Adaptive Primal-Dual Stochastic Algorithm for Inverse Problems

Antonin Chambolle, C. D., Matthias Ehrhardt, Carola-Bibiane Schönlieb and Junqi Tang

EDF Lab - University of Bath

SIAM Data Science 2022

Overview

Stochastic Primal-Dual Hybrid Gradient (SPDHG)

- stochastic version of the Chambolle-Pock algorithm
- very efficient on large-scale inverse problems

Problem: how to tune the free parameter in the step-sizes definition of SPDHG ?

Overview

Stochastic Primal-Dual Hybrid Gradient (SPDHG)

- stochastic version of the Chambolle-Pock algorithm
- very efficient on large-scale inverse problems

Problem: how to tune the free parameter in the step-sizes definition of SPDHG ?

Idea: online tuning.

Goals:

- provide a theoretical framework and convergence guarantees
- design effective update rules in practice

$$\min_{x \in X} F(Ax) + G(x) = \min_{x \in X} \sum_{i=1}^{n} F_i(A_i x) + G(x),$$

with

- $X,\ Y = Y_1 \times \ldots Y_n$ Hilbert spaces
- $A_i : X \to Y_i$ bounded linear operators
- $F_i : Y_i \to \bar{\mathbb{R}}$ and $G : X \to \bar{\mathbb{R}}$ convex

Saddle-point formulation:

$$\min_{x \in X} \sup_{y_1, \ldots, y_n} \sum_{i=1}^{n} \left(\langle A_i x, y_i \rangle - F_i^*(y_i) \right) + G(x)$$

SPDHG (Chambolle et al, 2018)

Input

- primal step-size τ and dual step-sizes σ_i for $1 \leq i \leq n$
- probabilities $p_i > 0$ for $1 \leq i \leq n$.

Initialize $x^0 = \bar{z}^0 = 0,\ y^0 = 0.$

Iterate

- $x^{k+1} = \mathrm{prox}_G^{\tau}(x - \bar{z}^k)$
- Pick an index i with probability p_i
- $y_i^{k+1} = \mathrm{prox}_{F_i^*}^{\sigma_i}(y_i^k + \sigma_i A_i x^k)$ and $y_j^{k+1} = y_j^k$ for $j \neq i$
- $\delta^k = A_i^*(y_i^{k+1} - y_i^k)$
- $\bar{z}^{k+1} = \bar{z}^k + (1 + p_i^{-1})\delta^k$

Convergence: when? how fast?

Convergence condition:

$$\tau\sigma_i\|A_i\|^2 < p_i, \quad 1 \leq i \leq n.$$

Convergence in the sense of Bregman distances (Chambolle et al, 2018); almost-sure convergence (Alacaoglu et al, 2020; Gutierrez et al, 2021).

Convergence: when? how fast?

Convergence condition:

$$\tau\sigma_i \|A_i\|^2 < p_i, \quad 1 \le i \le n.$$

Convergence in the sense of Bregman distances (Chambolle et al, 2018); almost-sure convergence (Alacaoglu et al, 2020; Gutierrez et al, 2021).

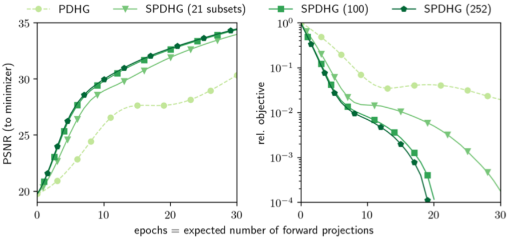


Figure: SPDHG is faster than PDHG on large-scale inverse problems (Positron Emission Tomography reconstruction - Ehrhardt et al, 2019)

How fast, again?

Convergence condition:

$$\tau \sigma_i \|A_i\|^2 < p_i, \quad 1 \leq i \leq n.$$

Admissible step-sizes: for $\gamma > 0$ and $0 < \beta < 1$,

$$\tau = \gamma * \beta \min \frac{p_i}{\|A_i\|}, \quad \sigma_i = \frac{\beta}{\gamma * \|A_i\|}.$$

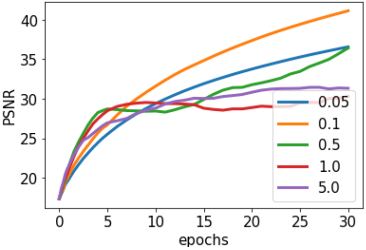


Figure: Impact of free parameter $\gamma > 0$ on convergence speed (Positron Emission Tomography reconstruction - Delplancke et al, 2020)

Adaptive step-sizes for primal-dual algorithms

For PDHG:

- primal-dual balancing (Goldstein et al, 2015, Malitsky and Pock, 2018)
- backtracking strategy (same)
- adapt to local smoothness (Vladarean et al, 2021)

For SPDHG: numerical experiments on MPI reconstruction (Zdun and Brandt, 2021), no proof of convergence.

Proposed algorithm: adaptive SPDHG

Input

- primal step-size τ^0 and dual step-sizes σ_i^0 for $1 \leq i \leq n$
- update rule
- probabilities $p_i > 0$ for $1 \leq i \leq n$.

Initialize $x^0 = \bar{z}^0 = 0$, $y^0 = 0$.

Iterate

- Determine $(\sigma_i^{k+1})_{1 \leq i \leq n}$, τ^{k+1} according to the update rule
- $x^{k+1} = \mathrm{prox}_G^{\tau^{k+1}}(x - \bar{z}^k)$
- Pick an index i with probability p_i
- $y_i^{k+1} = \mathrm{prox}_{F_i^*}^{\sigma_i^{k+1}}(y_i^k + \sigma_i A_i x^k)$ and $y_j^{k+1} = y_j^k$ for $j \neq i$
- $\delta^k = A_i^*(y_i^{k+1} - y_i^k)$
- $\bar{z}^{k+1} = \bar{z}^k + (1 + p_i^{-1})\delta^k$

Step-sizes assumptions (informal)

(i) the step-sizes at step $k+1$ depend only of the iterates up to step k,

(ii) the step-sizes product satisfy to a uniform version of SPDHG's convergence condition (upper-bound) and are not arbitrarily small (lower-bound),

(iii) the step-sizes sequences *do not vary too fast*.

Norm-inducing operators

Let H be a Hilbert space and $\mathbb{S}^+(H)$ the set of positive-definite bounded self-adjoint linear operators from H to H.

Induced norm on H: for all $M \in \mathbb{S}^+(H)$, define

$$\|u\|_M^2 = \langle Mu, u \rangle, \quad u \in H.$$

Partial order on $\mathbb{S}(H)$:

$$N \preccurlyeq M \quad \text{if} \quad \forall u \in H, \|u\|_N \leq \|u\|_M.$$

SPDHG step-sizes induce a metric on $X \times Y_i$

Define

$$M_i^k = \begin{pmatrix} \frac{1}{\tau^k}\mathrm{Id} & -\frac{1}{p_i}A_i \\ -\frac{1}{p_i}A_i^* & \frac{1}{p_i\sigma_i^k}\mathrm{Id} \end{pmatrix}, \quad N_i^k = \begin{pmatrix} \frac{1}{\tau^k}\mathrm{Id} & 0 \\ 0 & \frac{1}{p_i\sigma_i^k}\mathrm{Id} \end{pmatrix}.$$

Then, for fixed $\alpha > 0$, $0 < \beta < 1$:

- The condition $(1-\beta)N_i^k \preccurlyeq M_i^k$ for all i and k is equivalent to:

$$\tau^k \sigma_i^k \frac{\|A_i\|^2}{p_i} \leq \beta < 1, \quad \forall i,\, k$$

- The condition $\alpha \mathrm{Id} \preccurlyeq N_i^k$ for all i and k is equivalent to:

$$\tau^k \geq \alpha, \quad \sigma_i^k \geq \alpha, \quad \forall i,\, k$$

Quasi-decreasing sequence

We call a random sequence $(M^k)_{k\in\mathbb{N}}$ in $\mathbb{S}^+(H)$ uniformly almost surely quasi-decreasing if there exists a non-negative sequence $(\eta^k)_{k\in\mathbb{N}}$ such that $\sum_{k=1}^{\infty} \eta_k < \infty$ and a.s.

$$M^{k+1} \preccurlyeq (1+\eta^k) M^k, \quad k \in \mathbb{N}.$$

Step-sizes assumptions

(i) the step-size sequences at step $k+1$ are in the σ-algebra generated by the iterates up to step k,

(ii) there exists $\alpha > 0$ and $\beta \in (0,1)$ such that

$$(1-\beta)N_i^k \preccurlyeq M_i^k \quad \text{and} \quad \alpha \mathrm{Id} \preccurlyeq N_i^k,$$

(iii) the sequences $(M_i^k)_{k\in\mathbb{N}}$ and $(N_i^k)_{k\in\mathbb{N}}$ are uniformly a.s. quasi-decreasing.

Convergence result

Theorem: Let's assume that X and Y are separable, and that the set of saddle-points is non-empty. If the assumptions on the step-sizes are met, then the adaptive SPDHG algorithm almost surely converges to a saddle-point.

Idea: online primal-dual balancing

SPDHG: for $\gamma > 0$ and $0 < \beta < 1$,

$$\tau = \gamma * \beta \min \frac{p_i}{\|A_i\|}, \quad \sigma_i = \frac{\beta}{\gamma * \|A_i\|}.$$

Adaptive SPDHG:

$$\tau^k = \gamma^k * \beta \min \frac{p_i}{\|A_i\|}, \quad \sigma_i^k = \frac{\beta}{\gamma^k * \|A_i\|}.$$

Let (ϵ^k) be a non-negative sequence such that $\sum_k \epsilon_k < \infty$, e.g. $\epsilon^k = 1/k^2$. The assumptions of the theorem are satisfied if at each iteration:

- either $\gamma^{k+1} = (1 + \epsilon^{k+1})\gamma^k$ (increase the primal step-size),
- or $\gamma^{k+1} = \gamma^k/(1 + \epsilon^{k+1})$ (increase the dual step-sizes)

and the criterion for the choice above depends only of iterates' values up to step k.

Update rule (a)

At step k, let i be the updated index and define the residuals' norms

$$v_k = \|(x^{k-1} - x^k)/\tau^k - p_i^{-1} A_i^T (y_i^{k-1} - y_i^k)\|_1$$
$$d_k = p_i^{-1} \|(y^{k-1} - y^k)/\sigma^k - A_i (x^{k-1} - x^k)\|_1.$$

For some $\delta > 1$, the update rule is

- if $v^k < d^k/\delta$, then $\gamma^{k+1} = (1 + \epsilon^{k+1})\gamma^k$;
- if $v^k > \delta d^k$, then $\gamma^{k+1} = (1 + \epsilon^{k+1})\gamma^k$.

Comments:

- Equivalent to rule proposed in Goldstein et al for PDHG
- Computational overhead: $+50\%$ in theory because of computation $A_i(x^{k-1} - x^k)$. In practice, we replace the dual residual norm by a sub-sampled approximation, resulting in a $+5\%$ overhead, with identical results.

Update rule (b)

At step k, let i be the updated index and define

$$q^k = (x^{k-1} - x^k)/\tau^k - p_i^{-1} A_i^T (y_i^{k-1} - y_i^k)$$
$$w^k = \langle x^{k-1} - x^k, q^k \rangle / \|x^{k-1} - x^k\|_2 \|q^k\|_2$$

For some $c > 0$, the update rule is

- if $v^k < 0$, then $\gamma^{k+1} = (1 + \epsilon^{k+1})\gamma^k$;
- if $w^k > c$, then $\gamma^{k+1} = (1 + \epsilon^{k+1})\gamma^k$.

Comments:

- Used by Zdun and Brandt for SPDHG
- No computational overhead

Numerical experiments

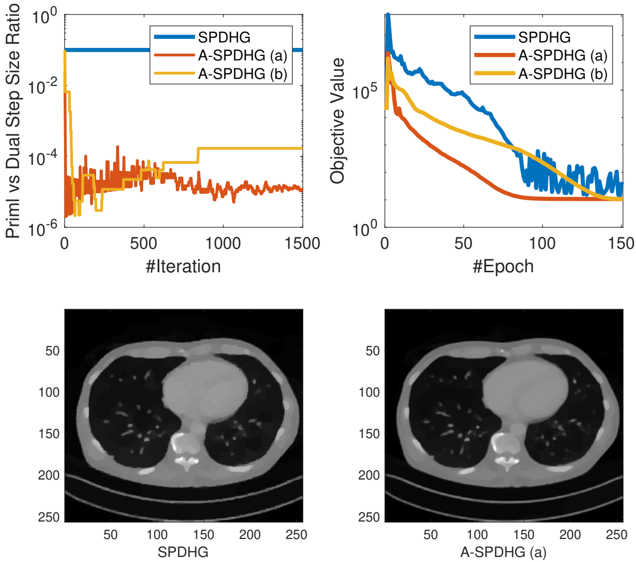
Setting: fanbeam Computerized Tomography (CT) measurements corrupted by Gaussian noise

$$\arg\min \frac{1}{2}\|Ax - y\|_2^2 + \lambda\|\nabla x\|_1$$

For different values of γ, we compare

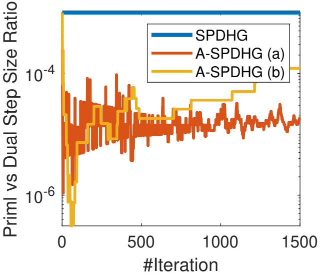
- SPDHG with fixed γ
- adaptive SPDHG with $\gamma^0 = \gamma$

$\gamma = 10^{-1}$

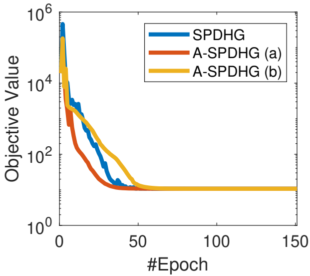


Priml vs Dual Step Size Ratio
#Iteration
SPDHG
A-SPDHG (a)
A-SPDHG (b)
Objective Value
#Epoch
SPDHG
A-SPDHG (a)

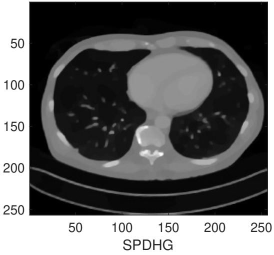
$\gamma = 10^{-3}$



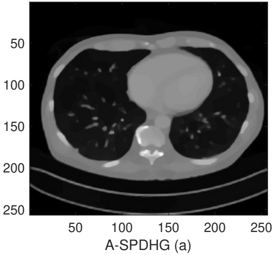
Priml vs Dual Step Size Ratio
10^{-4}
10^{-6}
0
500
1000
1500
#Iteration
SPDHG
A-SPDHG (a)
A-SPDHG (b)



Objective Value
10^{6}
10^{4}
10^{2}
10^{0}
0
50
100
150
#Epoch
SPDHG
A-SPDHG (a)
A-SPDHG (b)

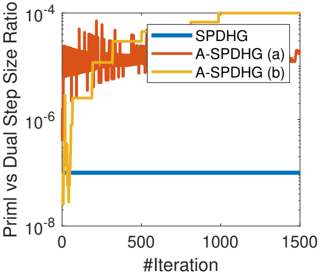


50
100
150
200
250
50
100
150
200
250
SPDHG

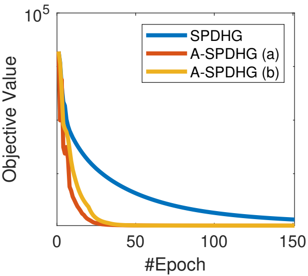


50
100
150
200
250
50
100
150
200
250
A-SPDHG (a)

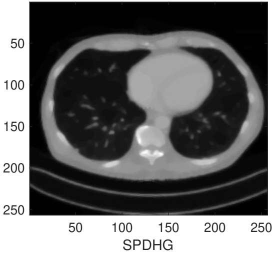
$\gamma = 10^{-7}$



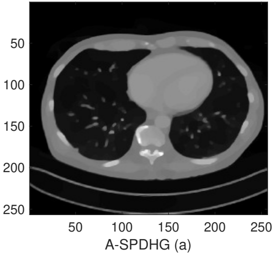
Priml vs Dual Step Size Ratio
10^-4
10^-6
10^-8
0
500
1000
1500
#Iteration
SPDHG
A-SPDHG (a)
A-SPDHG (b)



Objective Value
10^5
0
50
100
150
#Epoch
SPDHG
A-SPDHG (a)
A-SPDHG (b)

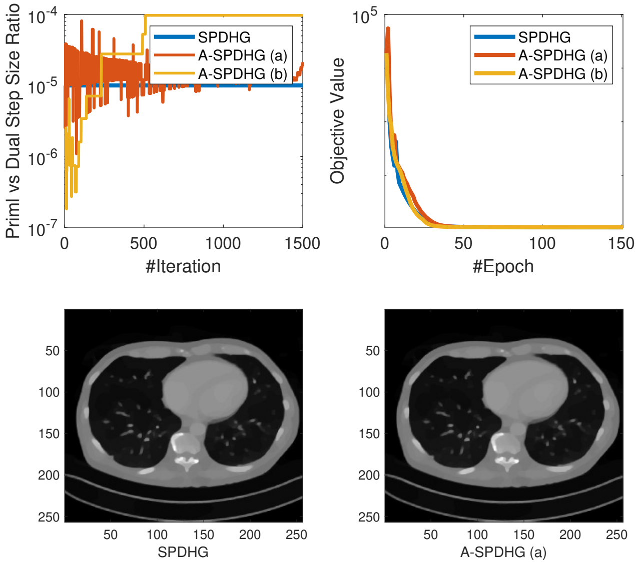


50
100
150
200
250
50
100
150
200
250
SPDHG



50
100
150
200
250
50
100
150
200
250
A-SPDHG (a)

$\gamma = 10^{-5}$



Priml vs Dual Step Size Ratio
10-4
10-5
10-6
10-7
0
500
1000
1500
#Iteration
SPDHG
A-SPDHG (a)
A-SPDHG (b)
Objective Value
105
0
50
100
150
#Epoch
SPDHG
A-SPDHG (a)
A-SPDHG (b)
50
100
150
200
250
SPDHG
50
100
150
200
250
A-SPDHG (a)

Take-home message

SPDHG offers excellent performance on large-scale inverse problems but depends on the tuning of a free parameter.

In turn, adaptive SPDHG offers:

- improved convergence speed
- convergence guarantees
- easy implementation and small overhead