Stochastic EM methods with Variance Reduction for Penalised PET Reconstructions

23 November 2021

Introduction

$\mathbf{Af} + \mathbf{r} = \mathbb{E}[\mathbf{g}]$

- Iterative methods are widely used in PET reconstruction
- EM-ML[1] and its variants are particularly prevalent

$$\mathbf{f}^{k+1} = \underset{\mathbf{f} \geq \mathbf{0}}{\operatorname{argmax}} \, \mathbb{E}_{\mathbf{G}|\mathbf{g},\mathbf{f}^k}[\log p(\mathbf{G}|\mathbf{f})]$$

image — measured data — complete data

- Explicit solution in each step
- Ordered subset (OS) methods improve the convergence in early iterations

[1] Shepp and Vardi, '82

Introduction

- Iterative methods are widely used in PET reconstruction
- EM and its variants are particularly prevalent

$$\mathbf{f}^{k+1} = \underset{\mathbf{f}\geq\mathbf{0}}{\operatorname{argmax}}\, \mathbb{E}_{\mathbf{G}_t|\mathbf{g}_t,\mathbf{f}^k}[\log p(\mathbf{G}_t|\mathbf{f})]$$

subset index

- Explicit solution in each step
- Ordered subset (OS) methods improve the convergence in early iterations

Two common issues with OSEM methods:

- Loss of convergence towards the maximising solution
  - Instead we enter a limit cycle behaviour
- Problems if there is a penalty (MAP-EM)

$$\mathbf{f}_{\text{map}} = \underset{\mathbf{f}\geq 0}{\operatorname{argmax}}\{\Phi(\mathbf{f}) := \mathcal{L}(\mathbf{f}) - \beta\,\mathcal{R}(\mathbf{f})\}$$

log likelihood ($\mathcal{L}(\mathbf{f})$), penalty ($\mathcal{R}(\mathbf{f})$), penalty strength (β)

  - Maximisation is no longer analytical and thus further approximations are needed

Two common issues with OSEM methods:

- Loss of convergence towards the maximising solution
  - Instead we enter a limit cycle behaviour
- Problems if there is a penalty (MAP-EM)

$$\mathbf{f}_{\text{map}}^{k+1} = \underset{\mathbf{f}\geq\mathbf{0}}{\operatorname{argmax}}\{\mathbb{E}_{\mathbf{G}|\mathbf{g},\mathbf{f}^k}[\log p(\mathbf{G}|\mathbf{f})] - \beta\,\mathcal{R}(\mathbf{f})\}$$

conditional expectation

penalty

penalty strength

  - Maximisation is no longer analytical and thus further approximations are needed

- Alternative: optimise using gradient ascent based methods (as discussed by Robbie)
- Instead we consider stochastic EM algorithms for MAP-EM which
  - Uses OS and **exponentially moving average** of the expected statistic
  - Employs **separable parabolic surrogates**[1] for the prior

[1] de Pierro '95, de Pierro and Yamagishi '01, Erdogan and Fessler '98, etc.

Online EM [2]

- Write MAP-EM as

depend on $\mathbf{f}$

$$\mathbf{f}^{k+1} = \underset{\mathbf{f}\geq\mathbf{0}}{\operatorname{argmax}} \left\{ \log(\mathbf{f})^\top s(\mathbf{f}^k) - \sum_{m=1}^{M} a_m^\top \mathbf{f} - \beta \mathcal{R}(\mathbf{f}) \right\},$$

- Here

full conditional statistic

$$s(\mathbf{f}^k) = \mathbb{E}_{\mathbf{G}|\mathbf{g},\mathbf{f}^k} \left[\log\Big(\sum_{m=1}^{M} g_{mn}\Big)_{n=1}^{N} \right] = \frac{1}{N_s} \sum_{t=1}^{N_s} \tau_t(\mathbf{f}^k)$$

where

$$\tau_t(\mathbf{f}) = N_s f \odot (\nabla \mathcal{L}_t(\mathbf{f}) + A_t^\top 1)$$

subset conditional statistic

Stochastic EM

Instead of $s(\mathbf{f}^k)$ we compute [2, 1, 3]

- **SEM**

$$\hat{s}^{k+1} = (1-\alpha_k)\,\hat{s}^k + \alpha_k \tau_{t_k}(\hat{\mathbf{f}}^k_{\mathrm{sem}})$$

- **SVREM**

$$\hat{s}^{k+1} = (1-\alpha)\,\hat{s}^k + \alpha\left(\tau_{t_k}(\hat{\mathbf{f}}^k_{\mathrm{svrem}}) - \tau_{t_k}(\hat{\mathbf{f}}^{\mathrm{anc}}) + s^{\mathrm{anc}}\right)$$

If $k \bmod \eta N_s = 0$, set $\mathbf{f}^{\mathrm{anc}} = \mathbf{f}^k_{\mathrm{svrem}}$ and update $s^{\mathrm{anc}} = s(\mathbf{f}^{\mathrm{anc}})$

- **SAGAEM**

$$\hat{s}^{k+1} = (1-\alpha)\,\hat{s}^k + \alpha\left(\tau_{t_k}(\hat{\mathbf{f}}^k_{\mathrm{sagaem}}) - \mathfrak{s}_{t_k} + \frac{1}{N_s}\sum_{t=1}^{N_s} \mathfrak{s}_t\right)$$

Draw $\tilde{t}_k \in [N_s]$, set $\mathfrak{s}_{\tilde{t}_k} = \tau_{\tilde{t}_k}(\hat{\mathbf{f}}^k_{\mathrm{sagaem}})$, keep the rest intact

Stochastic EM

Instead of $s(\mathbf{f}^k)$ we compute [2, 1, 3]

- **SEM**

$$\widehat{s}^{k+1} = (1-\alpha_k)\widehat{s}^k + \alpha_k \, \tau_{t_k}(\widehat{\mathbf{f}}^k_{\mathrm{sem}})$$

- **SVREM**

$$\widehat{s}^{k+1} = (1-\alpha)\widehat{s}^k + \alpha \left(\tau_{t_k}(\widehat{\mathbf{f}}^k_{\mathrm{svrem}}) - \tau_{t_k}(\widehat{\mathbf{f}}^{\mathrm{anc}}) + s^{\mathrm{anc}}\right)$$

If $k \mod \eta N_s = 0$, set $\mathbf{f}^{\mathrm{anc}} = \mathbf{f}^k_{\mathrm{svrem}}$ and update $s^{\mathrm{anc}} = s(\mathbf{f}^{\mathrm{anc}})$

- **SAGAEM**

$$\widehat{s}^{k+1} = (1-\alpha)\widehat{s}^k + \alpha \left(\tau_{t_k}(\widehat{\mathbf{f}}^k_{\mathrm{sagaem}}) - \mathfrak{s}_{t_k} + \frac{1}{N_s}\sum_{t=1}^{N_s} \mathfrak{s}_t\right)$$

Draw $\tilde{t}_k \in [N_s]$, set $\mathfrak{s}_{\tilde{t}_k} = \tau_{\tilde{t}_k}(\widehat{\mathbf{f}}^k_{\mathrm{sagaem}})$, keep the rest intact

Separable surrogates

- We consider (standard) priors of the form

$$\mathcal{R}(\mathbf{f}) = \frac{1}{2}\sum_{n=1}^{N}\sum_{j\in\mathcal{N}_n} w_{nj}\,\rho\big(f_n - f_j\big)$$

smooth, non decreasing function of $|f_n - f_j|$

- The issue with (explicit) maximisation with general priors is that the gradients are not spatially independent

- Instead of ρ use a parabolic surrogate [4]

$$\widehat{\rho}^k(f_n; f_j) = \gamma_\rho(f_n^k - f_j^k)\Big(\big(f_n - \tfrac{f_n^k + f_j^k}{2}\big)^2 + \big(f_j - \tfrac{f_n^k + f_j^k}{2}\big)^2\Big),$$

where $\gamma_\rho(f) = \frac{\rho(f)}{f}$

- The surrogate M-step for MAP-SEM/SVREM/SAGAEM is given by

$$\mathbf{f}^{k+1} = \underset{\mathbf{f} \geq \mathbf{0}}{\operatorname{argmax}} \left\{ \log(\mathbf{f})^\top \widehat{s}^{k+1} - \sum_{m=1}^{M} a_m^\top \mathbf{f} - \beta \widehat{\mathcal{R}}(\mathbf{f}; \mathbf{f}^k) \right\}$$

where

$$\widehat{\mathcal{R}}(\mathbf{f}; \mathbf{f}^k) = \frac{1}{2} \sum_{n=1}^{N} \sum_{j \in \mathcal{N}_n} w_{nj} \, \hat{\rho}^k \left(f_n; f_j \right)$$

- Explicit maximiser (root of the gradient is a quadratic polynomial with a single non-negative solution)

Explicit maximiser

Let $d_{nj} := w_{nj}\gamma_\rho(f_n^k - f_j^k)$ and

$$a_n = \widehat{s}_n^k, \quad b_n = \beta \sum_{j \in \mathcal{N}_n} d_{nj},$$

$$c_n = \beta f_n^k \sum_{j \in \mathcal{N}_n} d_{nj} + \beta \sum_{j \in \mathcal{N}_n} d_{nj} f_j^k - \sum_{m=1}^{M} a_{mn}$$

Then

$$f_n^{k+1} = \frac{c_n + \sqrt{c_n^2 + 8a_n b_n}}{4b_n}$$

UCL CENTRE FOR INVERSE PROBLEMS

Admissible potentials

	$\rho(t)$	$\rho'(x)$	$\gamma_\rho(x)$
quadratic	$\frac{x^2}{2}$	x	1
log cosh	$\delta^2 \log\cosh(x/\delta)$	$\delta \tanh(x/\delta)$	$\delta \frac{\tanh(x/\delta)}{x}$
hyperbola	$\delta\left(\sqrt{1+(x/\delta)^2}-1\right)$	$\frac{x}{\sqrt{1+(x/\delta)^2}}$	$\frac{1}{\sqrt{1+(x/\delta)^2}}$

Convergence for SAGA and SVRG

As a reminder, variance reduced gradient ascent methods obey

$$\mathbf{f}^{k+1} = \mathbf{P}_{\geq 0}\left(\mathbf{f}^k + \alpha \mathbf{D}_k(\mathbf{f}^k)\tilde{\nabla}_k\right)$$

Theorem

Let $d \in \mathbb{R}^N_{>0}$*, denote by* $L = \max_{t \in N_s} L_t$ *where* L_t *is the Lipschitz constant of sub-objective gradients* $\widetilde{\Phi}_t(\mathbf{f})$ *and by* $d_{\max} = \|d\|_\infty$*, and assume* $\operatorname{argmax}_{\mathbf{f} \geq \mathbf{0}} \Phi(\mathbf{f}) \neq \emptyset$*.*
Taking $\alpha \leq \frac{1}{3 L d_{\max}^{1/2}}$ *and* $\mathbf{D}_t(\mathbf{f}^k_{\text{saga}}) = \operatorname{diag}(d)$ *in the* SAGA *algorithm we have* $\widetilde{\Phi}(\mathbf{f}^k_{\text{saga}}) \to \Phi(\mathbf{f}^\star)$ *and* $\mathbf{f}^k_{\text{saga}} \to \mathbf{f}^\star$ *almost surely.*
Taking $\alpha \leq \frac{1}{4 L d_{\max}^{1/2}(\eta N_s + 2)}$ *and* $\mathbf{D}_t(\mathbf{f}^k_{\text{svrg}}) = \operatorname{diag}(d)$ *in the* SVRG *algorithm we have* $\mathbf{f}^k_{\text{svrg}} \to \mathbf{f}^\star$ *almost surely and* $\mathbb{E}[\Phi(\mathbf{f}^\star) - \widetilde{\Phi}(\mathbf{f}^{k\eta N_s}_{\text{svrg}})] = \mathcal{O}(1/k)$*.*

Convergence

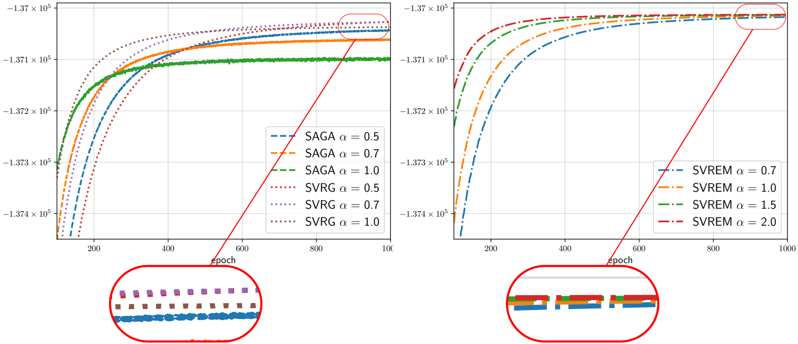
- Subset gradients Φ_t are in general not Lipschitz
- But, the assumptions are satisfied in physically realistic cases (everywhere non-zero backgrounds $\mathbf{r} > \mathbf{0}$) or with the used of a modified log-likelihood[2]
- Under current theory SVREM and SAGAEM will also converge under certain (but somewhat stronger) Lipschitz assumptions

[2] Ahn and Fessler, '03

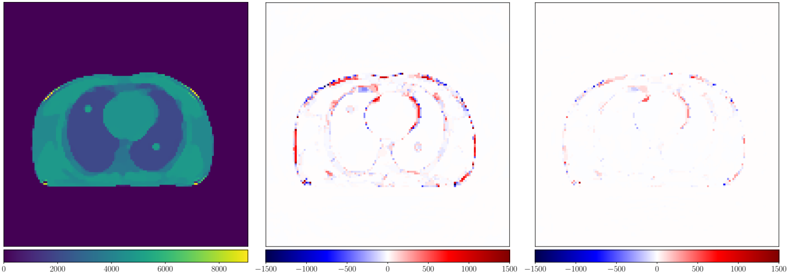
Experiments - XCAT Phantom

- XCAT torso phantom; 280 view scanner
- $\log\cosh$ prior with hand selected δ and penalty strength β
- Initialised with 5 epochs of OSEM
- Sinogram data pre-binned as OS. A subset index is then sampled at random in each iteration

Objective Value - 40 Subsets



SVREM Reconstruction Progression



(a) f^{1000}_{svrem} (b) $f^{1000}_{\text{svrem}} - f^{50}_{\text{svrem}}$ (c) $f^{1000}_{\text{svrem}} - f^{200}_{\text{svrem}}$

Figure: (a) SVREM reconstruction after 1000, and (b)-(c) pixel-wise differences of SVREM reconstructions after 200 and 50 epochs.

Quick and easy way heuristics to accelearate the convergence

- Using *SVRG without the outer loop* (and adjusting η)
- Nonlinear acceleration through extrapolation
  - Improves performance drastically on simple data
  - Inconsistent on more realistic data
- Nesterov, etc.

References

J. Chen, J. Zhu, Y. W. Teh, and T. Zhang,
Stochastic Expectation Maximization with Variance Reduction,
NeurIPS, (2018).

O. Cappé and E. Moulines,
Online Expectation-Maximization Algorithm for Latent Data Models,
Journal of the Royal Statistical Society: Series B, **71(3)** (2009).

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

J.-H. Chang, J. M. M. Anderson, and J. R. Votaw,
Regularized Image Reconstruction Algorithms for Positron Emission Tomography,
IEEE Trans. Med. Imag., **23(9)** (2004).