

Supplementary materials to paper “NAG-GS: Semi-Implicit, Accelerated and Robust Stochastic Optimizer”

Valentin Leplat · Daniil Merkulov · Aleksandr
Katrutsa · Daniel Bershatsky · Olga Tsymboi ·
Ivan Oseledets

the date of receipt and acceptance should be inserted later

1 Additional remarks related to theoretical background

An accelerated ODE has been presented in the main text Section 1.1, which relied on a specific spectral transformation. In this brief section, we add some useful insights:

- Equation (11) is a variant of the heavy ball model with variable damping coefficients in front of \ddot{x} and \dot{x} .
- Thanks to the scaling factor γ , both the convex case $\mu = 0$ and the strongly convex case $\mu > 0$ can be handled in a unified way.
- In the continuous time, one can solve easily (9) as follows: $\gamma(t) = \mu + (\gamma_0 - \mu)e^{-t}$, $t \geq 0$. Since $\gamma_0 > 0$, we have that $\gamma(t) > 0$ for all $t \geq 0$ and $\gamma(t)$ converges to μ exponentially and monotonically as $t \rightarrow +\infty$. In particular, if $\gamma_0 = \mu > 0$, then $\gamma(t) = \mu$ for all $t \geq 0$. We remark here the links between

Valentin Leplat, Corresponding author
Innopolis University
Innopolis, Russia
v.leplat@innopolis.ru

Daniil Merkulov
Skoltech, Moscow Institute of Physics and Technology
Moscow, Russia

Aleksandr Katrutsa
Skoltech, AIRI
Moscow, Russia

Daniel Bershatsky
Skoltech
Moscow, Russia

Olga Tsymboi
Moscow Institute of Physics and Technology, Sber AI Lab
Moscow, Russia

Ivan Oseledets
AIRI, Skoltech
Moscow, Russia

- the behavior of the scaling factor $\gamma(t)$ and the sequence $\{\gamma_k\}_{k=0}^\infty$ introduced by Nesterov in its analysis of optimal first-order methods in discrete-time, see [15].
- Authors from [12] prove the exponential decay property $\mathcal{L}(t) \leq e^{-t} \mathcal{L}_0$, $t > 0$ for a Taylored Lyapunov function $\mathcal{L}(t) := f(x(t)) - f(x^*) + \frac{\gamma(t)}{2} \|v(t) - x^*\|^2$ where $x^* \in \operatorname{argmin} f$ is a global minimizer of f . Again, we note the similarity between the Lyapunov function proposed here and the estimating sequence $\{\phi_k(x)\}_{k=0}^\infty$ of function f introduced by Nesterov in its optimal first-order methods analysis [15]. In [15], this sequence takes the form $\phi_k(x) = \phi_k^*(x) + \frac{\gamma_k}{2} \|v_k - x\|^2$ where $\gamma_{k+1} := (1 - \alpha_k)\gamma_k + \alpha_k\mu$ and $v_{k+1} := \frac{1}{\gamma_{k+1}}[(1 - \alpha_k)\gamma_k v_k + \alpha_k\mu y_k - \alpha_k \nabla f(y_k)]$ which stand for a forward Euler discretization respectively of (9) and second ODE of (10).

We ask the attentive reader to remember that this discussion mainly concerns the continuous time case. A second central part of our analysis was based on the methods of discretization of (10). Indeed, these discretizations ensure, together with the spectral transformation (8), the optimal convergence rates of the methods and their particular ability to handle noisy gradients.

Finally, we delve into a crucial insight motivating the proposition of an optimizer that exhibits robustness concerning the choice of the step size, enabling the utilization of a wide range of values for the step size (or learning rate). An established approach for analyzing Stochastic Gradient Descent (SGD) involves viewing it as a discretization of a continuous-time process, expressed as:

$$dx_t = -\nabla f(x_t)dt + \sqrt{\alpha\sigma^2}dB_t,$$

where B_t denotes the standard Brownian motion. This stochastic differential equation (SDE) is a variant of the well-known Langevin diffusion. Under mild regularity assumptions on f , it can be shown that the Markov process $(x_t)_{t \geq 0}$ is ergodic, with its unique invariant measure having a density proportional to $\exp(-f(x)/(\alpha\sigma^2))$ for any $\alpha > 0$ [18].

Building on this observation, existing research has shed light on the relationship between the invariant measure and algorithm parameters. The study [7] focused on the interplay of the invariant measure with step-size α and mini-batch size as a function of σ^2 . They concluded that the ratio of learning rate to batch size serves as the control parameter determining the width of the minima found by SGD.

Additionally, the work [8] explored sharp and flat minimizers, their impact on generalization, and the distinctions between large-batch and small-batch methods, especially in deep neural networks. Their key observations include:

- Sharp and Flat Minimizers: Flat minimizers exhibit slow function variation in a wide neighborhood, while sharp minimizers show rapid increases in a small neighborhood. Flat minimizers can be described with lower precision, contrasting with the higher precision needed for sharp minimizers.
- Effect on Generalization: Sharp minimizers negatively affect model generalization due to their large sensitivity in the training function. The Minimum Description Length (MDL) theory suggests that lower complexity models generalize better, making flat minimizers preferable.

- Large-batch vs. Small-batch Methods: Large-batch methods tend to converge to sharp minimizers, reducing generalization ability. Conversely, small-batch methods converge to flat minimizers, generally leading to better generalization.
- Observation in Deep Neural Networks: The loss function landscape in deep neural networks attracts large-batch methods towards regions with sharp minimizers, trapping them and impeding their escape from these basins of attraction.

Motivated by these insights and assuming a Gaussian noise process, it becomes evident that proposing a solver capable of handling a broad range of step size values is crucial. Specifically, allowing for a larger ratio α/b (with b as the mini-batch size) increases the likelihood of converging to wider local minima, ultimately enhancing the generalization performance of the trained model.

2 Proof of the main theoretical result (Theorem 2.1)

As briefly mentioned in Section 2.3 of the main text, the two key elements to come up with a maximum (constant) step size for Algorithm 1 are the study of the spectral radius of iteration matrix associated with NAG-GS scheme (Section 2.1) and the covariance matrix at stationarity (Section 2.1.1) w.r.t. all the significant parameters of the scheme. These parameters are the step size (integration step/time step) α , the convexity parameters $0 \leq \mu \leq L \leq \infty$ of the function $f(x)$, the variance of the noise σ^2 and the positive scaling parameter γ . Note that this theoretical study only concerns the case $f(x) = \frac{1}{2}x^\top Ax$.

Reproducibility.

- In Section 2.1, we start by determining the explicit formulation of the spectral radius of the iteration matrix $\rho(E(\alpha))$, specifically for the 2-dimensional quadratic case. This formulation allows us to derive the optimal step size α_c that minimizes $\rho(E(\alpha))$, resulting in the highest convergence rate for the NAG-GS method. Notably, Lemma 2.2 presents a crucial outcome for the asymptotic convergence analysis of NAG-GS, revealing that $\rho(E(\alpha))$ is a strictly monotonically increasing function of α within a certain interval, under mild assumptions.
- In Section 2.1.1, we conduct an in-depth analysis of the covariance matrix at stationarity, which enables us to establish sufficient conditions for α_c to ensure the asymptotic convergence of the NAG-GS method. The formal proof for this convergence is presented in Lemma 2.3 for the case of $n = 2$.
- In Section 2.3, we provide the formal proof of Theorem 1, which is enunciated in the main text. This theorem stated the asymptotic convergence of the NAG-GS method for dimensions $n > 2$.

2.1 Spectral radius analysis

Let us assume $f(x) = \frac{1}{2}x^\top Ax$ and since $A \in \mathbb{S}_+^n$ by hypothesis, it is diagonalizable and can be presented as $A = \text{diag}(\lambda_1, \dots, \lambda_n)$ without loss of generality, that is to

say, that we will consider a system of coordinates composed of the eigenvectors of matrix A . Let us note that $\mu = \lambda_1 \leq \dots \leq \lambda_n = L$.

For the following, we restrict the discussion to the case $n = 2$. In this setting, $y = (x, v) \in \mathbb{R}^4$ and the matrices M and N from the Gauss-Seidel splitting of G_{NAG} (7) are:

$$M = \begin{bmatrix} -I_{2 \times 2} & 0_{2 \times 2} \\ \mu/\gamma I_{2 \times 2} - A/\gamma & -\mu/\gamma I_{2 \times 2} \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -\mu/\gamma & 0 \\ 0 & \mu/\gamma - L/\gamma & 0 & -\mu/\gamma \end{bmatrix},$$

$$N = \begin{bmatrix} 0_{2 \times 2} & I_{2 \times 2} \\ 0_{2 \times 2} & 0_{2 \times 2} \end{bmatrix}$$

For the minimization of $f(x) = \frac{1}{2}x^\top Ax$, given the property of Brownian motion $\Delta W_k = W_{k+1} - W_k = \sqrt{\alpha_k} \eta_k$ where $\eta_k \sim \mathcal{N}(0, 1)$, (12) reads:

$$y_{k+1} = (I_{4 \times 4} - \alpha M)^{-1} (I_{4 \times 4} + \alpha N) y_k + (I_{4 \times 4} - \alpha M)^{-1} \begin{bmatrix} 0 \\ \sigma \sqrt{\alpha} \eta_k \end{bmatrix} \quad (1)$$

Since matrix M is lower-triangular, matrix $I_{4 \times 4} - \alpha M$ is as well and can be factorized as follows:

$$\begin{aligned} I_{4 \times 4} - \alpha M &= DT \\ &= \begin{bmatrix} (1 + \alpha)I_{2 \times 2} & 0_{2 \times 2} \\ 0_{2 \times 2} & (1 + \frac{\alpha\mu}{\gamma})I_{2 \times 2} \end{bmatrix} \begin{bmatrix} I_{2 \times 2} & 0_{2 \times 2} \\ \frac{\alpha(A - \mu I_{2 \times 2})}{\gamma(1 + \frac{\alpha\mu}{\gamma})} & I_{2 \times 2} \end{bmatrix} \end{aligned}$$

Hence $(I_{4 \times 4} - \alpha M)^{-1} = T^{-1}D^{-1}$ where D^{-1} can be easily computed. It remains to compute T^{-1} ; T can be decomposed as follows: $T = I_{4 \times 4} + Q$ with Q a nilpotent matrix such that $QQ = 0_{4 \times 4}$. For such decomposition, it is well known that:

$$T^{-1} = (I_{4 \times 4} + Q)^{-1} = I_{4 \times 4} - Q = \begin{bmatrix} I_{2 \times 2} & 0_{2 \times 2} \\ \frac{\alpha(\mu I_{2 \times 2} - A)}{\gamma(1 + \tau_k)} & I_{2 \times 2} \end{bmatrix} \quad (2)$$

where $\tau_k = \frac{\alpha\mu}{\gamma}$. Combining these results, (1) finally reads:

$$\begin{aligned} y_{k+1} &= \begin{bmatrix} \frac{1}{\alpha+1} & 0 & \frac{\alpha}{1+\alpha} & 0 \\ 0 & \frac{1}{\alpha+1} & 0 & \frac{\alpha}{1+\alpha} \\ 0 & 0 & \frac{1}{1+\tau} & 0 \\ 0 & \frac{\alpha(\mu-L)}{\gamma(\tau+1)(\alpha+1)} & 0 & \frac{\alpha^2(\mu-L)}{\gamma(1+\tau)(1+\alpha)} + \frac{1}{1+\tau} \end{bmatrix} y_k + \begin{bmatrix} 0 \\ \sigma \frac{\sqrt{\alpha}}{1+\tau} \eta_k \end{bmatrix} \\ &= E y_k + \begin{bmatrix} 0 \\ \sigma \frac{\sqrt{\alpha}}{1+\tau} \eta_k \end{bmatrix} \end{aligned} \quad (3)$$

with E denoting the iteration matrix associated with the NAG-GS method. Hence, (3) includes two terms: the first is the product of the iteration matrix times the current vector y_k and the second one features the effect of the noise. For the latter, it will be studied in Section 2.1.1 from the point of view of maximum step size for the NAG-GS

method through the key quantity of the covariance matrix. Let us focus on the first term. It is clear that in order to get the maximum contraction rate, we should look for α that minimizes the spectral radius of E . Since the spectral radius is the maximum absolute value of the eigenvalues of iteration matrix E , we start by computing them. Let us find the expression of $\lambda_i \in \sigma(E)$ for $1 \leq i \leq 4$ that satisfies $\det(E - \lambda I_{4 \times 4}) = 0$ as functions of the scheme's parameters. Solving

$$\begin{aligned} \det(E - \lambda I_{4 \times 4}) &= 0 \\ &\equiv \frac{(\gamma\lambda - \gamma + \alpha\lambda\mu)(\lambda + \alpha\lambda - 1)(\gamma - 2\gamma\lambda + \gamma\lambda^2 + \alpha^2\lambda^2\mu - \alpha\gamma\lambda - \alpha\lambda\mu + L\alpha^2\lambda + \alpha\gamma\lambda^2 + \alpha\lambda^2\mu - \alpha^2\lambda\mu)}{(\alpha + 1)^2(\gamma + \alpha\mu)^2} = 0 \end{aligned} \quad (4)$$

leads to the following eigenvalues:

$$\begin{aligned} \lambda_1 &= \frac{\gamma}{\gamma + \alpha\mu} \\ \lambda_2 &= \frac{1}{1 + \alpha} \\ \lambda_3 &= \frac{2\gamma + \alpha\gamma + \alpha\mu - L\alpha^2 + \alpha^2\mu}{2(\gamma + \alpha\gamma + \alpha\mu + \alpha^2\mu)} + \\ &\quad \frac{\alpha\sqrt{L^2\alpha^2 - 2L\alpha^2\mu - 2L\alpha\mu - 2\gamma L\alpha - 4\gamma L + \alpha^2\mu^2 + 2\alpha\mu^2 + 2\gamma\alpha\mu + \mu^2 + 2\gamma\mu + \gamma^2}}{2(\gamma + \alpha\gamma + \alpha\mu + \alpha^2\mu)} \\ \lambda_4 &= \frac{2\gamma + \alpha\gamma + \alpha\mu - L\alpha^2 + \alpha^2\mu}{2(\gamma + \alpha\gamma + \alpha\mu + \alpha^2\mu)} - \\ &\quad \frac{\alpha\sqrt{L^2\alpha^2 - 2L\alpha^2\mu - 2L\alpha\mu - 2\gamma L\alpha - 4\gamma L + \alpha^2\mu^2 + 2\alpha\mu^2 + 2\gamma\alpha\mu + \mu^2 + 2\gamma\mu + \gamma^2}}{2(\gamma + \alpha\gamma + \alpha\mu + \alpha^2\mu)} \end{aligned} \quad (5)$$

Let us first mention some general behavior of these eigenvalues. Given γ and μ positive, we observe that:

1. λ_1 and λ_2 are positive decreasing functions w.r.t. α . Moreover, for bounded γ and μ , we have $\lim_{\alpha \rightarrow \infty} |\lambda_1(\alpha)| = 0 = \lim_{\alpha \rightarrow \infty} |\lambda_2(\alpha)|$.
2. One can show that for $\alpha \in [\frac{\mu + \gamma - 2\sqrt{\gamma L}}{L - \mu}, \frac{\mu + \gamma + 2\sqrt{\gamma L}}{L - \mu}]$, functions $\lambda_3(\alpha)$ and $\lambda_4(\alpha)$ are complex values and one can easily show that both share the same absolute value. Note that the lower bound of the interval $\frac{\mu + \gamma - 2\sqrt{\gamma L}}{L - \mu}$ is negative as soon as $\gamma \in [2L - \mu - 2\sqrt{L^2 - \mu L}, 2L - \mu + 2\sqrt{L^2 - \mu L}] \subseteq \mathbb{R}_+$. Moreover, one can easily show that $\lim_{\alpha \rightarrow \infty} |\lambda_3(\alpha)| = 0$ and $\lim_{\alpha \rightarrow \infty} |\lambda_4(\alpha)| = \frac{L - \mu}{\mu} = \kappa(A) - 1$. The latter limit shows that eigenvalue λ_4 plays a central role in the convergence of the NAG-GS method since it is the one that can reach the value one and violate the convergence condition, as soon as $\kappa(A) > 2$. The analysis of λ_4 also allows us to come up with a good candidate for the step size α that minimizes the spectral radius of matrix E , especially and obviously at critical point $\alpha_{max} = \frac{\mu + \gamma + 2\sqrt{\gamma L}}{L - \mu}$ which is positive since $L \geq \mu$ by hypothesis. Note that the case $L \rightarrow \mu$ gives some preliminary hints that the maximum step size can be almost "unbounded" in some particular cases.

Now, let us study these eigenvalues in more detail, it seems that three different scenarios must be studied:

1. For any variant of Algorithm 1 for which $\gamma_0 = \mu$, then $\gamma = \mu$ for all $k \geq 0$ and therefore $\lambda_1(\alpha) = \lambda_2(\alpha)$. Moreover, at $\alpha = \frac{\mu + \gamma + 2\sqrt{\gamma L}}{L - \mu} = \frac{2\mu + 2\sqrt{\mu L}}{L - \mu}$, we can easily check that $|\lambda_1(\alpha)| = |\lambda_2(\alpha)| = |\lambda_3(\alpha)| = |\lambda_4(\alpha)|$. Therefore $\alpha = \frac{2\mu + 2\sqrt{\mu L}}{L - \mu}$ is the step size ensuring the minimal spectral radius and hence the maximum contraction rate. Figure 1 shows the evolution of the absolute values of the eigenvalues of iteration matrix E w.r.t. α for such a setting.
2. As soon as $\gamma < \mu$, one can easily show that $\lambda_1(\alpha) < \lambda_2(\alpha)$. Therefore the step size α with the minimal spectral radius is such that $|\lambda_4(\alpha)| = |\lambda_2(\alpha)|$. One can show that the equality holds for $\alpha = \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu}$. One can easily check that $\frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu} - \frac{\mu + \gamma + 2\sqrt{\gamma L}}{L - \mu} = (\mu - \gamma)^2 > 0$. Hence the second candidate for step size α will be bigger than the first one and the distance between them increases as the squared distance between γ and μ . Figure 2 shows the evolution of the absolute values of the eigenvalues of iteration matrix E w.r.t. α for this setting.
3. For $\gamma > \mu$: the analysis of this case gives the same results as the previous point. According to Algorithm 1, γ is either constant and equal to μ or decreasing to μ along iterations. Hence, the case $\gamma > \mu$ will be considered for the theoretical analysis when $\gamma \neq \mu$.

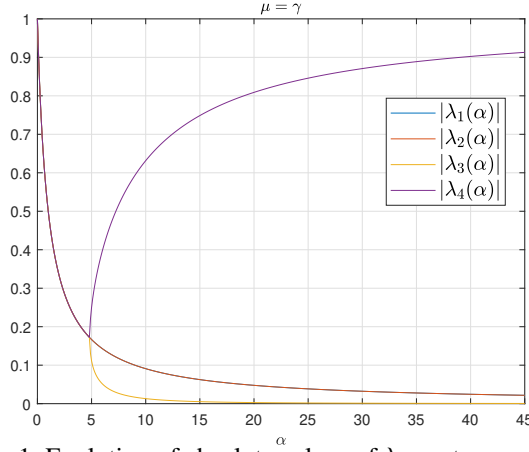


Fig. 1: Evolution of absolute values of λ_i w.r.t α ; $\mu = \gamma$.

As a first summary, the detailed analysis of the eigenvalues of iteration matrix E w.r.t. the significant parameters of the NAG-GS method leads us to come up with two candidates for the step size that minimize the spectral radius of E , hence ensuring the highest contraction rate possible. These results will be gathered with those obtained in Section 2.1.1 dedicated to the covariance matrix analysis.

Let us now look at the behavior of the dynamics in expectation; given the properties of the Brownian motion and by applying the Expectation operator \mathbb{E} on both sides of the system of SDE's (11), the resulting "averaged" equations identify with the "deterministic" setting studied by [12]. For such a setting, authors from [12]

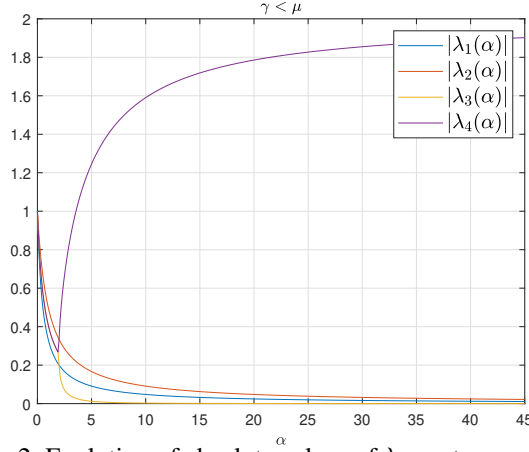


Fig. 2: Evolution of absolute values of λ_i w.r.t α ; $\gamma < \mu$.

demonstrated that, if $0 \leq \alpha \leq \frac{2}{\sqrt{\kappa(A)}}$, then a Gauss-Seidel splitting-based scheme for solving (10) is A -stable for quadratic objectives in the deterministic setting. We conclude this section by showing that the two candidates we derived above for step size are higher than the limit $\frac{2}{\sqrt{\kappa(A)}}$ given in [12, Theorem 1]. It can be intuitively understood in the case $L \rightarrow \mu$. However, we give a formal proof in Lemma 2.1.

Lemma 2.1 *Given $\gamma > 0$, and assuming $0 < \mu < L$, then for $\gamma = \mu$ and $\gamma > \mu$ the following inequalities respectively hold:*

$$\begin{aligned} \frac{2\mu + 2\sqrt{\mu L}}{L - \mu} &> \frac{2}{\sqrt{\kappa(A)}} \\ \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu} &> \frac{2}{\sqrt{\kappa(A)}} \end{aligned} \quad (6)$$

where $\kappa(A) = \frac{L}{\mu}$.

Proof Let us start for the case $\mu = \gamma$, hence first inequality from equation 6 becomes:

$$\begin{aligned} \frac{2\mu + 2\sqrt{L\mu}}{L - \mu} &> \frac{2}{\sqrt{L/\mu}} \\ &\equiv (\mu + \sqrt{L\mu})\sqrt{L/\mu} > (L - \mu) \\ &\equiv \sqrt{\mu L} + L > L - \mu \\ &\equiv \sqrt{\mu L} > -\mu \end{aligned}$$

which holds for any positive μ, L and satisfied by hypothesis. For the case $\gamma > \mu$, we have:

$$\begin{aligned} & \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu} > \frac{2}{\sqrt{L/\mu}} \\ \equiv & \sqrt{(\mu - \gamma)^2 + 4\gamma L} > \frac{2}{\sqrt{L/\mu}}(L - \mu) - \gamma - \mu \\ \equiv & (\mu - \gamma)^2 + 4\gamma L > (\mu + 2\sqrt{\frac{\mu}{L}}(\mu - L) + \gamma)^2 \\ \equiv & \gamma > \frac{-2\mu^2 + \mu^3/L + \mu^2\sqrt{\mu/L} + \mu L - \mu L\sqrt{\mu/L}}{-\mu - \sqrt{\mu/L}(\mu - L) + L} \end{aligned}$$

where second inequality hold since $L \geq \mu$ and last inequality holds since $-\mu - \sqrt{\mu/L}(\mu - L) + L > 0$ (one can easily check this by using $L > \mu$). It remains to show that:

$$\mu > \frac{-2\mu^2 + \mu^3/L + \mu^2\sqrt{\mu/L} + \mu L - \mu L\sqrt{\mu/L}}{-\mu - \sqrt{\mu/L}(\mu - L) + L}$$

which holds for any μ and L positive (technical details are skipped; it mainly consists of the study of a table of signs of a polynomial equation in μ).

Since $\gamma > \mu$ by hypothesis, therefore inequality

$$\gamma > \frac{-2\mu^2 + \mu^3/L + \mu^2\sqrt{\mu/L} + \mu L - \mu L\sqrt{\mu/L}}{-\mu - \sqrt{\mu/L}(\mu - L) + L}$$

holds for any μ and L positive as well, conditions satisfied by hypothesis. This concludes the proof.

Furthermore, let us note that both step size candidates, that are $\frac{2\mu+2\sqrt{\mu L}}{L-\mu}$ and $\frac{\mu+\gamma+\sqrt{(\mu-\gamma)^2+4\gamma L}}{L-\mu}$ respectively for the cases $\gamma = \mu$ and $\gamma > \mu$ show that NAG-GS method converges in the case $L \rightarrow \mu$ with a step size that tends to ∞ . This behavior cannot be anticipated by the upper-bound given by [12, Theorem 1]. Some simple numerical experiments are performed in Section 2.4 to support this theoretical result.

Finally, based on previous discussions, let us remark that for $\alpha \geq \frac{\mu+\gamma+\sqrt{(\mu-\gamma)^2+4\gamma L}}{L-\mu}$ when $\gamma \neq \mu$ or $\alpha \geq \frac{2\mu+2\sqrt{\mu L}}{L-\mu}$ when $\gamma = \mu$, we have $\rho(E(\alpha)) = |\lambda_4(\alpha)|$ and one can show that $\rho(E)$ is strictly monotonically increasing function of α for all $L > \mu > 0$ and $\gamma > 0$, see Lemma 2.2 for the formal proof.

Lemma 2.2 *Given $\gamma > 0$, and assuming $0 < \mu < L$, then for $\gamma = \mu$ and $\gamma > \mu$, the spectral radius $\rho(E(\alpha))$ is a strict monotonic increasing function of α for $\alpha \in [\alpha_c, \infty]$*

with $\alpha_c = \frac{2\mu+2\sqrt{\mu L}}{L-\mu}$ or $\alpha_c = \frac{\mu+\gamma+\sqrt{(\mu-\gamma)^2+4\gamma L}}{L-\mu}$.

Proof Let us first recall that on $[\alpha_c, \infty]$, the spectral radius $\rho(E(\alpha))$ is equal to $|\lambda_4|$, the expression of λ_4 as a function of parameters of interests for the convergence analysis of NAG-GS method was given in equation 5 and recalled here-under for convenience:

$$\lambda_4 = \frac{2\gamma + \alpha\gamma + \alpha\mu - L\alpha^2 + \alpha^2\mu}{2(\gamma + \alpha\gamma + \alpha\mu + \alpha^2\mu)} - \frac{\alpha\sqrt{L^2\alpha^2 - 2L\alpha^2\mu - 2L\alpha\mu - 2\gamma L\alpha - 4\gamma L + \alpha^2\mu^2 + 2\alpha\mu^2 + 2\gamma\alpha\mu + \mu^2 + 2\gamma\mu + \gamma^2}}{2(\gamma + \alpha\gamma + \alpha\mu + \alpha^2\mu)} \quad (7)$$

Let start by showing that λ_4 is negative on $[\alpha_c, \infty]$. Firstly, one can easily observe that the denominator of λ_4 is positive, secondly let us compute the values for α such that:

$$\begin{aligned} & 2\gamma + \alpha\gamma + \alpha\mu - L\alpha^2 + \alpha^2\mu - \\ & \alpha\sqrt{L^2\alpha^2 - 2L\alpha^2\mu - 2L\alpha\mu - 2\gamma L\alpha - 4\gamma L + \alpha^2\mu^2 + 2\alpha\mu^2 + 2\gamma\alpha\mu + \mu^2 + 2\gamma\mu + \gamma^2} = 0 \\ \equiv & -4\gamma^2 - 4\alpha\gamma(\mu + \gamma) + \alpha^2(\gamma^2 - 4\gamma L + 2\gamma\mu + \mu^2) - \alpha^2(\gamma^2 - 4\gamma L + 6\gamma\mu + \mu^2) = 0 \\ \equiv & (-4\gamma\mu)\alpha^2 - 4\gamma(\mu + \gamma)\alpha - 4\gamma^2 = 0 \end{aligned} \quad (8)$$

The expression above is negative as soon as $\alpha < -1$ or $\alpha > \frac{-\gamma}{\mu} < 0$ since $\gamma, \mu > 0$ by hypothesis. The latter is always satisfied since $\alpha \geq \alpha_c > 0$ by hypothesis. Therefore $\rho(E(\alpha)) = -\lambda_4$ for $\alpha \in [\alpha_c, \infty]$.

To show the monotonic increasing behavior of $\rho(E(\alpha))$ w.r.t. $\alpha \in [\alpha_c, \infty]$, it remains to show that:

$$\frac{d(\rho(E(\alpha)))}{d\alpha} = \frac{d(-\lambda_4)}{d\alpha} > 0. \quad (9)$$

To ease the analysis, let us decompose $-\lambda_4(\alpha) = t_1(\alpha) + t_2(\alpha)$ such that:

$$\begin{aligned} t_1(\alpha) &= -\frac{2\gamma + \alpha\gamma + \alpha\mu - L\alpha^2 + \alpha^2\mu}{2(\gamma + \alpha\gamma + \alpha\mu + \alpha^2\mu)} \\ t_2(\alpha) &= \frac{\alpha\sqrt{L^2\alpha^2 - 2L\alpha^2\mu - 2L\alpha\mu - 2\gamma L\alpha - 4\gamma L + \alpha^2\mu^2 + 2\alpha\mu^2 + 2\gamma\alpha\mu + \mu^2 + 2\gamma\mu + \gamma^2}}{2(\gamma + \alpha\gamma + \alpha\mu + \alpha^2\mu)} \end{aligned} \quad (10)$$

Let us now show that $\frac{dt_1(\alpha)}{d\alpha} > 0$ and $\frac{dt_2(\alpha)}{d\alpha} > 0$ for any $L > \mu > 0$. We first obtain:

$$\begin{aligned} \frac{dt_1(\alpha)}{d\alpha} &= \frac{(2\gamma + 2\mu + 4\alpha\mu)(2\gamma + \alpha\gamma + \alpha\mu - L\alpha^2 + \alpha^2\mu)}{(2\gamma + 2\alpha\gamma + 2\alpha\mu + 2\alpha^2\mu)^2} - \\ & \quad \frac{\gamma + \mu - 2L\alpha + 2\alpha\mu}{2\gamma + 2\alpha\gamma + 2\alpha\mu + 2\alpha^2\mu} \\ &= \frac{(L\alpha^2 + \gamma)(\gamma + \mu) + 2\alpha\gamma(L + \mu)}{2(\alpha + 1)^2(\gamma + \alpha\mu)^2} \end{aligned} \quad (11)$$

which is strictly positive since $L > \mu > 0$ and $\gamma > 0$ by hypothesis. Furthermore:

$$\begin{aligned} \frac{dt_2(\alpha)}{d\alpha} &= \\ & \frac{(\gamma + \mu)(L - \mu)(\alpha^3 L - 3\alpha\gamma) + \alpha^2(L(-\gamma^2 - \mu^2) + 2\gamma(L^2 - L\mu + \mu^2)) + \gamma(\gamma^2 - 2\gamma(2L - \mu) + \mu^2)}{2(\alpha + 1)^2(\alpha\mu + \gamma)^2\sqrt{\alpha^2(L^2 - 2L\mu + \mu^2) - 2\alpha(\gamma + \mu)(L - \mu) + \gamma^2 - 2\gamma(2L - \mu) + \mu^2}} \end{aligned} \quad (12)$$

The remaining demonstration is significantly long and technically heavy in the case $\gamma > \mu$. Then we limit the last part of the demonstration for the case $\mu = \gamma$ for

which we have shown previously than $\alpha_c = \frac{\mu + \gamma + 2\sqrt{\gamma L}}{L - \mu} = \frac{2\mu + 2\sqrt{\mu L}}{L - \mu}$. In practice, with respect to the NAG-GS method summarized by Algorithm 1, γ quickly decreases to μ and equality $\mu = \gamma$ holds for most of the iterations of the Algorithm. Hence, this case is more important to detail here. However, the reasoning explained herein ultimately leads to identical conclusions when considering the case where γ is greater than μ .

The first term of the numerator of (12) is positive as soon as $\alpha \geq \sqrt{\frac{3\gamma}{L}}$. In the case of $\mu = \gamma$, we determine the conditions under which the second term of the numerator of (12) is positive, that is:

$$\begin{aligned} \alpha^2(L(-2\mu^2) + 2\mu(L^2 - L\mu + \mu^2)) + \mu(2\mu^2 - 2\mu(2L - \mu)) &> 0 \\ \equiv \alpha^2(L(-2\mu^2) + 2\mu(L^2 - L\mu + \mu^2)) &> \mu(-2\mu^2 + 2\mu(2L - \mu)) \end{aligned} \quad (13)$$

First, one can see that:

$$\begin{aligned} (L(-2\mu^2) + 2\mu(L^2 - L\mu + \mu^2)) &> 0, \\ \mu(-2\mu^2 + 2\mu(2L - \mu)) &> 0 \end{aligned} \quad (14)$$

hold as soon as $L > \mu > 0$ which is satisfied by hypothesis. Therefore, the second term of the numerator of (12) is positive as soon as

$$\alpha > \sqrt{\frac{\mu(-2\mu^2 + 2\mu(2L - \mu))}{(L(-2\mu^2) + 2\mu(L^2 - L\mu + \mu^2))}} = \sqrt{\frac{2\mu}{L - \mu}} \quad (15)$$

which exists since $L > \mu > 0$ by hypothesis (the second root of equation 14 being negative). Finally, since $\alpha \in [\alpha_c, \infty]$ by hypothesis, $\frac{dt_2(\alpha)}{d\alpha}$ is positive as soon as:

$$\begin{aligned} \alpha_c &> \sqrt{\frac{3\mu}{L}} \\ \alpha_c &> \sqrt{\frac{2\mu}{L - \mu}} \end{aligned} \quad (16)$$

hold with $\alpha_c = \frac{2\mu + 2\sqrt{\mu L}}{L - \mu}$. One can easily show that both inequalities hold as soon as $L > \mu > 0$ which is satisfied by the hypothesis. This concludes the proof of the strict increasing monotonicity of $\rho(E(\alpha))$ w.r.t. α for $\alpha \in [\alpha_c, \infty]$ assuming $L > \mu > 0$ and $\gamma = \mu$.

2.1.1 Covariance analysis

In this section, we study the contribution to the computation of maximum step size for the NAG-GS method through the analysis of the covariance matrix at stationarity. Let us start by computing the covariance matrix C obtained at iteration $k + 1$ from Algorithm 1:

$$C_{k+1} = \mathbb{E}(y_{k+1}y_{k+1}^T) \quad (17)$$

By denoting $\xi_k = \begin{bmatrix} 0 \\ \sigma \frac{\sqrt{\alpha}}{1+\tau} \eta_k \end{bmatrix}$, let us replace y_{k+1} by its expression given in (3) and (17) writes:

$$\begin{aligned} C_{k+1} &= \mathbb{E}(y_{k+1} y_{k+1}^T) \\ &= \mathbb{E}((E y_k + \xi_k)(E y_k + \xi_k)^T) \\ &= \mathbb{E}(E y_k y_k^T E^T) + \mathbb{E}(\xi_k \xi_k^T) \end{aligned} \quad (18)$$

which holds since expectation operator $\mathbb{E}(\cdot)$ is a linear operator and by assuming statistical independence between ξ_k and $E y_k$. On the one hand, by using again the properties of linearity of \mathbb{E} and since E is seen as a constant by $\mathbb{E}(\cdot)$, one can show that $\mathbb{E}(E y_k y_k^T E^T) = E C_k E^T$. On the other hand, since $\eta_k \sim \mathcal{N}(0, 1)$, then (18) becomes:

$$C_{k+1} = E C_k E^T + Q \quad (19)$$

where $Q = \begin{bmatrix} 0_{2 \times 2} & 0_{2 \times 2} \\ 0_{2 \times 2} & \frac{\alpha_k \sigma^2}{(1+\tau_k)^2} I_{2 \times 2} \end{bmatrix}$. Let us now look at the limiting behavior of (19), that is $\lim_{k \rightarrow \infty} C_k$. Let be $C = \lim_{k \rightarrow \infty} C_k$ the covariance matrix reached in the asymptotic regime, also referred to as stationary regime. Applying the limit on both sides of Equation equation 19, C then satisfies

$$C = E C E^T + Q \quad (20)$$

Hence equation 20 is a particular case of discrete Lyapunov equation. For solving such equation, the vectorization operator denoted $\vec{\cdot}$ is applied on both sides on equation 20, this amounts to solve the following linear system:

$$(I_{4^2 \times 4^2} - E \otimes E) \vec{C} = \vec{Q} \quad (21)$$

where $A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}$ stands for the Kronecker product. The solution is given by:

$$C = \overleftarrow{(I_{4^2 \times 4^2} - E \otimes E)^{-1} \vec{Q}} \quad (22)$$

where $\overleftarrow{\cdot}$ stands for the un-vectorized operator.

Let us note that, even for the 2-dimensional case considered in this section, the dimension of matrix C rapidly growth and cannot be written in plain within this paper. For the following, we will keep its symbolic expression. The stationary matrix C quantifies the spreading of the limit of the sequence $\{y_k\}$, as a direct consequence of the Brownian motion effect. Now we look at the directions that maximize the scattering of the points, in other words, we are looking for the eigenvectors and the associated eigenvalues of C . Actually, the required information for the analysis of the step size is contained within the expression of the eigenvalues $\lambda_i(C)$. The obtained eigenvalues are rationale functions w.r.t. the parameters of the schemes, while their numerator

brings less interest for us (supported further), we will focus on their denominator. We obtained the following expressions:

$$\begin{aligned}\lambda_1(C) &= \frac{N_1(\alpha, \mu, L, \gamma, \sigma)}{D_1(\alpha, \mu, L, \gamma, \sigma)}, \\ \text{s.t. } D_1(\alpha, \mu, L, \gamma, \sigma) &= -L^2\alpha^3\mu - L^2\alpha^2\mu - \gamma L^2\alpha^2 + 2L\alpha^3\mu^2 + 4L\alpha^2\mu^2 + \\ &\quad 4\gamma L\alpha^2\mu + 2L\alpha\mu^2 + 8\gamma L\alpha\mu + 2\gamma^2 L\alpha + 4\gamma L\mu + 4\gamma^2 L\end{aligned} \quad (23)$$

$$\begin{aligned}\lambda_2(C) &= \frac{N_2(\alpha, \mu, L, \gamma, \sigma)}{D_2(\alpha, \mu, L, \gamma, \sigma)}, \\ \text{s.t. } D_2(\alpha, \mu, L, \gamma, \sigma) &= \alpha^3\mu^3 + 3\alpha^2\mu^3 + 3\gamma\alpha^2\mu^2 + 2\alpha\mu^3 + \\ &\quad 8\gamma\alpha\mu^2 + 2\gamma^2\alpha\mu + 4\gamma\mu^2 + 4\gamma^2\mu\end{aligned} \quad (24)$$

$$\begin{aligned}\lambda_3(C) &= \frac{N_3(\alpha, \mu, L, \gamma, \sigma)}{D_3(\alpha, \mu, L, \gamma, \sigma)}, \\ \text{s.t. } D_3(\alpha, \mu, L, \gamma, \sigma) &= \alpha^3\mu^3 + 3\alpha^2\mu^3 + 3\gamma\alpha^2\mu^2 + 2\alpha\mu^3 + \\ &\quad 8\gamma\alpha\mu^2 + 2\gamma^2\alpha\mu + 4\gamma\mu^2 + 4\gamma^2\mu\end{aligned} \quad (25)$$

$$\begin{aligned}\lambda_4(C) &= \frac{N_4(\alpha, \mu, L, \gamma, \sigma)}{D_4(\alpha, \mu, L, \gamma, \sigma)}, \\ \text{s.t. } D_4(\alpha, \mu, L, \gamma, \sigma) &= -L^2\alpha^3\mu - L^2\alpha^2\mu - \gamma L^2\alpha^2 + 2L\alpha^3\mu^2 + 4L\alpha^2\mu^2 + \\ &\quad 4\gamma L\alpha^2\mu + 2L\alpha\mu^2 + 8\gamma L\alpha\mu + 2\gamma^2 L\alpha + 4\gamma L\mu + 4\gamma^2 L\end{aligned} \quad (26)$$

One can observe that:

1. Given α, L, μ, γ positive, the denominators of eigenvalues λ_2 and λ_3 are positive as well, unlike eigenvalues λ_1 and λ_4 for which some vertical asymptotes may appear. The latter will be studied in more detail further. Note that, even if some eigenvalues share the same denominator, it is not the case for the numerator. This will be illustrated later in Figures 5 and 6 to ease the analysis.
2. Interestingly, the volatility of the noise defined by the parameter σ does not appear within the expressions of the denominators. It gives us a hint that these vertical asymptotes are due to the fact that spectral radius is getting close to 1 (discussed further in Section 2.2). Moreover, the parameter σ appears only within the numerators and based on intensive numerical tests, this parameter has a pure scaling effect onto the eigenvalues $\lambda_i(C)$ when studied w.r.t. α without modifying the trends of the curves.

Let us now study in more details the denominator of λ_1 and λ_4 and seek for critical step size as a function of γ, μ and L at which a vertical asymptote may appear by solving:

$$\begin{aligned}& -L^2\alpha^3\mu - L^2\alpha^2\mu - \gamma L^2\alpha^2 + 2L\alpha^3\mu^2 + 4L\alpha^2\mu^2 + \\ & 4\gamma L\alpha^2\mu + 2L\alpha\mu^2 + 8\gamma L\alpha\mu + 2\gamma^2 L\alpha + 4\gamma L\mu + 4\gamma^2 L = 0 \\ \equiv & \mu(2\mu - L)\alpha^3 + (\mu + \gamma)(4\mu - L)\alpha^2 + (2\mu^2 + 8\gamma\mu + 2\gamma^2)\alpha + 4\gamma(\mu + \gamma) = 0\end{aligned} \quad (27)$$

This polynomial equation in α has three roots:

$$\begin{aligned}\alpha_1 &= \frac{-\gamma - \mu}{\mu}, \\ \alpha_2 &= \frac{\mu + \gamma - \sqrt{\gamma^2 - 6\gamma\mu + \mu^2 + 4\gamma L}}{L - 2\mu}, \\ \alpha_3 &= \frac{\mu + \gamma + \sqrt{\gamma^2 - 6\gamma\mu + \mu^2 + 4\gamma L}}{L - 2\mu}.\end{aligned}\tag{28}$$

First, it is obvious that the first root α_1 is negative given γ, μ assumed nonnegative and therefore can be disregarded. Concerning α_2 and α_3 , those are real roots as soon as:

$$\begin{aligned}\gamma^2 - 6\gamma\mu + \mu^2 + 4\gamma L &\geq 0 \\ \equiv (\gamma - \mu)^2 - 4\gamma\mu + 4\gamma L &\geq 0 \\ \equiv (\gamma - \mu)^2 &\geq 4\gamma(\mu - L)\end{aligned}\tag{29}$$

which is always satisfied since $\gamma > 0$ and $0 < \mu < L$ by hypothesis.

Further, it is obvious that the study must include three scenarios:

1. Scenario 1: $L - 2\mu < 0$, or equivalently $\mu > L/2$. Given μ and γ positive by hypothesis, it implies that α_3 is negative and hence can be disregarded. It remains to check if α_2 can be positive, it amounts to verifying if

$$\begin{aligned}\mu + \gamma - \sqrt{\gamma^2 - 6\gamma\mu + \mu^2 + 4\gamma L} &< 0 \\ \equiv (\mu + \gamma)^2 &< \gamma^2 - 6\gamma\mu + \mu^2 + 4\gamma L \\ \equiv \mu &< \frac{L}{2}\end{aligned}$$

which never holds by hypothesis. Therefore, for the first scenario, there is no positive critical step size at which a vertical asymptote for the eigenvalues may appear.

2. Scenario 2: $L - 2\mu > 0$, or equivalently $\mu < L/2$. Obviously, α_3 is positive and hence shall be considered for the analysis of maximum step size for our NAG-GS method. It remains to check if α_2 is positive, that is to verify if the numerator can be negative. We have seen in the first scenario that α_2 is negative as soon as $\mu < \frac{L}{2}$ which is verified by hypothesis. Therefore, only α_3 is positive.
3. Scenario 3: $L - 2\mu = 0$. For such a situation, the critical step size is located at ∞ and can be disregarded as a potential limitation in our study.

In summary, a potentially critical and limiting step size only exists in the case $\mu < L/2$, or equivalently if $\kappa(A) > 2$. In this setting, the critical step size is positive and is equal to $\alpha_{\text{crit}} = \frac{\mu + \gamma + \sqrt{\gamma^2 - 6\gamma\mu + \mu^2 + 4\gamma L}}{L - 2\mu}$. Figures 3 to 4 display the evolution of the eigenvalues $\lambda_i(C)$ for $1 \leq i \leq 4$ w.r.t. to α for the two first scenarios, that are for $\mu > L/2$ and $\mu < L/2$. For the first scenario, the parameters σ, γ, μ and L have been respectively set to $\{1, 3/2, 1, 3/2\}$. For the second scenario, σ, γ, μ and L have been respectively set to $\{1, 3/2, 1, 3\}$. As expected, one can observe in Figure 3 that no vertical asymptote is present. Furthermore, one can observe $\lambda_i(C)$ seem to converge

to some limit point when $\alpha \rightarrow \infty$, numerically we report that this limit point is zero, for all the values of γ and σ considered.

Finally, again as expected by the results presented in this section, Figure 4 shows the presence of two vertical asymptotes for the eigenvalues λ_1 and λ_4 , and none for λ_2 and λ_3 . Moreover, the critical step size is approximately located at $\alpha = 6$ which aligns with analytical expression $\alpha_{\text{crit}} = \frac{\mu + \gamma + \sqrt{\gamma^2 - 6\gamma\mu + \mu^2 + 4\gamma L}}{L - 2\mu}$. Finally, one can observe that, after the vertical asymptotes, all the eigenvalues converge to some limit points, again numerically we report that this limit point is zero, for all the values of γ and σ considered.

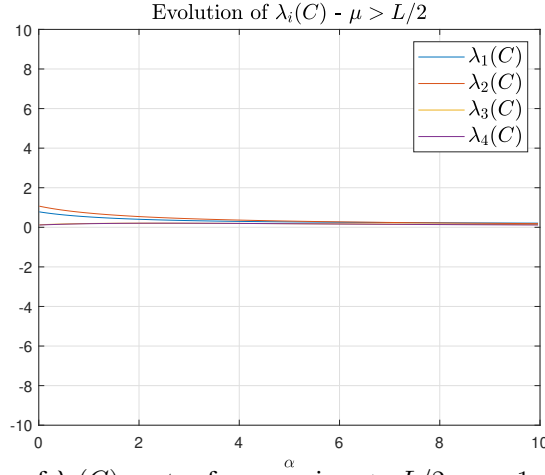


Fig. 3: Evolution of $\lambda_i(C)$ w.r.t α for scenario $\mu > L/2$; $\sigma = 1$, $\gamma = 3/2$, $\mu = 1$, $L = 3/2$.

2.2 A conclusion for the 2-dimensional case

In Section 2.1 and Section 2.1.1, several theoretical results have been derived for coming up with appropriate choices of constant step size for Algorithm 1. Key insights and interesting values for the step size have been discussed from the study of the spectral radius of iteration matrix E and through the analysis of the covariance matrix in the asymptotic regime. Let us summarize the theoretical results obtained:

- from the spectral radius analysis of iteration matrix E ; two scenarios have been highlighted, that are:
 1. $\gamma = \mu$: the step size α that minimizes the spectral radius of matrix E is $\alpha = \frac{2\mu + 2\sqrt{\mu L}}{L - \mu}$,
 2. $\gamma > \mu$: the step size α that minimizes the spectral radius of matrix E is $\alpha = \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu}$.

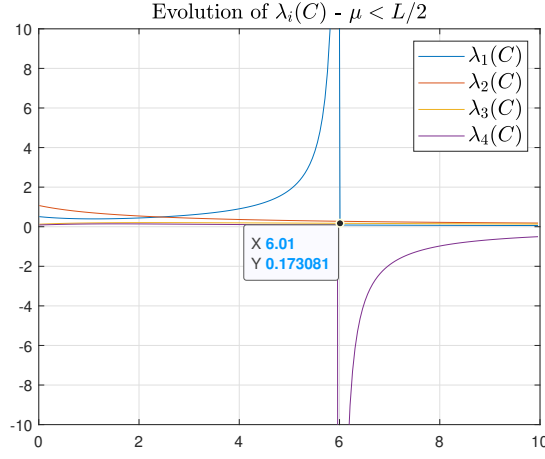


Fig. 4: Evolution of $\lambda_i(C)$ w.r.t α for scenario $\mu < L/2$; $\sigma = 1$, $\gamma = 3/2$, $\mu = 1$, $L = 3$.

- from the analysis of covariance matrix C at stationarity: in the case $L - 2\mu > 0$, or equivalently $\mu < L/2$, we have seen that there is a vertical asymptote for two eigenvalues of C at $\alpha_{\text{crit}} = \frac{\mu + \gamma + \sqrt{\gamma^2 - 6\gamma\mu + \mu^2 + 4\gamma L}}{L - 2\mu}$, leading to an intractable scattering of the limit points $\{y_k\}_{k \rightarrow \infty}$ generated by Algorithm 1. In the case $\mu > L/2$, there is no positive critical step size at which a vertical asymptote for the eigenvalues may appear.

Therefore, for quadratic functions such that $\mu > L/2$, we can safely choose either $\alpha = \frac{2\mu + 2\sqrt{\mu L}}{L - \mu}$ when $\gamma = \mu$ either $\alpha = \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu}$ when $\gamma > \mu$ to get the minimal spectral radius for iteration matrix E and hence the highest contraction rate for the NAG-GS method.

For quadratic functions such that $\mu < L/2$, we must show that the NAG-GS method is stable for both step sizes. Let us denote by $\alpha_c = \left\{ \frac{2\mu + 2\sqrt{\mu L}}{L - \mu}, \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu} \right\}$, two values of step size for the two scenarios $\gamma = \mu$ and $\gamma > \mu$. In Lemma 2.3, we show that NAG-GS is asymptotically convergent, or stable, for the 2-dimensional case under mild assumptions in the case $\mu < L/2$.

Lemma 2.3 *Given $\gamma > 0$, and assuming $0 < \mu < L/2$, then for $\gamma = \mu$ and $\gamma > \mu$ the following inequalities respectively hold:*

$$\begin{aligned} \frac{\mu + \gamma + \sqrt{\gamma^2 - 6\gamma\mu + \mu^2 + 4\gamma L}}{L - 2\mu} &> \frac{2\mu + 2\sqrt{\mu L}}{L - \mu} \\ \frac{\mu + \gamma + \sqrt{\gamma^2 - 6\gamma\mu + \mu^2 + 4\gamma L}}{L - 2\mu} &> \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu} \end{aligned} \quad (30)$$

Thus, in the 2-dimensional case, NAG-GS is asymptotically convergent (or stable) when choosing $\alpha_c = \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu}$ or $\alpha_c = \frac{2\mu + 2\sqrt{\mu L}}{L - \mu}$ respectively for the cases $\gamma > \mu$ and $\gamma = \mu$.

Proof In order to prove the asymptotic stability or convergence of NAG-GS for the 2-dimensional case within the set of assumptions detailed above, one must show that $\rho(E(\alpha_c)) < 1$ for the two choices of α_c .

Let us start by computing α such that $\rho(E(\alpha)) = 1$. As proved in Lemma 2.2, for $\alpha \in [\alpha_c, \infty]$, $\rho(E(\alpha)) = -\lambda_4$ with λ_4 given in (5), we then have to compute α such that:

$$-\lambda_4 = -\frac{2\gamma + \alpha\gamma + \alpha\mu - L\alpha^2 + \alpha^2\mu}{2(\gamma + \alpha\gamma + \alpha\mu + \alpha^2\mu)} + \frac{\alpha(L^2\alpha^2 - 2L\alpha^2\mu - 2L\alpha\mu - 2\gamma L\alpha - 4\gamma L + \alpha^2\mu^2 + 2\alpha\mu^2 + 2\gamma\alpha\mu + \mu^2 + 2\gamma\mu + \gamma^2)^{1/2}}{2(\gamma + \alpha\gamma + \alpha\mu + \alpha^2\mu)} = 1.$$

This leads to computing the roots of a quadratic polynomial equation in α , the positive root is:

$$\alpha = \frac{\gamma + \mu + \sqrt{4L\gamma + \gamma^2 - 6\gamma\mu + \mu^2}}{L - 2\mu}, \quad (31)$$

which not surprisingly identifies to α_{crit} from the covariance matrix analysis ¹.

Furthermore, as per Lemma 2.2, $\rho(E(\alpha))$ is strictly monotonically increasing function over the interval $[\alpha_c, \infty]$. Therefore, showing that $\rho(E(\alpha_c)) < 1$ is equivalent to show that α_c is strictly lower than $\alpha_{\text{crit}} := \frac{\gamma + \mu + \sqrt{4L\gamma + \gamma^2 - 6\gamma\mu + \mu^2}}{L - 2\mu}$.

Let us focus on the case $\gamma > \mu$; since $0 < \mu < L/2$ by hypothesis, the second inequality from equation 30 can be written as:

$$(L - \mu)(\gamma + \mu + \sqrt{(\gamma - \mu)^2 + 4\gamma(L - \mu)}) - (L - 2\mu)(\gamma + \mu + \sqrt{(\gamma - \mu)^2 + 4\gamma L}) > 0 \\ \equiv \gamma\mu + \mu^2 + (L - \mu)\sqrt{\gamma^2 + \mu^2 + \gamma(4L - 6\mu)} + (2\mu - L)\sqrt{(\gamma - \mu)^2 + 4\gamma L} > 0$$

Given $\gamma, \mu > 0$, it remains to show that:

$$(L - \mu)\sqrt{\gamma^2 + \mu^2 + \gamma(4L - 6\mu)} + (2\mu - L)\sqrt{(\gamma - \mu)^2 + 4\gamma L} > 0 \quad (32)$$

In order to show this, we study the conditions for γ such that the left-hand side of (32) is positive. With simple manipulations, one can show that canceling the left-hand side of (32) boils down to canceling the following quadratic polynomial:

$$(L - \mu)\sqrt{\gamma^2 + \mu^2 + \gamma(4L - 6\mu)} + (2\mu - L)\sqrt{(\gamma - \mu)^2 + 4\gamma L} = 0 \\ \equiv (-3\mu + 2L)\gamma^2 + (2\mu^2 - 8L\mu + 4L^2)\gamma + 2L\mu^2 - 3\mu^3 = 0$$

The two roots are:

$$\gamma_1 = \frac{-\mu^2 - 2L^2 - 2\sqrt{-2\mu^4 + L^4 - 4\mu L^3 + 4\mu^2 L^2 + \mu^3 L + 4\mu L}}{2L - 3\mu} \\ \gamma_2 = \frac{-\mu^2 - 2L^2 + 2\sqrt{-2\mu^4 + L^4 - 4\mu L^3 + 4\mu^2 L^2 + \mu^3 L + 4\mu L}}{2L - 3\mu},$$

¹ It explains why the critical α does not include σ , this singularity is due to the spectral radius reaching the value 1.

which are real and distinct as soon as:

$$\begin{aligned} & -2\mu^4 + L^4 - 4\mu L^3 + 4\mu^2 L^2 + \mu^3 L > 0 \\ & \equiv (L - 2\mu)(L - \mu)(-\mu^2 + L^2 - \mu L) > 0, \end{aligned}$$

which holds since $0 < \mu < L/2$ by hypothesis (one can easily show that $-\mu^2 + L^2 - \mu L$ is positive in such setting). Moreover, the denominator $2L - 3\mu$ is strictly positive since $0 < \mu < L/2$. One can check that γ_1 is negative for all $\gamma, L > 0$ and $0 < \mu < L/2$ (simply show that $-\mu^2 - 2L^2 + 4\mu L$ is negative) and can be disregarded since γ is positive by hypothesis. Therefore, proving that equation 32 holds is equivalent to show that:

$$\gamma > \frac{-\mu^2 - 2L^2 + 2\sqrt{(L - 2\mu)(L - \mu)(-\mu^2 + L^2 - \mu L)} + 4\mu L}{2L - 3\mu} \quad (33)$$

To achieve this, let us first show that

$$\begin{aligned} \mu & > \frac{-\mu^2 - 2L^2 + 2\sqrt{(L - 2\mu)(L - \mu)(-\mu^2 + L^2 - \mu L)} + 4\mu L}{2L - 3\mu} \\ & \equiv 0 > \mu^2 + \sqrt{(L - 2\mu)(L - \mu)(-\mu^2 + L^2 - \mu L)} - L^2 + \mu L \\ & \equiv -\mu^2 + L^2 - \mu L > (L - 2\mu)(L - \mu) \\ & \equiv \mu < \frac{2}{3}L, \end{aligned}$$

which holds by hypothesis. Since $\gamma > \mu$ by hypothesis, inequality equation 33 holds for any μ and L positive as well, conditions satisfied by hypothesis.

Finally, since $\frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu} > \frac{\mu + \gamma + 2\sqrt{\gamma L}}{L - \mu}$ for any $\gamma, \mu, L > 0$, then first inequality in equation 30 holds as well. This concludes the proof.

We conclude this section by discussing several important insights:

- Except for α_{crit} , we do not report significant information coming from the analysis of $\lambda_i(C)$ for the computation of the step size and the validity of the candidates for α that are from

$$\left\{ \frac{2\mu + 2\sqrt{\mu L}}{L - \mu}, \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu} \right\},$$

respectively for the cases $\gamma = \mu$ and $\gamma > \mu$.

- Concerning the effect of the volatility σ of the noise, we have mentioned earlier that the parameter σ appears only within the numerators $\lambda_i(C)$ and based on intensive numerical tests, this parameter has a pure scaling effect onto the eigenvalues $\lambda_i(C)$ when studied w.r.t. α without modifying the trends of the curves. For compliance purpose, Figures 5 and 6 respectively show the evolution of the numerators $N_i(\alpha, \mu, L, \gamma, \sigma)$ of eigenvalues expressions of C given in Equations equation 23 to equation 26 w.r.t. σ , for both scenarios $\mu < L/2$ and $\mu > L/2$. One can observe monotonic polynomial increasing behavior of $N_i(\alpha, \mu, L, \gamma, \sigma)$ w.r.t σ for all $1 \leq i \leq 4$.

- The theoretical analysis summarized in this section is valid for the 2-dimensional case, we show in Section 2.3 how to generalize our results for the n -dimensional case. This has no impact on our results.

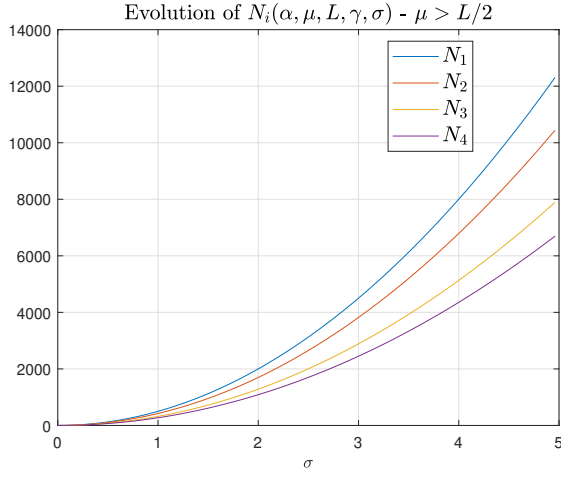


Fig. 5: Evolution of $N_i(\alpha, \mu, L, \gamma, \sigma)$ w.r.t σ for scenario $\mu > L/2$; $\gamma = 3/2$, $\mu = 1$, $L = 3/2$, $\alpha = \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu}$.

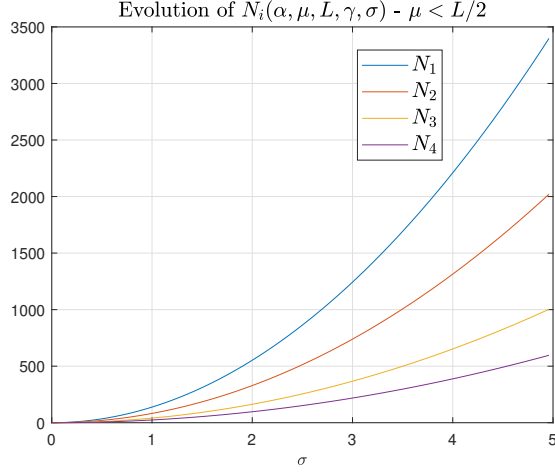


Fig. 6: Evolution of $N_i(\alpha, \mu, L, \gamma, \sigma)$ w.r.t σ for scenario $\mu < L/2$; $\gamma = 3/2$, $\mu = 1$, $L = 3$, $\alpha = \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu}$.

2.3 Extension to n -dimensional case

In this section, we show that we can easily extend the results obtained for the 2-dimensional case in Section 2.1, Section 2.1.1 and Section 2.2 to the n -dimensional case with $n > 2$. Let us start by recalling that for NAG transformation (7), the general SDE's system to solve for the quadratic case is:

$$\dot{y}(t) = \begin{bmatrix} -I_{n \times n} & I_{n \times n} \\ 1/\gamma(\mu I_{n \times n} - A) & -\mu/\gamma I_{n \times n} \end{bmatrix} y(t) + \begin{bmatrix} 0_{n \times 1} \\ \frac{dZ}{dt} \end{bmatrix}, \quad t > 0. \quad (34)$$

Let recall that $y = (x, v)$ with $x, v \in \mathbb{R}^n$, let n be even and let consider the permutation matrix P associated to permutation indicator π given here-under in two-line form:

$$\pi = \left[\begin{array}{cccc|cccc} (1 & 2) & (3 & 4) & \cdots & (n-1 & n) & (n+1 & n+2) & \cdots & (2n-1 & 2n) \\ (2 * 1 - 1 & 2 * 1) & (2 * 3 - 1 & 2 * 3) & \cdots & (2n-3 & 2n-2) & (3 & 4) & \cdots & (2n-1 & 2n) \end{array} \right]$$

where the bottom second-half part of π corresponds to the complementary of the bottom first half w.r.t. to the set $\{1, 2, \dots, 2n\}$ in the increasing order. For avoiding ambiguities, the ones element of P are at indices $(\pi(1, j), \pi(2, j))$ for $1 \leq j \leq 2n$. For such convention and since permutation matrix P associated to indicator π is orthogonal matrix, equation 34 can be equivalently written as follows:

$$\begin{aligned} \dot{y}(t) &= PP^T \begin{bmatrix} -I_{n \times n} & I_{n \times n} \\ 1/\gamma(\mu I_{n \times n} - A) & -\mu/\gamma I_{n \times n} \end{bmatrix} PP^T y(t) + \begin{bmatrix} 0_{n \times 1} \\ \dot{Z} \end{bmatrix}, \\ \equiv P^T \dot{y}(t) &= P^T \begin{bmatrix} -I_{n \times n} & I_{n \times n} \\ 1/\gamma(\mu I_{n \times n} - A) & -\mu/\gamma I_{n \times n} \end{bmatrix} PP^T y(t) + P^T \begin{bmatrix} 0_{n \times 1} \\ \dot{Z} \end{bmatrix}, \end{aligned} \quad (35)$$

Since we assumed w.l.o.g. $A = \text{diag}(\lambda_1, \dots, \lambda_n)$ with $\mu = \lambda_1 \leq \dots \leq \lambda_n = L$, one can easily see that Equation equation 35 has the structure:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{v}_1 \\ \dot{v}_2 \\ \vdots \\ \dot{x}_{2i-1} \\ \dot{x}_{2i} \\ \dot{v}_{2i-1} \\ \dot{v}_{2i} \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \\ \dot{v}_{n-1} \\ \dot{v}_n \end{bmatrix} = \begin{bmatrix} I_2 & -I_2 & 0 & 0 \\ 1/\gamma(\mu I_2 - A_1) & -\mu/\gamma I_2 & 0 & 0 \\ 0 & 0 & I_2 & -I_2 \\ 0 & 0 & 1/\gamma(\mu I_2 - A_i) & -\mu/\gamma I_2 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_2 \\ 0 & 0 & 0 & -I_2 \\ 1/\gamma(\mu I_2 - A_m) & -\mu/\gamma I_2 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ v_1 \\ v_2 \\ \vdots \\ x_{2i-1} \\ x_{2i} \\ v_{2i-1} \\ v_{2i} \\ \vdots \\ x_{n-1} \\ x_n \\ v_{n-1} \\ v_n \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \dot{Z}_1 \\ \dot{Z}_2 \\ \vdots \\ 0 \\ 0 \\ \dot{Z}_{2i-1} \\ \dot{Z}_{2i} \\ \vdots \\ 0 \\ 0 \\ \dot{Z}_{n-1} \\ \dot{Z}_n \end{bmatrix} \quad (36)$$

which boils down to $m = \frac{n}{2}$ independent 2-dimensional SDE's systems where $A_i = \text{diag}(\lambda_{2i-1}, \lambda_{2i})$ with $1 \leq i \leq m$ such that $\lambda_1 = \mu$ and $\lambda_n = L$.

Therefore, the m SDE's systems can be studied and theoretically solved independently with the schemes and the associated step sizes presented in previous sections. However, in practice, we will use a unique and general step size α to tackle the full SDE's system (34).

Let now use the "decoupled" structure given in (36) to come up with a general step size that will ensure the convergence of each system and hence the convergence of the full original system given in (34). Let us denote by α_i the step size for the i -th SDE's system with $1 \leq i \leq m = n/2$ minimizing the spectral radius of the system at hand. For convenience, let us consider the case $\gamma > \mu$, we apply the same method as detailed in Section 2.1 and Section 2.1.1 to compute the expression of α_i that minimizes $\rho(E_i(\alpha))$, we obtain:

$$\alpha_i = \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma\lambda_{2i}}}{\lambda_{2i} - \mu} \quad (37)$$

Finally, in Theorem 1, we show that choosing $\alpha_c := \alpha = \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu}$ ensures the convergence of NAG-GS method used to solve the SDE's system 34 in the n -dimensional case for $n > 2$. Theorem 1 is enunciated in Section 2.3 in the main text and the proof is given here-under.

Proof First, we recall that Lemma 2.3 in Section 2.2 provides the proof for the asymptotic convergence of NAG-GS method for $n = 2$ when choosing $\alpha := \alpha_c = \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu}$ for the case $\gamma > \mu$. In particular, it is shown that the spectral radius of the iteration matrix $\rho(E(\alpha_c))$ is strictly lower than 1 under consistent assumptions with the ones of Theorem 1 (see Lemma 2.3 for more details). The following steps of the proof show that choosing α_c also leads to the asymptotic convergence of the NAG-GS method for $n > 2$.

To do so, let us start by considering, w.l.o.g., the SDE's system in the form given by equation 36 and let $\alpha_i = \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma\lambda_{2i}}}{\lambda_{2i} - \mu}$ be the step size (given in Equation (37)) selected for solving the i -th SDE's system with $1 \leq i \leq m = n/2$, minimizing $\rho(E_i(\alpha))$, that is the spectral radius of the associated iteration matrix E_i . The result of Lemma 2.3 can be directly extended for each independent 2-dimensional SDE's system, in particular showing that $\rho(E_i(\alpha_i)) < 1$ for $1 \leq i \leq m = n/2$.

Therefore, to prove the convergence of the NAG-GS method by choosing a single step size α such that $0 < \alpha \leq \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu}$, it suffices to show that:

$$\alpha = \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu} \leq \min_{1 \leq i \leq m = n/2} \alpha_i \quad (38)$$

For proving that equation 38 holds, it sufficient to show that for any λ such that $0 < \mu \leq \lambda \leq L < \infty$ we have:

$$\frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu} \leq \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma \lambda}}{\lambda - \mu}. \quad (39)$$

which is equivalent to showing:

$$\begin{aligned} & \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu} - \frac{\mu + \gamma + \sqrt{(\mu - \gamma)^2 + 4\gamma \lambda}}{\lambda - \mu} \leq 0 \\ \equiv & \gamma \left(\frac{1}{L - \mu} - \frac{1}{\lambda - \mu} \right) + \mu \left(\frac{1}{L - \mu} - \frac{1}{\lambda - \mu} \right) + \frac{\sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu} - \frac{\sqrt{(\mu - \gamma)^2 + 4\gamma \lambda}}{\lambda - \mu} \leq 0 \end{aligned} \quad (40)$$

Since $0 < \mu \leq \lambda \leq L < \infty$ by hypothesis, one can easily show that first two terms of the last inequality are negative. It remains to show that:

$$\begin{aligned} & \frac{\sqrt{(\mu - \gamma)^2 + 4\gamma L}}{L - \mu} - \frac{\sqrt{(\mu - \gamma)^2 + 4\gamma \lambda}}{\lambda - \mu} \leq 0 \\ \equiv & (-\gamma^2 - 4\gamma\lambda + 2\gamma\mu - \mu^2)L^2 + (4\gamma\lambda^2 + 2\gamma^2\mu + 2\mu^3)L + \\ & \gamma^2\lambda^2 - 2\gamma^2\lambda\mu - 2\gamma\lambda^2\mu + \lambda^2\mu^2 - 2\lambda\mu^3 \leq 0 \end{aligned} \quad (41)$$

Note that we can easily show that the coefficient of L^2 is negative, hence last inequality is satisfied as soon as $L \leq \frac{-\gamma^2\lambda + 2\gamma^2\mu + 2\gamma\lambda\mu - \lambda\mu^2 + 2\mu^3}{\gamma^2 + 4\gamma\lambda - 2\gamma\mu + \mu^2}$ or $L \geq \lambda$. The latter condition is satisfied by hypothesis, this concludes the proof. Note that one can check that $\frac{-\gamma^2\lambda + 2\gamma^2\mu + 2\gamma\lambda\mu - \lambda\mu^2 + 2\mu^3}{\gamma^2 + 4\gamma\lambda - 2\gamma\mu + \mu^2} \leq \lambda$.

The theoretical results derived in these sections, along with the key insights, are validated in Section 2.4 through numerical experiments conducted for the NAG-GS method in the quadratic case.

2.4 Numerical tests for quadratic case

In this section, we report some simple numerical tests for the NAG-GS method (Algorithm 1) used to tackle the accelerated SDE's system given in (11) where:

- the objective function is $f(x) = (x - ce)^T A(x - ce)$ with $A \in \mathbb{S}_+^3$, e a all-ones vector of dimension 3 and c a positive scalar. For such a strongly convex setting, since the feasible set is $V = \mathbb{R}^3$, the minimizer x^* exists and is simply equal to ce . The matrix A is generated as follows: $A = QDQ^{-1}$ where matrix D is a diagonal matrix of size three and Q is a random orthogonal matrix. This test procedure allows us to specify the minimum and maximum eigenvalues of A that are respectively μ and L and hence it allows us to consider the two scenarios discussed in Section 2.1, that are $\mu > L/2$ and $\mu < L/2$.
- The noise volatility σ is set to 1, we report that this corresponds to a significant level of noise.
- Initial parameter γ_0 is set to μ .
- Different values for the step size α will be considered in order to empirically demonstrate the optimal choice α_c in terms of contraction rate, but also validate the critical values for step size in the case $\mu < L/2$ and, finally, highlight the effect of the step size in terms of scattering of the final iterates generated by NAG-GS around the minimizer of f .

From a practical point of view, we consider $m = 200000$ points. For each of them, the NAG-GS method is run for a maximum number of iterations to reach the stationarity, and the initial state x_0 is generated using normal Gaussian distribution. Since $f(x)$ is a quadratic function, the points are expected to converge to some Gaussian distribution around the minimizer $x^* = ce$. Furthermore, since the initial distribution is also Gaussian, the intermediate distributions (at each iteration of the NAG-GS method) are expected to be Gaussian as well. Therefore, to quantify the rate of convergence of the

NAG-GS method for different values of step size, we will monitor $\|\bar{x}^k - x^*\|$, that is the distance between the empirical mean of the distribution at iteration k and the minimizer x^* of f .

Figures 7 and 8 respectively show the evolution of $\|\bar{x}^k - x^*\|$ along iteration and the final distribution of points obtained by NAG-GS at stationarity for the scenario $\mu > L/2$, for the latter the points are projected onto the three planes to have a full visualization. As expected by the theory presented in Section 2.2, there is no critical α , hence one may choose arbitrary large values for step size while the NAG-GS method still converges. Moreover, the choice of $\alpha = \alpha_c$ gives the highest rate of convergence. Finally, one can observe that the distribution of limit points tightens more and more around the minimizer x^* of f as the chosen step increases, as expected by the analysis of Figure 3. Hence, one may choose a very large step size α so that the limit points converge to x^* almost surely but at the cost of a (much) slower convergence rate. Here comes the tradeoff between the convergence rate and the limit points scattering.

Finally, Figures 9 and 10 provide similar results for the scenario $\mu < L/2$. The theory outlined in Section 2.2 and Section 2.3 predicts a critical value of α that indicates when the convergence of NAG-GS is destroyed in such a scenario. In order to illustrate this gradually, different values of α have been chosen within the set $\{\alpha_c, \alpha_c/2, (\alpha_c + \alpha_{\text{crit}})/2, 0.98\alpha_{\text{crit}}\}$. First, one can observe that the choice of $\alpha = \alpha_c$ gives again the highest rate of convergence; see Figure 9. Moreover, one can see that for $\alpha \rightarrow \alpha_{\text{crit}}$, the convergence starts to fail, and the spreading of the limit points tends to infinity. We report that for $\alpha = \alpha_{\text{crit}}$, NAG-GS method diverges. Again, the theory derived in previous sections fully predicted these numerical results.

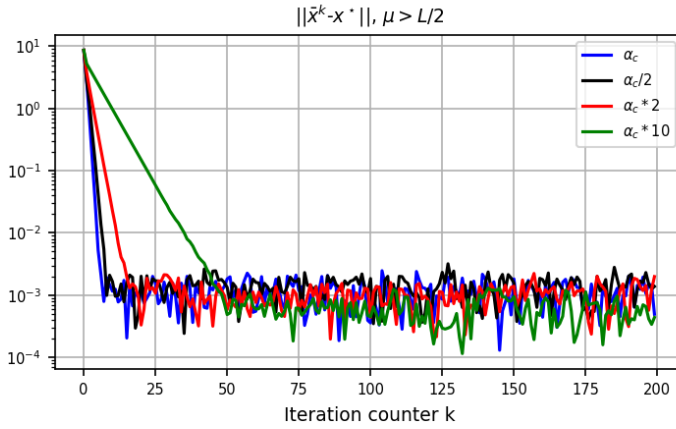


Fig. 7: Evolution of $\|\bar{x}^k - x^*\|$ along iteration for the scenario $\mu > L/2$; $c = 5$, $\gamma = \mu = 1$, $L = 1.9$ and $\sigma = 1$ for $\alpha \in \{\alpha_c, \alpha_c/2, 2\alpha_c, 10\alpha_c\}$ with $\alpha_c = \frac{2\mu + 2\sqrt{\mu L}}{L - \mu} = 5.29$.

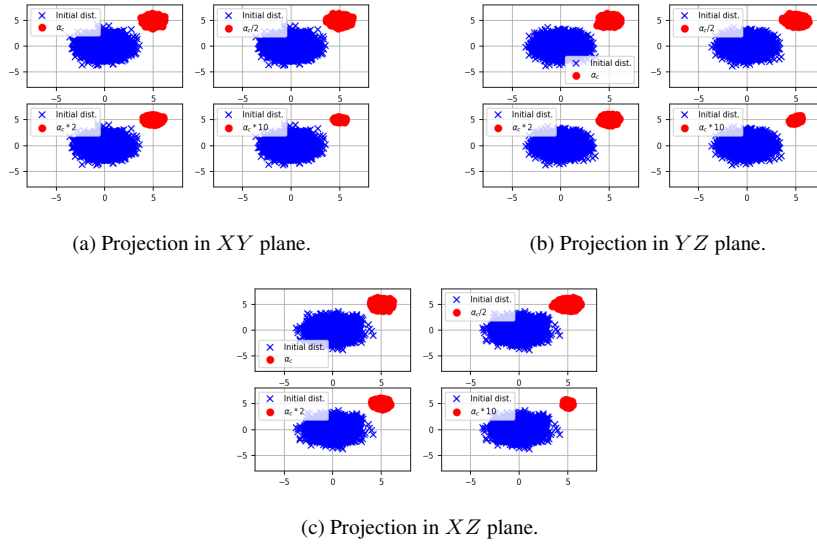


Fig. 8: Initial (blue crosses) and final (red circles) distributions of points generated by the NAG-GS method for the scenario $\mu > L/2$; $c = 5$, $\gamma = \mu = 1$, $L = 1.9$ and $\sigma = 1$ for $\alpha \in \{\alpha_c, \alpha_c/2, 2\alpha_c, 10\alpha_c\}$ with $\alpha_c = \frac{2\mu+2\sqrt{\mu L}}{L-\mu} = 5.29$.

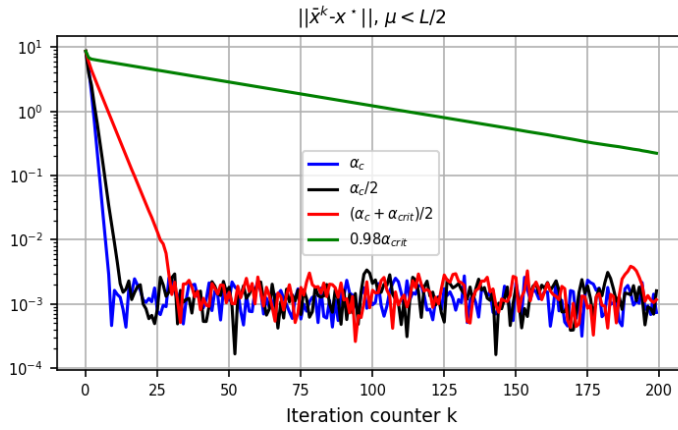


Fig. 9: Evolution of $\|\bar{x}^k - x^*\|$ along iteration for the scenario $\mu < L/2$; $c = 5$, $\gamma = \mu = 1$, $L = 3$ and $\sigma = 1$ for $\alpha \in \{\alpha_c, \alpha_c/2, (\alpha_c + \alpha_{crit})/2, 0.98\alpha_{crit}\}$ with $\alpha_c = \frac{2\mu+2\sqrt{\mu L}}{L-\mu} = 2.73$ and $\alpha_{crit} = \frac{\mu+\gamma+\sqrt{\gamma^2-6\gamma\mu+\mu^2+4\gamma L}}{L-2\mu} = 4.83$.

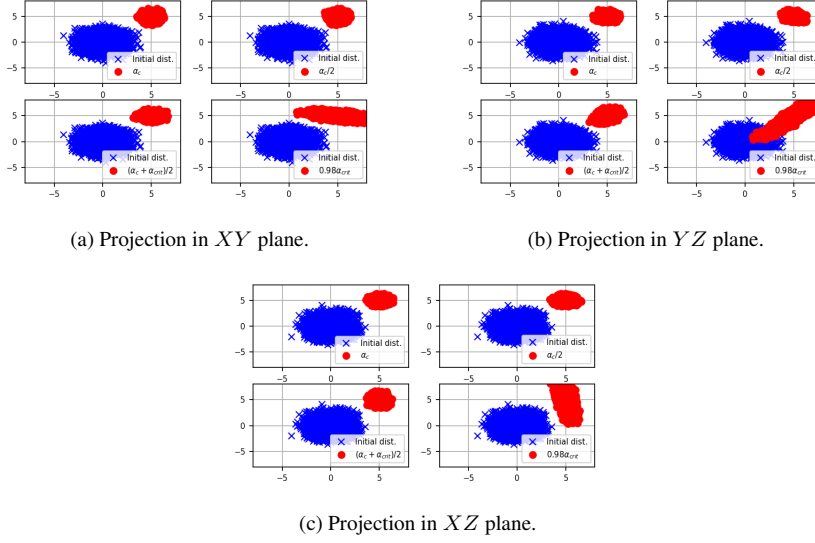


Fig. 10: Initial (blue crosses) and final (red circles) distributions of points generated by the NAG-GS method for scenario $\mu < L/2$; $c = 5$, $\gamma = \mu = 1$, $L = 3$ and $\sigma = 1$ for $\alpha \in \{\alpha_c, \alpha_c/2, (\alpha_c + \alpha_{\text{crit}})/2, 0.98\alpha_{\text{crit}}\}$ with $\alpha_c = \frac{2\mu + 2\sqrt{\mu L}}{L - \mu} = 2.73$ and $\alpha_{\text{crit}} = \frac{\mu + \gamma + \sqrt{\gamma^2 - 6\gamma\mu + \mu^2 + 4\gamma L}}{L - 2\mu} = 4.83$.

3 Fully-implicit scheme

In this section, we present an iterative method based on the NAG transformation G_{NAG} (7) along with a fully implicit discretization to tackle (4) in the stochastic setting, the resulting method shall be referred to as "NAG-FI" method. We propose the following discretization for (6) perturbed with noise; given step size $\alpha_k > 0$:

$$\begin{aligned} \frac{x_{k+1} - x_k}{\alpha_k} &= v_{k+1} - x_{k+1}, \\ \frac{v_{k+1} - v_k}{\alpha_k} &= \frac{\mu}{\gamma_k} (x_{k+1} - v_{k+1}) - \frac{1}{\gamma_k} A x_{k+1} + \sigma \frac{W_{k+1} - W_k}{\alpha_k}. \end{aligned} \quad (42)$$

As done for the NAG-GS method, from a practical point of view, we will use $W_{k+1} - W_k = \Delta W_k = \sqrt{\alpha_k} \eta_k$ where $\eta_k \sim \mathcal{N}(0, 1)$, by the properties of the Brownian motion.

In the quadratic case, that is $f(x) = \frac{1}{2} x^\top A x$, solving equation 42 is equivalent to solve:

$$\begin{bmatrix} x_k \\ v_k + \sigma \sqrt{\alpha_k} \eta_k \end{bmatrix} = \begin{bmatrix} (1 + \alpha_k)I & -\alpha_k I \\ \frac{\alpha_k}{\gamma_k} (A - \mu I) & (1 + \frac{\alpha_k \mu}{\gamma_k}) I \end{bmatrix} \begin{bmatrix} x_{k+1} \\ v_{k+1} \end{bmatrix} \quad (43)$$

where $\eta_k \sim \mathcal{N}(0, 1)$. Furthermore, ODE (8) from the main text is again discretized implicitly:

$$\frac{\gamma_{k+1} - \gamma_k}{\alpha_k} = \mu - \gamma_{k+1}, \quad \gamma_0 > 0. \quad (44)$$

As done for NAG-GS method, heuristically, for general $f \in \mathcal{S}_{L, \mu}^{1,1}$ with $\mu \geq 0$, we just replace Ax_{k+1} in equation 42 with $\nabla f(x_{k+1})$ and obtain the following NAG-FI scheme:

$$\begin{aligned} \frac{x_{k+1} - x_k}{\alpha_k} &= v_{k+1} - x_{k+1}, \\ \frac{v_{k+1} - v_k}{\alpha_k} &= \frac{\mu}{\gamma_k}(x_{k+1} - v_{k+1}) - \frac{1}{\gamma_k} \nabla f(x_{k+1}) + \sigma \frac{W_{k+1} - W_k}{\alpha_k}. \end{aligned} \quad (45)$$

From the first equation, we get $v_{k+1} = \frac{x_{k+1} - x_k}{\alpha_k} + x_{k+1}$ that we substitute within the second equation, we obtain:

$$x_{k+1} = \frac{v_k + \tau_k x_k - \frac{\alpha_k}{\gamma_k} \nabla f(x_{k+1}) + \sigma \sqrt{\alpha_k} \eta_k}{1 + \tau_k} \quad (46)$$

with $\tau_k = 1/\alpha_k + \mu/\gamma_k$.

Computing x_{k+1} is equivalent to computing a fixed point of the operator given by the right-hand side of equation 46. Hence, it is also equivalent to finding the root of the function:

$$g(u) = u - \left(\frac{v_k + \tau_k x_k - \frac{\alpha_k}{\gamma_k} \nabla f(u) + \sigma \sqrt{\alpha_k} \eta_k}{1 + \tau_k} \right) \quad (47)$$

with $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$. In order to compute the root of this function, we consider a classical Newton-Raphson procedure detailed in Algorithm 2. In Algorithm 2, $J_g(\cdot)$

Algorithm 2 Newton-Raphson method

Input: Choose the point $u_0 \in \mathbb{R}^n$, some $\alpha_k, \gamma_k, \tau_k > 0$, $\sigma \geq 0$, and $v_k, \eta_k \in \mathbb{R}^n$.

for $i = 0, 1, \dots$ **do**

 Compute $J_g(u_i) = I_n + \frac{\alpha_k}{\gamma_k(1+\tau_k)} \nabla^2 f(u_i)$

 Compute $g(u_i)$ using equation 47

 Set $u_{i+1} = u_i - [J_g(u_i)]^{-1} g(u_i)$

end for

denotes the Jacobian operator of function g equation 47 w.r.t. u , I_n denotes the identity matrix of size n and $\nabla^2 f$ denotes the Hessian matrix of objective function f . Please note that the iterative method outlined in Algorithm 2 exhibits a connection to the family of second-order methods called the Levenberg-Marquardt algorithm [11, 13] applied to the unconstrained minimization problem $\min_{x \in \mathbb{R}^n} f(x)$ for a twice-differentiable function f . Finally, Algorithm 3 summarizes the NAG-FI method.

By following a similar stability analysis as the one performed for NAG-GS, one can show that this method is unconditionally A-stable as expected by the theory of implicit schemes. In particular, one can show that eigenvalues of the iterations matrix

Algorithm 3 NAG-FI Method

Input: Choose the point $x_0 \in \mathbb{R}^n$, set $v_0 = x_0$, some $\sigma \geq 0$, $\mu \geq 0$, $\gamma_0 > 0$.

```

for  $k = 0, 1, \dots$  do
  Sample  $\eta_k \sim \mathcal{N}(0, 1)$ 
  Choose  $\alpha_k > 0$ 
  Set  $\tau_k = 1/\alpha_k + \mu/\gamma_k$ 
  Compute the root  $u$  of (47) by using Algorithm 2
  Set  $x_{k+1} = u$ 
  Set  $v_{k+1} = \frac{x_{k+1} - x_k}{\alpha_k} + x_{k+1}$ 
  Set  $\gamma_{k+1} := \frac{\gamma_k + \alpha_k \mu}{1 + \alpha_k}$ 
end for

```

are positive decreasing functions w.r.t. step size α , allowing then the choice of any positive value for α . Similarly, one can show that the eigenvalues of the covariance matrix at stationarity associated with the NAG-FI method are decreasing functions w.r.t. α that tends to 0 as soon as $\alpha \rightarrow \infty$. It implies that Algorithm 3 is theoretically able to generate iterates that converge to $\arg \min f$ almost surely, even in the stochastic setting with the potentially quadratic rate of converge. This theoretical result is quickly highlighted in Figure 11 that shows the final distribution of points generated by NAG-FI once used in test setup detailed in Section 2.4, in the most interesting and critical scenario $\mu < L/2$. As expected, α can be chosen as large as desired, we choose here $\alpha = 1000\alpha_c$. Moreover, for increasing α , the final distributions of points are more and more concentrated around x^* .

Therefore, the NAG-FI method constitutes a good basis for deriving efficient second-order methods for tackling stochastic optimization problems, which is hard to find in the current SOTA. Indeed, second-order methods and, more generally, some variants of preconditioned gradient methods have recently been proposed and used in the deep learning community for the training of NN, for instance. However, there is limited empirical success for such methods when used for training NN compared to well-tuned Stochastic Gradient Descent schemes, see [4, 20]. To the best of our knowledge, no theoretical explanations have been brought to formally support these empirical observations. This will be part of our future research directions.

Besides these nice preliminary theoretical results and numerical observations for small dimension problems, there is a limitation of the NAG-FI method that comes from the numerical feasibility for computing the root of the non-linear function (47) that can be very challenging in practice. We will try to address this issue in future works.

4 Convergence to the stationary distribution

Another way to study the convergence of the proposed algorithms is to consider the Fokker-Planck equation for the density function $\rho(t, x)$. We will consider the simple case of the scalar SDE for the stochastic gradient flow (similarly as in (11)). Here $f : \mathbb{R} \rightarrow \mathbb{R}$:

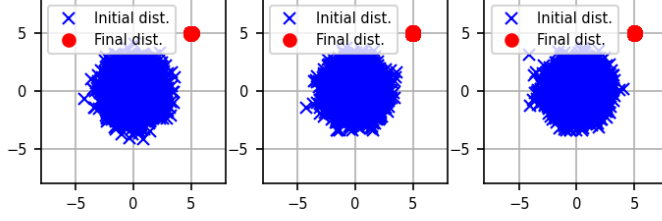


Fig. 11: Projection onto XY , YZ , and XZ planes (from left to right) of initial (blue crosses) and final (red circles) distributions of points generated by NAG-FI method - scenario $\mu < L/2$; $c = 5$, $\gamma = \mu = 1$, $L = 3$ and $\sigma = 1$ for $\alpha = 1000\alpha_c$ with $\alpha_c = \frac{2\mu+2\sqrt{\mu L}}{L-\mu} = 2.73$.

$$dx = -\nabla f(x)dt + dZ = -\nabla f(x)dt + \sigma dW, \quad x(0) \sim \rho(0, x).$$

It is well known, that the density function for $x(t) \sim \rho(t, x)$ satisfies the corresponding Fokker-Planck equation:

$$\frac{\partial \rho(t, x)}{\partial t} = \nabla (\rho(t, x) \nabla f(x)) + \frac{\sigma^2}{2} \Delta \rho(t, x) \quad (48)$$

For (48) one could write down the stationary (with $t \rightarrow \infty$) distribution

$$\rho^*(x) = \lim_{t \rightarrow \infty} \rho(t, x) = \frac{1}{Z} \exp \left(-\frac{2}{\sigma^2} f(x) \right), \quad Z = \int_{x \in V} \exp \left(-\frac{2}{\sigma^2} f(x) \right) dx. \quad (49)$$

It is useful to compare different optimization algorithms in terms of convergence in the probability space because it allows us to study the methods in the non-convex setting. We have to address two problems with this approach. Firstly, we need to specify some distance functional between current distribution $\rho_t = \rho(t, x)$ and stationary distribution $\rho^* = \rho^*(x)$. Secondly, we do not need to have access to the densities ρ_t, ρ^* themselves.

For the first problem, we will consider the following distance functionals between probability distributions in the scalar case:

- **Kullback-Leibler divergence.** Several studies dedicated to convergence in probability space are available [2, 6, 9]. We used the approach proposed in [17] to estimate KL divergence between continuous distributions based on their samples.

-
- **Wasserstein distance.** Wasserstein distance is relatively easy to compute for scalar densities. Also, it was shown, that the stochastic gradient process with a constant learning rate is exponentially ergodic in the Wasserstein sense [10].
 - **Kolmogorov-Smirnov statistics.** We used the two-sample Kolmogorov-Smirnov test for goodness of fit.

To the best of our knowledge, the explicit formula for the stationary distribution of Fokker-Planck equations for the ASG SDE (11) remains unknown. That is why we have decided to get samples from the empirical stationary distributions using Euler-Maruyama integration [14] with a small enough step size of corresponding SDE with a bunch of different independent initializations.

We tested two functions, which are presented in Figure 12. We initially generated 100 points uniformly in the function domain. Then we independently solved the initial value problem (9) for each of them with [14]. Results of the integration are presented in Figure 13. One can see, that in the relatively easy case (Figure 12a), NAG-GS converges faster, than gradient flow to its stationary distribution, see Figure 13a. At the same time, in the hard case (Figure 12b), NAG-GS is more robust to the large step size, see Figure 13b.

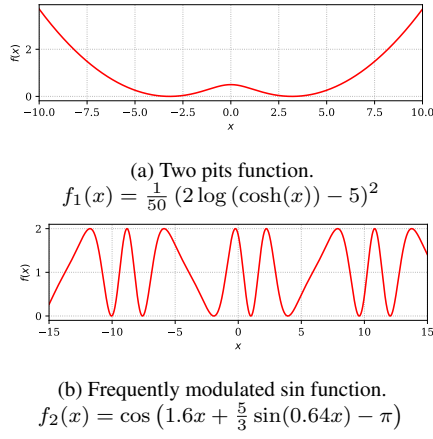


Fig. 12: Non convex scalar functions to test

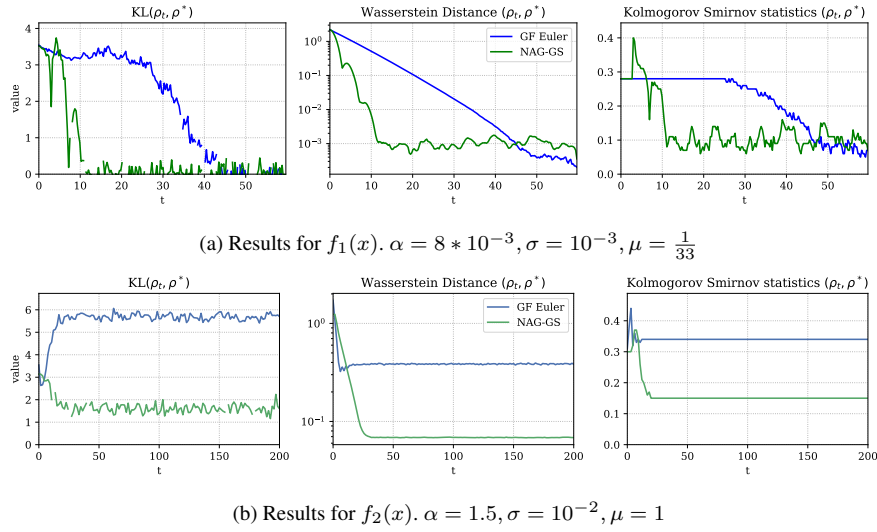


Fig. 13: Convergence in probabilities of Euler integration of Gradient Flow (GF Euler) and NAG-GS for the non-convex scalar problems.

5 Additional insights from experimental evaluation

In this section, we provide additional experimental details. In particular, we discuss our experimental setup a little bit more and give some insights about NAG-GS as well. Our computational resources are limited to a single Nvidia DGX-1 with 8 GPUs Nvidia V100. Almost all experiments were carried out on a single GPU. The only exception is for the training of ResNet50 on ImageNet, which used all 8 GPUs.

5.1 Phase diagrams

Our theoretical analysis of NAG-GS in the convex case includes some conditions on the optimizer parameters α , γ , and μ . In particular, it is required that $\mu > 0$ and $\gamma \geq \mu$. To bring some insights about these remarks in the non-convex setting and inspired by [19], we experimentally study the convergence regions of NAG-GS and sketch out the phase diagrams of convergence for different projection planes, see Figure 14.

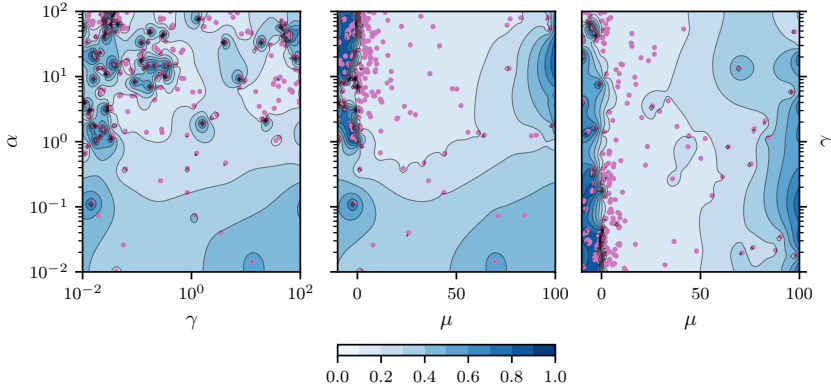


Fig. 14: Landscapes of classification error for ResNet-20 model trained on CIFAR-10 with NAG-GS after projections onto $\alpha - \gamma$, $\alpha - \mu$ and $\gamma - \mu$ planes (from left to right). Hyperparameter optimization algorithm samples learning rate α from $[10^{-2}, 10^2]$, factor γ from $[10^{-2}, 10^2]$, and factor μ from $[-10, 90]$. Hyperparameters α and γ are sampled from log-uniform distribution, and hyperparameter μ is sampled from a uniform distribution.

We consider the same setup as in Section 3.4 in the main text, a paragraph about the ResNet-20 model, and use hyper-optimization library OPTUNA [1]. Our preliminary experiments on RoBERTa show that α should be of magnitude 10^{-1} . With the estimate of the Hessian spectrum of ResNet-20, we define the following search space:

$$\alpha \sim \text{LogUniform}(10^{-2}, 10^2), \quad \gamma \sim \text{LogUniform}(10^{-2}, 10^2), \quad \mu \sim \text{Uniform}(-10, 100).$$

We sample a fixed number of triples and train the ResNet-20 model on CIFAR-10. The objective function is a top-1 classification error.

Table 1: The comparison of a single step duration for different optimizers on RESNET-20 on CIFAR-10. ADAM-like optimizers have in twice larger state than SGD with momentum or NAG-GS.

OPTIMIZER	MEAN, S	VARIANCE, S	REL. MEAN	REL. VARIANCE
SGD	0.458	0.008	1.0	1.0
NAG-GS	1.648	0.045	3.6	5.5
SGD-M	3.374	0.042	7.4	5.2
SGD-MW	3.512	0.037	17.7	4.7
ADAMW	5.208	0.102	11.4	12.6
ADAM	7.919	0.169	17.3	20.8

We report a convergence almost everywhere within the projected search space onto α - γ plane (see Figure 14). The analysis of projections onto α - μ and γ - μ planes brings different conclusions: there are regions of convergence for negative μ for some $\alpha < \alpha_{th}$ and $\gamma > \gamma_{th}$. Also, there is a subdomain of negative μ comparable to a domain of positive μ in the sense of the target metrics. Moreover, the majority of sampled points are located in the vicinity of the band $\lambda_{min} < \mu < \lambda_{max}$.

5.2 Implementation Details

In our work, we implemented NAG-GS in PyTorch [16] and JAX [5, 3]. Both implementations are used in our experiments and available online². According to Algorithm 1, the size of the NAG-GS state equals the number of optimization parameters, which makes NAG-GS comparable to SGD with momentum. It is worth noting that Adam-like optimizers have a twice larger state than NAG-GS. The arithmetic complexity of NAG-GS is linear $O(n)$ in the number of parameters. Table 1 shows a comparison of the computational efficiency of standard optimizers used in practice. Although forward pass and gradient computations usually make the main contribution to the training step, there is a setting where the efficiency of gradient updates is essential (e.g. batch size or a number of intermediate activations are small with respect to a number of parameters).

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2019)
2. Arnold, A., Markowich, P., Toscani, G., Unterreiter, A.: On convex sobolev inequalities and the rate of convergence to equilibrium for fokker-planck type equations (2001)
3. Babuschkin, I., Baumli, K., Bell, A., Bhupatiraju, S., Bruce, J., Buchlovsky, P., Budden, D., Cai, T., Clark, A., Danihelka, I., Fantacci, C., Godwin, J., Jones, C., Hemsley, R., Hennigan, T., Hessel, M., Hou, S., Kapturowski, S., Keck, T., Kemaev, I., King, M., Kunesch, M., Martens, L., Merzic,

² <https://github.com/naggsopt/naggs>

- H., Mikulik, V., Norman, T., Quan, J., Papamakarios, G., Ring, R., Ruiz, F., Sanchez, A., Schneider, R., Sezen, E., Spencer, S., Srinivasan, S., Wang, L., Stokowiec, W., Viola, F.: The DeepMind JAX Ecosystem (2020). URL <http://github.com/deepmind>
4. Botev, A., Ritter, H., Barber, D.: Practical Gauss-Newton optimisation for deep learning. In: Proceedings of the 34th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 70, pp. 557–565. PMLR (2017)
5. Bradbury, J., Frostig, R., Hawkins, P., Johnson, M.J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., Zhang, Q.: JAX: Composable Transformations of Python+NumPy Programs (2018). URL <http://github.com/google/jax>
6. Chewi, S., Le Gouic, T., Lu, C., Maunu, T., Rigollet, P.: Svdg as a kernelized wasserstein gradient flow of the chi-squared divergence. *Advances in Neural Information Processing Systems* **33**, 2098–2109 (2020)
7. Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., Storkey, A.: Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623* (2017)
8. Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On large-batch training for deep learning: Generalization gap and sharp minima. In: *International Conference on Learning Representations* (2017). URL <https://openreview.net/forum?id=HloyRlYgg>
9. Lambert, M., Chewi, S., Bach, F., Bonnabel, S., Rigollet, P.: Variational inference via wasserstein gradient flows. *arXiv preprint arXiv:2205.15902* (2022)
10. Latz, J.: Analysis of stochastic gradient descent in continuous time. *Statistics and Computing* **31**(4), 1–25 (2021)
11. Levenberg, K.: A method for the solution of certain non-linear problems in least squares. *Quart. Appl. Math.* **2** (1944)
12. Luo, H., Chen, L.: From differential equation solvers to accelerated first-order methods for convex optimization. *Mathematical Programming* pp. 1–47 (2021)
13. Marquardt, D.W.: An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics* **11**(2), 431–441 (1963)
14. Maruyama, G.: Continuous markov processes and stochastic equations. *Rendiconti del Circolo Matematico di Palermo* **4**(1), 48–90 (1955)
15. Nesterov, Y.: *Lectures on Convex optimization*, vol. 137. Springer Optimization and Its Applications (2018)
16. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: *Automatic Differentiation in PyTorch* (2017)
17. Pérez-Cruz, F.: Kullback-leibler divergence estimation of continuous distributions. In: *2008 IEEE international symposium on information theory*, pp. 1666–1670. IEEE (2008)
18. Roberts, G., Stramer, O.: Langevin diffusions and metropolis-hastings algorithms. *Methodology And Computing In Applied Probability* **4**, 337–357 (2002)
19. Velikanov, M., Kuznedelev, D., Yarotsky, D.: A view of mini-batch sgd via generating functions: conditions of convergence, phase transitions, benefit from negative momenta (2022). DOI 10.48550/arxiv.2206.11124
20. Zeiler, M.D.: Adadelata: An adaptive learning rate method (2012). DOI 10.48550/ARXIV.1212.5701