A Randomized Algorithm for Convex Optimization and Medical Imaging Applications

Matthias J. Ehrhardt

Institute for Mathematical Innovation
University of Bath, UK

March 8, 2019

Joint work with:

Mathematics:	A. Chambolle, Paris
	P. Richtárik, Edinburgh and KAUST
	C. Schönlieb, Cambridge
PET imaging:	P. Markiewicz, UCL
	J. Schott, UCL

Main Aim and Outline

Main aim:

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

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

Main Aim and Outline

Main aim:

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

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

Outline:

1) From Inverse Problems to Optimization (**Why?**)
2) Randomized Algorithm for Convex Optimization (**How?**)
3) Application: Medical Imaging (PET)

From Inverse Problems to Optimization

What is an inverse problem? Inverse to what?

Forward problem: given u, compute $\mathbf{A}u = v$. Evaluate $\mathbf{A}$

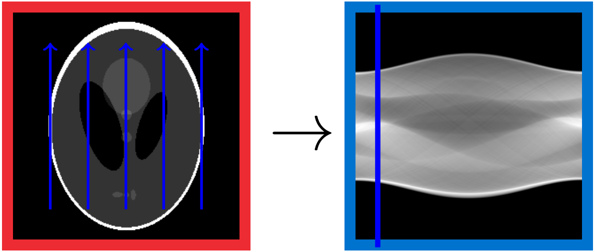
- $\mathbf{A} : U \to V$ (non-)linear operator between spaces U and V

What is an inverse problem? Inverse to what?

Forward problem: given u, compute $\mathbf{A}u = v$. Evaluate $\mathbf{A}$

- $\mathbf{A} : U \to V$ (non-)linear operator between spaces U and V
- Example: Radon / X-ray transform (used in CT, PET, ...)

$$\mathbf{A}u(L) = \int_L u(r)dr$$

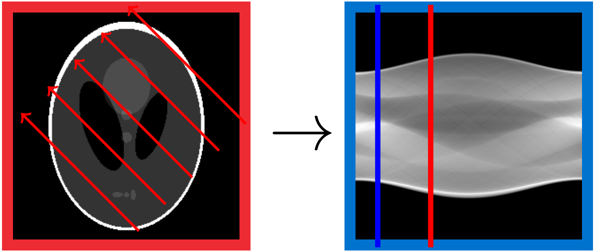


What is an inverse problem? Inverse to what?

Forward problem: given u, compute $\mathbf{A}u = v$. Evaluate $\mathbf{A}$

- $\mathbf{A} : U \to V$ (non-)linear operator between spaces U and V
- Example: Radon / X-ray transform (used in CT, PET, ...)

$$\mathbf{A}u(L) = \int_L u(r)dr$$

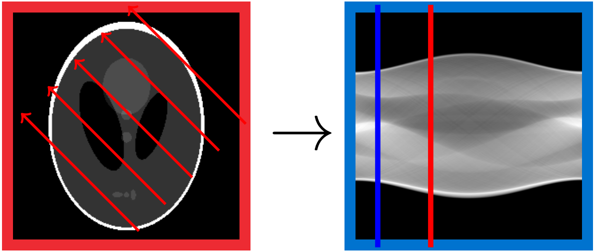


What is an inverse problem? Inverse to what?

Forward problem: given u, compute $\mathbf{A}u = v$. Evaluate $\mathbf{A}$

- $\mathbf{A} : U \to V$ (non-)linear operator between spaces U and V
- Example: Radon / X-ray transform (used in CT, PET, ...)

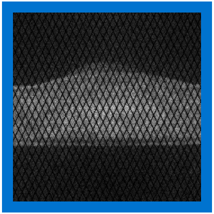
$$\mathbf{A}u(L) = \int_L u(r)dr$$



Inverse problem: given v, solve $\mathbf{A}u = v$. "Invert" $\mathbf{A}$

What is the problem with inverse problems?

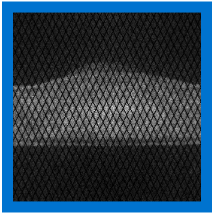
- PET example: $\mathbf{A}u(L) = \int_L u(r)dr$

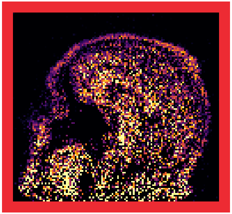


$\longrightarrow$

What is the problem with inverse problems?

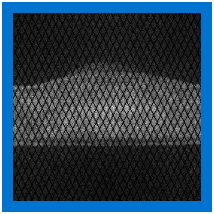
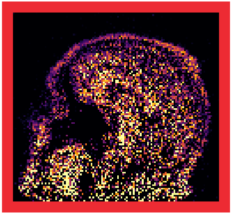
- PET example: $\mathbf{A}u(L) = \int_L u(r)dr$





What is the problem with inverse problems?

- PET example: $\mathbf{A}u(L) = \int_L u(r)dr$

 $\longrightarrow$ 

Definition (Hadamard, 1902): We call an inverse problem $\mathbf{A}u = v$ **well-posed** if

(1) a solution u^* **exists**

(2) the solution u^* is **unique**

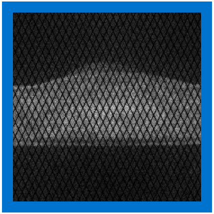
(3) u^* depends **continuously** on data v.

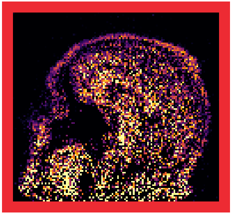
Otherwise, it is called **ill-posed**.

Jacques Hadamard

What is the problem with inverse problems?

- PET example: $\mathbf{A}u(L) = \int_L u(r)dr$





Definition (Hadamard, 1902): We call an inverse problem $\mathbf{A}u = v$ **well-posed** if

(1) a solution u^* **exists**

(2) the solution u^* is **unique**

(3) u^* depends **continuously** on data v.

Otherwise, it is called **ill-posed**.

Jacques Hadamard

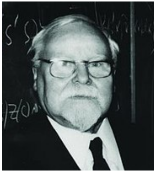
Most interesting problems are ill-posed, in particular (3) is violated.

A way to solve inverse problems

Tikhonov regularization (1943)

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

$$\begin{aligned} u_\lambda &= \arg\min_u \left\{ \|\mathbf{A}u - v\|^2 + \lambda \|u\|^2 \right\} \\ &= (\mathbf{A}^*\mathbf{A} + \lambda I)^{-1}\mathbf{A}^* v \end{aligned}$$



Andrey Tikhonov

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 Reconstruction with TV

$$u_\lambda \in \arg\min_u \left\{ \sum_{i=1}^N \mathrm{KL}(\mathbf{A}_i u + r_i; b_i) + \lambda \|\mathbf{D}u\|_1 + \imath_{\geq 0}(u) \right\}$$

- Kullback–Leibler divergence
$$\mathrm{KL}(y; b) = \begin{cases} y - b + b \log\left(\frac{b}{y}\right) & \text{if } y > 0 \\ \infty & \text{else} \end{cases}$$
- Non-smooth regularization, e.g. total variation Rudin, Osher, Fatemi 1992, Burger and Osher 2013, ... $(\mathbf{D} = \nabla)$
or directional total variation E and Betcke 2016, E et al. 2016
- Constraint
$$\imath_{\geq 0}(u) = \begin{cases} 0, & \text{if } u_i \geq 0 \text{ for all } i \\ \infty, & \text{else} \end{cases}$$

PET Reconstruction with TV

$$u_\lambda \in \arg\min_u \left\{ \sum_{i=1}^N \mathrm{KL}(\mathbf{A}_i u + r_i; b_i) + \lambda \|\mathbf{D} u\|_1 + \imath_{\geq 0}(u) \right\}$$

- Kullback–Leibler divergence
$$\mathrm{KL}(y; b) = \begin{cases} y - b + b \log\left(\frac{b}{y}\right) & \text{if } y > 0 \\ \infty & \text{else} \end{cases}$$
- Non-smooth regularization, e.g. total variation Rudin, Osher, Fatemi 1992, Burger and Osher 2013, ... $(\mathbf{D} = \nabla)$
or directional total variation E and Betcke 2016, E et al. 2016
- Constraint
$$\imath_{\geq 0}(u) = \begin{cases} 0, & \text{if } u_i \geq 0 \text{ for all } i \\ \infty, & \text{else} \end{cases}$$

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

PET Reconstruction with TGV

$$u_\lambda \in \arg\min_u \left\{ \sum_{i=1}^{N} \mathrm{KL}(\mathbf{A}_i u + r_i; b_i) + \lambda \mathrm{TGV}(u) + \imath_{\geq 0}(u) \right\}$$

- Total generalized variation Bredies, Kunisch, Pock 2010

$$\mathrm{TGV}(u) = \min_v \|\nabla u - v\|_1 + \beta \|\nabla v\|_1$$

PET Reconstruction with TGV

$$u_\lambda \in \arg\min_u \left\{ \sum_{i=1}^N \mathrm{KL}(\mathbf{A}_i u + r_i; b_i) + \lambda \mathrm{TGV}(u) + \imath_{\geq 0}(u) \right\}$$

- Total generalized variation Bredies, Kunisch, Pock 2010

$$\mathrm{TGV}(u) = \min_v \|\nabla u - v\|_1 + \beta \|\nabla v\|_1$$

$$u_\lambda, v_\lambda \in \arg\min_{u,v} \left\{ D(\mathbf{A}u, b) + \lambda \|\nabla u - v\|_1 + \lambda\beta \|\nabla v\|_1 + \imath_{\geq 0}(u) \right\}$$

PET Reconstruction with TGV

$$u_\lambda \in \arg\min_u \left\{ \sum_{i=1}^{N} \mathrm{KL}(\mathbf{A}_i u + r_i; b_i) + \lambda \mathrm{TGV}(u) + \imath_{\geq 0}(u) \right\}$$

- Total generalized variation Bredies, Kunisch, Pock 2010

$$\mathrm{TGV}(u) = \min_v \|\nabla u - v\|_1 + \beta \|\nabla v\|_1$$

$$u_\lambda, v_\lambda \in \arg\min_{u,v} \left\{ D(\mathbf{A}u, b) + \lambda \|\nabla u - v\|_1 + \lambda\beta \|\nabla v\|_1 + \imath_{\geq 0}(u) \right\}$$

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

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

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

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

Problem 1: The functions f_i, g are non-smooth but "simple"
Problem 2: n is large and/or $\mathbf{B}_i x$ expensive

Optimization

Subgradient

If f is convex and smooth, then for all $x, y \in X$ we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

Subgradient

If f is convex and smooth, then for all $x, y \in X$ we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

Extend definition to non-differentiable functions:

Definition: $f : X \mapsto \mathbb{R} \cup \{\infty\}$ is **subdifferentiable** at $x \in X$ if there exists a **subgradient** $p \in X$ such that for all $y \in X$

$$f(y) \geq f(x) + \langle p, y - x \rangle$$

holds. The set of all subgradients at $x \in X$ is called the **subdifferential** and denoted by $\partial f(x)$.

Example: $f(x) = |x|$

$$\partial f(x) = \begin{cases} \{1\} & \text{if } x > 0 \\ [-1, 1] & \text{if } x = 0 \\ \{-1\} & \text{if } x < 0 \end{cases}$$

Proximal Operators: A **gradient descent** point of view

(Sub-)Gradient descent: $p \in \partial f(x) \quad (= \{\nabla f(x)\}$ if f is diff.)

$$x^+ = x - p$$

Proximal Operators: A **gradient descent** point of view

(Sub-)Gradient descent: $p \in \partial f(x) \quad (= \{\nabla f(x)\}$ if f is diff.)

$$x^+ = x - p$$

Implicit (Sub-)Gradient descent: $p^+ \in \partial f(x^+)$

$$x^+ = x - p^+ \in x - \partial f(x^+)$$

Proximal Operators: A **gradient descent** point of view

(Sub-)Gradient descent: $p \in \partial f(x) \quad (= \{\nabla f(x)\}$ if f is diff.)

$$x^+ = x - p$$

Implicit (Sub-)Gradient descent: $p^+ \in \partial f(x^+)$

$$\begin{aligned} & x^+ = x - p^+ \in x - \partial f(x^+) \\ \Leftrightarrow \quad & x \in (I + \partial f)x^+ \end{aligned}$$

Proximal Operators: A **gradient descent** point of view

(Sub-)Gradient descent: $p \in \partial f(x) \quad (= \{\nabla f(x)\}$ if f is diff.)

$$x^+ = x - p$$

Implicit (Sub-)Gradient descent: $p^+ \in \partial f(x^+)$

$$\begin{aligned} & x^+ = x - p^+ \in x - \partial f(x^+) \\ \Leftrightarrow \quad & x \in (I + \partial f)x^+ \\ \Leftrightarrow \quad & x^+ = (I + \partial f)^{-1} x \qquad =: \operatorname{prox}_f(x) \end{aligned}$$

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

$$\operatorname{prox}_f(x) := (I + \partial f)^{-1}(x)\,.$$

prox_f has *many* names:
prox / proximal / proximity / resolvent operator

Proximal Operators: A **minimization** point of view

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

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

Proximal Operators: A **minimization** point of view

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

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

Proposition: $(I + \partial f)^{-1}(x) = \arg\min_z \left\{ \frac{1}{2}\|z - x\|^2 + f(z) \right\}$

Proximal Operators: A **minimization** point of view

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

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

Proposition: $(I + \partial f)^{-1}(x) = \arg\min_z \left\{ \frac{1}{2}\|z - x\|^2 + f(z) \right\}$

"Proof":

$$\begin{aligned}
& x^+ = \arg\min_z \left\{ \frac{1}{2}\|z - x\|^2 + f(z) \right\} \\
\Leftrightarrow \quad & 0 \in \partial \left\{ \frac{1}{2}\|x^+ - x\|^2 + f(x^+) \right\} \\
\Leftrightarrow \quad & 0 \in x^+ - x + \partial f(x^+) \\
\Leftrightarrow \quad & x \in (I + \partial f)x^+ \\
\Leftrightarrow \quad & x^+ = (I + \partial f)^{-1} x
\end{aligned}$$

Proximal operator: properties and examples

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

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:

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

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

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

Proximal operator: properties and examples

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

Many rules: e.g.

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

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

Examples:

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

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

- $f = \imath_C$: $\mathrm{prox}_f(x) = \mathrm{proj}_C(x)$
- $f = \imath_{\geq 0}$: $\mathrm{prox}_f(x)_i = \max(x_i, 0)$

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

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

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

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

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

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

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

Select $\mathbb{S}^{k+1} \subset \{1, \dots, n\}$ randomly.

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

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

- 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, E, Richtárik, Schönlieb '18

Convergence Guarantees

Step Size Condition with ESO[1]

Tall matrix $\mathbf{C} = [\mathbf{C}_1; \dots; \mathbf{C}_n]$, $\mathbf{C}^* h = \sum_{i=1}^n \mathbf{C}_i^* h_i$

Definition (Expected Separable Overapproximation, ESO):
Random subset $\mathbb{S} \subset \{1, \dots, n\}$. The ESO parameters v_i fulfill the ESO inequality if for all h

$$\mathbb{E}_{\mathbb{S}} \left\| \sum_{i \in \mathbb{S}} \mathbf{C}_i^* h_i \right\|^2 \leq \sum_{i=1}^n p_i v_i \|h_i\|^2 \, .$$

[1] Qu, Richtárik, Zhang '14

Step Size Condition with ESO[1]

Tall matrix $\mathbf{C} = [\mathbf{C}_1; \dots; \mathbf{C}_n]$, $\mathbf{C}^* h = \sum_{i=1}^n \mathbf{C}_i^* h_i$

Definition (Expected Separable Overapproximation, ESO): Random subset $\mathbb{S} \subset \{1, \dots, n\}$. The ESO parameters v_i fulfill the ESO inequality if for all h

$$\mathbb{E}_{\mathbb{S}} \left\| \sum_{i \in \mathbb{S}} \mathbf{C}_i^* h_i \right\|^2 \leq \sum_{i=1}^n p_i v_i \|h_i\|^2 .$$

Example (Full Sampling): $\mathbb{S} = \{1, \dots, n\}, p_i = 1,\ v_i = \|\mathbf{C}\|^2$

$$LHS = \|\mathbf{C}^* h\|^2$$

[1] Qu, Richtárik, Zhang '14

Step Size Condition with ESO[1]

Tall matrix $\mathbf{C} = [\mathbf{C}_1; \dots; \mathbf{C}_n]$, $\mathbf{C}^* h = \sum_{i=1}^n \mathbf{C}_i^* h_i$

Definition (Expected Separable Overapproximation, ESO): Random subset $\mathbb{S} \subset \{1, \dots, n\}$. The ESO parameters v_i fulfill the ESO inequality if for all h

$$\mathbb{E}_{\mathbb{S}} \left\| \sum_{i \in \mathbb{S}} \mathbf{C}_i^* h_i \right\|^2 \le \sum_{i=1}^n p_i v_i \|h_i\|^2 .$$

Example (Full Sampling): $\mathbb{S} = \{1, \dots, n\}, p_i = 1,\ v_i = \|\mathbf{C}\|^2$

$$LHS = \|\mathbf{C}^* h\|^2 \le \|\mathbf{C}^*\|^2 \|h\|^2$$

[1] Qu, Richtárik, Zhang '14

Step Size Condition with ESO[1]

Tall matrix $\mathbf{C} = [\mathbf{C}_1; \ldots; \mathbf{C}_n]$, $\mathbf{C}^* h = \sum_{i=1}^n \mathbf{C}_i^* h_i$

Definition (Expected Separable Overapproximation, ESO): Random subset $\mathbb{S} \subset \{1, \ldots, n\}$. The ESO parameters v_i fulfill the ESO inequality if for all h

$$\mathbb{E}_{\mathbb{S}} \left\| \sum_{i \in \mathbb{S}} \mathbf{C}_i^* h_i \right\|^2 \leq \sum_{i=1}^n p_i v_i \|h_i\|^2 .$$

Example (Full Sampling): $\mathbb{S} = \{1, \ldots, n\}, p_i = 1,\ v_i = \|\mathbf{C}\|^2$

$$LHS = \|\mathbf{C}^* h\|^2 \leq \|\mathbf{C}^*\|^2 \|h\|^2 = \sum_{i=1}^n \|\mathbf{C}^*\|^2 \|h_i\|^2$$

[1] Qu, Richtárik, Zhang '14

Step Size Condition with ESO[1]

Tall matrix $\mathbf{C} = [\mathbf{C}_1; \dots; \mathbf{C}_n]$, $\mathbf{C}^* h = \sum_{i=1}^n \mathbf{C}_i^* h_i$

Definition (Expected Separable Overapproximation, ESO): Random subset $\mathbb{S} \subset \{1, \dots, n\}$. The ESO parameters v_i fulfill the ESO inequality if for all h

$$\mathbb{E}_{\mathbb{S}} \left\| \sum_{i \in \mathbb{S}} \mathbf{C}_i^* h_i \right\|^2 \le \sum_{i=1}^n p_i v_i \|h_i\|^2 .$$

Example (Full Sampling): $\mathbb{S} = \{1, \dots, n\}, p_i = 1,\ v_i = \|\mathbf{C}\|^2$

$$LHS = \|\mathbf{C}^* h\|^2 \le \|\mathbf{C}^*\|^2 \|h\|^2 = \sum_{i=1}^n \|\mathbf{C}^*\|^2 \|h_i\|^2$$

Example (Serial Sampling): $\mathbb{S} = \{i\},\ v_i = \|\mathbf{C}_i\|^2$

$$LHS = \sum_{i=1}^n p_i \|\mathbf{C}_i^* h_i\|^2$$

[1] Qu, Richtárik, Zhang '14

Step Size Condition with ESO[1]

Tall matrix $\mathbf{C} = [\mathbf{C}_1; \ldots; \mathbf{C}_n]$, $\mathbf{C}^* h = \sum_{i=1}^n \mathbf{C}_i^* h_i$

Definition (Expected Separable Overapproximation, ESO): Random subset $\mathbb{S} \subset \{1, \ldots, n\}$. The ESO parameters v_i fulfill the ESO inequality if for all h

$$\mathbb{E}_{\mathbb{S}} \left\| \sum_{i \in \mathbb{S}} \mathbf{C}_i^* h_i \right\|^2 \leq \sum_{i=1}^n p_i v_i \|h_i\|^2 .$$

Example (Full Sampling): $\mathbb{S} = \{1, \ldots, n\}, p_i = 1,\ v_i = \|\mathbf{C}\|^2$

$$LHS = \|\mathbf{C}^* h\|^2 \leq \|\mathbf{C}^*\|^2 \|h\|^2 = \sum_{i=1}^n \|\mathbf{C}^*\|^2 \|h_i\|^2$$

Example (Serial Sampling): $\mathbb{S} = \{i\},\ v_i = \|\mathbf{C}_i\|^2$

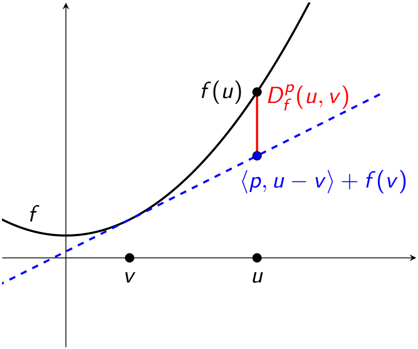
$$LHS = \sum_{i=1}^n p_i \|\mathbf{C}_i^* h_i\|^2 \leq \sum_{i=1}^n p_i \|\mathbf{C}_i^*\|^2 \|h_i\|^2$$

[1] Qu, Richtárik, Zhang '14

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

Theorem: Chambolle, E, Richtárik, Schönlieb '18
Let $(x^\sharp, y^\sharp)$ be a saddle point, $\theta = 1$ and choose σ_i, τ such that there exist ESO parameters v_i of $\mathbf{C} = [\mathbf{C}_1; \dots, \mathbf{C}_n]$ with $\mathbf{C}_i = \sqrt{\sigma_i \tau}\mathbf{B}_i$ which satisfy

$$v_i < p_i .$$

Then

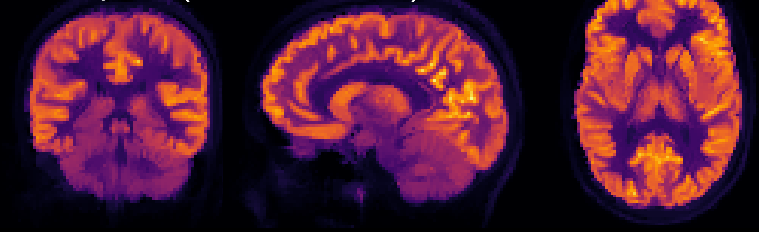
- Almost surely: $D_g^{r^\sharp}(x^k, x^\sharp) + D_{f^*}^{q^\sharp}(y^k, y^\sharp) \to 0$
- Rate for ergodic sequence $(x_K, y_K) = \frac{1}{K}\sum_{k=1}^K (x^k, y^k)$

$$\mathbb{E}\left\{D_g^{r^\sharp}(x_K, x^\sharp) + D_{f^*}^{q^\sharp}(y_K, y^\sharp)\right\} \le \frac{C}{K}$$

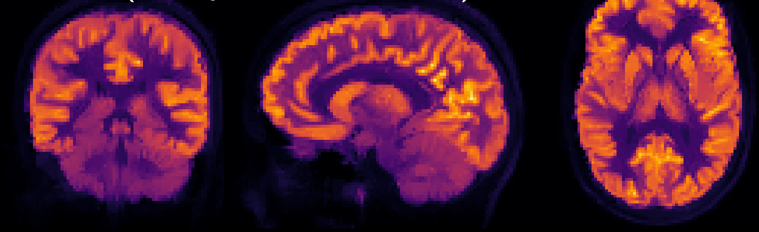
Applications

Convergence to Saddle Point (dTV): Sanity Check

saddle point (3000 iter PDHG)

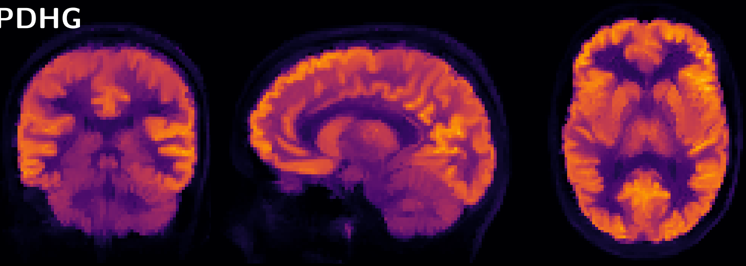


SPDHG (100 epochs, 100 subsets)

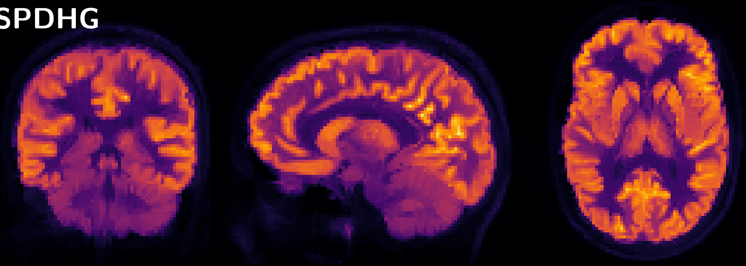


Faster than PDHG (dTV), 100 epochs

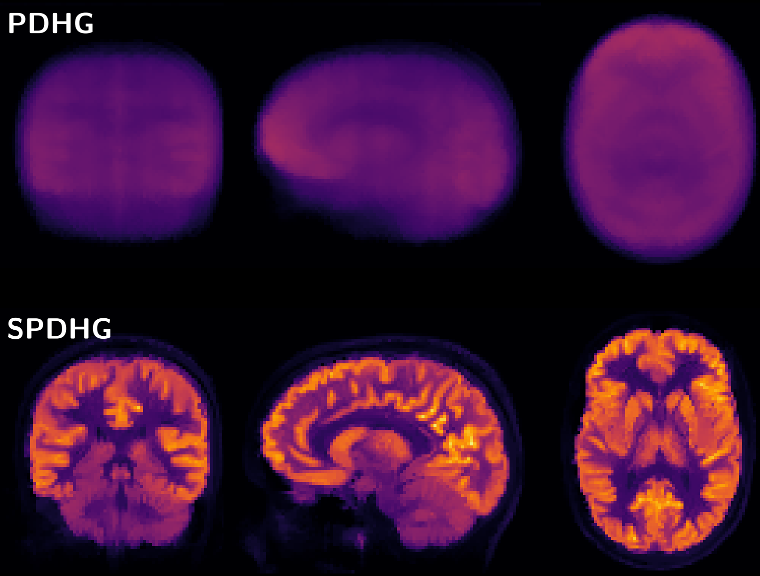
PDHG



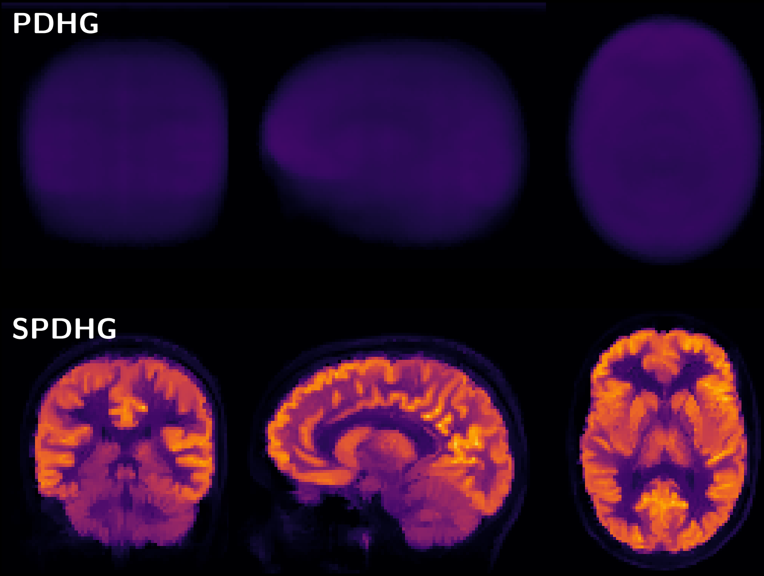
SPDHG



Faster than PDHG (dTV), 10 epochs

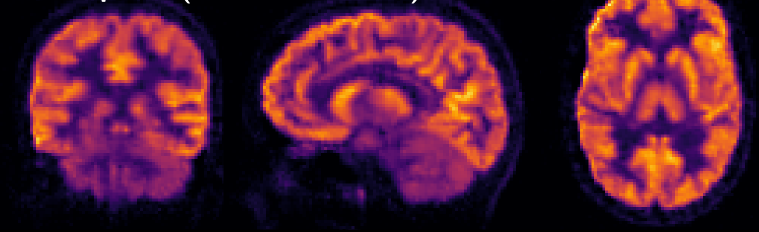


Faster than PDHG (dTV), 5 epochs

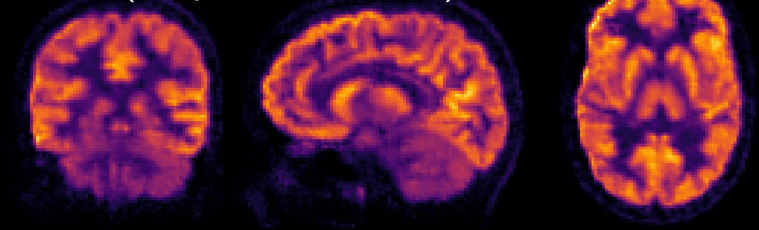


Convergence to Saddle Point (TGV): Sanity Check

saddle point (3000 iter PDHG)

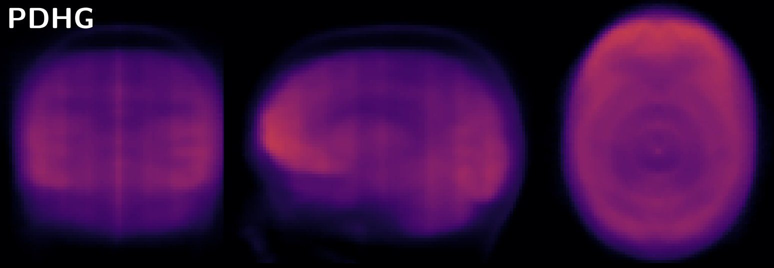


SPDHG (10 epochs, 252 subsets)

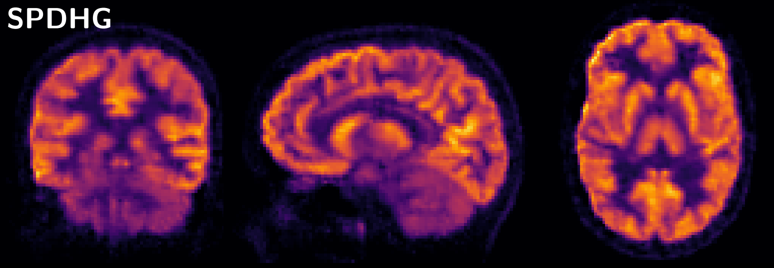


Faster than PDHG (TGV), 10 epochs

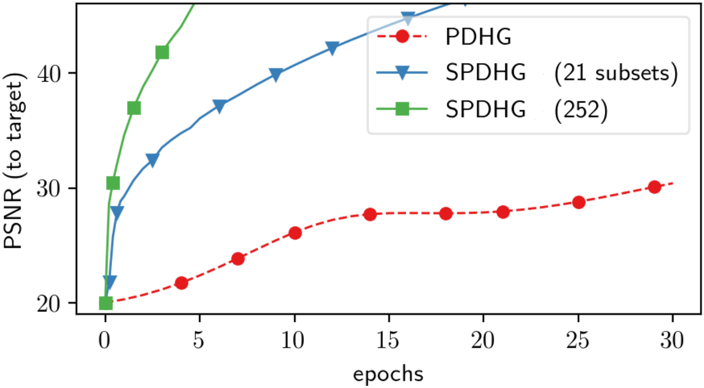
PDHG



SPDHG

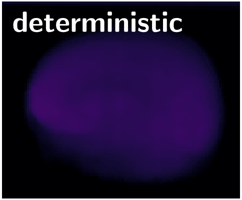


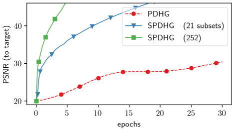
Quantitative results



Conclusions and Outlook

- **Randomized** optimisation for cost functionals with "separable structure"
- **Generalisation** of PDHG
- Convergence for **arbitrary sampling**
- **Much faster** PET reconstruction: making advanced models feasible for clinical data





Not shown today:

- Convergence theorems: 1) $\mathcal{O}(1/k^2)$ acceleration, 2) linear convergence

Future work:

- almost sure convergence of iterates
- sampling: 1) optimal, 2) adaptive
- non-convex extension with gradients