region relative to $A \quad \leftrightarrow \quad r(\theta) \leq r$

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_*) \leq r[\mathcal{L}(\theta_t) - \mathcal{L}(\theta_*)]$$

Superlinear convergence if $r(\theta_*) = 0$.

Summary

EM is extremely widely-used for EF models with missing data.

- Gaussian mixtures, student t, hidden Markov models, and so on.
- But previous non-asymptotic analyses show same rate as GD.

Main result is is a convergence rate of EM in terms of KL divergence:

- Based on showing EM for EFs is mirror descent with $\alpha = 1$.
- Invariant to parameterization.
  - No dependence on Lipschitz constant (which is often ∞).

The paper gives many results beyond the basic setting:

- Adding a conjugate prior (still parameterization-invariant).
- Linear/superlinear local convergence rates.
  - Depending on ratio of missing information.
- Approximate M-steps, and cases where M-step is not in the EF.