Randomized Image Reconstruction

Matthias J. Ehrhardt

Department of Mathematical Sciences, University of Bath, UK

12 September, 2022

Main Aim and Outline

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

- proper, convex and lower semi-continuous
- non-smooth
- n is large and/or $\mathbf{B}_i x$ expensive

Outline:

1) **Why?** Inverse Problems and Optimization
2) **How?** Randomized Algorithm for Convex Optimization
3) Applications: PET, motion corrected CT, parallel MRI

Inverse Problems and Optimization

A way to solve inverse problems

Variational regularization

Approximate a solution u^* of $\mathbf{A}u = v$ via

$$u_\lambda = \arg\min_u \left\{ D(\mathbf{A}u, v) + \lambda R(u) \right\}$$

- **data fit** D: quantify fit of prediction $\mathbf{A}u$ to data v. Usually a "divergence", i.e. $D(x, y) \geq 0$ and $D(x, y) = 0$ iff $x = y$

$$D(x, y) = \|x - y\|_2^2, \|x - y\|_1, \int x - y + y \log(y/x), \ldots$$

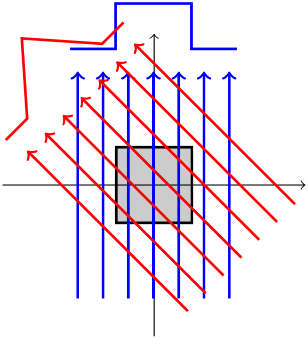
- **regularizer** R: penalize unwanted features, ensures stability

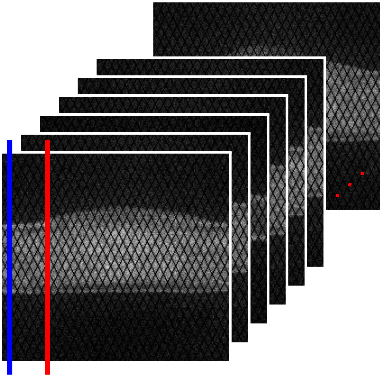
$$R(x) = \|x\|_2^2, \|x\|_1, \mathrm{TV}(x) = \|\nabla x\|_1, \mathrm{TGV}, \ldots$$

PET Modelling

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

- data $b_i \in \mathbb{N}$
- forward model $a_i^T u$
- background $r_i > 0$ (scatter, randoms)
- amount of data: 2D $N = 86k$, 3D $N = 355M$





PET Reconstruction[1]

$$u_\lambda \in \arg\min_u \left\{ \sum_{i=1}^N \mathrm{KL}(a_i^T u + r_i) + \lambda \mathcal{R}(u) + \imath_+(u) \right\}$$

- 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[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" $\mathbb{S}_1, \dots, \mathbb{S}_m$

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

- 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

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

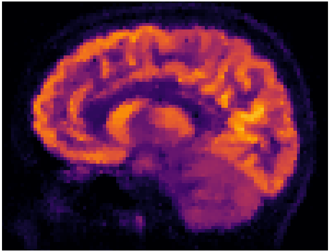
[1] Brune '10, Brune et al. '10, Setzer et al. '10, Müller et al. '11, Anthoine et al. '12, Knoll et al. '16, Ehrhardt et al. '16, Hohage and Werner '16, Schramm et al. '17, Rasch et al. '17, Ehrhardt et al. '17, Mehranian et al. '17 and many, many more

PET Reconstruction with TV

Total variation (TV)

Rudin, Osher, Fatemi '92

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

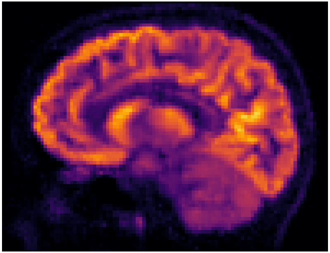


$$\min_u \left\{ \sum_{j=1}^{m} \mathcal{D}_j(\mathbf{A}_j u) + \lambda \|\nabla u\|_1 + \imath_+(u) \right\}$$

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

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

PET Reconstruction with TGV



Total generalized variation (TGV)

Bredies, Kunisch, Pock '10

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

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

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

$n = m + 2$

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

Motion corrected CT reconstruction

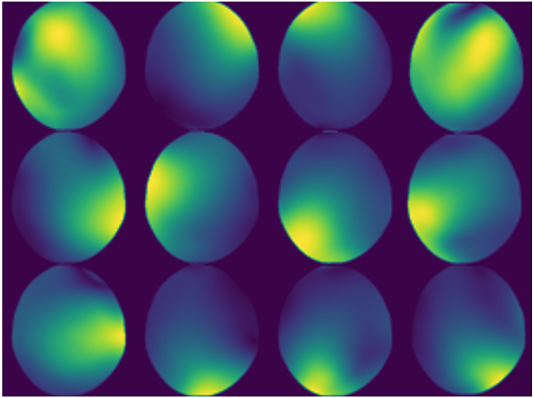
$$\min_u \sum_{i=1}^{n} \|A M_i u - b_i\|^2 + R(u)$$

- Here $n = 10$ motion gates
- No motion correction: $M_i = I$

e.g. Delplancke, Thielemans, Ehrhardt '21

Parallel MRI

$$\min_u \sum_{i=1}^{n} \|SF C_i u - b_i\|^2 + R(u)$$



Pruessmann et al. '99

Observations

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

- **Proper:** Extended valued $f : X \mapsto \mathbb{R} \cup \{\infty\}$ and $f \not\equiv \infty$
- **Convex:** e.g. C convex $\Rightarrow \imath_C$ convex
- **Lower semi-continuous (lsc):** $x_k \to x$ then

$$f(x) \leq \liminf_{k\to\infty} f(x_k)$$

  - continuous $\Rightarrow$ lsc
  - C closed $\Rightarrow \imath_C$ lsc
- $f(z) = \sum_i f_i(z_i)$ is "**separable**". Not separable in x.

Observations

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

- **Proper:** Extended valued $f : X \mapsto \mathbb{R} \cup \{\infty\}$ and $f \not\equiv \infty$
- **Convex:** e.g. C convex $\Rightarrow \imath_C$ convex
- **Lower semi-continuous (lsc):** $x_k \to x$ then

$$f(x) \le \liminf_{k\to\infty} f(x_k)$$

  - continuous $\Rightarrow$ lsc
  - C closed $\Rightarrow \imath_C$ lsc
- $f(z) = \sum_i f_i(z_i)$ is "**separable**". Not separable in x.

Problem 1: The functions f_i, g are non-smooth

Problem 2: n is large and/or $\mathbf{B}_i x$ expensive

Optimization

Proximal operator: properties and examples

Definition: The **proximal operator** of f is defined as

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

Many rules: e.g.

Proposition: Let f be separable, i.e. $f(x) = \sum_i f_i(x_i)$. Then

$$[\operatorname{prox}_f(x)]_i = \operatorname{prox}_{f_i}(x_i)\,.$$

Examples:

- $f(x) = \frac{1}{2}\|x\|_2^2$: $\operatorname{prox}_f(x) = \frac{1}{2}x$
- $f(x) = \|x\|_1$:

$$[\operatorname{prox}_f(x)]_i = \begin{cases} x_i - 1 & \text{if } x_i > 1 \\ 0 & |x_i| \leq 1 \\ x_i + 1 & \text{if } x_i < -1 \end{cases}$$

- $f = \imath_{\geq 0}$: $[\operatorname{prox}_f(x)]_i = \max(x_i, 0)$

Proximal operator: properties and examples

Definition: The **proximal operator** of f is defined as

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

Many rules: e.g.

Proposition: Let f be separable, i.e. $f(x) = \sum_i f_i(x_i)$. Then

$$[\operatorname{prox}_f(x)]_i = \operatorname{prox}_{f_i}(x_i)\,.$$

Examples:

- $f(x) = \frac{1}{2}\|x\|_2^2$: $\quad \operatorname{prox}_f(x) = \frac{1}{2}x$
- $f(x) = \|x\|_1$:

$$[\operatorname{prox}_f(x)]_i = \begin{cases} x_i - 1 & \text{if } x_i > 1 \\ 0 & |x_i| \leq 1 \\ x_i + 1 & \text{if } x_i < -1 \end{cases}$$

- $f = \imath_{\geq 0}$: $\quad [\operatorname{prox}_f(x)]_i = \max(x_i, 0)$

Problem: What is the proximal operator of $f(x) = \|\mathbf{C}x\|_1$?

The way out: Saddle Point Problems

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

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

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

The way out: Saddle Point Problems

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

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

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

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 Problems

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

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

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

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

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

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

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

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

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

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

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

$$(1)\ x^{k+1} = \mathrm{prox}_{\tau g}(x^k - \tau \textstyle\sum_{i=1}^n \mathbf{B}_i^* \bar{y}_i^k)$$

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

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

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

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

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

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

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

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

Select $i^{k+1} \in \{1, \ldots, n\}$ with probability (p_i).

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

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

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

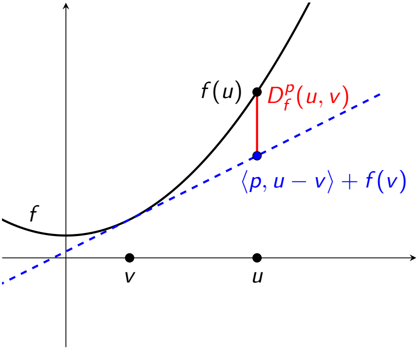
[1] Chambolle, Ehrhardt, Richtárik, Schönlieb '18

Convergence Guarantees

Bregman Distance

Definition: The **Bregman distance** of f is defined as

$$D_f^p(u,v) = f(u) - f(v) - \langle p, u - v \rangle\,, \qquad p \in \partial f(v).$$



Convergence of SPDHG

Let $\theta = 1$ and choose σ_i, τ such that $\sigma_i \tau \|\mathbf{B}_i\|^2 < p_i$.

Theorem: Chambolle, Ehrhardt, Richtárik, Schönlieb '18

Let $(x^\sharp, y^\sharp)$ be a saddle point. Then

- Almost surely: $D_g^{p^\sharp}(x^k, x^\sharp) + D_{f^*}^{q^\sharp}(y^k, y^\sharp) \to 0$
- Rate for ergodic sequence $(\hat{x}^K, \hat{y}^K) = \frac{1}{K}\sum_{k=1}^K (x^k, y^k)$

$$\mathbb{E}\left\{D_g^{p^\sharp}(\hat{x}^K x^\sharp) + D_{f^*}^{q^\sharp}(\hat{y}^K, y^\sharp)\right\} \leq \frac{C}{K}$$

Theorem: Gutiérrez, Delplancke, Ehrhardt '21, Alacaoglu, Fercoq, Cevher '22

There exists a saddle point $(x^\sharp, y^\sharp)$ such that almost surely

$$(x^k, y^k) \to (x^\sharp, y^\sharp).$$

Step-size condition of SPDHG

$$\sigma_i \tau \|\mathbf{B}_i\|^2 < p_i.$$

- Is a large-product $\sigma_i \tau$ good? Empirically yes

Step-size condition of SPDHG

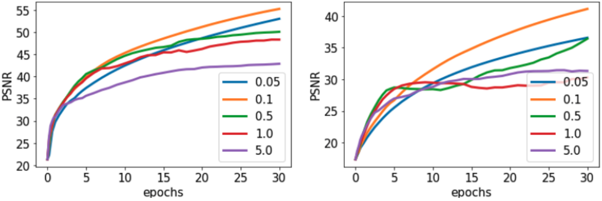
$$\sigma_i \tau \|\mathbf{B}_i\|^2 < p_i.$$

- Is a large-product $\sigma_i \tau$ good? Empirically yes
- Is upper bound tight? No, e.g. for PDHG $\sigma \tau \|\mathbf{B}\|^2 < 4/3$ is sometimes possible. Also empirically noticed for SPDHG in Schramm and Holler '22

Step-size condition of SPDHG

$$\sigma_i \tau \|\mathbf{B}_i\|^2 < p_i.$$

- Is a large-product $\sigma_i \tau$ good? Empirically yes
- Is upper bound tight? No, e.g. for PDHG $\sigma \tau \|\mathbf{B}\|^2 < 4/3$ is sometimes possible. Also empirically noticed for SPDHG in Schramm and Holler '22
- Is the ratio σ_i / τ important? Yes Delplancke et al. '20

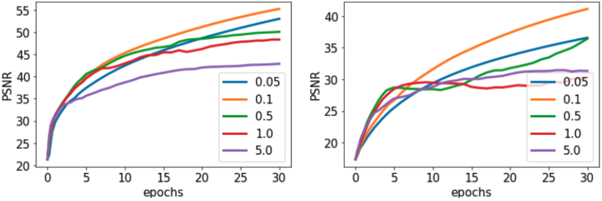


(a) synthetic data (b) real data

Step-size condition of SPDHG

$$\sigma_i \tau \|\mathbf{B}_i\|^2 < p_i.$$

- Is a large-product $\sigma_i \tau$ good? Empirically yes
- Is upper bound tight? No, e.g. for PDHG $\sigma \tau \|\mathbf{B}\|^2 < 4/3$ is sometimes possible. Also empirically noticed for SPDHG in Schramm and Holler '22
- Is the ratio σ_i / τ important? Yes Delplancke et al. '20



(a) synthetic data (b) real data

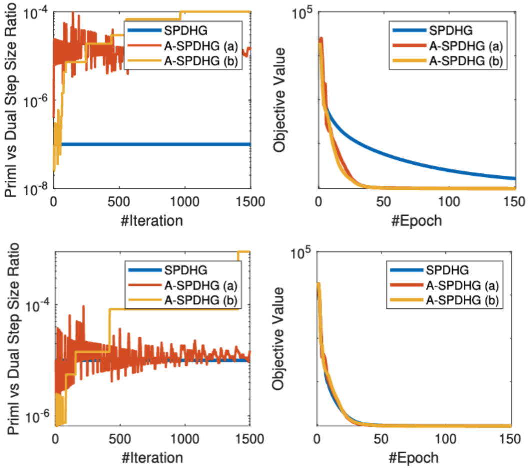
- How to choose the ratio σ_i / τ? Open question

Adaptive step-sizes

- Idea: let σ and τ vary with iterations
- PDHG: a bit of theory + emprical results Goldstein et al. '15
- SPDHG: empirical results for MPI Zdun and Brandt '21

Adaptive step-sizes

- Idea: let σ and τ vary with iterations
- PDHG: a bit of theory + emprical results Goldstein et al. '15
- SPDHG: empirical results for MPI Zdun and Brandt '21
- SPDHG: theory + numerics for CT preprint to be submitted



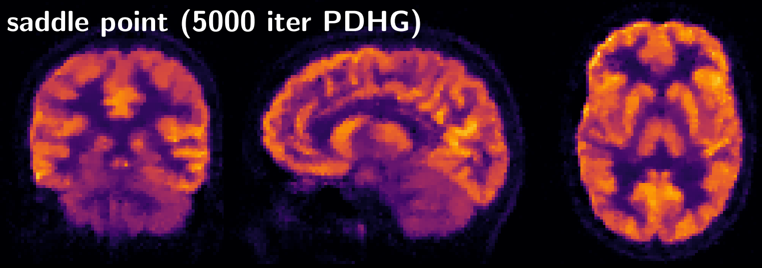
Descent or primal-dual?

- KL not “smooth”: gradient with **large Lipschitz constant** depending on background and data
- variance-reduced SGD (like SAGA or SVRG) currently **cannot do linesearch**, so stepsizes difficult to choose for PET
- both need more **memory** than e.g. gradient descent
- **ratio** of step-sizes in primal-dual algorithms difficult to choose
- see next talk!

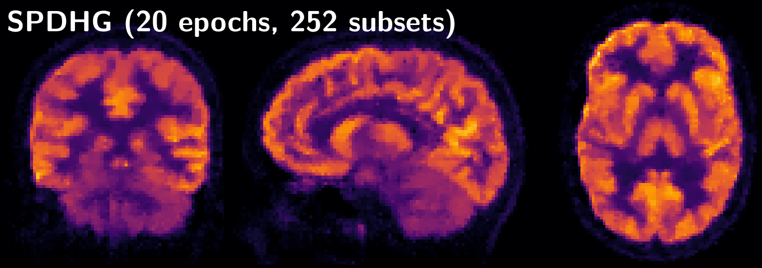
Applications

Sanity Check: Convergence to Saddle Point (TV)

saddle point (5000 iter PDHG)



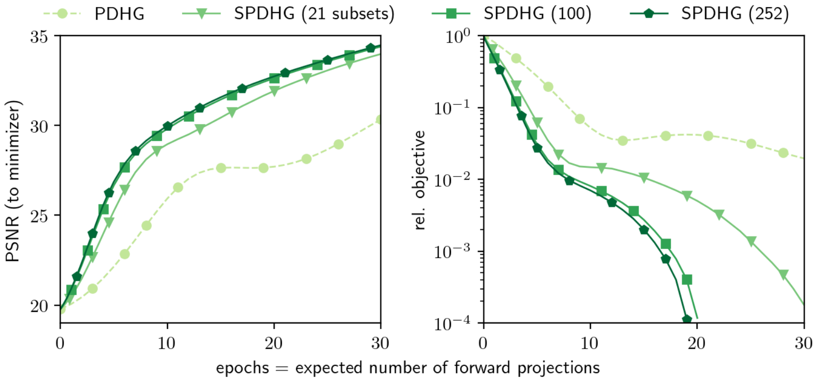
SPDHG (20 epochs, 252 subsets)



Ehrhardt, Markiewicz, Schönlieb '19

More subsets are faster

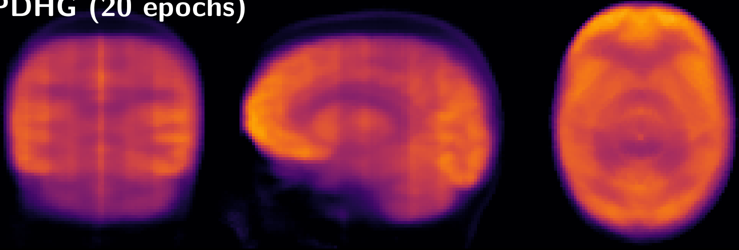
$m = 1, 21, 100, 252$



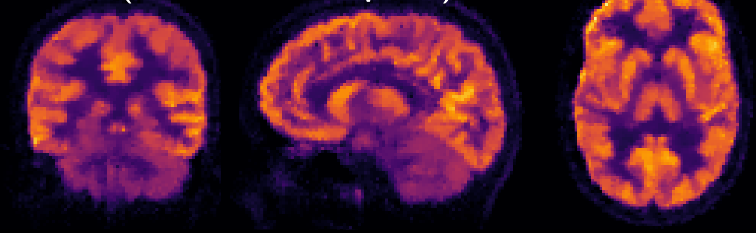
Ehrhardt, Markiewicz, Schönlieb '19

Faster than PDHG, TV

PDHG (20 epochs)



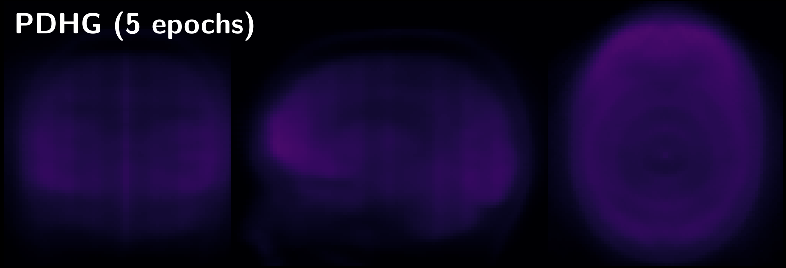
SPDHG (252 subsets, 20 epochs)



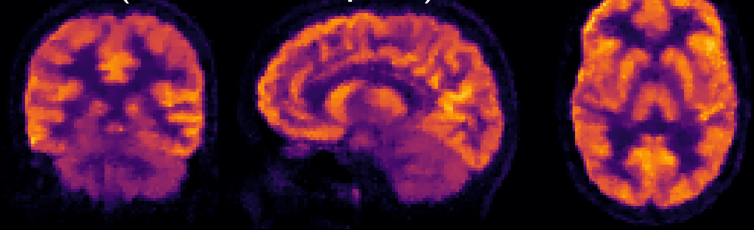
Ehrhardt, Markiewicz, Schönlieb '19

Faster than PDHG, TV

PDHG (5 epochs)



SPDHG (252 subsets, 5 epochs)

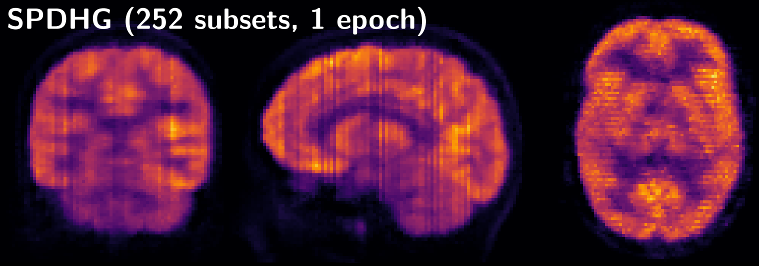


Ehrhardt, Markiewicz, Schönlieb '19

Faster than PDHG, TV

PDHG (1 epoch)

SPDHG (252 subsets, 1 epoch)

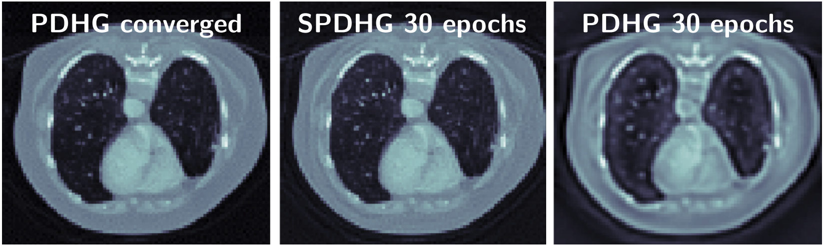


Ehrhardt, Markiewicz, Schönlieb '19

Motion corrected CT reconstruction

$$\min_u \sum_{i=1}^{n} \|A M_i u - b_i\|^2 + R(u)$$

- Here $n = 10$ motion gates

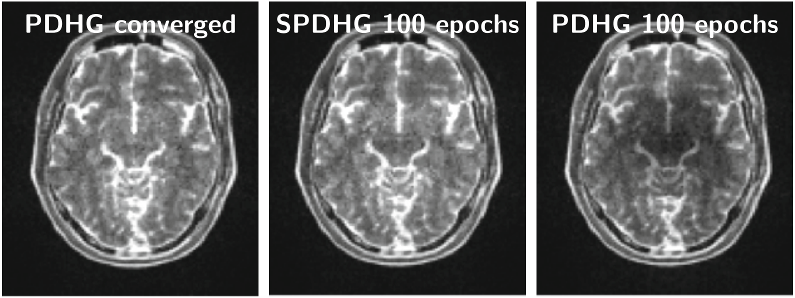


Delplancke, Thielemans, Ehrhardt '21

Parallel MRI

$$\min_u \sum_{i=1}^{n} \|SF C_i u - b_i\|^2 + R(u)$$

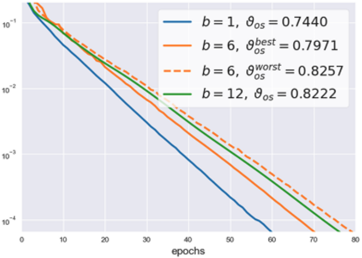
- Here $n = 8$ coils



Gutiérrez, Delplancke, Ehrhardt '21

Parallel MRI

$$\min_{x} \sum_{i=1}^{n} \|SF\,C_i x - b_i\|^2 + R(x)$$



(a) Best v worst error $\mathbf{e}_b$

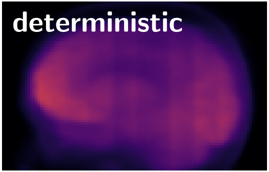
(b) Best Partition

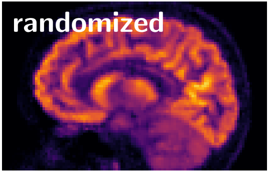
(c) Worst Partition

Gutiérrez, Delplancke, Ehrhardt ’22

Conclusions and Outlook

- **Randomized** optimisation for cost functionals with "separable structure"
- **Generalisation** of PDHG and its convergence results
- **Speeds up** PET, parallel MRI, motion-corrected CT





Current/future work:

- sampling: adaptive
- step-sizes: adaptive, tighter bound, ratio