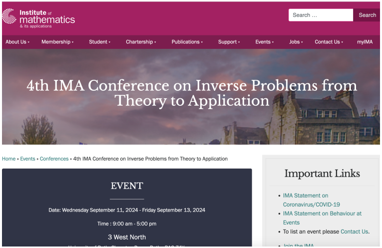
Stochastic Optimisation for Large-Scale Inverse Problems

Matthias J. Ehrhardt

Department of Mathematical Sciences, University of Bath, UK

22 May, 2024

IMA Inverse Problems: 11-13 September 2024



Invited Speakers

Coralia Cartis (Oxford)
Marcelo Pereyra (Heriot Watt)
Olga Hernandez (Eindhoven University of Technology)
Rob Scheichl (Heidelberg)

Main Aim and Outline

$$x^\sharp \in \arg\min_x \left\{ \sum_{i=1}^{\ell} f_i(A_i x) + \sum_{i=1}^{m} g_i(x) + \sum_{i=1}^{n} h_i(x) \right\}$$

- proper, convex and lower semi-continuous
- ℓ, m, n large and/or $A_i x$ expensive

Outline:

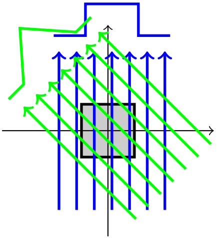
1) **Why?** Inverse Problems and Optimization
2) **How?** Randomized Algorithms for Convex Optimization
3) **So what?** Applications: PET, CT, ...

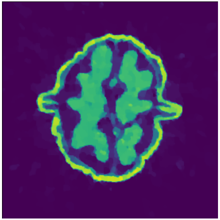
CT Reconstruction with TV

Total variation (TV)

Rudin, Osher, Fatemi '92

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





$$\min_u \left\{ \sum_{j=1}^{s} \|K_j u - b_j\|^2 + \lambda \|Du\|_1 + \imath_+(u) \right\}$$

$$\min_x \left\{ g(x) + \sum_{i=1}^{n} h_i(x) \right\}$$

$x = u,\ \ell = 0,\ m = 1,\ n = s$

$f = 0$

$g(x) = \lambda \|Dx\|_1 + \imath_+(x)$

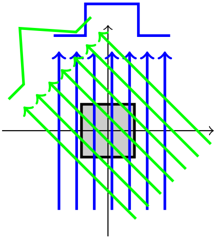
$h_i = \|K_i \cdot - b_i\|^2 \quad i \in [n]$

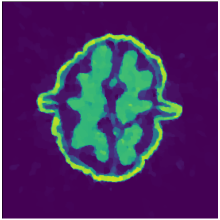
CT Reconstruction with TV: alternative

Total variation (TV)

Rudin, Osher, Fatemi '92

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





$$\min_u \left\{ \sum_{j=1}^{s} \|K_j u - b_j\|^2 + \lambda \|Du\|_1 + \imath_+(u) \right\}$$

$$\min_x \left\{ f(Ax) + g(x) + \sum_{i=1}^{n} h_i(x) \right\}$$

$x = u,\ \ell = 1,\ m = 1,\ n = s$

$f(y) = \lambda \|y\|_1,\ A = D$

$g(x) = \imath_+(x)$

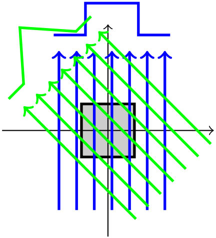
$h_i(x) = \|K_i x - b_i\|^2 \quad i \in [n]$

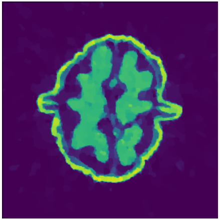
CT Reconstruction with TV: subsets

Total variation (TV)

Rudin, Osher, Fatemi '92

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





$$\min_u \left\{ \sum_{j=1}^{s} \|K_j u - b_j\|^2 + \lambda \|Du\|_1 + \imath_+(u) \right\}$$

$$\min_x \left\{ f(Ax) + g(x) + \sum_{i=1}^{n} h_i(x) \right\}$$

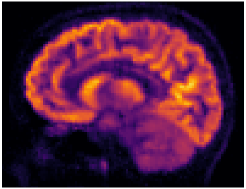
$x = u,\ \ell = 1,\ m = 1,\ n = ?$
$f(y) = \lambda\|y\|_1,\ A = D$
$g(x) = \imath_+(x)$
$h_i(x) = \sum_{j \in S_i} \|K_j x - b_j\|^2$

PET Reconstruction with TGV

Total generalized variation (TGV)

Bredies, Kunisch, Pock '10

$$\mathcal{R}(u) = \min_v \{\|Du - v\|_1 + \beta\|Dv\|_1\}$$



$$\min_{u,v} \left\{ \sum_{j=1}^{s} \mathcal{D}_j(K_j u) + \lambda\|Du - v\|_1 + \lambda\beta\|Dv\|_1 + \imath_+(u) \right\}$$

$$\min_x \left\{ \sum_{i=1}^{\ell} f_i(A_i x) + g(x) \right\}$$

$x = (u; v)$, $\ell = s+2$, $m = 1$, $n = 0$
$f_i = \mathcal{D}_i$, $A_i = (K_i, 0)$, $i \in [s]$
$f_{\ell-1} = \lambda\|\cdot\|_1$, $A_{n-1} = (D, -I)$
$f_\ell = \lambda\beta\|\cdot\|_1$, $A_n = (0, D)$
$g(x) = \imath_+(u)$

Motion corrected CT reconstruction

$$\min_u \left\{ \sum_{i=1}^{s} \|K M_i u - b_i\|^2 + \mathcal{R}(u) \right\}$$

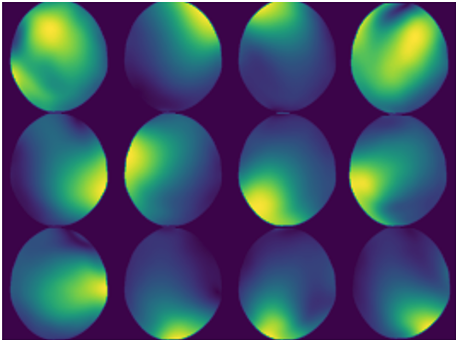
- M_i motion transformation
- here $s = 10$ motion gates; computations are a bottleneck
- No motion correction: $M_i = I$

e.g. Delplancke, Thielemans, Ehrhardt '21

Parallel MRI

$$\min_u \left\{ \sum_{i=1}^{s} \|SFC_i u - b_i\|^2 + \mathcal{R}(u) \right\}$$

- C_i sensitivity map for ith MR coil, $s = 12$



Pruessmann et al. '99

Designing Optimisation Algorithms

Building blocks for Convex Optimisation

Template:

$$\min_x \{f(Ax) + g(x) + h(x)\}$$

- h: convex and smooth: gradient descent

$$x^+ = x - \tau \nabla h(x)$$

Building blocks for Convex Optimisation

Template:

$$\min_x \{f(Ax) + g(x) + h(x)\}$$

- h: convex and smooth: gradient descent

$$x^+ = x - \tau \nabla h(x)$$

- g: convex and prox-friendly: proximal point algorithm

$$x^+ = \mathrm{prox}_{\tau g}(x) = \arg\min_z \left\{ \frac{1}{2}\|z - x\|^2 + \tau g(z) \right\}$$

Building blocks for Convex Optimisation

Template:

$$\min_x \{f(Ax) + g(x) + h(x)\}$$

- h: convex and smooth: gradient descent

$$x^+ = x - \tau \nabla h(x)$$

- g: convex and prox-friendly: proximal point algorithm

$$x^+ = \operatorname{prox}_{\tau g}(x) = \arg\min_z \left\{ \frac{1}{2}\|z - x\|^2 + \tau g(z) \right\}$$

- f: convex, prox-friendly, but $f \circ A$ is not: split f and A
  $f(Ax) = f^{**}(Ax) = \sup_y \langle Ax, y \rangle - f^*(x)$

Building blocks for Convex Optimisation

Template:

$$\min_x \{f(Ax) + g(x) + h(x)\}$$

- h: convex and smooth: gradient descent

$$x^+ = x - \tau \nabla h(x)$$

- g: convex and prox-friendly: proximal point algorithm

$$x^+ = \text{prox}_{\tau g}(x) = \arg\min_z \left\{ \frac{1}{2}\|z - x\|^2 + \tau g(z) \right\}$$

- f: convex, prox-friendly, but $f \circ A$ is not: split f and A
  $f(Ax) = f^{**}(Ax) = \sup_y \langle Ax, y \rangle - f^*(x)$

  Dual: $\min_y \{f^*(y) + (g + h)^*(-A^* y)\}$

Building blocks for Convex Optimisation

Template:

$$\min_x \{f(Ax) + g(x) + h(x)\}$$

- h: convex and smooth: gradient descent

$$x^+ = x - \tau \nabla h(x)$$

- g: convex and prox-friendly: proximal point algorithm

$$x^+ = \text{prox}_{\tau g}(x) = \arg\min_z \left\{ \frac{1}{2}\|z - x\|^2 + \tau g(z) \right\}$$

- f: convex, prox-friendly, but $f \circ A$ is not: split f and A
  $f(Ax) = f^{**}(Ax) = \sup_y \langle Ax, y\rangle - f^*(x)$

  Dual: $\min_y \{f^*(y) + (g + h)^*(-A^* y)\}$
  Primal-Dual: $\min_x \max_y \{\langle Ax, y\rangle - f^*(y) + g(x) + h(x)\}$

Building Algorithms

Template: $\min_x \{f(Ax) + g(x) + h(x)\}$

New algorithms are designed by mix-and-match:

Proximal Gradient Descent ($f = 0$): Combettes and Wajs '05

$$x^+ = \text{prox}_{\tau g}(x - \tau \nabla h(x))$$

Building Algorithms

Template: $\min_x \{f(Ax) + g(x) + h(x)\}$

New algorithms are designed by mix-and-match:

Proximal Gradient Descent ($f = 0$): Combettes and Wajs '05

$$x^+ = \text{prox}_{\tau g}(x - \tau \nabla h(x))$$

Primal-Dual Hybrid Gradient ($h = 0$) Chambolle and Pock '11

$$x^+ = \text{prox}_{\tau g}(x - \tau A^* y)$$

$$\overline{x} = x + \theta(x^+ - x)$$

$$y^+ = \text{prox}_{\sigma f^*}(y + \sigma A\overline{x})$$

Building Algorithms

Template: $\min_x \{f(Ax) + g(x) + h(x)\}$

New algorithms are designed by mix-and-match:

Proximal Gradient Descent $(f = 0)$: Combettes and Wajs '05

$$x^+ = \text{prox}_{\tau g}(x - \tau \nabla h(x))$$

Primal-Dual Hybrid Gradient $(h = 0)$ Chambolle and Pock '11

$$x^+ = \text{prox}_{\tau g}(x - \tau A^* y)$$

$$\overline{x} = x + \theta(x^+ - x)$$

$$y^+ = \text{prox}_{\sigma f^*}(y + \sigma A\overline{x})$$

Primal-Dual Three Operator Splitting (PD3O) Yan '18

$$x^+ = \text{prox}_{\tau g}(x - \tau A^* y - \tau \nabla h(x))$$

$$\overline{x} = x + \theta(x^+ - x) + \tau(\nabla h(x^+) - \nabla h(x))$$

$$y^+ = \text{prox}_{\sigma f^*}(y + \sigma A\overline{x})$$

Revisiting Gradient Descent: SGD and its variants

GD $(f = 0, g = 0)$

$$x^+ = x - \tau \nabla h(x)$$

Revisiting Gradient Descent: SGD and its variants

GD $(f = 0, g = 0)$

$$x^+ = x - \tau \sum_{i=1}^{n} \nabla h_i(x)$$

Revisiting Gradient Descent: SGD and its variants

GD $(f = 0, g = 0)$

$$x^+ = x - \tau \sum_{i=1}^{n} \nabla h_i(x)$$

SGD and variants $(f = 0, g = 0)$

Uniformly at random select j

$$x^+ = x - \tau \tilde{\nabla}^j h(x)$$

Revisiting Gradient Descent: SGD and its variants

GD $(f = 0, g = 0)$

$x^+ = x - \tau \sum_{i=1}^{n} \nabla h_i(x)$

SGD and variants $(f = 0, g = 0)$

Uniformly at random select j

$x^+ = x - \tau \tilde{\nabla}^j h(x)$

- SGD: randomly choose j,

$$\tilde{\nabla}^j h(x) = n \nabla h_j(x)$$

nonconvergence for fixed τ, "slow" convergence for carefully decreasing τ Robbins and Monro '51

Revisiting Gradient Descent: SGD and its variants

GD $(f = 0, g = 0)$

$$x^+ = x - \tau \sum_{i=1}^n \nabla h_i(x)$$

SGD and variants $(f = 0, g = 0)$

Uniformly at random select j

$$x^+ = x - \tau \tilde{\nabla}^j h(x)$$

- SGD: randomly choose j,

$$\tilde{\nabla}^j h(x) = n \nabla h_j(x)$$

  nonconvergence for fixed τ, "slow" convergence for carefully decreasing τ Robbins and Monro '51

- SAGA/SVRG: randomly choose j,

$$\tilde{\nabla}^j h(x) = n(\nabla h_j(x) - g_j) + g$$

  g historic gradient, g_j historic stochastic gradient Defazio et al. '14, Johnsen and Zhang '13, SAGA converges for $\tau \leq 1/(3nL_{\max})$

Revisiting Gradient Descent: SGD and its variants

GD $(f = 0, g = 0)$

$$x^+ = x - \tau \sum_{i=1}^{n} \nabla h_i(x)$$

SGD and variants $(f = 0, g = 0)$

Uniformly at random select j

$$x^+ = x - \tau \tilde{\nabla}^j h(x)$$

- SGD: randomly choose j,

$$\tilde{\nabla}^j h(x) = n \nabla h_j(x)$$

  nonconvergence for fixed τ, "slow" convergence for carefully decreasing τ Robbins and Monro '51
- SAGA/SVRG: randomly choose j,

$$\tilde{\nabla}^j h(x) = n(\nabla h_j(x) - g_j) + g$$

  g historic gradient, g_j historic stochastic gradient Defazio et al. '14, Johnsen and Zhang '13, SAGA converges for $\tau \leq 1/(3nL_{\max})$
- Similar algorithms exist for $\sum_i g_i(x)$ Bianchi '16, Traore et al. '23

Revisiting PDHG

PDHG:

$$x^+ = \mathrm{prox}_{\tau g}(x - \tau A^* y)$$
$$\overline{x} = x^+ + \theta(x^+ - x)$$
$$y^+ = \mathrm{prox}_{\sigma f^*}(y + \sigma A\overline{x})$$

Revisiting PDHG

PDHG:

$$x^+ = \text{prox}_{\tau g}(x - \tau A^* y)$$
$$\overline{x} = x^+ + \theta(x^+ - x)$$
$$y^+ = \text{prox}_{\sigma f^*}(y + \sigma A\overline{x})$$

PDHG (dual extrapolation):

$$y^+ = \text{prox}_{\sigma f^*}(y + \sigma A x)$$
$$\overline{y} = y^+ + \theta(y^+ - y)$$
$$x^+ = \text{prox}_{\tau g}(x - \tau A^* \overline{y})$$

Revisiting PDHG

PDHG:

$$x^+ = \text{prox}_{\tau g}(x - \tau A^* y)$$
$$\bar{x} = x^+ + \theta(x^+ - x)$$
$$y^+ = \text{prox}_{\sigma f^*}(y + \sigma A\bar{x})$$

PDHG (dual extrapolation):

$$y^+ = \text{prox}_{\sigma f^*}(y + \sigma A x)$$
$$\bar{y} = y^+ + \theta(y^+ - y)$$
$$x^+ = \text{prox}_{\tau g}(x - \tau A^* \bar{y})$$

PDHG (dual extrapolation with $f = \sum_i f_i$):

$$y_i^+ = \text{prox}_{\sigma f_i^*}(y_i + \sigma A_i x), i = 1, \ldots, \ell$$
$$\bar{y}_i = y_i^+ + \theta(y_i^+ - y_i), i = 1, \ldots, \ell$$
$$x^+ = \text{prox}_{\tau g}(x - \tau \textstyle\sum_{i=1}^{\ell} A_i^* \bar{y}_i)$$

From PDHG to SPDHG

PDHG (dual extrapolation with $f = \sum_i f_i$):

$$y_i^+ = \text{prox}_{\sigma f_i^*}(y_i + \sigma A_i x), i = 1, \dots, \ell$$

$$\overline{y}_i = y_i^+ + \theta(y_i^+ - y_i), i = 1, \dots, \ell$$

$$x^+ = \text{prox}_{\tau g}(x - \tau \sum_{i=1}^{\ell} A_i^* \overline{y}_i)$$

From PDHG to SPDHG

PDHG (dual extrapolation with $f = \sum_i f_i$):

$$y_i^+ = \text{prox}_{\sigma f_i^*}(y_i + \sigma A_i x), i = 1, \ldots, \ell$$

$$\overline{y}_i = y_i^+ + \theta(y_i^+ - y_i), i = 1, \ldots, \ell$$

$$x^+ = \text{prox}_{\tau g}(x - \tau \textstyle\sum_{i=1}^{\ell} A_i^* \overline{y}_i)$$

Stochastic PDHG (SPDHG): Chambolle, Ehrhardt, Richtárik, Schönlieb '18

Uniform at randomly select j

$$y_i^+ = \text{prox}_{\sigma f_i^*}(y_i + \sigma A_i x), i = j$$

$$\overline{y}_i = y_i^+ + \theta\ell(y_i^+ - y_i), i = j; \overline{y}_i = y_i \text{ else}$$

$$x^+ = \text{prox}_{\tau g}(x - \tau \textstyle\sum_{i=1}^{\ell} A_i^* \overline{y}_i)$$

- convergence for $\sigma\tau < 1/(\ell \max_i \|A_i\|^2)$, $\theta = 1$
  Chambolle, Ehrhardt, Richtárik, Schönlieb '18, Gutiérrez, Delplancke, Ehrhardt '21, Alacaoglu, Fercoq, Cevher '22

SPDHG as SAGA

Stochastic PDHG (SPDHG): Chambolle, Ehrhardt, Richtárik, Schönlieb '18

Uniform at randomly select j

$y_j^+ = \text{prox}_{\sigma f_j^*}(y_j + \sigma A_j x)$

$\overline{y}_i = y_i^+ + \theta\ell(y_i^+ - y_i), i = j; \overline{y}_i = y_i$ else

$x^+ = \text{prox}_{\tau g}(x - \tau \sum_{i=1}^{\ell} A_i^* \overline{y}_i)$

SPDHG as SAGA

Stochastic PDHG (SPDHG): Chambolle, Ehrhardt, Richtárik, Schönlieb '18

Uniform at randomly select j

$$y_j^+ = \operatorname{prox}_{\sigma f_j^*}(y_j + \sigma A_j x)$$

$$\overline{y}_i = y_i^+ + \theta\ell(y_i^+ - y_i), i = j; \overline{y}_i = y_i \text{ else}$$

$$x^+ = \operatorname{prox}_{\tau g}(x - \tau \textstyle\sum_{i=1}^{\ell} A_i^* \overline{y}_i)$$

SPDHG as SAGA (new):

Uniform at randomly select j

$$y_j^+ = \operatorname{prox}_{\sigma f_j^*}(y_j + \sigma A_j x)$$

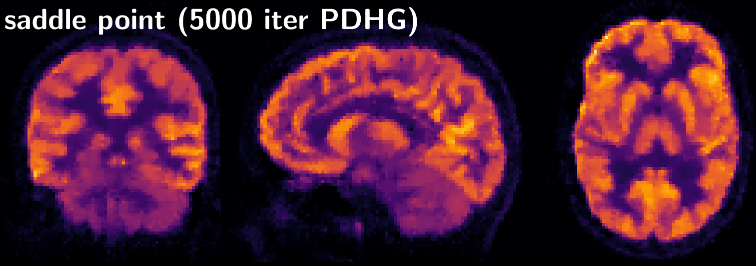
$$\tilde{\nabla}^j = (1 + \theta\ell) A_j^*(y_j^+ - y_j) + \textstyle\sum_{i=1}^{\ell} A_i^* y_i$$

$$x^+ = \operatorname{prox}_{\tau g}(x - \tau \tilde{\nabla}^j)$$

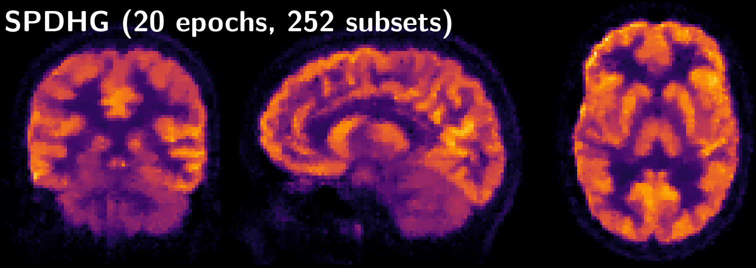
- essentially SAGA version of SPDHG
- for $\sigma = 1$, step size bound $\tau < 1/(\ell \max_i \|A_i\|^2)$ 3× larger

PET: Sanity Check, Convergence to Saddle Point (TV)

saddle point (5000 iter PDHG)



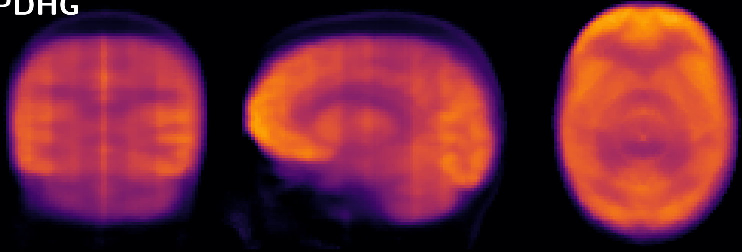
SPDHG (20 epochs, 252 subsets)



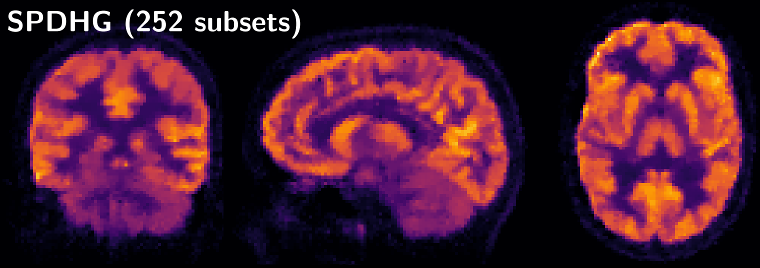
Ehrhardt, Markiewicz, Schönlieb '19

PET: Faster than PDHG, TV, 20 epochs

PDHG



SPDHG (252 subsets)

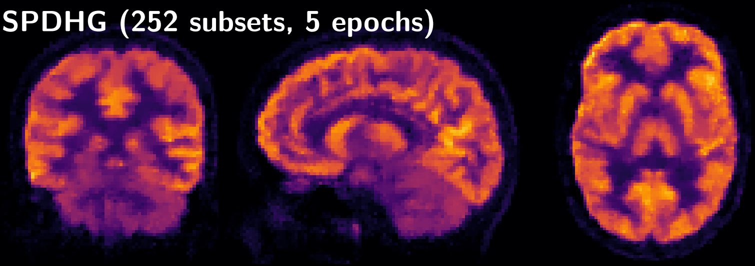


Ehrhardt, Markiewicz, Schönlieb '19

PET:Faster than PDHG, TV, 5 epochs

PDHG

SPDHG (252 subsets, 5 epochs)

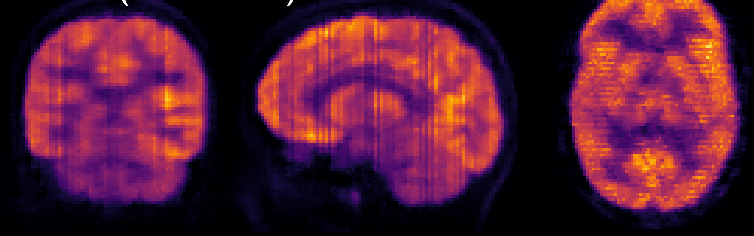


Ehrhardt, Markiewicz, Schönlieb '19

PET:Faster than PDHG, TV, 1 epochs

PDHG

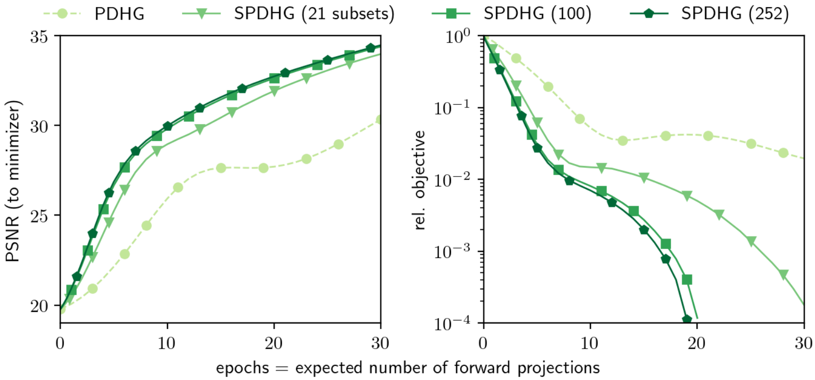
SPDHG (252 subsets)



Ehrhardt, Markiewicz, Schönlieb '19

PET, More subsets are faster

$\ell = 1, 21, 100, 252$



Ehrhardt, Markiewicz, Schönlieb '19

Step-size condition of SPDHG

$$\sigma\tau < 1/(\ell \max_i \|A_i\|^2)$$

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

Step-size condition of SPDHG

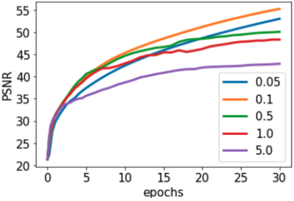
$$\sigma\tau < 1/(\ell \max_i \|A_i\|^2)$$

- Is a large-product $\sigma\tau$ good? Empirically yes
- Is upper bound tight? No, e.g. for PDHG $\sigma\tau\|A\|^2 < 4/3$ is possible Ma et al. '23 (and in fact optimal). Also empirically noticed for SPDHG, e.g. Schramm and Holler '22

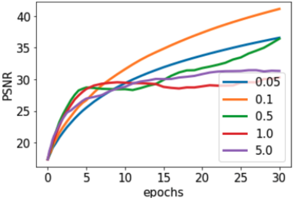
Step-size condition of SPDHG

$$\sigma\tau < 1/(\ell \max_i \|A_i\|^2)$$

- Is a large-product $\sigma\tau$ good? Empirically yes
- Is upper bound tight? No, e.g. for PDHG $\sigma\tau\|A\|^2 < 4/3$ is possible Ma et al. '23 (and in fact optimal). Also empirically noticed for SPDHG, e.g. Schramm and Holler '22
- Is the ratio σ/τ important? Yes Delplancke et al. '20



(a) synthetic data

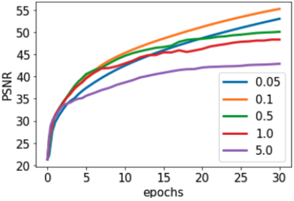


(b) real data

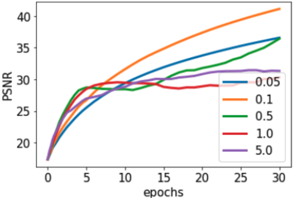
Step-size condition of SPDHG

$$\sigma\tau < 1/(\ell \max_i \|A_i\|^2)$$

- Is a large-product $\sigma\tau$ good? Empirically yes
- Is upper bound tight? No, e.g. for PDHG $\sigma\tau\|A\|^2 < 4/3$ is possible Ma et al. '23 (and in fact optimal). Also empirically noticed for SPDHG, e.g. Schramm and Holler '22
- Is the ratio σ/τ important? Yes Delplancke et al. '20



(a) synthetic data



(b) real data

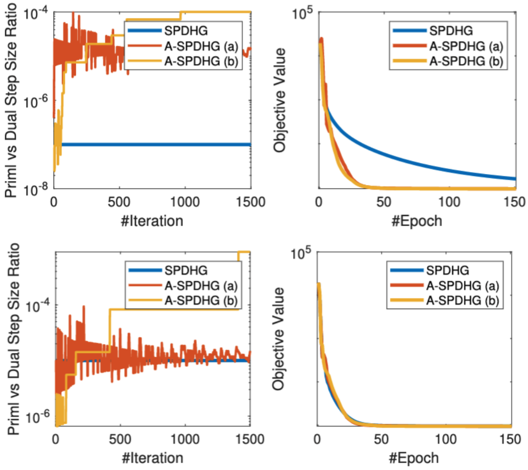
- How to choose the ratio σ/τ? Open question

Adaptive step-sizes

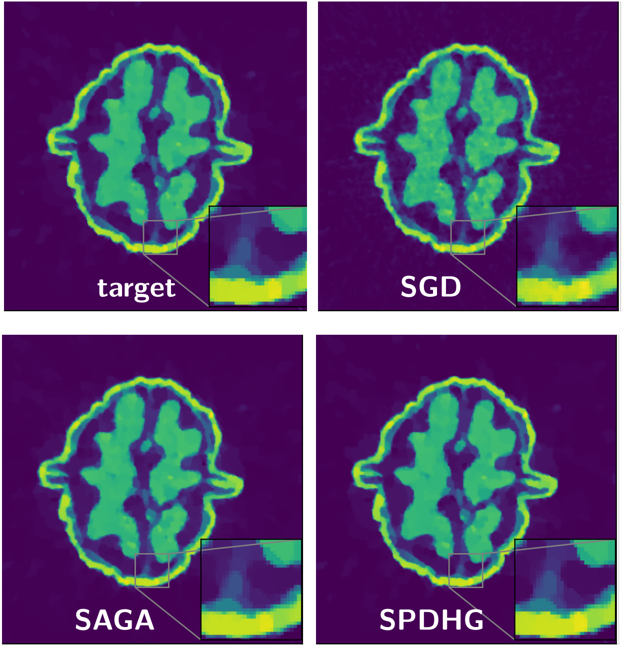
- Idea: let σ and τ vary with iterations
- PDHG: a bit of theory + emprical results Goldstein et al. '15
- SPDHG: empirical results for MPI Zdun and Brandt '21

Adaptive step-sizes

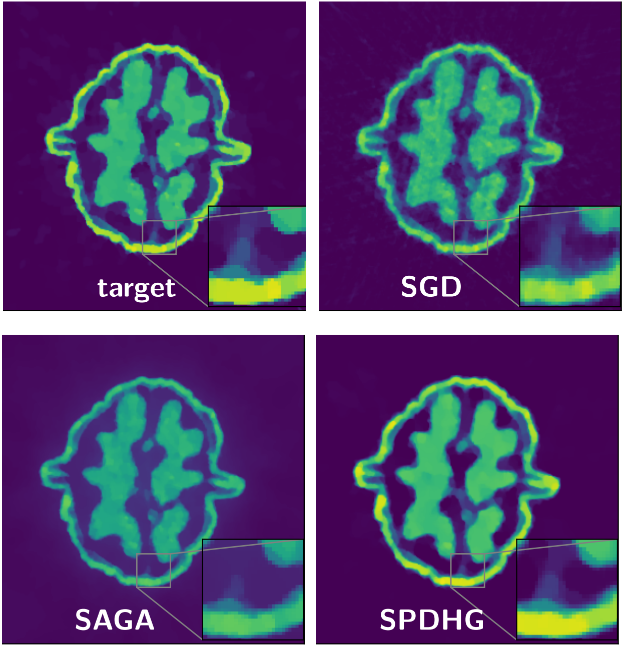
- Idea: let σ and τ vary with iterations
- PDHG: a bit of theory + emprical results Goldstein et al. '15
- SPDHG: empirical results for MPI Zdun and Brandt '21
- SPDHG: theory + numerics for CT Chambolle, Ehrhardt et al. '24



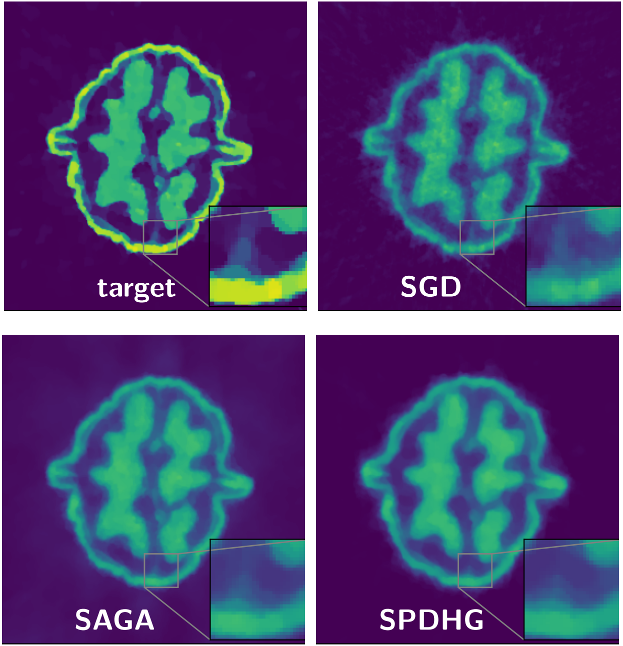
CT: 10 epochs Ehrhardt, Kereta, Liang, Tang '24 (to be submitted)



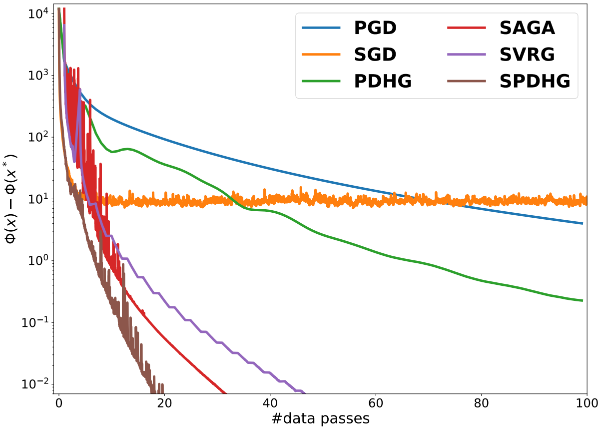
CT: 3 epochs Ehrhardt, Kereta, Liang, Tang '24 (to be submitted)



CT: 1 epoch Ehrhardt, Kereta, Liang, Tang ’24 (to be submitted)

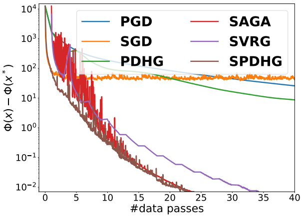


CT: Quantitative Comparison

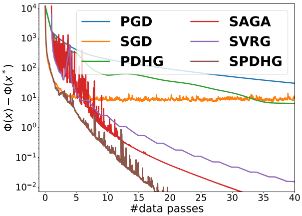


Ehrhardt, Kereta, Liang, Tang '24 (to be submitted)

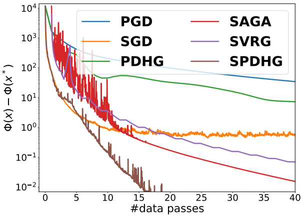
CT: Quantitative Comparison, Noise



high noise



medium noise (shown)

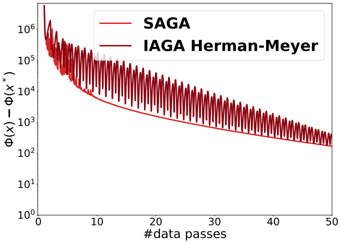


low noise

- Speed seems to depend on noise in the data
- Gradient based methods more effected

Ehrhardt, Kereta, Liang, Tang '24 (to be submitted)

CT: Random v Deterministic

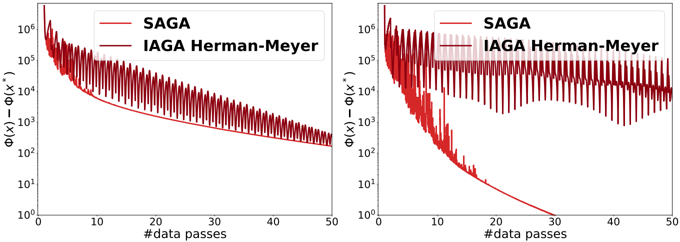


30 subsets

- similar convergence for 30 subsets (similar to literature)

Herman and Meyer ’93, Ehrhardt, Kereta, Liang, Tang ’24 (to be submitted)

CT: Random v Deterministic



30 subsets **240 subsets**

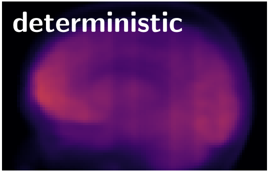
- similar convergence for 30 subsets (similar to literature)
- big difference for 240 subsets

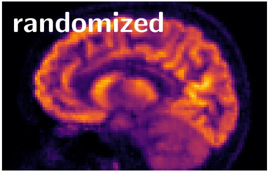
Herman and Meyer '93, Ehrhardt, Kereta, Liang, Tang '24 (to be submitted)

Conclusions and Outlook

Conclusions:

- **Zoo** of stochastic algorithms exists (gets larger and larger)
- **Randomness** seems important in general and not just mathematical convenience
- **Speeds up** reconstruction of inverse problems; e.g. PET, listmode PET (randomize over events), CT, parallel MRI, motion-corrected CT, magnetic particle imaging





Future directions:

- Tighter analysis
- Inverse problems specific analysis
- Learned algorithms