On Stochastic Variance Reduction for Penalised PET Reconstruction

Zeljko Kereta [1], Robert Twyman [2], Simon Arridge [1], Kris Thielemans [2], Bangti Jin [1]

[1] Computer Science Department, University College London

[2] Institute of Nuclear Medicine, University College London

Advanced Image Reconstruction Methods

12 September 2022

Introduction

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

- Penalised PET reconstruction is concerned with

$$\mathbf{f}_{\text{map}} = \underset{\mathbf{f} \quad 0}{\operatorname{argmax}} f \quad (\mathbf{f}) := L(\mathbf{f}) \quad R(\mathbf{f}) g$$

log likelihood

penalty

penalty strength

- Iterative optimisation is widely used in the reconstruction
- Using all the data can be costly and slow - **ordered subsets** (mini batching) dramatically improve convergence in early iterations

Motivation

- Standard subset methods experience loss of convergence towards the maximising solution
- This often leads to limit cycle behaviour and significant variability between successive image updates
- Convergence can be achieved by a relaxed step size sequence
- Alternatively, we can utilise variance reduction algorithms

Preconditioned Gradient Ascent

- Update equation via diagonally preconditioned gradient ascent methods

$$\mathbf{f}^{k+1} = \mathbf{P}_0\left[\mathbf{f}^k + \alpha_k D_{t_k}(\mathbf{f}^k)\, g_{t_k}(\mathbf{f}^k)\right]$$

stepsize

diagonal, non-negative preconditioner

gradient estimator

- $\mathbf{P}_0$ is a non-negativity projection; t_k is a subset index

Example: SGD/OSEM

- $r_t(\mathbf{f}) = \mathbf{A}_t^{>}\left(\frac{\mathbf{g}_m}{\mathbf{A}_t\mathbf{f}+\mathbf{r}_m} \quad 1\right)$ $\frac{}{M}\, r\, R(\mathbf{f}),\ D_t(\mathbf{f}) = \text{diag}\ \frac{\mathbf{f}}{\mathbf{A}_t^{>}1}$

Stochastic Variance Reduction

Common features

- Gradient estimator $\theta_{t_k}(\mathbf{f}^k)$ is computed by using both a new subset gradient $r_{t_k}(\mathbf{f}^k)$ and a history of stored gradients
- Each update has approximately the same computational cost as one subset gradient update, but the memory is increased
- Use of a constant step-size

Stochastic Variance Reduction

- **SAGA** -replace stored subset gradient with the current one

$$\nabla f_{t_k}(\mathbf{f}^k) = M\left(r_{t_k}(\mathbf{f}^k) - \mathbf{g}_{t_k}\right) + \sum_{t=1}^{N_S} \mathbf{g}_t$$

$$\text{update } \mathbf{g}_{t_k} = r_{t_k}(f^k); \text{ and for } t \neq t_k \text{ keep } \mathbf{g}_t$$

- **SVRG** -periodically recompute subset gradient history at current image

$$\nabla f_{t_k}(\mathbf{f}^k) = M\left(r_{t_k}(\mathbf{f}^k) - r_{t_k}(\mathbf{f}^{\text{anc}})\right) + G$$

$$\text{If } k \bmod N_S = 0 \text{ set } f^{\text{anc}} = f^k \text{ and update } G = r(f^{\text{anc}})$$

Convergence for SAGA and SVRG

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 $\nabla_t(f)$ and by $d_{max} = \|d\|_1$, and assume $\operatorname{argmax}_{f \geq 0} \ell(f) \neq \emptyset$.

Taking $\frac{1}{3Ld_{max}^{1/2}}$ and $D_t(f^k_{saga}) = \operatorname{diag}(d)$ in the SAGA algorithm we have $\ell(f^k_{saga}) \to \ell(f^\star)$ and $f^k_{saga} \to f^\star$ almost surely.

Taking $\frac{1}{4Ld_{max}^{1/2}(N_s+2)}$ and $D_t(f^k_{svrg}) = \operatorname{diag}(d)$ in the SVRG algorithm we have $f^k_{svrg} \to f^\star$ almost surely and $E[\ell(f^\star) - \ell(f^{k\,N_s}_{svrg})] = O(1/k)$.

NB: Subset gradients ∇_t are in general not Lipschitz. But, the assumptions are satisfied in physically realistic cases (everywhere non-zero backgrounds $r > 0$) or with the used of a modified log-likelihood

Experiments - XCAT Phantom

- GE Discovery 690 Scanner
- XCAT torso phantom with a 1cm diameter, 1cm long hot lesion (2.5:1 contrast)
- Data binned into 288 projection angles
- Image reconstruction algorithms based on STIR
- BSREM used for comparison

50 million event simulated data

Figure: Photon emission distribution

50 million event simulated data

Figure: Attenuation distribution cm-1

50 million event simulated data

Figure: Initial OSEM image

Experiments - Algorithm Setup

- Relative difference prior

$$R(x) = \sum_{i=1}^{N_v} \sum_{l \in N_i} p \frac{(f_i - f_l)^2}{(f_i + f_l) + 2|f_i - f_l|}$$

- Preconditioner $D(f) = \mathrm{diag}(\frac{f+}{A^{>} 1})$, with small
- Algorithms warm started with 1 epoch of OSEM (24 subsets)
- ROI drawn over the lung lesion, and we compute

$$\text{ROI Percentage Error} = 100\% \frac{\hat{k}}{\hat{k}};$$

Experiments - Investigations

Investigate the impact of

- Choice of preconditioner: variable $D(f^k)$, frozen at initial point $D(f_0)$, after 5 epochs $D(f_{5M})$ or after 10 epochs $D(f_{10M})$
- Constant ($_k = 1$), and decaying ($_k = \frac{1}{k=M+1}$) stepsizes
- The number of subsets
- Subsets are pre-binned and then randomly sampled - other methodologies were tested but are not shown

Number of subsets

Figure: BSREM

Number of subsets

Figure: SAGA

Number of subsets

Figure: SVRG

Role of the preconditioner

Figure: SAGA

Role of the preconditioner

Figure: SVRG

Role of the preconditioner

Figure: ROI outside the torso

Variance Reduction in MAP-EM

- Traditionally, PET was optimised through an MLEM (maximum-likelihood expectation maximisation) algorithm

$$f^{k+1} = \underset{f \quad 0}{\operatorname{argmax}} E_{G|g;f}[\log p(G|f)]$$

- The issue with (explicit) maximisation with general priors (MAP-EM) is that the gradients are not spatially independent, and thus there is no closed-form maximiser

Online EM

- Write MAP-EM as

depend on f

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

- Here

full conditional statistic

$$T(f^{k}) = \frac{1}{N_s} \sum_{t=1}^{N_s} {}_t(f^{k})$$

where

$${}_t(f) = N_s \operatorname{diag}(f)(rL \; {}_t(f) + A_t^{\top} 1)$$

subset conditional statistic

Stochastic EM

Instead of $T(f^k)$ we compute

- SEM

$$\theta^{k+1} = (1 - \gamma_k)\theta^k + \gamma_k t_k(\theta^k)$$

- SVREM

$$\theta^{k+1} = (1 - \gamma_k)\theta^k + \gamma_k t_k(\theta^k) - \gamma_k t_k(\theta^{anc}) + s^{anc}$$

If $k \bmod N_s = 0$; set $f^{anc} = f^k$ and update $s^{anc} = T(f^{anc})$

- SAGAEM

$$\theta^{k+1} = (1 - \gamma_k)\theta^k + \gamma_k t_k(\theta^k) - s_{t_k} + \frac{1}{N_s}\sum_{t=1}^{N_s} s_t$$

Draw $t_k \in [N_s]$; set $s_{t_k} = t_{t_k}(\theta^k)$; keep the rest intact

Stochastic EM

Instead of $T(f^k)$ we compute

- SEM

$$\theta^{k+1} = (1 - \eta_k)\theta^k + \eta_k \, \hat{\theta}_{t_k}(h^k)$$

- SVREM

$$\theta^{k+1} = (1 - \eta_k)\theta^k + \eta_k \left[\hat{\theta}_{t_k}(h^k) - \hat{\theta}_{t_k}(h^{anc}) + s^{anc} \right]$$

If $k \bmod N_s = 0$; set $f^{anc} = f^k$ and update $s^{anc} = T(f^{anc})$

- SAGAEM

$$\theta^{k+1} = (1 - \eta_k)\theta^k + \eta_k \left[\hat{\theta}_{t_k}(h^k) - s_{t_k} + \frac{1}{N_s} \sum_{t=1}^{N_s} s_t \right]$$

Draw $t_k \in [N_s]$; set $s_{t_k} = \hat{\theta}_{t_k}(h^k)$; keep the rest intact

Separable surrogates

- We consider (standard) priors of the form

$$R(f) = \frac{1}{2} \sum_{n=1}^{N} \sum_{j \in 2N_n} w_{nj} \, f_n \, f_j$$

and use parabolic surrogates

$$b^k(f_n; f_j) = (f_n^k \; f_j^k) \; f_n \; \frac{f_n^k + f_j^k}{2}^2 + f_j \; \frac{f_n^k + f_j^k}{2}^2 ;$$

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

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

$$f^{k+1} = \operatorname*{argmax}_{f \geq 0} \left\{ \log(f)^{\top} h^{k+1} - \sum_{m=1}^{M} a_m^{\top} f - \hat{R}(f; f^k) \right\}$$

where

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

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

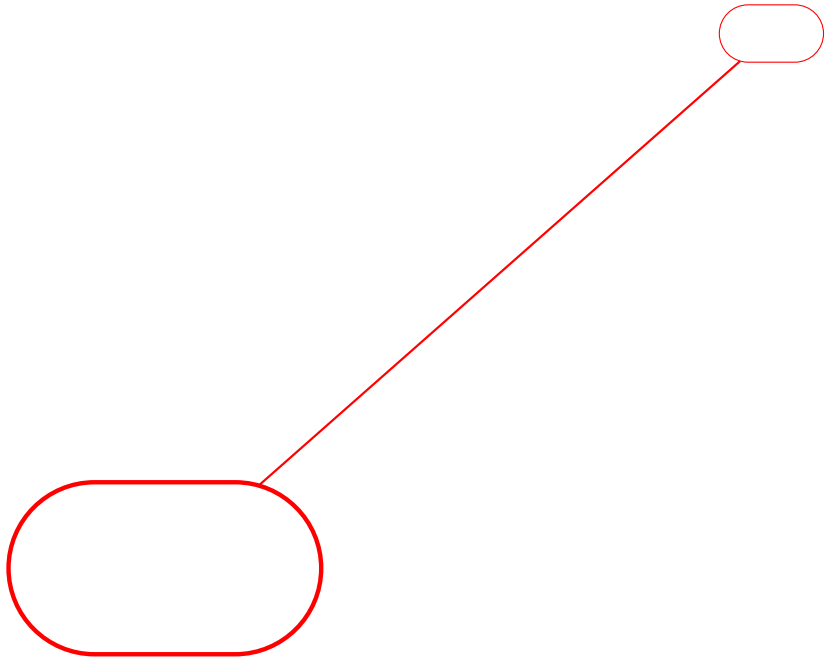
Admissible potentials

	(t)	$^0(x)$	(x)
quadratic	$\frac{x^2}{2}$	x	1
log cosh	$p^{\ 2}\ \log\cosh(x=\ )$	$\tanh(x=\ )$	$\frac{\tanh(x=\ )}{x}$
hyperbola	$\frac{}{1+(\ x=\ )^2}\quad 1$	$p\ \frac{x}{1+(\ x=\ )^2}$	$p\ \frac{1}{1+(\ x=\ )^2}$

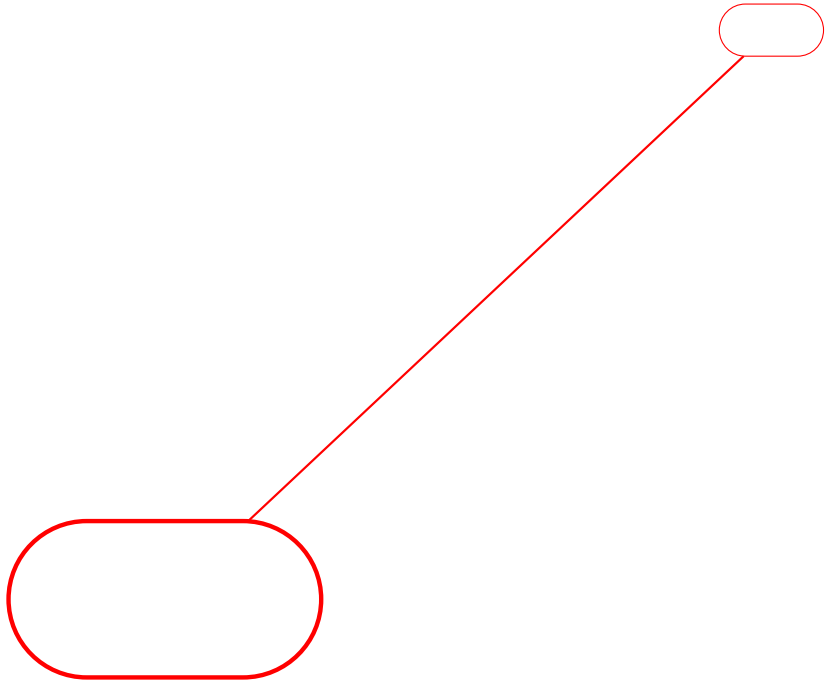
Experiments - XCAT Phantom

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

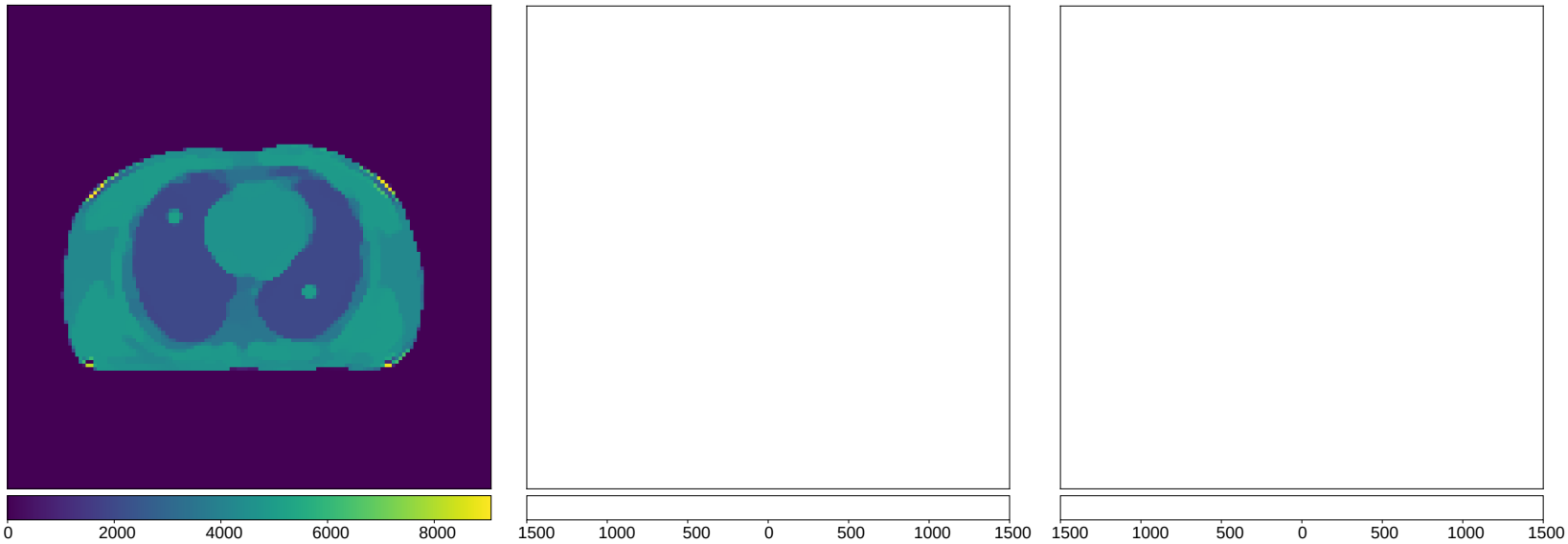
Objective Value - 40 Subsets



Objective Value - 40 Subsets



SVREM Reconstruction Progression



(a) f^{1000}_{svrem} (b) f^{1000}_{svrem} f^{50}_{svrem} (c) f^{1000}_{svrem} f^{200}_{svrem}

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