Faster PET Reconstruction with Non-Smooth Priors by Randomization

Matthias J. Ehrhardt

Institute for Mathematical Innovation
University of Bath, UK

September 5, 2019

Joint with:	Mathematics:	A. Chambolle, Ecolé Polytechnique, France
		P. Richtárik, KAUST, Saudi Arabia
		C. Schönlieb, Cambridge, UK
	PET imaging:	P. Markiewicz, UCL, UK
		J. Schott, UCL, UK

Institute for Mathematical Innovation | UNIVERSITY OF BATH

EPSRC
Engineering and Physical Sciences Research Council

Outline

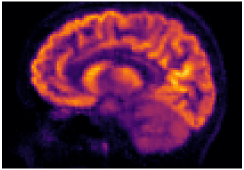
1) PET Reconstruction via Optimization (**Why?**)

$$\sum_{i=1}^{n} f_i(\mathbf{B}_i x) + g(x)$$

2) Randomized Algorithm for Convex Optimization (**How?**)

non-smooth
n large
$\mathbf{B}_i x$ expensive

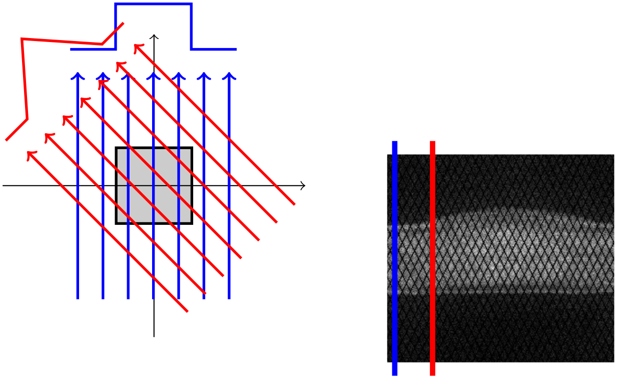
3) Numerical Evaluation: PET imaging



PET Modelling

$$b_i \sim \text{Poisson}(a_i^T u + r_i)$$

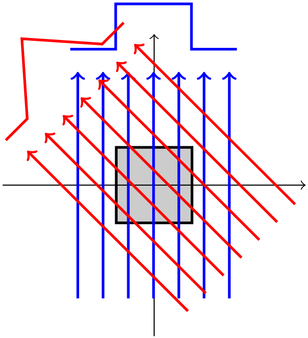
- data $b_i \in \mathbb{N}$
- forward model $a_i^T u \approx \gamma_i \int_{L_i} u$ (x-ray transform)
- multiplicative correction $\gamma_i > 0$ (attenuation, normalisation)
- background $r_i > 0$ (scatter, randoms)

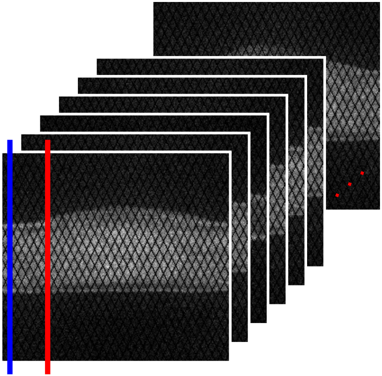


PET Modelling

$$b_i \sim \text{Poisson}(a_i^T u + r_i)$$

- data $b_i \in \mathbb{N}$
- forward model $a_i^T u \approx \gamma_i \int_{L_i} u$ (x-ray transform)
- multiplicative correction $\gamma_i > 0$ (attenuation, normalisation)
- background $r_i > 0$ (scatter, randoms)
- number of data / rays: 2D $N = 86k$, 3D $N = 355M$





PET Reconstruction[1]

$$u_\lambda \in \arg\min_u \left\{ \sum_{j=1}^m \mathcal{D}_j(\mathbf{A}_j u + r_j) + \lambda \mathcal{R}(u) + \imath_+(u) \right\}$$

- Partition data in "subsets" $\mathbb{S}_1, \ldots, \mathbb{S}_m$

$$\mathcal{D}_j(y) := \sum_{i \in \mathbb{S}_j} \mathrm{KL}(y_i; b_i)$$

- Kullback–Leibler divergence

$$\mathrm{KL}(y; b) = \begin{cases} y - b + b \log\left(\frac{b}{y}\right) & \text{if } y > 0 \\ \infty & \text{else} \end{cases}$$

- Regularizer $\mathcal{R}$, see next page
- Constraint

$$\imath_+(u) = \begin{cases} 0, & \text{if } u_i \geq 0 \text{ for all } i \\ \infty, & \text{else} \end{cases}$$

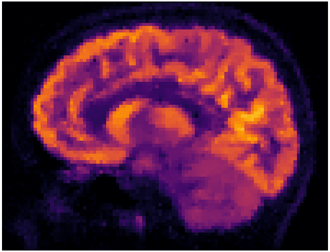
[1] Brune '10, Brune et al. '10, Setzer et al. '10, Müller et al. '11, Anthoine et al. '12, Knoll et al. '16, Ehrhardt et al. '16, Hohage and Werner '16, Schramm et al. '17, Rasch et al. '17, Ehrhardt et al. '17, Mehranian et al. '17 and many, many more

PET Reconstruction with TV

Total variation (TV)

Rudin, Osher, Fatemi 1992

$$\mathcal{R}(u) = \|\nabla u\|_1$$

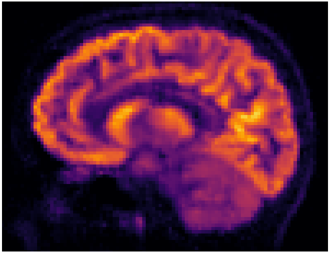


$$\min_u \left\{ \sum_{j=1}^{m} \mathcal{D}_j(\mathbf{A}_j u) + \lambda \|\nabla u\|_1 + \imath_+(u) \right\}$$

$$\min_x \left\{ \sum_{i=1}^{n} f_i(\mathbf{B}_i x) + g(x) \right\}$$

$$\begin{array}{lll} n = m + 1 & g(x) = \imath_+(x) & \\ \mathbf{B}_i = \mathbf{A}_i & f_i = \mathcal{D}_i & i \in [m] \\ \mathbf{B}_n = \nabla & f_n = \lambda \| \cdot \|_1 & \end{array}$$

PET Reconstruction with TGV



Total generalized variation (TGV)

Bredies, Kunisch, Pock 2010

$$\mathcal{R}(u) = \min_v \|\nabla u - v\|_1 + \beta \|\mathbf{D} v\|_1$$

$$\min_{u,v} \left\{ \sum_{j=1}^{m} \mathcal{D}_j(\mathbf{A}_j u) + \lambda \|\nabla u - v\|_1 + \lambda\beta \|\mathbf{D} v\|_1 + \imath_+(u) \right\}$$

$$\min_x \left\{ \sum_{i=1}^{n} f_i(\mathbf{B}_i x) + g(x) \right\}$$

$n = m + 2$

$x = (u; v)$	$g(x) = \imath_+(u)$
$\mathbf{B}_i = (\mathbf{A}_i, 0)$	$f_i = \mathcal{D}_i \quad i \in [m]$
$\mathbf{B}_{n-1} = (\nabla, -\mathbf{I})$	$f_{n-1} = \lambda \| \cdot \|_1$
$\mathbf{B}_n = (0, \mathbf{D})$	$f_n = \lambda\beta \| \cdot \|_1$

Observations

$$x^\sharp \in \arg\min_x \left\{ \sum_{i=1}^n f_i(\mathbf{B}_i x) + g(x) \right\}$$

- **Proper** $f : X \mapsto \mathbb{R} \cup \{\infty\}$ and $f \not\equiv \infty$, **convex** and **lower semi-continuous (lsc)** $x_k \to x$ then $f(x) \leq \liminf_{k\to\infty} f(x_k)$
- $f(z) = \sum_i f_i(z_i)$ is "**separable**". Not separable in x.
- f_i, g are **non-smooth** but **proximal operator** has closed-form

  $$\mathrm{prox}_f^{\mathbf{T}}(x) = \arg\min_z \left\{ \frac{1}{2}\|z - x\|_{\mathbf{T}}^2 + f(z) \right\}, \quad \|x\|_{\mathbf{T}}^2 := \langle \mathbf{T}^{-1}x, x \rangle$$

  Note: $\mathrm{prox}_f^{\tau \mathbf{I}} = \mathrm{prox}_{\tau f}^1$

Observations

$$x^\sharp \in \arg\min_x \left\{ \sum_{i=1}^n f_i(\mathbf{B}_i x) + g(x) \right\}$$

- **Proper** $f : X \mapsto \mathbb{R} \cup \{\infty\}$ and $f \not\equiv \infty$, **convex** and **lower semi-continuous (lsc)** $x_k \to x$ then $f(x) \le \liminf_{k\to\infty} f(x_k)$
- $f(z) = \sum_i f_i(z_i)$ is "**separable**". Not separable in x.
- f_i, g are **non-smooth** but **proximal operator** has closed-form

  $$\mathrm{prox}_f^{\mathbf{T}}(x) = \arg\min_z \left\{ \tfrac{1}{2}\|z - x\|_{\mathbf{T}}^2 + f(z) \right\}, \quad \|x\|_{\mathbf{T}}^2 := \langle \mathbf{T}^{-1}x, x \rangle$$

  Note: $\mathrm{prox}_f^{\tau \mathbf{I}} = \mathrm{prox}_{\tau f}^1$

Problem 1: Cannot compute $\mathrm{prox}_{f_i \circ \mathbf{B}_i}$
Problem 2: n is large and/or $\mathbf{B}_i x$ expensive

Algorithm

The way out: Saddle Point Problem

$$x^\sharp \in \arg\min_x \left\{ \sum_{i=1}^n f_i(\mathbf{B}_i x) + g(x) \right\}$$

- $f(y) := \sum_i f_i(y_i)$, $\mathbf{B} = [\mathbf{B}_1; \ldots; \mathbf{B}_n]$

$$x^\sharp \in \arg\min_x \{f(\mathbf{B}x) + g(x)\}$$

The way out: Saddle Point Problem

$$x^\sharp \in \arg\min_x \left\{ \sum_{i=1}^n f_i(\mathbf{B}_i x) + g(x) \right\}$$

- $f(y) := \sum_i f_i(y_i)$, $\mathbf{B} = [\mathbf{B}_1; \ldots; \mathbf{B}_n]$

$$x^\sharp \in \arg\min_x \{ f(\mathbf{B}x) + g(x) \}$$

Definition: The **convex conjugate** of f is given by

$$f^*(y) := \sup_z \langle z, y \rangle - f(z).$$

Theorem: Let f be proper, convex and lsc, then

$$f(z) = (f^*)^*(z) = \sup_y \langle z, y \rangle - f^*(y).$$

The way out: Saddle Point Problem

$$x^\sharp \in \arg\min_x \left\{ \sum_{i=1}^n f_i(\mathbf{B}_i x) + g(x) \right\}$$

- $f(y) := \sum_i f_i(y_i)$, $\mathbf{B} = [\mathbf{B}_1; \ldots; \mathbf{B}_n]$

$$x^\sharp \in \arg\min_x \{f(\mathbf{B}x) + g(x)\}$$

Definition: The **convex conjugate** of f is given by

$$f^*(y) := \sup_z \langle z, y\rangle - f(z).$$

Theorem: Let f be proper, convex and lsc, then

$$f(z) = (f^*)^*(z) = \sup_y \langle z, y\rangle - f^*(y).$$

$$(x^\sharp, y^\sharp) \in \arg\min_x \sup_y \left\{ \langle \mathbf{B}x, y\rangle - f^*(y) + g(x) \right\}$$

Primal-Dual Hybrid Gradient (PDHG) Algorithm[1]

Given $x^0, y^0, \overline{y}^0 = y^0$
Iterate

(1) $x^{k+1} = \mathrm{prox}_g^{\mathbf{T}}(x^k - \mathbf{T}\mathbf{B}^*\overline{y}^k)$

(2) $y^{k+1} = \mathrm{prox}_{f^*}^{\mathbf{S}}(y^k + \mathbf{S}\mathbf{B}x^{k+1})$

(3) $\overline{y}^{k+1} = y^{k+1} + \theta(y^{k+1} - y^k)$

- evaluation of $\mathbf{B}$ and $\mathbf{B}^*$
- proximal operator
- convergence: $\theta = 1, \|\mathbf{S}^{1/2}\mathbf{B}\mathbf{T}^{1/2}\|^2 < 1$, cf. $\sigma\tau\|\mathbf{B}\|^2 < 1$

[1] Pock, Cremers, Bischof, Chambolle '09, Esser, Zhang, Chan '10, Chambolle and Pock '11

Primal-Dual Hybrid Gradient (PDHG) Algorithm[1]

Given $x^0, y^0, \overline{y}^0 = y^0$
Iterate

(1) $x^{k+1} = \mathrm{prox}_g^{\mathbf{T}}(x^k - \mathbf{T}\mathbf{B}^*\overline{y}^k)$

(2) $y^{k+1} = \mathrm{prox}_{f^*}^{\mathbf{S}}(y^k + \mathbf{S}\mathbf{B}x^{k+1})$

(3) $\overline{y}^{k+1} = y^{k+1} + \theta(y^{k+1} - y^k)$

- evaluation of $\mathbf{B}$ and $\mathbf{B}^*$
- proximal operator
- convergence: $\theta = 1, \|\mathbf{S}^{1/2}\mathbf{B}\mathbf{T}^{1/2}\|^2 < 1$, cf. $\sigma\tau\|\mathbf{B}\|^2 < 1$
- $z^k = \mathbf{B}^* y^k$

[1] Pock, Cremers, Bischof, Chambolle '09, Esser, Zhang, Chan '10, Chambolle and Pock '11

Primal-Dual Hybrid Gradient (PDHG) Algorithm[1]

Given $x^0, y^0, z^0 = \bar{z}^0 = \mathbf{B}^* y^0$, e.g. all equal 0

Iterate

(1) $x^{k+1} = \text{prox}_g^{\mathbf{T}}(x^k - \mathbf{T}\bar{z}^k)$

(2) $y^{k+1} = \text{prox}_{f^*}^{\mathbf{S}}(y^k + \mathbf{S}\mathbf{B}x^{k+1})$

(3) $z^{k+1} = \mathbf{B}^* y^{k+1}$

(4) $\bar{z}^{k+1} = z^{k+1} + \theta(z^{k+1} - z^k)$

- evaluation of $\mathbf{B}$ and $\mathbf{B}^*$
- proximal operator
- convergence: $\theta = 1, \|\mathbf{S}^{1/2}\mathbf{B}\mathbf{T}^{1/2}\|^2 < 1$, cf. $\sigma\tau\|\mathbf{B}\|^2 < 1$
- $z^k = \mathbf{B}^* y^k$

[1] Pock, Cremers, Bischof, Chambolle '09, Esser, Zhang, Chan '10, Chambolle and Pock '11

Primal-Dual Hybrid Gradient (PDHG) Algorithm[1]

Given $x^0, y^0, z^0 = \overline{z}^0 = \mathbf{B}^* y^0$, e.g. all equal 0

Iterate

(1) $x^{k+1} = \text{prox}_g^{\mathbf{T}}(x^k - \mathbf{T}\overline{z}^k)$

(2) $y_i^{k+1} = \text{prox}_{f_i^*}^{\mathbf{S}_i}(y_i^k + \mathbf{S}_i \mathbf{B}_i x^{k+1}) \quad i = 1, \ldots, n$

(3) $z^{k+1} = \sum_{i=1}^n \mathbf{B}_i^* y_i^{k+1}$

(4) $\overline{z}^{k+1} = z^{k+1} + \theta(z^{k+1} - z^k)$

- $f(y) := \sum_i f_i(y_i)$, $[\text{prox}_{f^*}(y)]_i = \text{prox}_{f_i^*}(y_i)$
- $\mathbf{B} = [\mathbf{B}_1; \ldots; \mathbf{B}_n]^T$, $\mathbf{B}^* y = \sum_{i=1}^n \mathbf{B}_i^* y_i$
- $z^k = \sum_{i=1}^n \mathbf{B}_i^* y_i^k$

[1] Pock, Cremers, Bischof, Chambolle '09, Esser, Zhang, Chan '10, Chambolle and Pock '11

Stochastic PDHG Algorithm[1]

Given $x^0, y^0, z^0 = \overline{z}^0 = \mathbf{B}^* y^0$, e.g. all equal 0

Iterate

(1) $x^{k+1} = \text{prox}_g^{\mathbf{T}}(x^k - \mathbf{T}\overline{z}^k)$

(2) $y_i^{k+1} = \text{prox}_{f_i^*}^{\mathbf{S}_i}(y_i^k + \mathbf{S}_i \mathbf{B}_i x^{k+1}) \quad i = 1, \ldots, n$

(3) $z^{k+1} = \sum_{i=1}^n \mathbf{B}_i^* y_i^{k+1}$

(4) $\overline{z}^{k+1} = z^{k+1} + \theta(z^{k+1} - z^k)$

[1] Chambolle, E, Richtárik, Schönlieb '18

Stochastic PDHG Algorithm[1]

Given $x^0, y^0, z^0 = \overline{z}^0 = \mathbf{B}^* y^0$, e.g. all equal 0
Iterate

(1) $x^{k+1} = \operatorname{prox}_g^{\mathbf{T}}(x^k - \mathbf{T}\overline{z}^k)$

Select $j \in \{1, \dots, n\}$ with probability p_j.

(2) $y_i^{k+1} = \begin{cases} \operatorname{prox}_{f_i^*}^{\mathbf{S}}(y_i^k + \mathbf{S}_i \mathbf{B}_i x^{k+1}) & i = j \\ y_i^k & \text{else} \end{cases}$

(3) $z^{k+1} = \sum_{i=1}^n \mathbf{B}_i^* y_i^{k+1}$

(4) $\overline{z}^{k+1} = z^{k+1} + \theta(z^{k+1} - z^k)$

[1] Chambolle, E, Richtárik, Schönlieb '18

Stochastic PDHG Algorithm[1]

Given $x^0, y^0, z^0 = \overline{z}^0 = \mathbf{B}^* y^0$, e.g. all equal 0

Iterate

(1) $x^{k+1} = \mathrm{prox}_g^{\mathbf{T}}(x^k - \mathbf{T}\overline{z}^k)$

Select $j \in \{1, \ldots, n\}$ with probability p_j.

(2) $y_i^{k+1} = \begin{cases} \mathrm{prox}_{f_i^*}^{\mathbf{S}}(y_i^k + \mathbf{S}_i \mathbf{B}_i x^{k+1}) & i = j \\ y_i^k & \text{else} \end{cases}$

(3) $z^{k+1} = \sum_{i=1}^n \mathbf{B}_i^* y_i^{k+1} = z^k + \mathbf{B}_j^* \left(y_j^{k+1} - y_j^k\right)$

(4) $\overline{z}^{k+1} = z^{k+1} + \theta(z^{k+1} - z^k)$

[1] Chambolle, E, Richtárik, Schönlieb '18

Stochastic PDHG Algorithm[1]

Given $x^0, y^0, z^0 = \overline{z}^0 = \mathbf{B}^* y^0$, e.g. all equal 0
Iterate

(1) $x^{k+1} = \text{prox}_g^{\mathbf{T}}(x^k - \mathbf{T}\overline{z}^k)$

Select $j \in \{1, \dots, n\}$ with probability p_j.

(2) $y_i^{k+1} = \begin{cases} \text{prox}_{f_i^*}^{\mathbf{S}}(y_i^k + \mathbf{S}_i \mathbf{B}_i x^{k+1}) & i = j \\ y_i^k & \text{else} \end{cases}$

(3) $z^{k+1} = \sum_{i=1}^n \mathbf{B}_i^* y_i^{k+1} = z^k + \mathbf{B}_j^* \left(y_j^{k+1} - y_j^k \right)$

(4) $\overline{z}^{k+1} = z^{k+1} + \frac{\theta}{p_j}(z^{k+1} - z^k)$

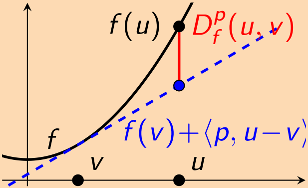
unbiased: $\mathbb{E}_j \frac{\theta}{p_j}(z^{k+1} - z^k) =$ deterministic update with all data

[1] Chambolle, E, Richtárik, Schönlieb '18

Convergence of SPDHG

Definition:
The **Bregman distance** of f at u, v and $p \in \partial f(v)$ is defined as

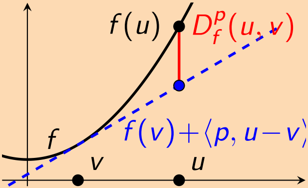
$$D_f^p(u,v) = f(u) - \Big[f(v) + \langle p, u - v \rangle \Big] .$$



Convergence of SPDHG

Definition:
The **Bregman distance** of f at u, v and $p \in \partial f(v)$ is defined as

$$D_f^p(u,v) = f(u) - \Big[f(v) + \langle p, u - v\rangle\Big] \,.$$



Theorem: Chambolle, E, Richtárik, Schönlieb '18
Let $(x^\sharp, y^\sharp)$ be a saddle point, $\theta = 1$ and choose $\mathbf{S}_i, \mathbf{T} := \min_i \mathbf{T}_i$ such that $\|\mathbf{S}_i^{1/2}\mathbf{B}_i\mathbf{T}_i^{1/2}\|^2 < p_i, i = 1, \ldots, n$.

Then almost surely

$$D_g^{r^\sharp}(x^k, x^\sharp) + D_{f^*}^{q^\sharp}(y^k, y^\sharp) \to 0$$

and the ergodic sequence $(X^k, Y^k) = \frac{1}{k}\sum_{j=1}^k (x^j, y^j)$ converges with **rate**

$$\mathbb{E}\left\{D_g^{r^\sharp}(X^k, x^\sharp) + D_{f^*}^{q^\sharp}(Y^k, y^\sharp)\right\} \leq \frac{C}{k} \,.$$

Step-sizes and Preconditioning

Theorem: E, Markiewicz, Schönlieb '19

Let $\rho < 1$ and $\gamma > 0$. Then $\|\mathbf{S}_i^{1/2}\mathbf{B}_i\mathbf{T}_i^{1/2}\|^2 < p_i$ is satisfied by

$$\mathbf{S}_i = \frac{\gamma\rho}{\|\mathbf{B}_i\|}\mathbf{I}\,, \quad \mathbf{T}_i = \frac{\rho p_i}{\gamma\|\mathbf{B}_i\|}\mathbf{I}\,.$$

If $\mathbf{B}_i \geq 0$, then the step-size condition is also satisfied for

$$\mathbf{S}_i = \mathrm{diag}\left(\frac{\gamma\rho}{\mathbf{B}_i 1}\right)\,, \quad \mathbf{T}_i = \mathrm{diag}\left(\frac{\rho p_i}{\gamma\mathbf{B}_i^T 1}\right)\,.$$

Step-sizes and Preconditioning

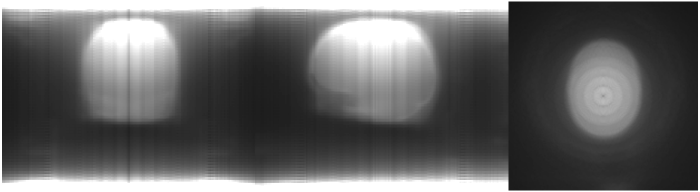
Theorem: E, Markiewicz, Schönlieb '19

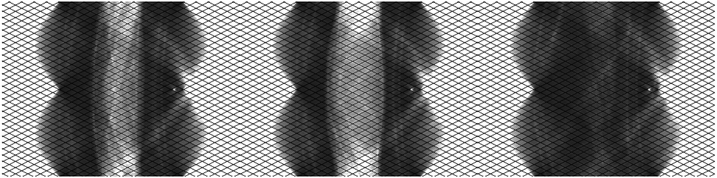
Let $\rho < 1$ and $\gamma > 0$. Then $\|\mathbf{S}_i^{1/2}\mathbf{B}_i\mathbf{T}_i^{1/2}\|^2 < p_i$ is satisfied by

$$\mathbf{S}_i = \frac{\gamma\rho}{\|\mathbf{B}_i\|}\mathbf{I}\,, \quad \mathbf{T}_i = \frac{\rho p_i}{\gamma\|\mathbf{B}_i\|}\mathbf{I}\,.$$

If $\mathbf{B}_i \geq 0$, then the step-size condition is also satisfied for

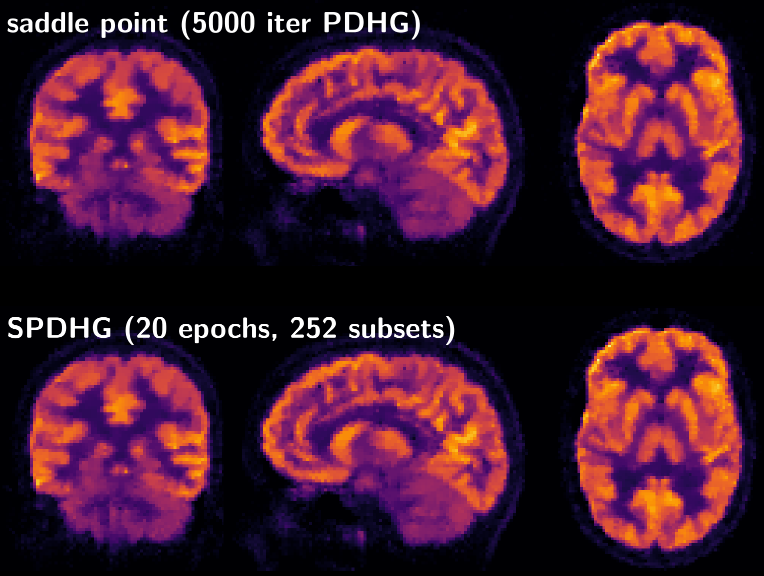
$$\mathbf{S}_i = \mathrm{diag}\left(\frac{\gamma\rho}{\mathbf{B}_i 1}\right)\,, \quad \mathbf{T}_i = \mathrm{diag}\left(\frac{\rho p_i}{\gamma \mathbf{B}_i^T 1}\right)\,.$$





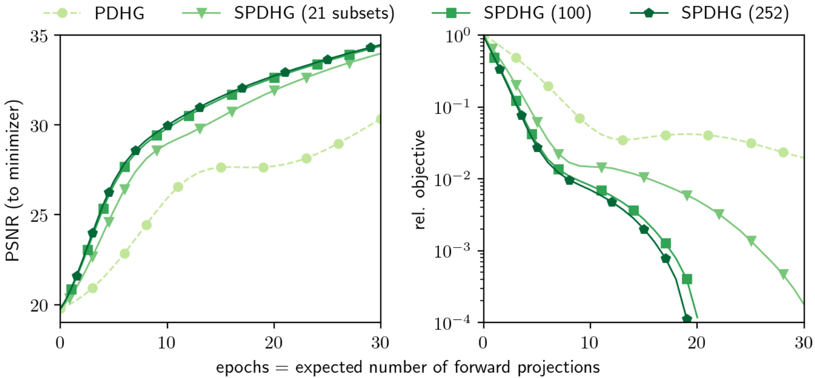
Applications

Sanity Check: Convergence to Saddle Point (TV)



More subsets are faster

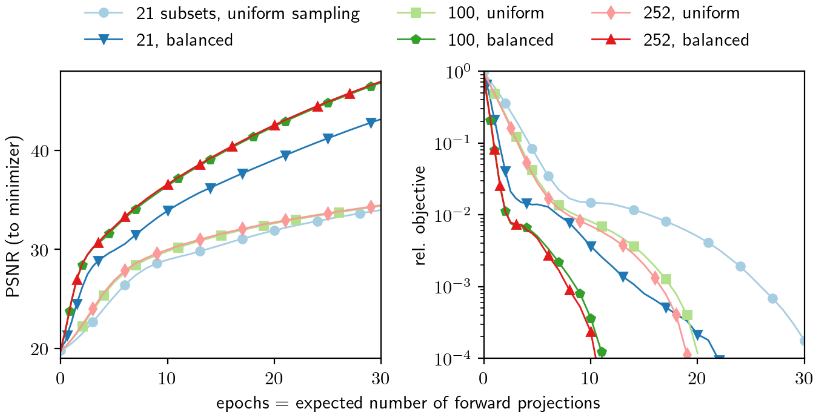
$$m = 1, 21, 100, 252$$



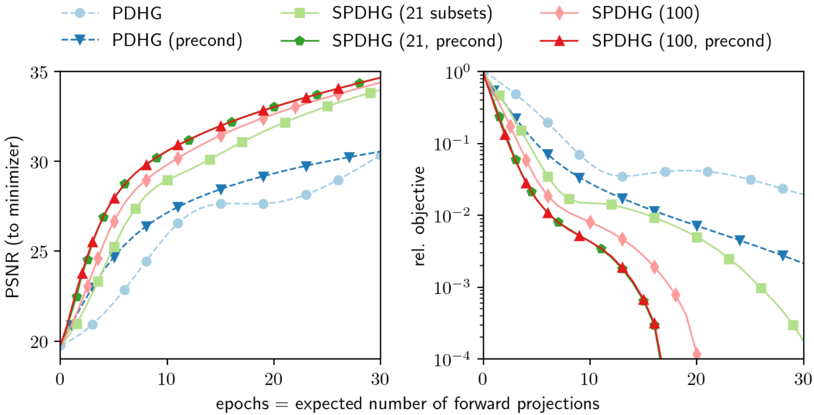
"Balanced sampling" is faster

$$\text{uniform sampling: } p_i = 1/n$$

$$\text{balanced sampling: } p_i = \begin{cases} \frac{1}{2m} & i < n \\ \frac{1}{2} & i = n \end{cases}$$

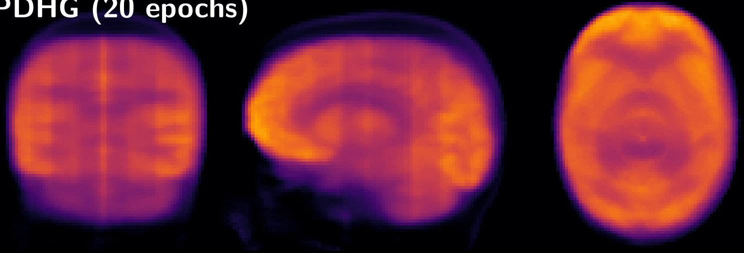


Preconditioning is faster

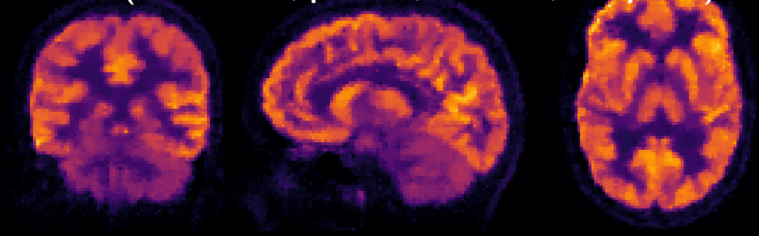


Faster than PDHG, TV

PDHG (20 epochs)

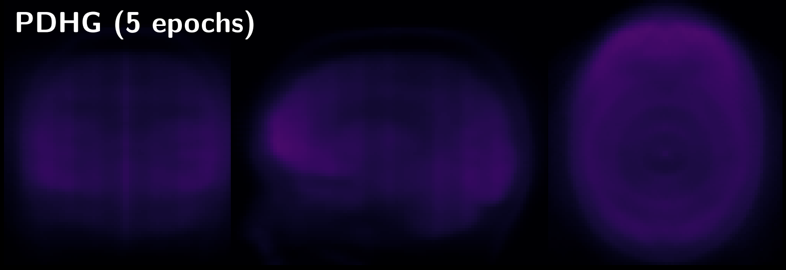


SPDHG (252 subsets, precond, balanced, 20 epochs)

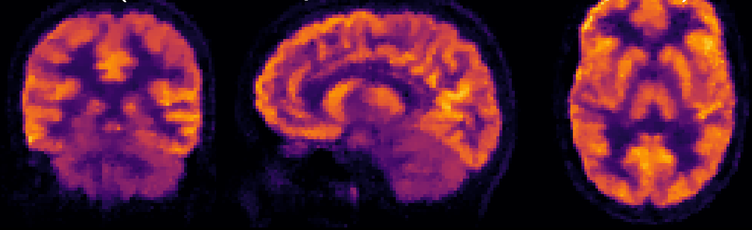


Faster than PDHG, TV

PDHG (5 epochs)



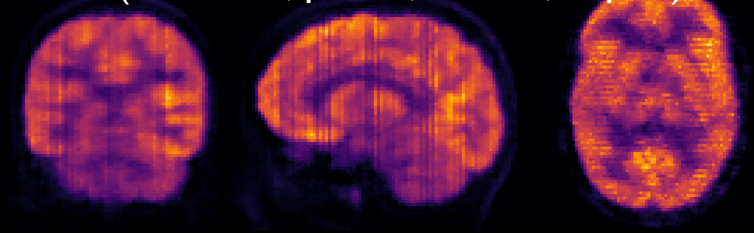
SPDHG (252 subsets, precond, balanced, 5 epochs)



Faster than PDHG, TV

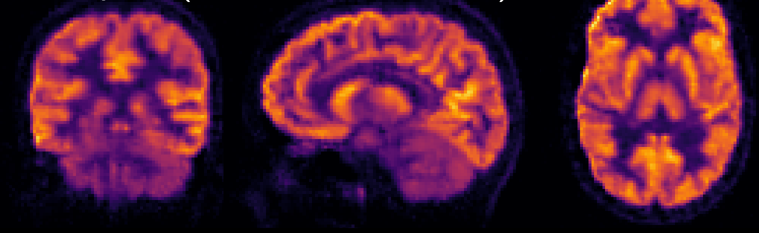
PDHG (1 epoch)

SPDHG (252 subsets, precond, balanced, 1 epoch)

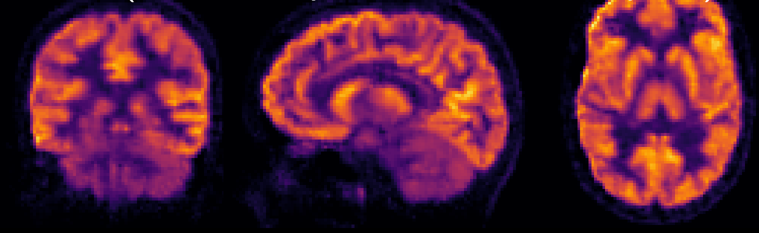


Total Generalized Variation

saddle point (PDHG, 5000 iterations)

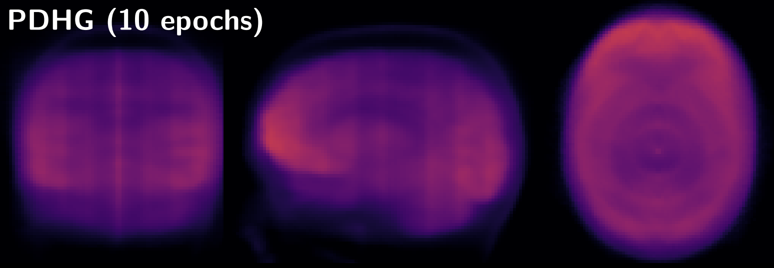


SPDHG (252 subsets, precond, balanced, 10 epochs)

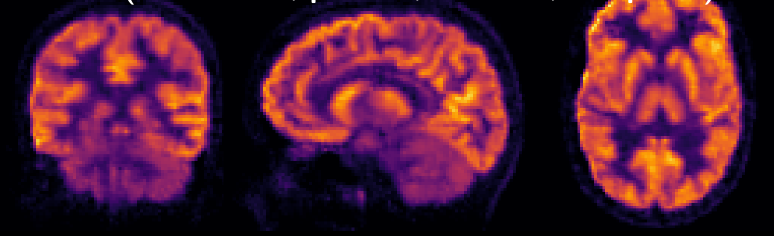


Total Generalized Variation

PDHG (10 epochs)



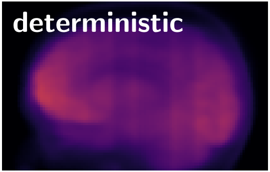
SPDHG (252 subsets, precond, balanced, 10 epochs)

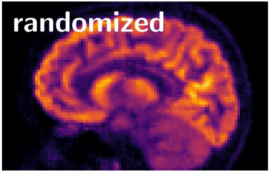


Conclusions and Outlook

Summary:

- **Randomized** optimization for cost functionals with "separable structure"
- **Generalisation** of PDHG ($n = 1$)
- **Randomization**, **preconditioning** and **balanced sampling** all speed up SPDHG
- **Much faster** PET reconstruction: advanced models feasible for clinical data





Future work:

- almost sure convergence of iterates
- biased extrapolation
- sampling: 1) optimal, 2) adaptive

Conclusions and Outlook

Summary:

- **Randomized** optimization for cost functionals with "separable structure"
- **Generalisation** of PDHG ($n = 1$)
- **Randomization**, **preconditioning** and **balanced sampling** all speed up SPDHG
- **Much faster** PET reconstruction: advanced models feasible for clinical data

Future work:

- almost sure convergence of iterates
- biased extrapolation
- sampling: 1) optimal, 2) adaptive

Job Opportunity:

- PostDoc in Bath, UK on imaging and machine learning; talk to me!