Faster PET Reconstruction with Non-Smooth Anatomical Priors by Randomization and Preconditioning

Matthias J. Ehrhardt

Institute for Mathematical Innovation
University of Bath, UK

November 4, 2019

Joint work with:
Mathematics: Chambolle (Paris), Richtárik (KAUST), Schönlieb (Cambridge)
PET imaging: Markiewicz, Schott (both UCL)

Institute for Mathematical Innovation | UNIVERSITY OF BATH | EPSRC Engineering and Physical Sciences Research Council | THE FARADAY INSTITUTION

Outline

1) PET reconstruction via Optimization

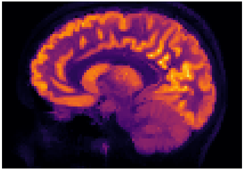
$$\sum_{i=1}^{n} f_i(\mathbf{B}_i x) + g(x)$$

2) Randomized Algorithm for Convex Optimization

non-smooth

$\mathbf{B}_i x$ expensive

3) Numerical Evaluation: clinical PET imaging



PET Reconstruction[1]

$$u_\lambda \in \arg\min_u \left\{ \sum_{i=1}^N \mathrm{KL}(b_i; \mathbf{A}_i u + r_i) + \lambda \mathcal{R}(u; v) + \imath_+(u) \right\}$$

- **Kullback–Leibler** divergence

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

- Nonnegativity **constraint**

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

- **Regularizer**: e.g. $\mathcal{R}(u; v) = \mathrm{TV}(u)$

[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[1]

$$u_\lambda \in \arg\min_u \left\{ \sum_{i=1}^N \mathrm{KL}(b_i; \mathbf{A}_i u + r_i) + \lambda \mathcal{R}(u; v) + \imath_+(u) \right\}$$

- **Kullback–Leibler** divergence

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

- Nonnegativity **constraint**

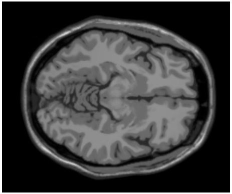
$$\imath_+(u) = \begin{cases} 0, & \text{if } u_i \geq 0 \text{ for all } i \\ \infty, & \text{else} \end{cases}$$

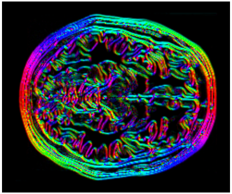
- **Regularizer**: e.g. $\mathcal{R}(u; v) = \mathrm{TV}(u)$

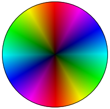
How to incorporate MRI information into $\mathcal{R}$?

[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

Directional Total Variation



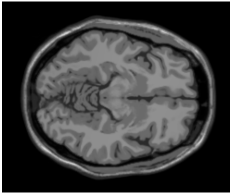


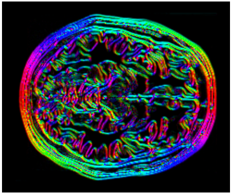


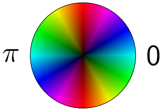
Let $\|\nabla v\| = 1$. Then u and v have **Parallel Level Sets** iff

$$u \sim v \quad \Leftrightarrow \quad \nabla u \parallel \nabla v \quad \Leftrightarrow \quad \nabla u - \langle \nabla u, \nabla v \rangle \nabla v = 0$$

Directional Total Variation







Let $\|\nabla v\| = 1$. Then u and v have **Parallel Level Sets** iff

$$u \sim v \quad \Leftrightarrow \quad \nabla u \parallel \nabla v \quad \Leftrightarrow \quad \nabla u - \langle \nabla u, \nabla v \rangle \nabla v = 0$$

Definition: The **Directional Total Variation (dTV)** of u is

$$\mathrm{dTV}(u) := \sum_i \|[\mathbf{I} - \xi_i \xi_i^T] \nabla u_i\|, \quad 0 \leq \|\xi_i\| \leq 1$$

Ehrhardt and Betcke '16, related to Kaipio et al. '99, Bayram and Kamasak '12

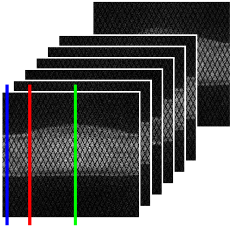
- If $\xi_i = 0$, then $\mathrm{dTV} = \mathrm{TV}$.
- $\xi_i = \frac{\nabla v_i}{\|\nabla v_i\|_\eta}, \quad \|\nabla v_i\|_\eta^2 = \|\nabla v_i\|^2 + \eta^2, \quad \eta > 0$

PET Reconstruction

Partition data in **subsets** $\mathbb{S}_j$:

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

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

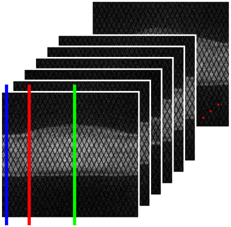


PET Reconstruction

Partition data in **subsets** $\mathbb{S}_j$:

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

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



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

$n = m + 1 \quad g(x) = \imath_+(x)$

$\mathbf{B}_i = \mathbf{A}_i \quad f_i = \mathcal{D}_i \quad i = 1, \dots, m$

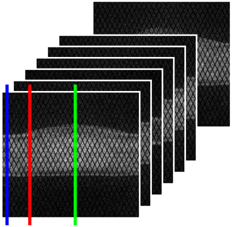
$\mathbf{B}_n = \mathbf{D}\nabla \quad f_n = \lambda \| \cdot \|_1$

PET Reconstruction

Partition data in **subsets** $\mathbb{S}_j$:

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

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



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

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

- f_i, g are **non-smooth** but can compute **proximal operator**

$$\mathrm{prox}_f(x) := \arg\min_z \left\{ \frac{1}{2}\|z - x\|^2 + f(z) \right\} .$$

- **Cannot compute** proximal operator of $f_i \circ \mathbf{B}_i$
- $\mathbf{B}_i x$ is **expensive** to compute

Optimization

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

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

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

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

(3) $\overline{y}_i^{k+1} = y_i^{k+1} + \theta(y_i^{k+1} - y_i^k) \quad i = 1, \ldots, n$

- evaluation of $\mathbf{B}_i$ and $\mathbf{B}_i^*$
- proximal operator: $\mathrm{prox}_f^{\mathbf{S}}(x) := \arg\min_z \left\{ \frac{1}{2} \|z - x\|_{\mathbf{S}}^2 + f(z) \right\}$
- convergence: $\theta = 1, \mathbf{C}_i = \mathbf{S}_i^{1/2} \mathbf{B}_i \mathbf{T}^{1/2}$

$$\left\| \begin{pmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_n \end{pmatrix} \right\|^2 < 1$$

[1] Pock, Cremers, Bischof, Chambolle '09, Chambolle and Pock '11

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

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

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

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

(3) $\overline{y}_i^{k+1} = y_i^{k+1} + \theta(y_i^{k+1} - y_i^k) \quad i = 1, \ldots, n$

- evaluation of $\mathbf{B}_i$ and $\mathbf{B}_i^*$
- proximal operator: $\mathrm{prox}_f^{\mathbf{S}}(x) := \arg\min_z \left\{ \frac{1}{2}\|z - x\|_{\mathbf{S}}^2 + f(z) \right\}$
- convergence: $\theta = 1, \mathbf{C}_i = \mathbf{S}_i^{1/2} \mathbf{B}_i \mathbf{T}^{1/2}$

$$\left\| \begin{pmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_n \end{pmatrix} \right\|^2 < 1$$

[1] Pock, Cremers, Bischof, Chambolle '09, Chambolle and Pock '11

Stochastic PDHG Algorithm[1]

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

(1) $x^{k+1} = \operatorname{prox}_g^{\mathbf{T}}(x^k - \mathbf{T}\sum_{i=1}^n \mathbf{B}_i^* \overline{y}_i^k)$

Select $j^{k+1} \in \{1, \ldots, n\}$ randomly.

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

(3) $\overline{y}_i^{k+1} = \begin{cases} y_i^{k+1} + \frac{\theta}{p_i}(y_i^{k+1} - y_i^k) & i = j^{k+1} \\ y_i^{k+1} & \text{else} \end{cases}$

- probabilities $p_i := \mathbb{P}(i = j^{k+1}) > 0$ (**proper** sampling)
- Compute $\sum_{i=1}^n \mathbf{B}_i^* \overline{y}_i^k$ using only $\mathbf{B}_i^*$ for $i = j^{k+1}$ + **memory**

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

Stochastic PDHG Algorithm[1]

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

(1) $x^{k+1} = \operatorname{prox}_g^{\mathbf{T}}(x^k - \mathbf{T}\sum_{i=1}^n \mathbf{B}_i^* \overline{y}_i^k)$

Select $j^{k+1} \in \{1, \ldots, n\}$ randomly.

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

(3) $$\overline{y}_i^{k+1} = \begin{cases} y_i^{k+1} + \frac{\theta}{p_i}(y_i^{k+1} - y_i^k) & i = j^{k+1} \\ y_i^{k+1} & \text{else} \end{cases}$$

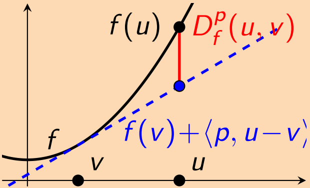
- probabilities $p_i := \mathbb{P}(i = j^{k+1}) > 0$ (**proper** sampling)
- Compute $\sum_{i=1}^n \mathbf{B}_i^* \overline{y}_i^k$ using only $\mathbf{B}_i^*$ for $i = j^{k+1}$ + **memory**
- evaluation of $\mathbf{B}_i$ and $\mathbf{B}_i^*$ **only** for $i = j^{k+1}$.

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

Convergence of SPDHG

Definition: Let $p \in \partial f(v)$. The **Bregman distance** of f is defined as

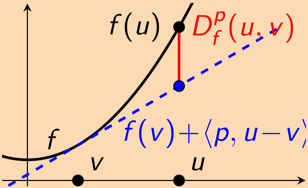
$$D_f^p(u, v) = f(u) - \left[f(v) + \langle p, u - v \rangle \right] .$$



Convergence of SPDHG

Definition: Let $p \in \partial f(v)$. The **Bregman distance** of f 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, choose $\theta = 1$ and step sizes $\mathbf{S}_i, \mathbf{T} := \min_i \mathbf{T}_i$ such that

$$\left\| \mathbf{S}_i^{1/2} \mathbf{B}_i \mathbf{T}_i^{1/2} \right\|^2 < p_i \quad i = 1, \ldots, n.$$

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

Step-sizes and Preconditioning

Theorem: E, Markiewicz, Schönlieb '18

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

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

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

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

Step-sizes and Preconditioning

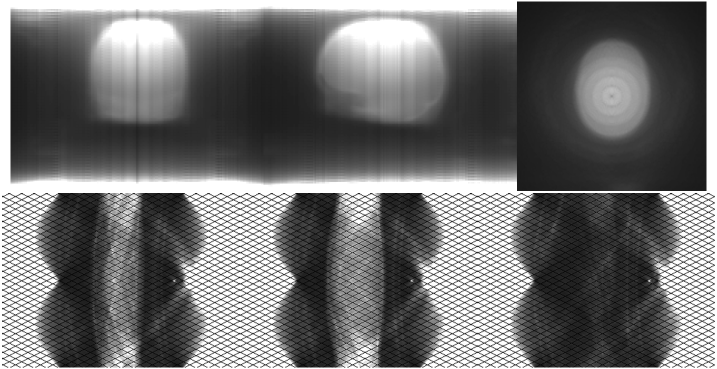
Theorem: E, Markiewicz, Schönlieb '18

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

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

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

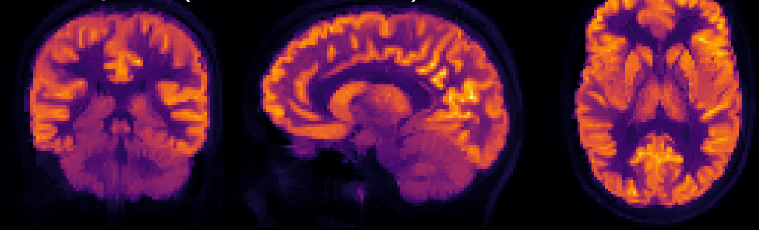
$$\mathbf{S}_i = \operatorname{diag}\left(\frac{\rho}{\mathbf{B}_i 1}\right), \quad \mathbf{T}_i = \operatorname{diag}\left(\frac{p_i}{\mathbf{B}_i^T 1}\right).$$



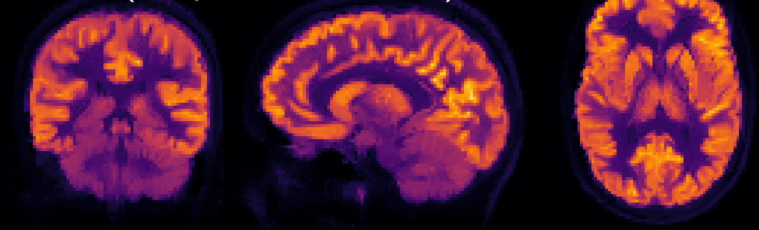
Application

Sanity Check: Convergence to Saddle Point (dTV)

saddle point (5000 iter PDHG)

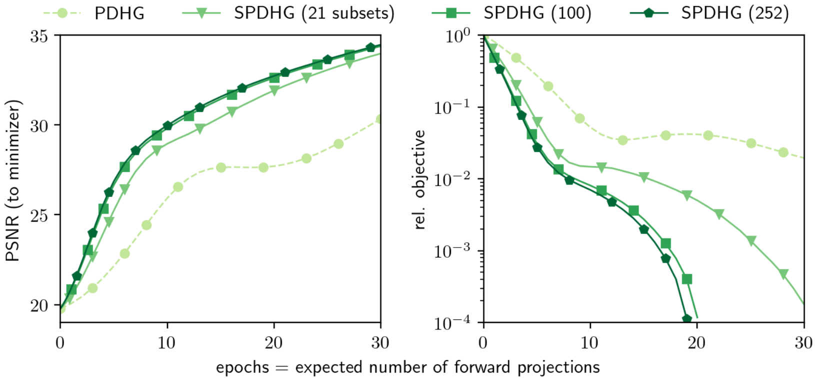


SPDHG (20 epochs, 252 subsets)



More subsets are faster

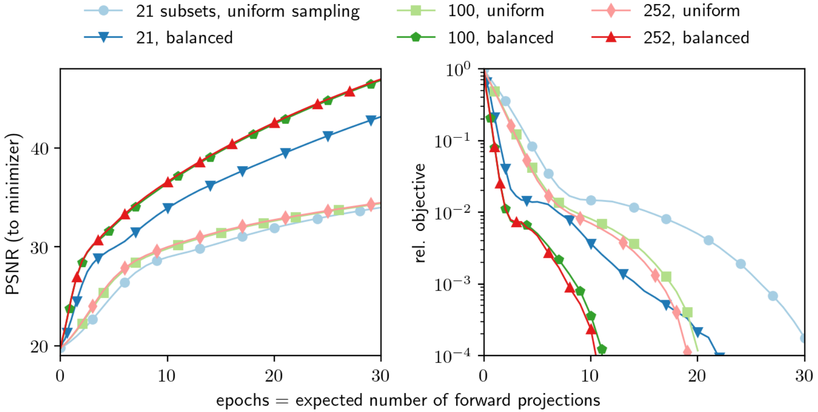
Number of **subsets**: $m = 1, 21, 100, 252$



"Balanced sampling" is faster

uniform sampling: $p_i = 1/n$

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

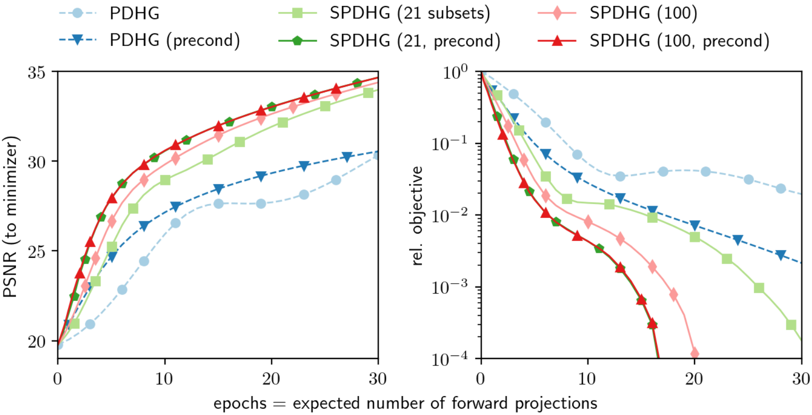


Preconditioning is faster

Scalar step sizes: $\mathbf{S}_i = \frac{\rho}{\|\mathbf{B}_i\|}\mathbf{I}\,, \quad \mathbf{T}_i = \frac{p_i}{\|\mathbf{B}_i\|}\mathbf{I}$

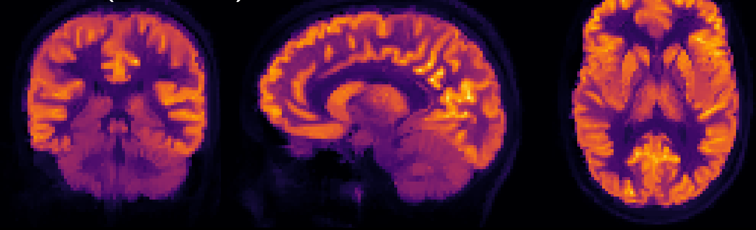
Preconditioned (vector-valued) step sizes:

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

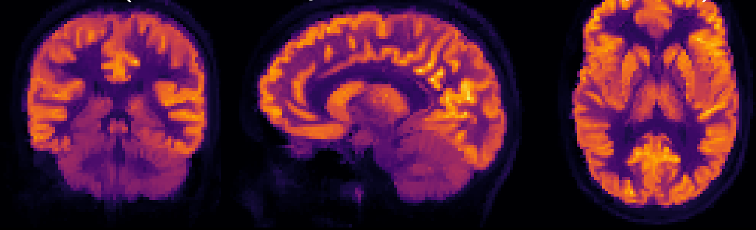


FDG

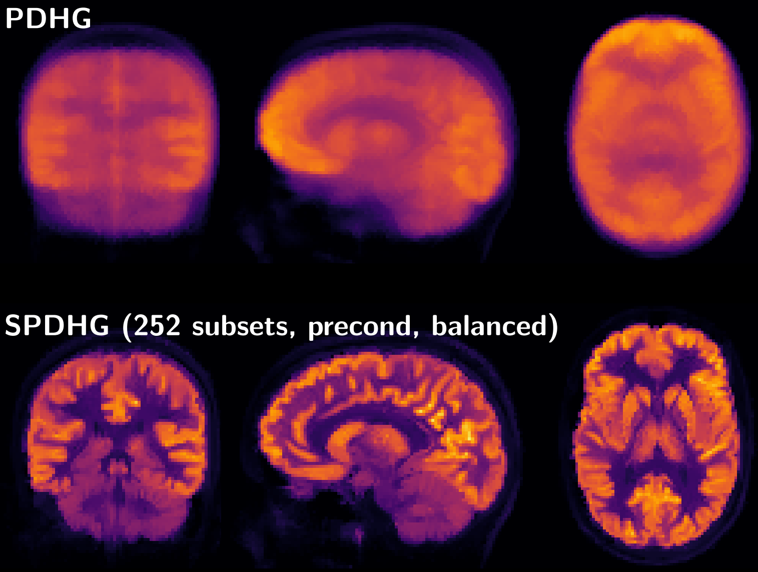
PDHG (5000 iter)



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

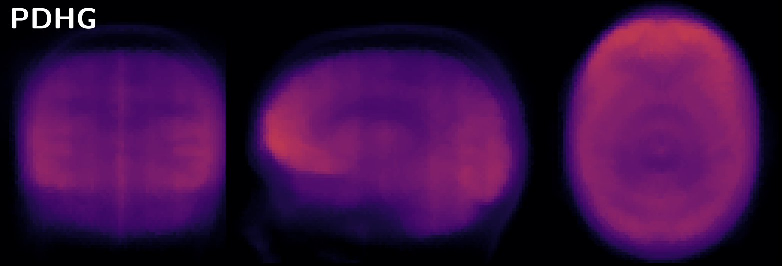


FDG, 20 epochs

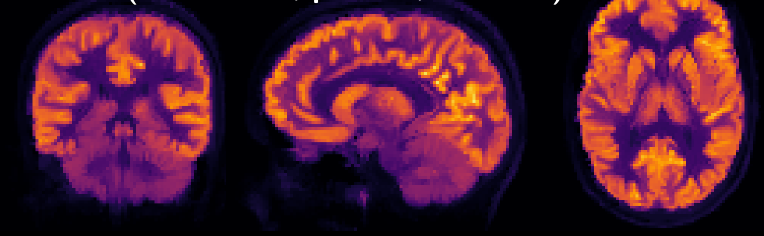


FDG, 10 epochs

PDHG

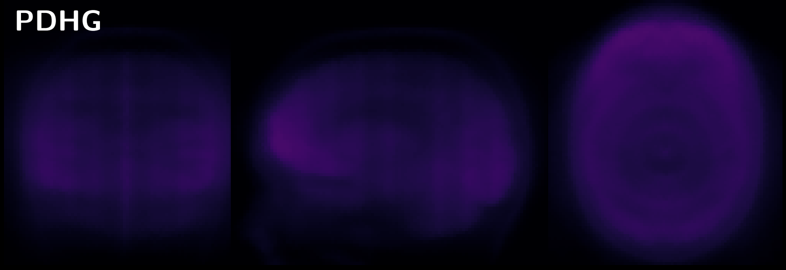


SPDHG (252 subsets, precond, balanced)

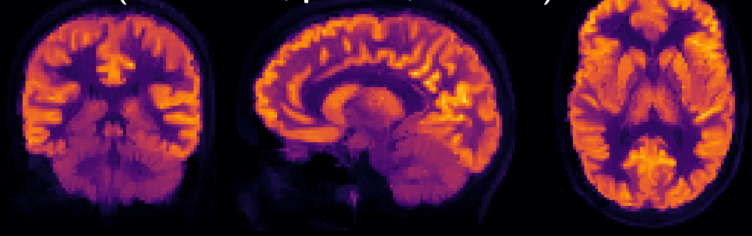


FDG, 5 epochs

PDHG



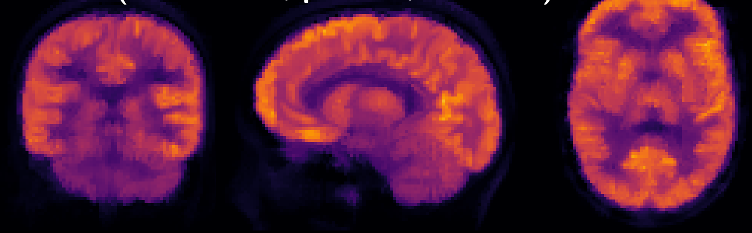
SPDHG (252 subsets, precond, balanced)



FDG, 1 epoch

PDHG

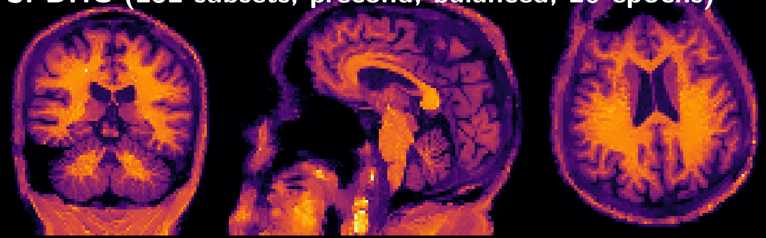
SPDHG (252 subsets, precond, balanced)



Florbetapir

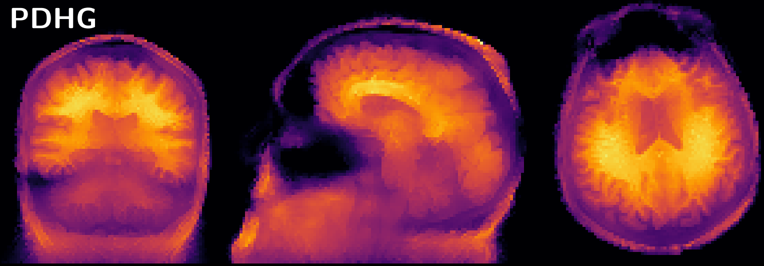
PDHG (5000 iter)

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

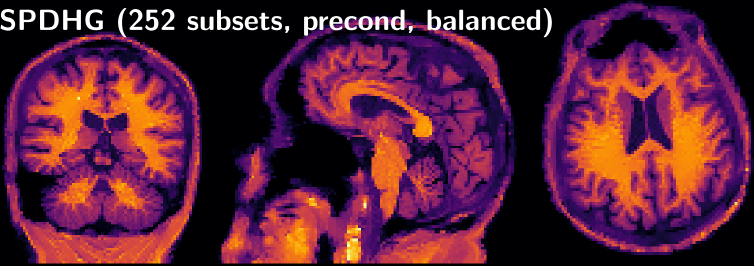


Florbetapir, 20 epochs

PDHG

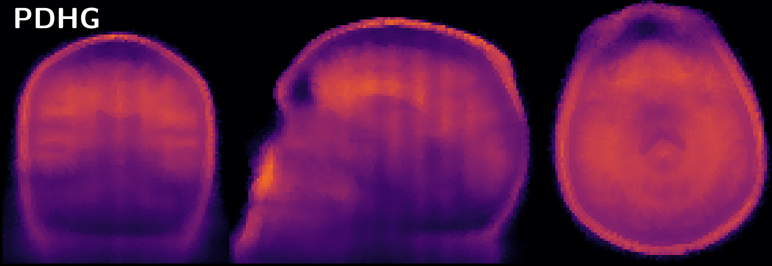


SPDHG (252 subsets, precond, balanced)

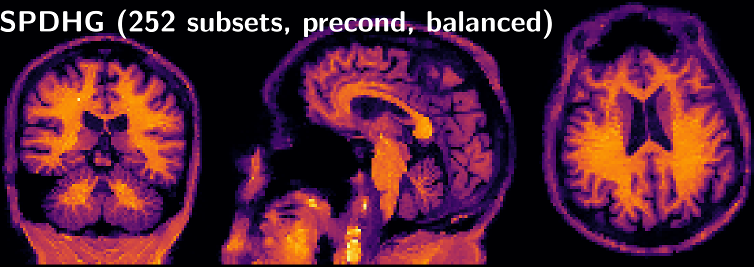


Florbetapir, 10 epochs

PDHG

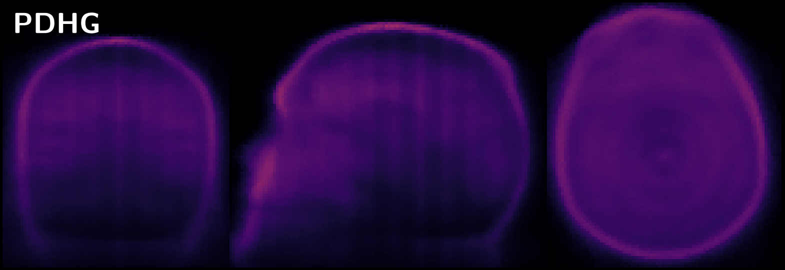


SPDHG (252 subsets, precond, balanced)

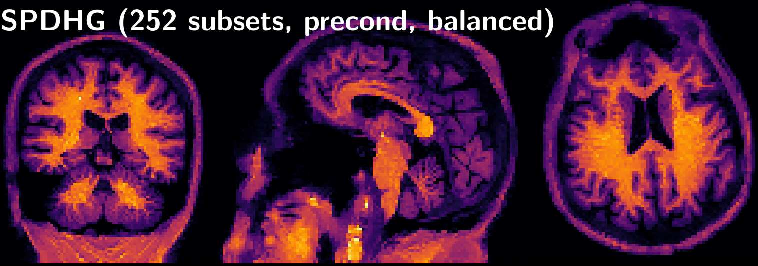


Florbetapir, 5 epochs

PDHG

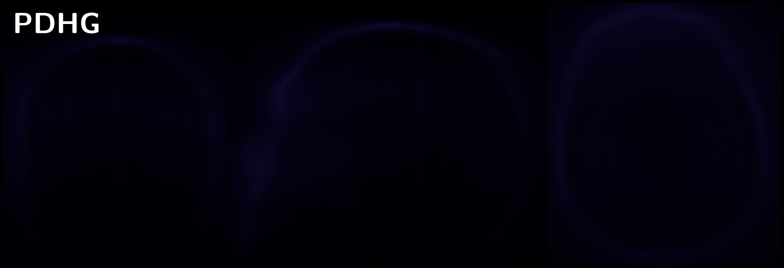


SPDHG (252 subsets, precond, balanced)

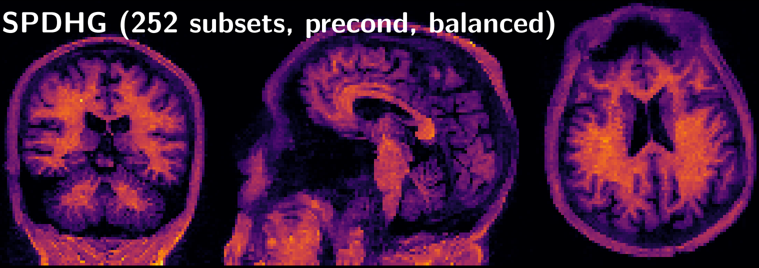


Florbetapir, 1 epoch

PDHG



SPDHG (252 subsets, precond, balanced)



Conclusions and Outlook

Summary:

- **Randomized** optimization which exploits "separable structure"
- More subsets, balanced sampling and preconditioning **all speed up**
- **only 5-20 epochs** needed for advanced models on clinical data

Future work:

- **evaluation** in concrete situations (with Addenbrookes' Cambridge)
- **sampling**: 1) optimal, 2) adaptive