Nonconvex Optimization for Latent Data Models: Algorithms, Analysis and Applications

Belhal Karimi

Ph.D. Defense - September 19th 2019

Under The Supervision of **Marc Lavielle** and **Eric Moulines**

INVENTEURS DU MONDE NUMÉRIQUE

ECOLE POLYTECHNIQUE

IP PARIS

A Principled Approach

- **Observe** the world
- **Design a model** of the world
- **Learn this model** of the world

+ Convex −	
Easy to implement	Small set of applications
Algorithms with strong guarantees	Much higher computation and memory costs

+ Nonconvex −	
Countless number of problems and applications	Requires new programming languages and skills
Going back to simpler methods in $\mathcal{O}(dn)$	Weak theory. No general recipe for hyper parameters tuning

Challenges: **Explosion of data** **and** **Complex Models**

New **Methods** and **Theory** for modern **Applications**

Statistical Learning in Latent Data Models

Supervised Learning

- **Sensors:** Observe input-output pairs of random variables (x, y) in (X, Y) from unknown distribution $\mathcal{P}$
- **Modeling Phase:** Define a model $M_\theta : \mathsf{X} \to \mathsf{Y}$ of parameter $\theta \in \mathbb{R}^d$, called the predictor
- **Performance Measuring:** Measure the performance of the model using a *loss* function $\ell : \mathsf{Y} \times \mathsf{Y} \mapsto \mathbb{R}$ where $\ell(y, y')$ is the loss incurred when the true output is y whereas y' is predicted
- **Training Phase:** Train your model on your observations as follows

Data fit term Regularization

$$\arg\min_{\theta \in \mathbb{R}^d} \overline{\mathcal{L}}(\boldsymbol{\theta}) = \arg\min_{\theta \in \mathbb{R}^d} \{\mathcal{L}(\boldsymbol{\theta}) + \mathrm{R}(\boldsymbol{\theta})\}$$

$$\mathcal{L}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \ell\left(y_i, M_{\boldsymbol{\theta}}\left(x_i\right)\right)$$

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\ell\left(y, M_{\boldsymbol{\theta}}(x)\right)\right]$$

Optimization at the heart of (Supervised) **Learning**

Latent Data Models

Mathematical Formulation

- Models where the input-output relationship is not completely characterized by the observed $(x, y) \in \mathsf{X} \times \mathsf{Y}$

- Dependence on a set of unobserved latent variables

$$z \in \mathsf{Z} \subset \mathbb{R}^m$$

- Loss function ℓ to accept a third argument as follows:

$$\ell(y, M_\theta(x)) = \int_{\mathsf{Z}} \ell(z, y, M_\theta(x)) \mathrm{d}z$$

Some examples

- **Missing Data**

Latent Structure = Missingness of the Data

- **Mixed Effects Models**

Latent Structure = Interindividual Variability

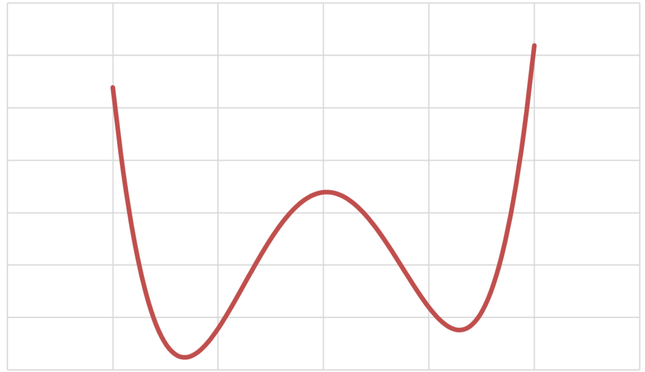
- **Mixture Models**

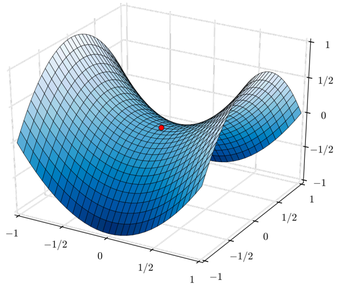
Latent Structure = Mixture Components

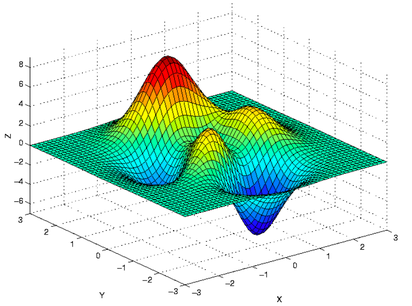
Nonconvex Optimization

- **Convexity:** $\mathcal{L}(\theta)$ is **non-convex**

$$\mathcal{L}(\lambda\theta + (1-\lambda)\vartheta) \leq \lambda\mathcal{L}(\theta) + (1-\lambda)\mathcal{L}(\vartheta)$$



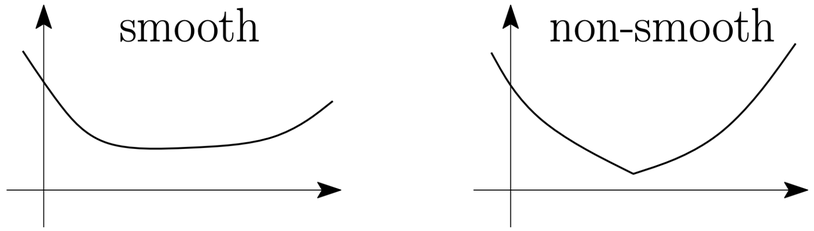




- **Smoothness:**

A function $\mathcal{L} : \mathbb{R}^d \mapsto \mathbb{R}$ is L-smooth if and only if it is twice differentiable and its gradient is L-Lipschitz-continuous, i.e. for all $(\theta, \vartheta) \in \mathbb{R}^d \times \mathbb{R}^d$:

$$\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\vartheta)\| \leq L\|\theta - \vartheta\|$$



- **Examples of non-convex problems:**

PCA **Mixture Models** **Neural Networks**

Nonlinear regression **MLE with Latent Variables**

Why is Non-convex optimization is hard?

- **How to solve those non-convex problems**

Either same techniques as in the ***convex*** *case*: SGD, Mini-batching, SVRG, Momentum

Or other methods for non-convex problems: Alternating Minimization Procedures or Majorization-Minimization

- **Varieties of theoretical convergence results:**

Convergence to **stationary** point, to **local** minimum

Local Convergence to **global** minima

Global Convergence to **global** minima

- **Stability condition to analyze training algorithms:**

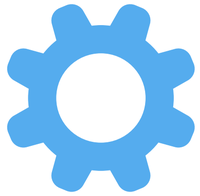
Convex Objective ‣ Can efficiently find the optimal solution θ^* $\quad |\mathcal{L}(\theta) - \mathcal{L}(\theta^*)|$ or $\left\|\theta - \theta^*\right\|^2$

Non-convex Objective ‣ Characterization of the stationarity $\quad \|\nabla \mathcal{L}(\theta)\|^2$

θ^* is said to be $\mathcal{E}$-stationary if $\|\nabla \mathcal{L}(\theta^*)\|^2 \leq \varepsilon$

A stochastic algorithm achieves $\mathcal{E}$-stationarity in $\Gamma > 0$ iterations if $\mathbb{E}\left[\|\nabla \mathcal{L}(\theta^{(\Gamma)})\|^2\right] \leq \varepsilon$

Contributions Of This Thesis



- Develop Improved **Methods**
  - Faster **Convergence**
  - **Scaling** to large datasets
  - Easy To **Implement**
  - **R Package**

- **Derive** Statistical Guarantees
  - Convergence Behavior
  - Global analysis
  - Rates

- **Apply** methods to models of interests in **Machine Learning** and **Check** assumptions
  - From Logistic Regression to MDP
  - Real and Simulated Datasets

Agenda

Risk Minimization

$$\mathcal{L}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \ell\left(y_i, M_{\boldsymbol{\theta}}\left(x_i\right)\right) \text{ or } \mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\ell\left(y, M_{\boldsymbol{\theta}}(x)\right)\right]$$

1

- **Incremental** Updates
- **Constrained** Minimization
- Applications to **Logistic Regression** and **Bayesian Deep Learning**

2

- **Online** Updates
- **Biased** Stochastic Approximation
- Applications to **online EM** and **Reinforcement Learning**

Maximum Likelihood

$$\log p(y, \theta) = \sum_{i=1}^{n} \log p_i(y_i, \theta)$$

3

- **EM** Algorithm
- **Exponential Family** Model
- **Incremental** Updates
- Applications to **Mixture** and **Topic Modeling**

4

- Applications to **Pharmacology**
- Stochastic **Approximation** of EM
- **Mixed Effects** Models
- 2 **Variants**

Non-convex Risk Minimization

Agenda

Risk Minimization

$$\mathcal{L}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \ell\left(y_i, M_{\boldsymbol{\theta}}\left(x_i\right)\right) \text{ or } \mathcal{L}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\ell\left(y, M_{\theta}(x)\right)\right]$$

1

- **Incremental** Updates
- **Constrained** Minimization
- Applications to **Logistic Regression** and **Bayesian Deep Learning**

2

- **Online** Updates
- **Biased** Stochastic Approximation
- Applications to **online EM** and **Reinforcement Learning**

Maximum Likelihood

$$\log p(y, \theta) = \sum_{i=1}^{n} \log p_i(y_i, \theta)$$

3

- **EM** Algorithm
- **Exponential Family** Model
- **Incremental** Updates
- Applications to **Mixture** and **Topic Modeling**

4

- Applications to **Pharmacology**
- Stochastic **Approximation** of EM
- **Mixed Effects** Models
- 2 **Variants**

Large-Scale ML

Constrained Optimization of Finite Sum

$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta})$$

- Large finite-sum objective function
- Convex and compact Θ subset of $\mathbb{R}^d$
- Function $\mathcal{L}_i : \mathbb{R}^d \to \mathbb{R}$ is bounded from below and (possibly) non-convex and non-smooth

Some examples

- Logistic Regression {-1/1} binary outputs

$$\mathcal{L}_i(\theta) := \log(1 + e^{-y_i \theta^\top x_i})$$

- Variational Inference

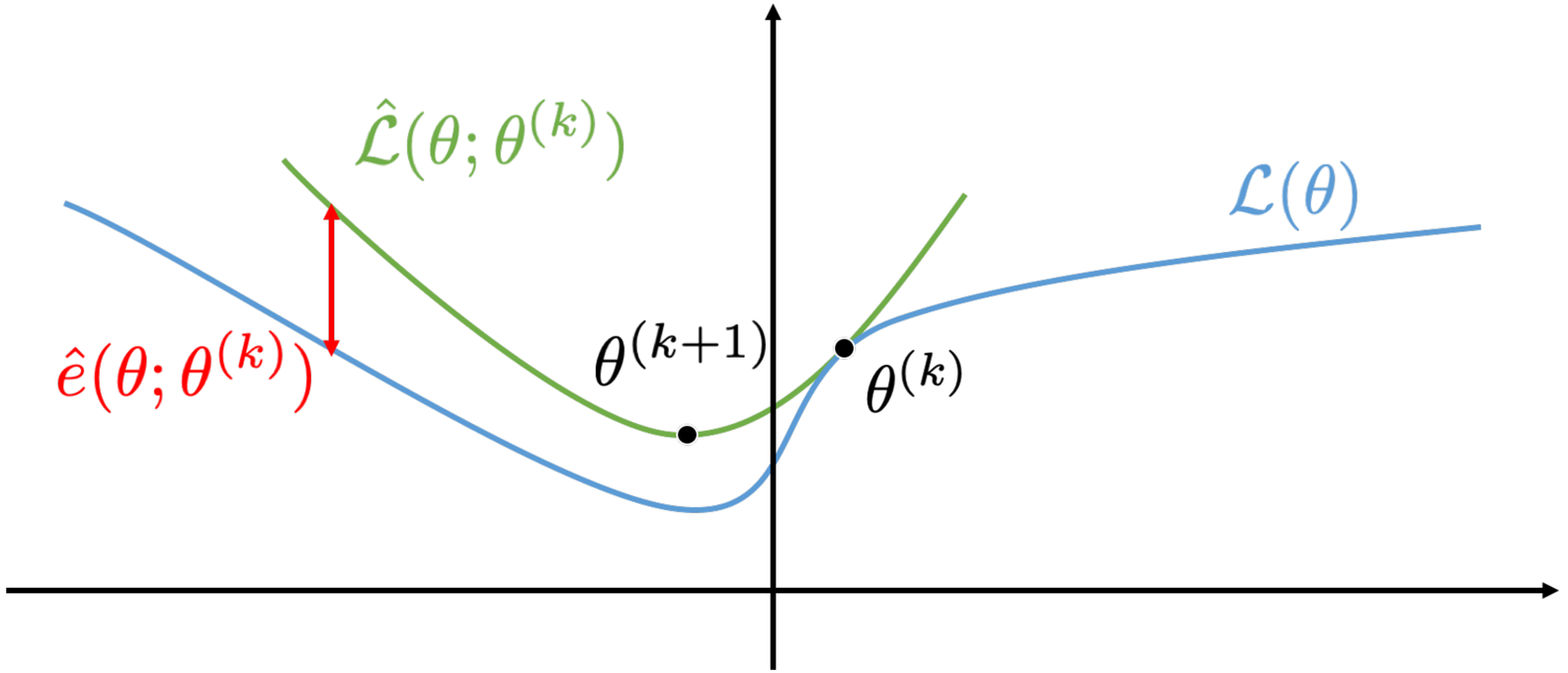
$$\mathcal{L}_i(\theta) := \mathrm{KL}\left(q(w;\theta) \| p_i\left(w | y_i, x_i\right)\right)$$

- Maximum likelihood estimation

$$\mathcal{L}_i(\theta) := -\log p_i(y_i, \theta)$$

Majorization-Minimization Principle

[Lange, 2013]



- Iteratively minimize locally tight upper bounds on the objective
- Drives the objective function downwards
- Examples: the proximal gradient algorithm **[Beck and Teboulle, 2009]**, the EM algorithm **[McLachlan and Krishnan, 2007]** and variational inference **[Wainwright and Jordan, 2008]**.
- The approximation error $\hat{e}(\theta,\bar{\theta})$ at $\bar{\theta}$ plays a key role in the analysis

MISO Algorithm

[Mairal, 2015]

Algorithm 1 MISO algorithm

Initialization: given an initial parameter estimate $\hat{\theta}^{(0)}$, for all $i \in [\![1, n]\!]$ compute a surrogate function $\vartheta \to \widehat{\mathcal{L}}_i(\hat{\theta}^{(0)}; \vartheta)$.

Iteration k: given the current estimate $\hat{\theta}^{(k)}$:

1. Pick i_k uniformly from $[\![1, n]\!]$.
2. Update $\mathcal{A}_i^{k+1}(\boldsymbol{\theta})$ as:

$$\mathcal{A}_i^{k+1}(\boldsymbol{\theta}) = \begin{cases} \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(k)}), & \text{if } i = i_k \\ \mathcal{A}_i^k(\boldsymbol{\theta}), & \text{otherwise.} \end{cases}$$

3. Set $\hat{\theta}^{(k+1)} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{A}_i^{k+1}(\boldsymbol{\theta})$.

- Extension to Latent Data Models?
- How do those surrogates functions look like when there exists a dependence on a latent variable
- Can we derive a general algorithm?

Examples

- For smooth function $\mathcal{L}_i(\theta)$

$$\hat{\mathcal{L}}_i(\cdot, \bar{\theta}) : \theta \mapsto \mathcal{L}_i(\bar{\theta}) + \nabla \mathcal{L}_i(\bar{\theta})^\top (\theta - \bar{\theta}) + \frac{L}{2} \|\theta - \bar{\theta}\|_2^2$$

Leading to Gradient Descent Algorithm

Existing Results

- **Nonconvex problems:** Almost Sure Convergence
- **Convex problems rates on** $\mathcal{L}(\theta^{(k)}) - \mathcal{L}^*$ **:**
  - $\mathcal{O}(nL/k)$ rate for convex objective
  - $\mathcal{O}((1 - \mu/(nL))^k)$ rate for strongly convex objective

Intractable Surrogates

The EM algorithm

- Complete likelihood, i.e., joint likelihood of the observations and the latent data: $f_i(z_i, y_i, \theta)$
- Likelihood of the observations: $g_i(y_i, \theta) := \int_{\mathsf{Z}} f_i(z_i, y_i, \theta)\mu_i(\mathrm{d}z_i)$
- Posterior distribution: $p_i(z_i, \theta)$
- Objective function: $\mathcal{L}_i(\theta) := -\log g_i(y_i, \theta)$

$$\mathcal{L}_i(\theta) := -\log g_i(y_i, \theta) = -\log \int_{\mathsf{Z}} f_i(z_i, y_i, \theta)\mu_i(\mathrm{d}z_i) = -\log \int_{\mathsf{Z}} f_i(z_i, y_i, \theta)\frac{p_i(z_i, \vartheta)}{p_i(z_i, \vartheta)}\mu_i(\mathrm{d}z_i)$$

Jensen Inequality $\longrightarrow$

$$\leq \int_{\mathsf{Z}} \log \frac{p_i(z_i, \vartheta)}{f_i(z_i, y_i, \theta)} p_i(z_i, \vartheta)\mu_i(\mathrm{d}z_i)$$

- Kullback Leibler (KL) surrogate [Neal and Hinton, 1998]

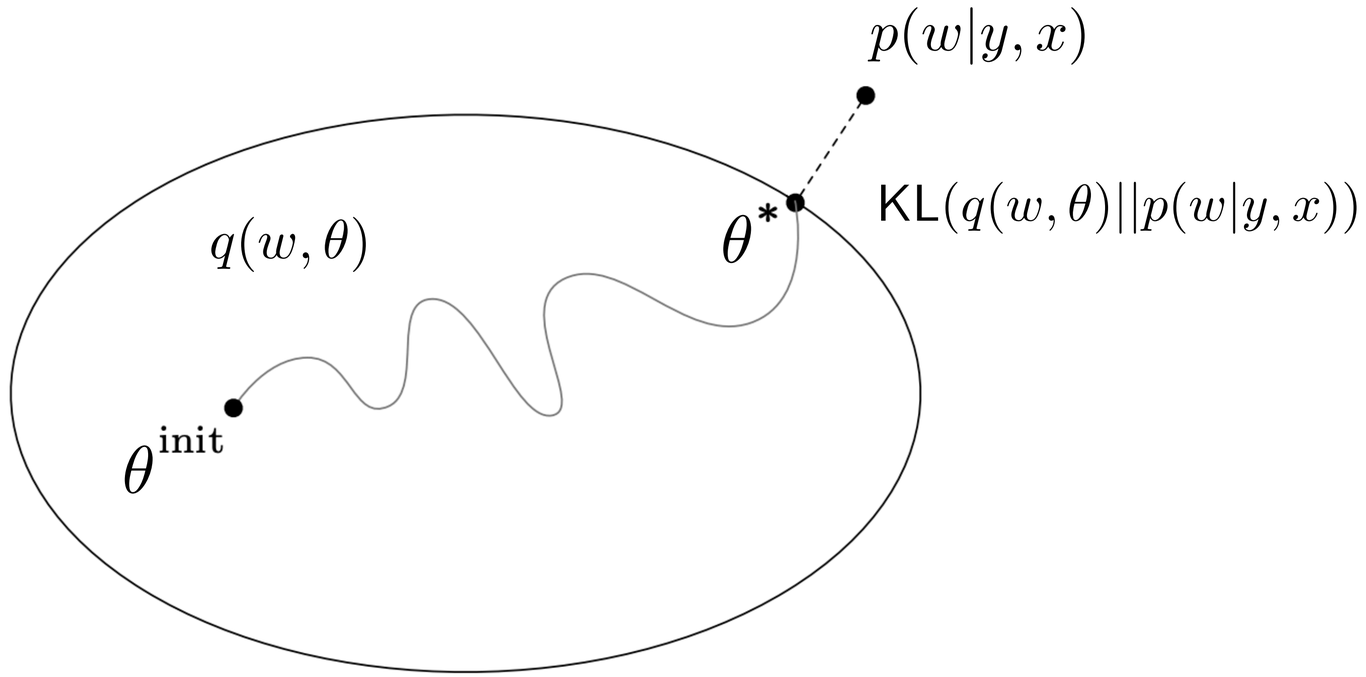
$$\hat{\mathcal{L}}_i(\theta, \vartheta) := \int_{\mathsf{Z}} \log \frac{p_i(z_i, \vartheta)}{f_i(z_i, y_i, \theta)} p_i(z_i, \vartheta)\mu_i(\mathrm{d}z_i) = \mathsf{KL}(p_i(z_i, \vartheta)||p_i(z_i, \theta)) + \mathcal{L}_i(\theta)$$

Intractable KL term

Intractable Surrogates

Variational Inference (VI)

- Input-output pairs $((x_i, y_i), 1 \leq i \leq n)$ and w a global latent variable with a prior distribution $\pi(w)$
- We want to minimize the KL between the variational candidate $q(w, \theta)$ and the true posterior $p(w|y, x)$

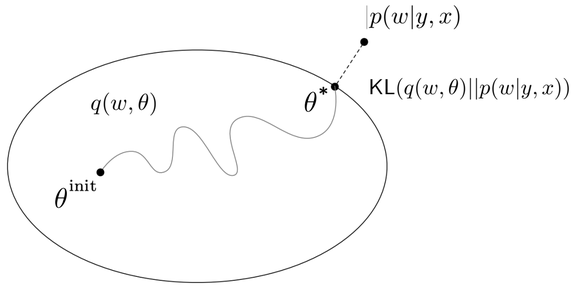


- KL term is intracable: VI optimizes the Evidence Lower Bound (ELBO)

$$\mathcal{L}(\theta) := -\mathbb{E}_{q(w;\theta)}\big[\log p(y|x, w)\big] + \mathbb{E}_{q(w;\theta)}\big[\log q(w;\theta)/\pi(w)\big]$$

- ELBO is a lower bound of the incomplete log likelihood.
- Maximizing ELBO minimizes the KL
- ‘Data term’ fits to the data and ‘KL’ term fits to the prior

Intractable Surrogates



Variational Inference (VI)

- Input-output pairs $((x_i, y_i), 1 \leq i \leq n)$ and w a global latent variable with a prior distribution $\pi(w)$
- We want to minimize the KL between the variational candidate $q(w, \theta)$ and the true posterior $p(w|y, x)$
- Individual Objective function:

$$\mathcal{L}_i(\theta) := -\mathbb{E}_{q(w;\theta)}\big[\log p(y_i|x_i, w)\big] + \frac{1}{n}\mathbb{E}_{q(w;\theta)}\big[\log q(w;\theta)/\pi(w)\big]$$

- Quadratic Surrogate function:

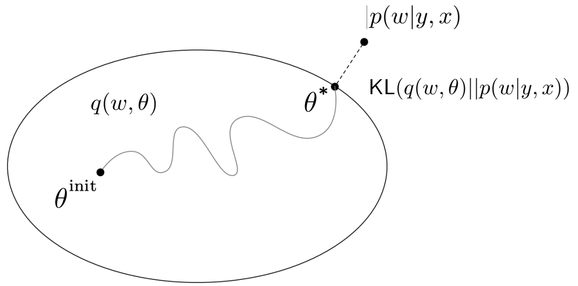
$$\hat{\mathcal{L}}_i(\cdot, \vartheta) : \theta \mapsto \mathcal{L}_i(\vartheta) + (\nabla r_i(\vartheta) + \nabla d_i(\vartheta))^\top(\theta - \vartheta) + \frac{L}{2}\|\theta - \vartheta\|_2^2$$

- Reparametrization trick [Blundell+, 2015]:

Let $t : \mathbb{R}^d \times \Theta \mapsto \mathbb{R}^d$ be a differentiable function w.r.t. ϑ s.t. $w = t(z, \vartheta) \sim q(\cdot, \vartheta)$ and $z \sim \mathcal{N}_p(0, I)$

$$\nabla r_i(\vartheta) = \mathbb{E}_{z\sim\mathcal{N}_p(0,I)}[\mathsf{J}_\theta^t(z, \vartheta)\nabla_w \log p(y_i|x_i, w)|_{w=t(z,\vartheta)}]$$

Intractable Surrogates



Variational Inference (VI)

- Input-output pairs $((x_i, y_i), 1 \leq i \leq n)$ and w a global latent variable with a prior distribution $\pi(w)$
- We want to minimize the KL between the variational candidate $q(w, \theta)$ and the true posterior $p(w|y, x)$
- Individual Objective function:

$$\mathcal{L}_i(\theta) := -\mathbb{E}_{q(w;\theta)}\big[\log p(y_i|x_i, w)\big] + \frac{1}{n}\mathbb{E}_{q(w;\theta)}\big[\log q(w;\theta)/\pi(w)\big]$$

- Quadratic Surrogate function:

$$\hat{\mathcal{L}}_i(\cdot, \vartheta) : \theta \mapsto \mathcal{L}_i(\vartheta) + (\nabla r_i(\vartheta) + \nabla d_i(\vartheta))^\top(\theta - \vartheta) + \frac{L}{2}\|\theta - \vartheta\|_2^2$$

- Reparametrization trick [Blundell+, 2015]:

Let $t : \mathbb{R}^d \times \Theta \mapsto \mathbb{R}^d$ be a differentiable function w.r.t. ϑ s.t. $w = t(z, \vartheta) \sim q(\cdot, \vartheta)$ and $z \sim \mathcal{N}_p(0, I)$

$$\nabla r_i(\vartheta) = \mathbb{E}_{z\sim\mathcal{N}_p(0,I)}[\mathsf{J}_\theta^t(z, \vartheta)\nabla_w \log p(y_i|x_i, w)|_{w=t(z,\vartheta)}]$$

Intractable gradient term

MISSO Algorithm

[Karimi+, 2019]

Algorithm 2 MISSO algorithm

Initialization: $\hat{\theta}^{(0)}$; a sequence of non-negative numbers $\{M_{(k)}\}_{k=0}^{\infty}$.

For all $i \in [\![1, n]\!]$, draw $M_{(0)}$ samples from $p_i(\cdot; \hat{\theta}^{(0)})$ and $\widetilde{\mathcal{A}}_i^0(\theta) := \widetilde{\mathcal{L}}_i(\theta; \hat{\theta}^{(0)}, \{z_{i,m}^{(0)}\}_{m=1}^{M_{(k)}})$.

Iteration k: given the current estimate $\hat{\theta}^{(k)}$:

1. Pick a function index i_k uniformly on $[\![1, n]\!]$.
2. Draw $M_{(k)}$ Monte-Carlo samples from $p_i(\cdot; \hat{\theta}^{(k)})$.
3. Update the individual surrogate functions recursively as:

$$\widetilde{\mathcal{A}}_i^{k+1}(\theta) = \begin{cases} \widetilde{\mathcal{L}}_i(\theta; \hat{\theta}^{(k)}, \{z_{i,m}^{(k)}\}_{m=1}^{M_{(k)}}), & \text{if } i = i_k \\ \widetilde{\mathcal{A}}_i^k(\theta), & \text{otherwise.} \end{cases} \tag{13}$$

4. Set $\hat{\theta}^{(k+1)} \in \arg\min_{\theta \in \Theta} \widetilde{\mathcal{L}}^{(k+1)}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \widetilde{\mathcal{A}}_i^{k+1}(\theta)$.

Class of Surrogate Functions

- Set of latent variables $(z_i \in \mathsf{Z}^m, 1 \leq i \leq n)$
- $\mathcal{P}_i = \{p_i(z_i, \theta); \theta \in \Theta\}$ Family of probability densities with respect to μ_i
- There exists function $r_i(\theta, \vartheta, z_i)$ such that:

$$\hat{\mathcal{L}}_i(\theta, \vartheta) := \int_{\mathsf{Z}} r_i(\theta, \vartheta, z_i) p_i(z_i; \vartheta) \mu_i(\mathrm{d}z_i)$$

- $\hat{\mathcal{L}}_i(\theta, \vartheta)$ fully defined by the pair $(r_i(\theta, \vartheta, z_i), p_i(z_i; \vartheta))$

Minimization by Incremental Stochastic Surrogate Optimization (MISSO) method:

$$\tilde{\mathcal{L}}_i(\theta, \vartheta, \{z_m\}_{m=1}^{M}) := \frac{1}{M} \sum_{m=1}^{M} r_i(\theta, \vartheta, z_m)$$

where $\{z_m\}_{m=1}^{M}$ is the Monte Carlo batch sampled from $p_i(z_i; \vartheta)$ either directly or using MCMC

Analysis for Constrained Optimization

- **Constrained** optimization, consider the following **stationarity measure**:

$$g(\overline{\boldsymbol{\theta}}) := \inf_{\boldsymbol{\theta} \in \Theta} \frac{\mathcal{L}'(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}})}{\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|} \quad \text{and} \quad g(\overline{\boldsymbol{\theta}}) = g_+(\overline{\boldsymbol{\theta}}) - g_-(\overline{\boldsymbol{\theta}})$$

where $g_+(\bar{\theta}) := \max\{0, g(\bar{\theta})\}$ and $g_-(\bar{\theta}) := -\min\{0, g(\bar{\theta})\}$ denote the **positive** and **negative** part of $g(\overline{\boldsymbol{\theta}})$,respectively.

- $\bar{\theta}$ is a stationary point if and only if $g_-(\bar{\theta}) = 0$ **[Fletcher+, 2002]**.
- Furthermore, suppose that the sequence $\{\theta^{(k)}\}_{k \geq 0}$ has a limit point $\bar{\theta}$ that is a stationary point, then one has:

$$\lim_{k \to \infty} g_-(\theta^{(k)}) = 0$$

Global Convergence

Assumptions

(S1) Upper bounding surrogate $\hat{\mathcal{L}}_i(\theta, \vartheta) \geq \mathcal{L}_i(\theta)$ with **equality** if $\theta = \vartheta$

(S2) The **approximation error** $\widehat{e}(\theta; \{\vartheta_i\}_{i=1}^n) := \dfrac{1}{n}\sum_{i=1}^{n} \hat{\mathcal{L}}_i(\theta, \vartheta_i) - \mathcal{L}(\theta)$

is defined on $\Theta\epsilon = \{\theta \in \mathbb{R}^d, \inf_{\theta' \in \Theta} \|\theta - \theta'\| < \epsilon\}$ and satisfies: $\left\|\nabla \widehat{e}(\theta; \{\vartheta_i\}_{i=1}^n)\right\|^2 \leq 2L\, \widehat{e}(\theta; \{\vartheta_i\}_{i=1}^n)$

(H1) $r_i(\theta, \vartheta, z_i)$ is **convex** and **lower bounded**

(H2) There exists the following **constants**:

$$C_{\mathsf{r}} := \sup_{\vartheta \in \Theta} \sup_{M > 0} \frac{1}{\sqrt{M}} \mathbb{E}_{\vartheta} \left[\sup_{\theta \in \Theta} \left| \sum_{m=1}^{M} \left\{ r_i(\theta; \vartheta, z_{i,m}) - \hat{\mathcal{L}}_i(\theta, \vartheta) \right\} \right| \right]$$

$$C_{\mathsf{gr}} := \sup_{\vartheta \in \Theta} \sup_{M > 0} \sqrt{M} \mathbb{E}_{\vartheta} \left[\sup_{\theta \in \Theta} \left| \frac{1}{M} \sum_{m=1}^{M} \frac{\widehat{\mathcal{L}}_i'(\theta, \theta - \vartheta; \vartheta) - r_i'(\theta, \theta - \vartheta; \vartheta, z_{i,m})}{\|\vartheta - \theta\|} \right|^2 \right]$$

Global Convergence

Non-Asymptotic Analysis

Theorem

Under **(S1)**, **(S2)**, **(H1)**, **(H2)** and define the following quantity:

$$\Delta_{(K_{\max})} := 2nL\mathbb{E}\left[\widetilde{\mathcal{L}}^{(0)}\left(\hat{\boldsymbol{\theta}}^{(0)}\right) - \widetilde{\mathcal{L}}^{(K_{\max})}\left(\hat{\boldsymbol{\theta}}^{(K_{\max})}\right)\right] + \sum_{k=0}^{K_{\max}-1} \frac{4LC_r}{\sqrt{M_{(k)}}}$$

Then we have the following bounds:

$$\mathbb{E}\left[\left\|\nabla \hat{e}^{(K)}\left(\hat{\theta}^{(K)}\right)\right\|^2\right] \leq \frac{\Delta_{(K_{\max})}}{K_{\max}}$$

$$\mathbb{E}\left[g_-\left(\hat{\theta}^{(K)}\right)\right] \leq \sqrt{\frac{\Delta_{(K_{\max})}}{K_{\max}}} + \frac{C_{gr}}{K_{\max}} \sum_{k=0}^{K_{\max}-1} M_{(k)}^{-1/2}$$

Asymptotic Analysis

Theorem

Also, assume $\sum_{k=0}^{\infty} M_{(k)}^{-1/2} < \infty$ (non decreasing sequence)

- $\lim_{k\to\infty} g_-\left(\hat{\theta}^{(k)}\right) = 0$
- $\lim_{k\to\infty} \mathcal{L}\left(\hat{\theta}^{(k)}\right) = \mathcal{L}^*$

Remarks

- $\Delta_{(K_{\max})}$ **is finite for any** $K_{\max} \in \mathbb{N}$
- **MISO as a special case of MISSO**

$$C_{\mathrm{r}} = C_{\mathrm{gr}} = 0$$

- **Non-asymptotic rate of**

$$\mathbb{E}[g_-^{(K)}] \leq \mathcal{O}(\sqrt{nL/K_{\max}})$$

- **MISSO sequence** $\{\theta^{(k)}\}_{k\geq 0}$ **satisfies an** ***asymptotic stationary*** **point condition**

Numerical Applications

Logistic Regression with Missing Covariates

- $y = (y_i, 1 \leq i \leq n)$ vector of binary responses and $z_i = (z_{i,p}) \in \mathbb{R}^d$ covariates
- z_i is not fully observed:
  - $z_{i,mis}$ missing values and $z_{i,obs}$ observed
- $z = (z_i, 1 \leq i \leq n) \sim \mathcal{N}(\beta, \Omega)$ where $\beta \in \mathbb{R}^d$
- Logit model

$$p_i(y_i|z_i, \delta) = \frac{\exp(-y_i \boldsymbol{\delta}^\top \bar{z}_i)}{1 + \exp(-\boldsymbol{\delta}^\top \bar{z}_i)}$$

- Estimate $\theta = (\delta, \beta, \Omega)$

EM Surrogate

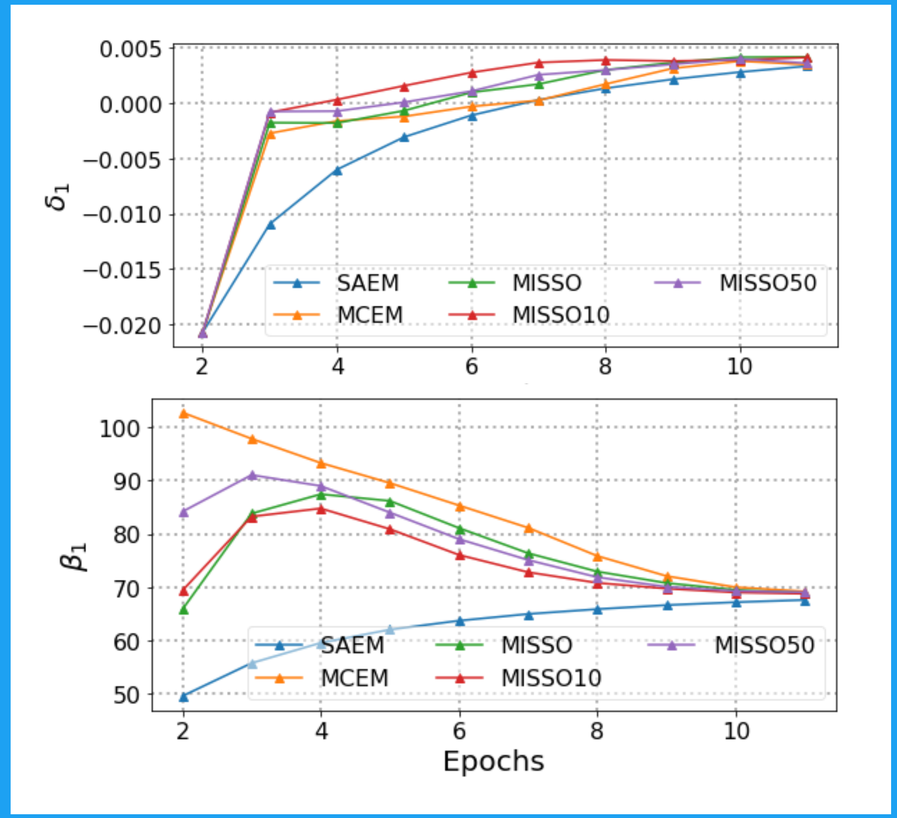
$$\hat{\mathcal{L}}_i(\theta, \vartheta) = -\int_Z \log p_i(y_i|z_{i,mis}, z_{i,obs}, \delta)\, p_i(z_i, \vartheta)\mu_i(\mathrm{d}z_{i,mis}) - \int_Z \log p_i(z_{i,mis}, \beta, \Omega)\, p_i(z_i, \vartheta)\mu_i(\mathrm{d}z_{i,mis})$$

$$\beta^{(k)} = \frac{1}{M_{(k)}} \sum_{m=1}^{M_{(k)}} \frac{1}{n} \sum_{i=1}^{n} z_{i,m}^{(k)} \quad \Omega^{(k)} = \frac{1}{M_{(k)}} \sum_{m=1}^{M_{(k)}} \frac{1}{n} \sum_{i=1}^{n} z_{i,m}^{(k)} (z_{i,m}^{(k)})^\top - \beta^{(k)} (\beta^{(k)})^\top$$

Quasi-newton for $\delta^{(k)}$

Experiments - TraumaBase dataset

- TraumaBase (http://traumabase.eu) dataset:
  - **15 trauma centers** in France
  - Measurements from the initial to last stage of trauma.
  - **6384** patients and **d=16** quantitative variables
- Predict binary response: severe trauma or not.



$$\hat{\mathcal{L}}_i(\cdot, \vartheta) : \theta \mapsto \mathcal{L}_i(\vartheta) + (\nabla r_i(\vartheta) + \nabla d_i(\vartheta))^\top (\theta - \vartheta) + \frac{L}{2}\|\theta - \vartheta\|_2^2$$

Numerical Applications

Fitting Bayesian LeNet5 on MNIST

- **MNIST:** $N = 60\,000$ handwritten digits, 28×28 images, $d = 784$

- **Weight prior:** $p(w) = \mathcal{N}(0, I)$
  $p(y_i|x_i, w) = \mathrm{Softmax}(f(x_i, w))$ where f is a NN

MISSO Update

- **Means**

$$\mu_\ell^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \mu_\ell^{(\tau_i^k)} - \frac{\gamma}{n}\sum_{i=1}^{n} \hat{\delta}_{\mu_\ell, I}^{(k)}$$

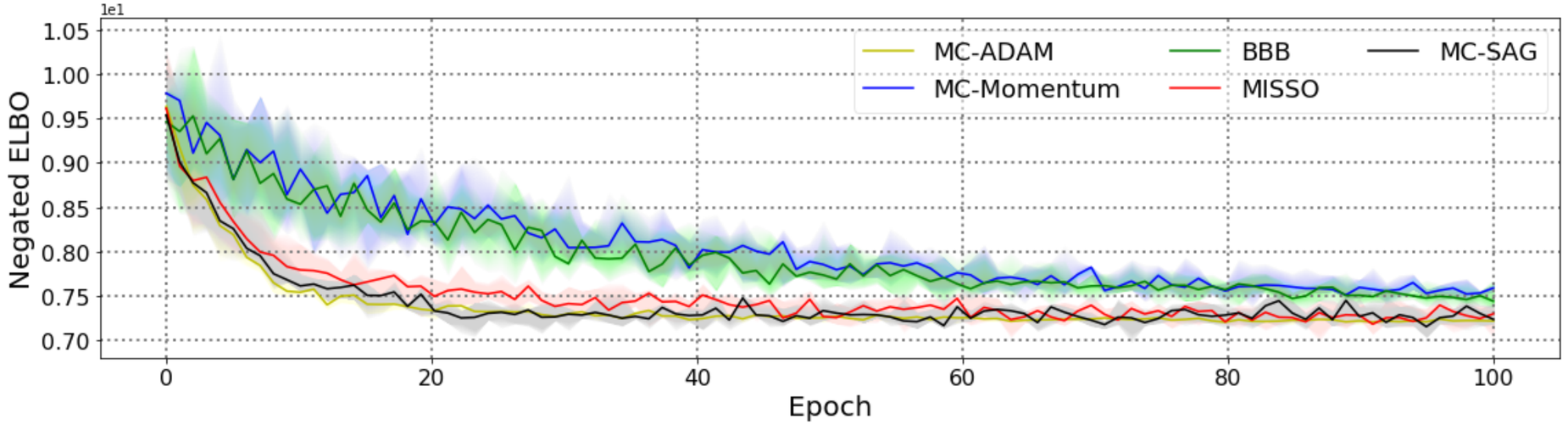
where:

$$\hat{\delta}_{\mu_\ell, i_k}^{(k)} = -\frac{1}{M_{(k)}}\sum_{m=1}^{M_{(k)}} \nabla_w \log p(y_{i_k}|x_{i_k}, w)\Big|_{w=t(\theta^{(k-1)}, z_m^{(k)})} + \nabla_{\mu_\ell} d(\theta^{(k-1)})$$

and:

$$d(\theta) = n^{-1}\sum_{\ell=1}^{d} \left(-\log(\sigma) + (\sigma^2 + \mu_\ell^2)/2 - 1/2\right)$$

- **Variational candidate:** for any layer: $q(w_\ell, \theta_\ell)$ is a Gaussian distribution $\mathcal{N}(\mu_\ell, \sigma^2 I)$



(Incremental Variational Inference) Negated ELBO versus epochs elapsed for fitting the Bayesian LeNet-5 on MNIST using different algorithms. The solid curve is obtained from averaging over 5 independent runs of the methods, and the shaded area represents the standard deviation.

Agenda

Risk Minimization

$$\mathcal{L}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \ell\left(y_i, M_{\boldsymbol{\theta}}\left(x_i\right)\right) \text{ or } \mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\ell\left(y, M_{\boldsymbol{\theta}}(x)\right)\right]$$

1

- Incremental Updates
- Constrained Minimization
- Applications to Logistic Regression and Bayesian Deep Learning

2

- **Online** Updates
- **Biased** Stochastic Approximation
- Applications to **online EM** and **Reinforcement Learning**

Maximum Likelihood

$$\log p(y, \theta) = \sum_{i=1}^{n} \log p_i(y_i, \theta)$$

3

- **EM** Algorithm
- **Exponential Family** Model
- **Incremental** Updates
- Applications to **Mixture** and **Topic Modeling**

4

- Applications to **Pharmacology**
- Stochastic **Approximation** of EM
- **Mixed Effects** Models
- 2 **Variants**

Stochastic Approximation Scheme

[Robbins and Monro, 1951]

- Consider a smooth and possibly non-convex Lyapunov function $V : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ of which we want to find a **stationary point**

- SA scheme is a stochastic process:

$$\theta^{(k+1)} = \theta^{(k)} - \gamma_{k+1} H_{\theta^{(k)}}(X_{k+1}), \quad k \in \mathbb{N}$$

- The **drift term** $H_{\theta^{(k)}}(X_{k+1})$ depends on an **i.i.d.~random element** X_{k+1} and the mean-field satisfies

$$h(\theta^{(k)}) = \mathbb{E}\big[H_{\theta^{(k)}}(X_{k+1})|\mathcal{F}_k\big] = \nabla V(\theta^{(k)}),$$

where $\mathcal{F}_k$ if the filtration generated by $\{\theta^{(0)}, \{X_m\}_{m \le k}\}$

- Here, **SA** scheme is better known as the **SGD** method.

This Work

We relax a few restrictions of the classical SA:

- The **mean field** $h(\theta) \neq \nabla V(\theta)$

$\Longrightarrow$ relevant to **non-gradient** method where the gradient is hard to compute, e.g., online EM.

- $\{X_k\}_{k \ge 1}$ is not i.i.d.~and form a **state-dependent Markov chain**

$\Longrightarrow$ relevant to **SGD with non-iid noise** and **policy gradient**. E.g., $\theta^{(k)}$ controls the policy in a Markov Decision Process

Biased SA Scheme

Prior Work

$$e_{k+1} = H_{\theta^{(k)}}(X_{k+1}) - h(\theta^{(k)})$$

- **Case 1:** $\{e_k\}_{k\geq 1}$ **is Martingale difference**

$$\mathbb{E}\left[e_{k+1}|\mathcal{F}_k\right] = 0$$

- **Asymptotic Analysis:** with smooth $h(\cdot)$ [Robbins and Monro, 1951], [Benveniste+, 1990], [Borkar, 2009]

- **Non-asymptotic Analysis:** focus on $h(\theta) = \nabla V(\theta)$
  - Convex case: rate of $\mathcal{O}(1/n)$ in [Moulines and Bach, 2011], biased SA also studied in TD learning [Dalal+, 2018].
  - Non-convex case: [Ghadimi and Lan, 2013], [Bottou+, 2018] studied convergence with martingale noise.

- **Case 2:** $\{e_k\}_{k\geq 1}$ **is state-controlled Markov noise**

$$\mathbb{E}\left[e_{k+1}|\mathcal{F}_k\right] = P_{\theta^{(k)}} H_{\theta^{(k)}}(X_k) - h(\theta^{(k)}) \neq 0$$

- **Asymptotic Analysis:** studied with $h(\theta) = \nabla V(\theta)$ [Kushner and Yin, 2003], similar biased SA setting in [Tadìc and Doucet, 2017]

- **Non-asymptotic Analysis:** rather poor…
  - [Sun+, 2018] and [Duchi+, 2012] assumed gradient drift term and **state-independent** Markov chain.
  - [Bhandari+, 2018] studied a similar setting but focuses on linear SA with convex Lyapunov function.

Algorithm Behavior Analysis

Stopping Criterion

- We adopt a stopping rule similar to [Ghadimi and Lan, 2013], which is typical for **non-convex** problems.
- Fix any $k \geq 1$ and let $K \in \{0, \ldots, k\}$ be a discrete random variable (independent of $\{\mathcal{F}_k, k \in \mathbb{N}\}$) with

$$\mathbb{P}(K = \ell) = \left(\sum_{i=0}^{k} \gamma_{i+1}\right)^{-1} \gamma_{\ell+1}$$

where K serves as the terminating iteration for the SA scheme.

- Assume K is distributed according the prob. distribution above and study the estimator $\theta^{(K)}$.

- Usually **Best Iterate** $\min_k \mathbb{E}[||\nabla \mathcal{L}(\theta^{(k)})||^2] \leq (1/K) \sum_k \mathbb{E}[||\nabla \mathcal{L}(\theta^{(k)})||^2]$
- **Stochastic Optimization** => **Difficult** to compute its gradient and find the iteration that minimizes it
- **Practical solution:** Pick a random termination point to derive new complexity results

Main Results

Case 1: Martingale Difference Noise

(A1) There exists constants $c_0 \geq 0, c_1 > 0$ such that

$$c_0 + c_1 \langle \nabla V(\theta), h(\theta) \rangle \geq \|h(\theta)\|^2$$

(A2) Lyapunov fonction V is L-smooth **(non-convex)**

$$\|\nabla V(\theta) - \nabla V(\vartheta)\| \leq L\|\theta - \vartheta\|$$

(A3) $\{e_k\}_{k\geq 1}$ Martingale difference sequence such that

$$\mathbb{E}[e_{k+1}|\mathcal{F}_k] = 0$$

$$\mathbb{E}[\|e_{k+1}\|^2|\mathcal{F}_k] \leq \sigma_0^2 + \sigma_1^2\|h(\theta^{(k)})\|^2$$

For instance when X_k is i.i.d. similar to SGD setting

Non-Asymptotic Analysis [Karimi+, COLT 2019]

Theorem

Assume **(A1) (A2) (A3)** and $\gamma_{k+1} \leq (2c_1 L(1+\sigma_1^2))^{-1}$

Let $V_{0,k} := \mathbb{E}[V(\theta^{(0)}) - V(\theta^{(k+1)})]$, then for all $k \geq 0$

$$\mathbb{E}[\|h(\theta^{(K)})\|^2] \leq \frac{2c_1\left(V_{0,k} + \sigma_0^2 L \sum_{\ell=0}^{k} \gamma_{\ell+1}^2\right)}{\sum_{\ell=0}^{k} \gamma_{\ell+1}} + 2c_0$$

K is in $\{0, \cdots, k\}$ and distributed according to

$$\mathbb{P}(K = \ell) = \left(\sum_{i=0}^{k} \gamma_{i+1}\right)^{-1} \gamma_{\ell+1}$$

If $\gamma_k = (2c_1 L(1+\sigma_1^2)\sqrt{k})^{-1}$
Then the SA scheme finds within k iterations an

$$\mathcal{O}(c_0 + \log k/\sqrt{k})$$ stationary point

Main Results

Case 2: State-dependent Markov Noise

(A4) There exists constants $d_0 \geq 0, d_1 > 0$ such that

$$d_0 + d_1\|h(\theta)\| \geq \|\nabla V(\theta)\|$$

The mean field is **indirectly related** to $\nabla V(\theta)$
When $c_0 = d_0 = 0$ then the SA scheme is un-biased

(A5) There exists a Borel measure function $\hat{H} : \Theta \times \mathsf{X} \to \Theta$

$$\hat{H}_\theta(x) - P_\theta \hat{H}_\theta(x) = H_\theta(x) - h(\theta), \ \forall\ \theta \in \Theta, x \in \mathsf{X}.$$

Existence of a solution to Poisson equation

(A6) We assume $\|\hat{H}_\theta(x)\| \leq L_{PH}^{(0)}$, $\|P_\theta \hat{H}_\theta(x)\| \leq L_{PH}^{(0)}$

$$\sup_{x \in \mathsf{X}} \|P_\theta \hat{H}_\theta(x) - P_\vartheta \hat{H}_\vartheta(x)\| \leq L_{PH}^{(1)} \|\theta - \vartheta\|$$

Lipschitzness of $\hat{H}_\theta(x)$ satisfied if $H_\theta(x)$, P_θ are too

(A7) it holds $\|H_\theta(x) - h(\theta)\| \leq \sigma$

Uniformly bounded noise

Non-Asymptotic Analysis [Karimi+, COLT 2019]

Theorem

Assume **(A1) (A2)** and **(A4)-(A7)** and

$$\gamma_{k+1} \leq \gamma_k,\ \gamma_k \leq a\gamma_{k+1},\ \gamma_k - \gamma_{k+1} \leq a'\gamma_k^2,\ \gamma_1 \leq 0.5\big(c_1(L + C_h)\big)^{-1}$$

Then:

$$\mathbb{E}[\|h(\theta^{(K)})\|^2] \leq \frac{2c_1\big(V_{0,k} + C_{0,k} + (\sigma_0^2 L + C_\gamma)\sum_{\ell=0}^k \gamma_{\ell+1}^2\big)}{\sum_{\ell=0}^k \gamma_{\ell+1}} + 2c_0$$

Key idea is to use decompose noise using the Poisson equation in **(A5)**

Tightness of both Theorems

Lemma

Assume **(A3)-(A7)** and $h(\theta) = \nabla V(\theta)$, $c_0 = 0$ then:

$$\mathbb{E}[\|h(\theta^{(K)})\|^2] \geq \frac{V_{0,k} + C_{\mathrm{lb}}\sum_{\ell=0}^k \gamma_{\ell+1}^2}{\sum_{\ell=0}^k \gamma_{\ell+1}}$$

If $\gamma_k = c/\sqrt{k}$ then $\mathbb{E}[\|h(\theta^{(K)})\|^2] = \Omega(\log k/\sqrt{k})$

Applications: Online EM algorithm

Problem Setting

- Given a stream of i.i.d. data $\{Y_k\}_{k\geq 1} \qquad Y_k \sim \pi$
- We have vectors of data Y that are observed and Z that are latent
- We want to optimize the incomplete likelihood $g(y,\theta)$ w.r.t. θ
- **Exponential Family Distribution:**

$$f(z,y;\theta) := h(z,y)\exp\left(\langle S(z,y)|\phi(\theta)\rangle - \psi(\theta)\right)$$

Regularized Online EM [Cappé and Moulines, 2009]

- **E-step:** $\hat{s}_{k+1} = \hat{s}_k + \gamma_{k+1}\left(\bar{s}(y_{k+1},\theta^{(k)}) - \hat{s}_k\right)$ where $\bar{s}(y,\theta) = \mathbb{E}_\theta[S(z,y)|y]$
  - E-step is an SA update with drift term $H_{\hat{s}_k}(y_{k+1}) = \hat{s}_k - \bar{s}(y_{k+1},\bar{\theta}(\hat{s}_k))$ and $h(\hat{s}_k) = \hat{s}_k - \mathbb{E}_\pi[\bar{s}(y_{k+1},\bar{\theta}(\hat{s}_k))]$
- **M-step:** $\theta^{(k+1)} = \bar{\theta}(\hat{s}_{k+1})$ where $\bar{\theta}(s) := \arg\max_{\theta\in\Theta} \langle s|\phi(\theta)\rangle - \psi(\theta) - R(\theta)$

Lyapunov Function

$$V(s) := \mathbb{E}_\pi\left[\frac{\pi(y)}{g(y,\bar{\theta}(s))}\right] + R(\bar{\theta}(s))$$

Each step decreases the KL to the target distribution $\langle \nabla_s V(s), h(s)\rangle \leq 0$

Applications: Online EM algorithm

Fitting Gaussian Mixture Model (GMM)

- Fit $\{Y_k\}_{k\geq 1}$ in a GMM with $\theta = (\{w_m\}_{m=1}^{M-1}, \{\mu_m\}_{m=1}^{M})$
- The incomplete likelihood reads:

$$g(y;\theta) \propto \Big(1 - \sum_{m=1}^{M-1} \omega_m\Big)\exp\left(-\frac{(y-\mu_M)^2}{2}\right) + \sum_{m=1}^{M-1} \omega_m \exp\left(-\frac{(y-\mu_m)^2}{2}\right)$$

- And the regularizer: $R(\theta) = \epsilon \sum_{m=1}^{M} \left\{\mu_m^2/2 - \log(\omega_m)\right\} - \epsilon \log\left(1 - \sum_{m=1}^{M-1} \omega_m\right)$ with $\epsilon > 0$
- Assume: **(A8)** The samples are i.i.d. and $|Y_k| \leq \overline{Y}$

Corollary

Assume **(A8)** and $\gamma_k = (2c_1 L(1+\sigma_1^2)\sqrt{k})^{-1}$

Then, the ro-EM for GMM satisfies

$$\mathbb{E}[\|\nabla V(\hat{s}_K)\|^2] = \mathcal{O}(\log k/\sqrt{k})$$

Where the expectation is total (over K and observation law)

Applications: Policy Gradient algorithm

Markov Decision Process (MDP)

- MDP (S, A, R, P):
  - S is a finite set of state (state-space)
  - A is a finite set of action (action-space)
  - $R : S \times A \to [0, R_{\max}]$ is a reward function
  - P is the transition model, i.e., given an action $a \in A$, $P^a = \{P^a_{s,s'}\}$ is a matrix and $P^a_{s,s'}$ is the probability of transiting from the **s**-th state to the **s'**-th state upon taking action **a**
- We consider **parametric policy**: $\Pi_\theta(a, s)$: probability of taking action **a** in state **s**
- $\{(S_t, A_t)\}_{t\geq 1}$ forms a Markov chain with the following *transition probability*:

$$Q_\theta((s, a); (s', a')) := \Pi_\theta(a'; s')P^a_{s,s'}$$ whose *invariant distribution* is noted $v_\theta(s, a)$

Policy Optimization Problem

- **Goal:** Find a policy θ that maximizes the average reward defined as

$$J(\theta) := \sum_{s\in S, a\in A} v_\theta(s, a)R(s, a)$$

Applications: Policy Gradient algorithm

Online Policy Gradient algorithm

- We use a gradient algorithm to maximize the reward
- What is the gradient of $J(\theta)$? It can be verified [Sutton and Barto, 2018] that:

$$\nabla J(\theta) = \lim_{T\to\infty} \mathbb{E}_\theta\left[R(S_T, A_T) \sum_{i=0}^{T-1} \nabla \log \Pi_\theta(A_{T-i}; S_{T-i})\right]$$

- We use a numerically stable estimator [Baxter and Bartlett, 2001] by introducing a discount factor (thus biased estimator):

$$\widehat{\nabla}_T J(\theta) := R(S_T, A_T) \sum_{i=0}^{T-1} \lambda^i \nabla \log \Pi_\theta(A_{T-i}; S_{T-i})$$

- We update the policy on-the-fly with an online policy gradient update [Baxter and Bartlett, 2001]

$$G_{k+1} = \lambda G_k + \nabla \log \Pi_{\theta^{(k)}}(A_{k+1}; S_{k+1})$$

$$\theta^{(k+1)} = \theta^{(k)} + \gamma_{k+1} G_{k+1} R(S_{k+1}, A_{k+1})$$

- SA interpretation:

$$H_{\theta^{(k)}}(S_{k+1}, A_{k+1}, G_{k+1}) = G_{k+1} R(S_{k+1}, A_{k+1}) \quad \text{and} \quad h(\theta) = \lim_{T\to\infty} \mathbb{E}_{\Pi_\theta}\left[\widehat{\nabla}_T J(\theta)\right]$$

Applications: Policy Gradient algorithm

Convergence Analysis of the online PG algorithm

state-action feature vector

- Focus on **exponential family policy**: $\Pi_\theta(a;s) = \{\sum_{a' \in A} \exp(\langle \theta, x(s,a') \rangle)\}^{-1} \exp(\langle \theta, x(s,a) \rangle)$
- **(A9)** we have $\|x(s,a)\| \leq \bar{b}$
- **(A10)** $\{(S_t, A_t)\}_{t \geq 1}$ is ergodic with $\|Q_\theta^k - 1(v_\theta)^\top\| \leq \rho^k K_R$.The invariant distribution v_θ and its Jacobian are Lipschitz.
- We can proof smoothness of the objective function $J(\theta)$ and **(A1):** $(1-\lambda)^2\Gamma^2 + 2\langle \nabla J(\theta), h(\theta) \rangle \geq \|h(\theta)\|^2$

Corollary

Assume **(A9)-(A10)** and $\gamma_k = (2c_1 L(1+C_h)\sqrt{k})^{-1}$

Then, the online policy gradient finds a policy such that

$$\mathbb{E}[\|\nabla J(\theta^{(K)})\|^2] = \mathcal{O}\left((1-\lambda)^2\Gamma^2 + c(\lambda)\log k/\sqrt{k}\right)$$

where $c(\lambda) = \mathcal{O}(\frac{1}{1-\max(\rho,\lambda)})^2$

Where the expectation is total (over K and the space-action Markov chain)

- Shows a *variance-bias trade-off with* $\lambda \in (0,1)$
- *While setting* $\lambda \to 1$ *reduces the bias, it increases the variance in the converging term with* $c(\lambda) = \mathcal{O}\left((1-\lambda)^{-1}\right)$

Fast MLE Algorithms

Maximum Likelihood Estimation (MLE)

- The MLE problem is, given a model $g(Y, \theta)$ and some actual data Y, find the parameter θ which makes the data most likely:

$$\theta^{ML} := \arg\max_{\theta} \; g(Y, \theta)$$

- This problem is an **optimization problem**, which we could use any imaginable tool to solve
- In practice, it's often **hard** to get expressions for the **derivatives** needed by **gradient** methods
- **Expectation-Maximization (EM)** method is one popular and powerful way of proceeding, but not the only way. **It takes advantage of the latent data to complete the observations.**

EM Algorithm

[Dempster, Laird and Rubin, 1977]

- **E-step:** Given $\theta^{(k-1)}$ compute the surrogate quantity

$$Q(\theta, \theta^{(k-1)}) = \mathbb{E}_{p(z|y,\theta^{(k-1)})} \left[\log f(z, y, \theta)\right]$$

$$= \sum_{i=1}^{n} \mathbb{E}_{p(z_i|y_i,\theta^{(k-1)})} \left[\log f\left(z_i, y_i; \theta\right)\right]$$

- **M-step:** Maximize w.r.t. the parameter

$$\theta^{(k)} = \arg\max_{\theta \in \Theta} Q(\theta, \theta^{(k-1)})$$

Expectation

Conditional Distribution

Large Sum

Agenda

Risk Minimization

$$\mathcal{L}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \ell\left(y_i, M_{\boldsymbol{\theta}}\left(x_i\right)\right) \text{ or } \mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\ell\left(y, M_{\boldsymbol{\theta}}(x)\right)\right]$$

1

- Incremental Updates
- Constrained Minimization
- Applications to **Logistic Regression** and **Bayesian Deep Learning**

2

- Online Updates
- Biased Stochastic Approximation
- Applications to online **EM** and **Reinforcement Learning**

Maximum Likelihood

$$\log p(y, \theta) = \sum_{i=1}^{n} \log p_i(y_i, \theta)$$

3

- **EM** Algorithm
- **Exponential Family** Model
- **Incremental** Updates
- Applications to **Mixture** and **Topic Modeling**

4

- Applications to Pharmacology
- Stochastic Approximation of EM
- Mixed Effects Models
- 2 Variants

EM Method for Exponential Family

Updates

- **Exponential family:**

$$f\left(z_i, y_i; \theta\right) = h\left(z_i, y_i\right) \exp\left(\left\langle S\left(z_i, y_i\right) | \phi(\theta) \right\rangle - \psi(\theta)\right)$$

- **E-step**:

$$\bar{\mathbf{s}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \bar{\mathbf{s}}_i(\theta)$$

where:

$$\bar{\mathbf{s}}_i(\theta) = \int_{\mathbf{Z}} S\left(z_i, y_i\right) p\left(z_i | y_i; \theta\right) \mu\left(\mathrm{d}z_i\right)$$

- **M-step**:

$$\theta = \bar{\theta}(\bar{s}) = \arg\min_{\theta \in \Theta} \{ R(\theta) + \psi(\theta) - \langle s | \phi(\theta) \rangle \}$$

Limitations

- Even though the EM has appealing features:
  - Monotone in likelihood
  - Invariant w.r.t. parametrization
  - Numerically stable (well defined set)
- It is not applicable with the sheer size of today's data
- Approaches based on Stochastic Optimization:
  - [Neal and Hinton, 1998]: Incremental EM (iEM)
  - [Cappé and Moulines, 2009]: Online EM (sEM)
  - [Chen+, 2018]: Variance Reduced EM (sEM-VR)

Stochastic Optimization for EM Methods

General Formulation [Karimi+, NeurIPS 2019]

- Stochastic EM:

sE-step: $\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} - \gamma_{k+1}\left(\hat{\mathbf{s}}^{(k)} - \mathcal{S}^{(k+1)}\right)$

where γ_k is the stepsize and $\mathcal{S}^{(k+1)}$ is a proxy for $\bar{\mathbf{s}}\left(\boldsymbol{\theta}^{(k)}\right)$

- **M-step**:

$$\theta^{(k+1)} = \bar{\theta}(\hat{\mathbf{s}}^{(k+1)}) = \arg\min_{\theta\in\Theta}\{R(\theta) + \psi(\theta) - \left\langle \hat{\mathbf{s}}^{(k+1)} | \phi(\theta)\right\rangle\}$$

- We simplify the notations:

$$\bar{\mathbf{s}}_i^{(k)} := \bar{\mathbf{s}}_i\left(\boldsymbol{\theta}^{(k)}\right) = \int_{\mathsf{Z}} S\left(z_i, y_i\right) p\left(z_i | y_i; \hat{\boldsymbol{\theta}}^{(k)}\right) \mu\left(\mathrm{d}z_i\right)$$

$$\bar{\mathbf{s}}^{(k)} := \bar{\mathbf{s}}\left(\boldsymbol{\theta}^{(k)}\right) = \frac{1}{n}\sum_{i=1}^{n} \bar{\mathbf{s}}_I^{(k)}$$

$\ell(k) := m\lfloor k/m \rfloor$ First iteration number of the current epoch

(*iEM* [*NH*, 1998])	$\boldsymbol{\mathcal{S}}^{(k+1)} = \boldsymbol{\mathcal{S}}^{(k)} + \frac{1}{n}\left(\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(\tau_{i_k}^k)}\right)$	[1]
(*sEM* [*CM*, 2009])	$\boldsymbol{\mathcal{S}}^{(k+1)} = \bar{\mathbf{s}}_{i_k}^{(k)}$	[2]
(*sEM − VR* [*CZTZ.*, 2018])	$\boldsymbol{\mathcal{S}}^{(k+1)} = \bar{\mathbf{s}}^{(\ell(k))} + \left(\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(\ell(k))}\right)$	[3]
(*fiEM* [*KLMW.*, 2019])	$\boldsymbol{\mathcal{S}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + \left(\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(t_{i_k}^k)}\right)$	
	$\overline{\boldsymbol{\mathcal{S}}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + n^{-1}\left(\bar{\mathbf{s}}_{j_k}^{(k)} - \bar{\mathbf{s}}_{j_k}^{(t_{j_k}^k)}\right).$	[4]

Algorithm 3 sEM algorithms

Initialization: initializations $\hat{\boldsymbol{\theta}}^{(0)} \leftarrow 0$, $\hat{\mathbf{s}}^{(0)} \leftarrow \bar{\mathbf{s}}^{(0)}$, $K_{\max} \leftarrow$ max. iteration number.

Set the terminating iteration number, $K \in \{0, \ldots, K_{\max} - 1\}$, as a discrete r.v. with:

$$P(K = k) = \frac{\gamma_k}{\sum_{\ell=0}^{K_{\max}-1} \gamma_\ell}. \tag{42}$$

Iteration k: Given the current state of the chain $\psi_i^{(t-1)}$:

1. Draw index $i_k \in [\![1, n]\!]$ uniformly (and $j_k \in [\![1, n]\!]$ for fiEM).
2. Compute the surrogate sufficient statistics $\boldsymbol{\mathcal{S}}^{(k+1)}$ using [1] or [2] or [3] or [4]
3. Compute $\hat{\mathbf{s}}^{(k+1)}$ via the sE-step
4. Compute $\boldsymbol{\theta}^{(k+1)}$ via the M-step

Return: $\boldsymbol{\theta}^{(K)}$.

Global Convergence of iEM

Lemma

Under **(A1)-(A4),** define $e_i\left(\boldsymbol{\theta};\boldsymbol{\theta}'\right) := Q_i\left(\boldsymbol{\theta};\boldsymbol{\theta}'\right) - \mathcal{L}_i(\boldsymbol{\theta})$
We have

$$\left\|\nabla e_i\left(\boldsymbol{\theta};\boldsymbol{\theta}'\right) - \nabla e_i\left(\overline{\boldsymbol{\theta}};\boldsymbol{\theta}'\right)\right\| \leq \mathrm{L}_e\|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\|$$

where $\mathrm{L}_e := C_\phi C_{\mathrm{Z}} \mathrm{L}_p + C_{\mathrm{S}} \mathrm{L}_\phi$

Theorem

Under **(A1)-(A4)** for the iEM **[1]** for any $K_{\max} \geq 1$

$$\mathbb{E}\left[\left\|\nabla\overline{\mathcal{L}}\left(\boldsymbol{\theta}^{(K)}\right)\right\|^2\right] \leq n\frac{2L_e}{K_{\max}}\mathbb{E}\left[\overline{\mathcal{L}}\left(\boldsymbol{\theta}^{(0)}\right) - \overline{\mathcal{L}}\left(\boldsymbol{\theta}^{(K_{\max})}\right)\right]$$

where L_e is defined above and K is a uniform random variable on $[0, K_{\max} - 1]$ and independent of the $\{i_k\}_{k=0}^{K_{\max}}$

Through the Lens of MISO…

Stochastic EM as Scaled Gradient Methods

Through the Lens of Stochastic Approximation…

$$\min_{\mathbf{s}\in S} V(\mathbf{s}) := \overline{\mathcal{L}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) = \mathrm{R}(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\overline{\boldsymbol{\theta}}(\mathbf{s}))$$

Lemma

Under **(A1)-(A4),** we have

$$\left\|\overline{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \overline{\mathbf{s}}_i\left(\overline{\boldsymbol{\theta}}\left(\mathbf{s}'\right)\right)\right\| \leq \mathbf{L_s}\left\|\mathbf{s}-\mathbf{s}'\right\|$$

$$\left\|\nabla V(\mathbf{s}) - \nabla V\left(\mathbf{s}'\right)\right\| \leq \mathbf{L_V}\left\|\mathbf{s}-\mathbf{s}'\right\|$$

where $\mathrm{L_s} := C_{\mathrm{Z}}\mathrm{L}_p\mathrm{L}_\theta$ and $\mathrm{L}_V := v_{\max}\left(1+\mathrm{L_s}\right) + \mathrm{L}_B C_{\mathrm{S}}$

Theorem (sEM-VR)

There exists a constant $\mu \in (0,1)$ such that if

$$\overline{L}_v := \max(L_V, L_s) \qquad \gamma = \frac{\mu v_{\min}}{\overline{L}_v n^{2/3}} \qquad m = \frac{n}{2\mu^2 v_{\min}^2 + \mu}$$

Then:

$$\mathbb{E}\left[\left\|\nabla V\left(\hat{s}^{(K)}\right)\right\|^2\right] \leq n^{\frac{2}{3}}\frac{2\overline{L}_v}{\mu K_{\max}}\frac{v_{\max}^2}{v_{\min}^2}\mathbb{E}\left[V\left(\hat{s}^{(0)}\right) - V\left(\hat{s}^{(K_{\max})}\right)\right]$$

Theorem (fiEM)

There exists a constant $\mu \in (0,1)$ such that if

$$\overline{L}_v := \max(L_V, L_s) \qquad \gamma = \frac{v_{\min}}{\alpha\overline{L}_v n^{2/3}} \qquad \alpha := \max(6, 1+4v_{\min})$$

Then:

$$\mathbb{E}\left[\left\|\nabla V\left(\hat{\boldsymbol{s}}^{(K)}\right)\right\|^2\right] \leq n^{\frac{2}{3}}\frac{\alpha^2\overline{L}_v}{K_{\max}}\frac{v_{\max}^2}{v_{\min}^2}\mathbb{E}\left[V\left(\hat{\boldsymbol{s}}^{(0)}\right) - V\left(\hat{\boldsymbol{s}}^{(K_{\max})}\right)\right]$$

Numerical Applications

Gaussian Mixture Models (GMM)

- Fit a GMM model to a set of n observations
- Each of M components with unit variance
- The complete log likelihood reads:

$$\log f\left(z_i, y_i; \boldsymbol{\theta}\right) = \sum_{m=1}^{M} 1_{\{m\}}\left(z_i\right)\left[\log\left(\omega_m\right) - \mu_m^2/2\right] + \sum_{m=1}^{M} 1_{\{m\}}\left(z_i\right)\mu_m y_i + \text{ constant}$$

$$\theta := (\omega, \mu) \qquad \omega = \{\omega_m\}_{m=1}^{M-1} \qquad \mu = \{\mu_m\}_{m=1}^{M}$$

- Penalization used:

$$\mathrm{R}(\boldsymbol{\theta}) = \frac{\delta}{2}\sum_{m=1}^{M}\mu_m^2 - \log \mathrm{Dir}(\boldsymbol{\omega}; M, \epsilon)$$

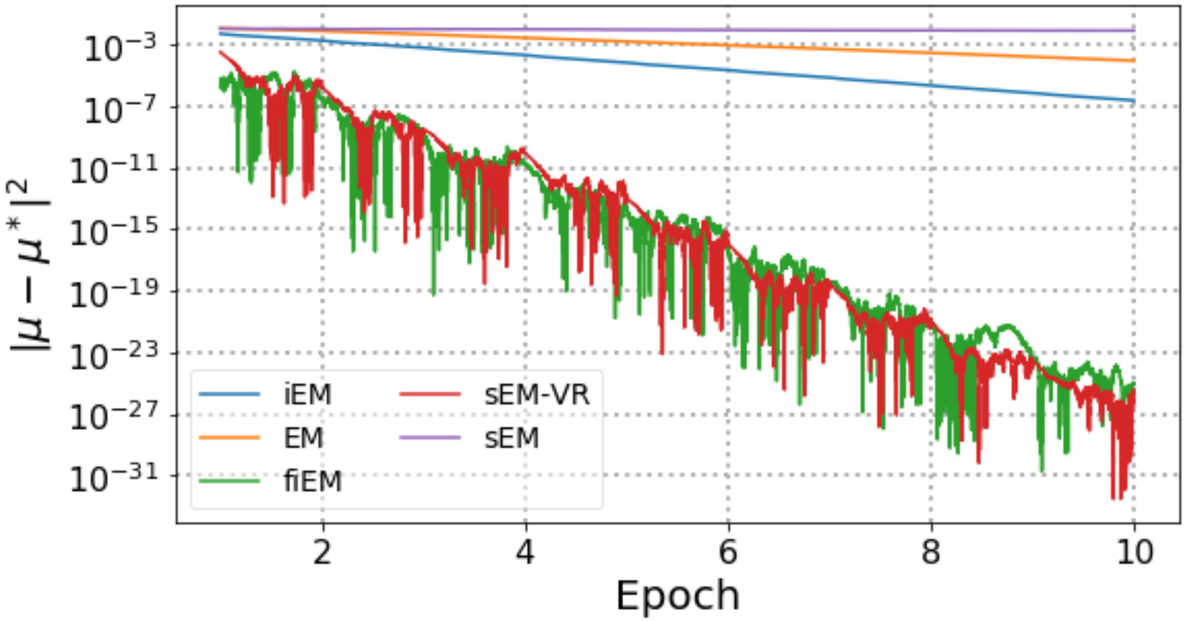
Experiments

- **Numerical:** M=2 and $\mu_1 = -\mu_2 = 0.5$

- **Fixed sample size:** size $n = 10^4$ and run to get μ^*

Stepsize for sEM $\gamma_k = 3/(k+10)$

Stepsize for sEM-VR and fiEM prop. to $1/n^{2/3}$



Numerical Applications

Gaussian Mixture Models (GMM)

- Fit a GMM model to a set of n observations
- Each of M components with unit variance
- The complete log likelihood reads:

$$\log f\left(z_i, y_i; \boldsymbol{\theta}\right) = \sum_{m=1}^{M} 1_{\{m\}}\left(z_i\right)\left[\log\left(\omega_m\right) - \mu_m^2/2\right] + \sum_{m=1}^{M} 1_{\{m\}}\left(z_i\right)\mu_m y_i + \text{ constant}$$

$$\theta := (\omega, \mu) \qquad \omega = \{\omega_m\}_{m=1}^{M-1} \qquad \mu = \{\mu_m\}_{m=1}^{M}$$

- Penalization used:

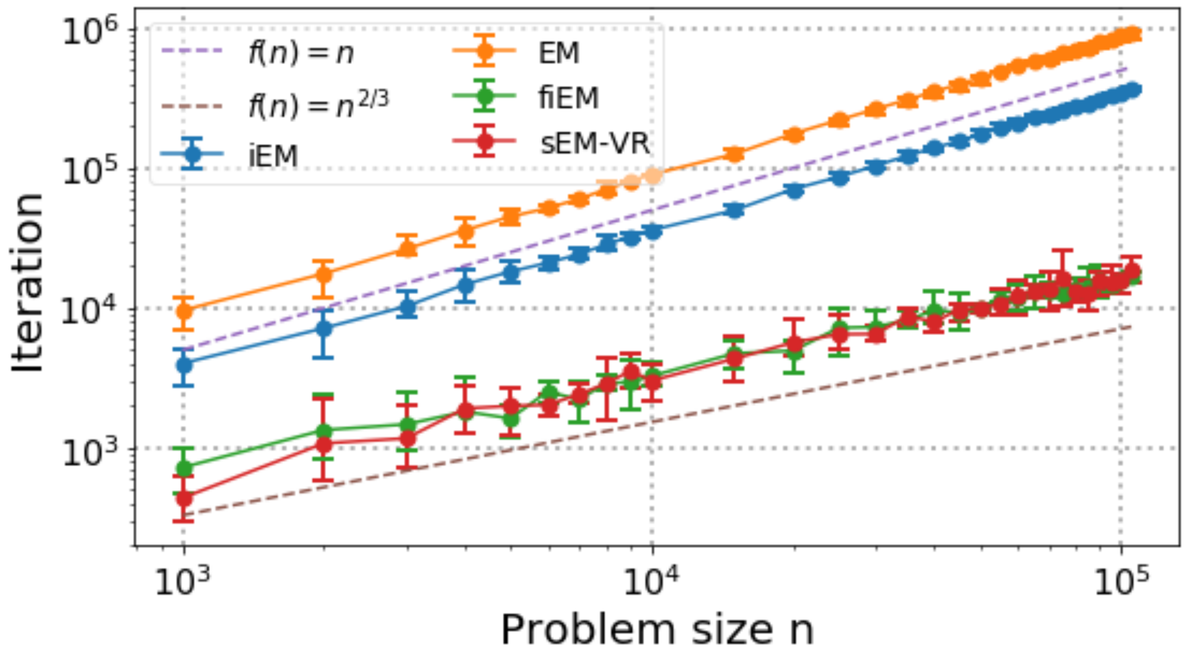
$$\mathrm{R}(\boldsymbol{\theta}) = \frac{\delta}{2}\sum_{m=1}^{M}\mu_m^2 - \log \mathrm{Dir}(\boldsymbol{\omega}; M, \epsilon)$$

Experiments

- **Numerical:** M=2 and $\mu_1 = -\mu_2 = 0.5$

- **Fixed sample size:** size $n = 10^4$ and run to get μ^*
  Stepsize for sEM $\gamma_k = 3/(k+10)$
  Stepsize for sEM-VR and fiEM prop. to $1/n^{2/3}$

- **Varying sample size:** nb. Iterations required to reach a precision of 10^{-3} from $n = 10^3$ to $n = 10^5$



Numerical Applications

Probabilistic Latent Semantic Analysis

- Consider D documents with terms from a vocabulary of size V.

- The goal of pLSA is to classify the documents into K topics which is modeled as a latent variable associated with each token $z_i \in [1, K]$

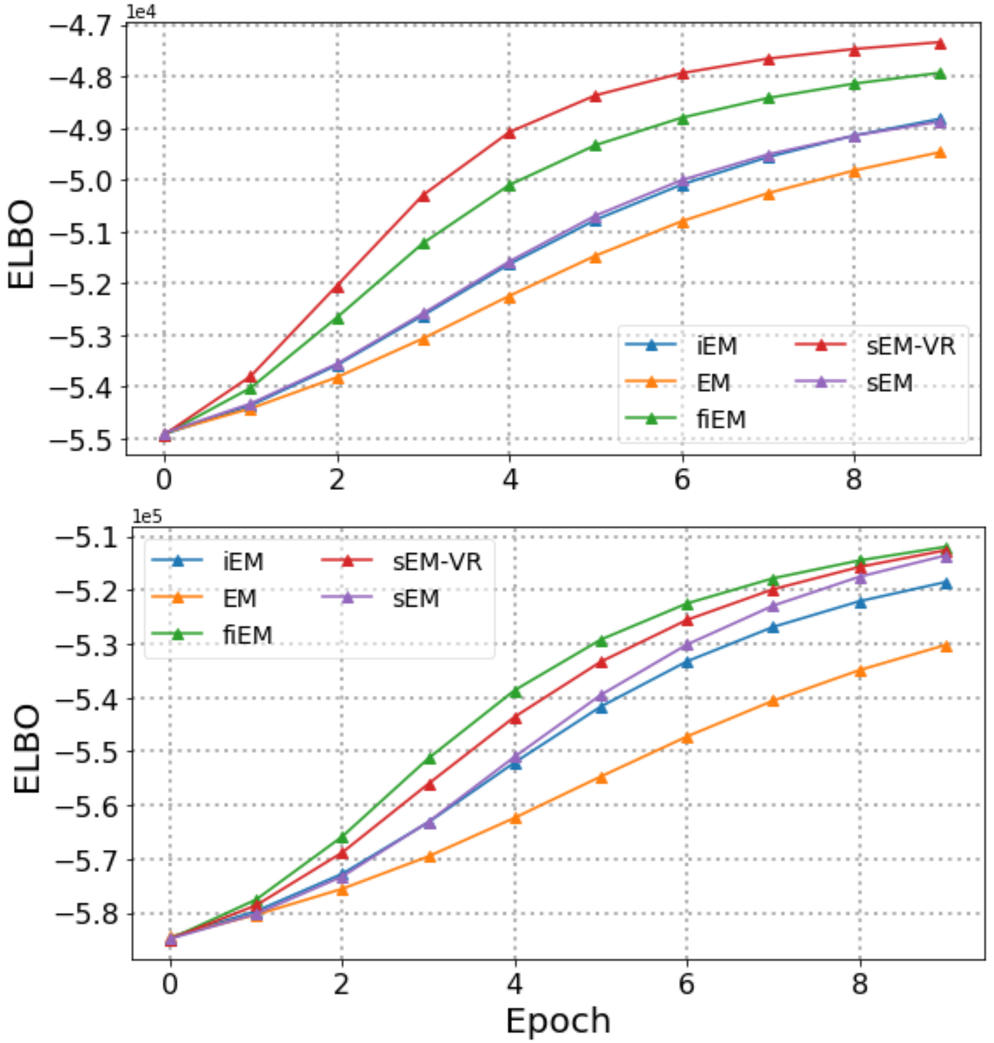
$$\log f(z_i, y_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \sum_{d=1}^{D} \log(\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})}) \mathbb{1}_{\{k,d\}}(z_i, y_i^{(\mathrm{d})}) + \sum_{k=1}^{K} \sum_{v=1}^{V} \log(\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})}) \mathbb{1}_{\{k,v\}}(z_i, y_i^{(\mathrm{w})})$$

- Penalization used:

$$\mathrm{R}(\boldsymbol{\theta}^{(\mathrm{t}|\mathrm{d})}, \boldsymbol{\theta}^{(\mathrm{w}|\mathrm{t})}) = -\log \mathrm{Dir}(\boldsymbol{\theta}^{(\mathrm{t}|\mathrm{d})}; K, \alpha') - \log \mathrm{Dir}(\boldsymbol{\theta}^{(\mathrm{w}|\mathrm{t})}; V, \beta')$$

$$\boldsymbol{\theta} := (\boldsymbol{\theta}^{(\mathrm{t}|\mathrm{d})}, \boldsymbol{\theta}^{(\mathrm{w}|\mathrm{t})})$$

Experiments



Agenda

Risk Minimization

$$\mathcal{L}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \ell\left(y_i, M_{\boldsymbol{\theta}}\left(x_i\right)\right) \text{ or } \mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\ell\left(y, M_{\boldsymbol{\theta}}(x)\right)\right]$$

1

- Incremental Updates
- Constrained Minimization
- Applications to Logistic Regression and Bayesian Deep Learning

2

- Online Updates
- Biased Stochastic Approximation
- Applications to online EM and Reinforcement Learning

Maximum Likelihood

$$\log p(y, \theta) = \sum_{i=1}^{n} \log p_i(y_i, \theta)$$

3

- EM Algorithm
- Exponential Family Model
- Incremental Updates
- Applications to Mixture and Topic Modeling

4

- Applications to **Pharmacology**
- Stochastic **Approximation** of EM
- **Mixed Effects** Models
- 2 **Variants**

Pharmacology and Mixed Effects Modeling

- **Pharmacokinetics:** Evolution of drug in the body
  - Outcome can be plasma drug concentration
- **Pharmacodynamics:** Reaction of the body to a drug
  - Evolution of number of seizures after treatment

Settings and Notations

- **Population approach:** Consider n individuals. Vector of measurements for each individual $y_i = (y_{ij}, 1 \leq j \leq n_i)$
- **Latent Data Model:** The latent variables are called 'individual parameters' ψ_i
- **Parametrized Hierarchical model:**

$$y_i \sim p(y_i|\psi_i, \theta) \qquad \psi_i \sim p(\psi_i, \theta)$$

- **Mixed Effects Model:** The individual parameters can be decomposed as follows $\psi_i = G(\beta, \eta_i)$

β : Population parameter η_i : Random effects $\eta_i \sim \mathcal{N}(0, \Omega)$

Continuous data model

- Continuous, nonlinear and mixed effects model

$$y_{ij} = f(t_{ij}; \psi_i) + \epsilon_{ij}$$

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \qquad \psi_i = \beta + \eta_i \Rightarrow \psi_i \sim \mathcal{N}(\beta, \Omega)$$
$$\theta = (\beta, \Omega, \sigma)$$

Non Continuous data model

- No analytical relationship between observations and individual parameters
- **Repeated time-to-event models**

$$\mathbb{P}(T_{ij} > t | T_{i,j-1} = t_{i,j-1}) = e^{-\int_{t_{i,j-1}}^{t} h(u,\psi_i)\mathrm{d}u}$$

EM Algorithm

Updates

- **E-step:** Given $\theta^{(k-1)}$:

$$Q(\theta, \theta^{(k-1)}) = \sum_{i=1}^{n} \mathbb{E}_{p(\psi_i | y_i, \theta^{(k-1)})} [\log f(\psi_i, y_i, \theta)]$$

- **M-step:** Maximize w.r.t. the parameter

$$\theta^{(k)} = \arg\max_{\theta \in \Theta} Q(\theta, \theta^{(k-1)})$$

SAEM Algorithm

Updates [Delyon+, 1999]

- **E-step:** Given $\theta^{(k-1)}$:
  - **Simulation Step:** $\psi_i^{(k)} \sim p(\psi_i | y_i, \theta^{(k-1)})$
  - **Stochastic Approximation of** $Q(\theta, \theta^{(k-1)})$ **:**

$$Q^{(k)}(\theta) = Q^{(k-1)}(\theta) + \gamma_k \left(\sum_{i=1}^{n} \log f(\psi_i^{(k)}; y_i, \theta) - Q^{(k-1)}(\theta) \right)$$

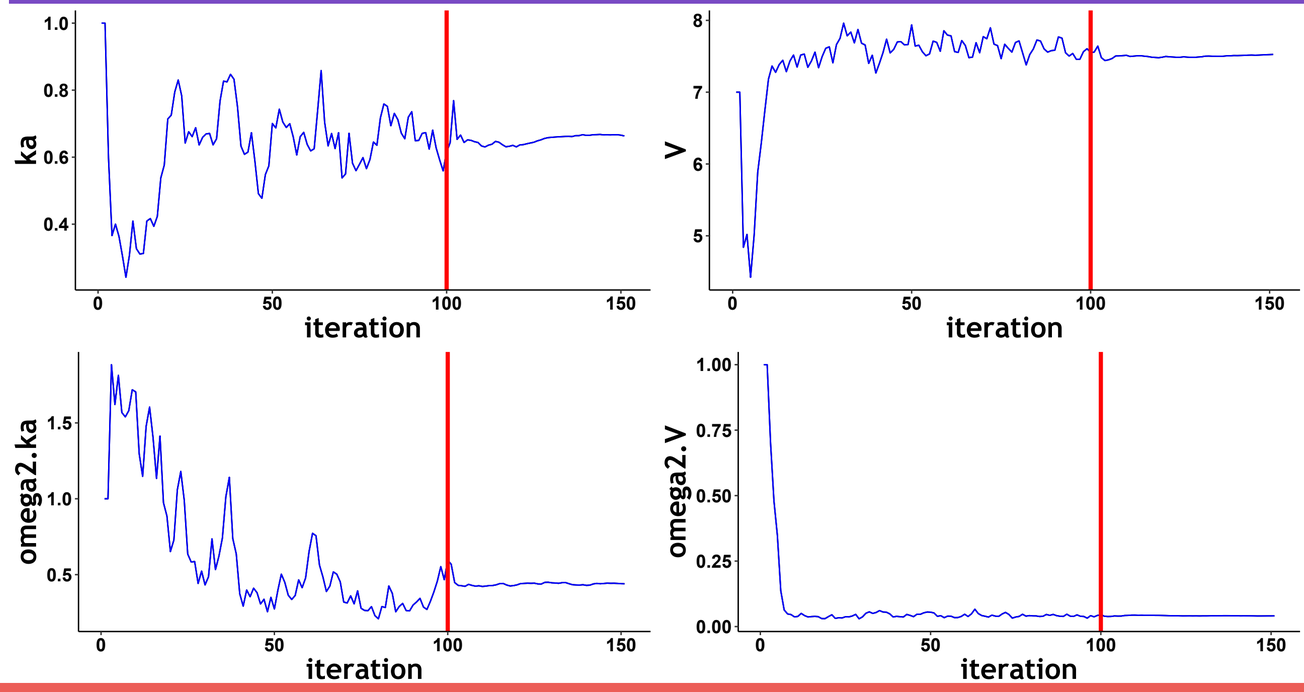
- **M-step:** Maximize w.r.t. the parameter

$$\theta^{(k)} = \arg\max_{\theta \in \Theta} Q^{(k)}(\theta)$$

Speed up the sampling procedure to speed up the MLE algorithm

How is it behaving?

- For a simple pharmacokinetics (PK) model
- In practice the stepwise is equal to 1 during the first K1 iterations then decreases in $1/k^{\alpha}$



Posterior Sampling Procedures

Metropolis-Hastings (MH)

- Sampling a candidate from a proposal

$$\psi^c \sim q(\psi|\psi^{(t-1)})$$

- Compute MH ratio

$$\alpha(\psi_i^{(t-1)}, \psi_i^c) = \frac{p(\psi_i^c|y_i)}{p(\psi_i^{(t-1)}|y_i)} \frac{q_i(\psi_i^{(t-1)}|\psi_i^c)}{q_i(\psi_i^c|\psi_i^{(t-1)})}$$

- Accept or reject with probability $\min(1, \alpha(\psi_i^c, \psi_i^{(t-1)}))$

Can Be SLOW

Metropolis Adjusted Langevin [Roberts+, 1998]

- Using the gradient of the target distribution

$$\psi_i^c \sim \mathcal{N}(\psi_i^{(t)} - \gamma_t \nabla \log \pi(\psi_i^{(t)}), 2\gamma_t)$$

- Special case of RWM [Ma+, 2015] with covariance matrix that is diagonal and isotropic

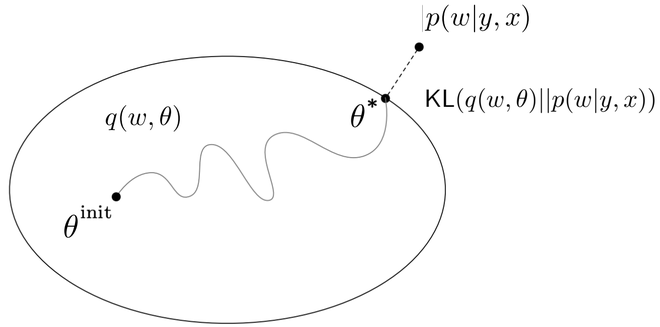
COSTLY and NEED TUNING

The No-U-Turn Samplers [Hoffman+, 2011]

- Adaptive and automated Hamiltonian Monte Carlo
- Using Hamiltonian dynamics
- State-of-the-art sampler

VERY COSTLY

The Variational MCMC [de Freitas, 2013]



Need to design **efficient** and easy-to-**implement** proposal based on the **covariance structure** of the latent variables

Efficient Metropolis-Hastings Procedure

[Karimi and Lavielle, BAYSM 2018]

- **Maximum A Posteriori**

$$\hat{\psi}_i = \arg\max_{\psi_i} p(\psi_i|y_i,\theta) = \arg\max_{\psi_i} p(y_i|\psi_i,\theta)p(\psi_i,\theta)$$

- **General Data Models**
  - Compute the Laplace Approximation of the incomplete likelihood

$$g(y_i,\theta) = \int e^{\log f(y_i,\psi_i,\theta)} \mathrm{d}\psi_i$$

  - Taylor expansion of the complete log likelihood around the MAP

$$\log p(\hat{\psi}_i|y_i,\theta) \approx -\frac{p}{2}\log 2\pi - \frac{1}{2}\log\left(\left|-\nabla^2 \log p(y_i,\hat{\psi}_i,\theta)\right|\right)$$

Proposal

$$\mathcal{N}(\mu_i,\Gamma_i) \qquad \mu_i = \hat{\psi}_i$$

$$\Gamma_i = -\nabla^2 \log p(y_i,\hat{\psi}_i,\theta)^{-1} = -\left(\nabla^2 \log p(y_i|\hat{\psi}_i,\theta) + \Omega^{-1}\right)^{-1}$$

- **Continuous Data Models**
  - Taylor expansion of the structural model around the MAP

$$f(\psi_i) \approx f(\hat{\psi}_i) + \mathsf{J}_f^{\psi_i}(\hat{\psi}_i)(\psi_i - \hat{\psi}_i)$$

  - Relation between observation and individual parameter is now linear and the posterior is a tractable Normal distribution

Proposal

$$\mathcal{N}(\mu_i,\Gamma_i) \qquad \mu_i = \hat{\psi}_i$$

$$\Gamma_i = \left(\frac{\mathsf{J}_f^{\psi_i}(\hat{\psi}_i)^\top \mathsf{J}_f^{\psi_i}(\hat{\psi}_i)}{\sigma^2} + \Omega^{-1}\right)^{-1}$$

f-SAEM Algorithm

[Karimi+, CSDA 2019]

- **E-step:** Given $\theta^{(k-1)}$:
  - **Simulation Step:** $\psi_i^{(k)} \sim p(\psi_i|y_i, \theta^{(k-1)})$
  - **Stochastic Approximation of** $Q(\theta, \theta^{(k-1)})$ **:**

$$Q^{(k)}(\theta) = Q^{(k-1)}(\theta) + \gamma_k \left(\sum_{i=1}^{n} \log f(\psi_i^{(k)}; y_i, \theta) - Q^{(k-1)}(\theta) \right)$$

- **M-step:** Maximize w.r.t. the parameter

$$\theta^{(k)} = \arg\max_{\theta \in \Theta} Q^{(k)}(\theta)$$

nlme-MH Algorithm

Algorithm 5.3 The nlme-IMH algorithm

Initialization: Initialize the chain sampling $\psi_i^{(0)}$ from some initial distribution ξ_i .

- Compute the MAP estimate:

$$\hat{\psi}_i = \arg\max_{\psi_i \in \mathbb{R}^p} p_i(\psi_i|y_i, \theta) \ .$$

- Compute the covariance matrix Γ_i using the corresponding proposal.

Iteration t: Given the current state of the chain $\psi_i^{(t-1)}$:

1. Sample a candidate ψ_i^c from a the independent proposal $\mathcal{N}(\hat{\psi}_i, \Gamma_i)$ denoted $q_i(\cdot|\hat{\psi}_i)$.
2. Compute the MH ratio:

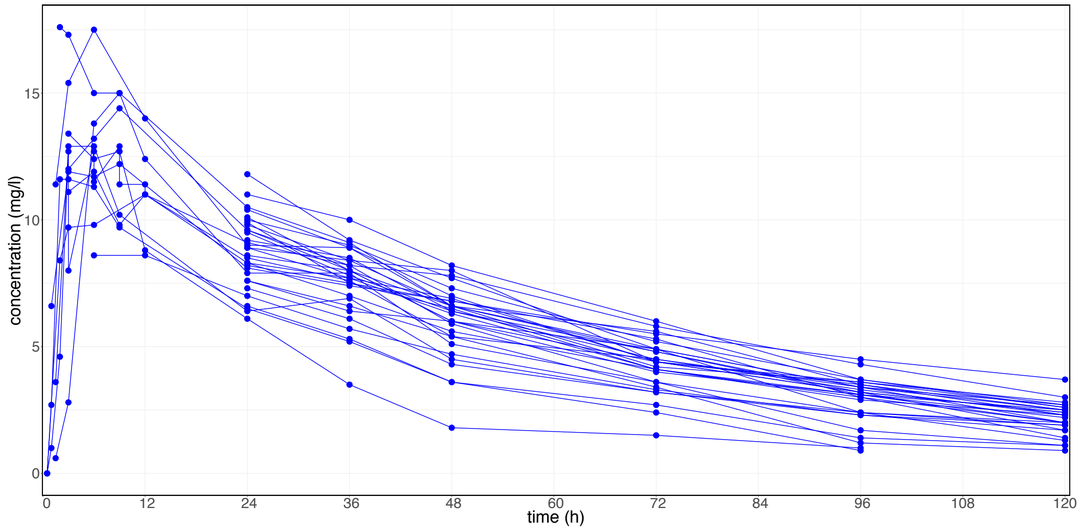
$$\alpha(\psi_i^{(t-1)}, \psi_i^c) = \frac{p_i(\psi_i^c|y_i, \theta)}{p_i(\psi_i^{(t-1)}|y_i, \theta)} \frac{q_i(\hat{\psi}_i|\psi_i^c)}{q_i(\psi_i^c|\hat{\psi}_i)} \ .$$

3. Set $\psi_i^{(t)} = \psi_i^c$ with probability $\min(1, \alpha(\psi_i^c, \psi_i^{(t-1)}))$ (otherwise, keep $\psi_i^{(t)} = \psi_i^{(t-1)}$).

Numerical Applications

Warfarin Data

- 32 healthy volunteers received a 1.5 mg/kg single oral dose of warfarin, an anticoagulant normally used in the prevention of thrombosis



- **Goal:** fit a PK model on this dataset
  - Compute population parameters
  - Obtain a fitted model for predictive tasks

PK Model

- Continuous nonlinear mixed effects model

$$y_{ij} = f(t_{ij}, \psi_i) + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

- Lognormally distributed individual parameters

$$\log(\psi_i) \sim \mathcal{N}(\log(\psi_{\text{pop}}), \omega_\psi^2)$$

- One-compartment PK model for oral administration, assuming first-order absorption and linear elimination processes:

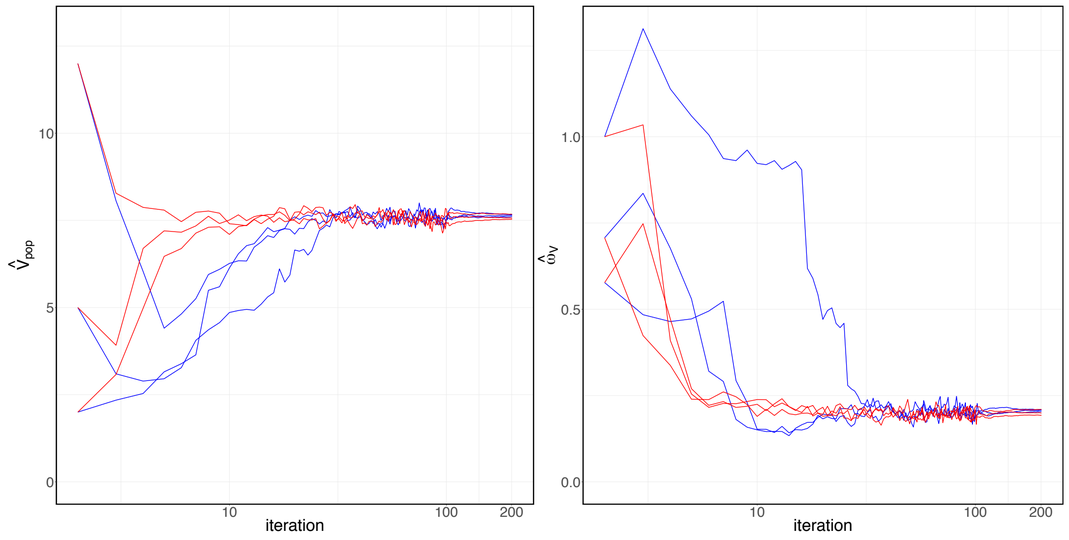
$$f(t, ka, V, k) = \frac{D\, ka}{V(ka - k)}(\mathrm{e}^{-ka\, t} - \mathrm{e}^{-k\, t})$$

$$\psi_i = (ka_i, V_i, k_i) \qquad \theta = (ka_{\text{pop}}, \cdots, \omega_{ka}, \cdots, \sigma)$$

Numerical Applications

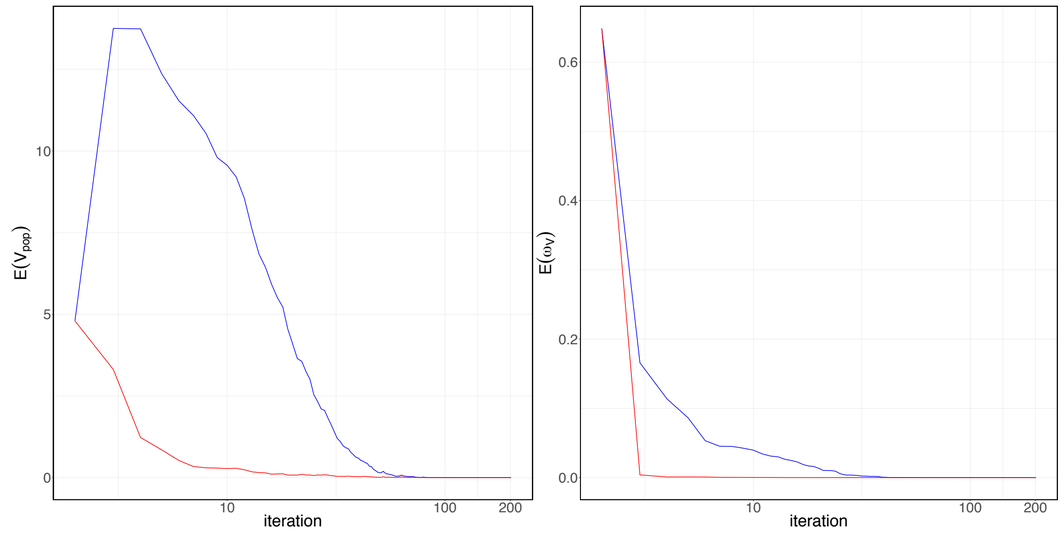
Warfarin Data: MLE Convergence

- **Single run MLE**
  - Three different initialization
  - K1 = 100, K2 = 100
  - 6 MCMC transitions per iteration



- **Monte Carlo Study**
  - M = 50 synthetic datasets
  - M runs of K iterations to obtain vector of ML estimates

$$E_k(\ell) = \frac{1}{M} \sum_{m=1}^{M} \left(\theta_k^{(m)}(\ell) - \theta_K^{(m)}(\ell) \right)^2$$



- Estimation of V_{pop} and ω_V
- Reference (RWM) in Blue and f-SAEM in Red

Numerical Applications

Time-to-event Data Model

- **Time-to-event Data Model**

$$\mathbb{P}(T_{ij} > t | T_{i,j-1} = t_{i,j-1}) = e^{-\int_{t_{i,j-1}}^{t} h(u,\psi_i)\mathrm{d}u}$$

- **Weibull model** for time-to-event data. Hazard function is defined as

$$h(t,\psi_i) = \frac{\beta_i}{\lambda_i}\left(\frac{t}{\lambda_i}\right)^{\beta_i - 1}$$

- Two parameters are independent and log normally distributed

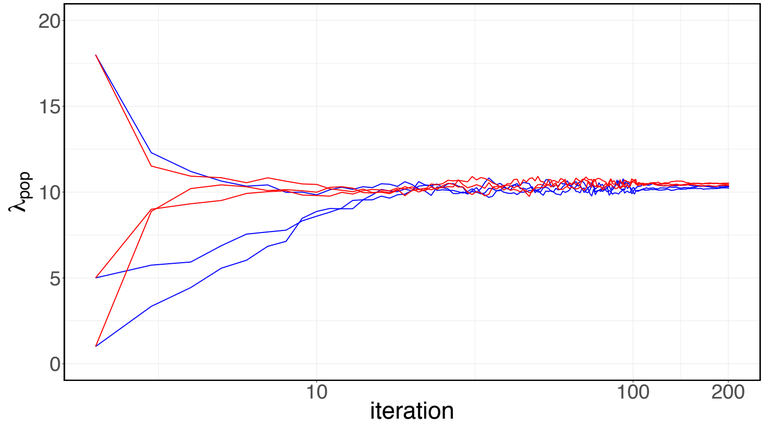
$$\log(\lambda_i) \sim \mathcal{N}(\log(\lambda_{\text{pop}}), \omega_\lambda^2)$$
$$\log(\beta_i) \sim \mathcal{N}(\log(\beta_{\text{pop}}), \omega_\beta^2)$$

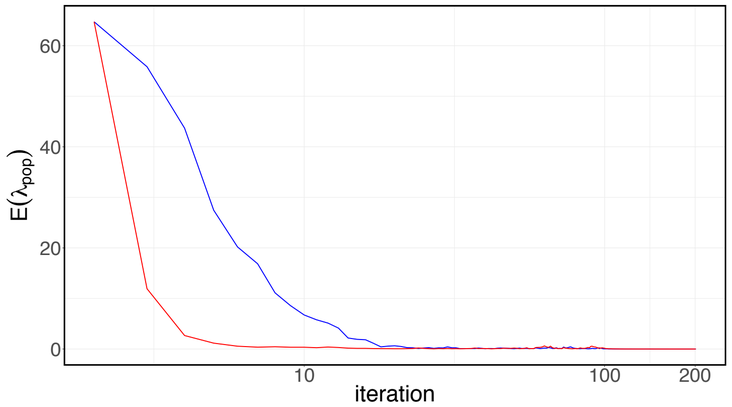
- Vector of parameters to estimate

$$\theta = (\lambda_{\text{pop}}, \beta_{\text{pop}}, \omega_\lambda, \omega_\beta)$$

Experiments

- Synthetic data: n= 100 individuals and right censoring time of 20
- Posterior sampling: comparison between reference RWM in blue and nlme-IMH in red





iSAEM Algorithm

Updates for exponential family

- **E-step:** Given $\theta^{(k-1)}$:
  - **Simulation Step:** $\psi_i^{(k)} \sim p(\psi_i|y_i, \theta^{(k-1)})$
  - **Stochastic Approximation of** $\bar{s}(y, \bar{\theta}(\hat{s}_{k-1}))$**:**

$$\hat{s}_i^{(k)} = \hat{s}_i^{(k-1)} + \gamma_k \left(S(\psi_i^{(k)}, y_i) - \hat{s}_i^{(k-1)} \right)$$

- **M-step:** Maximization function

$$\theta^{(k)} = \bar{\theta}(\hat{s}_k) \qquad \hat{s}_k = \left(\hat{s}_1^{(k)}, \cdots, \hat{s}_n^{(k)} \right)$$

$$\bar{\theta}(s) := \arg\max_{\theta \in \Theta} \langle s | \phi(\theta) \rangle - \psi(\theta) - R(\theta)$$

Incremental Updates

- **Simulation step:** Sample latent variables only for a mini batch of indices sampled uniformly

$$I_k \sim \{A \subset [1, n], \mathrm{card}(A) = p\}$$

- **Update:** Minibatch of sufficient statistics component are updated. The others remain unchanged

$$\hat{s}_i^{(k)} = \begin{cases} \hat{s}_i^{(k-1)} + \gamma_k (S_i(\psi_i^{(k)}, y_i) - \hat{s}_i^{(k-1)}) & \text{if } i \in I_k. \\ \hat{s}_i^{(k-1)} & \text{otherwise.} \end{cases}$$

- **Maximization** remains unchanged

Numerical Applications

Gaussian Mixture Models (GMM)

- Fit a GMM model to a set of n observations
- Each of M components with unit variance
- The complete log likelihood reads:

$$\log f\left(z_i, y_i; \boldsymbol{\theta}\right) = \sum_{m=1}^{M} 1_{\{m\}}\left(z_i\right)\left[\log\left(\omega_m\right) - \mu_m^2/2\right] + \sum_{m=1}^{M} 1_{\{m\}}\left(z_i\right)\mu_m y_i + \text{ constant}$$

$$\theta := (\omega, \mu)$$

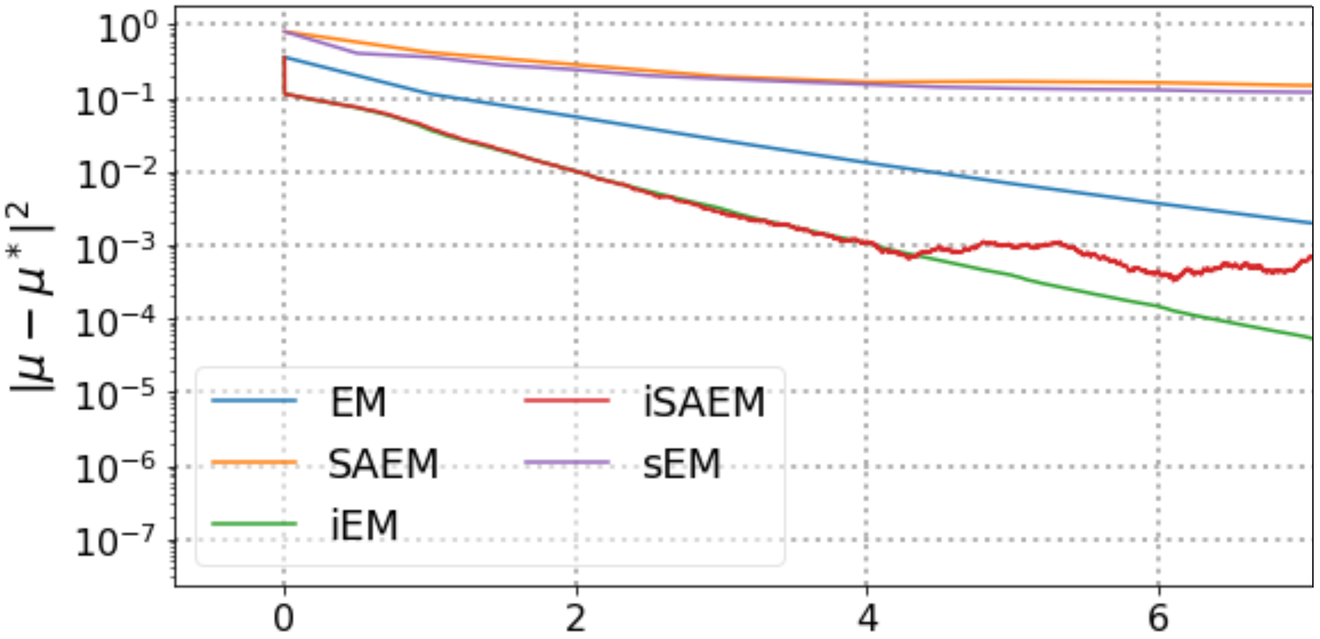
$$\omega = \{\omega_m\}_{m=1}^{M-1}$$

$$\mu = \{\mu_m\}_{m=1}^{M}$$

- Penalization used: $\mathrm{R}(\boldsymbol{\theta}) = \frac{\delta}{2}\sum_{m=1}^{M}\mu_m^2 - \log \mathrm{Dir}(\boldsymbol{\omega}; M, \epsilon)$

Experiments

- Numerical: GMM with M=2 and $\mu_1 = -\mu_2 = 0.5$
- **Fixed sample size:** size $n = 10^3$ and run to get μ^*
  Stepsize for sEM $\gamma_k = 3/(k+10)$
  Stepsize for iSAEM $\gamma_k = 1/k^{0.6}$
- Compare to iEM, sEM and Batch EM



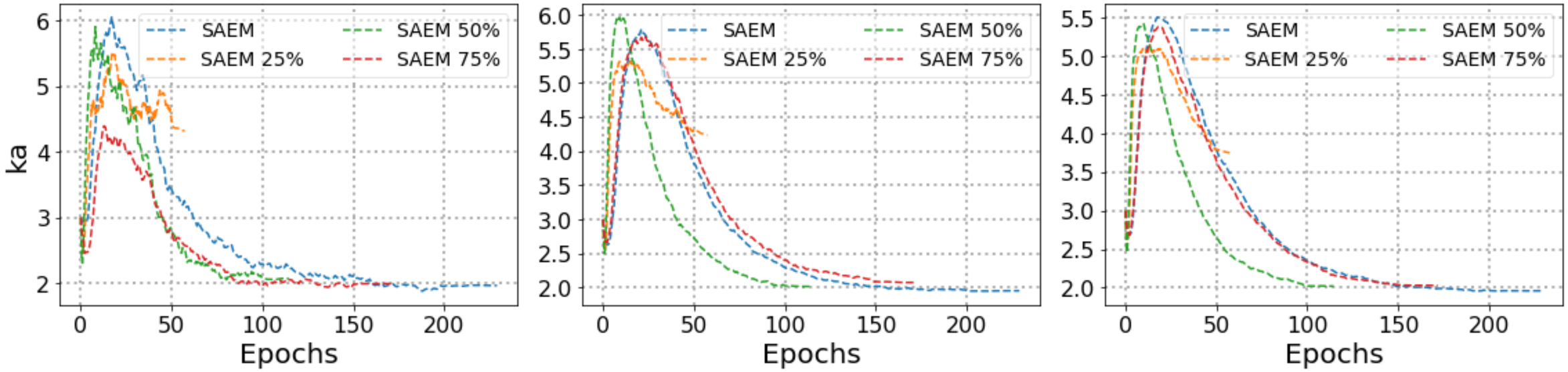
Numerical Applications

One-compartment PK Model

$$f(t, ka, V, k) = \frac{D\, ka}{V(ka - k)}(\mathrm{e}^{-ka\, t} - \mathrm{e}^{-k\, t}) \qquad \log(\psi_i) \sim \mathcal{N}(\log(\psi_{\text{pop}}), \omega_\psi^2) \qquad \theta = (ka_{\text{pop}}, \cdots, \omega_{ka}, \cdots, \sigma)$$

Synthetic data

- Simulate observations for $n = 10^3$ individuals with $n_i = 5$
- Stepsizes set to $\gamma_k = 1$ for K1 = 200 and $\gamma_k = 1/k$ for K2 = 50
- Run for three different size of MC batch size: 1, 10 and 20



Conclusion

Take-Aways

- We derived **incremental** and **online** methods for the optimization problem in machine learning.
  - For either **Empirical** or **Expected** risk minimization problems
  - When the objective function is a likelihood or not
  - For latent data models

- We conducted **finite-time analysis** of these methods for **nonconvex** loss functions and **non necessarily gradient** methods.

- **Applications** to several models of interest in machine learning with **rigorous verification** of the assumptions.

Perspectives

- Incremental algorithms: choice of the indices at each iteration.
  - **Optimal sampling strategies**: **[Le Roux+, 2012]** or **[Horvath and Richatrik, 2018]**.
  - **Optimal mini-batch size** of stochastic and incremental algorithms. See **[Gower+, 2019]** (variance-cost trade off).

- **Interplay** between the Monte Carlo batch and the mini-batch of indices drawn at each iteration (**bias-variance trade off**).

- **Complexity** of $\mathcal{O}(n/\epsilon)$ was found for the MISO method. $\mathcal{O}(n^{2/3}/\epsilon)$ for **quadratic** surrogates in **[Qian+, 2019]**.

Publications

- ***On the Global Convergence of (Fast) Incremental Expectation Maximization Methods***, Belhal Karimi, Hoi-To Wai, Eric Moulines, Marc Lavielle, accepted at NeurIPS 2019.

- ***Non-asymptotic Analysis of Biased Stochastic Approximation Scheme***, Belhal Karimi, Blazej Miasojedow, Eric Moulines and Hoi-To Wai, Proceedings of Conference on Learning Theory, COLT 2019.

- ***Efficient Metropolis-Hastings sampling for nonlinear mixed effects models***, Belhal Karimi and Marc Lavielle, Proceedings of BAYSM 2018.

- ***f-SAEM: A fast Stochastic Approximation of the EM algorithm***, Belhal Karimi, Marc Lavielle and Eric Moulines, Computational Statistics and Data Analysis, CSDA 2019.

- ***A Doubly Stochastic Surrogate Optimization Scheme for Non-convex Finite-sum Problems***, Belhal Karimi, Eric Moulines and Hoi-To Wai, Work in Progress 2019.





Thank You !