An SDE for Modeling SAM: Theory and Insights

Enea Monzio Compagnoni [1] **Luca Biggio** [2] **Antonio Orvieto** [2] **Frank Norbert Proske** [3] **Hans Kersting** [4] **Aurelien Lucchi** [1]

Abstract

We study the SAM (Sharpness-Aware Minimization) optimizer which has recently attracted a lot of interest due to its increased performance over more classical variants of stochastic gradient descent. Our main contribution is the derivation of continuous-time models (in the form of SDEs) for SAM and two of its variants, both for the full-batch and mini-batch settings. We demonstrate that these SDEs are rigorous approximations of the real discrete-time algorithms (in a weak sense, scaling linearly with the learning rate). Using these models, we then offer an explanation of why SAM prefers flat minima over sharp ones – by showing that it minimizes an implicitly regularized loss with a Hessian-dependent noise structure. Finally, we prove that SAM is attracted to saddle points under some realistic conditions. Our theoretical results are supported by detailed experiments.

1. Introduction

Optimization plays a fundamental role in the performance of machine learning models. The core problem it addresses is the minimization of the following optimization problem:

$$\min_{x\in\mathbb{R}^d}\left[f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x)\right], \tag{1}$$

where $f, f_i : \mathbb{R}^d \to \mathbb{R}$ for $i = 1, \dots, N$. In machine learning, f is an empirical risk (or loss) function where f_i are the contributions due to the i-th data point. In this notation, $x \in \mathbb{R}^d$ is a vector of trainable parameters and N is the size of the dataset.

Solving Eq. (1) is typically achieved via Gradient Descent (GD) methods that, starting from a given estimate x_0, iteratively update an estimate x_k as follows,

$$x_{k+1} = x_k - \eta\nabla f(x_k), \tag{2}$$

where $\eta > 0$ is the learning rate. Since $\nabla f(x)$ requires computing the average of the N gradients $\nabla f_i(x)$ (which is computationally expensive for large datasets where $N \gg 1$), it is common to instead replace $\nabla f(x_k)$ with a gradient estimated on a subset γ_k of size $B \geq 1$ of the dataset which is called a *mini-batch*. The resulting algorithm is known as Stochastic Gradient Descent (SGD) whose update is

$$x_{k+1} = x_k - \eta\nabla f_{\gamma_k}(x_k), \tag{3}$$

where $\{\gamma_k\}$ are modelled here as i.i.d. random variables uniformly distributed and taking value in $\{1, \cdots, N\}$.

Recently, Foret et al. (2021) proposed a stochastic optimizer known as Sharpness-Aware Minimization (SAM), which yields significant performance gains in various fields such as computer vision and natural language processing (Bahri et al., 2022; Foret et al., 2021). The general idea behind SAM is to seek parameters in low-loss regions that have a flatter curvature, which has been shown to improve the generalization of the model (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017; Dziugaite & Roy, 2017; Jiang et al., 2019). For a radius $\rho > 0$, the iteration of SAM is

$$x_{k+1} = x_k - \eta\nabla f_{\gamma_k}\left(x_k + \rho\frac{\nabla f_{\gamma_k}(x_k)}{\|\nabla f_{\gamma_k}(x_k)\|}\right). \tag{4}$$

Numerous works have studied SAM and proposed variants such as ESAM (Du et al., 2022), ASAM (Kwon et al., 2021), GSAM (Zhuang et al., 2022), as well as Random SAM and Variational SAM (Ujváry et al., 2022). Since SAM is hard to treat theoretically, (Andriushchenko & Flammarion, 2022) introduced USAM which is more easily analyzable as it drops the gradient normalization in Eq. (4), thus yielding the following update:

$$x_{k+1} = x_k - \eta\nabla f_{\gamma_k}\left(x_k + \rho\nabla f_{\gamma_k}(x_k)\right). \tag{5}$$

[1]Department of Mathematics & Computer Science, University of Basel, Basel, Switzerland [2]Department of Computer Science, ETH Zürich, Zürich, Switzerland [3]Department of Mathematics, University of Oslo, Oslo, Norway [4]Inria, Ecole Normale Supérieure PSL Research University, Paris, France. Correspondence to: Enea Monzio Compagnoni <enea.monziocompagnoni@unibas.ch>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Before analyzing the full version of SAM, we first take a smaller step toward it by considering a variant with a deterministic normalization factor. We call the resulting algorithm DNSAM (Deterministic Normalization SAM), whose update step is

$$x_{k+1} = x_k - \eta \nabla f_{\gamma_k} \left(x_k + \rho \frac{\nabla f_{\gamma_k}(x_k)}{\|\nabla f(x_k)\|} \right). \quad (6)$$

We will demonstrate both theoretically and empirically that DNSAM is a better proxy of SAM than USAM. However, we do not claim that DNSAM is an algorithm to be used in practice as its update requires the calculation of the full gradient of the loss.

Following the theoretical framework of (Li et al., 2017), our work provides the first *formal* derivation of the SDEs of DNSAM, USAM, and SAM. Formally, such continuous-time models are weak approximations (i.e. approximations in distribution) of their respective discrete-time models. Importantly, SDE models are not meant to be used as practical implementations since they have to be discretized, giving rise to their discrete-time counterparts. Instead, continuous-time models have typically been used in the recent literature to derive novel insights about the discrete algorithms, see e.g. (Su et al., 2014; Li et al., 2017).

We make the following contributions:

1. **Small ρ regime.** If $\rho = \mathcal{O}(\eta)$, we show that USAM, DNSAM, and SAM essentially behave like SGD.
2. **Moderate ρ regime.** For $\rho = \mathcal{O}(\sqrt{\eta})$, we derive an SDE model of USAM (7), of DNSAM (10), and of SAM (11). These can be interpreted as the SDE of SGD on an implicitly regularized loss and with an additional *implicit curvature-induced* noise. Leveraging these results, we demonstrate that the additional noise is driven by the Hessian of the loss so that the noise of the processes is larger in sharp minima. This is a key factor that leads SAM and its variants to prefer flatter minima where the additional noise decreases. However, while larger values of ρ increase the noise of the process, it also amplifies the implicit bias of the optimizer toward critical points independently of whether they are minima, saddles, or maxima.
3. Both in the full and mini-batch versions, USAM and SAM have very different implicit regularizations.
4. USAM might be attracted by saddles if ρ is too large. Differently, for any $\rho > 0$, DNSAM and SAM might be attracted by saddles but eventually, escape them after a long time. Thus, DNSAM is a more reliable model to theoretically study SAM than USAM.
5. **Empirical validation.** We empirically validate the proposed SDEs on several models and landscapes commonly studied in the optimization and machine learning communities.

In order to gain further insights from these continuous-time models, we also study their behaviors on quadratic losses. The latter are commonly used to model the landscape in the proximity of a critical point (Ge et al., 2015; Levy, 2016; Jin et al., 2017; Poggio et al., 2017; Mandt et al., 2017b), including several recent works that studied SAM (Bartlett et al., 2022; Wen et al., 2023). This leads us to the following important observations:

1. **ODE - Pitfalls.** After noticing that the ODE of SAM and DNSAM coincide, we derive precise conditions under which SAM and USAM converge to the origin even when it is a saddle or a maximum.
2. **SDE - Pitfalls.** We derive the stationary distribution of the USAM SDE and find that even this model is attracted by saddles under the same condition on ρ as found for the ODE [1]. In contrast to USAM, we find that the dynamics of DNSAM is more complex: while a certain region centered at the origin behaves like an attractor, the origin itself repulses the dynamics away as the volatility rapidly increases to infinity. This behavior of DNSAM is consistent with what was empirically reported in (Kaddour et al., 2022) about SAM being able to get stuck around saddles. To the best of our knowledge, this is the first time that these potential pitfalls are formally demonstrated.
3. **Empirical validation.** We empirically validate our claims for the quadratic loss as well as other models.

2. Related Work

Theoretical Understanding of SAM The current understanding of the dynamics of SAM and USAM is still limited. Prior work includes the recent work by (Bartlett et al., 2022) that shows that, for convex quadratics, SAM converges to a cycle oscillating between the sides of the minimum in the direction with the largest curvature. For the non-quadratic case, they also show that the dynamics drifts towards wider minima. A concurrent work by (Wen et al., 2023) makes similar findings to (Bartlett et al., 2022) as well as provides further insights regarding which notion of sharpness SAM regularizes. Interestingly, the behavior of full-batch and mini-batch SAM is intrinsically different. The former minimizes the largest eigenvalue of the hessian of the loss, while the latter tries to uniformly reduce the magnitude of the trace of the hessian of the loss. More interestingly, (Wen et al., 2023) show how the dynamics of 1-SAM can be divided into two phases. The first phase follows the gradient flow with respect to the loss until the dynamics approaches a manifold of minimizers. In the second phase, the dynamics is driven towards parts of the landscape with a lower trace of the hessian of the loss. (Rangwani et al., 2022) showed that

[1] Of course, the SDE does not point-wise converge to the origin but rather oscillates around it with a certain variance.

USAM could in some cases escape saddles faster than SGD. We however note that their analysis is not completely formal as it relies on prior results by (Daneshmand et al., 2018) which were specifically derived for SGD, not for USAM. On our side, we here provide a more complete and rigorous description that shows that USAM can be much slower than SGD at escaping a saddle. Finally, a concurrent work (Kim et al., 2023) *informally* derived an SDE for USAM around critical points, which relies on approximating the objective function by a quadratic function. We remark that the authors did not formally derive any guarantee showing the SDE closely approximates the true discrete-time algorithm. In contrast, we formally and empirically demonstrate the validity of our SDEs. In addition, our SDEs and analyses *do not require* the quadratic approximation assumption made by (Kim et al., 2023) and are instead valid for the entire trajectory of an optimizer, including, of course, around critical points.

ODE Approximations Continuous-time models in the form of (stochastic) differential equations are a well-established tool to study discrete-time optimizers; see e.g. Helmke & Moore (1994) and Kushner & Yin (2003). In machine learning, such models have lately received increasing interest to study both deterministic and stochastic optimizers. A notable reference is the work by Su et al. (2014) that derives a second-order ODE to model Nesterov's accelerated gradient. ODE models have also been used recently to study SAM. This includes the work of Wen et al. (2023, Section 4.2) discussed above, as well as Andriushchenko & Flammarion (2022, Appendix B.1). Importantly we highlight two significant differences with our work. First, our analysis focuses on the stochastic setting for which we derive SDEs. Second, the ODE representations used in Wen et al. (2023) only hold formally in the limit $\rho \to 0$, which is not the case in practical settings where $\rho > 0$. In contrast, our analysis allows for significantly larger values of ρ, more precisely $\rho = \mathcal{O}(\sqrt{\eta})$. Last but not least, neither of these papers empirically validates the ODEs they derived.

SDE Approximations of Stochastic Optimizers. For *stochastic* optimizers, Li et al. (2017; 2019) derived an SDE that provably approximates SGD (in the weak sense, i.e. in distribution). The validity of this SDE model was experimentally tested in (Li et al., 2021). Similar results are derived for ASGD by (An et al., 2020), and for RMSprop and Adam by Malladi et al. (2022). In this paper, we derive an SDE approximation for SAM, DNSAM, and USAM. The proof technique employed in our work (as well as in An et al. (2020); Malladi et al. (2022)) relies on the theoretical framework established by Li et al. (2017; 2019) (which itself requires Assumption A.3 to hold). SDE approximations have also been derived for different types of noise. This includes heavy-tailed noise that is shown to be a good model for the noise of SGD in Simsekli et al. (2019), although the evidence is still somewhat contested (Panigrahi et al., 2019; Xie et al., 2021; Li et al., 2021). Zhou et al. (2020) also derived a Lévy-driven stochastic differential equation to model the non-gaussianity of the noise, which however does not enjoy any type of known theoretical approximation guarantee. Finally, fractional Brownian noise, a generalization of Brownian noise that allows for correlation, was used by (Lucchi et al., 2022).

Applications of SDE Approximations. Continuous-time models are valuable analysis tools to study and design new optimization methods. For instance, one concrete application of such models is the use of *stochastic optimal control* to select the learning rate (Li et al., 2017; 2019) or the batch size (Zhao et al., 2022). In addition, *scaling rules* to adjust the optimization hyperparameters w.r.t. the batch size can be recovered from SDE models (Malladi et al., 2022). Apart from these algorithmic contributions, SDE approximation can be useful to better understand stochastic optimization methods. In this regard, (Jastrzebski et al., 2018) analyzed the factors influencing the minima found by SGD, and (Orvieto & Lucchi, 2019) derived convergence bounds for mini-batch SGD and SVRG. (Smith et al., 2020) used an SDE model to distinguish between "noise-dominated" and "curvature-dominated" regimes of SGD. Yet another example is the study of *escape times* of SGD from minima of different sharpness (Xie et al., 2021). Moreover, (Li et al., 2020) and (Kunin et al., 2021) studied the *dynamics* of the SDE approximation under some symmetry assumptions. Finally, SDEs can be studied through the lens of various tools in the field of stochastic calculus, e.g. the Fokker–Planck equation gives access to the stationary distribution of a stochastic process. Such tools are for instance valuable in the field of Bayesian machine learning (Mandt et al., 2017a). For additional references, see (Kushner & Yin, 2003; Ljung et al., 2012; Chen et al., 2015; Mandt et al., 2015; Chaudhari & Soatto, 2018; Zhu et al., 2019; He et al., 2018; An et al., 2020).

3. Formal Statements & Insights: The SDEs

In this section, we present the general formulations of the SDEs of USAM, DNSAM, and SAM. Due to the technical nature of the analysis, we refer the reader to the Appendix for the complete formal statements and proofs. For didactic reasons, we provide simplified versions under mild additional assumptions in the main paper.

Definition 3.1 (Weak Approximation)**.** Let G denote the set of continuous functions $\mathbb{R}^d \to \mathbb{R}$ of at most polynomial growth, i.e. $g \in G$ if there exists positive integers $\kappa_1, \kappa_2 > 0$ such that $|g(x)| \leq \kappa_1 \left(1 + |x|^{2\kappa_2}\right)$, for all $x \in \mathbb{R}^d$. Then, we say that a continuous-time stochastic process $\{X_t : t \in [0, T]\}$ is an order α weak approximation

of a discrete stochastic process $\{x_k : k = 0, \dots, N\}$ if for every $g \in G$, there exists a positive constant C, independent of η, such that $\max_{k=0,\dots,N} |\mathbb{E}g(x_k) - \mathbb{E}g(X_{k\eta})| \leq C\eta^\alpha$.

This definition comes from the field of numerical analysis of SDEs, see (Mil'shtein, 1986). Consider the case where $g(x) = \|x\|^j$, then the bound limits the difference between the j-th moments of the discrete and the continuous process.

In Theorem A.7 (USAM), Theorem A.11 (DNSAM), and Theorem A.16 (SAM), we prove that if $\rho = \mathcal{O}(\eta)$ (small ρ regime), the SDE of SGD (Eq. (18)) is also an order 1 weak approximation for USAM, DNSAM, and SAM. In contrast, in the more realistic moderate ρ regime where $\rho = \mathcal{O}(\sqrt{\eta})$, Eq. (18) is no longer an order 1 weak approximation for any the models we analyze. Under such a condition, we recover more insightful SDEs.

3.1. USAM SDE

Theorem 3.2 (USAM SDE - Informal Statement of Theorem A.4)**.** *Under sufficient regularity conditions and $\rho = \mathcal{O}(\sqrt{\eta})$ the solution of the following SDE is an order 1 weak approximation of the discrete update of USAM* (5)*:*

$$dX_t = -\nabla \tilde{f}^{USAM}(X_t)dt + \sqrt{\eta\left(\Sigma^{SGD}(X_t) + \rho\left(\tilde{\Sigma}(X_t) + \tilde{\Sigma}(X_t)^\top\right)\right)}dW_t, \tag{7}$$

where $\tilde{\Sigma}(x)$ is defined as

$$\mathbb{E}\left[\left(\nabla f(x) - \nabla f_\gamma(x)\right)\cdot \left(\mathbb{E}\left[\nabla^2 f_\gamma(x)\nabla f_\gamma(x)\right] - \nabla^2 f_\gamma(x)\nabla f_\gamma(x)\right)^\top\right] \tag{8}$$

and $\tilde{f}^{USAM}(x) := f(x) + \frac{\rho}{2}\mathbb{E}\left[\|\nabla f_\gamma(x)\|_2^2\right]$.

To have a more direct comparison with the SDE of SGD (Eq. (18)), we prove Corollary 3.3, a consequence of Theorem 3.2 that provides a more interpretable SDE for USAM.

Corollary 3.3 (Informal Statement of Corollary A.6)**.** *Under the assumptions of Theorem* (3.2) *and assuming a constant gradient noise covariance, i.e. $\nabla f_\gamma(x) = \nabla f(x) + Z$ such that Z is a noise vector that does not depend on x, the solution of the following SDE is the order 1 weak approximation of the discrete update of USAM* (5)*:*

$$dX_t = -\nabla \tilde{f}^{USAM}(X_t)dt + \left(I_d + \rho\nabla^2 f(X_t)\right)\left(\eta\Sigma^{SGD}(X_t)\right)^{1/2}dW_t, \tag{9}$$

where $\tilde{f}^{USAM}(x) := f(x) + \frac{\rho}{2}\|\nabla f(x)\|_2^2$.

Corollary 3.3 shows that the dynamics of USAM is equivalent to that of SGD on a regularized loss and with an additional noise component that depends on the curvature of the landscape (captured by the term $\nabla^2 f$).

3.2. DNSAM: A step towards SAM

Theorem 3.4 (DNSAM SDE - Informal Statement of Theorem A.9)**.** *Under sufficient regularity conditions, assuming a constant gradient noise covariance, i.e. $\nabla f_\gamma(x) = \nabla f(x) + Z$ such that Z is a noise vector that does not depend on x, and $\rho = \mathcal{O}(\sqrt{\eta})$ the solution of the following SDE is the order 1 weak approximation of the discrete update of DNSAM* (6)*:*

$$dX_t = -\nabla \tilde{f}^{DNSAM}(X_t)dt + \left(I_d + \rho\frac{\nabla^2 f(X_t)}{\|\nabla f(X_t)\|_2}\right)\left(\eta\Sigma^{SGD}(X_t)\right)^{\frac{1}{2}}dW_t \tag{10}$$

and $\tilde{f}^{DNSAM}(x) = f(x) + \rho\|\nabla f(x)\|_2$.

Theorem 3.4 shows that similarly to USAM, the dynamics of DNSAM is equivalent to that of SGD on a regularized loss with an additional noise component that depends on the curvature of the landscape. However, we notice that, unlike USAM, the volatility component explodes near critical points.

3.3. SAM SDE

Theorem 3.5 (SAM SDE - Informal Statement of Theorem A.12)**.** *Under sufficient regularity conditions and $\rho = \mathcal{O}(\sqrt{\eta})$ the solution of the following SDE is the order 1 weak approximation of the discrete update of SAM* (4)*:*

$$dX_t = -\nabla \tilde{f}^{SAM}(X_t)dt + \sqrt{\eta\left(\Sigma^{SGD}(X_t) + \rho\left(\hat{\Sigma}(X_t) + \hat{\Sigma}(X_t)^\top\right)\right)}dW_t \tag{11}$$

where $\hat{\Sigma}(x)$ is defined as

$$\mathbb{E}\left[\left(\nabla f(x) - \nabla f_\gamma(x)\right)\cdot \left(\mathbb{E}\left[\frac{\nabla^2 f_\gamma(x)\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right] - \frac{\nabla^2 f_\gamma(x)\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right)^\top\right] \tag{12}$$

and $\tilde{f}^{SAM}(x) := f(x) + \rho\mathbb{E}\left[\|\nabla f_\gamma(x)\|_2\right]$.

To have a more direct comparison with the SDE of SGD (Eq. (18)), we derive a corollary of Theorem 3.5 that provides a more insightful SDE for SAM.

Corollary 3.6 (Informal Statement of Corollary A.15)**.** *Under the assumptions of Theorem 3.5 and assuming a constant gradient noise covariance, i.e. $\nabla f_\gamma(x) = \nabla f(x) + Z$ such that Z is a noise vector that does not depend on x, the solution of the following SDE is an order 1 weak approximation of the discrete update of SAM* (4)

$$dX_t = -\nabla \tilde{f}^{SAM}(X_t)dt + \sqrt{\eta\left(\Sigma^{SGD}(X_t) + \rho H_t\left(\bar{\Sigma}(X_t) + \bar{\Sigma}(X_t)^\top\right)\right)}dW_t \tag{13}$$

where $H_t := \nabla^2 f(X_t)$ *and* $\bar{\Sigma}(x)$ *is defined as*

$$\mathbb{E}\left[(\nabla f(x) - \nabla f_\gamma(x)) \cdot \left(\mathbb{E}\left[\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right] - \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right)^\top\right], \tag{14}$$

and $\tilde{f}^{SAM}(x) := f(x) + \rho\mathbb{E}\left[\|\nabla f_\gamma(x)\|_2\right]$.

We note that the regularization term of SAM is the *expected* norm of some gradient. While one can of course use sampling in order to simulate the SDE in Eq. (13), it results in an additional computational cost, which is worth highlighting.

3.4. Comparison: USAM vs (DN)SAM

The analyses of the SDEs we derived (Eq. (9), Eq. (13), and Eq. (10)) reveal that all three algorithms are implicitly minimizing a regularized loss with an additional injection of noise (in addition to the SGD noise). While the regularized loss is $\frac{\rho}{2}\|\nabla f(x)\|_2^2$ for USAM, it is $\rho\|\nabla f(x)\|_2$ (not squared) for DNSAM, and $\rho\mathbb{E}\left[\|\nabla f_\gamma(x)\|_2\right]$ for SAM. Therefore, when the process is closer to a stationary point, the regularization is stronger for (DN)SAM while it is the opposite when it is far away.

Regarding the additional noise, we observe that it is *curvature-dependent* as the Hessian appears in the expression of all volatility terms. Note that the sharper the minimum, the larger the noise contribution from the Hessian. This phenomenon is more extreme for DNSAM where the volatility is scaled by the inverse of the norm of the gradient which drives the volatility to explode as it approaches a critical point. Based on the SAM SDE, it is clear that the diffusion term is more regular than that of DNSAM (in the sense that the denominator does not vanish). Therefore, SAM is intrinsically less volatile than DNSAM near a critical point. In contrast, we note that the SAM dynamics exhibits more randomness than USAM which in turn is more noisy than SGD. These theoretical insights are validated experimentally in Section 5. Therefore, it is intuitive that SAM and its variants are more likely to stop or oscillate in a flat basin and more likely to escape from sharp minima than SGD.

We conclude with a discussion of the role of ρ. On one hand, larger values of ρ increase the variance of the process. On the other hand, they also increase the marginal importance of the factor $\frac{\rho}{2}\|\nabla f(x)\|_2^2$ (USAM) and $\rho\|\nabla f(x)\|_2$ (DNSAM), and $\rho\mathbb{E}\left[\|\nabla f_\gamma(x)\|_2\right]$ (SAM), which for sufficiently large ρ might overshadow the marginal relevance of minimizing f and thus implicitly bias the optimizer toward any point with zero gradients, including saddles and maxima. We study this pitfall in detail for the quadratic case in the next section and verify it experimentally in Section 5 for other models as well. See Table 2 and Table 3 for a detailed summary.

4. Behavior Near Saddles - Theory

In this section, we leverage the ODEs (modeling the full-batch algorithms) and SDEs (modeling the mini-batch algorithms) to study the behavior of SAM and its variants near critical points. We especially focus on saddle points that have been a subject of significant interest in machine learning (Jin et al., 2017; 2021; Daneshmand et al., 2018). We consider a quadratic loss (which as mentioned earlier is a common model to study saddle points) of the form $f(x) = \frac{1}{2}x^\top Hx$. W.l.o.g. we assume that the Hessian matrix H is diagonal [2] and denote the eigenvalues of H by $(\lambda_1, \dots, \lambda_d)$ where $\lambda_1 \geq \lambda_1 \geq \cdots \geq \lambda_d$. If there are negative eigenvalues, we denote by λ_* the largest negative eigenvalue.

4.1. USAM ODE

We study the deterministic dynamics of USAM on a quadratic which is defined as

$$dX_t = -H\left(I_d + \rho H\right)X_t dt \Rightarrow X_t^j = X_0^j e^{-\lambda_j(1+\rho\lambda_j)t}. \tag{15}$$

Therefore, it is obvious (see Lemma C.1) that, if all the eigenvalues of H are positive, for all $\rho > 0$, we have that $X_t^j \overset{t\to\infty}{\to} 0, \quad \forall j \in \{1, \dots, d\}$. In particular, we notice that, since $e^{-\lambda_j(1+\rho\lambda_j)t} < e^{-\lambda_j t}$, such convergence to 0 is faster for the flow of USAM than for the gradient flow. More interestingly, if ρ is *too large*, the following result states that the deterministic dynamics of USAM might be attracted by a saddle or even a maximum.

Lemma 4.1 (Informal Statement of Lemma C.2)**.** *Let* H *have at least one strictly negative eigenvalue. Then, for all* $\rho > -\frac{1}{\lambda_*}, \quad X_t^j \overset{t\to\infty}{\to} 0, \quad \forall j \in \{1, \dots, d\}$.

Therefore, if ρ is not chosen appropriately, USAM might converge to $0 \in \mathbb{R}^d$, even if it is a saddle point or a maximum, which is very *undesirable*. Of course, we observe that if $\rho < \frac{1}{\lambda_*}$, USAM will diverge from the saddle (or maximum), which is *desirable*. Interestingly, we also notice that since $e^{-\lambda_j(1+\rho\lambda_j)t} < e^{-\lambda_j t}$, USAM will escape the saddle but more slowly than the gradient flow.

4.2. USAM SDE

Based on Corollary A.6, if we assume that $\Sigma^{\text{SGD}} = \varsigma^2 I_d$, the SDE of USAM on a quadratic is given by

$$dX_t = -H\left(I_d + \rho H\right)X_t dt + \left[(I_d + \rho H)\sqrt{\eta}\varsigma\right]dW_t. \tag{16}$$

Theorem 4.2 (Stationary distribution - Theorem C.3 and Theorem C.4)**.** *If all the eigenvalues of* H *are positive, i.e.* 0 *is a minimum, we have that for any* $\rho > 0$, $\forall i \in \{1, \cdots, d\}$,

[2] Recall that symmetric matrices can be diagonalized.

the stationary distribution of Eq. (16) *is*

$$P(x) = \sqrt{\frac{\lambda_i}{\pi\eta\varsigma^2}\frac{1}{1+\rho\lambda_i}}\exp\left[-\frac{\lambda_i}{\eta\varsigma^2}\frac{1}{1+\rho\lambda_i}x^2\right]. \quad (17)$$

If there exists a negative eigenvalue, this formula does not, in general, parametrize a probability distribution. However, if $\rho > -\frac{1}{\lambda_*}$*, Eq.* (17) *is still the stationary distribution of Eq.*(16)*,* $\forall i \in \{1, \cdots, d\}$.

Theorem 4.2 states that in case the origin is a saddle (or a maximum) and ρ is small enough, the stationary distribution of USAM is divergent at infinity, meaning that the process will escape the bad critical point, which is desirable. In such a case, the escape from the saddle is however slower than SGD as the variance in the direction of negative eigenvalues, e.g. the escape directions, is smaller. However, if ρ is too large, then the dynamics of the USAM SDE will oscillate around the origin even if this is a saddle or a maximum, which is undesirable. This is consistent with the results derived for the SDE of USAM in Section 4.1. There, we found that under the very same condition on ρ, the USAM ODE converges to 0 even when it is a saddle or a maximum.

4.3. SAM ODE

We now provide insights on the dynamics of the SAM ODE on a quadratic with Hessian H.

Lemma 4.3 (Lemma C.6)**.** *For all* $\rho > 0$*, if* H *is PSD (Positive Semi-Definite), the origin is (locally) asymptotically stable. Additionally, if* H *is not PSD and* $\|HX_t\| \leq -\rho\lambda_*$*, then the origin is still (locally) asymptotically stable.*

Lemma 4.3 demonstrates that USAM and SAM have completely different behaviors. For USAM, Lemma 4.1 shows that selecting ρ small enough would prevent the convergence towards a saddle or a maximum. In contrast, Lemma 4.3 shows that for any value of ρ, if the dynamics of SAM is close enough to any critical point, i.e. enters an attractor, it is attracted by it. We also observe that if $\rho \to 0$, this attractor reduces to a single point, i.e. the critical point itself.

To the best of our knowledge, this is the first time that these phenomena are formally demonstrated. Importantly, these theoretical insights are consistent with the experimental results of (Kaddour et al., 2022) that show how SAM might get stuck around saddles.

Finally, by comparing the ODE of USAM (Eq. (15)) with that of SAM (Eq. (149)), we observe that the dynamics of SAM is equivalent to that of USAM where the radius ρ has been scaled by $\frac{1}{\|HX_t\|}$. In a way, while USAM has a fixed radius ρ, SAM has an *time-dependent* radius $\frac{\rho}{\|HX_t\|}$ which is smaller than ρ if the dynamics is far from the origin ($\|HX_t\| > 1$) and larger when it is close to it ($\|HX_t\| < 1$). Therefore, SAM converges to the origin slower than USAM when it is far from it and it becomes faster as it gets closer.

4.4. (DN)SAM SDE

We now provide insights on the dynamics of the DNSAM SDE on a quadratic with Hessian H.

Observation 4.4 (Details in C.7). We observe that for all $\rho > 0$, there exists an $\epsilon > 0$ such that if $\|HX_t\| \in (\epsilon, -\rho\lambda_*)$, the dynamics of X_t is attracted towards the origin. If the eigenvalues are all positive, the condition becomes $\|HX_t\| \in (\epsilon, \infty)$. On the contrary, if $\|HX_t\| < \epsilon$, then the dynamics is pushed away from the origin.

This insight suggests that if DNSAM is initialized close enough to a quadratic saddle, it is attracted toward it, but is also repulsed by it if it gets too close. This is due to the explosion of the volatility next to the origin. We expect that this observation extends to SAM as well, and it remains to be shown theoretically in future work. In the next section, we experimentally verify that the dynamics gets cyclically pulled to 0 and pushed away from it, not only for the quadratic saddle but also for that of other models.

5. Experiments

The main goal of this experimental section is two-fold: 1) to verify the validity of the theorems derived in Section 3, and 2) to validate the claims made about the behavior of SAM and its variants near saddle points. The latter requires us to use models, for which saddle points are known to be present (Safran & Shamir, 2018), including for instance linear autoencoders (Kunin et al., 2019).

5.1. Empirical Validation of the SDEs

We first experimentally validate the results of Corollary 3.3, Corollary 3.6, and Theorem 3.4. To do so, we use two different test functions ($g(x)$ in Def. (3.1)), which are $g_1(x) := \|x\| + \|\nabla f(x)\|$ and $g_2(x) := f(x)$. We test on four models. The first model is a convex quadratic landscape. The second task is a classification one on the Iris Database (Dua & Graff, 2017) using a linear MLP with 1 hidden layer. The third is a classification task on the Breast Cancer Database (Dua & Graff, 2017) using a nonlinear MLP with 1 hidden layer. The fourth is a Teacher-Student model where the Teacher is a linear MLP with 20 hidden layers and the Student is a nonlinear MLP with 20 hidden layers. Figure 2 uses the first metric $g_1(x)$ to measure the maximum absolute error (across the whole trajectory) of the SDEs of SGD, USAM, DNSAM, and SAM in approximating the respective discrete algorithms. Additionally, we plot the same error if we were to use the SDE of SGD to model/approximate the discrete iterates of SAM and its variants. We observe that when $\rho = \eta$, the absolute error is small in all cases, meaning that all the discrete iterates and

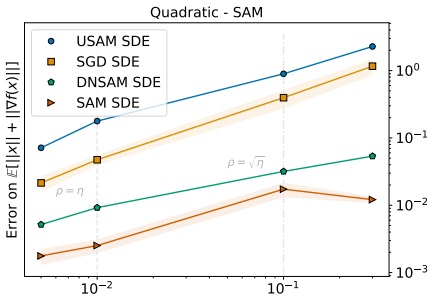

Figure 1. Comparison in terms of $g_1(x)$ with respect to ρ.

SDEs behave essentially in the same way. This supports our claim that if $\rho = \mathcal{O}(\eta)$, the SDE of SGD is a good model for USAM, DNSAM, and SAM (Theorem A.7, Theorem A.11, and Theorem A.16). When $\rho = \sqrt{\eta}$, we see that the derived SDEs correctly approximate the respective discrete algorithms, while the SDE of SGD has a significantly larger relative error, which validates the results of Corollary 3.3, Theorem 3.4, and Corollary 3.6. Although we do not have any theoretical guarantee for larger ρ, we observe empirically that the modeling error is still rather low. Finally, Figure 3 shows the evolution of the metric $g_2(x) := f(x)$ for the different algorithms. We notice that all the SDEs are matching the respective algorithms. In Appendix D.1.1, we provide evidence that failing to include the correct diffusion terms in the USAM SDE Eq. (7) and the DNSAM SDE Eq. (10) leads to less accurate models.

Finally, Figure 1 shows that, in the Quadratic case, DNSAM results in a much closer approximation to SAM than other SDEs. Based on this observation and the analyses of Section 4, we conclude that DNSAM is a better proxy to theoretically study SAM than USAM. It however remains not advised to employ DNSAM as a practical algorithm since its update rule requires the calculation of the full gradient, see Eq. (6).

Interplay between noise, curvature, ρ, and suboptimality Next, we check how the parameter ρ and the curvature (measured by the trace operator of the Hessian) influence the noise of the stochastic process and its suboptimality. These insights substantiate the intuition that SAM and its variants are more likely to escape sharp minima faster than SGD.

First of all, we fix the value of ρ as well as a diagonal Hessian H with random positive eigenvalues. Then, we study the loss for SGD, USAM, DNSAM, and SAM as they optimize a convex quadratic loss of increasingly larger curvature (i.e. larger Hessian magnitude). We observe that DNSAM exhibits a loss that is orders of magnitude larger than that of SGD, with more variance, and even more so as the curvature increases. Note that SAM behaves similarly to DNSAM, but with less variance. Finally, USAM exhibits a similar pattern but less pronounced. All the observations

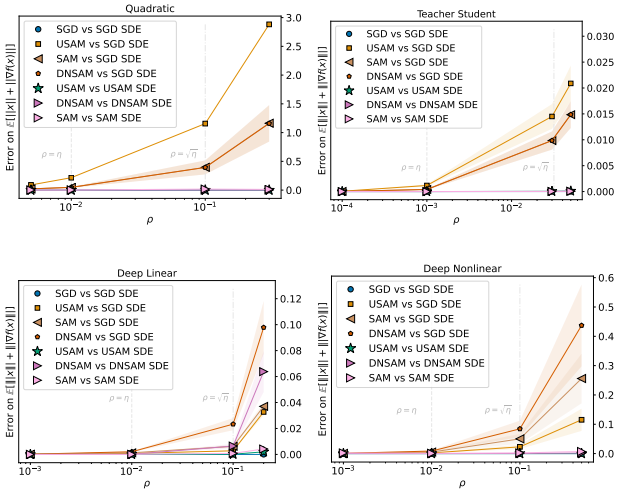

Figure 2. Comparison in terms of $g_1(x)$ with respect to ρ - Quadratic (top left); Teacher-Student (top right); Deep linear class (lower left); Deep Nonlinear class (lower right).

are consistent with the insights gained from the covariance matrices in the respective SDEs. For details, we refer the reader to Figure 8, Figure 9, and Figure 10 in Appendix.

In a similar experiment, we fix the Hessian as above and study the loss as we increase ρ. Once again, we observe that DNSAM exhibits a larger loss with more variance, and this is more and more clear as ρ increases. Observations similar to the above ones can be done for SAM and USAM. For details, we refer the reader to Figure 11, 12 and Figure 13 in Appendix.

Finally, we note that SAM and its variant exhibit an additional implicit curvature-induced noise compared to SGD. This leads to increased suboptimality as well as a higher likelihood to escape sharp minima. We provide an additional justification for this phenomenon in Observation C.5.

5.2. Behavior Near Saddles

In this section, we study the behavior of SAM and USAM (full batch versions), and of PSAM, DNSAM, and PUSAM (perturbed gradient versions) near saddle points. See Table 1 for more details.

Quadratic Landscape We first empirically verify the insight gained in Section 4.4 — the dynamics of DNSAM is attracted to the origin, but if it gets too close, it gets repulsed away. For a quadratic saddle, in Figure 4 we show the distribution of 10^5 trajectories after $5 \cdot 10^4$ iterations. These are distributed symmetrically around the origin but the concentration is lower close to it. While this is intuitive for the convex case (see Figure 14 in Appendix), it is surprising for the saddle case: our insights are fully verified. The second and third images of Figure 4 show that all the trajectories

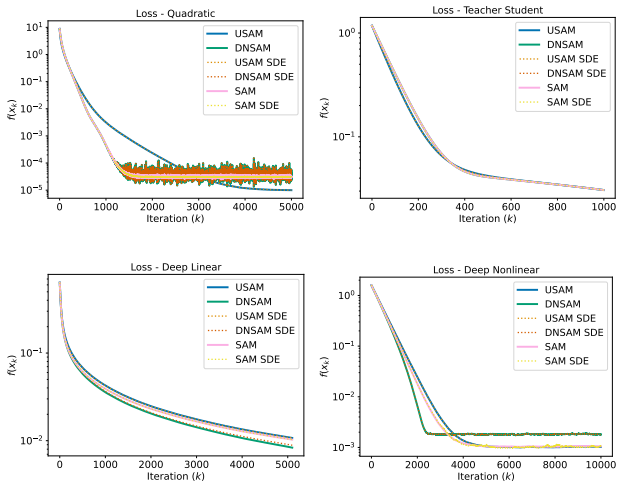

Figure 3. Comparison in terms of $g_2(x)$ with respect to time - Quadratic (top left); Teacher-Student (top right); Deep linear class (lower left); Deep Nonlinear class (lower right).

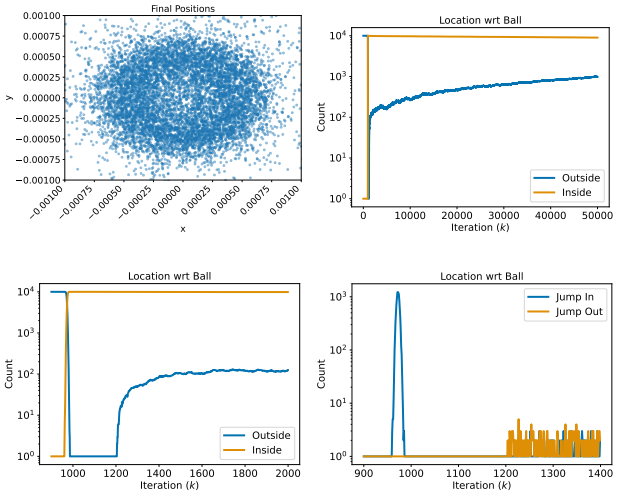

Figure 4. Quadratic Saddle - Top Left: Distribution points around the origin are scarcer near to the origin; Top Right: Number of trajectories outside a small ball around the origin increases over time; Lower Left: All the trajectories eventually enter the ball and then start exiting it; Lower Right: There is a constant oscillation of points in and out of the ball.

are initialized outside of a certain ball around the origin and then they get pulled inside it. Then, we see that the number of points outside this ball increased again and the right-most image shows the number of points jumping in and out of it. This shows that there is a cyclical dynamics towards and away from the origin. Of course, all the points eventually escape the saddle, but much more slowly than what would happen under the dynamics of SGD where the trajectories would not even get close to the origin in the first place. In Figure 15 in Appendix, we show the behavior of several optimizers when initialized in an escaping direction from the saddle and we observe that full-batch SAM is attracted by the saddle while the others are able to escape it. Interestingly, PSAM is anyway slower than SGD in escaping. Figure 15 in Appendix shows that full-batch SAM and PSAM cannot escape the saddle if it is too close to it, while DNSAM can if it is close enough to enjoy a spike in volatility. More details are in the Appendix D.2.

Linear Autoencoder Inspired by the insights gained so far, we study the behavior of SAM when it is initialized close to the saddle present at the origin of the linear autoencoder introduced by (Kunin et al., 2019). The top-left of Figure 5 shows the evolution of the loss as we optimize it with SAM starting from different starting points closer and closer to the saddle in the origin. The scalar σ parametrizes how close the initialization is to the origin. We observe that when SAM starts sufficiently far from it ($\sigma \geq 0.005$), it optimizes immediately, while the closer it is initialized to it, the more it stays around it, up to not being able to move at all ($\sigma \leq 0.001$). Regarding DNSAM, in the top-right figure, we observe the same behavior, apart from one case: if it is initialized sufficiently close to the origin, instead of getting stuck there, it jumps away following a spike in volatility. Differently, PSAM behaves more like SAM and is slower in escaping if σ is lower. The bottom-right of Figure 5 shows the comparison with other optimizers: SAM does not optimize the loss while the other optimizers do. These findings are consistent with those observed in Figure 15 in Appendix for the quadratic landscape. In Figure 16 in Appendix, we show a similar result for a saddle landscape studied in (Lucchi et al., 2022). More details are in Appendix D.3 and Appendix D.4, respectively. In both these experiments, we observe the suboptimality patterns forecasted by our theory.

6. Discussion

6.1. Future Work

Inspired by (Malladi et al., 2022), it would be interesting to study possible scaling rules for SAM and its variants, thus shedding light on the interplay of the learning rate η, the batch size B, and ρ. Another direction could be to use our SDEs to study the role of ρ in balancing the optimization speed and generalization properties of SAM. The insights gained could be useful in improving SAM in terms of optimization speed and generalization. Finally, we expect that the application of more analytical tools to the SDEs of SAM and DNSAM will lead to further insights into SAM. It would for instance be of particular interest to revisit claims made about other optimizers via their SDE models (see "Applications of SDE approximations" in Section 2). Hopefully, this will help to demystify the high performance

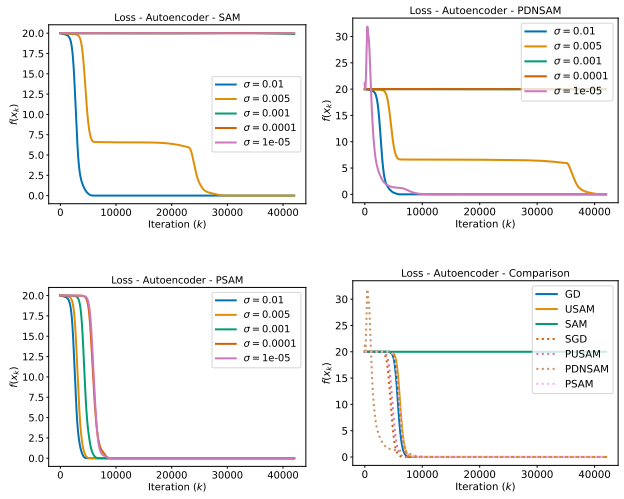

Figure 5. Autoencoder - Top-Left: SAM does not escape the saddle if it is too close to it. Top-Right: DNSAM escapes if it is extremely close to the origin thanks to a volatility spike. Bottom-Left: Like SAM, PSAM does not escape if too close to the origin. Bottom-Right: DNSAM is the fastest to escape, while SAM is stuck.

of SAM on large-scale ML problems.

6.2. Limitations

We highlight that modeling discrete-time algorithms via SDEs relies on Assumption A.3. Furthermore, this setup cannot fully capture the regime of large learning rates. As observed in (Li et al., 2021), a large η or the lack of certain conditions on ∇f and on the noise covariance matrix might lead to an approximation failure. However, the authors claim that this failure could be avoided by increasing the order of the weak approximation. Additionally, most of our discussions are focused on the case where $\rho = \mathcal{O}(\sqrt{\eta})$, which is not the only interesting setup, as some authors use $\rho < \eta$. Finally, since our work is more theoretical in nature, we did not aim at conducting SOTA experiments but rather focused on improving the understanding of the dynamics of SAM. Thus, we analyzed relatively simple models and landscapes that are relevant to the optimization and machine learning community.

6.3. Conclusion

We proposed new continuous-time models (in the form of SDEs) for the SAM optimizer and two variants. While the USAM variant was introduced in prior work (Andriushchenko & Flammarion, 2022), the DNSAM variant we introduce is a step between USAM and SAM, allowing us to gain further insights into the role of the normalization. We formally proved (and experimentally verified) that these SDEs approximate their real discrete-time counterparts; see Theorems 3.2–3.6 for the theory and Section 5.1 for the experiments. An interesting side aspect of our analysis is that DNSAM appears to be a better surrogate model to describe the dynamics of SAM than USAM: SAM and DNSAM share common behaviors around saddles, they have more similar noise structures, and experiments support these claims. Of course, by no means does this paper intend to propose DNSAM as a new practical optimizer: it is instead meant to be used for theoretical analyses.

The SDEs we derived explicitly decompose the learning dynamics (in the parameter space) into a deterministic drift and a stochastic diffusion coefficient which in itself reveals some novel insights: The drift coefficient – by the definition of $\tilde{f}$ in Theorems 3.2–3.6 – exposes how the ascent parameter ρ impacts the average dynamics of SAM and its variants. The diffusion coefficient, on the other hand, increases with the Hessian of the loss – thereby implying that SAM and its variants are noisier in sharp minima. This could be interpreted as an implicit bias towards flat minima (as sharp minima will be more unstable due to the noise).

The continuous-time SDE models allow the application of tools from stochastic calculus (e.g. integration and differentiation) to study the behavior of SAM. As a start in this direction, we proved that the flow of USAM gets stuck around saddles if ρ is too large. In contrast, SAM oscillates around saddles if initialized close to them but eventually slowly escapes them thanks to the additional noise. Importantly, our claims are substantiated by experiments on several models and invite further investigation to prevent a costly waste of computation budget near saddle points.

Acknowledgement We would like to thank the reviewers for their feedback which greatly helped us improve this manuscript. Frank Proske acknowledges the financial support of the Norwegian Research Council (project number 274410). Enea Monzio Compagnoni and Aurelien Lucchi acknowledge the financial support of the Swiss National Foundation, SNF grant No 207392. Hans Kersting thanks the European Research Council for support through the ERC grant 724063.

References

An, J., Lu, J., and Ying, L. Stochastic modified equations for the asynchronous stochastic gradient descent. *Information and Inference: A Journal of the IMA*, 9(4):851–873, 2020.

Andriushchenko, M. and Flammarion, N. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning*, pp. 639–668. PMLR, 2022.

Bahri, D., Mobahi, H., and Tay, Y. Sharpness-aware minimization improves language model generalization. *ACL 2022*, 2022.

Bartlett, P. L., Long, P. M., and Bousquet, O. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. *arXiv preprint arXiv:2210.01513*, 2022.

Chaudhari, P. and Soatto, S. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–10. IEEE, 2018.

Chen, C., Ding, N., and Carin, L. On the convergence of stochastic gradient mcmc algorithms with high-order integrators. *Advances in neural information processing systems*, 28, 2015.

Daneshmand, H., Kohler, J., Lucchi, A., and Hofmann, T. Escaping saddles with stochastic gradients. In *International Conference on Machine Learning*, pp. 1155–1164. PMLR, 2018.

Du, J., Yan, H., Feng, J., Zhou, J. T., Zhen, L., Goh, R. S. M., and Tan, V. Y. Efficient sharpness-aware minimization for improved training of neural networks. *ICLR 2022*, 2022.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Uncertainty in Artificial Intelligence*, 2017.

Folland, G. B. Higher-order derivatives and taylor's formula in several variables. *Preprint*, pp. 1–4, 2005.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *ICLR 2021*, 2021.

Gardiner, C. W. et al. *Handbook of stochastic methods*, volume 3. springer Berlin, 1985.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pp. 797–842, 2015.

Gyöngy, I. and Martínez, T. On stochastic differential equations with locally unbounded drift. *Czechoslovak Mathematical Journal, No. 4, p. 763–783, Vol. 51*, 2001.

He, L., Meng, Q., Chen, W., Ma, Z.-M., and Liu, T.-Y. Differential equations for modeling asynchronous algorithms. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pp. 2220–2226. AAAI Press, 2018.

Helmke, U. and Moore, J. B. *Optimization and Dynamical Systems*. Springer London, 1st edition, 1994.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three factors influencing minima in sgd. *ICANN 2018*, 2018.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2019.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pp. 1724–1732. PMLR, 2017.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *J. ACM*, 68 (2), 2021.

Kaddour, J., Liu, L., Silva, R., and Kusner, M. J. When do flat minima optimizers work? *Advances in Neural Information Processing Systems*, 35:16577–16595, 2022.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *ICLR 2017*, 2017.

Kim, H., Park, J., Choi, Y., and Lee, J. Stability analysis of sharpness-aware minimization. *arXiv preprint arXiv:2301.06308*, 2023.

Kunin, D., Bloom, J., Goeva, A., and Seed, C. Loss landscapes of regularized linear autoencoders. In *International Conference on Machine Learning*, pp. 3560–3569. PMLR, 2019.

Kunin, D., Sagastuy-Brena, J., Ganguli, S., Yamins, D. L., and Tanaka, H. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations*, 2021.

Kushner, H. and Yin, G. G. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Kwon, J., Kim, J., Park, H., and Choi, I. K. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pp. 5905–5914. PMLR, 2021.

Levy, K. Y. The power of normalization: Faster evasion of saddle points. *arXiv preprint arXiv:1611.04831*, 2016.

Li, Q., Tai, C., and Weinan, E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pp. 2101–2110. PMLR, 2017.

Li, Q., Tai, C., and Weinan, E. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, 20(1):1474–1520, 2019.

Li, Z., Lyu, K., and Arora, S. Reconciling modern deep learning with traditional optimization analyses: the intrinsic learning rate. In *Advances in Neural Information Processing Systems*, 2020.

Li, Z., Malladi, S., and Arora, S. On the validity of modeling SGD with stochastic differential equations (SDEs). In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Ljung, L., Pflug, G., and Walk, H. *Stochastic approximation and optimization of random systems*, volume 17. Birkhäuser, 2012.

Lucchi, A., Proske, F., Orvieto, A., Bach, F., and Kersting, H. On the theoretical properties of noise correlation in stochastic optimization. In *Advances in Neural Information Processing Systems*, 2022.

Malladi, S., Lyu, K., Panigrahi, A., and Arora, S. On the SDEs and scaling rules for adaptive gradient algorithms. In *Advances in Neural Information Processing Systems*, 2022.

Mandt, S., Hoffman, M. D., Blei, D. M., et al. Continuous-time limit of stochastic gradient descent revisited. *NIPS-2015*, 2015.

Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic gradient descent as approximate Bayesian inference. *J. Mach. Learn. Res.*, 18(1):4873–4907, 2017a. ISSN 1532-4435.

Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic gradient descent as approximate bayesian inference. *JMLR 2017*, 2017b.

Mao, X. *Stochastic differential equations and applications*. Elsevier, 2007.

Mil'shtein, G. Weak approximation of solutions of systems of stochastic differential equations. *Theory of Probability & Its Applications*, 30(4):750–766, 1986.

Orvieto, A. and Lucchi, A. Continuous-time models for stochastic optimization algorithms. *Advances in Neural Information Processing Systems*, 32, 2019.

Panigrahi, A., Somani, R., Goyal, N., and Netrapalli, P. Non-Gaussianity of stochastic gradient noise. *SEDL workshop at NeurIPS 2019*, 2019.

Poggio, T., Kawaguchi, K., Liao, Q., Miranda, B., Rosasco, L., Boix, X., Hidary, J., and Mhaskar, H. Theory of deep learning iii: explaining the non-overfitting puzzle. *arXiv preprint arXiv:1801.00173*, 2017.

Rangwani, H., Aithal, S. K., Mishra, M., and Babu, R. V. Escaping saddle points for effective generalization on class-imbalanced data. *NeurIPS 2022*, 2022.

Risken, H. Fokker-planck equation. In *The Fokker-Planck Equation*, pp. 63–95. Springer, 1996.

Safran, I. and Shamir, O. Spurious local minima are common in two-layer relu neural networks. In *International Conference on Machine Learning*, pp. 4433–4441. PMLR, 2018.

Sagun, L., Evci, U., Guney, V. U., Dauphin, Y., and Bottou, L. Empirical analysis of the hessian of over-parametrized neural networks. *ICLR 2018 Workshop Track*, 2018. URL `https://openreview.net/forum?id=rJrTwxbCb`.

Simsekli, U., Sagun, L., and Gurbuzbalaban, M. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, 2019.

Smith, S., Elsen, E., and De, S. On the generalization benefit of noise in stochastic gradient descent. In *International Conference on Machine Learning*, 2020.

Su, W., Boyd, S., and Candes, E. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, 2014.

Ujváry, S., Telek, Z., Kerekes, A., Mészáros, A., and Huszár, F. Rethinking sharpness-aware minimization as variational inference. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

Wen, K., Ma, T., and Li, Z. How does sharpness-aware minimization minimize sharpness? *ICLR 2023*, 2023.

Xie, Z., Sato, I., and Sugiyama, M. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2021.

Zhao, J., Lucchi, A., Proske, F. N., Orvieto, A., and Kersting, H. Batch size selection by stochastic optimal control. In *Has it Trained Yet? NeurIPS 2022 Workshop*, 2022.

Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S. C. H., et al. Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Advances in Neural Information Processing Systems*, 33:21285–21296, 2020.

Zhu, Z., Wu, J., Yu, B., Wu, L., and Ma, J. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. *ICML 2019*, 2019.

Zhuang, J., Gong, B., Yuan, L., Cui, Y., Adam, H., Dvornek, N., Tatikonda, S., Duncan, J., and Liu, T. Surrogate gap minimization improves sharpness-aware training. *ICML 2022*, 2022.

A. Theoretical Framework - SDEs

In the subsequent proofs, we will make repeated use of Taylor expansions in powers of η. To simplify the presentation, we introduce the shorthand that whenever we write $\mathcal{O}(\eta^\alpha)$, we mean that there exists a function $K(x) \in G$ such that the error terms are bounded by $K(x)\eta^\alpha$. For example, we write

$$b(x+\eta) = b_0(x) + \eta b_1(x) + \mathcal{O}\left(\eta^2\right)$$

to mean: there exists $K \in G$ such that

$$|b(x+\eta) - b_0(x) - \eta b_1(x)| \leq K(x)\eta^2.$$

Additionally, let us introduce some notation:

- A multi-index is $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ such that $\alpha_j \in \{0, 1, 2, \ldots\}$
- $|\alpha| := \alpha_1 + \alpha_2 + \cdots + \alpha_n$
- $\alpha! := \alpha_1!\alpha_2! \cdots \alpha_n!$
- For $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, we define $x^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$
- For a multi-index β, $\partial_\beta^{|\beta|} f(x) := \frac{\partial^{|\beta|}}{\partial_{x_1}^{\beta_1} \partial_{x_2}^{\beta_2} \cdots \partial_{x_n}^{\beta_n}} f(x)$
- We also denote the partial derivative with respect to x_i by ∂_{e_i}.

Lemma A.1 (Lemma 1 (Li et al., 2017)). *Let $0 < \eta < 1$. Consider a stochastic process $X_t, t \geq 0$ satisfying the SDE*

$$dX_t = b(X_t)\, dt + \eta^{\frac{1}{2}} \sigma(X_t)\, dW_t$$

with $X_0 = x \in \mathbb{R}^d$ and b, σ together with their derivatives belong to G. Define the one-step difference $\Delta = X_\eta - x$, then we have

1. $\mathbb{E}\Delta_i = b_i\eta + \frac{1}{2}\left[\sum_{j=1}^d b_j \partial_{e_j} b_i\right]\eta^2 + \mathcal{O}\left(\eta^3\right) \quad \forall i = 1, \ldots, d;$
2. $\mathbb{E}\Delta_i\Delta_j = \left[b_i b_j + \sigma\sigma^T_{(ij)}\right]\eta^2 + \mathcal{O}\left(\eta^3\right) \quad \forall i, j = 1, \ldots, d;$
3. $\mathbb{E}\prod_{j=1}^s \Delta_{(i_j)} = \mathcal{O}\left(\eta^3\right)$ *for all* $s \geq 3, i_j = 1, \ldots, d.$

All functions above are evaluated at x.

Theorem A.2 (Theorem 2 and Lemma 5, (Mil'shtein, 1986)). *Let the assumptions in Theorem A.4 hold and let us define $\bar{\Delta} = x_1 - x$ to be the increment in the discrete-time algorithm. If in addition there exists $K_1, K_2, K_3, K_4 \in G$ so that*

1. $\left|\mathbb{E}\Delta_i - \mathbb{E}\bar{\Delta}_i\right| \leq K_1(x)\eta^2, \quad \forall i = 1, \ldots, d;$
2. $\left|\mathbb{E}\Delta_i\Delta_j - \mathbb{E}\bar{\Delta}_i\bar{\Delta}_j\right| \leq K_2(x)\eta^2, \quad \forall i, j = 1, \ldots, d;$
3. $\left|\mathbb{E}\prod_{j=1}^s \Delta_{i_j} - \mathbb{E}\prod_{j=1}^s \bar{\Delta}_{i_j}\right| \leq K_3(x)\eta^2, \quad \forall s \geq 3, \quad \forall i_j \in \{1, \ldots, d\};$
4. $\mathbb{E}\prod_{j=1}^3 \left|\bar{\Delta}_{i_j}\right| \leq K_4(x)\eta^2, \quad \forall i_j \in \{1, \ldots, d\}.$

Then, there exists a constant C so that for all $k = 0, 1, \ldots, N$ we have

$$|\mathbb{E}g(X_{k\eta}) - \mathbb{E}g(x_k)| \leq C\eta.$$

Before starting, we remind the reader that the order 1 weak approximation SDE of SGD (see Li et al. (2017; 2019)) is given by

$$dX_t = -\nabla f(X_t)dt + \sqrt{\eta}\left(\Sigma^{\text{SGD}}(X_t)\right)^{\frac{1}{2}} dW_t \tag{18}$$

where $\Sigma^{\text{SGD}}(x)$ is the SGD covariance matrix defined as

$$\mathbb{E}\left[\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)^T\right]. \tag{19}$$

A.1. Formal Derivation - USAM

The next result is inspired by Theorem 1 of (Li et al., 2017) and is derived under some regularity assumption on the function f.

Assumption A.3. Assume that the following conditions on f, f_i and their gradients are satisfied:

- $\nabla f, \nabla f_i$ satisfy a Lipschitz condition: there exists $L > 0$ such that

$$|\nabla f(x) - \nabla f(y)| + \sum_{i=1}^{n} |\nabla f_i(x) - \nabla f_i(y)| \leq L|x - y|;$$

- f, f_i and its partial derivatives up to order 7 belong to G;
- $\nabla f, \nabla f_i$ satisfy a growth condition: there exists $M > 0$ such that

$$|\nabla f(x)| + \sum_{i=1}^{n} |\nabla f_i(x)| \leq M(1 + |x|);$$

We will consider the stochastic process $X_t \in \mathbb{R}^d$ defined by

$$dX_t = -\nabla \tilde{f}^{\text{USAM}}(X_t)dt + \sqrt{\eta}\left(\Sigma^{\text{SGD}}(X_t) + \rho\left(\tilde{\Sigma}(X_t) + \tilde{\Sigma}(X_t)^\top\right)\right)^{\frac{1}{2}} dW_t \tag{20}$$

where

$$\Sigma^{\text{SGD}}(x) := \mathbb{E}\left[\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)^T\right].$$

is the usual covariance of SGD, while

$$\tilde{\Sigma}(x) := \mathbb{E}\left[\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)\left(\mathbb{E}\left[\nabla^2 f_\gamma(x)\nabla f_\gamma(x)\right] - \nabla^2 f_\gamma(x)\nabla f_\gamma(x)\right)^\top\right] \tag{21}$$

and

$$\tilde{f}^{\text{USAM}}(x) := f(x) + \frac{\rho}{2}\mathbb{E}\left[\|\nabla f_\gamma(x)\|_2^2\right].$$

In the following, we will use the notation

$$\Sigma^{\text{USAM}}(x) := \left(\Sigma^{\text{SGD}}(X_t) + \rho\left(\tilde{\Sigma}(X_t) + \tilde{\Sigma}(X_t)^\top\right)\right) \tag{22}$$

Theorem A.4 (Stochastic modified equations)**.** *Let $0 < \eta < 1, T > 0$ and set $N = \lfloor T/\eta \rfloor$. Let $x_k \in \mathbb{R}^d, 0 \leq k \leq N$ denote a sequence of USAM iterations defined by Eq.* (5)*. Additionally, let us take*

$$\rho = \mathcal{O}\left(\eta^{\frac{1}{2}}\right). \tag{23}$$

Consider the stochastic process X_t defined in Eq. (20) *and fix some test function $g \in G$ and suppose that g and its partial derivatives up to order 6 belong to G.*
Then, under Assumption A.3, there exists a constant $C > 0$ independent of η such that for all $k = 0, 1, \ldots, N$, we have

$$|\mathbb{E}g\left(X_{k\eta}\right) - \mathbb{E}g\left(x_k\right)| \leq C\eta^1.$$

That is, the SDE (20) *is an order* 1 *weak approximation of the SAM iterations* (5)*.*

Lemma A.5. *Under the assumptions of Theorem A.4, let $0 < \eta < 1$ and consider $x_k, k \geq 0$ satisfying the USAM iterations* (5)

$$x_{k+1} = x_k - \eta \nabla f_{\gamma_k}\left(x_k + \rho \nabla f_{\gamma_k}(x_k)\right)$$

with $x_0 = x \in \mathbb{R}^d$. Additionally, we define $\partial_{e_i}\tilde{f}^{USAM}(x) := \partial_{e_i} f(x) + \rho \mathbb{E}\left[\sum_j \partial^2_{e_i+e_j} f_\gamma(x) \partial_{e_j} f_\gamma(x)\right]$. From the definition the one-step difference $\bar{\Delta} = x_1 - x$, then we have

1. $\mathbb{E}\bar{\Delta}_i = -\partial_{e_i}\tilde{f}^{USAM}(x)\eta + \mathcal{O}(\eta\rho^2) \quad \forall i = 1, \ldots, d.$
2. $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j = \partial_{e_i}\tilde{f}^{USAM}(x)\partial_{e_j}\tilde{f}^{USAM}(x)\eta^2 + \Sigma^{USAM}_{(ij)}\eta^2 + \mathcal{O}\left(\eta^3\right) \quad \forall i, j = 1, \ldots, d.$
3. $\mathbb{E}\prod_{j=1}^s \bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right) \quad \forall s \geq 3, \quad i_j \in \{1, \ldots, d\}.$

and all the functions above are evaluated at x.

Together with many other proofs in this Section, the following one relies on a Taylor expansion. The truncated terms are multiplied by a mixture of terms in η and ρ. Therefore, a careful balancing of the relative size of these two quantities is needed as reflected in Equation (23).

Proof of Lemma A.5. Since the first step is to evaluate $\mathbb{E}\Delta_i = \mathbb{E}\left[-\partial_{e_i} f_\gamma(x + \rho\nabla f_\gamma(x))\eta\right]$, we start by analyzing $\partial_{e_i} f_\gamma(x + \rho\nabla f_\gamma(x))$, that is the partial derivative in the direction $e_i := (0, \cdots, 0, \underset{i-th}{1}, 0, \cdots, 0)$. Then, we have that

$$\partial_{e_i} f_\gamma(x + \rho\nabla f_\gamma(x)) = \partial_{e_i} f_\gamma(x) + \sum_{|\alpha|=1} \partial^2_{e_i+\alpha} f_\gamma(x)\rho\partial_\alpha f_\gamma(x) + \mathcal{R}^{\partial_{e_i} f_\gamma(x)}_{x,1}(\rho\nabla f_\gamma(x)), \tag{24}$$

where the residual is defined in Eq. (4) of (Folland, 2005). Therefore, for some constant $c \in (0, 1)$, it holds that

$$\mathcal{R}^{\partial_{e_i} f_\gamma(x)}_{x,1}(\rho\nabla f_\gamma(x)) = \sum_{|\alpha|=2} \frac{\partial^3_{e_i+\alpha} f_\gamma(x + c\rho\nabla f_\gamma(x))\rho^2 \left(\nabla f_\gamma(x)\right)^\alpha}{\alpha!}. \tag{25}$$

Therefore, we can rewrite it as

$$\partial_{e_i} f_\gamma(x + \rho\nabla f_\gamma(x)) = \partial_{e_i} f_\gamma(x) + \rho\sum_j \partial^2_{e_i+e_j} f_\gamma(x)\partial_{e_j} f_\gamma(x) + \rho^2\left[\sum_{|\alpha|=2} \frac{\partial^3_{e_i+\alpha} f_\gamma(x + c\rho\nabla f_\gamma(x))\left(\nabla f_\gamma(x)\right)^\alpha}{\alpha!}\right] \tag{26}$$

Now, we observe that

$$K_i(x) := \left[\sum_{|\alpha|=2} \frac{\partial^3_{e_i+\alpha} f_\gamma(x + c\rho\nabla f_\gamma(x)) \left(\nabla f_\gamma(x)\right)^\alpha}{\alpha!} \right] \tag{27}$$

is a finite sum of products of functions that by assumption are in G. Therefore, $K_i(x) \in G$ and$\bar{K}_i(x) = \mathbb{E}\left[K_i(x)\right] \in G$. Based on these definitions, we rewrite Eq. (26) as

$$\partial_{e_i} f_\gamma(x + \rho\nabla f_\gamma(x)) = \partial_{e_i} f_\gamma(x) + \rho \sum_j \partial^2_{e_i+e_j} f_\gamma(x) \partial_{e_j} f_\gamma(x) + \rho^2 K_i(x). \tag{28}$$

which implies that

$$\mathbb{E}\left[\partial_{e_i} f_\gamma(x + \rho\nabla f_\gamma(x))\right] = \partial_{e_i} f(x) + \rho\mathbb{E}\left[\sum_j \partial^2_{e_i+e_j} f_\gamma(x) \partial_{e_j} f_\gamma(x) \right] + \rho^2 \bar{K}_i(x). \tag{29}$$

Let us now remember that

$$\partial_{e_i} \tilde{f}^{\text{USAM}}(x) = \partial_{e_i} \left(f(x) + \frac{\rho}{2}\mathbb{E}\left[\|\nabla f_\gamma(x)\|_2^2\right] \right) = \partial_{e_i} f(x) + \rho\mathbb{E}\left[\sum_j \partial^2_{e_i+e_j} f_\gamma(x) \partial_{e_j} f_\gamma(x) \right] \tag{30}$$

Therefore, by using Eq. (29), Eq. (30), and the assumption (23) we have that $\forall i = 1, \ldots, d$

$$\mathbb{E}\bar{\Delta}_i = -\partial_{e_i} \tilde{f}^{\text{USAM}}(x)\eta + \eta\rho^2 \bar{K}_i(x) = -\partial_{e_i} \tilde{f}^{\text{USAM}}(x)\eta + \mathcal{O}\left(\eta^2\right). \tag{31}$$

Additionally, we have that

$$\begin{aligned}
\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j =& \text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) + \mathbb{E}\bar{\Delta}_i\mathbb{E}\bar{\Delta}_j \\
&\overset{(31)}{=} \text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) + \partial_{e_i} \tilde{f}^{\text{USAM}} \partial_{e_j} \tilde{f}^{\text{USAM}} \eta^2 + \eta^2\rho^2 (\partial_{e_i} \tilde{f}^{\text{USAM}} \bar{K}_j(x) + \partial_{e_j} \tilde{f}^{\text{USAM}} \bar{K}_i(x)) + \eta^2\rho^4 \bar{K}_i(x)\bar{K}_j(x) \\
&= \text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) + \partial_{e_i} \tilde{f}^{\text{USAM}} \partial_{e_j} \tilde{f}^{\text{USAM}} \eta^2 + \mathcal{O}\left(\eta^2\rho^2\right) + \mathcal{O}\left(\eta^2\rho^4\right) \\
&= \partial_{e_i} \tilde{f}^{\text{USAM}} \partial_{e_j} \tilde{f}^{\text{USAM}} \eta^2 + \text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) + \mathcal{O}\left(\eta^2\rho^2\right) + \mathcal{O}\left(\eta^2\rho^4\right) \quad \forall i, j = 1, \ldots, d
\end{aligned} \tag{32}$$

Let us now recall the expression (21) of $\tilde{\Sigma}$ and the expression (22) of Σ^{USAM}. Then, we automatically have that

$$\text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) = \eta^2 \left(\Sigma^{\text{SGD}}_{i,j}(x) + \rho \left[\tilde{\Sigma}_{i,j}(x) + \tilde{\Sigma}_{i,j}(x)^\top \right] + \mathcal{O}(\rho^2) \right) = \eta^2 \Sigma^{\text{USAM}}_{i,j}(x) + \mathcal{O}(\eta^2\rho^2) \tag{33}$$

Therefore, remembering Eq. (32) and Eq. (23) we have

$$\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j = \partial_{e_i} \tilde{f}^{\text{USAM}} \partial_{e_j} \tilde{f}^{\text{USAM}} \eta^2 + \Sigma^{\text{USAM}}_{i,j} \eta^2 + \mathcal{O}\left(\eta^3\right), \quad \forall i, j = 1, \ldots, d \tag{34}$$

Finally, with analogous considerations, it is obvious that under our assumptions

$$\mathbb{E}\prod_{j=1}^{s} \bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^s\right) \quad \forall s \geq 3, \quad i_j \in \{1, \ldots, d\}$$

which in particular implies that

$$\mathbb{E}\prod_{j=1}^{3} \bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right), \quad i_j \in \{1, \ldots, d\}.$$

□

Additional Insights from Lemma A.5. Let us notice that $\nabla f_\gamma(x)$ is dominated by a factor $M(1+|x|)$, if all $\partial^3_{e_i+\alpha} f_\gamma(x)$ are limited by a common constant L, for some positive constant C we have that

$$|K_i(x)| = \rho^2 \left| \sum_{|\alpha|=2} \frac{\partial^3_{e_i+\alpha} f_\gamma(x + c\rho\nabla f_\gamma(x)) \left(\nabla f_\gamma(x)\right)^\alpha}{\alpha!} \right| \tag{35}$$

$$\leq \rho^2 CL \|\nabla f_\gamma(x) \nabla f_\gamma(x)^\top\|_F^2 \leq \rho^2 CLd^2 M^2 (1+|x|)^2 \tag{36}$$

Therefore, $K_i(x)$ does not only lay in G, but has at most quadratic growth.

Proof of Theorem A.4. To prove this result, all we need to do is check the conditions in Theorem A.2. As we apply Lemma A.1, we make the following choices:

- $b(x) = -\nabla \tilde{f}^{\text{USAM}}(x)$;
- $\sigma(x) = \Sigma^{\text{USAM}}(x)^{\frac{1}{2}}$.

First of all, we notice that $\forall i = 1, \ldots, d$, it holds that

- $\mathbb{E}\bar{\Delta}_i \overset{\text{1. Lemma A.5}}{=} -\partial_{e_i} \tilde{f}^{\text{USAM}}(x)\eta + \mathcal{O}(\eta^2)$;
- $\mathbb{E}\Delta_i \overset{\text{1. Lemma A.1}}{=} -\partial_{e_i} \tilde{f}^{\text{USAM}}(x)\eta + \mathcal{O}\left(\eta^2\right)$.

Therefore, we have that for some $K_1(x) \in G$

$$\left|\mathbb{E}\Delta_i - \mathbb{E}\bar{\Delta}_i\right| \leq K_1(x)\eta^2, \quad \forall i = 1, \ldots, d. \tag{37}$$

Additionally,we notice that $\forall i, j = 1, \ldots, d$, it holds that

- $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j \overset{\text{2. Lemma A.5}}{=} \partial_{e_i}\tilde{f}^{\text{USAM}}\partial_{e_j}\tilde{f}^{\text{USAM}}\eta^2 + \Sigma^{\text{USAM}}_{i,j}\eta^2 + \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\Delta_i\Delta_j \overset{\text{2. Lemma A.1}}{=} \partial_{e_i}\tilde{f}^{\text{USAM}}\partial_{e_j}\tilde{f}^{\text{USAM}}\eta^2 + \Sigma^{\text{USAM}}_{i,j}\eta^2 + \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_2(x) \in G$

$$\left|\mathbb{E}\Delta_i\Delta_j - \mathbb{E}\bar{\Delta}_i\bar{\Delta}_j\right| \leq K_2(x)\eta^2, \quad \forall i, j = 1, \ldots, d \tag{38}$$

Additionally, we notice that $\forall s \geq 3, \forall i_j \in \{1, \ldots, d\}$, it holds that

- $\mathbb{E}\prod_{j=1}^s \bar{\Delta}_{i_j} \overset{\text{3. Lemma A.5}}{=} \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\prod_{j=1}^s \Delta_{i_j} \overset{\text{3. Lemma A.1}}{=} \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_3(x) \in G$

$$\left|\mathbb{E}\prod_{j=1}^s \Delta_{i_j} - \mathbb{E}\prod_{j=1}^s \bar{\Delta}_{i_j}\right| \leq K_3(x)\eta^2. \tag{39}$$

Additionally, for some $K_4(x) \in G, \forall i_j \in \{1, \ldots, d\}$

$$\mathbb{E}\prod_{j=1}^{3}\left|\bar{\Delta}_{(i_j)}\right| \overset{\text{3. Lemma A.5}}{\leq} K_4(x)\eta^2. \tag{40}$$

Finally, Eq. (37), Eq. (38), Eq. (39), and Eq. (40) allow us to conclude the proof.

□

Corollary A.6. *Let us take the same assumptions of Theorem A.4. Additionally, let us assume that the dynamics is near the minimizer. In this case, the noise structure is such that the stochastic gradient can be written as* $\nabla f_\gamma(x) = \nabla f(x) + Z$ *such that* Z *is the noise that does not depend on* x*. Therefore, the SDE* (20) *becomes*

$$dX_t = -\nabla \tilde{f}^{USAM}(X_t)dt + \left(I_d + \rho\nabla^2 f(X_t)\right)\left(\eta\Sigma^{SGD}(X_t)\right)^{\frac{1}{2}} dW_t \tag{41}$$

where

$$\Sigma^{SGD}(x) := \mathbb{E}\left[\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)^T\right]$$

is the usual covariance of SGD, and

$$\tilde{f}^{USAM}(x) = f(x) + \frac{\rho}{2}\|\nabla f(x)\|_2^2.$$

Proof of Corollary A.6. Based on our assumption on the noise structure, we can rewrite Eq. (21) of the matrix $\tilde{\Sigma}$ as

$$\tilde{\Sigma}(x) = \mathbb{E}\left[\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)\left(\mathbb{E}\left[\nabla^2 f_\gamma(x)\nabla f_\gamma(x)\right] - \nabla^2 f_\gamma(x)\nabla f_\gamma(x)\right)^\top\right] \tag{42}$$

$$= \nabla^2 f(x)\mathbb{E}\left[\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)\left(\nabla f(x) - \nabla f_\gamma(x)\right)^\top\right] \tag{43}$$

Therefore, the Eq. (22) of the covariance Σ^{USAM} becomes

$$\Sigma^{\text{USAM}}(x) = \left(I_d + 2\rho\nabla^2 f(x)\right)\Sigma^{\text{SGD}}(X_t) \tag{44}$$

which implies that

$$\left(\Sigma^{\text{USAM}}(x)\right)^{\frac{1}{2}} \approx \left(I_d + \rho\nabla^2 f(x)\right)\left(\Sigma^{\text{SGD}}(X_t)\right)^{\frac{1}{2}}. \tag{45}$$

Finally, we have that

$$\tilde{f}^{\text{USAM}}(x) := f(x) + \frac{\rho}{2}\mathbb{E}\left[\|\nabla f_\gamma(x)\|_2^2\right] = f(x) + \frac{\rho}{2}\mathbb{E}\left[\|\nabla f(x)\|_2^2 + Z^2 + 2Z\nabla f_\gamma(x)\right] \tag{46}$$

$$= f(x) + \frac{\rho}{2}\|\nabla f(x)\|_2^2 + \frac{\rho}{2}\mathbb{E}\left[Z^2\right] \tag{47}$$

Since the component $\mathbb{E}\left[Z^2\right]$ is independent on x, we ignore it and conclude that

$$\tilde{f}^{\text{USAM}}(x) = f(x) + \frac{\rho}{2}\|\nabla f(x)\|_2^2.$$

□

A.1.1. USAM IS SGD IF $\rho = \mathcal{O}(\eta)$

The following result is inspired by Theorem 1 of (Li et al., 2017).

Theorem A.7 (Stochastic modified equations)**.** *Let $0 < \eta < 1, T > 0$ and set $N = \lfloor T/\eta \rfloor$. Let $x_k \in \mathbb{R}^d, 0 \leq k \leq N$ denote a sequence of USAM iterations defined by Eq.* (5)*. Additionally, let us take*

$$\rho = \mathcal{O}\left(\eta^1\right). \tag{48}$$

Define $X_t \in \mathbb{R}^d$ as the stochastic process satisfying the SDE

$$dX_t = -\nabla f\left(X_t\right) dt + \left(\eta \Sigma^{SGD}\left(X_t\right)\right)^{1/2} dW_t \tag{49}$$

Such that $X_0 = x_0$ and

$$\Sigma^{SGD}(x) := \mathbb{E}\left[\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)^T\right]$$

Fix some test function $g \in G$ and suppose that g and its partial derivatives up to order 6 belong to G. Then, under Assumption A.3, there exists a constant $C > 0$ independent of η such that for all $k = 0, 1, \ldots, N$, we have

$$\left|\mathbb{E}g\left(X_{k\eta}\right) - \mathbb{E}g\left(x_k\right)\right| \leq C\eta^1.$$

That is, the SDE (49) *is an order* 1 *weak approximation of the USAM iterations* (5)*.*

Lemma A.8. *Under the assumptions of Theorem A.7, let $0 < \eta < 1$. Consider $x_k, k \geq 0$ satisfying the USAM iterations*

$$x_{k+1} = x_k - \eta \nabla f_{\gamma_k}\left(x_k + \rho \nabla f_{\gamma_k}(x_k)\right)$$

with $x_0 = x \in \mathbb{R}^d$. From the definition the one-step difference $\bar{\Delta} = x_1 - x$, then we have

1. $\mathbb{E}\bar{\Delta}_i = -\partial_{e_i} f(x)\eta + \mathcal{O}(\eta^2) \quad \forall i = 1, \ldots, d.$
2. $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j = \partial_{e_i} f \partial_{e_j} f \eta^2 + \Sigma^{SGD}_{(ij)}\eta^2 + \mathcal{O}\left(\eta^3\right) \quad \forall i, j = 1, \ldots, d.$
3. $\mathbb{E}\prod_{j=1}^s \bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right) \quad \forall s \geq 3, \quad i_j \in \{1, \ldots, d\}.$

All functions above are evaluated at x.

Proof of Lemma A.8. First of all, we write that

$$\partial_{e_i} f_\gamma(x + \rho \nabla f_\gamma(x)) = \partial_{e_i} f_\gamma(x) + \mathcal{R}^{\partial_{e_i} f_\gamma(x)}_{x,0}(\rho \nabla f_\gamma(x)), \tag{50}$$

where the residual is defined in Eq. (4) of (Folland, 2005). Therefore, for some constant $c \in (0, 1)$, it holds that

$$\mathcal{R}^{\partial_{e_i} f_\gamma(x)}_{x,0}(\rho \nabla f_\gamma(x)) = \sum_{|\alpha|=1} \frac{\partial^2_{e_i+\alpha} f_\gamma(x + c\rho \nabla f_\gamma(x)) \rho^1 \left(\nabla f_\gamma(x)\right)^\alpha}{\alpha!}. \tag{51}$$

Let us now observe that $\mathcal{R}^{\partial_{e_i} f_\gamma(x)}_{x,0}(\rho \nabla f_\gamma(x))$ is a finite sum of products of functions in G and that, therefore, it lies in G. Additionally, given its expression Eq. (51), we can factor out a common ρ and have that $K(x) = \rho K_1(x)$ for some function $K_1(x) \in G$. Therefore, we rewrite Eq. (50) as

$$\partial_{e_i} f_\gamma(x + \rho \nabla f_\gamma(x)) = \partial_{e_i} f_\gamma(x) + \rho K_1(x). \tag{52}$$

First of all, we notice that if we define $\bar{K}_1(x) = \mathbb{E}\left[K_1(x)\right]$, also $\bar{K}_1(x) \in G$. Therefore, it holds that

$$\mathbb{E}\left[\partial_{e_i} f_\gamma(x + \rho \nabla f_\gamma(x))\right] \overset{(52)}{=} \partial_{e_i} f(x) + \rho \bar{K}_1(x) \tag{53}$$

Therefore, using assumption (48), $\forall i = 1, \dots, d$, we have that

$$\mathbb{E}\bar{\Delta}_i = -\partial_{e_i} f(x)\eta + \eta\rho\bar{K}_i(x) = -\partial_{e_i} f(x)\eta + \mathcal{O}\left(\eta^2\right) \tag{54}$$

Additionally, by keeping in mind the definition of the covariance matrix Σ, We immediately have

$$\begin{aligned}
\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j &\overset{(52)}{=} Cov(\bar{\Delta}_i, \bar{\Delta}_j) + \mathbb{E}\bar{\Delta}_i\mathbb{E}\bar{\Delta}_j \\
&= \Sigma^{\text{SGD}}_{(ij)}\eta^2 + \partial_{e_i} f \partial_{e_j} f \eta^2 + \eta^2\rho(\partial_{e_i} f \bar{K}_j(x) + \partial_{e_j} f \bar{K}_i(x)) + \eta^2\rho^2 \bar{K}_i(x)\bar{K}_j(x) \\
&= \Sigma^{\text{SGD}}_{(ij)}\eta^2 + \partial_{e_i} f \partial_{e_j} f \eta^2 + \mathcal{O}\left(\eta^2\rho\right) + \mathcal{O}\left(\eta^2\rho^2\right) \\
&= \partial_{e_i} f \partial_{e_j} f \eta^2 + \Sigma^{\text{SGD}}_{(ij)}\eta^2 + \mathcal{O}\left(\eta^3\right) \quad \forall i, j = 1, \dots, d
\end{aligned} \tag{55}$$

Finally, with analogous considerations, it is obvious that under our assumptions

$$\mathbb{E}\prod_{j=1}^{s} \bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right) \quad \forall s \geq 3, \quad i_j \in \{1, \dots, d\}.$$

□

Proof of Theorem A.7. To prove this result, all we need to do is check the conditions in Theorem A.2. As we apply Lemma A.1, we make the following choices:

- $b(x) = -\nabla f\left(x\right)$,
- $\sigma(x) = \Sigma^{\text{SGD}}(X_t)^{\frac{1}{2}}$;

First of all, we notice that $\forall i = 1, \dots, d$, it holds that

- $\mathbb{E}\bar{\Delta}_i \overset{\text{1. Lemma A.8}}{=} -\partial_{e_i} f(x)\eta + \mathcal{O}\left(\eta^2\right)$;
- $\mathbb{E}\Delta_i \overset{\text{1. Lemma A.1}}{=} -\partial_{e_i} f(x)\eta + \mathcal{O}\left(\eta^2\right)$.

Therefore, we have that for some $K_1(x) \in G$

$$\left|\mathbb{E}\Delta_i - \mathbb{E}\bar{\Delta}_i\right| \leq K_1(x)\eta^2, \quad \forall i = 1, \dots, d. \tag{56}$$

Additionally,we notice that $\forall i, j = 1, \dots, d$, it holds that

- $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j \overset{\text{2. Lemma A.8}}{=} \partial_{e_i} f \partial_{e_j} f \eta^2 + \Sigma^{\text{SGD}}_{(ij)}\eta^2 + \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\Delta_i\Delta_j \overset{\text{2. Lemma A.1}}{=} \partial_{e_i} f \partial_{e_j} f \eta^2 + \Sigma^{\text{SGD}}_{(ij)}\eta^2 + \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_2(x) \in G$

$$\left|\mathbb{E}\Delta_i\Delta_j - \mathbb{E}\bar{\Delta}_i\bar{\Delta}_j\right| \leq K_2(x)\eta^2, \quad \forall i, j = 1, \dots, d \tag{57}$$

Additionally, we notice that $\forall s \geq 3, \forall i_j \in \{1, \dots, d\}$, it holds that

- $\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j} \overset{\text{3. Lemma A.8}}{=} \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\prod_{j=1}^{s}\Delta_{i_j} \overset{\text{3. Lemma A.1}}{=} \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_3(x) \in G$

$$\left|\mathbb{E}\prod_{j=1}^{s}\Delta_{i_j} - \mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j}\right| \leq K_3(x)\eta^2. \tag{58}$$

Additionally, for some $K_4(x) \in G, \forall i_j \in \{1,\ldots,d\}$

$$\mathbb{E}\prod_{j=1}^{3}\left|\bar{\Delta}_{(i_j)}\right| \overset{\text{3. Lemma A.8}}{\leq} K_4(x)\eta^2. \tag{59}$$

Finally, Eq. (56), Eq. (57), Eq. (58), and Eq. (59) allow us to conclude the proof.

□

A.2. Formal Derivation - DNSAM

We now derive an SDE model for the DNSAM iteration given in (6) which we prove to be a 1-order weak approximation of such a discrete iteration. The following result is inspired by Theorem 1 of (Li et al., 2017). We will consider the stochastic process $X_t \in \mathbb{R}^d$ defined as the solution of the SDE

$$dX_t = -\nabla\tilde{f}^{\text{DNSAM}}(X_t)dt + \left(I_d + \rho\frac{\nabla^2 f(X_t)}{\|\nabla f(X_t)\|_2}\right)\left(\eta\Sigma^{\text{SGD}}(X_t)\right)^{\frac{1}{2}}dW_t \tag{60}$$

where the regularized loss is

$$\tilde{f}^{\text{DNSAM}}(x) = f(x) + \rho\|\nabla f(x)\|_2,$$

the covariance matrix is

$$\Sigma^{DNSAM}(x) := \Sigma^{SGD}(x)\left(I_d + 2\rho\frac{\nabla^2 f(x)}{\|\nabla f(x)\|}\right) \tag{61}$$

and

$$\Sigma^{\text{SGD}}(x) := \mathbb{E}\left[\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)^T\right]$$

is the usual covariance of SGD.

Theorem A.9 (Stochastic modified equations)**.** *Let $0 < \eta < 1, T > 0$ and set $N = \lfloor T/\eta \rfloor$. Let $x_k \in \mathbb{R}^d, 0 \leq k \leq N$ denote a sequence of DNSAM iterations defined by Eq.* (6)*. Additionally, let us assume that the noise structure is such that the stochastic gradient can be written as $\nabla f_\gamma(x) = \nabla f(x) + Z$ and*

$$\rho = \mathcal{O}\left(\eta^{\frac{1}{2}}\right). \tag{62}$$

Consider the stochastic process X_t defined in Eq. (60) *and fix some test function $g \in G$ and suppose that g and its partial derivatives up to order 6 belong to G.*
Then, under Assumption A.3, there exists a constant $C > 0$ independent of η such that for all $k = 0, 1, \ldots, N$, we have

$$\left|\mathbb{E}g\left(X_{k\eta}\right) - \mathbb{E}g\left(x_k\right)\right| \leq C\eta^1.$$

That is, the SDE (60) *is an order* 1 *weak approximation of the DNSAM iterations* (6)*.*

Lemma A.10. *Under the assumptions of Theorem A.9, let* $0 < \eta < 1$ *and consider* $x_k, k \geq 0$ *satisfying the DNSAM iterations* (6)

$$x_{k+1} = x_k - \eta \nabla f_{\gamma_k}\left(x_k + \rho \frac{\nabla f_{\gamma_k}(x_k)}{\|\nabla f(x_k)\|}\right)$$

with $x_0 = x \in \mathbb{R}^d$. *Additionally, we define* $\partial_{e_i} \tilde{f}^{DNSAM}(x) := \partial_{e_i} f(x) + \rho \frac{\sum_j \partial^2_{e_i+e_j} f(x) \partial_{e_j} f(x)}{\|\nabla f(x)\|}$. *From the definition the one-step difference* $\bar{\Delta} = x_1 - x$, *and we indicate with* $\bar{\Delta}_i$ *the* i*-th component of such difference. Then, we have*

1. $\mathbb{E}\bar{\Delta}_i = -\partial_{e_i} \tilde{f}^{DNSAM}(x)\eta + \mathcal{O}(\eta\rho^2) \quad \forall i = 1, \ldots, d;$
2. $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j = \partial_{e_i} \tilde{f}^{DNSAM}(x)\partial_{e_j} \tilde{f}^{DNSAM}(x)\eta^2 + \Sigma^{DNSAM}_{(ij)}\eta^2 + \mathcal{O}\left(\eta^3\right) \quad \forall i, j = 1, \ldots, d;$
3. $\mathbb{E}\prod_{j=1}^{s} \bar{\Delta_{i_j}} = \mathcal{O}\left(\eta^3\right) \quad \forall s \geq 3, \quad i_j \in \{1, \ldots, d\}.$

and all the functions above are evaluated at x.

Proof of Lemma A.10. Since the first step is to evaluate $\mathbb{E}\Delta_i = -\mathbb{E}\left[\partial_{e_i} f_\gamma\left(x + \frac{\rho}{\|\nabla f(x)\|}\nabla f_\gamma(x)\right)\eta\right]$, we start by analyzing $\partial_{e_i} f_\gamma\left(x + \frac{\rho}{\|\nabla f(x)\|}\nabla f_\gamma(x)\right)$, that is the partial derivative in the direction $e_i := (0, \cdots, 0, \underset{i-th}{1}, 0, \cdots, 0)$. Under the noise assumption $\nabla f_\gamma(x) = \nabla f(x) + Z$, we have that $\nabla^2 f_\gamma(x) = \nabla^2 f(x)$. Then, we have that

$$\partial_{e_i} f_\gamma\left(x + \frac{\rho}{\|\nabla f(x)\|}\nabla f_\gamma(x)\right) = \partial_{e_i} f_\gamma(x) + \sum_{|\alpha|=1} \partial^2_{e_i+\alpha} f(x) \rho \frac{\partial_\alpha f_\gamma(x)}{\|\nabla f(x)\|} + \mathcal{R}^{\partial_{e_i} f_\gamma(x)}_{x,1}\left(\rho \frac{\nabla f_\gamma(x)}{\|\nabla f(x)\|}\right), \tag{63}$$

where the residual is defined in Eq. (4) of (Folland, 2005). Therefore, for some constant $c \in (0, 1)$, it holds that

$$\mathcal{R}^{\partial_{e_i} f_\gamma(x)}_{x,1}\left(\rho \frac{\nabla f_\gamma(x)}{\|\nabla f(x)\|}\right) = \sum_{|\alpha|=2} \frac{\partial^3_{e_i+\alpha} f_\gamma\left(x + c\rho \frac{\nabla f_\gamma(x)}{\|\nabla f(x)\|}\right) \rho^2 \left(\frac{\nabla f_\gamma(x)}{\|\nabla f(x)\|}\right)^\alpha}{\alpha!}. \tag{64}$$

Combining the last two equations, we obtain

$$\partial_{e_i} f_\gamma\left(x + \frac{\rho}{\|\nabla f(x)\|}\nabla f_\gamma(x)\right) = \partial_{e_i} f_\gamma(x) + \frac{\rho}{\|\nabla f(x)\|} \sum_{|\alpha|=1} \partial^2_{e_i+\alpha} f(x) \partial_\alpha f_\gamma(x) \tag{65}$$

$$+ \rho^2 \sum_{|\alpha|=2} \frac{\partial^3_{e_i+\alpha} f_\gamma\left(x + c\rho \frac{\nabla f_\gamma(x)}{\|\nabla f(x)\|}\right) \left(\frac{\nabla f_\gamma(x)}{\|\nabla f(x)\|}\right)^\alpha}{\alpha!}. \tag{66}$$

Now, we observe that

$$K_i(x) := \left[\sum_{|\alpha|=2} \frac{\partial^3_{e_i+\alpha} f_\gamma\left(x + c\rho \frac{\nabla f(x)}{\|\nabla f(x)\|}\right) \left(\frac{\nabla f_\gamma(x)}{\|\nabla f(x)\|}\right)^\alpha}{\alpha!}\right] \tag{67}$$

is a finite sum of products of functions that by assumption are in G. Therefore, $K_i(x) \in G$ and $\bar{K}_i(x) = \mathbb{E}\left[K_i(x)\right] \in G$. Based on these definitions, we rewrite Eq. (65) as

$$\partial_{e_i} f_\gamma\left(x + \frac{\rho}{\|\nabla f(x)\|}\nabla f_\gamma(x)\right) = \partial_{e_i} f_\gamma(x) + \frac{\rho}{\|\nabla f(x)\|} \sum_{|\alpha|=1} \partial^2_{e_i+\alpha} f(x) \partial_\alpha f_\gamma(x) + \rho^2 K_i(x). \tag{68}$$

which implies that

$$\mathbb{E}\left[\partial_{e_i} f_\gamma\left(x + \frac{\rho}{\|\nabla f(x)\|}\nabla f_\gamma(x)\right)\right] = \partial_{e_i} f(x) + \rho \frac{\sum_j \partial^2_{e_i+e_j} f(x) \partial_{e_j} f(x)}{\|\nabla f(x)\|} + \rho^2 \bar{K}_i(x), \tag{69}$$

where we used the unbiasedness property of the stochastic gradients: $\mathbb{E}\nabla f_\gamma(x) = \nabla f(x)$.

Let us now remember that by definition

$$\partial_{e_i}\tilde{f}^{\text{DNSAM}}(x) = \partial_{e_i} f(x) + \rho\frac{\sum_j \partial^2_{e_i+e_j} f(x)\partial_{e_j} f(x)}{\|\nabla f(x)\|}. \tag{70}$$

Therefore, by using Eq. (69), Eq. (70), and the assumption (62) we have that $\forall i = 1, \ldots, d$,

$$\mathbb{E}\bar{\Delta}_i = -\partial_{e_i}\tilde{f}^{\text{DNSAM}}(x)\eta + \eta\rho^2\bar{K}_i(x) = -\partial_{e_i}\tilde{f}^{\text{DNSAM}}(x)\eta + \mathcal{O}\left(\eta^2\right). \tag{71}$$

We now observe that the covariance matrix of the difference between the drift $\nabla f(x) + \rho\frac{\nabla^2 f(x)\nabla f(x)}{\|\nabla f(x)\|}$ of the SDE (60) and the gradient $\nabla f_\gamma\left(x + \frac{\rho}{\|\nabla f(x)\|}\nabla f_\gamma(x)\right) = \nabla f_\gamma(x) + \rho\frac{\nabla^2 f(x)\nabla f_\gamma(x)}{\|\nabla f(x)\|} + \rho^2 K(x)$ in the discrete algorithm (6) is

$$\begin{aligned}
\bar{\Sigma} := \mathbb{E}\Bigg[&\left(\nabla f(x) + \rho\frac{\nabla^2 f(x)\nabla f(x)}{\|\nabla f(x)\|} - \nabla f_\gamma(x) - \rho\frac{\nabla^2 f(x)\nabla f_\gamma(x)}{\|\nabla f(x)\|} - \rho^2 K(x)\right) && (72)\\
&\left(\nabla f(x) + \rho\frac{\nabla^2 f(x)\nabla f(x)}{\|\nabla f(x)\|} - \nabla f_\gamma(x) - \rho\frac{\nabla^2 f(x)\nabla f_\gamma(x)}{\|\nabla f(x)\|} - \rho^2 K(x)\right)^\top\Bigg] && (73)\\
&= \Sigma^{SGD}\left(I_d + 2\rho\frac{\nabla^2 f(x)}{\|\nabla f(x)\|}\right) + \mathcal{O}(\rho^2) = \Sigma^{DNSAM}(x) + \mathcal{O}(\rho^2). && (74)
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j =& \text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) + \mathbb{E}\bar{\Delta}_i\mathbb{E}\bar{\Delta}_j\\
&\overset{(71)}{=} \text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) + \partial_{e_i}\tilde{f}^{\text{DNSAM}}\partial_{e_j}\tilde{f}^{\text{DNSAM}}\eta^2 + \eta^2\rho^2(\partial_{e_i}\tilde{f}^{\text{DNSAM}}\bar{K}_j(x) + \partial_{e_j}\tilde{f}^{\text{DNSAM}}\bar{K}_i(x)) + \eta^2\rho^4\bar{K}_i(x)\bar{K}_j(x)\\
&= \text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) + \partial_{e_i}\tilde{f}^{\text{DNSAM}}\partial_{e_j}\tilde{f}^{\text{DNSAM}}\eta^2 + \mathcal{O}\left(\eta^2\rho^2\right) + \mathcal{O}\left(\eta^2\rho^4\right)\\
&= \partial_{e_i}\tilde{f}^{\text{DNSAM}}\partial_{e_j}\tilde{f}^{\text{DNSAM}}\eta^2 + \text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) + \mathcal{O}\left(\eta^2\rho^2\right) + \mathcal{O}\left(\eta^2\rho^4\right) \quad \forall i, j = 1, \ldots, d. \qquad (75)
\end{aligned}$$

By the definitions of Σ^{DNSAM} and of $\bar{\Sigma}(x)$, we have

$$\text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) = \eta^2\bar{\Sigma}_{i,j}(x) = \eta^2\Sigma^{\text{DNSAM}}_{i,j}(x) + \mathcal{O}(\eta^2\rho^2). \tag{76}$$

Therefore, remembering Eq. (75) and Eq. (62) we have

$$\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j = \partial_{e_i}\tilde{f}^{\text{DNSAM}}\partial_{e_j}\tilde{f}^{\text{DNSAM}}\eta^2 + \Sigma^{\text{DNSAM}}_{i,j}\eta^2 + \mathcal{O}\left(\eta^3\right), \quad \forall i, j = 1, \ldots, d. \tag{77}$$

Finally, with analogous considerations, it is obvious that under our assumptions

$$\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^s\right) \quad \forall s \geq 3, \quad i_j \in \{1, \ldots, d\}$$

which in particular implies that

$$\mathbb{E}\prod_{j=1}^{3}\bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right), \quad i_j \in \{1, \ldots, d\}.$$

□

Proof of Theorem A.9. To prove this result, all we need to do is check the conditions in Theorem A.2. As we apply Lemma A.1, we make the following choices:

- $b(x) = -\nabla \tilde{f}^{\text{DNSAM}}(x)$;
- $\sigma(x) = \Sigma^{\text{DNSAM}}(x)^{\frac{1}{2}}$.

First of all, we notice that $\forall i = 1, \dots, d$, it holds that

- $\mathbb{E}\bar{\Delta}_i \overset{\text{1. Lemma A.10}}{=} -\partial_{e_i}\tilde{f}^{\text{DNSAM}}(x)\eta + \mathcal{O}(\eta^2)$;
- $\mathbb{E}\Delta_i \overset{\text{1. Lemma A.1}}{=} -\partial_{e_i}\tilde{f}^{\text{DNSAM}}(x)\eta + \mathcal{O}\left(\eta^2\right)$.

Therefore, we have that for some $K_1(x) \in G$

$$\left|\mathbb{E}\Delta_i - \mathbb{E}\bar{\Delta}_i\right| \leq K_1(x)\eta^2, \quad \forall i = 1, \dots, d. \tag{78}$$

Additionally,we notice that $\forall i, j = 1, \dots, d$, it holds that

- $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j \overset{\text{2. Lemma A.10}}{=} \partial_{e_i}\tilde{f}^{\text{DNSAM}}\partial_{e_j}\tilde{f}^{\text{DNSAM}}\eta^2 + \Sigma^{\text{DNSAM}}_{i,j}\eta^2 + \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\Delta_i\Delta_j \overset{\text{2. Lemma A.1}}{=} \partial_{e_i}\tilde{f}^{\text{DNSAM}}\partial_{e_j}\tilde{f}^{\text{DNSAM}}\eta^2 + \Sigma^{\text{DNSAM}}_{i,j}\eta^2 + \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_2(x) \in G$

$$\left|\mathbb{E}\Delta_i\Delta_j - \mathbb{E}\bar{\Delta}_i\bar{\Delta}_j\right| \leq K_2(x)\eta^2, \quad \forall i, j = 1, \dots, d. \tag{79}$$

Additionally, we notice that $\forall s \geq 3, \forall i_j \in \{1, \dots, d\}$, it holds that

- $\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j} \overset{\text{3. Lemma A.10}}{=} \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\prod_{j=1}^{s}\Delta_{i_j} \overset{\text{3. Lemma A.1}}{=} \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_3(x) \in G$

$$\left|\mathbb{E}\prod_{j=1}^{s}\Delta_{i_j} - \mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j}\right| \leq K_3(x)\eta^2. \tag{80}$$

Additionally, for some $K_4(x) \in G, \forall i_j \in \{1, \dots, d\}$

$$\mathbb{E}\prod_{j=1}^{3}\left|\bar{\Delta}_{(i_j)}\right| \overset{\text{3. Lemma A.10}}{\leq} K_4(x)\eta^2. \tag{81}$$

Finally, Eq. (78), Eq. (79), Eq. (80), and Eq. (81) allow us to conclude the proof.

□

A.2.1. DNSAM IS SGD IF $\rho = \mathcal{O}(\eta)$

The following result is inspired by Theorem 1 of (Li et al., 2017). We will consider the stochastic process $X_t \in \mathbb{R}^d$ defined as the solution of the SDE

$$dX_t = -\nabla f\left(X_t\right)dt + \left(\eta\Sigma^{\text{SGD}}\left(X_t\right)\right)^{1/2}dW_t \tag{82}$$

Such that $X_0 = x_0$ and

$$\Sigma^{\text{SGD}}(x) := \mathbb{E}\left[\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)^T\right]$$

Theorem A.11 (Stochastic modified equations). *Let $0 < \eta < 1, T > 0$ and set $N = \lfloor T/\eta \rfloor$. Let $x_k \in \mathbb{R}^d, 0 \leq k \leq N$ denote a sequence of DNSAM iterations defined by Eq.* (6)*. Additionally, let us take*

$$\rho = \mathcal{O}\left(\eta^1\right). \tag{83}$$

Consider the stochastic process X_t defined in Eq. (82) *and fix some test function $g \in G$ and suppose that g and its partial derivatives up to order 6 belong to G. Then, under Assumption A.3, there exists a constant $C > 0$ independent of η such that for all $k = 0, 1, \ldots, N$, we have*

$$|\mathbb{E}g\left(X_{k\eta}\right) - \mathbb{E}g\left(x_k\right)| \leq C\eta^1.$$

That is, the SDE (82) *is an order* 1 *weak approximation of the SAM iterations* (4)*.*

Proof. The proof is completely similar to that of Theorem A.7 presented before and of Theorem A.16 presented later. □

A.3. Formal Derivation - SAM

The following result is inspired by Theorem 1 of (Li et al., 2017). We will consider the stochastic process $X_t \in \mathbb{R}^d$ defined as the solution of the SDE

$$dX_t = -\nabla \tilde{f}^{\text{SAM}}(X_t)dt + \sqrt{\eta}\left(\Sigma^{\text{SGD}}(X_t) + \rho\left(\hat{\Sigma}(X_t) + \hat{\Sigma}(X_t)^\top\right)\right)^{\frac{1}{2}} dW_t \tag{84}$$

where

$$\Sigma^{\text{SGD}}(x) := \mathbb{E}\left[\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)^T\right]$$

is the usual covariance of SGD, while

$$\hat{\Sigma}(x) := \mathbb{E}\left[\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)\left(\mathbb{E}\left[\frac{\nabla^2 f_\gamma(x)\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right] - \frac{\nabla^2 f_\gamma(x)\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right)^\top\right] \tag{85}$$

and

$$\tilde{f}^{\text{SAM}}(x) := f(x) + \rho\mathbb{E}\left[\|\nabla f_\gamma(x)\|_2\right].$$

In the following, we will use the notation

$$\Sigma^{\text{SAM}}(x) := \left(\Sigma^{\text{SGD}}(X_t) + \rho\left(\hat{\Sigma}(X_t) + \hat{\Sigma}(X_t)^\top\right)\right). \tag{86}$$

Theorem A.12 (Stochastic modified equations). *Let $0 < \eta < 1, T > 0$ and set $N = \lfloor T/\eta \rfloor$. Let $x_k \in \mathbb{R}^d, 0 \leq k \leq N$ denote a sequence of SAM iterations defined by Eq.* (4)*. Additionally, let us take*

$$\rho = \mathcal{O}\left(\eta^{\frac{1}{2}}\right). \tag{87}$$

Consider the stochastic process X_t defined in Eq. (84) *and fix some test function $g \in G$ and suppose that g and its partial derivatives up to order 6 belong to G.*
Then, under Assumption A.3, there exists a constant $C > 0$ independent of η such that for all $k = 0, 1, \ldots, N$, we have

$$|\mathbb{E}g\left(X_{k\eta}\right) - \mathbb{E}g\left(x_k\right)| \leq C\eta^1.$$

That is, the SDE (84) *is an order* 1 *weak approximation of the SAM iterations* (4)*.*

Remark A.13. Denote by b and σ the (Borel measurable) drift and diffusion coefficient in (84), respectively. Suppose that there exists a non-negative function $F \in L^{d+1}([0,\infty)\times\mathbb{R}^d)$ such that

$$\|b(t,x)\| \le K + F(t,x) \tag{88}$$

$dt\times dx - a.e.$ for some constant $K \ge 0$. Further, assume that $\sigma\sigma^\top$ is strongly elliptic, that is there exists a $\delta\in(0,1)$ such that for all $t\ge 0$, $x\in\mathbb{R}^d$

$$\delta I_d \le \sigma\sigma^\top(t,x) \le \delta^{-1} I_d, \tag{89}$$

where $I_d \in \mathbb{R}^{d\times d}$ is the unit matrix. Then there exists a (global) weak solution to (84). Moreover, if there is a (global) weak solution to (84), $b\in L^{2d+2}_{loc}([0,\infty)\times\mathbb{R}^d)$, σ is locally Lipschitz continuous uniformly in time and if (89) holds, then there exists a unique strong solution to (84). See (Gyöngy & Martínez, 2001).

Lemma A.14. *Under the assumptions of Theorem A.12, let* $0<\eta<1$ *and consider* $x_k, k\ge 0$ *satisfying the USAM iterations* (4)

$$x_{k+1} = x_k - \eta\nabla f_{\gamma_k}\left(x_k + \rho\frac{\nabla f_{\gamma_k}(x_k)}{\|\nabla f_\gamma(x_k)\|}\right)$$

with $x_0 = x\in\mathbb{R}^d$. *Additionally, we define* $\partial_{e_i}\tilde{f}^{SAM}(x) := \partial_{e_i}f(x) + \rho\mathbb{E}\left[\frac{\sum_j \partial^2_{e_i+e_j}f_\gamma(x)\partial_{e_j}f_\gamma(x)}{\|\nabla f_\gamma(x)\|}\right]$. *From the definition the one-step difference* $\bar\Delta = x_1 - x$, *then we have*

1. $\mathbb{E}\bar\Delta_i = -\partial_{e_i}\tilde{f}^{SAM}(x)\eta + \mathcal{O}(\eta\rho^2) \quad \forall i = 1,\dots,d;$
2. $\mathbb{E}\bar\Delta_i\bar\Delta_j = \partial_{e_i}\tilde{f}^{SAM}(x)\partial_{e_j}\tilde{f}^{SAM}(x)\eta^2 + \Sigma^{SAM}_{(ij)}\eta^2 + \mathcal{O}\left(\eta^3\right) \quad \forall i,j = 1,\dots,d;$
3. $\mathbb{E}\prod_{j=1}^s \bar\Delta_{i_j} = \mathcal{O}\left(\eta^3\right) \quad \forall s\ge 3, \quad i_j\in\{1,\dots,d\}.$

and all the functions above are evaluated at x.

Proof of Lemma A.14. Since the first step is to evaluate $\mathbb{E}\Delta_i = -\mathbb{E}\left[\partial_{e_i}f_\gamma\left(x+\frac{\rho}{\|\nabla f_\gamma(x)\|}\nabla f_\gamma(x)\right)\eta\right]$, we start by analyzing $\partial_{e_i}f_\gamma\left(x+\frac{\rho}{\|\nabla f_\gamma(x)\|}\nabla f_\gamma(x)\right)$, that is the partial derivative in the direction $e_i := (0,\cdots,0,\underset{i-th}{1},0,\cdots,0)$. Then, we have that

$$\partial_{e_i}f_\gamma\left(x+\frac{\rho}{\|\nabla f_\gamma(x)\|}\nabla f_\gamma(x)\right) = \partial_{e_i}f_\gamma(x) + \sum_{|\alpha|=1}\partial^2_{e_i+\alpha}f_\gamma(x)\rho\frac{\partial_\alpha f_\gamma(x)}{\|\nabla f_\gamma(x)\|} + \mathcal{R}^{\partial_{e_i}f_\gamma(x)}_{x,1}\left(\rho\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|}\right) \tag{90}$$

Where the residual is defined in Eq. (4) of (Folland, 2005). Therefore, for some constant $c\in(0,1)$, it holds that

$$\mathcal{R}^{\partial_{e_i}f_\gamma(x)}_{x,1}\left(\rho\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|}\right) = \sum_{|\alpha|=2}\frac{\partial^3_{e_i+\alpha}f_\gamma\left(x+c\rho\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|}\right)\rho^2\left(\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|}\right)^\alpha}{\alpha!}. \tag{91}$$

Therefore, we can rewrite it as

$$\partial_{e_i}f_\gamma\left(x+\frac{\rho}{\|\nabla f_\gamma(x)\|}\nabla f_\gamma(x)\right) = \partial_{e_i}f_\gamma(x) + \frac{\rho}{\|\nabla f_\gamma(x)\|}\sum_{|\alpha|=1}\partial^2_{e_i+\alpha}f_\gamma(x)\partial_\alpha f_\gamma(x) \tag{92}$$

$$+\rho^2\sum_{|\alpha|=2}\frac{\partial^3_{e_i+\alpha}f_\gamma\left(x+c\rho\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|}\right)\left(\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|}\right)^\alpha}{\alpha!}. \tag{93}$$

Now, we observe that

$$K_i(x) := \left[\sum_{|\alpha|=2} \frac{\partial^3_{e_i+\alpha} f_\gamma \left(x + c\rho \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|} \right) \left(\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|} \right)^\alpha}{\alpha!} \right] \tag{94}$$

is a finite sum of products of functions that by assumption are in G. Therefore, $K_i(x) \in G$ and $\bar{K}_i(x) = \mathbb{E}\left[K_i(x)\right] \in G$. Based on these definitions, we rewrite Eq. (92) as

$$\partial_{e_i} f_\gamma \left(x + \frac{\rho}{\|\nabla f_\gamma(x)\|} \nabla f_\gamma(x) \right) = \partial_{e_i} f_\gamma(x) + \frac{\rho}{\|\nabla f_\gamma(x)\|} \sum_{|\alpha|=1} \partial^2_{e_i+\alpha} f_\gamma(x) \partial_\alpha f_\gamma(x) + \rho^2 K_i(x). \tag{95}$$

which implies that

$$\mathbb{E}\left[\partial_{e_i} f_\gamma \left(x + \frac{\rho}{\|\nabla f_\gamma(x)\|} \nabla f_\gamma(x) \right) \right] = \partial_{e_i} f(x) + \rho \mathbb{E}\left[\frac{\sum_j \partial^2_{e_i+e_j} f_\gamma(x) \partial_{e_j} f_\gamma(x)}{\|\nabla f_\gamma(x)\|} \right] + \rho^2 \bar{K}_i(x). \tag{96}$$

Let us now remember that

$$\partial_{e_i} \tilde{f}^{\text{SAM}}(x) = \partial_{e_i} \left(f(x) + \rho \mathbb{E}\left[\|\nabla f_\gamma(x)\|_2 \right] \right) = \partial_{e_i} f(x) + \rho \mathbb{E}\left[\frac{\sum_j \partial^2_{e_i+e_j} f_\gamma(x) \partial_{e_j} f_\gamma(x)}{\|\nabla f_\gamma(x)\|} \right] \tag{97}$$

Therefore, by using Eq. (96), Eq. (97), and the assumption (87) we have that $\forall i = 1, \ldots, d$

$$\mathbb{E}\bar{\Delta}_i = -\partial_{e_i} \tilde{f}^{\text{SAM}}(x)\eta + \eta\rho^2 \bar{K}_i(x) = -\partial_{e_i} \tilde{f}^{\text{SAM}}(x)\eta + \mathcal{O}\left(\eta^2\right). \tag{98}$$

Additionally, we have that

$$\begin{aligned}
\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j =& \text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) + \mathbb{E}\bar{\Delta}_i \mathbb{E}\bar{\Delta}_j \\
&\overset{(98)}{=} \text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) + \partial_{e_i} \tilde{f}^{\text{SAM}} \partial_{e_j} \tilde{f}^{\text{SAM}} \eta^2 + \eta^2\rho^2 (\partial_{e_i} \tilde{f}^{\text{SAM}} \bar{K}_j(x) + \partial_{e_j} \tilde{f}^{\text{SAM}} \bar{K}_i(x)) + \eta^2\rho^4 \bar{K}_i(x) \bar{K}_j(x) \\
&= \text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) + \partial_{e_i} \tilde{f}^{\text{SAM}} \partial_{e_j} \tilde{f}^{\text{SAM}} \eta^2 + \mathcal{O}\left(\eta^2\rho^2\right) + \mathcal{O}\left(\eta^2\rho^4\right) \\
&= \partial_{e_i} \tilde{f}^{\text{SAM}} \partial_{e_j} \tilde{f}^{\text{SAM}} \eta^2 + \text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) + \mathcal{O}\left(\eta^2\rho^2\right) + \mathcal{O}\left(\eta^2\rho^4\right) \quad \forall i, j = 1, \ldots, d.
\end{aligned} \tag{99}$$

Let us now recall the expression (85) of $\hat{\Sigma}$ and the expression (86) of Σ^{SAM}. Then, we automatically have that

$$\text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) = \eta^2 \left(\Sigma^{\text{SGD}}_{i,j}(x) + \rho \left[\hat{\Sigma}_{i,j}(x) + \hat{\Sigma}_{i,j}(x)^\top \right] + \mathcal{O}(\rho^2) \right) = \eta^2 \Sigma^{\text{SAM}}_{i,j}(x) + \mathcal{O}(\eta^2\rho^2). \tag{100}$$

Therefore, remembering Eq. (99) and Eq. (87) we have

$$\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j = \partial_{e_i} \tilde{f}^{\text{SAM}} \partial_{e_j} \tilde{f}^{\text{SAM}} \eta^2 + \Sigma^{\text{SAM}}_{i,j} \eta^2 + \mathcal{O}\left(\eta^3\right), \quad \forall i, j = 1, \ldots, d. \tag{101}$$

Finally, with analogous considerations, it is obvious that under our assumptions

$$\mathbb{E} \prod_{j=1}^{s} \bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^s\right) \quad \forall s \geq 3, \quad i_j \in \{1, \ldots, d\}$$

which in particular implies that

$$\mathbb{E} \prod_{j=1}^{3} \bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right), \quad i_j \in \{1, \ldots, d\}.$$

□

Proof of Theorem A.12. To prove this result, all we need to do is check the conditions in Theorem A.2. As we apply Lemma A.1, we make the following choices:

- $b(x) = -\nabla \tilde{f}^{\text{SAM}}(x)$;
- $\sigma(x) = \Sigma^{\text{SAM}}(x)^{\frac{1}{2}}$.

First of all, we notice that $\forall i = 1, \ldots, d$, it holds that

- $\mathbb{E}\bar{\Delta}_i \overset{\text{1. Lemma A.14}}{=} -\partial_{e_i}\tilde{f}^{\text{SAM}}(x)\eta + \mathcal{O}(\eta^2)$;
- $\mathbb{E}\Delta_i \overset{\text{1. Lemma A.1}}{=} -\partial_{e_i}\tilde{f}^{\text{SAM}}(x)\eta + \mathcal{O}\left(\eta^2\right)$.

Therefore, we have that for some $K_1(x) \in G$

$$\left|\mathbb{E}\Delta_i - \mathbb{E}\bar{\Delta}_i\right| \leq K_1(x)\eta^2, \quad \forall i = 1, \ldots, d. \tag{102}$$

Additionally,we notice that $\forall i, j = 1, \ldots, d$, it holds that

- $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j \overset{\text{2. Lemma A.14}}{=} \partial_{e_i}\tilde{f}^{\text{SAM}}\partial_{e_j}\tilde{f}^{\text{SAM}}\eta^2 + \Sigma^{\text{SAM}}_{i,j}\eta^2 + \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\Delta_i\Delta_j \overset{\text{2. Lemma A.1}}{=} \partial_{e_i}\tilde{f}^{\text{SAM}}\partial_{e_j}\tilde{f}^{\text{SAM}}\eta^2 + \Sigma^{\text{SAM}}_{i,j}\eta^2 + \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_2(x) \in G$

$$\left|\mathbb{E}\Delta_i\Delta_j - \mathbb{E}\bar{\Delta}_i\bar{\Delta}_j\right| \leq K_2(x)\eta^2, \quad \forall i, j = 1, \ldots, d. \tag{103}$$

Additionally, we notice that $\forall s \geq 3, \forall i_j \in \{1, \ldots, d\}$, it holds that

- $\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j} \overset{\text{3. Lemma A.14}}{=} \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\prod_{j=1}^{s}\Delta_{i_j} \overset{\text{3. Lemma A.1}}{=} \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_3(x) \in G$

$$\left|\mathbb{E}\prod_{j=1}^{s}\Delta_{i_j} - \mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j}\right| \leq K_3(x)\eta^2. \tag{104}$$

Additionally, for some $K_4(x) \in G$, $\forall i_j \in \{1, \ldots, d\}$

$$\mathbb{E}\prod_{j=1}^{3}\left|\bar{\Delta}_{(i_j)}\right| \overset{\text{3. Lemma A.5}}{\leq} K_4(x)\eta^2. \tag{105}$$

Finally, Eq. (102), Eq. (103), Eq. (104), and Eq. (105) allow us to conclude the proof.

□

Corollary A.15. *Let us take the same assumptions of Theorem* (A.12)*. Additionally, let us assume that the dynamics is near the minimizer. In this case, the noise structure is such that the stochastic gradient can be written as* $\nabla f_\gamma(x) = \nabla f(x) + Z$ *such that* Z *is the noise that does not depend on* x*. In this case, the SDE* (84) *becomes*

$$dX_t = -\nabla \tilde{f}^{SAM}(X_t)dt + \sqrt{\eta\left(\Sigma^{SGD}(X_t) + \rho H_t\left(\bar{\Sigma}(X_t) + \bar{\Sigma}(X_t)^\top\right)\right)}dW_t$$

where $H_t := \nabla^2 f(X_t)$ *and* $\bar{\Sigma}(x)$ *is defined as*

$$\mathbb{E}\left[\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)\left(\mathbb{E}\left[\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right] - \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right)^\top\right],$$

and $\tilde{f}^{SAM}(x) := f(x) + \rho\mathbb{E}\left[\|\nabla f_\gamma(x)\|_2\right].$

Proof of Corollary A.15. It follows immediately by substituting the expression for the perturbed gradient.

□

A.3.1. SAM IS SGD IF $\rho = \mathcal{O}(\eta)$

The following result is inspired by Theorem 1 of (Li et al., 2017). We will consider the stochastic process $X_t \in \mathbb{R}^d$ defined as the solution of the SDE

$$dX_t = -\nabla f\left(X_t\right)dt + \left(\eta\Sigma^{\text{SGD}}\left(X_t\right)\right)^{1/2}dW_t \tag{106}$$

Such that $X_0 = x_0$ and

$$\Sigma^{\text{SGD}}(x) := \mathbb{E}\left[\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)^T\right]$$

Theorem A.16 (Stochastic modified equations)**.** *Let* $0 < \eta < 1, T > 0$ *and set* $N = \lfloor T/\eta \rfloor$*. Let* $x_k \in \mathbb{R}^d, 0 \le k \le N$ *denote a sequence of SAM iterations defined by Eq.* (4)*. Additionally, let us take*

$$\rho = \mathcal{O}\left(\eta^1\right). \tag{107}$$

Consider the stochastic process X_t *defined in Eq.* (106) *and fix some test function* $g \in G$ *and suppose that* g *and its partial derivatives up to order 6 belong to* G*. Then, under Assumption A.3, there exists a constant* $C > 0$ *independent of* η *such that for all* $k = 0, 1, \ldots, N$*, we have*

$$\left|\mathbb{E}g\left(X_{k\eta}\right) - \mathbb{E}g\left(x_k\right)\right| \le C\eta^1.$$

That is, the SDE (106) *is an order* 1 *weak approximation of the SAM iterations* (4)*.*

Lemma A.17. *Under the assumptions of Theorem A.16, let* $0 < \eta < 1$*. Consider* $x_k, k \ge 0$ *satisfying the SAM iterations*

$$x_{k+1} = x_k - \eta\nabla f_{\gamma_k}\left(x_k + \rho\frac{\nabla f_{\gamma_k}(x_k)}{\|\nabla f_{\gamma_k}(x_k)\|}\right)$$

with $x_0 = x \in \mathbb{R}^d$*. From the definition the one-step difference* $\bar{\Delta} = x_1 - x$*, then we have*

1. $\mathbb{E}\bar{\Delta}_i = -\partial_{e_i} f(x)\eta + \mathcal{O}(\eta^2) \quad \forall i = 1, \ldots, d.$
2. $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j = \partial_{e_i} f\partial_{e_j} f\eta^2 + \Sigma^{SGD}_{(ij)}\eta^2 + \mathcal{O}\left(\eta^3\right) \quad \forall i, j = 1, \ldots, d.$
3. $\mathbb{E}\prod_{j=1}^s \bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right) \quad \forall s \ge 3, \quad i_j \in \{1, \ldots, d\}.$

All functions above are evaluated at x*.*

Proof of Lemma A.17. First of all, we write that

$$\partial_{e_i} f_\gamma \left(x + \rho \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|} \right) = \partial_{e_i} f_\gamma(x) + \mathcal{R}_{x,0}^{\partial_{e_i} f_\gamma(x)} \left(\rho \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|} \right), \tag{108}$$

where the residual is defined in Eq. (4) of (Folland, 2005). Therefore, for some constant $c \in (0,1)$, it holds that

$$\mathcal{R}_{x,0}^{\partial_{e_i} f_\gamma(x)} \left(\rho \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|} \right) = \sum_{|\alpha|=1} \frac{\partial_{e_i+\alpha}^2 f_\gamma \left(x + c\rho \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|} \right) \rho^1 \left(\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|} \right)^\alpha}{\alpha!}. \tag{109}$$

Let us now observe that $\mathcal{R}_{x,0}^{\partial_{e_i} f_\gamma(x)} \left(\rho \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|} \right)$ is a finite sum of products of functions in G and that, therefore, it lies in G. Additionally, given its expression Eq. (109), we can factor out a common ρ and have that $K(x) = \rho K_1(x)$ for some function $K_1(x) \in G$. Therefore, we rewrite Eq. (108) as

$$\partial_{e_i} f_\gamma \left(x + \rho \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|} \right) = \partial_{e_i} f_\gamma(x) + \rho K_1(x). \tag{110}$$

First of all, we notice that if we define $\bar{K}_1(x) = \mathbb{E}\left[K_1(x)\right]$, also $\bar{K}_1(x) \in G$. Therefore, it holds that

$$\mathbb{E}\left[\partial_{e_i} f_\gamma \left(x + \rho \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|} \right) \right] \overset{(110)}{=} \partial_{e_i} f(x) + \rho \bar{K}_1(x) \tag{111}$$

Therefore, using assumption (107), $\forall i = 1, \dots, d$, we have that

$$\mathbb{E}\bar{\Delta}_i = -\partial_{e_i} f(x)\eta + \eta\rho\bar{K}_i(x) = -\partial_{e_i} f(x)\eta + \mathcal{O}\left(\eta^2\right) \tag{112}$$

Additionally, by keeping in mind the definition of the covariance matrix Σ, We immediately have

$$\begin{aligned}
\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j &\overset{(110)}{=} Cov(\bar{\Delta}_i, \bar{\Delta}_j) + \mathbb{E}\bar{\Delta}_i\mathbb{E}\bar{\Delta}_j \\
&= \Sigma_{(ij)}^{\text{SGD}}\eta^2 + \partial_{e_i} f \partial_{e_j} f \eta^2 + \eta^2\rho(\partial_{e_i} f \bar{K}_j(x) + \partial_{e_j} f \bar{K}_i(x)) + \eta^2\rho^2 \bar{K}_i(x)\bar{K}_j(x) \\
&= \Sigma_{(ij)}^{\text{SGD}}\eta^2 + \partial_{e_i} f \partial_{e_j} f \eta^2 + \mathcal{O}\left(\eta^2\rho\right) + \mathcal{O}\left(\eta^2\rho^2\right) \\
&= \partial_{e_i} f \partial_{e_j} f \eta^2 + \Sigma_{(ij)}^{\text{SGD}}\eta^2 + \mathcal{O}\left(\eta^3\right) \quad \forall i, j = 1, \dots, d
\end{aligned} \tag{113}$$

Finally, with analogous considerations, it is obvious that under our assumptions

$$\mathbb{E}\prod_{j=1}^{s} \bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right) \quad \forall s \geq 3, \quad i_j \in \{1, \dots, d\}.$$

□

Proof of Theorem A.16. To prove this result, all we need to do is check the conditions in Theorem A.2. As we apply Lemma A.1, we make the following choices:

- $b(x) = -\nabla f(x)$,
- $\sigma(x) = \Sigma^{\text{SGD}}(X_t)^{\frac{1}{2}}$;

First of all, we notice that $\forall i = 1, \dots, d$, it holds that

- $\mathbb{E}\bar{\Delta}_i \overset{\text{1. Lemma A.17}}{=} -\partial_{e_i} f(x)\eta + \mathcal{O}\left(\eta^2\right)$;

- $\mathbb{E}\Delta_i \overset{\text{1. Lemma A.1}}{=} -\partial_{e_i} f(x)\eta + \mathcal{O}\left(\eta^2\right)$.

Therefore, we have that for some $K_1(x) \in G$

$$\left|\mathbb{E}\Delta_i - \mathbb{E}\bar{\Delta}_i\right| \leq K_1(x)\eta^2, \quad \forall i = 1, \ldots, d. \tag{114}$$

Additionally,we notice that $\forall i, j = 1, \ldots, d$, it holds that

- $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j \overset{\text{2. Lemma A.17}}{=} \partial_{e_i} f \partial_{e_j} f \eta^2 + \Sigma^{\text{SGD}}_{(ij)}\eta^2 + \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\Delta_i\Delta_j \overset{\text{2. Lemma A.1}}{=} \partial_{e_i} f \partial_{e_j} f \eta^2 + \Sigma^{\text{SGD}}_{(ij)}\eta^2 + \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_2(x) \in G$

$$\left|\mathbb{E}\Delta_i\Delta_j - \mathbb{E}\bar{\Delta}_i\bar{\Delta}_j\right| \leq K_2(x)\eta^2, \quad \forall i, j = 1, \ldots, d \tag{115}$$

Additionally, we notice that $\forall s \geq 3, \forall i_j \in \{1, \ldots, d\}$, it holds that

- $\mathbb{E}\prod_{j=1}^{s} \bar{\Delta}_{i_j} \overset{\text{3. Lemma A.17}}{=} \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\prod_{j=1}^{s} \Delta_{i_j} \overset{\text{3. Lemma A.1}}{=} \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_3(x) \in G$

$$\left|\mathbb{E}\prod_{j=1}^{s} \Delta_{i_j} - \mathbb{E}\prod_{j=1}^{s} \bar{\Delta}_{i_j}\right| \leq K_3(x)\eta^2. \tag{116}$$

Additionally, for some $K_4(x) \in G$, $\forall i_j \in \{1, \ldots, d\}$

$$\mathbb{E}\prod_{j=1}^{3} \left|\bar{\Delta}_{(i_j)}\right| \overset{\text{3. Lemma A.17}}{\leq} K_4(x)\eta^2. \tag{117}$$

Finally, Eq. (114), Eq. (115), Eq. (116), and Eq. (117) allow us to conclude the proof.

□

B. Random SAM

Following (Ujváry et al., 2022) (Algorithm 2), we define Random SAM (RSAM) as the following discrete algorithm

$$x_{k+1} = x_k - \eta\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Sigma)}\nabla f_{\gamma_k}\left(x_k + \epsilon\right). \tag{118}$$

As a first attempt, we focus on the case where $\Sigma = \sigma^2 I_d$.

B.1. Formal Derivation - RSAM

We will consider the stochastic process $X_t \in \mathbb{R}^d$ defined by

$$dX_t = -\nabla \tilde{f}^{\text{RSAM}}(X_t)dt + \sqrt{\eta}\left(\Sigma^{\text{SGD}}(X_t) + \frac{\sigma^2}{2}\left(\tilde{\Sigma}(X_t) + \tilde{\Sigma}(X_t)^\top\right)\right)^{\frac{1}{2}} dW_t \tag{119}$$

where

$$\Sigma^{\text{SGD}}(x) := \mathbb{E}\left[\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)\left(\nabla f\left(x\right) - \nabla f_\gamma\left(x\right)\right)^T\right]$$

is the usual covariance of SGD, while

$$\tilde{\Sigma}(x) := \mathbb{E}\left[\left(\nabla f\left(x\right) - \nabla f_{\gamma}\left(x\right)\right)\left(\mathbb{E}\left[\nabla^3 f_{\gamma}(x)[I_d]\right] - \nabla^3 f_{\gamma}(x)[I_d]\right)^{\top}\right] \tag{120}$$

and

$$\tilde{f}^{\text{RSAM}}(x) := f(x) + \frac{\sigma^2}{2} Tr(\nabla^2 f(x)).$$

In the following, we will use the notation

$$\Sigma^{\text{RSAM}}(x) := \left(\Sigma^{\text{SGD}}(X_t) + \frac{\sigma^2}{2}\left(\tilde{\Sigma}(X_t) + \tilde{\Sigma}(X_t)^{\top}\right)\right). \tag{121}$$

Theorem B.1 (Stochastic modified equations)**.** *Let $0 < \eta < 1, T > 0$ and set $N = \lfloor T/\eta \rfloor$. Let $x_k \in \mathbb{R}^d, 0 \leq k \leq N$ denote a sequence of RSAM iterations defined by Eq.* (118)*. Additionally, let us take*

$$\sigma = \mathcal{O}\left(\eta^{\frac{1}{3}}\right). \tag{122}$$

Consider the stochastic process X_t defined in Eq. (119) *and fix some test function $g \in G$ and suppose that g and its partial derivatives up to order 6 belong to G.*
Then, under Assumption A.3, there exists a constant $C > 0$ independent of η such that for all $k = 0, 1, \ldots, N$, we have

$$\left|\mathbb{E}g\left(X_{k\eta}\right) - \mathbb{E}g\left(x_k\right)\right| \leq C\eta^1.$$

That is, the SDE (119) *is an order* 1 *weak approximation of the RSAM iterations* (118)*.*

Lemma B.2. *Under the assumptions of Theorem B.1, let $0 < \eta < 1$ and consider $x_k, k \geq 0$ satisfying the RSAM iterations* (118)

$$x_{k+1} = x_k - \eta \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\Sigma)} \nabla f \gamma_k\left(x_k + \epsilon\right).$$

where $\Sigma = \sigma^2 I_d$ and $x_0 = x \in \mathbb{R}^d$. From the definition the one-step difference $\bar{\Delta} = x_1 - x$, then we have

1. $\mathbb{E}\bar{\Delta}_i = -\partial_{e_i} \tilde{f}^{RSAM}(x)\eta + \mathcal{O}(\eta\sigma^3) \quad \forall i = 1, \ldots, d.$
2. $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j = \partial_{e_i}\tilde{f}^{RSAM}(x)\partial_{e_j}\tilde{f}^{RSAM}(x)\eta^2 + \Sigma^{RSAM}_{(ij)}\eta^2 + \mathcal{O}\left(\eta^3\right) \quad \forall i, j = 1, \ldots, d.$
3. $\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right) \quad \forall s \geq 3, \quad i_j \in \{1, \ldots, d\}.$

and all the functions above are evaluated at x.

Proof of Lemma B.2. We perform a Taylor expansion of $\partial_i f(\cdot)$ around x_k

$$x^i_{k+1} = x^i_k - \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Sigma)}\left[\eta\partial_i f\left(x_k\right) - \eta\sum_j \partial^2_{ij} f\left(x_k\right)\epsilon^j_k - \frac{\eta}{2}\sum_{j,l}\partial^3_{ijl} f\left(x_k\right)\epsilon^j_k\epsilon^l_k + \mathcal{O}\left(\eta\left\|\epsilon_k\right\|^3\right)\right],$$

and we notice that the term $\frac{\eta}{2}\sum_{j,l}\partial^3_{ijl} f\left(x_k\right)\epsilon^j_k\epsilon^l_k$ is equal to $\frac{\eta}{2}\partial_i\sum_{jl}\partial^2_{jl} f\left(x_k\right)\epsilon^j_k\epsilon^l_k$ due to Clairaut's theorem (assuming that f has continuous fourth-order partial derivatives). By exploiting that ϵ_k has mean zero and covariance $\sigma^2 I_d$, we have that

$$\mathbb{E}\left[x_{k+1} - x_k\right] = -\eta\nabla\tilde{f}^{\text{RSAM}}\left(x_k\right) + \mathcal{O}\left(\eta\mathbb{E}\left[\left\|\epsilon_k\right\|^3\right]\right) = -\eta\nabla\tilde{f}^{\text{RSAM}}\left(x_k\right) + \mathcal{O}\left(\eta\sigma^3\right),$$

where the modified loss $\tilde{f}^{\text{RSAM}}$ is given by

$$\tilde{f}^{\text{RSAM}}(x) := f(z) + \frac{\sigma^2}{2}\operatorname{Tr}\left(\nabla^2 f(x)\right).$$

Therefore, using (122), we have that $\forall i = 1, \dots, d$

$$\mathbb{E}\bar{\Delta}_i = -\partial_{e_i}\tilde{f}^{\text{RSAM}}(x)\eta + \mathcal{O}\left(\eta^2\right). \tag{123}$$

Additionally, we have that

$$\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j = \partial_{e_i}\tilde{f}^{\text{RSAM}}\partial_{e_j}\tilde{f}^{\text{RSAM}}\eta^2 + \text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) + \mathcal{O}\left(\eta^3\right) \quad \forall i, j = 1, \dots, d. \tag{124}$$

Let us now recall the expression (120) of $\tilde{\Sigma}$ and the expression (121) of Σ^{RSAM}. Then, we automatically have that

$$\text{Cov}(\bar{\Delta}_i, \bar{\Delta}_j) = \eta^2\left(\Sigma^{\text{SGD}}_{i,j}(x) + \frac{\sigma^2}{2}\left[\tilde{\Sigma}_{i,j}(x) + \tilde{\Sigma}_{i,j}(x)^\top\right] + \mathcal{O}(\sigma^3)\right) = \eta^2\Sigma^{\text{RSAM}}_{i,j}(x) + \mathcal{O}(\eta^2\sigma^3) \tag{125}$$

Therefore, remembering Eq. (124) and Eq. (122) we have

$$\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j = \partial_{e_i}\tilde{f}^{\text{RSAM}}\partial_{e_j}\tilde{f}^{\text{RSAM}}\eta^2 + \Sigma^{\text{SAM}}_{i,j}\eta^2 + \mathcal{O}\left(\eta^3\right), \quad \forall i, j = 1, \dots, d \tag{126}$$

Finally, with analogous considerations, it is obvious that under our assumptions

$$\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^s\right) \quad \forall s \geq 3, \quad i_j \in \{1, \dots, d\}$$

which in particular implies that

$$\mathbb{E}\prod_{j=1}^{3}\bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right), \quad i_j \in \{1, \dots, d\}.$$

□

Proof of Theorem B.1. To prove this result, all we need to do is check the conditions in Theorem A.2. As we apply Lemma A.1, we make the following choices:

- $b(x) = -\nabla\tilde{f}^{\text{RSAM}}(x)$;
- $\sigma(x) = \Sigma^{\text{RSAM}}(x)^{\frac{1}{2}}$.

First of all, we notice that $\forall i = 1, \dots, d$, it holds that

- $\mathbb{E}\bar{\Delta}_i \overset{\text{1. Lemma B.2}}{=} -\partial_{e_i}\tilde{f}^{\text{RSAM}}(x)\eta + \mathcal{O}(\eta^2)$;
- $\mathbb{E}\Delta_i \overset{\text{1. Lemma A.1}}{=} -\partial_{e_i}\tilde{f}^{\text{RSAM}}(x)\eta + \mathcal{O}\left(\eta^2\right)$.

Therefore, we have that for some $K_1(x) \in G$

$$\left|\mathbb{E}\Delta_i - \mathbb{E}\bar{\Delta}_i\right| \leq K_1(x)\eta^2, \quad \forall i = 1, \dots, d. \tag{127}$$

Additionally,we notice that $\forall i, j = 1, \dots, d$, it holds that

- $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j \overset{\text{2. Lemma B.2}}{=} \partial_{e_i}\tilde{f}^{\text{RSAM}}\partial_{e_j}\tilde{f}^{\text{RSAM}}\eta^2 + \Sigma^{\text{RSAM}}_{i,j}\eta^2 + \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\Delta_i\Delta_j \overset{\text{2. Lemma A.1}}{=} \partial_{e_i}\tilde{f}^{\text{RSAM}}\partial_{e_j}\tilde{f}^{\text{RSAM}}\eta^2 + \Sigma^{\text{RSAM}}_{i,j}\eta^2 + \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_2(x) \in G$

$$\left|\mathbb{E}\Delta_i\Delta_j - \mathbb{E}\bar{\Delta}_i\bar{\Delta}_j\right| \leq K_2(x)\eta^2, \quad \forall i,j = 1,\dots,d \tag{128}$$

Additionally, we notice that $\forall s \geq 3, \forall i_j \in \{1,\dots,d\}$, it holds that

- $\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j} \overset{\text{3. Lemma B.2}}{=} \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\prod_{j=1}^{s}\Delta_{i_j} \overset{\text{3. Lemma A.1}}{=} \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_3(x) \in G$

$$\left|\mathbb{E}\prod_{j=1}^{s}\Delta_{i_j} - \mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j}\right| \leq K_3(x)\eta^2. \tag{129}$$

Additionally, for some $K_4(x) \in G, \forall i_j \in \{1,\dots,d\}$

$$\mathbb{E}\prod_{j=1}^{3}\left|\bar{\Delta}_{(i_j)}\right| \overset{\text{3. Lemma B.2}}{\leq} K_4(x)\eta^2. \tag{130}$$

Finally, Eq. (127), Eq. (128), Eq. (129), and Eq. (130) allow us to conclude the proof.

□

C. Convergence Analysis: Quadratic Loss

C.1. ODE USAM

Let us study the quadratic loss function $f(x) = x^\top Hx$ where H is a diagonal matrix of eigenvalues $(\lambda_1,\dots,\lambda_d)$ such that $\lambda_1 \geq \lambda_1 \geq \cdots \geq \lambda_d$. Under the dynamics of the ODE of USAM, we have that

$$dX_t = -H\left(X_t + \rho H X_t\right)dt = -H\left(I_d + \rho H\right)X_t dt, \tag{131}$$

which, for the single component gives us the following dynamics

$$dX_t^j = -\lambda_j(1+\rho\lambda_j)X_t^j dt \tag{132}$$

whose solution is

$$X_t^j = X_0^j e^{-\lambda_j(1+\rho\lambda_j)t}. \tag{133}$$

Lemma C.1. *For all $\rho > 0$, if all the eigenvalues of H are positive, then*

$$X_t^j \overset{t\to\infty}{\to} 0, \quad \forall j \in \{1,\dots,d\} \tag{134}$$

Proof of Lemma C.1. For each $j \in \{1,\cdots,d\}$, we have that

$$X_t^j = X_0^j e^{-\lambda_j(1+\rho\lambda_j)}.$$

Therefore, since the exponent is always negative, $X_t^j \to 0$ as $t \to \infty$.

□

Lemma C.2. *Let H have at least one strictly negative eigenvalue and let λ_* be the largest negative eigenvalue of H. Then, for all $\rho > -\frac{1}{\lambda_*}$,*

$$X_t^j \overset{t\to\infty}{\to} 0, \quad \forall j \in \{1, \dots, d\}. \tag{135}$$

Proof of Lemma C.2. For each $j \in \{1, \cdots, d\}$, we have that

$$X_t^j = X_0^j e^{-\lambda_j(1+\rho\lambda_j)}.$$

Therefore, if $\lambda_j > 0$, the exponent is always negative for each value of $\rho > 0$. Therefore, $X_t^j \to 0$ as $t \to \infty$. Differently, if $\lambda_j < 0$, the exponent $-\lambda_j(1 + \rho\lambda_j)$ is negative only if $\rho > -\frac{1}{\lambda_*}$ where λ_* is the largest negative eigenvalue of H. Therefore, if$\rho > -\frac{1}{\lambda_*}$, $X_0^j \to 0$ if $t \to \infty$.

□

C.2. SDE USAM - Stationary Distribution

Let us consider the noisy quadratic model $f(x) = \frac{1}{2}x^\top Hx$, where H is a symmetric matrix. Then, based on Theorem (A.4) in the case where $\Sigma(x) = \varsigma I_d$, the corresponding SDE is give by

$$dX_t = -H\left(I_d + \rho H\right) X_t dt + \left[(I_d + \rho H)\sqrt{\eta}\varsigma\right] dW_t. \tag{136}$$

Theorem C.3 (Stationary distribution - PSD Case.)**.** *For any $\rho > 0$, the stationary distribution of Eq.* (136) *is*

$$P\left(x, \infty \mid \rho\right) = \sqrt{\frac{\lambda_i}{\pi\eta\varsigma^2}\frac{1}{1+\rho\lambda_i}}\exp\left[-\frac{\lambda_i}{\eta\varsigma^2}\frac{1}{1+\rho\lambda_i}x^2\right] \tag{137}$$

where $(\lambda_1, \dots, \lambda_d)$ are the eigenvalues of H and $\lambda_i > 0, \forall i \in \{1, \cdots, d\}$.

More interestingly, if ρ is too large, this same conclusion holds even for a saddle point.

Theorem C.4 (Stationary distribution - Indefinite Case.)**.** *Let $(\lambda_1, \dots, \lambda_d)$ are the eigenvalues of H such that there exists at least one which is strictly negative. If $\rho > -\frac{1}{\lambda_*}$ where λ_* is the largest negative eigenvalue of H, then the stationary distribution of Eq.* (136) *is*

$$P\left(x, \infty \mid \rho\right) = \sqrt{\frac{\lambda_i}{\pi\eta\varsigma^2}\frac{1}{1+\rho\lambda_i}}\exp\left[-\frac{\lambda_i}{\eta\varsigma^2}\frac{1}{1+\rho\lambda_i}x^2\right] \tag{138}$$

Proof of Theorem C.3. Note that Eq. (136) is a linear SDE, and that drift and diffusion matrices are co-diagonalizable: Let $H = U\Lambda U^\top$ be one eigenvalue decomposition of H, with $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$. If we plug this in, we get

$$dX_t = -U\left(\Lambda + \rho\Lambda^2\right)U^\top X_t dt + U\left[(I_d + \rho\Lambda)\sqrt{\eta}\varsigma\right]U^\top dW_t.$$

Let us multiply the LHS with $U^\top$, then

$$d(U^\top X_t) = -\left(\Lambda + \rho\Lambda^2\right)(U^\top X_t)dt + \left[(I_d + \rho\Lambda)\sqrt{\eta}\varsigma\right]U^\top dW_t.$$

Finally, note that $U^\top dW_t = dW_t$ in law, so we can write

$$d(U^\top X_t) = -\left(\Lambda + \rho\Lambda^2\right)(U^\top X_t)dt + \left[(I_d + \rho\Lambda)\sqrt{\eta}\varsigma\right]dW_t.$$

This means that the coordinates of the vector $Y = U^\top X$ evolve independently

$$dY_t = -\left(\Lambda + \rho\Lambda^2\right) Y_t dt + \left[(I_d + \rho\Lambda)\sqrt{\eta}\varsigma\right] dW_t,$$

since Λ is diagonal. Therefore for the i-th component Y_i we can write

$$dY_{i,t} = -\left(\lambda_i + \rho\lambda_i^2\right) Y_{i,t} dt + \left[(1 + \rho\lambda_i)\sqrt{\eta}\varsigma\right] dW_{i,t}. \tag{139}$$

Note that this is a simple one-dimensional Ornstein–Uhlenbeck process $dY_t = -\theta Y_t dt + \sigma dW_t$ ($\theta > 0, \sigma \neq 0$) with parameters

$$\theta = \lambda_i \left(1 + \rho\lambda_i\right) > 0 \quad \text{and} \quad \sigma = (1 + \rho\lambda_i)\sqrt{\eta}\varsigma > 0 \tag{140}$$

Therefore, from Section 4.4.4 of (Gardiner et al., 1985), we get that

$$\mathbb{E}[Y_t] = e^{-\theta t} Y_0, \quad \text{Var}(Y_t) = \frac{\sigma^2}{2\theta}(1 - e^{-2\theta t}). \tag{141}$$

In our case we have that

$$\mathbb{E}[Y_t] = e^{-\theta t} Y_0 \to 0 \quad \text{and} \quad \text{Var}(Y_t) = \frac{\sigma^2}{2\theta}(1 - e^{-2\theta t}) \to \frac{\sigma^2}{2\theta} = \frac{\eta\varsigma^2}{2\lambda_i}(1 + \rho\lambda_i). \tag{142}$$

Additionally, using the Fokker–Planck equation, see Section 5.3 of (Risken, 1996), we have the following formula for the stationary distribution of each eigendirection. Indeed, let us recall that for $D := \frac{\sigma^2}{2}$, the probability density function is

$$P\left(x, t \mid x', t', \rho\right) = \sqrt{\frac{\theta}{2\pi D\left(1 - e^{-2\theta(t-t')}\right)}} \exp\left[-\frac{\theta}{2D}\frac{\left(x - x' e^{-\theta\left(t-t'\right)}\right)^2}{1 - e^{-2\theta(t-t')}}\right]. \tag{143}$$

Therefore, the stationary distribution is

$$\begin{aligned} P\left(x, \infty \mid \rho\right) &= \sqrt{\frac{\theta}{2\pi D}} \exp\left[-\frac{\theta}{2D} x^2\right] \\ &= \sqrt{\frac{\theta}{\pi\sigma^2}} \exp\left[-\frac{\theta}{\sigma^2} x^2\right] \\ &= \sqrt{\frac{\lambda_i}{\pi\eta\varsigma^2}\frac{1}{1 + \rho\lambda_i}} \exp\left[-\frac{\lambda_i}{\eta\varsigma^2}\frac{1}{1 + \rho\lambda_i} x^2\right]. \end{aligned} \tag{144}$$

To conclude,

$$Y_{i,\infty} \sim \mathcal{N}\left(0, \frac{\eta\varsigma^2}{\lambda_i}(1 + \rho\lambda_i)\right). \tag{145}$$

Since all of the eigenvalues are positive, this distribution has more variance than SGD on each direction. □

Since the proof of Theorem C.4 is perfectly similar to that of Theorem C.3, we skip it. Additionally, a very analogous result holds true even if all the eigenvalues are strictly negative and thus the quadratic has a single maximum as a critical point. From these results, we understand that under certain circumstances, USAM might be attracted not only by the minimum but possibly also by a saddle or a maximum. This is fully consistent with the results derived for the ODE of USAM in Lemma C.1 and Lemma C.2.

Observation C.5 (Suboptimality under the Stationary Distribution – comparison to SGD). In the special case where the stochastic process has reached stationarity, one can approximate the loss landscape with a quadratic loss (Jastrzebski et al., 2018). By further assuming that $\Sigma^{SGD}(x) = H$ (see e.g.(Sagun et al., 2018; Zhu et al., 2019)), Theorem (A.4) implies that for USAM

$$dX_t = -H\left(I_d + \rho H\right) X_t dt + \left[(I_d + \rho H)\sqrt{\eta}\sqrt{H}\right] dW_t. \tag{146}$$

Up to a change of variable, we assume H to be diagonal and therefore

$$\mathbb{E}_{USAM}\left[f\right] = \frac{1}{2}\sum_{i=1}^{d}\lambda_i \mathbb{E}[X_i^2] = \frac{\eta}{4}\sum_{i=1}^{d}\lambda_i(1+\rho\lambda_i)^2 = \frac{\eta}{4}\left(Tr(H) + 2\rho Tr(H^2) + \rho^2 Tr(H^3)\right) \gg \mathbb{E}_{SGD}\left[f\right], \tag{147}$$

where subscripts indicate that f is being optimized with SGD and USAM, respectively.
Regarding DNSAM, Theorem (A.9) implies that

$$dX_t = -H\left(I_d + \frac{\rho H}{\|HX_t\|}\right) X_t dt + \sqrt{\eta}\sqrt{H}\left(I_d + \frac{\rho H}{\|HX_t\|}\right) dW_t. \tag{148}$$

Therefore, we argue that DNSAM has to have a suboptimality with respect to SGD which is even larger than that of USAM. Intuitively, when $\|HX_t\| < 1$, the variance of DNSAM is larger than that of USAM. Therefore, its suboptimality has to be larger as well. Finally, the behavior of SAM is close to that of DNSAM, but less pronounced because the denominator can never get too close to 0 due to the noise injection.

C.3. ODE SAM

W.l.o.g, we take H to be diagonal and if it has negative eigenvalues, we denote the largest negative eigenvalue with λ_*. Let us recall that the ODE of SAM for the quadratic loss is given by

$$dX_t = -H\left(I_d + \frac{\rho H}{\|HX_t\|}\right) X_t dt \tag{149}$$

Lemma C.6. *For all $\rho > 0$, if H is PSD, the origin is (locally) asymptotically stable. Additionally, if H is not PSD, if $\|HX_t\| \leq -\rho\lambda_*$, then the origin is still (locally) asymptotically stable.*

Proof of Lemma C.6. Let $V(x) := \frac{x^\top K x}{2}$ be the Lyapunov function, where K is a diagonal matrix with positive eigenvalues $(k_1, \cdots, k_d)$. Therefore, we have

$$V(X_t) = \frac{1}{2}\sum_{i=1}^{d} k_i \left(X_t^i\right)^2 > 0 \tag{150}$$

and

$$\dot{V}(X_t) = \sum_{i=1}^{d} k_i X_t^i \dot{X}_t^i = \sum_{i=1}^{d} k_i(-\lambda_i)\left(1 + \frac{\rho\lambda_i}{\|HX_t\|}\right) X_t^i X_t^i dt = -\sum_{i=1}^{d} k_i\lambda_i\left(1 + \frac{\rho\lambda_i}{\|HX_t\|}\right)\left(X_t^i\right)^2 dt. \tag{151}$$

Let us analyze the terms

$$k_i\lambda_i\left(1 + \frac{\rho\lambda_i}{\|HX_t\|}\right)\left(X_t^i\right)^2.$$

When $\lambda_i > 0$, these quantities are all positive and the proof is concluded. However, if there exists $\lambda_i < 0$, these quantities are positive only if $\left(1 + \frac{\rho\lambda_i}{\|HX_t\|}\right) \leq 0$, that is if $\|HX_t\| \leq -\rho\lambda_i$. Therefore, a sufficient condition for $\dot{V}(X_t) \leq 0$ is that

$$\|HX_t\| \leq -\rho\lambda_i, \quad \forall i \text{ s.t. } \lambda_i < 0.$$

Based on Theorem 1.1 of (Mao, 2007), we conclude that if $\|HX_t\| \leq -\rho$ where λ_* is the largest negative eigenvalue of H, $V(X_t) > 0$ and $\dot{V}(X_t) \leq 0$, and that therefore the dynamics of X_t is bounded inside this compact set and cannot diverge. □

From this result, we understand that the dynamics of the ODE of USAM might converge to a saddle or even a maximum if it gets too close to it.

C.4. SDE DNSAM

W.l.o.g, we take H to be diagonal and if it has negative eigenvalues, we denote the largest negative eigenvalue with λ_*. Based on Eq. (152) and in the case where $\Sigma^{\text{SGD}} = \varsigma^2 I_d$, the SDE of DNSAM for the quadratic loss is given by

$$dX_t = -H\left(I_d + \frac{\rho H}{\|HX_t\|}\right) X_t dt + \sqrt{\eta}\varsigma\left(I_d + \frac{\rho H}{\|HX_t\|}\right) dW_t \tag{152}$$

Observation C.7. For all $\rho > 0$, there exists an $\epsilon > 0$ such that if $\|HX_t\| \in (\epsilon, -\rho\lambda_*)$, the dynamics of X_t is attracted towards the origin. If the eigenvalues are all positive, the condition is $\|HX_t\| \in (\epsilon, \infty)$. On the contrary, if $\|HX_t\| < \epsilon$, then the dynamics is pushed away from the origin.

Formal calculations to support Observation C.7. Let $V(t,x) := e^{-t\frac{x^\top Kx}{2}}$ be the Lyapunov function, where K is a diagonal matrix with strictly positive eigenvalues $(h_1, \cdots, h_d)$. Therefore, we have

$$V(X_t) = e^{-t}\frac{1}{2}\sum_{i=1}^{d} k_i \left(X_t^i\right)^2 > 0 \tag{153}$$

and

$$\begin{aligned} LV(t,X_t) &= -e^{-t}\frac{1}{2}\sum_{i=1}^{d} k_i \left(X_t^i\right)^2 + e^{-t}\sum_{i=1}^{d} k_i(-\lambda_i)\left(1 + \frac{\rho\lambda_i}{\|HX_t\|}\right)\left(X_t^i\right)^2 + e^{-t}\frac{\eta\varsigma^2}{2}\sum_{i=1}^{d} k_i\left(1 + \frac{\rho\lambda_i}{\|HX_t\|}\right)^2 \\ &= -e^{-t}\left(\frac{1}{2}\sum_{i=1}^{d} k_i \left(X_t^i\right)^2 + \sum_{i=1}^{d} k_i\lambda_i\left(1 + \frac{\rho\lambda_i}{\|HX_t\|}\right)\left(X_t^i\right)^2 - \frac{\eta\varsigma^2}{2}\sum_{i=1}^{d} k_i\left(1 + \frac{\rho\lambda_i}{\|HX_t\|}\right)^2\right) \end{aligned} \tag{154}$$

Let us analyze the terms

$$k_i\lambda_i\left(1 + \frac{\rho\lambda_i}{\|HX_t\|}\right)\left(X_t^i\right)^2.$$

When $\lambda_i > 0$, these quantities are all positive. When $\lambda_i < 0$, these quantities are positive only if $\left(1 + \frac{\rho\lambda_i}{\|HX_t\|}\right) \leq 0$, that is if $\|HX_t\| \leq -\rho\lambda_i$. Let us now assume that

$$\|HX_t\| \leq -\rho\lambda_i, \quad \forall i \text{ s.t. } \lambda_i < 0.$$

that is, $\|HX_t\| \leq -\rho\lambda_*$ such that λ_*. Then, we observe that

- If $\|HX_t\| \to 0$, $LV(t,X_t) \geq 0$
- If ς is small enough, for $\|HX_t\| \approx -\rho\lambda_*$, $LV(t,X_t) \leq 0$

Given that all functions and functionals involved are continuous, there exists $\epsilon > 0$ such that

- If $\|HX_t\| < \epsilon$, $LV(t,X_t) \geq 0$
- If ς is small enough, for $\|HX_t\| \in (\epsilon, -\rho\lambda_*)$, $LV(t,X_t) \leq 0$

□

Based on Theorem 2.2 of (Mao, 2007), we understand that if the dynamics is sufficiently close to the origin, it gets pulled towards it, but if gets too close, it gets repulsed from it. If there is no negative eigenvalue, the same happens but the dynamics can never get close to the minimum.

D. Experiments

In this section, we provide additional details regarding the validation that the SDEs we proposed indeed weakly approximate the respective algorithms. We do so on a quadratic landscape, on a classification task with a deep linear model, a binary classification task with a deep nonlinear model, and a gregression task with a teacher-student model. Since our SDEs prescribe the calculation of the Hessian of the whole neural network at each iteration step, this precludes us from testing our theory on large-scale models.

D.1. SDE Validation

Quadratic In this paragraph, we provide the details of the Quadratic experiment. We optimize the loss function $f(x) = \frac{1}{2}x^\top Hx$ of dimension $d = 20$. The Hessian H is a random SPD matrix generated using the standard Gaussian matrix $A \in \mathbb{R}^{d\times 2d}$ as $H = AA^\top/(2d)$. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \Sigma)$ where $\Sigma = \sigma I_d$ and $\sigma = 0.01$. We use $\eta = 0.01$, $\rho \in \{0.001, 0.01, 0.1, 0.5\}$. The results are averaged over 3 experiments.

Deep Linear Classification In this paragraph, we provide the details of the Deep Linear Classification experiment. This is a classification task on the Iris Database (Dua & Graff, 2017). The model is a Linear MLP with 1 hidden layer with a width equal to the number of features and we optimize the cross-entropy loss function. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \Sigma)$ where $\Sigma = \sigma I_d$ and $\sigma = 0.01$. We use $\eta = 0.01$, $\rho \in \{0.001, 0.01, 0.1, 0.2\}$. The results are averaged over 3 experiments.

Deep Nonlinear Classification In this paragraph, we provide the details of the Deep Nonlinear Classification experiment. This is a binary classification task on the Breast Cancer Database (Dua & Graff, 2017). The model is a Nonlinear MLP with 1 hidden layer with a width equal to the number of features, sigmoid activation function, and we optimize the ℓ^2-regularized logistic loss with parameter $\lambda = 0.1$. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \Sigma)$ where $\Sigma = \sigma I_d$ and $\sigma = 0.01$. We use $\eta = 0.01$, $\rho \in \{0.001, 0.01, 0.1, 0.5\}$. The results are averaged over 3 experiments.

Deep Teacher-Student Model In this paragraph, we provide the details of the Teacher-Student experiment. This is a regression task where the database is generated by the Teacher model based on random inputs in $\mathbb{R}^5$ and output in $\mathbb{R}$. The Teacher model is a deep linear MLP with 20 hidden layers with 10 nodes, while the Student is a deep nonlinear MLP with 20 hidden layers and 10 nodes. We optimize the MSE loss. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \Sigma)$ where $\Sigma = \sigma I_d$ and $\sigma = 0.001$. We use $\eta = 0.001$, $\rho \in \{0.0001, 0.001, 0.03, 0.05\}$. The results are averaged over 3 experiments.

D.1.1. Importance of the Additional Noise.

In this section, we empirically test the importance of using the correct diffusion terms the USAM SDE Eq. (9) and the DNSAM SDE Eq. (10). Let us introduce two new SDEs where the diffusion term is the one of the SGD SDE in Eq. (18) rather than that from the correct SDEs:

$$dX_t = -\nabla \tilde{f}^{\text{USAM}}(X_t)dt + \sqrt{\eta\left(\Sigma^{\text{SGD}}(X_t)\right)}dW_t, \quad \text{where} \quad \tilde{f}^{\text{USAM}}(x) := f(x) + \frac{\rho}{2}\|\nabla f(x)\|_2^2. \tag{155}$$

$$dX_t = -\nabla \tilde{f}^{\text{DNSAM}}(X_t)dt + \sqrt{\eta\left(\Sigma^{\text{SGD}}(X_t)\right)}dW_t, \quad \text{where} \quad \tilde{f}^{\text{DNSAM}}(x) := f(x) + \rho\|\nabla f(x)\|_2. \tag{156}$$

In Figure 6 we observe how approximating USAM with the SGD SDE (Eq. (18)) brings a large error in all four cases. Introducing the correct drift but excluding the correct covariance, i.e. using Eq. (155), reduces the error, but the best performer is the complete USAM SDE Eq. (7). From Figure 7, the same observations hold for DNSAM.

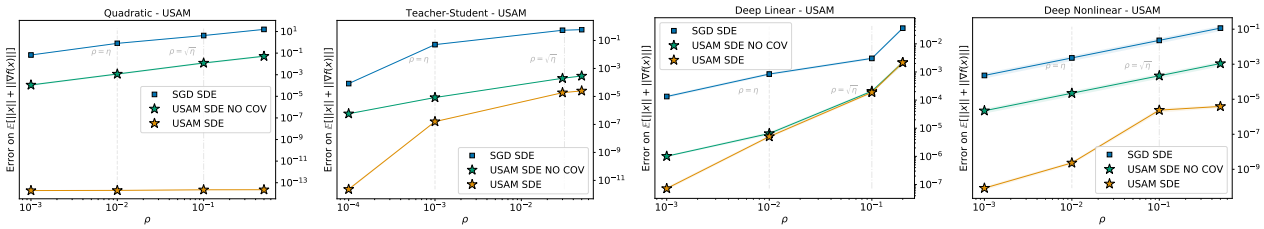

Figure 6. USAM - Comparison in terms of $g_1(x)$ with respect to ρ - Quadratic (left); Teacher-Student (center-left); Deep linear class (center-right); Deep Nonlinear class (right).

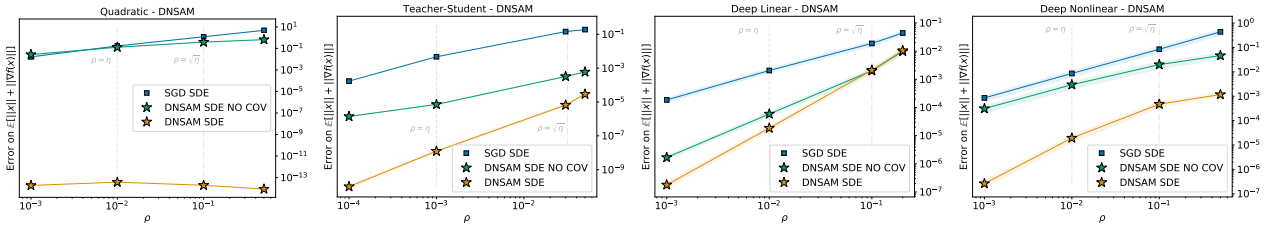

Figure 7. DNSAM - Comparison in terms of $g_1(x)$ with respect to ρ - Quadratic (left); Teacher-Student (center-left); Deep linear class (center-right); Deep Nonlinear class (right).

From these experiments, we understand that naively adding noise to the ODEs of USAM and SAM does not provide SDEs with sufficient approximation power. This is consistent with our proofs.

D.2. Quadratic Landscape

Interplay Between Hessian, ρ, and the Noise. In this paragraph, we provide additional details regarding the interplay between the Hessian, ρ, and the noise. In the first experiment represented in Figure 8, we fix $\rho = \sqrt{\eta}$, where $\eta = 0.001$ is the learning rate. Then, we fix the Hessian $H \in \mathbb{R}^{100\times 100}$ to be diagonal with random positive eigenvalues. Then, we select the scaling factors $\sigma \in \{1, 2, 4\}$. For each value of σ, we optimize the quadratic loss with SGD and DNSAM where the hessian is scaled up by a factor σ. The starting point is $x_0 = (0.02, \cdots, 0.02)$ and the number of iterations is 20000. The results are averaged over 5 runs.

In the second experiment represented in Figure 11, we fix the Hessian H with random positive eigenvalues. then, we select $\rho = \sqrt{\eta}$, where $\eta = 0.001$ is the learning rate. Then, we select the scaling factors $\sigma \in \{1, 2, 4\}$. For each value of σ, we optimize the quadratic loss with SGD and DNSAM where the hessian is fixed and ρ is scaled up by a factor σ. The starting point is $x_0 = (0.02, \cdots, 0.02)$ and the number of iterations is 20000. The results are averaged over 5 runs.

The very same setup holds for the experiments carried out for USAM and is represented in Figure 9 and Figure 13.

The very same setup holds for the experiments carried out for SAM and is represented in Figure 10 and Figure 12.

Stationary Distribution Convex Case In this paragraph, we provide the details of the experiment about the dynamics of the SDE of DNSAM in the quadratic convex case of dimension 2. The hessian is diagonal with both eigenvalues equal to 1. We select $\rho = \sqrt{\eta}$, where $\eta = 0.001$ is the learning rate. In the first image on the left of Figure 14, we show the distribution of 10^5 trajectories all starting at $(0.02, 0.02)$ after $5 \cdot 10^4$ iterations. In the second image, we plot the number of trajectories that at a certain time are inside a ball of radius 0.007, e.g. close to the origin. As we can see in greater detail in the third image, all of them are initialized outside such a ball, then they get attracted inside, and around the 600-th iteration they get repulsed out of it. We highlight that the proportion of points inside/outside the ball is relatively stable. In the fourth image, we count the number of trajectories that are jumping in or out of such a ball. All of the trajectories enter the ball between the 400-th and 500-th iteration, and then they start jumping in and out after the iteration 600. We conclude that this experimental evidence are supporting the claim that the origin attracts the dynamics, but repulses it at the moment that the

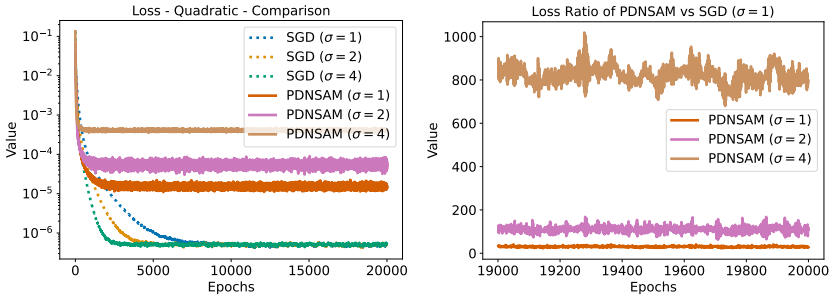

Figure 8. Role of the Hessian - Left: Comparison between SGD and DNSAM for fixed rho and larger Hessians. Right: Ratio between the Loss of DNSAM for different scaling of the Hessian by the loss of the unscaled case of SGD.

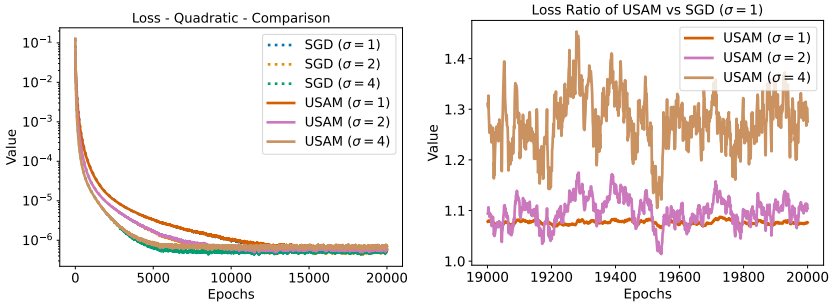

Figure 9. Role of the Hessian - Left: Comparison between SGD and USAM for fixed rho and larger Hessians. Right: Ratio between the Loss of USAM for different scaling of the Hessian by the loss of the unscaled case of SGD.

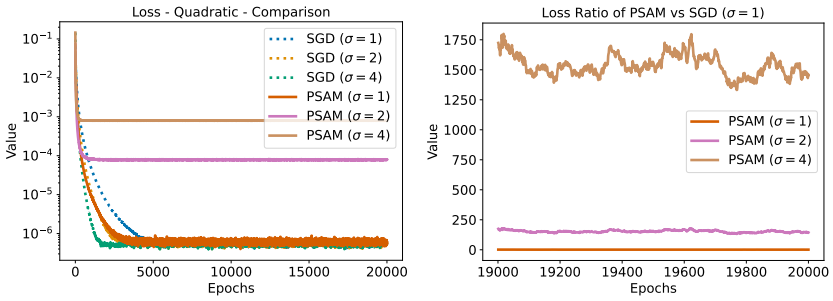

Figure 10. Role of the Hessian - Left: Comparison between SGD and SAM for fixed rho and larger Hessians. Right: Ratio between the Loss of PSAM for different scaling of the Hessian by the loss of the unscaled case of SGD.

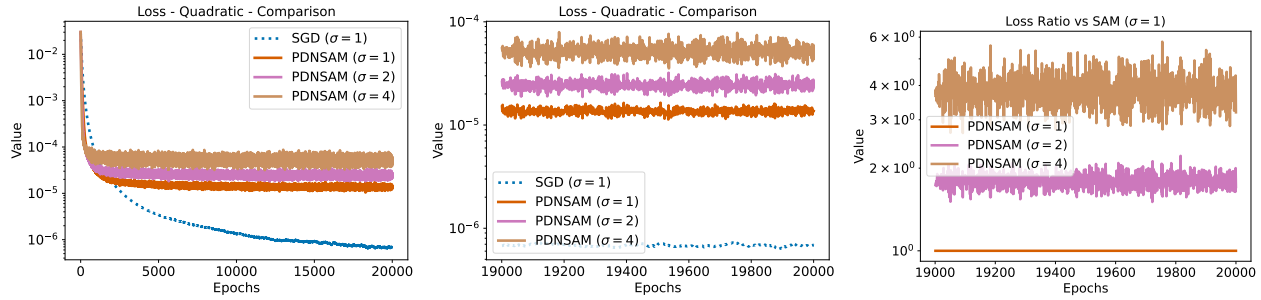

Figure 11. Role of ρ - Left: Comparison between SGD and DNSAM for fixed hessian and larger ρ values. Center: Zoom at convergence. Right: Ratio between the Loss of DNSAM for different scaling of the Hessian by the loss of the unscaled case of DNSAM.

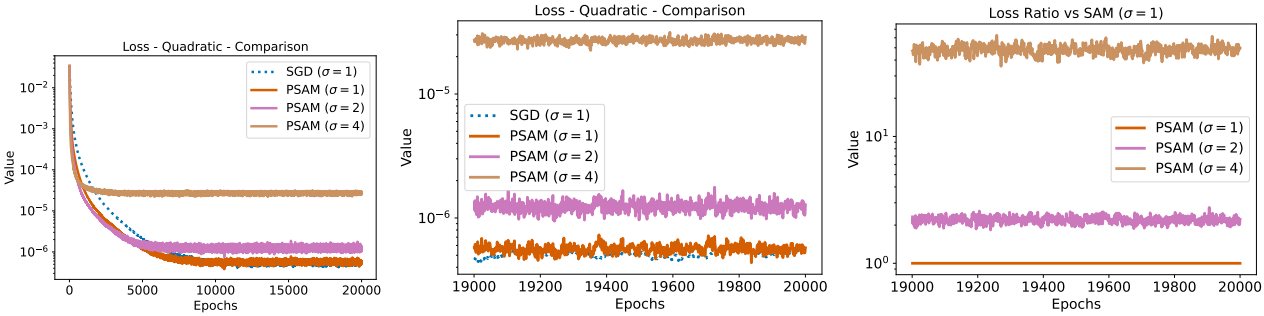

Figure 12. Role of ρ - Left: Comparison between SGD and PSAM for fixed hessian and larger ρ values. Center: Zoom at convergence. Right: Ratio between the Loss of PSAM for different scaling of the Hessian by the loss of the unscaled case of PSAM.

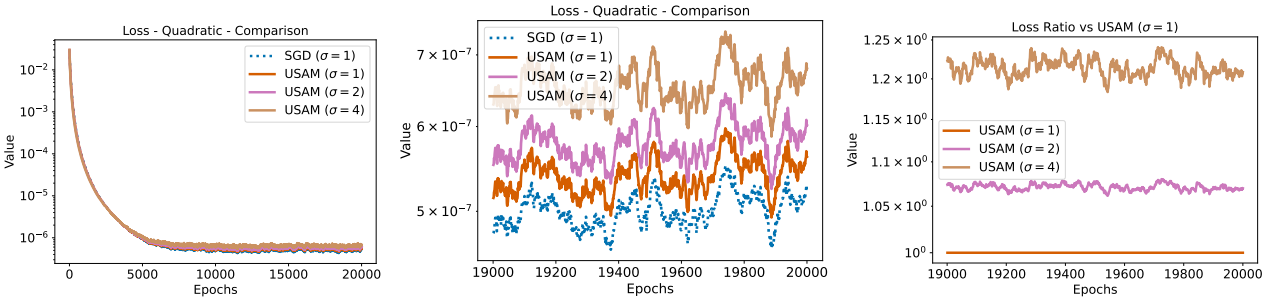

Figure 13. Role of ρ - Left: Comparison between SGD and USAM for fixed hessian and larger ρ values. Center: Zoom at convergence. Right: Ratio between the Loss of USAM for different scaling of the Hessian by the loss of the unscaled case of USAM.

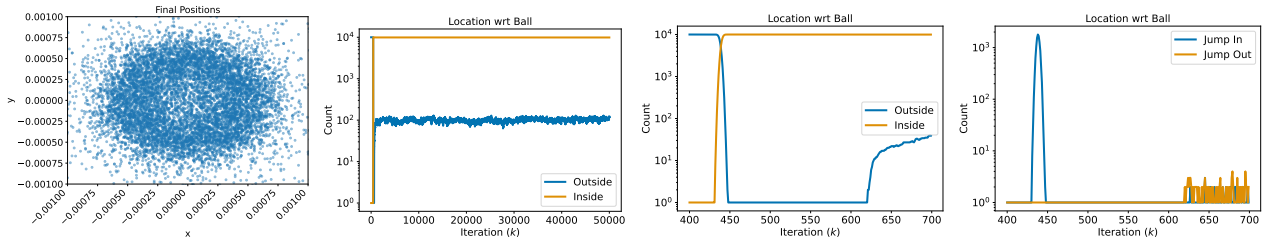

Figure 14. Convex Quadratic - Left: Distribution points around the origin is scarcer near to the origin; Center-Left: Number of trajectories outside a small ball around the origin increases over time; Center-Right: All the trajectories eventually enter the ball and then start exiting it; Right: There is a constant oscillation of points in and out of the ball.

trajectories get too close to it.

Stationary Distribution Saddle Case In this paragraph, we provide the details of the experiment about the dynamics of the SDE of DNSAM in the quadratic saddle case of dimension 2. The hessian is diagonal with eigenvalues equal to 1 and -1. We select $\rho = \sqrt{\eta}$, where $\eta = 0.001$ is the learning rate. In the first image on the left of Figure 4, we show the distribution of 10^5 trajectories all starting at $(0.02, 0.02)$ after $5 \cdot 10^4$ iterations. In the second image, we plot the number of trajectories that at a certain time are inside a ball of radius 0.007, e.g. close to the origin. As we can see in greater detail in the third image, all of them are initialized outside such a ball, then they get attracted inside, and around the 1200-th iteration they get repulsed out of it. We highlight that the proportion of points outside the ball is stably increasing, meaning that the trajectories are slowly escaping from the saddle. In the fourth image, we count the number of trajectories that are jumping in or out of such a ball. All of the trajectories enter the ball between the 950-th and 1000-th iteration, and then they start jumping in and out after the iteration 1200. We conclude that this experimental evidence are supporting the claim that the origin attracts the dynamics, but repulses it at the moment that the trajectories get too close to it, even when this is a saddle.

Escaping the Saddle - Low Dimensional In this paragraph, we provide details for the Escaping the Saddle experiment in dimension $d = 2$. As in the previous experiment, the saddle is a quadratic of dimension 2 and its hessian is diagonal with eigenvalues equal to 1 and -1. We select $\rho = \sqrt{\eta}$, where $\eta = 0.001$ is the learning rate. We initialize the GD, USAM, SAM, SGD, PUSAM, DNSAM, and PSAM in the point $x_0 = (0, 0.01)$, e.g. in the direction of the fastest escape from the saddle. In the left of Figure 15, we observe that GD and USAM manage to escape the saddle while SAM remains stuck. We highlight that running for more iterations would not change this as SAM is oscillating across the origin. In the second figure, we observe that the stochastic optimizers escape the saddle quicker than their deterministic counterpart and even PSAM and DNSAM manage to escape. Results are averaged over 3 runs.

Escaping the Saddle - High Dimensional In this paragraph, we provide details for the Escaping the Saddle experiment in dimension $d = 400$. We fix the Hessian $H \in \mathbb{R}^{400\times 400}$ to be diagonal with random positive eigenvalues. To simulate a saddle, we flip the sign of the smallest 10 eigenvalues. We select $\rho = \sqrt{\eta}$, where $\eta = 0.001$ is the learning rate. We study the optimization dynamics of SAM, PSAM, and DNPSAM as we initialize the process closer and closer to the saddle in the origin. The starting point $x_0 = (1, \cdots, 1)$ is scaled with factors $\sigma \in \{10^0, 10^{-4}, 10^{-8}\}$ and we notice that the one scaled with $\sigma = 1$ escapes slowly from the saddle. The one scaled with $\sigma = 10^{-8}$ experiences a sharp spike in volatility and jumps away from the origin and ends up escaping the saddle faster than the previous case. Finally, the one scaled with $\sigma = 10^{-4}$ stays trapped in the saddle. Results are represented in Figure 15. Results are averaged over 3 runs.

D.3. Linear Autoencoder

In this paragraph, we provide additional details regarding the Linear Autoencoder experiment. In this experiment, we approximate the Identity matrix of dimension 20 as the product of two square matrices $W1$ and $W2$. As described in (Kunin et al., 2019), there is a saddle of the loss function around the origin. Inspired by the results obtained for the quadratic landscape, we test if SAM and its variants struggle to escape this saddle as well. To do this, we initialize the two matrices with entries normally distributed. We select $\rho = \sqrt{\eta}$, where $\eta = 0.001$ is the learning rate. Then, we study the dynamics of the optimization process in case we scale the matrices by a factor $\sigma \in \{10^{-2}, 10^{-3}, 5 \cdot 10^{-3}, 10^{-4}, 10^{-5}\}$. As we can see from the first image of Figure 5, initializing SAM far from the origin, that is $\sigma = 0.01$, allows SAM to optimize the

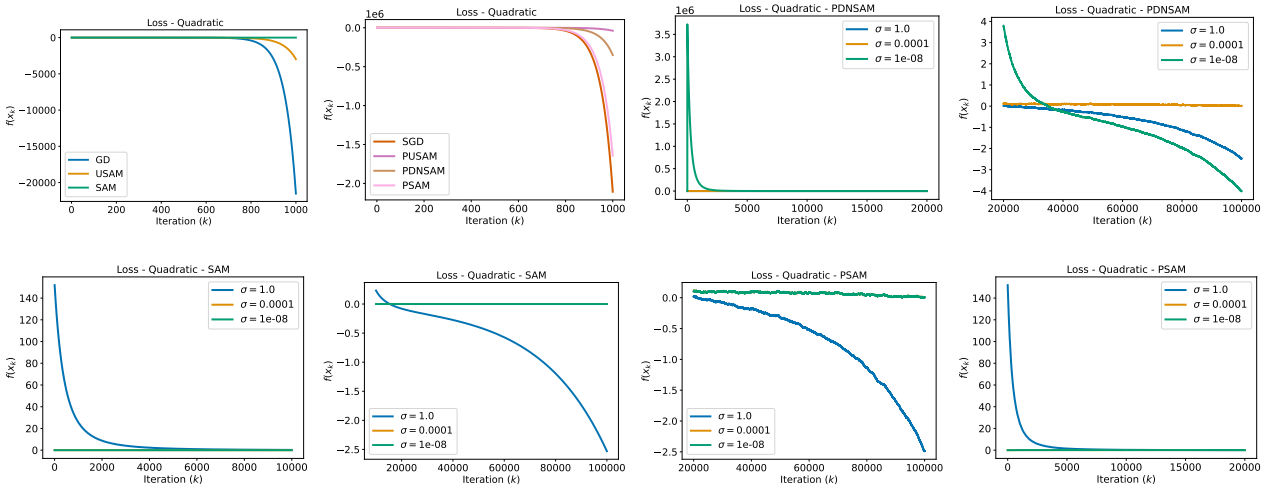

Figure 15. Escaping the Saddle - Top-Left: Comparison between GD, SAM, and USAM at escaping from a quadratic saddle. Top-Center-Left: Comparison between PGD, PSAM, DNSAM, and PUSAM at escaping from a quadratic saddle. Top-Center-Right: If close enough to the origin, DNSAM escapes from the origin immediately due to a volatility spike. Top-Right: The DNSAM initialized far from the origin starts escaping from it and the PSAM which jumped away from it efficiently escapes the saddle.
Bottom: Both SAM and PSAM get stuck in the origin if initialized too close to it.

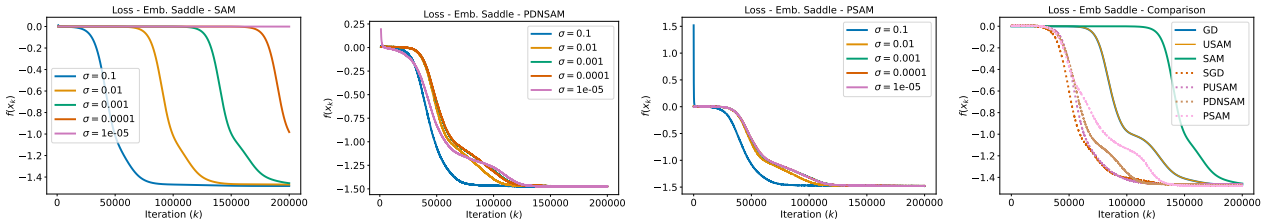

Figure 16. Embedded Saddle - Left: SAM does not escape the saddle if it is too close to it. Center-Left: DNSAM always escapes it, but more slowly if initialized closer to the origin. If extremely close, it recovers speed thanks to a volatility spike. Center-Right: Similarly to SAM, PSAM gets progressively slower the closer it gets initialized to the origin. Right: SAM is stuck while the other optimizers manage to escape.

loss. Decreasing σ implies that SAM becomes slower and slower at escaping the loss up to not being able to escape it anymore. The second image shows that the very same happens if we use DNSAM. However, if the process is initialized extremely close to the origin, that is $\sigma = 10^{-5}$, then the process enjoys a volatility spike that pushes it away from the origin. This allows the process to escape the saddle quickly and effectively. In the third image, we observe that similarly to SAM, PSAM becomes slower at escaping if σ is lower. In the fourth image, we compare the behavior of GD, USAM, SAM, PGD, PUSAM, DNSAM, and PSAM where $\sigma = 10^{-5}$. We observe that DNSAM is the fastest algorithm to escape the saddle, followed by the others. As expected, SAM does not escape. Results are averaged over 3 runs.

D.4. Embedded Saddle

In this paragraph, we provide additional details regarding the Embedded Saddle experiment. In this experiment, we optimize a regularized quadratic d-dimensional landscape $L(x) = \frac{1}{2}x^\top Hx + \lambda \sum_{i=1}^{d} x_i^4$. As described in (Lucchi et al., 2022), if H is not PSD, there is a saddle of the loss function around the origin and local minima away from it. We fix the Hessian $H \in \mathbb{R}^{400\times 400}$ to be diagonal with random positive eigenvalues. To simulate a saddle, we flip the sign of the smallest 10 eigenvalues. The regularization parameter is fixed at $\lambda = 0.001$. We use $\eta = 0.005$, $\rho = \sqrt{\eta}$, run for 200000 and average over 3 runs. In Figure 16 we see the very same observations we had for the Autoencoder.

Name	Algorithm	Theorem for SDE
SGD	$x_{k+1} = x_k - \eta \nabla f_{\gamma_k}(x_k)$	Theorem 1 (Li et al., 2017)
SAM	$x_{k+1} = x_k - \eta \nabla f_{\gamma_k}\left(x_k + \rho \frac{\nabla f_{\gamma_k}(x_k)}{\|\nabla f_{\gamma_k}(x_k)\|}\right)$	Theorem 3.5
USAM	$x_{k+1} = x_k - \eta \nabla f_{\gamma_k}\left(x_k + \rho \nabla f_{\gamma_k}(x_k)\right)$	Theorem 3.2
DNSAM	$x_{k+1} = x_k - \eta \nabla f_{\gamma_k}\left(x_k + \rho \frac{\nabla f_{\gamma_k}(x_k)}{\|\nabla f_{\gamma_k}(x_k)\|}\right)$	Not Available
PGD	$x_{k+1} = x_k - \eta \nabla f(x_k) + \eta Z$	Theorem 1 (Li et al., 2017)
PSAM	$x_{k+1} = x_k - \eta \nabla f\left(x_k + \rho \frac{\nabla f(x_k)+Z}{\|\nabla f(x_k)+Z\|}\right) + \eta Z$	Theorem 3.6
PUSAM	$x_{k+1} = x_k - \eta \nabla f\left(x_k + \rho \nabla f(x_k) + Z\right) + \eta Z$	Theorem 3.3
PDNSAM	$x_{k+1} = x_k - \eta \nabla f\left(x_k + \rho \frac{\nabla f(x_k)+Z}{\|\nabla f(x_k)\|}\right) + \eta Z$	Theorem 3.4

Table 1. Comparison of algorithms for methods analyzed in the paper. The learning rate is η, the radius is ρ, and $Z \sim \mathcal{N}(0, \Sigma)$ is the injected noise.

Name	Drift Term
SGD	$-\nabla f(x)$
SAM	$-\nabla \left(f(x) + \rho \mathbb{E}\left[\|\nabla f_\gamma(x)\|_2\right]\right)$
USAM	$-\nabla \left(f(x) + \frac{\rho}{2} \mathbb{E}\left[\|\nabla f_\gamma(x)\|_2^2\right]\right)$
PGD	$-\nabla f(x)$
PSAM	$-\nabla \left(f(x) + \rho \mathbb{E}\left[\|\nabla f_\gamma(x)\|_2\right]\right)$
PUSAM	$-\nabla \left(f(x) + \frac{\rho}{2} \|\nabla f(x)\|_2^2\right)$
(P)DNSAM	$-\nabla \left(f(x) + \rho \|\nabla f(x)\|_2\right)$

Table 2. Comparison of the drift terms of the SDEs for methods analyzed in the paper.

Name	Diffusion Term	$\tilde{\Sigma}(x)$
SGD	$\sqrt{\eta} \left(\Sigma(x)\right)^{\frac{1}{2}}$	
SAM	$\sqrt{\eta \left(\Sigma(x) + \rho \left(\tilde{\Sigma}(x) + \tilde{\Sigma}(x)^\top\right)\right)}$	$\mathbb{E}\left[\left(\nabla f(x) - \nabla f_\gamma(x)\right) \cdot \left(\mathbb{E}\left[\frac{H_\gamma(x)\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right] - \frac{H_\gamma(x)\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right)^\top\right]$
USAM	$\sqrt{\eta \left(\Sigma(x) + \rho \left(\tilde{\Sigma}(x) + \tilde{\Sigma}(x)^\top\right)\right)}$	$\mathbb{E}\left[\left(\nabla f(x) - \nabla f_\gamma(x)\right) \left(\mathbb{E}\left[H_\gamma(x)\nabla f_\gamma(x)\right] - H_\gamma(x)\nabla f_\gamma(x)\right)^\top\right]$
PGD	$\sqrt{\eta\Sigma(x)}$	
PSAM	$\sqrt{\eta \left(\Sigma(x) + \rho \left(\bar{\Sigma}(x) + \bar{\Sigma}(x)^\top\right)\right)}$	$H(x)\mathbb{E}\left[\left(\nabla f(x) - \nabla f_\gamma(x)\right) \cdot \left(\mathbb{E}\left[\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right] - \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right)^\top\right]$
PUSAM	$\left(I_d + \rho H(x)\right) \left(\eta\Sigma(x)\right)^{1/2}$	
(P)DNSAM	$\left(I_d + \rho \frac{H(x)}{\|\nabla f(x)\|_2}\right) \left(\eta\Sigma(x)\right)^{\frac{1}{2}}$	

Table 3. Comparison of the diffusion terms of SDEs for methods analyzed in the paper. The matrix $\Sigma(x)$ is equal to $\Sigma(x)^{\text{SGD}}$ and $H(x) = \nabla^2 f(x)$.