Homeomorphic-Invariance of EM: Non-Asymptotic Convergence in KL Divergence

Frederik Kunstner[1] Raunak Kumar[2] Mark Schmidt[1]

University of British Columbia Cornell University

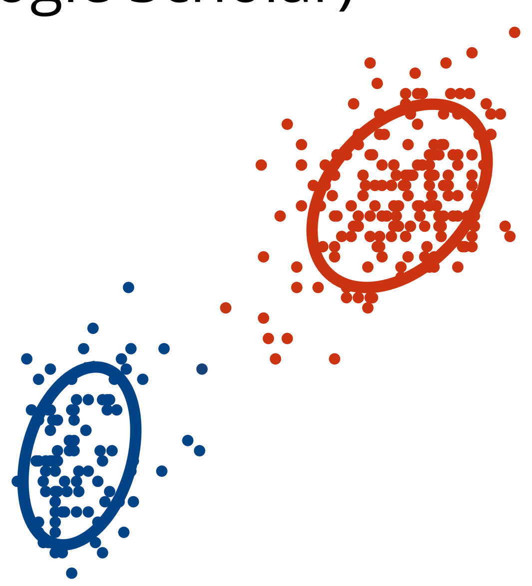
EM is a fundamental algorithm for probabilistic models

The default to deal with missing data or latent variables

The EM paper (Dempster et al., 1977) is one of the most cited works of the last century ($>$ 60'000 on Google Scholar)

Most famous application in ML:
clustering with mixtures of Gaussians

The "missing data":
which cluster generated each point



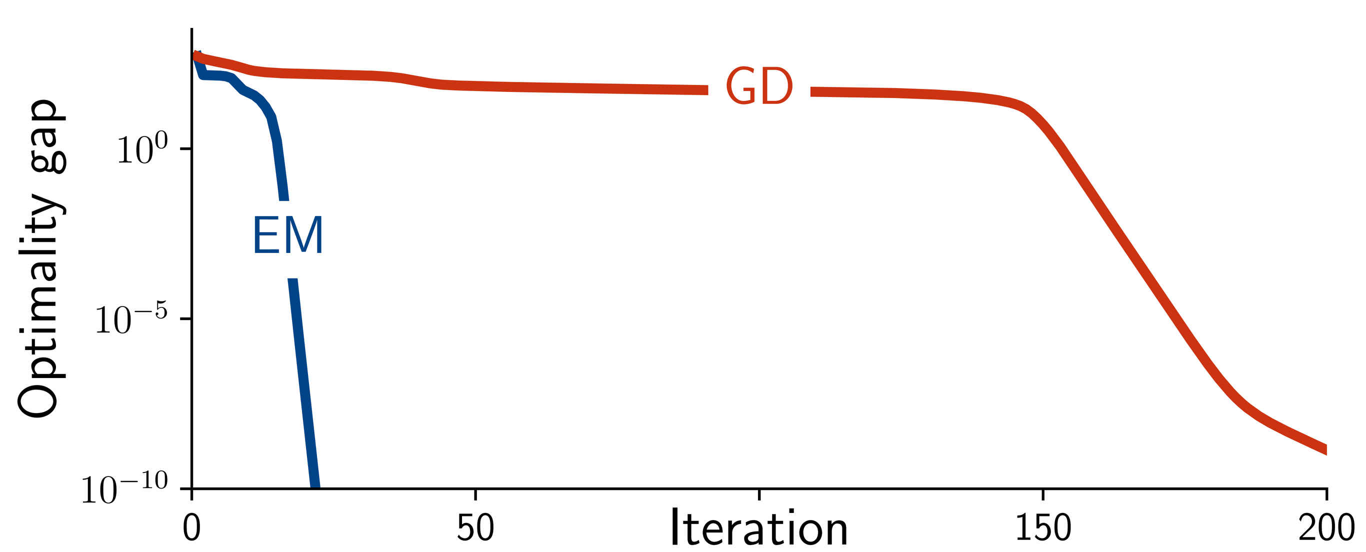
Does maximum likelihood with observed data x and missing or latent variables z by averaging over possible missing data

$$\mathcal{L}(\theta) = -\log p(x \mid \theta) = -\log \int p(x, z \mid \theta)\, dz$$

But we do not have a good understanding of its performance

Previous work: "EM does at least as well as gradient descent"

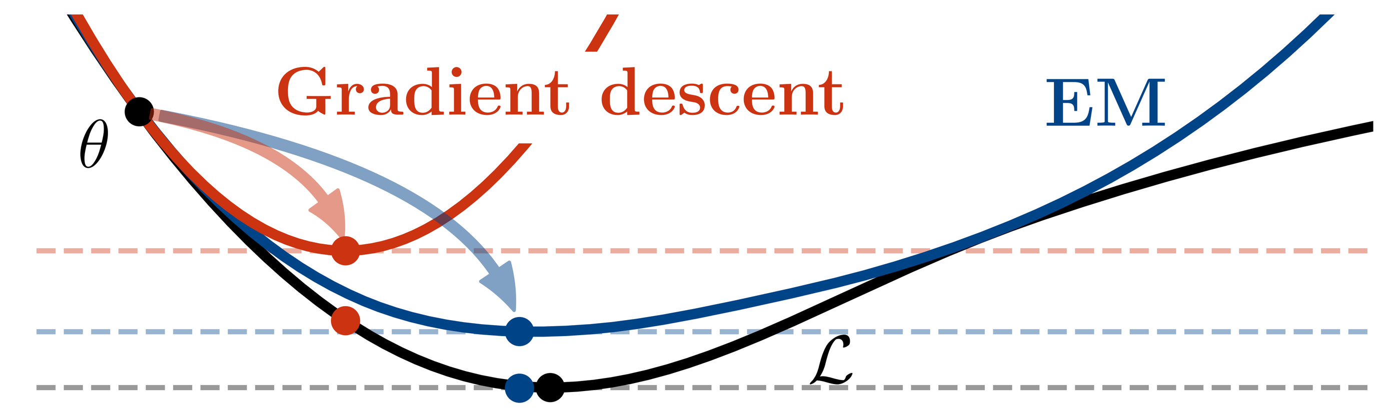
Quite the understatement, even for the toy problem above (gradient descent also needs a step-size, here with grid-search)



Our understanding of optimization for probabilistic models is limited. Need to understand EM to develop better methods!

The standard approach assumes the objective is smooth

"the bound optimized by EM is bounded by a quadratic"



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

Standard results for gradient descent on non-convex functions

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

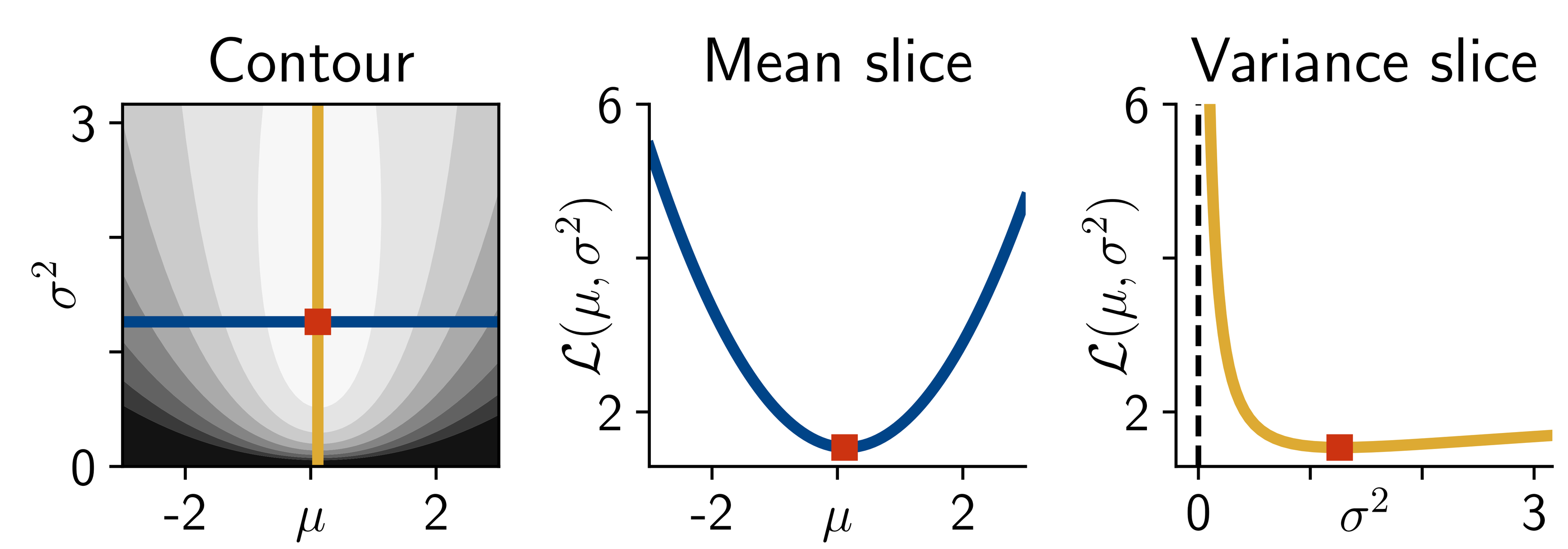
But this doesn't hold even for mixture of Gaussians

The gradient descent analysis:

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

For many problems including mixtures of Gaussians, the objective function is not smooth

Fitting a Gaussian $\mathcal{N}(\mu, \sigma^2)$ is already not smooth
Loss diverges when $\sigma^2 \to 0$, cannot be bounded by a quadratic



Our approach: analysis in KL divergence

For **exponential family models** of the form

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

the bound optimized by EM (and the algorithm) can be written as **mirror descent** (relative to the log-partition function A)

$$\theta_{t+1} = \min_{\theta} \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta - \theta_t \rangle + D_A(\theta, \theta_t)$$

The Bregman divergence D_A is the KL between models

$$D_A(\theta, \theta_t) = \mathrm{KL}[p_{\theta_t} \,\|\, p_\theta]$$

Gives **convergence to stationary points in KL divergence**

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

The KL analysis:

- ✓ **is parametrization invariant**
  only the probability distributions matter
- ✓ **has no unknown constant**
  $L = 1$, no hyperparameter
- ✓ **works for mixture of Gaussians**
  and most applications of EM with a closed-form M step

The analysis extends to common variations

- Using a prior to ensure valid updates
- Convergence to local min. in convex region
- Linear rate depending on ratio of missing information
- Approximate M-steps

more details in the paper!