The Limitation and Practical Acceleration of Stochastic Gradient Algorithms in Inverse Problems

Junqi Tang

University of Edinburgh

ICASSP 2019

Joint work with Karen Egiazarian and Mike Davies

Many inverse problems involve solving convex composite optimization tasks:

$$x^\star \in \arg\min_{x \in \mathcal{X}} \left\{ F(x) := \frac{1}{n} \sum_{i=1}^{n} \bar{f}(a_i, b_i, x) + \lambda g(x) \right\}, \tag{1}$$

Data fidelity term $f(x) := \frac{1}{n} \sum_{i=1}^{n} \bar{f}(a_i, b_i, x)$, regularization $g(x)$.
In imaging inverse problems:

- $x \in \mathbb{R}^d \rightarrow$ **vectorized image**,
  $A = [a_1; a_2; ...; a_n] \in \mathbb{R}^{n \times d} \rightarrow$ **the forward model/measurements** ,
  $b = [b_1; b_2; ...; b_n] \in \mathbb{R}^n \rightarrow$ **the observations**.

- Example: Total-Variation regularized least-squares

$$F(x) := \|Ax - b\|_2^2 + \lambda \|Dx\|_1. \tag{2}$$

($D \rightarrow$ discrete gradient operator.)

First-order optimization:

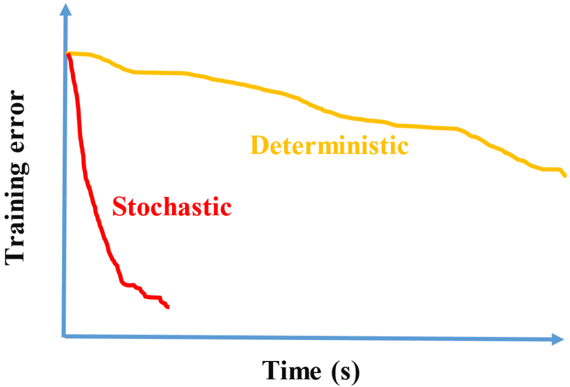
- **Deterministic gradients** $\rightarrow$ FISTA, PDHG, GFB, TOS, etc.
- **Stochastic gradients** $\rightarrow$ SGD, SVRG, SAG, Katyusha,..., etc

First-order optimization:

- **Deterministic gradients** $\rightarrow$ large per-iteration cost scales with n
- **Stochastic gradients**
  $\rightarrow$ small per-iteration cost
  $\rightarrow$ Optimal convergence rate via variance-reduction + momentum

Success of Stochastic Optimization in Machine Learning

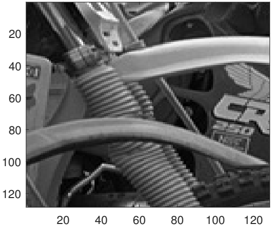
Stochastic gradient methods are almost always preferred than detereministic methods in machine learning practice.

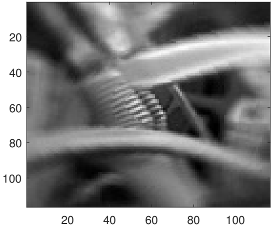


A Deblurring Experiment

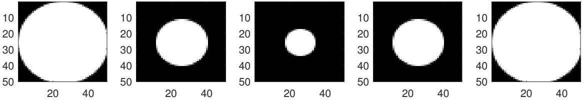
where stochastic gradient methods fail to be efficient

We consider a non-uniform deblurring task:





where the size of the blur kernel is space-varying:



edge $\rightarrow$ **central** $\rightarrow$ **edge**

A Deblurring Experiment

where stochastic gradient methods fail to be efficient

Deblur with TV regularization

$$F(x) := \|Ax - b\|_2^2 + \lambda \|Dx\|_1. \quad (3)$$

FISTA beats the best stochastic algorithms (with 10% random subsampling in each iteration).

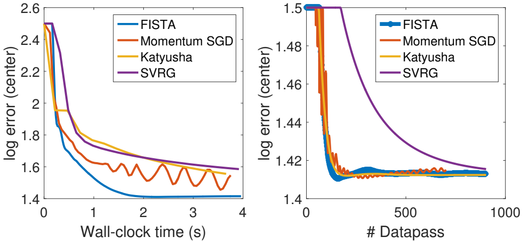


Figure: the estimation error of the central part (100 by 100) of the image.

A Deblurring Experiment

where stochastic gradient methods fail to be efficient

(at least) two pitfalls of stochastic gradient methods in imaging inverse problems:

- Fundamental limitation : for some tasks we indeed cannot expect significant benefit from stochastic gradient methods
- Inefficiency regarding the proximal operators : Compared to FISTA, the stochastic gradient methods typically need to compute the proximal operator much more often.
  And..
  – the proximal operator may be non-trivial to compute.
  – we may have multiple non-smooth regularization terms.

A Deblurring Experiment

where stochastic gradient methods fail to be efficient

To move forward

- Fundamental limitation
  we need to:
  $\rightarrow$ identify whether a inverse problem is suitable for stochastic gradient methods.
  $\rightarrow$ find the best sampling scheme to maximize the potential of stochastic methods.
- Inefficiency regarding the proximal operators
  we need to:
  $\rightarrow$ choose/design appropriately the algorithmic framework.

Stochastic Acceleration Factor

For a given a minibatch index $[S_0, S_1, S_2, ..., S_K]$ such that $S_1 \cup S_2 \cup ... \cup S_K = [n]$ and:

$$f_{S_k}(x) = \frac{K}{2n} \sum_{i \in S_k} f_i(x), \quad \nabla f_{S_k}(x) := \frac{K}{n} \sum_{i \in S_k} \nabla f_i(x), \tag{4}$$

while $k \in [K]$.

Assumption

(Smoothness of the Full-Batch and the Mini-Batches.)

f is L_f-smooth and each f_{S_k} is L_b-smooth, that is:

$$f(x) - f(y) - \nabla f(y)^T (x - y) \leq \frac{L_f}{2} \|x - y\|_2^2, \quad \forall x, y \in \mathcal{X}, \tag{5}$$

and

$$f_{S_k}(x) - f_{S_k}(y) - \nabla f_{S_k}(y)^T (x - y) \leq \frac{L_b}{2} \|x - y\|_2^2, \quad \forall x, y \in \mathcal{X}, \tag{6}$$

Stochastic Acceleration Factor

Definition

(SA Factor Function.) For a given minibatch index $\bar{S} = [S_1, ...S_K]$, the Stochastic Acceleration (SA) factor function is defined as:

$$\Upsilon(A, \bar{S}, K) = \frac{KL_f}{L_b} \quad (7)$$

Definition

(The class of optimal deterministic gradient algorithms.)
A deterministic gradient method $\mathcal{A}_{\text{full}}$ is called optimal if for any $s \geq 1$, the update of s-th iteration $x^s_{\mathcal{A}_{\text{full}}}$ satisfies:

$$F(x^s_{\mathcal{A}_{\text{full}}}) - F^\star \leq \frac{C_1 L_f \|x^0 - x^\star\|_2^2}{s^2}, \tag{8}$$

for some positive constant C_1.

Stochastic Acceleration Factor

Motivation

Definition

(The class of optimal stochastic gradient algorithms.)

A stochastic gradient method $\mathcal{A}_{\text{stoc}}$ is called optimal if for any $s \geq 1$ and $K \geq 1$, after a number of $s \cdot K$ stochastic gradient evaluations, the output of the algorithm $x^s_{\mathcal{A}_{\text{stoc}}}$ satisfies:

$$\mathbb{E}F(x^s_{\mathcal{A}_{\text{stoc}}}) - F^\star \leq \frac{C_2[F(x^0) - F^\star]}{s^2} + \frac{C_3 L_b \|x^0 - x^\star\|_2^2}{Ks^2}, \tag{9}$$

for some positive constants C_2 and C_3.

Stochastic Acceleration Factor

A motivating theorem

Theorem (informal)

(A motivating theorem for SA factor function.) Denote an optimal deterministic gradient algorithm $\mathcal{A}_{\text{full}}$, and an optimal stochastic gradient algorithm $\mathcal{A}_{\text{stoc}}$. For some sufficiently large dimension d, there exists a worst case choice of objective F, such that:

$$\frac{\mathbb{E}F(x^s_{\mathcal{A}_{\text{stoc}}}) - F^\star}{F(x^s_{\mathcal{A}_{\text{full}}}) - F^\star} \geq c_0 \cdot \frac{L_b}{KL_f} \tag{10}$$

for some positive constant c_0 which do not depend on L_b, L_f and K.

(A upper bound can also be shown which is also scale with the ratio $\frac{L_b}{KL_f}$)

Stochastic Acceleration Factor

Definition

Definition

(SA Factor Function.) For a given minibatch index $\bar{S} = [S_1, ...S_K]$, the Stochastic Acceleration (SA) factor function is defined as:

$$\Upsilon(A, \bar{S}, K) = \frac{KL_f}{L_b} \quad (11)$$

Stochastic Acceleration Factor

Examples for regularized Least-squares regression

Consider the least squares loss function with different types of forward operator:

$$f(x) = \|Ax - b\|_2^2 = \frac{1}{K}\sum_{k=1}^{K} f_{S_k}(x), \tag{12}$$

$$f_{S_k}(x) := K\|A_{S_k}x - b_{S_k}\|_2^2, \tag{13}$$

Interleaving sampling:

$$f_{S_k}(x) := \frac{K}{n}\sum_{i=1}^{\lfloor n/K \rfloor} f_{k+iK}(x) = K\sum_{i=1}^{\lfloor n/K \rfloor}(a_{k+iK}^T x - b_{k+iK}) \tag{14}$$

Stochastic Acceleration Factor

Examples for regularized Least-squares regression

- space-varying deblurring ($A_{\mathrm{blur}} \in \mathbb{R}^{262144\times 262144}$, $g(x) = \|x\|_{TV}$)
- compressed sensing random matrix ($A_{\mathrm{rand}} \in \mathbb{R}^{500\times 2000}$, $g(x) = \|x\|_1$)
- X-ray CT ($A_{\mathrm{CT}} \in \mathbb{R}^{91240\times 65536}$, $g(x) = \|x\|_{TV}$)
- RCV1 dataset ($A_{\mathrm{rcv1}} \in \mathbb{R}^{20242\times 47236}$, $g(x) = \|x\|_1$)
- magic04 dataset ($A_{\mathrm{magic04}} \in \mathbb{R}^{19000\times 50}$, $g(x) = \|x\|_1$)

Stochastic Acceleration Factor

Examples for regularized Least-squares regression

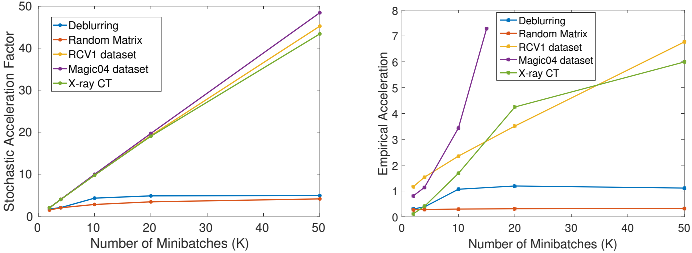


Figure: Left: Stochastic Acceleration (SA) factor of inverse problems with different forward operators.
Right: Empirical observation comparing the objective gap convergence of Katyusha and FISTA algorithm in 15 epochs.

Stochastic Primal-Dual Three-Operator Splitting

Tackling the inefficiency on proximal operators in stochastic optimization

Consider now a generic composite minimization task with two regularization terms (with a linear operator):

$$x^\star \in \arg\min_{x \in \mathbb{R}^d} \left\{ F(x) := f(x) + \lambda g(Dx) + \mu h(x) \right\}, \tag{15}$$

The saddle-point formulation can be written as:

$$[x^\star, y^\star] = \min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^r} f(x) + h(x) + y^T Dx - \lambda g^*(y) \tag{16}$$

Tackling the inefficiency on proximal operators in stochastic optimization

To move forward

- Fundamental limitation
- **Inefficiency regarding the proximal operators**
  we need to:
  $\rightarrow$ choose/design appropriately the algorithmic framework.

Accelerated Primal-Dual SGD

Tackling the inefficiency on proximal operators in stochastic optimization

Initialization: $x^0 = v^0 = v^{-1} \in \mathrm{dom}(g)$, the step size sequences $\alpha_{(.)}, \eta_{(.)}, \theta_{(.)}$, $l = 0$, and a balanced sampling partition $\bar{S}$.

Outer loop (Momentum) (t = 1, 2, 3, ... N):

$x^t \leftarrow \frac{(3t-2)v^{t-1}+tx^{t-1}-(2t-4)v^{t-2}}{2t+2}$, $x_0 \leftarrow x^t$, $z_0 \leftarrow x^t$, $y_0 \leftarrow Dx_0$

Inner loop (k = 1, 2, 3, ... K):

$l \leftarrow l + 1$, Pick $i \in [1, 2, ...K]$ uniformly at random

Dual Ascent $\rightarrow y_{k+1} = \mathrm{prox}_{\lambda g^*}^{\alpha_l}(y_k + \alpha_l D z_k)$

Primal Descent $\rightarrow x_{k+1} = \mathrm{prox}_{\gamma h}^{\eta_l}\left(x_k - \eta_l(D^T y_{k+1} + \nabla f_{S_i}(x_k))\right)$

Inner-loop Momentum $\rightarrow z_{k+1} = x_{k+1} + \theta_l(x_{k+1} - x_k)$

$v^t \leftarrow x_K$

Return x^t

Space-Varying Deblurring Experiment

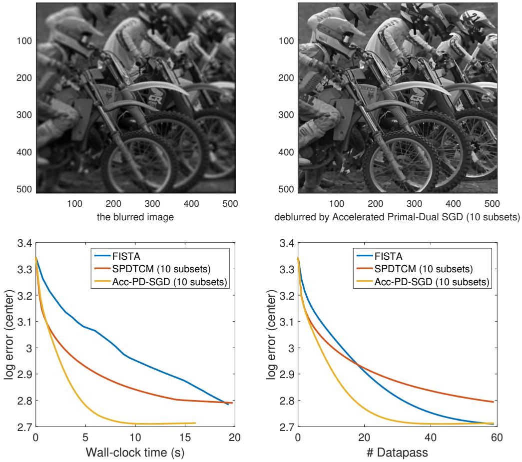


Figure: The estimation error plot for the deblurring experiment with TV-regularization. Image: Kodim05, with an additive Guassian noise (variance 1).

X-Ray CT reconstruction

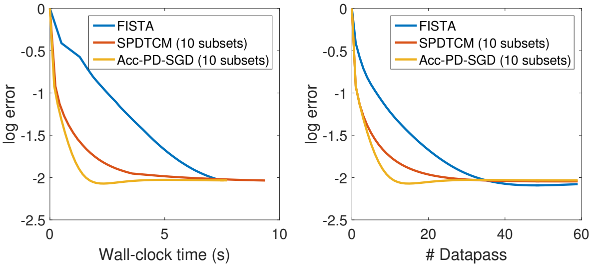


Figure: The estimation error plot for the X-ray CT image reconstruction experiment with TV-regularization. Measurement SNR : $\log_{10} \frac{\|Ax^\dagger\|_2^2}{\|w\|_2^2} \approx 3.16$

X-Ray CT reconstruction

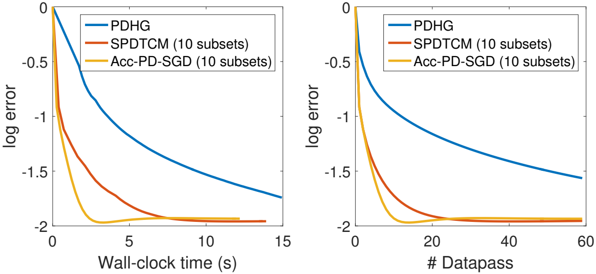


Figure: The estimation error plot for a noisy X-ray CT image reconstruction experiment with TV-regularization and ℓ_1 regularization on Haar-wavelet basis. Measurement SNR : $\log_{10} \frac{\|Ax^\dagger\|_2^2}{\|w\|_2^2} \approx 2.86$

X-Ray CT reconstruction

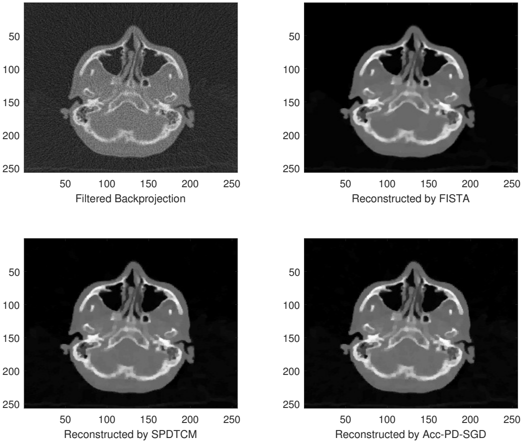


Figure: The reconstructed images by the compared algorithms with TV-regularization.

X-Ray CT reconstruction

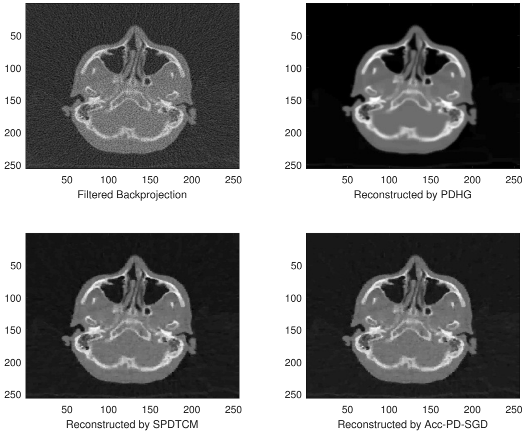


Figure: The reconstructed images by the compared algorithms at termination using joint TV-ℓ_1 regularization.

Summary

Take-home messages:

- For some inverse problems we cannot expect too much benefit from randomized algorithms.
- We can effectively characterize this fundamental limitation via the SA factor function
- We propose an accelerated stochastic primal-dual framework for efficiently handle the proximal operators.

On-going works:

- Understand the connection between inherent structure of forward model A and SA factor function.
- Design the optimal sampling scheme for SGD via the SA factor function.
- Extensions to plug-and-play algorithms.