Nesterov Accelerated Shuffling Gradient Method for Convex Optimization

Trang H. Tran[1] · Katya Scheinberg [1] · Lam M. Nguyen[2]

[1] Cornell University, School of Operations Research and Information Engineering [2] IBM Research, Thomas J. Watson Research Center

* Correspondence to Lam M. Nguyen, `LamNguyen.MLTD@ibm.com`.

ICML | 2022

Problem Statement

We consider the following **finite-sum minimization:**

$$\min_{w\in\mathbb{R}^d}\left\{F(w) := \frac{1}{n}\sum_{i=1}^{n} f(w;i)\right\}, \qquad (1)$$

where $f(\cdot;i):\mathbb{R}^d\to\mathbb{R}$ is **a Lipschitz smooth function** for $i\in[n] := \{1,\dots,n\}$, and F is **convex**. Assume that we have access to the first order oracle of $f(\cdot;i)$. Below are some common sampling schemes:

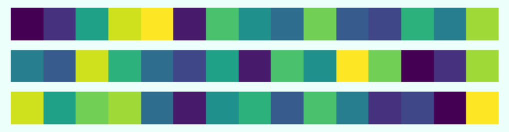
Regular (Standard) Scheme: Uniformly at random: at each iteration i_t of epoch t, sample an index uniformly at random from $[n]$.

Shuffling Schemes:

Incremental Gradient: use a fixed permutation $\{1,\dots,n\}$ for all epochs.

Shuffle Once: random shuffle one permutation and use it for all epochs.

Random Reshuffling: random shuffle a new permutation at every epoch.



Nesterov Accelerated Shuffling Gradient

Algorithm 1: Nesterov Accelerated Shuffling Gradient (NASG) Method

Initialization: Choose an initial point $\tilde{x}_0, \tilde{y}_0 \in \mathbb{R}^d$.

for $t = 1, 2, \cdots, T$ **do**

  Set $y_0^{(t)} := \tilde{y}_{t-1}$;

  Generate any permutation $\pi^{(t)}$ of $[n]$ (either deterministic or random);

  for $i = 1, \cdots, n$ **do**

    Update $y_i^{(t)} := y_{i-1}^{(t)} - \eta_i^{(t)}\nabla f(y_{i-1}^{(t)}; \pi^{(t)}(i))$;

  end for

  Set $\tilde{x}_t := y_n^{(t)}$;

  Update $\tilde{y}_t := \tilde{x}_t + \gamma_t(\tilde{x}_t - \tilde{x}_{t-1})$;

end for

Comparison with deterministic NAG:

- Inner loop of deterministic NAG

  for $i = 1, \cdots, n$ **do**

    Update $y_i^{(t)} := y_{i-1}^{(t)} - \eta_i^{(t)}\nabla f(y_0^{(t)}; \pi^{(t)}(i))$; ← fixed point

  end for

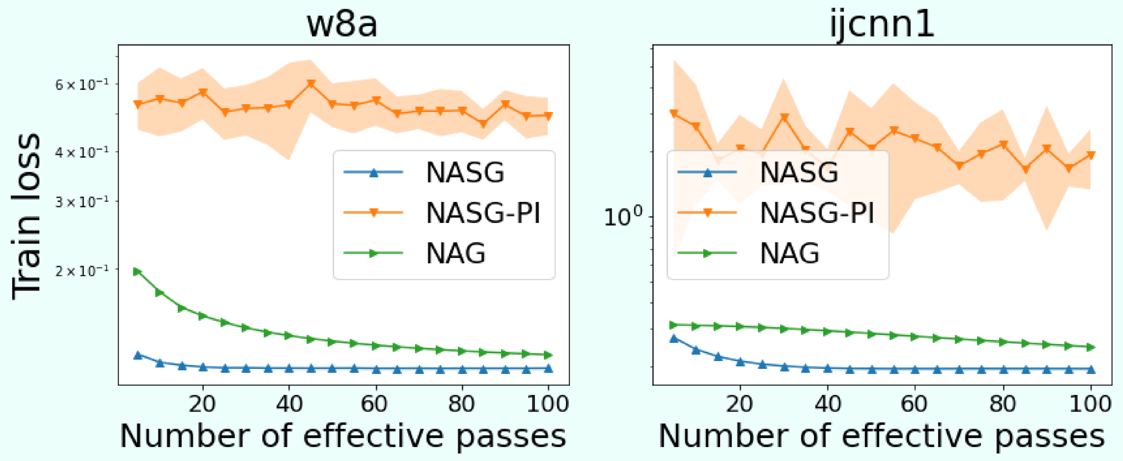
- Inner loop of stochastic NASG

  for $i = 1, \cdots, n$ **do**

    Update $y_i^{(t)} := y_{i-1}^{(t)} - \eta_i^{(t)}\nabla f(y_{i-1}^{(t)}; \pi^{(t)}(i))$; ← moving continuously

  end for

Our binary classification experiments for `w8a` and `ijcnn1` datasets show our motivation. NASG-PI is the stochastic version that applies Nesterov momentum per iteration, while our method is per epoch.



Assumptions

Problem (1) satisfies:

(a) (Bounded below and convexity for F) We assume the existence of a minimizer for F, and F is convex.

(b) (L-smoothness) $f(\cdot;i)$ is L-smooth for all $i\in[n]$: , i.e., there exists $L>0$:

$$\forall w, w' \in \operatorname{dom}(F)\ \|\nabla f(w;i) - \nabla f(w';i)\| \le L\|w - w'\|. \qquad (2)$$

We let x_* be any minimizer of F and consider the variance of F at x_*:

$$\sigma_*^2 := \frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x_*;i)\|^2 \in [0,+\infty). \qquad (3)$$

In addition, we assume either (c1) or (c2):

(c1) (Individual convexity) $f(\cdot;i)$ is convex for all $i\in[n]$.

(c2) (Generalized bounded variance) There exist two finite constants $\Theta, \sigma \ge 0$:

$$\forall w \in \operatorname{dom}(F): \frac{1}{n}\sum_{i=1}^{n}\|\nabla f(w;i) - \nabla F(w)\|^2 \le \Theta\|\nabla F(w)\|^2 + \sigma^2. \qquad (4)$$

Main results

Theorem 1 - Unified Schemes (Informal)

We assume Assumption (a) and (b) with either (c1) or (c2) is satisfied. Let $\Delta := \|\tilde{x}_0 - x_*\|^2$ with the initial point $\tilde{x}_0$ and the minimizer x_*. With an appropriate choice of the learning rate, $F(\tilde{x}_T) - F(x_*)$ is upper bounded by

either $\mathcal{O}\left(\frac{\sigma_*^2/L + L\Delta}{T}\right)$, for individual convexity (c1)

or $\mathcal{O}\left(\frac{\sigma^2/(\Theta L) + L\Theta^{1/3}\Delta}{T}\right)$, for generalized bounded variance (c2)

The convergence rate of NASG is better than the current state-of-the-art rate in term of T for convex problems with general shuffling-type strategies [1, 3].

Theorem 2 - Randomized Schemes (Informal)

Suppose that Assumption (a), (b) and (c1) hold. Let $\Delta := \|\tilde{x}_0 - x_*\|^2$ with the initial point $\tilde{x}_0$ and the minimizer x_*. With an appropriate choice of the learning rate and randomized shuffling schemes, we have

$$\mathbb{E}[F(\tilde{x}_T) - F(x_*)] \le \mathcal{O}\left(\frac{\sigma_*^2/L}{nT} + \frac{L\Delta}{T}\right)$$

This rate has a factor of n improved, and is better than the corresponding rate for randomized schemes in the literature for convex problems [1, 3]. In the table below, we show **the complexity to reach an ϵ-accurate solution** x that satisfies $F(x) - F(x_*) \le \epsilon$ (or $\mathbb{E}[F(x) - F(x_*)] \le \epsilon$ in random case).

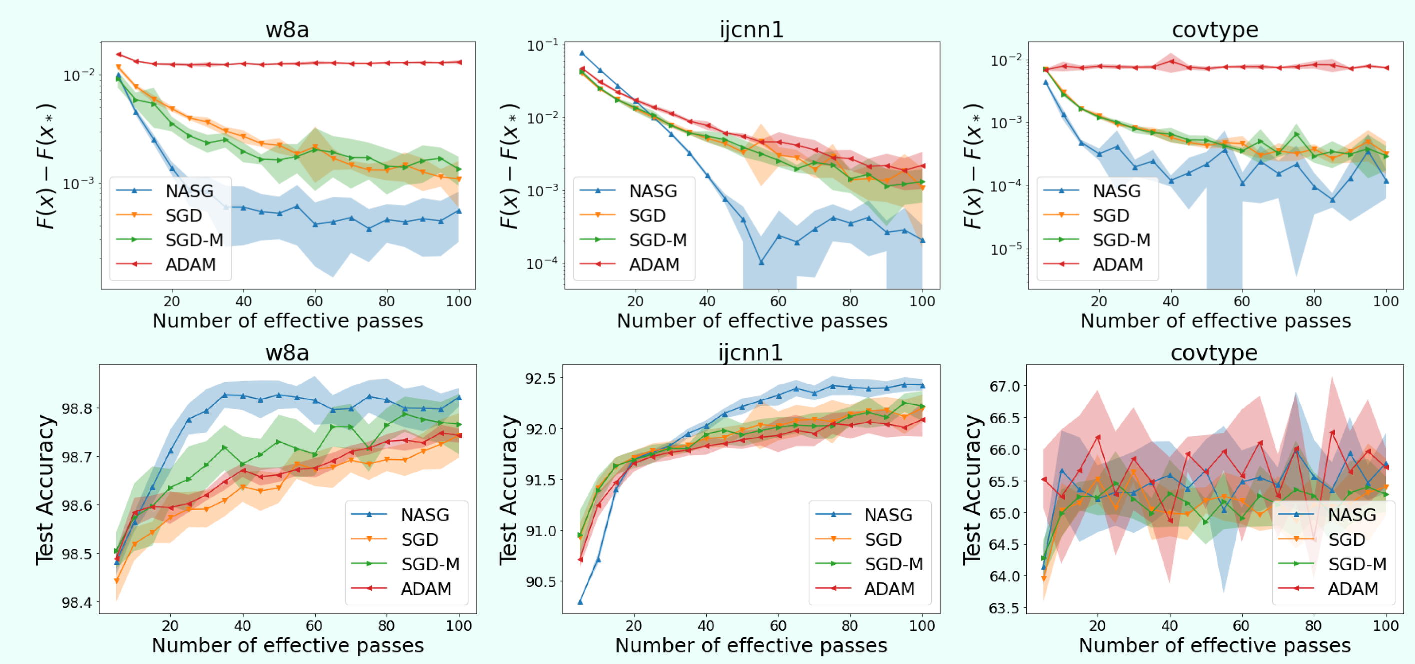
Algorithms	Complexity	References
Standard SGD(1)	$\mathcal{O}\left(\frac{\Delta_0^2+G^2}{\epsilon^2}\right)$ (1)	[2, 4]
SGD - Unified Schemes	$\mathcal{O}\left(\frac{nL\Delta}{\epsilon} + \frac{n\sqrt{L}\sigma_*\Delta}{\epsilon^{3/2}}\right)$	[1, 3]
SGD - Randomized Schemes	$\mathcal{O}\left(\frac{nL\Delta}{\epsilon} + \frac{\sqrt{nL}\sigma_*\Delta}{\epsilon^{3/2}}\right)$	[1, 3]
NASG - Unified Schemes	$\mathcal{O}\left(\frac{nL\Delta}{\epsilon} + \frac{n\sigma_*^2}{L\epsilon}\right)$	Theorem 1
NASG - Randomized Schemes	$\mathcal{O}\left(\frac{nL\Delta}{\epsilon} + \frac{\sigma_*^2}{L\epsilon}\right)$	Theorem 2

(1) Standard results for SGD often use bounded domain that $\|x - x_*\|^2 \le \Delta_0$ for each iterate x and/or bounded gradient that $\mathbb{E}[\|\nabla f(x;i)\|] \le G^2$.

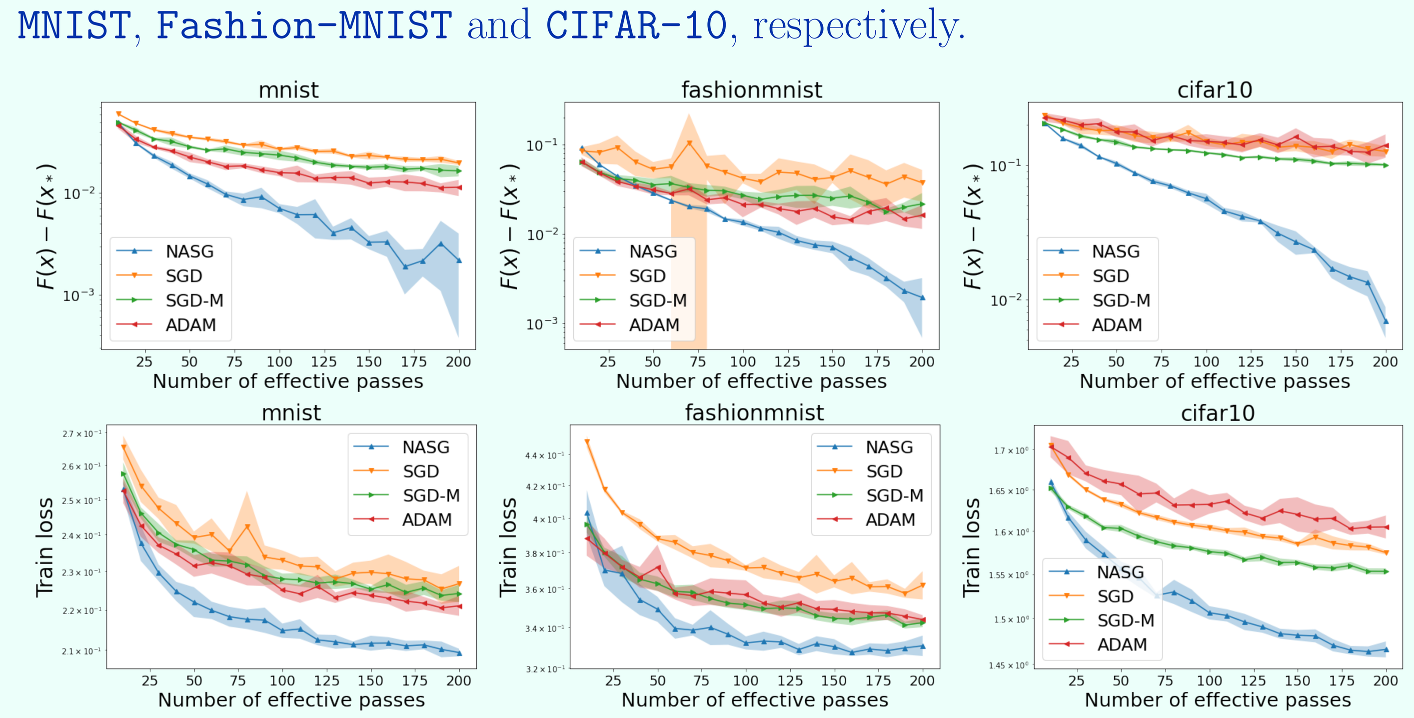
Experiments

We test **NASG method with SGD algorithm, SGD with momentum and ADAM**. Our tests have shown encouraging results for NASG.

(Convex Binary Classification). For the first experiment, we choose a binary classification problem. Below, we show comparisons of loss residual $F(x) - F(x_*)$ (top) and test accuracy (bottom) produced by first-order methods for `w8a`, `ijcnn1` and `covtype` datasets, respectively.



(Convex and Non-convex Image Classification). We test four methods for the second problem: training a neural network to classify images. Our figure below compares the loss residual $F(x) - F(x_*)$ (convex setting, top) and train loss $F(x)$ (non-convex setting, bottom) produced by first-order methods for `MNIST`, `Fashion-MNIST` and `CIFAR-10`, respectively.



Key References

[1] Mishchenko, K., Khaled Ragab Bayoumi, A., and Richtárik P. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems, 33, 2020.*

[2] Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM J. on Optimization, 19(4):1574–1609, 2009.*

[3] Nguyen, L. M., Tran-Dinh, Q., Phan, D. T., Nguyen, P. H., and van Dijk, M. A unified convergence analysis for shuffling-type gradient methods. *Journal of Machine Learning Research, 22(207):1–44.*

[4] Shamir, O. and Zhang, T. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. *Proceedings of the 30th International Conference on Machine Learning, PMLR.*

[5] Tran, T. H., Nguyen, L. M., and Tran-Dinh, Q. SMG: A shuffling gradient-based method with momentum. *Proceedings of the 38th International Conference on Machine Learning, PMLR.*