A Stochastic Path-Integrated Differential EstimatoR EM Algorithm

Gersende Fort[1], Eric Moulines[2] and Hoi-To Wai[3]

CNRS, France[1], École Polytechnique, France[2], Chinese University of Hong Kong, Hong-Kong[3]

(smooth) regularized Empiricak Risk Minimization

- Given a set of n observations $\{y_i, i \in [n]\}$.
- Goal: solve the possibly non-convex problem

$$\arg\min_{\Theta} F, \qquad F(\boldsymbol{\theta}) := -\frac{1}{n}\sum_{i=1}^{n} \log g(y_i, \boldsymbol{\theta}) + \mathrm{R}(\boldsymbol{\theta})$$

when $\Theta \subseteq \mathbb{R}^d$ and intractable positive function g

$$g(y_i, \boldsymbol{\theta}) = \int_{\mathsf{Z}} f(z, y_i; \boldsymbol{\theta}) \mu(dz)$$

Setting: Maximum Likelihood Estimation in latent variable models from exponential family

- parametric model, $\boldsymbol{\theta} \mapsto g(y_i, \boldsymbol{\theta})$ is the likelihood of y_i,
- latent variable $z \in \mathsf{Z}$; $f(z, y_i; \boldsymbol{\theta})$ is complete likelihood.
- Focus on the **Exponential Family Distribution**:

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

- Tool: **Expectation Maximization (EM)** (Dempster et al., 1977) algorithm which takes advantage of the latent structure.

EM for Exponential Family

★ **E-operation**:

$$\bar{\mathbf{s}}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \bar{\mathbf{s}}_i(\boldsymbol{\theta}), \quad \bar{\mathbf{s}}_i(\boldsymbol{\theta}) := \int_{\mathsf{Z}} S(z, y_i) p(z|y_i; \boldsymbol{\theta}) \mu(\mathrm{d}z)$$

where $z \mapsto p(z|y_i; \boldsymbol{\theta}) \propto f(y_i, z; \boldsymbol{\theta})$ is the a posteriori distribution of the latent variable z.

★ **M-operation**: for any s,

$$\mathsf{T}(s) := \arg\min_{\boldsymbol{\theta}\in\Theta} \left\{\mathrm{R}(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \left\langle s \,|\, \phi(\boldsymbol{\theta}) \right\rangle\right\}$$

Batch EM Algorithm: given $\hat{\mathbf{s}}^{(0)}$, repeat for $k \geq 0$,

1. (M-step) $\tau^{(k+1)} = \mathsf{T}(\hat{\mathbf{s}}^{(k)})$
2. (E-step) $\hat{\mathbf{s}}^{(k+1)} = \bar{\mathbf{s}}(\tau^{(k+1)}) = \bar{\mathbf{s}} \circ \mathsf{T}(\hat{\mathbf{s}}^{(k)})$

- Fixed points:

$$\{s : h(s) = 0\}, \qquad h(s) := \bar{\mathbf{s}} \circ \mathsf{T}(s) - s$$

For large n, the E-step is computationally expensive!

Variance reduced Stochastic Approximation within EM

Idea: Take advantage of the **Stochastic Approximation (SA)** context

$$h(s) = 0 \Longleftrightarrow \mathbb{E}\left[\bar{\mathbf{s}}_I \circ \mathsf{T}(s) - s\right] = 0, \qquad I \text{ uniform on } [n]$$

combined with **Variance Reduction (vr)**

$$h(s) = 0 \Longleftrightarrow \mathbb{E}\left[\bar{\mathbf{s}}_I \circ \mathsf{T}(s) - s + V\right] = 0, \text{with } \mathbb{E}[V] = 0.$$

vr-SA within EM (g[al] form):
- constant stepsize γ, mini-batch $\mathcal{B}_{t,k+1}$ in $[n]$,
- possibly nested loops (outer $\#t$, inner $\#k$),

$$\hat{\mathbf{s}}^{(t,k+1)} = \hat{\mathbf{s}}^{(t,k)} + \gamma\left(\bar{\mathbf{s}}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\hat{\mathbf{s}}^{(t,k)}) - \hat{\mathbf{s}}^{(t,k)} + \mathcal{S}^{(t,k+1)}\right)$$

★ SPIDER-EM. Init: $\mathcal{S}^{(t,0)} = \bar{\mathbf{s}} \circ \mathsf{T}(\hat{\mathbf{s}}^{(t-1,k_{\mathrm{in}}-1)})$

$$\mathcal{S}^{(t,k+1)} = \mathcal{S}^{(t,k)} - \bar{\mathbf{s}}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\hat{\mathbf{s}}^{(t,k-1)})$$

inspired by Fang et al., 2018; see also Nguyen et al. 2017, Wang et al. 2019

- online-EM: (Cappé & Moulines, 2009) unique outer, $\mathcal{S}^{(k+1)} = 0$.
- sEM-VR: (Chen et al., 2018) Init: $\mathcal{S}^{(t,0)} = \bar{\mathbf{s}} \circ \mathsf{T}(\hat{\mathbf{s}}^{(t-1,k_{\mathrm{in}})})$

$$\mathcal{S}^{(t,k+1)} = \mathcal{S}^{(t,0)} - \bar{\mathbf{s}}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\hat{\mathbf{s}}^{(t-1,k_{\mathrm{in}})})$$

- fiEM: (Karimi et al., 2019) unique outer, a memory $\mathsf{S}_i^{(k)}$ for $i \in [n]$,

$$\mathcal{S}^{(k+1)} = \overline{\mathcal{S}}^{(k+1)} - \mathsf{S}^{(k)}_{\mathcal{B}_{k+1}}, \qquad \overline{\mathcal{S}}^{(k+1)} = \overline{\mathcal{S}}^{(k)} + n^{-1}\sum_{i\in\mathcal{B}'_{k+1}} \left(\mathsf{S}_i^{(k+1)} - \mathsf{S}_i^{(k)}\right), \qquad \mathsf{S}_i^{(k+1)} = \mathsf{S}_i^{(k)} \mathbb{1}_{i\notin\mathcal{B}'_{k+1}} + \bar{\mathbf{s}}_i \circ \mathsf{T}(\hat{\mathbf{s}}^{k}) \mathbb{1}_{i\in\mathcal{B}'_{k+1}}$$

SPIDER-EM

- Termination numbers $\tau \sim \mathcal{U}([k_{\mathrm{out}}])$, $\xi \sim \mathcal{U}([k_{\mathrm{in}}]) - 1$
- Initialization: $\hat{\mathbf{s}}^{(1,0)} = \hat{\mathbf{s}}^{(1,-1)} = \hat{\mathbf{s}}^{\mathrm{init}}$, $\mathcal{S}^{(1,0)} = \bar{\mathbf{s}} \circ \mathsf{T}(\hat{\mathbf{s}}^{\mathrm{init}})$
- For $t = 1, \ldots, \tau$:
  - For $k = 0, \ldots, k_{\mathrm{in}} - 2$:
    - Sample $\mathcal{B}_{t,k+1}$ in $[n]$
    - $\mathcal{S}^{(t,k+1)} = \mathcal{S}^{(t,k)} - \bar{\mathbf{s}}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\hat{\mathbf{s}}^{t,k-1})$
    - $\hat{\mathbf{s}}^{(t,k+1)} = \hat{\mathbf{s}}^{(t,k)} + \gamma\left(\mathcal{S}^{(t,k+1)} + \bar{\mathbf{s}}_{\mathcal{B}_{t,k+1}} \circ \mathsf{T}(\hat{\mathbf{s}}^{t,k}) - \hat{\mathbf{s}}^{(t,k)}\right)$
  - $\hat{\mathbf{s}}^{(t+1,-1)} = \hat{\mathbf{s}}^{(t,k_{\mathrm{in}}-1)}$
  - $\mathcal{S}^{(t+1,0)} = \bar{\mathbf{s}} \circ \mathsf{T}(\hat{\mathbf{s}}^{(t+1,-1)})$
  - $\hat{\mathbf{s}}^{(t+1,0)} = \hat{\mathbf{s}}^{(t+1,-1)} + \gamma\left(\mathcal{S}^{(t+1,0)} - \hat{\mathbf{s}}^{(t+1,-1)}\right)$
- **Return**: $\hat{\mathbf{s}}^{(\tau,\xi-1)}$

SPIDER-EM: state of the art complexity bounds

- **Assumptions.** (i) $\mathsf{T}(s)$ exists and is unique; (ii) ϕ, ψ, R and F are C^1; (iii) $\nabla(\phi \circ \mathsf{T})(s)$ is a symmetric matrix with eigenvalues uniformly bounded in $[v_{\min}, v_{\max}]$ such that $v_{\min} > 0$; (iv) for all i, $\bar{\mathbf{s}}_i \circ \mathsf{T}$ is globally Lipschitz with constant L_i; (v) the function $s \mapsto \nabla(F \circ \mathsf{T})(s)$ is globally Lipschitz with constant $L_{\dot{W}}$
- **Property.** The **Lyapunov** function $W := F \circ \mathsf{T}$ has a globally Lipschitz gradient and $\|\nabla W(s)\|^2 \leq v_{\max}^2 \|h(s)\|^2$.
- **Upper bounds for convergence.** Minibatch size: b.

Theorem Set $L^2 := n^{-1}\sum_i L_i^2$. Choose $\gamma := \alpha/L$ where $\alpha \in (0, v_{\min}/\mu_\star(k_{\mathrm{in}}, \mathrm{b}))$ and $\mu_\star(k_{\mathrm{in}}, \mathrm{b}) := v_{\max}\sqrt{k_{\mathrm{in}}/\mathrm{b}} + L_{\dot{W}}/(2L)$.

$$\mathbb{E}\left[\|h(\hat{\mathbf{s}}^{\tau,\xi-1})\|^2\right] \leq \left(\frac{1}{k_{\mathrm{in}}} + \frac{\alpha^2}{\mathrm{b}}\right) \frac{2L}{\alpha\{v_{\min} - \alpha\mu_\star(k_{\mathrm{in}}, \mathrm{b})\}} \frac{1}{k_{\mathrm{out}}} \left(\mathbb{E}\left[W(\hat{\mathbf{s}}^{\mathrm{init}})\right] - \min W\right)$$

- **Complexity bounds.** $\mathcal{K}_{\mathrm{CE}}$: total nbr of conditional number of per-sample conditional expectations $\bar{\mathbf{s}}_i$ evaluations; $\mathcal{K}_{\mathrm{opt}}$: total nbr of optimization steps $\mathsf{T}(\hat{\mathbf{s}})$.

SPIDER-EM
- Find b, k_{in} and k_{out} as a function of n, ϵ s.t.

$$\mathbb{E}\left[\|h(\hat{\mathbf{s}}^{\tau,\xi-1})\|^2\right] \leq \epsilon$$

- Optimal strategy:

$$\mathrm{b} = O(\sqrt{n}), \quad k_{\mathrm{in}} = O(\sqrt{n}), \quad k_{\mathrm{out}} = O(\epsilon^{-1}/\sqrt{n})$$

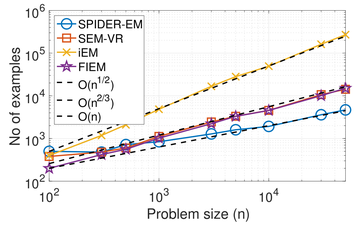
Comparison

	$\mathcal{K}_{\mathrm{CE}}$	$\mathcal{K}_{\mathrm{opt}}$	Optimal $\mathcal{K}_{\mathrm{CE}}$
EM	$n + nk_{\max}$	$1 + k_{\max}$	–
online-EM	$n + \mathrm{b}k_{\max}$	$1 + k_{\max}$	ϵ^{-2}
fiEM	$n + 2\mathrm{b}k_{\max}$	$1 + k_{\max}$	$\epsilon^{-1}n^{2/3}$
sEM-vr	$n + 2nk_{\mathrm{out}} + bk_{\mathrm{in}}k_{\mathrm{out}}$	$1 + k_{\mathrm{in}}k_{\mathrm{out}}$	$\epsilon^{-1}n^{2/3}$
SPIDER-EM	$n + nk_{\mathrm{out}} + 2\mathrm{b}k_{\mathrm{in}}k_{\mathrm{out}}$	$1 + k_{\mathrm{in}}k_{\mathrm{out}}$	$\epsilon^{-1}\sqrt{n}$

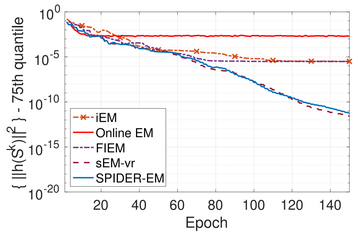
$k_{\max} = k_{\mathrm{in}}k_{\mathrm{out}}$ for algorithms with a single outer loop

Fitting a Gaussian Mixture Model (GMM)

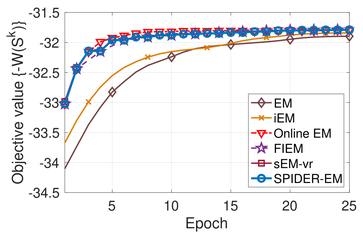
▶ Synthetic data Efficiency against the problem size n. Dataset is a GMM in $\mathbb{R}$ with 2 components. $\boldsymbol{\theta}$ collects the means.
- Median estimated nbr $\mathcal{K}_{\mathrm{CE}} - n$ (estimated over 50 runs) needed to reach an accuracy of $2.5e-5$.



▶ MNIST dataset $n = 6e4$ images classified into 12 classes. $\boldsymbol{\theta}$: parameters of a GMM in $\mathbb{R}^{20}$ (784 pixels reduced to 20 features through PCA)
- Quantile 0.75 of the distribution of $\|h(\hat{\mathbf{s}}_{t,-1})\|^2$



- Evolution of the objective fct W



References

O. Cappé and E. Moulines. On-line expectation-maximization algorithm for latent data models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(3):593–613, 2009.

J. Chen, J. Zhu, Y.W. Teh, and T. Zhang. Stochastic expectation maximization with variance reduction. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7967–7977. 2018.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Roy. Stat. Soc. B Met.*, 39(1):1–38, 1977.

C. Fang, C.J. Li, Z. Lin, and T. Zhang. SPIDER: Near-Optimal Non-Convex Optimization via Stochastic Path-Integrated Differential Estimator. In *NeurIPS 31*. 2018.

B. Karimi, H.-T. Wai, E. Moulines, and M. Lavielle. On the Global Convergence of (Fast) Incremental Expectation Maximization Methods. In *NeurIPS 32*. 2019.

L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, 2017.

Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh. SpiderBoost and Momentum: Faster Stochastic Variance Reduction Algorithms. In *NeurIPS 32*. 2019.

Advances in Neural Information Processing Systems 34