The Perturbed Prox-Preconditioned SPIDER algorithm for EM-based large scale learning

Gersende Fort (speaker)
(CNRS, Institut de Mathématiques de Toulouse, France)
Eric Moulines
(Ecole Polytechnique, CMAP, France)

SSP 2021

In the paper

- A novel EM algorithm: `Perturbed-Prox-Preconditioned-SPIDER-EM`
- Adapted to the large scale learning setting – large number of examples n
- Stochastic EM: it combines
  - the Stochastic Approximation method
  - a variance reduction technique
- Built on `SPIDER-EM` (Fort et al, 2020) – state of the art among the incremental EM's.

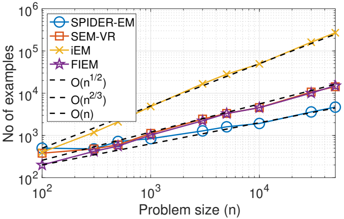


Figure: Nbr of processed examples required to reach convergence, as a function of the problem size n. From Fort et al. (2020, NeurIPS)

Optimization problem at hand

- Solve on $\Theta \subseteq \mathbb{R}^d$ the minimization problem

$$\mathrm{argmin}_{\theta \in \Theta} F(\theta)$$

$$F(\theta) \stackrel{\mathrm{def}}{=} -\sum_{i=1}^{n} \log \int_{\mathsf{Z}} h_i(z_i) \, \exp\left(\langle \mathsf{s}_i(z_i), \phi(\theta) \rangle\right) \mathrm{d}\mu(z_i) + \mathsf{R}(\theta), \qquad h_i(z) > 0$$

- In Statistical Learning:
  - minimization of the (penalized) negative log-likelihood in *latent variable* models.
  - observations $Y_1, \cdots, Y_n$; latent variables $Z_1, \cdots, Z_n$. $h_i \leftarrow h_{Y_i}$; $\mathsf{s}_i \leftarrow \mathsf{s}_{Y_i}$.
  - finite sum setting when the observations are independent.
  - the complete data likelihood of the pair $\#i$: (Y_i, Z_i) is from the Curved exponential family
  - An example ? inference in Gaussian Mixtures Models, inference in mixtures of densities from the curved exponential family, inference in logistic regression models, $\cdots$.

What EM would do

Objective function:

$$-\sum_{i=1}^{n} \log \int_{\mathsf{Z}} h_i(z_i) \, \exp\left(\langle \mathsf{s}_i(z_i), \phi(\theta) \rangle\right) \, \mathsf{d}\mu(z_i) + \mathsf{R}(\theta),$$

Repeat for $t = 0, \dots$

- **Expectation Step:**
  - for $i = 1, \cdots, n$, compute the expectation of the *sufficient statistics* s_i under the conditional distribution of the latent variables given the observations

$$\bar{\mathsf{s}}_i(\theta) = \int_{\mathsf{Z}} \mathsf{s}_i(z) \, \frac{p_i(z;\theta)}{\int p_i(u;\theta) \mathsf{d}\mu(u)} \, \mathsf{d}\mu(z) \qquad p_i(z;\theta) \propto h_i(z_i) \, \exp\left(\langle \mathsf{s}_i(z_i), \phi(\theta) \rangle\right)$$

  - compute the sum

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

- **Optimization Step:** update the parameter

$$\theta_{t+1} = \mathsf{T}(\bar{\mathsf{s}}(\theta_t)) \qquad \mathsf{T}(s) \stackrel{\text{def}}{=} \operatorname{argmin}_{\theta \in \Theta} \ \mathsf{R}(\theta) - \langle s, \phi(\theta) \rangle$$

Intractable !! $\rightarrow$ a novel *incremental EM*

What incremental EM's do

- based on the observation that **EM is** *equivalent to* **find the root of**

$$\mathsf{h}(s) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} \bar{\mathsf{s}}_i \circ \mathsf{T}(s) - s = \mathbb{E}\left[\bar{\mathsf{s}}_I \circ \mathsf{T}(s) - s\right]$$

- designed to address the finite sum setting
- use a Stochastic Approximation update mechanism

$$\widehat{S}_{t+1} = \widehat{S}_t + \gamma_{t+1} H_{t+1} \qquad H_{t+1} \approx \mathsf{h}(\widehat{S}_t)$$

Key observation for the definition of the field H_{t+1}

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

What 3P-SPIDER-EM does

- As for *Incremental EM's*: a stochastic approximation of the full sum

$$\frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t+1}} \bar{\mathsf{s}}_i \circ \mathsf{T}(\widehat{S}_t) - \widehat{S}_t$$

- **(new)** An approximation of the conditional expectations, possibly random

$$\hat{\mathsf{s}}_i^t \approx \bar{\mathsf{s}}_i \circ \mathsf{T}(\widehat{S}_t)$$

- The same definition of the *control variate* V as in SPIDER-EM Fort et al., 2020
- **(new)** Constraint on the updated statistics : $\theta \in \Theta \to \bar{\mathsf{s}} \circ \mathsf{T}(s) \in \mathcal{S}$

$$\widehat{S}_{t+1} = \mathrm{Prox}_{B_{t+1}, \gamma_{t+1} g} \left(\widehat{S}_t + \gamma_{t+1} H_{t+1} \right) \qquad \mathrm{Prox}_{B,g}(s) \stackrel{\mathrm{def}}{=} \mathrm{argmin}_u \, g(u) + \frac{1}{2} \|u - s\|^2_B$$

For the convergence analysis: 3P-SPIDER-EM does not satisfy the descent property of EM:

$$F \circ \mathsf{T}(\widehat{S}_{t+1}) \leq F \circ \mathsf{T}(\widehat{S}_t)$$

but 3P-SPIDER-EM is related to a *preconditioned gradient* algorithm

$$\nabla (F \circ \mathsf{T}) = -B(s) \, \mathsf{h}(s) \qquad B_{t+1} \stackrel{\mathrm{def}}{=} B(\widehat{S}_t)$$

Perturbed-Prox-Preconditioned-SPIDER-EM (Stochastic Path Integrated Differential EstimatoR)

$\widehat{S}_{1,0} = \widehat{S}_{1,-1} = \widehat{S}_{\mathrm{init}} \qquad \mathsf{S}_{1,0} = \bar{\mathsf{s}} \circ \mathsf{T}(\widehat{S}_{1,-1}) + \mathcal{E}_1$

for $t = 1, \cdots, k_{\mathrm{out}}$ **do**

  for $k = 0, \ldots, k_{\mathrm{in}} - 1$ **do**

    Sample a mini batch $\mathcal{B}_{t,k+1}$ of size b from $\{1, \cdots, n\}$

    $\mathsf{S}_{t,k+1} = \mathsf{S}_{t,k} + \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{t,k+1}} \left(\hat{\mathsf{s}}_i^{t,k} - \hat{\mathsf{s}}_i^{t,k-1} \right)$

    $\widehat{S}_{t,k+1/2} = \widehat{S}_{t,k} + \gamma_{t,k+1} \left(\mathsf{S}_{t,k+1} - \widehat{S}_{t,k} \right)$

    $\widehat{S}_{t,k+1} = \mathrm{Prox}_{B_{t,k}, \gamma_{t,k+1} g} \left(\widehat{S}_{t,k+1/2} \right)$

  end for

  $\widehat{S}_{t+1,-1} = \widehat{S}_{t,k_{\mathrm{in}}}$

  $\mathsf{S}_{t+1,0} = \bar{\mathsf{s}} \circ \mathsf{T}(\widehat{S}_{t+1,-1}) + \mathcal{E}_{t+1}$ $\mathcal{E}_{t+1}$: a possible error

  $\widehat{S}_{t+1,-1/2} = \widehat{S}_{t+1,-1} + \gamma_{t+1,0} \left(\mathsf{S}_{t+1,0} - \widehat{S}_{t+1,-1} \right)$

  $\widehat{S}_{t+1,0} = \mathrm{Prox}_{B_{t+1,-1}, \gamma_{t+1,0} g} \left(\widehat{S}_{t+1,-1/2} \right)$

end for

Convergence Analysis

- **Non-convex** optimization problem: find the root of $s \mapsto \mathsf{h}(s)$
  - Explicit control in expectation of the algorithm stopped at a random termination time

$$\mathbb{E}\left[\|\mathsf{h}(\widehat{S}_{\tau,K})\|^2\right] \qquad (\tau, K) \sim \mathcal{U}\left([1, \cdots, k_{\mathrm{out}}] \times [0, \cdots, k_{\mathrm{in}} - 1]\right)$$

  - With conditions on the perturbations $\bar{\mathsf{s}}_i \circ \mathsf{T}(\widehat{S}_{t,k}) - \hat{\mathsf{s}}_i^{t,k}$, which are satisfied when Monte Carlo approximation of $\bar{\mathsf{s}}_i$.

- Technical difficulties for the proof:
  - A biased control variate

$$\mathbb{E}\left[\mathsf{S}_{t,k+1} | \mathrm{past}_{t,k}\right] \neq n^{-1} \sum_{i=1}^{n} \bar{\mathsf{s}}_i \circ \mathsf{T}(\widehat{S}_{t,k})$$

  - A possibly biased approximation $\bar{\mathsf{s}}_i \circ \mathsf{T}(\widehat{S}_{t,k}) - \hat{\mathsf{s}}_i^{t,k}$
  - The proximal operator

See the companion paper "The Perturbed Prox-Preconditioned SPIDER algorithm: non-asymptotic convergence bounds", SSP 2021

Application: The logistic regression model

- n observations $Y_i \in \{-1, 1\}$

$$\mathbb{P}(Y_i = 1 | Z_i) = \frac{1}{\exp\left(-\langle X_i, Z_i \rangle\right)} \qquad Z_i \sim \mathcal{N}_d(\theta, \sigma^2 I)$$

- Unknown: the expectation θ of the individual predictors Z_i (latent variables).
- Ridge penalized ML estimator.
- $\bar{\mathsf{s}}_i(\theta)$ is an intractable expectation $\rightarrow$ MCMC sampler.

Numerical illustrations: from the MNIST data set. Class 1 contains 12873 images (labels 1 and 3); class -1 contains 12116 images (labels 7 and 8)

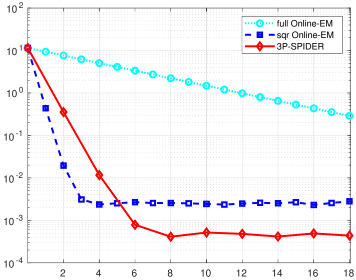
Application: 3P-SPIDER-EM compared to Online EM

Displayed: the variation, estimated by MC over 25 independent runs

$$\mathbb{E}\left[\frac{\|\widehat{S}_{t,k}-\widehat{S}_{t,k-1}\|^2}{\gamma_{t,k}^2}\right]$$

a kind of distance to the roots of h

vs the number of epochs. Compared to *full* OnlineEM and *sqr* Online EM.



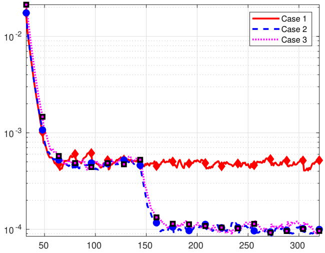
Conclusion: Despite the proximal step and the MCMC approximations of $\bar{s}_i'$s, 3P-SPIDER-EM improve on classical Incremental EM.

Application: approximation $\hat{\mathsf{s}}_i^{t,k}$ and strategy for $\gamma_{t,0}$

Displayed: the variation, estimated by MC over 25 independent runs

$$\mathbb{E}\left[\frac{\|\widehat{S}_{t,k}-\widehat{S}_{t,k-1}\|^2}{\gamma_{t,k}^2}\right]$$

vs the number of epochs. When the number of MC points is increased (see `Case 1` and `Cases 2,3`); when $\gamma_{t,0}=0$ or not (see `Case 2` and `Case 3`).



Conclusion: The efficiency of 3P-SPIDER-EM depends on the quality of the approximations of the $\bar{\mathsf{s}}_i$'s; the strategy for $\gamma_{t,0}$ is not clear ($\rightarrow$ an error at the same level as the MCMC approximations here)