# NAG-GS: Semi-Implicit, Accelerated and Robust Stochastic Optimizer

Valentin Leplat[1], Daniil Merkulov[2,3], Aleksandr Katrutsa[2,4], Daniel Bershatsky[2], Olga Tsymboi[2,5], and Ivan Oseledets[4,2]

## Introduction and Motivation

We consider the unconstrained minimization problem of a smooth convex function:

$$\min_{x\in\mathbb{R}^n} f(x)$$

### Main Contributions:

- We present a stochastic extension of the algorithm based on Gauss-Seidel discretization of the ODE related to the accelerated gradient method.
- We provide an asymptotic convergence analysis for strongly convex quadratic objectives and identify the maximum feasible learning rate.
- We demonstrate experimentally that NAG-GS converges faster in the initial epochs and achieves similar or better final test accuracy on logistic regression, VGG-11, ResNet-20, and Transformer models.

## Proposed Method

### Accelerated Stochastic Gradient Flow:

$$\frac{dx}{dt} = v - x, \qquad \dot{\gamma}(t) = \mu - \gamma(t)$$
$$\frac{dv}{dt} = \frac{\mu}{\gamma}(x - v) - \frac{1}{\gamma}\nabla f(x) + \sigma\frac{dW}{dt}, \qquad \gamma(0) = \gamma_0 > 0$$

where $\mu$ is the strong convexity parameter, and $W$ is a standard $n$-dimensional Brownian motion.

### Gauss-Seidel Discretization:

$$x_{k+1} = (1 - a_k)x_k + a_k v_k,$$
$$v_{k+1} = (1 - b_k)v_k + b_k x_{k+1} - \mu^{-1} b_k \nabla\tilde{f}(x_{k+1}),$$

where $a_k$ and $b_k$ are step size parameters, and $\nabla\tilde{f}(x_{k+1})$ is the possibly noisy gradient.

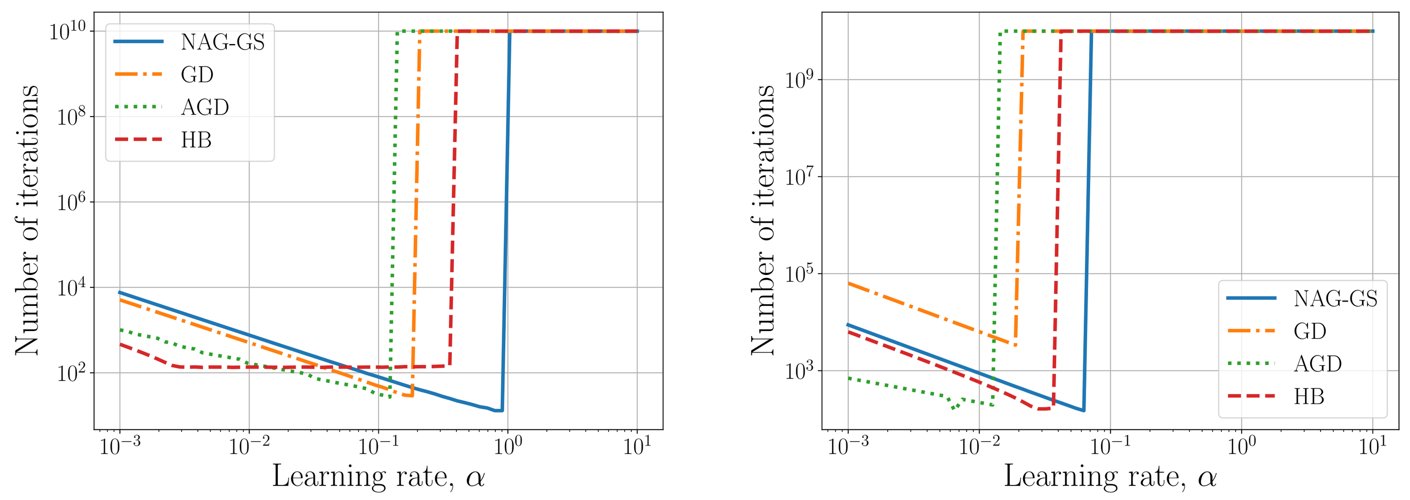
**Theorem:** *For $f(x) = \frac{1}{2}x^\top Ax$, with $A$ symmetric positive definite, and assuming $0 < \mu = \lambda_1 \leq \ldots \leq \lambda_n = L < \infty$, and given $\gamma \geq \mu$, if $0 < \alpha \leq \frac{\mu+\gamma+\sqrt{(\mu-\gamma)^2+4\gamma L}}{L-\mu}$, then the NAG-GS method converges.*

**Algorithm:** Nesterov Accelerated Gradient with Gauss–Seidel splitting (NAG-GS)

**Input:** Initial point $x_0$, parameters $\mu \geq 0$, $\gamma_0 > 0$
Set $v_0 := x_0$
**for** $k = 1, 2, \ldots$ **do**
  Choose step size $\alpha_k > 0$
  Set $a_k := \alpha_k(\alpha_k + 1)^{-1}$
  Update $\gamma_{k+1} := (1 - a_k)\gamma_k + a_k\mu$
  Update $x_{k+1} := (1 - a_k)x_k + a_k v_k$
  Set $b_k := \alpha_k\mu(\alpha_k\mu + \gamma_{k+1})^{-1}$
  Update $v_{k+1} := (1 - b_k)v_k + b_k x_{k+1} - \mu^{-1} b_k \nabla\tilde{f}(x_{k+1})$
**end**

## Experiments

### Strongly convex quadratics




(a) $\mu = 1$, $L = 10$ (b) $\mu = 10^{-1}$, $L = 100$

Figure 1: Dependence of the number of iterations needed for convergence on the learning rate. NAG-GS is more robust with respect to the learning rate than gradient descent (GD) and accelerated gradient descent (AGD). The number of iterations $10^{10}$ indicates the divergence.

### ResNet-20 on CIFAR-10

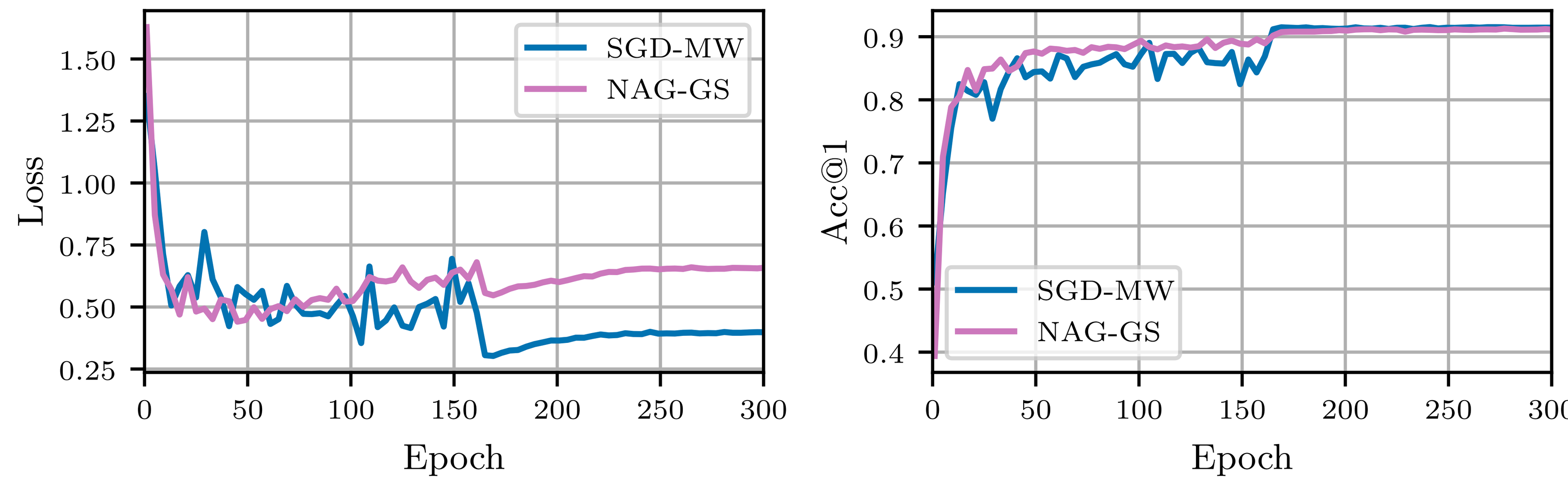



Figure 2: NAG-GS outperforms SGD-MW uniformly in the first 150 epochs and provides the same accuracy further.

### VGG-11 on CIFAR-10

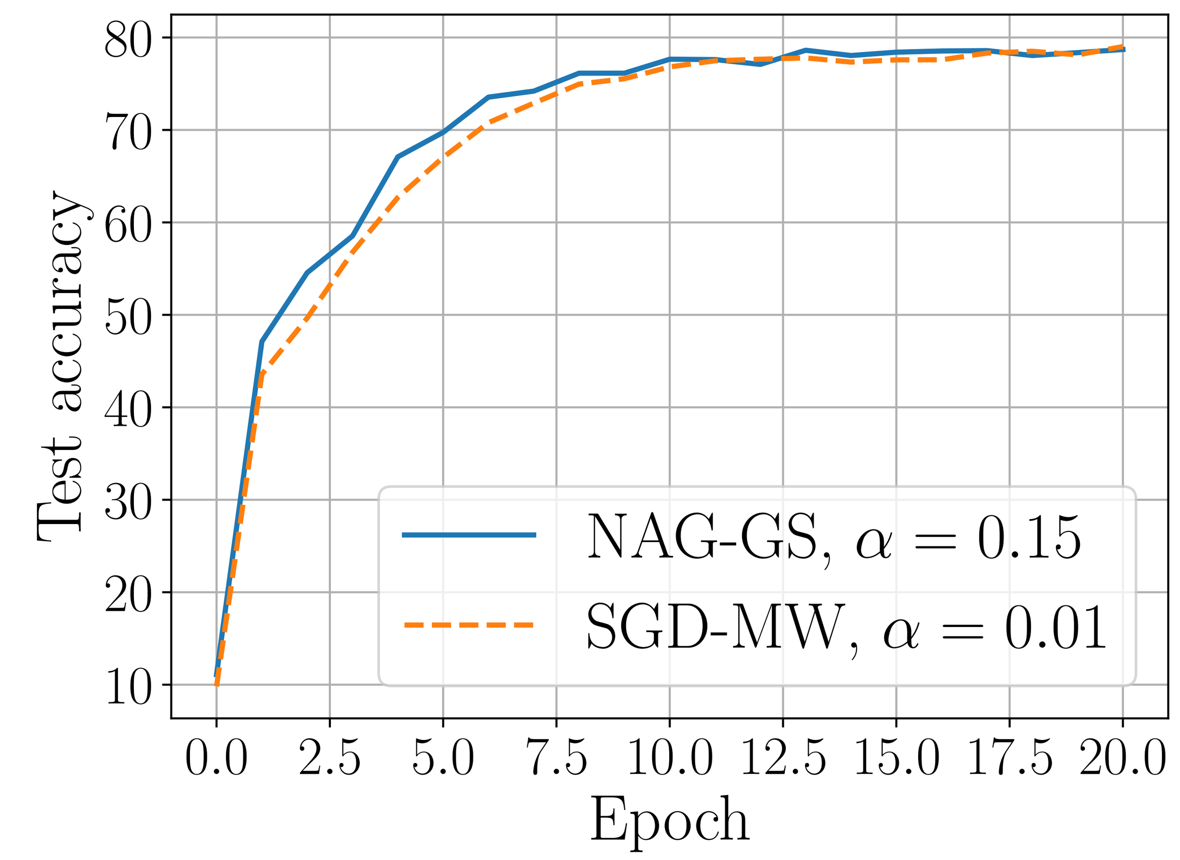



Figure 3: Comparison of the convergence of NAG-GS and SGD-MW with the best learning rates. NAG-GS gives a higher test accuracy faster than SGD-MW (see 1–10 epochs) while converging to a similar test accuracy in the middle of training.

### Vision Transformer on food101

| Stage | NAG-GS | AdamW |
|---|---|---|
| After 1 epoch | **0.8419** | 0.8269 |
| After 25 epochs | **0.8606** | 0.8324 |

Table 1: Test accuracies of NAG-GS and AdamW for Vision Transformer model fine-tuned on `food101` dataset. The NAG-GS outperforms AdamW after the presented number of epochs.

### RoBERTa on GLUE benchmark

| Optimizer | CoLA | MNLI | MRPC | QNLI | QQP |
|---|---|---|---|---|---|
| AdamW | **61.60** | **87.56** | 88.24 | **92.62** | **91.69** |
| NAG-GS | **61.60** | 87.24 | **90.69** | 92.59 | 91.01 |

| Optimizer | RTE | SST2 | STS-B | WNLI |
|---|---|---|---|---|
| AdamW | **78.34** | **94.95** | **90.68** | **56.34** |
| NAG-GS | 77.97 | 94.50 | 90.21 | **56.34** |

Table 2: Note that NAG-GS has lower computational complexity and memory requirements than AdamW.

[1] Innopolis University, [2] Skoltech, [3] Moscow Institute of Physics and Technology, [4] Artificial Intelligence Research Institute, [5] Sber AI Lab, Moscow, Russia.