Implicit Bias in Noisy-SGD: With Applications to Differentially Private Training

Tom Sander
Meta AI (FAIR) & École polytechnique*

Maxime Sylvestre
Université Paris Dauphine

Alain Durmus
École polytechnique*

Abstract

Training Deep Neural Networks (DNNs) with small batches using Stochastic Gradient Descent (SGD) yields superior test performance compared to larger batches. The specific noise structure inherent to SGD is known to be responsible for this implicit bias. DP-SGD, used to ensure differential privacy (DP) in DNNs' training, adds Gaussian noise to the clipped gradients. Surprisingly, large-batch training still results in a significant decrease in performance, which poses an important challenge because strong DP guarantees necessitate the use of massive batches. We first show that the phenomenon extends to Noisy-SGD (DP-SGD without clipping), suggesting that the stochasticity (and not the clipping) is the cause of this implicit bias, *even with additional isotropic Gaussian noise.* We theoretically analyse the solutions obtained with continuous versions of Noisy-SGD for the Linear Least Square and Diagonal Linear Network settings, and reveal that the implicit bias is indeed amplified by the additional noise. Thus, the performance issues of large-batch DP-SGD training are rooted in the same underlying principles as SGD, offering hope for potential improvements in large batch training strategies.

1 Introduction

In Machine Learning, the Gradient Descent (GD) algorithm is used to minimize an empirical loss function by iteratively updating the model parameters in the direction opposite to the gradient. Its stochastic variant, Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951; Duflo, 1996) uses a random subset of the training data at each step, known as a mini-batch, to estimate the true gradient. It enables the training of machine learning models on vast datasets or with extremely large models, where the computation of the full gradient becomes too computationally intensive. Especially in Deep Learning, SGD or its variant has proven to be a crucial tool for training DNNs, delivering exceptional performance across various domains including computer vision (He et al., 2016), natural language processing (Devlin et al., 2018; Touvron et al., 2023), and speech recognition (Amodei et al., 2016).

SGD can achieve better performance than GD under a fixed compute budget (*i.e.,* number of epochs), as well

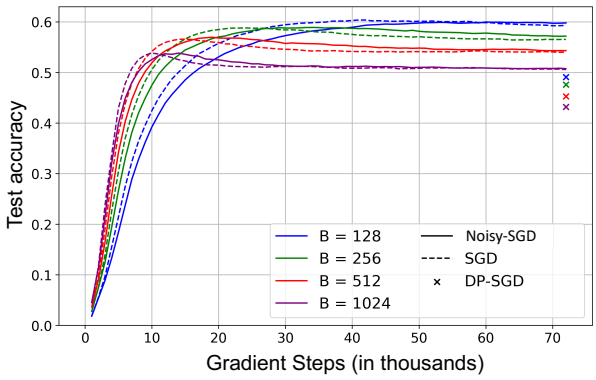

Figure 1: Training from scratch on ImageNet for $S = 72k$ steps, using a constant learning rate, with different batch sizes B. The effective noise σ/B is constant within DP-SGD and Noisy-SGD experiments. The crosses for DP-SGD are obtained from Sander et al. (2023). We observe a similar phenomenon for the non-clipped version (Noisy-SGD), *i.e.*, small batches perform better that larger ones, suggesting that clipping is not solely responsible. Even with isotropic noise added with greater magnitude than the gradients, SGD's implicit bias persists: the natural noise structure in SGD is robust to Gaussian perturbations.

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. *Centre de Mathématiques Appliquées CMAP, CNRS UMR 7641, École Polytechnique, Institut Polytechnique de Paris. Correspondence at tomsander@meta.com.

as at a fixed number of steps budget (*i.e.,* number of updates) (Keskar et al., 2016; Smith et al., 2020; Masters and Luschi, 2018). Consequently, not only does SGD offer significant computational resource savings, but its inherent stochastic nature introduces randomness into the algorithm which facilitates faster convergence and enhance generalization performance, by allowing the iterates to escape from unfavorable local minima. On simpler, overparameterized model architectures, the distinctive noise structure of SGD is recognized as a factor responsible for yielding superior solutions compared to Gradient Descent, which is referred to as its "implicit bias" (HaoChen et al., 2020).

DNNs possess the ability to grasp the general statistical patterns and trends within their training data distribution —such as grammar rules for language models— but also to memorize specific and precise details about individual data points, including sensitive information like credit card numbers (Carlini et al., 2019, 2021). This capability raises concerns regarding privacy, as accessing a trained model may lead to the exposure of training data, potentially jeopardizing privacy. One possible solution to address this issue is Differential Privacy (DP) (Dwork et al., 2006), which theoretically controls the amount of information learned from each individual training sample.

Differentially Private Stochastic Gradient Descent (DP-SGD) (Abadi et al., 2016) offers a robust framework by providing strict DP guarantees for DNNs. The gradient of each training sample is clipped and Gaussian noise is subsequently introduced to the sum before the update. Efforts to enhance the trade-off between privacy and utility in DP-SGD training have been associated with the use of exceedingly large batches (Yu et al., 2021; De et al., 2022; Yu et al., 2023). A recent observation by Sander et al. (2023) has introduced an intriguing challenge: under the same Gaussian noise, small batches perform considerably better than large batches, akin to the behavior of regular SGD. It represents a bottleneck because large-batch training is essential for achieving robust privacy guarantees.

DP-SGD distinguishes itself from SGD by two factors: (1) the application of per-sample gradient clipping and (2) the addition of isotropic Gaussian noise to the sum per batch. We first observe a persistent manifestation of the implicit bias associated with DP-SGD observed by Sander et al. (2023) when we remove the gradient clipping component, as depicted in Figure 1. Therefore, the implicit bias of SGD does persist even with Gaussian noise with greater magnitude than the gradients (see Figure 2). However, the implicit bias in SGD is conventionally understood to stem from its inherent noise geometry (HaoChen et al., 2020). In this context, we theoretically investigate the impact of changing the noise structure of SGD on its implicit bias, in the Linear Least Square and Diagonal Linear network (DLN) settings. Our precise contributions are the following:

1. We show that the performance drop observed for large batch training in DP-SGD persists without clipping when training a DNN on ImageNet
2. For Linear Least Squares, we show how Noisy-SGD alters the limiting distributions: the different implicit bias compared to SGD can be controlled by the amount of additional noise
3. For DLNs, we observe that Noisy-SGD can even exhibit an enhanced implicit bias in comparison to SGD. Using continuous modelings, we theoretically demonstrate that a favorable implicit bias indeed exists: It leads to the same solution as SGD, albeit with a distinct effective initialization.

It proves that the gradient's noise geometry is robust to Gaussian perturbations. Our work also suggests that enhancing the performance of large batch training with DP-SGD, which is crucial for achieving robust privacy-utility trade-off, can be accomplished by adopting strategies and techniques developed for managing large batches in non-private settings.

2 Background and Related Work

2.1 Differential Privacy

General Introduction $\mathcal{M}$ is a mechanism that takes as input a dataset D and outputs a machine learning model $\theta \sim \mathcal{M}(D)$.

Definition 1 (Approximate Differential Privacy)**.** *A randomized mechanism $\mathcal{M}$ satisfies (ε, δ)-DP (Dwork et al., 2006) if, for any pair of datasets D and D' that differ by one sample and for all subset $R \subset \mathbf{Im}(\mathcal{M})$,*

$$\mathbb{P}(\mathcal{M}(D) \in R) \leq \mathbb{P}(\mathcal{M}(D') \in R) \exp(\varepsilon) + \delta. \quad (1)$$

Differential Privacy (DP) safeguards against the ability of any potential adversary to infer information about the dataset $\mathcal{D}$ once they've observed the output of the algorithm. In machine learning, this concept implies that if someone acquires the model's parameter $\boldsymbol{\theta}$ trained with a DP algorithm $\mathcal{M}$, then the training data is, by provable guarantees, difficult to reconstruct or infer (Balle et al., 2022; Guo et al., 2022, 2023).

DP-SGD (Chaudhuri et al., 2011; Abadi et al., 2016) is the most popular DP algorithm to train DNNs. It first selects samples uniformly at random with probability $q = B/N$. For $C > 0$, define the clipping function for any $X \in \mathbb{R}^d$ by $\text{clip}_C(X) = C \cdot X / \|X\|$

if $\|X\| \geq C$ and $\text{clip}_C(X) = X$ otherwise. DP-SGD clips the per sample gradients, aggregates them and adds Gaussian noise. Given model parameters $\boldsymbol{\theta}_k$, DP-SGD defines the update $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta_k \widetilde{\mathbf{g}}_k$ where η_k is the step size and $\widetilde{\mathbf{g}}_k$ is given by:

$$\widetilde{\mathbf{g}}_k := \frac{1}{B} \sum_{i \in \mathcal{B}_k} \text{clip}_C \left(\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}_k) \right) + \mathcal{N}\left(0, C^2 \frac{\sigma^2}{B^2} \mathbf{I}\right). \quad (2)$$

$\ell_i(\boldsymbol{\theta})$ represents the individual loss function computed for the sample $\mathbf{x}_i$. The privacy analysis of DP-SGD hinges on the combination of multiple steps. A notably robust analytical framework to account for (ε, δ) is founded on Rényi differential privacy (Mironov, 2017). The theoretical analysis of the convergence properties of DP-SGD has been studied for the convex, strongly convex, and nonconvex settings (Bassily et al., 2014; Wang et al., 2017; Feldman et al., 2018).

Necessity of Large batches for DP-SGD When it comes to training machine learning models with DP-SGD, there is necessarily a privacy-performance trade-off. However, it can be improved by employing very large batch sizes (Anil et al., 2021; Li et al., 2021; De et al., 2022; Yu et al., 2023).

One practical underlying reason is that the magnitude of the effective noise introduced into the average clipped gradient is determined by σ/B (*cf.* Equation (2)). To mitigate this noise, there are two potential approaches: reducing σ or increasing B. While decreasing σ might initially appear appealing, Rényi differential privacy accounting techniques suggest that as σ decreases too much, the privacy parameter ε experiences an exponential increase (Dwork and Rothblum, 2016; Bun and Steinke, 2016; Mironov et al., 2019; Sander et al., 2023). Therefore, increasing the batch size emerges as a critical strategy in DP-SGD training, as it is the most effective means of reducing the effective noise and accelerating the convergence process.

While it is true that employing a larger batch size in DP-SGD results in reduced effective noise, it is important to note that DP training can experience a substantial drop in performance when using larger batch sizes. As demonstrated by Sander et al. (2023), when training an image classifier with DP-SGD on the ImageNet dataset, keeping the number of steps S fixed and maintaining a constant effective noise level of σ/B, the model's performance exhibited a notable decrease when they increased the batch size (refer to Figure 1).

2.2 Implicit Bias of SGD

This section introduces the implicit bias of SGD through the use of Stochastic Differential Equations, and is largely based on Pillaud-Vivien and Pesm (2022). Let $X \in \mathbf{R}^{nxd}$ and $Y \in \mathbf{R}^n$ be the training features and labels, matrices that represent $(x_i, y_i)_{1 \leq i \leq n}$, the set of input-label training pairs. Let $\boldsymbol{\theta} \in \mathbb{R}^d$ the model's parameters and $\bar{X} := X/\sqrt{n}$ the normalized features.

General introduction We consider generic predictors h and the square loss. The empirical risk is:

$$R_n(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} (h(\boldsymbol{\theta}, x_i) - y_i)^2 \quad (3)$$

At step t, the update with learning rate γ is:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \gamma \nabla_\theta R_n(\boldsymbol{\theta_{t-1}}) + \gamma \varepsilon_t(\boldsymbol{\theta}_{t-1}) \quad (4)$$

where $\varepsilon_t(\boldsymbol{\theta})$ is the noise term, that depends on the example(s) used to estimate the gradient. If only example with index i_t is used:

$$\varepsilon_t(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} r_j(\theta) \nabla_\theta h(\theta, x_j) - r_{i_t}(\theta) \nabla_\theta h(\theta, x_{i_t}) \quad (5)$$

With $r_i(\theta) = h(\theta, x_i) - y_i$. We refer to Wojtowytsch (2021) for additional references on the noise. It leads to a first notable characteristic of the gradient noise:

- **The geometry**. SGD noise lies in a linear subspace of dimension at most n spanned by the gradients: $\varepsilon(\boldsymbol{\theta_t}) \in \text{span}\{\nabla h(\boldsymbol{\theta}_t, x_1), ..., \nabla h(\boldsymbol{\theta}_t, x_n)\}$, which is a strict subspace of R^d.

Training overparametrized (*i.e.* $d > n$) models with small batches can lead to better generalization performance compared to large batch training (Keskar et al., 2016; Smith et al., 2020; Masters and Luschi, 2018). In this case, SGD is not only beneficial in terms of computational complexity, but also induces a bias that is beneficial to the performance. *In the overparametrized case*, a second characteristic of the gradient noise is:

- **The scale**. The noise vanishes near optimal solutions: $\varepsilon(\boldsymbol{\theta^*}) = 0$.

The fact that SGD converges towards a particular interpolator is refered to as its implicit bias (Pesme et al., 2021). "Implicit" because no explicit regularization term is added; the regularization comes from the stochastic noise of estimating the gradient at each step using a mini-natch of samples (Zhang et al., 2017).

Stochastic Differential Equations At constant step size, SGD is a homogeneous Markov chain (Meyn and Tweedie, 2012). Studying the continuous time

counterparts of numerical optimization methods is a well-established field in applied mathematics, as it helps to study the limiting distribution of the iterates. Due to its stochastic nature, SGD cannot be modeled as a deterministic flow. A natural approach is to use stochastic differential equations (SDEs) (Øksendal, 2003) to represent its dynamics:

$$d\theta_t = b(t, \theta_t)dt + \sigma(t, \theta_t)dB_t \tag{6}$$

where B is a standard Brownian motion. The drift term b corresponds to the negative gradient of the risk function, and the noise term represents the stochasticity of SGD, which must meet $\sigma_t\sigma_t^T = \gamma\mathbb{E}[\varepsilon_t\varepsilon_t^T|\theta_t]$ and $\sigma_t \in \text{span}\{\nabla h(\boldsymbol{\theta}_t, x_1), ..., \nabla h(\boldsymbol{\theta}_t, x_n)\}$ (Li et al., 2019).

Linear Least Square. We minimize in θ:

$$L(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(\langle\theta, x_i\rangle - y_i)^2 \tag{7}$$

In the overparametrized case, even if there is an infinite number of interpolators, GD and SGD will both converge to θ^{LS}, the solution with the smallest l_2 norm, leading to the same implicit bias (Zhang et al., 2017).

Diagonal Linear Networks (DLN) The implicit bias of SGD appears for more complex architectures. For instance, Chizat and Bach (2020) have studied a two-layer neural networks trained with the Logistic Loss, and Pesme et al. (2021) a sparse regression using a 2-layer DLN. In this work, we focus on the latter. DLN corresponds to a toy neural network that has garnered significant interest in the research community (Vaškevičius et al., 2019; Woodworth et al., 2020; HaoChen et al., 2020). This attention stems from the fact that it is an informative simplification of nonlinear models. The forward pass is $\langle u \odot v, x\rangle$ where $\odot$ is a term by term multiplication, which can equivalently be written $\langle w_+^2 - w_-^2, x_i\rangle$, for $u, v, w_-, w_+ \in \mathbb{R}^d$. The the goal is to minimize in $w = (w_+, w_-)$ the loss:

$$L(w) := \sum_{i=1}^{n}(\langle w_+^2 - w_-^2, x_i\rangle - y_i)^2 =: \sum_{i=1}^{n}(\langle\beta, x_i\rangle - y_i)^2$$

It is similar to the linear least square problem but this time the optimization is (non convex) in w. Considering GD's gradient flow from the initialisation $w_{0,\pm} = \alpha$, Woodworth et al. (2020) have shown that $\beta := w_+^2 - w_-^2$ follows a mirror flow (Beck and Teboulle, 2003) on the loss L and with the hyperbolic entropy:

$$\phi_\alpha(\beta) = \frac{1}{4}\sum_{i=1}^{d}\beta_i \text{ arcsinh}(\frac{\beta_i}{2\alpha_i^2}) - \sqrt{\beta_i^2 + 4\alpha_i^4} \tag{8}$$

as a potential. It means that $\beta = w_+^2 - w_-^2$ follows the dynamic $d\nabla\phi_\alpha(\beta_t) = -\nabla_\beta R(\beta)dt$. It converges to the limit $\beta_\infty^\alpha := \arg\min_{\beta s.t. X\beta = Y}\phi_\alpha(\beta)$.

For large initialisations α, ϕ_α is close to the l_2-norm, which means that the recovered solution has a small l_2-norm. On the other hand, for smaller values of α, the potential aligns with the l_1-norm. As a result, the retrieved solution will inherently possess some sparsity. In the case of a sparse regression, using small values for α thus results in an improved implicit bias.

Expanding upon the GD analysis, Pesme et al. (2021) introduced a continuous modeling approach for SGD. Their research demonstrated that, starting from the same initialization α, the stochastic process converges with high probability towards $\beta_\infty^{\alpha_\infty}$:

$$\beta_\infty^{\alpha_\infty} = \underset{\beta s.t. X\beta = Y}{\arg\min}\ \phi_{\alpha_\infty}(\beta), \tag{9}$$

which is similar to GD, but with a smaller effective α:

$$\alpha_t = \alpha \odot \exp\left(-2\gamma\text{diag}(\bar{X}^T\bar{X})\int_0^t L(\beta_s)ds\right) \tag{10}$$

Thus, SGD leads to sparser solutions than GD, which is a proof of the implicit bias of SGD for DLNs.

3 Implicit bias of noisy-SGD

Notations. $(B_{t,\pm})_t$ and $(\tilde{B}_{t,\pm})_t$ are standardized Brownian motions on $\mathbb{R}^n$ and $\mathbb{R}^d$ respectively.

3.1 Noisy SGD

DP-SGD (*cf.* Equation 2) differentiates itself from SGD through (1) the utilization of per-sample gradient clipping and (2) the incorporation of isotropic Gaussian noise into the batch-wise gradient sum. To study the reason behind the implicit bias of DP-SGD observed in Sander et al. (2023) (*i.e.*, that small batches perform better than large batches at fixed effective noise σ/B), we examine if a similar phenomenon exists without clipping (Noisy-SGD), thus using the following noisy gradient update:

$$\widetilde{\mathbf{g}}_k := \frac{1}{B}\sum_{i\in\mathcal{B}_k}\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_k) + \mathcal{N}\left(0, \frac{\sigma^2}{B^2}\mathbf{I}\right) \tag{11}$$

We demonstrate in Figure 1 the persistence of this small-batch training superiority even in the absence of clipping, and even when confronted with Gaussian noise that surpasses the magnitude of the gradients (as depicted in Figure 2; see Section 4.1 for experimental

details). This piques our interest in delving into the theoretical aspects of Noisy-SGD for simple architectures in order to explore the extent to which the natural gradient geometry remains robust to Gaussian perturbations. The implications of studying Noisy-SGD as a proxy for DP-SGD are discussed in Section 6.

3.2 Linear Least Square

The simplest model to study is Linear Least Square. We verify that noisy versions of SGD leads to similar solutions than SGD and GD.

Warm-up: Underparametrized setting ($n > d$) The stochastic noise $\varepsilon(\theta)$ (*cf.* Equation 4) does not vanish near optimal solutions. For fixed $\epsilon > 0$, we consider the following SDE as the continuous approximation of SGD (Ali et al., 2020):

$$d\theta_t = -\bar{X}^T(\bar{X}\theta_t - Y)dt + \sqrt{\gamma}\epsilon\bar{X}^T dW_t \tag{12}$$

We notice that it respects the characteristics highlighted in Section 2.2. This Orsnten-Uhlenbeck process has a limit, which can be solved through its characteristic Lyapunov equation (see Appendix A). With $X^\dagger$ the pseudo-inverse of X and $\theta^{LS} := X^\dagger Y$:

$$\theta_\infty \sim \mathcal{N}\left(\theta^{LS}, \frac{\gamma\epsilon^2}{2} I_d\right) \tag{13}$$

We now consider adding isotropic Gaussian noise on top of the gradient's structured noise of SGD:

$$d\theta_t = -\bar{X}^T(\bar{X}\theta_t - Y)dt + \sqrt{\gamma}\epsilon\bar{X}^T dW_t + \sigma d\tilde{W}_t \tag{14}$$

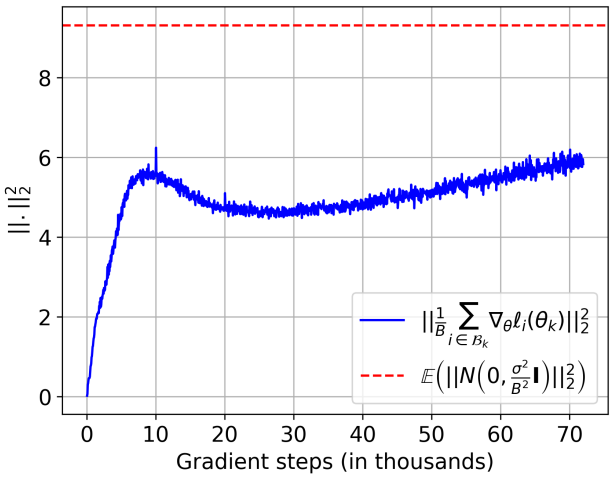

Figure 2: Noisy-SGD on ImageNet. We compare the norm of the mini-batch gradient to the one of the Gaussian noise when training with Noisy-SGD on ImageNet, for $B = 128$ and the same set-up as in Figure 1. The noise magnitude is greater than the gradients.

Solving the characteristic Lyapunov equation (see Appendix A), the limiting distribution becomes:

$$\boxed{\theta_\infty \sim \mathcal{N}\left(\theta^{LS}, \frac{\gamma\epsilon^2}{2} I_d + \frac{\sigma^2}{4}(\bar{X}^T\bar{X})^{-1}\right)} \tag{15}$$

We observe that adding isotropic noise changes the limiting distribution: its shape depends on the training data. This simple example shows that geometry variation of the noise in the linear least square setting implies a controlled variation of the limiting process.

Over parametrized case ($d > n$): As described in Section 2.2, GD and SGD will in this case both converge to θ^{LS} and therefore have the same implicit bias. SGD's dynamic can be studied through the following SDE (Ali et al., 2020), where this time the gradient noise vanishes near the interpolators:

$$d\theta_t = -\bar{X}^T(\bar{X}\theta_t - Y)dt + \sqrt{\gamma}\|X\theta_t - Y\|_2\bar{X}^T dB_t \tag{16}$$

which also respects SGD's specificities. θ_t converges to θ^{LS} for $\gamma \leq 1/Tr(\bar{X}^T\bar{X})$. If we add a constant spherical Gaussian noise at each step, the iterates will not converge. We thus look at the impact of adding a noise with a similar scale as the gradient noise:

$$\begin{aligned} d\beta_t = &-\bar{X}^T(\bar{X}\beta_t - Y)dt + \sqrt{\gamma}\|X\beta_t - Y\|_2\bar{X}^T dB_t \\ &+ \sqrt{\gamma}\|X\beta_t - Y\|_2\sigma d\tilde{B}_t \end{aligned}$$

We show that we can control the difference between the SDE and the noisy SDE through $\eta_t = \|\theta_t - \beta_t\|_2^2$:

Theorem 1. *If $\gamma \leq 1/Tr(\bar{X}^T\bar{X})$ then*

$$\boxed{\mathbf{E}(\eta_t) \leq \gamma d\sigma^2 \int_0^t L(\beta_t)ds} \tag{17}$$

Proof. Apply Itô's formula to η_t, see Appendix A □

Here, we have highlighted that the variation from the original SDE will be contingent on factors such as dimension, step size, noise magnitude, and convergence rate. As a result, the solution obtained from a noisy variant of SGD may closely resemble that of traditional SGD, thus leading to a similar implicit bias. The extent of this similarity depends on the parameters used.

3.3 Noisy-SGD training of DLNs

For a 2-layer DLN sparse regression, SGD-induced noise steers the optimization dynamics towards advantageous regions that exhibit better generalization than GD (*i.e.*, more sparse) Pesme et al. (2021). We investigate the impact of the addition of Gaussian noise.

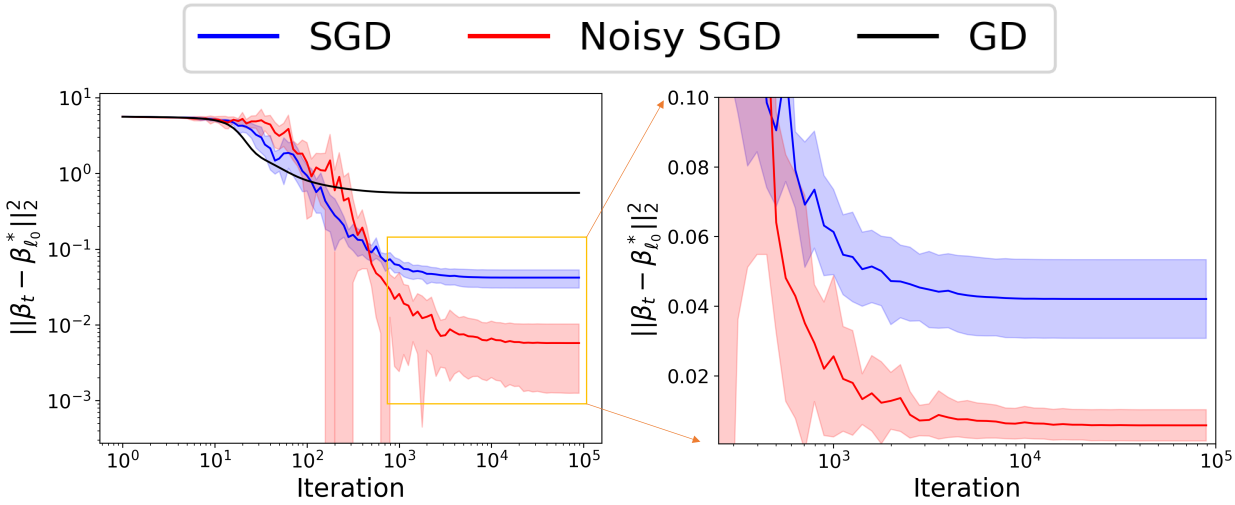

Figure 3: Diagonal Linear Network: Implicit Bias of GD, SGD and Noisy SGD ($\sigma = 0.5$ in Equation 18). Shaded areas represent one standard deviation over 5 runs. (Left) Compared to GD with the same initialisation $\alpha = 0.1$, SGD attain solutions closer to the sparse $\beta^*_{l_0}$, as expected from Pesme et al. (2021). Moreover, we observe that Noisy-SGD has a better implicit bias than SGD: the gradient noise structure is enhanced by perturbations. (Right) In absolute terms, Noisy-SGD does not even showcase more variance than SGD (near convergence).

Continuous model for Noisy-SGD Adding constant Gaussian noise at each gradient step as in Equation 11 inevitably makes the process diverge for overparametrized DLNs. However, adding Gaussian noise with comparable magnitude to that of the gradients still allows to study the impact of geometry deformation, while maintaining a vanishing noise, enabling convergence. Pesme et al. (2021) have shown that the SGD update can be written $w_{t+1,\pm} = -\gamma \nabla_{w_{t,\pm}} L(w_t) \pm \gamma[X^T \zeta_{i_t}(\beta_t)] \odot w_{t,\pm}$, where ζ_{i_t} is detailed in Appendix B. We thus consider the following noisy update:

$$\begin{aligned} w_{t+1,\pm} = &- \gamma \nabla_{w_{t,\pm}} L(w_t) \pm \gamma[X^T \zeta_{i_t}(\beta_t)] \odot w_{t,\pm} \\ &\pm \gamma \sigma_t Z_{t,\pm} \odot w_{t,\pm}, \end{aligned} \tag{18}$$

with $Z_{t,\pm} \sim \mathcal{N}_d(0,1)$ and $\sigma_t \in \mathbb{R}^{\mathbb{R}}$. We consider here $\sigma_t = 2\sigma\sqrt{L(w_t)}$ for $\sigma \geq 0$, and show more general results in Appendix B. The corresponding SDE is:

$$\begin{aligned} dw_{t,\pm} = &\mp [\bar{X}^T r(w_t)] \odot w_{t,\pm} dt \\ &+ 2\sqrt{\gamma L(w_t)} w_{t,\pm} \odot \bar{X}^T dB_{t,\pm} \\ &+ 2\sigma\sqrt{\gamma L(w_t)} w_{t,\pm} \odot d\tilde{B}_{t,\pm} \end{aligned} \tag{19}$$

where $r(w) = \bar{X}(w_+^2 - w_-^2 - \beta^*)$ and β^* is an interpolator. We notice that compared to the SDE of Pesme et al. (2021), it only adds the term $2\sigma\sqrt{\gamma L(w_t)} w_{t,\pm} \odot d\tilde{B}_{t,\pm}$. See section 6 for a discussion on the impact of modeling with a decreasing noise, and appendix B for detailed computation of the link between equations (18) and (19).

We now demonstrate that despite the Gaussian perturbation, β follows a stochastic mirror flow similar to the one in Pesme et al. (2021), detailed in Section 2.2:

Proposition 1. *Let $(w_{t,\pm})_{t\geq 0}$ be defined as in equation (19) from initialisation α. Then $(\beta_t = w_{t,+}^2 - w_{t,-}^2)_{t\geq 0}$ follows a stochastic mirror flow defined by:*

$$\begin{aligned} d\nabla\phi_{\alpha_t^\beta}(\beta_t) = &- \nabla L(\beta_t) dt + \sqrt{\gamma L(\beta_t)} \bar{X}^T dB_t \\ &+ \sqrt{\gamma L(\beta_t)} \sigma d\tilde{B}_t \end{aligned} \tag{20}$$

where $\tilde{B}_s = (\tilde{B}_{s,+} + \tilde{B}_{s,+})/2$, $B_s = (B_{s,+} + B_{s,+})/2$,

$$\begin{aligned} \alpha_t^\beta = &\alpha \exp\left(-2\gamma\sigma^2 \int_0^t L(\beta_s) ds\right) \\ &\odot \exp\left(-2\gamma \mathrm{diag}(\bar{X}^T \bar{X}) \int_0^t L(\beta_s) ds\right) \end{aligned} \tag{21}$$

Proof. See appendix B. □

The continuous version of SGD analysed by Pesme et al. (2021) had led to the following mirror flow:

$$d\nabla\phi_{\alpha_t}(\beta_t) = -\nabla L(\beta_t) dt + \sqrt{\gamma L(\beta_t)} \bar{X}^T dB_t \tag{22}$$

With $\alpha_t = \alpha \exp\left(-2\gamma \mathrm{diag}(\bar{X}^T \bar{X}) \int_0^t L(\beta_s) ds\right)$. Our resulting SDE for Noisy-SGD (*cf.* Equation 20) differs by two aspects: the additional spherical noise $\sqrt{\gamma L(\beta_t)} \sigma d\tilde{B}_t$ and a smaller effective value for α.

Solution. If β follows the original mirror flow, than the process converges with high probability towards $\arg\min_{\theta \text{ s.t. } X\theta = Y} \phi_{\alpha_\infty}$. This formulation is possible only because the KKT conditions are respected by the limit vector, as the updates remain confined to $\mathrm{span}(x_1, ..., x_n)$. However, this is no longer the case for Noisy-SGD. We still show a comparable outcome.

Theorem 2. *Let $(w_t)_{t\geq 0}$ follow the SDE (19) from initialisation α. Then for any p close to 0 there is a constant C s.t. for any step size $\gamma \leq C$, with probability at least $1-p$, $\int_0^\infty \beta_s ds$ converges and β_t converges with high probability to an interpolator β_∞.*

Proof. We give here an idea of the proof which is along the lines of the proof found in Pesme et al. (2021). We first note that by arbitrarily changing $\bar{X}^T$ in SDE (22), the process still converges. In our case:

$$dw_{t,\pm} = -\nabla_{w_\pm} L(w_t)dt + 2\sqrt{\gamma L(w_t)} w_{t,\pm} \circledcirc [A dB_{t,\pm}]$$

where this time B_t is valued in $\mathbb{R}^{d+n}$ and $A := \left(\bar{X}|\sigma I_d\right)$. The whole proof is given in Appendix B □

However, because the corresponding mirror flow updates no longer lie in $\text{span}(x_1, ..., x_n)$ (*cf.* Equation 20), the KKT conditions of the minimization problem are no longer verified by β_∞, and

$$\beta_\infty \neq \underset{\beta \text{ s.t. } X\beta = Y}{\arg\min} \phi_{\alpha_\infty^\beta}(\beta) \tag{23}$$

Nevertheless, we show that β_∞ satisfy the KKT conditions for a small perturbation of that problem:

Proposition 2. *Under the assumptions of Theorem 2, on the set of convergence of β_t, there exists r_∞ s.t.:*

$$\boxed{\beta_\infty = \underset{X\beta = Y}{\arg\min} \phi_{\alpha_\infty}(\beta) - \langle r_\infty, \beta \rangle} \tag{24}$$

Proof. Let us consider P the orthogonal projection over $\text{span}(x_1, ..., x_n)$. Integrating (20) and decomposing the right hand side as a term $v \in \text{span}(x_1, ..., x_n)$ and a term in the orthogonal, we get:

$$\nabla \phi_{\alpha_\infty}(\beta_\infty) = v + \int_0^\infty \sigma \sqrt{\gamma L(\beta_s)}(1-P) dB s \tag{25}$$

Let us define $r_\infty = \int_0^\infty \sigma \sqrt{\gamma L(\beta_s)}(1-P)dBs$, and $\tilde{\beta}_0 = \nabla \phi_{\alpha_\infty}^{-1}(r_\infty)$, which is well defined by strong convexity of ϕ_{α_∞} (Refer to Appendix B). We have:

$$\nabla \phi_{\alpha_\infty}(\beta_\infty) - \nabla \phi_{\alpha_\infty}(\tilde{\beta}_0) \in \text{span}(x_1, ..., x_n) \tag{26}$$

Thus, by defining D the following Bregman divergence:

$$D_{\phi_{\alpha_\infty}}(\beta, \tilde{\beta}_0) = \phi_{\alpha_\infty}(\beta) - \phi_{\alpha_\infty}(\tilde{\beta}_0) - \langle \nabla \phi_{\alpha_\infty}(\tilde{\beta}_0), \beta - \tilde{\beta}_0 \rangle$$

The gradient taken at β_∞ lies in $\text{span}(x_1, ..., x_n)$, insuring that the KKT conditions of the following problem are verified and thus:

$$\boxed{\beta_\infty = \underset{X\beta = Y}{\arg\min} D_{\phi_{\alpha_\infty}}(\beta, \tilde{\beta}_0)} \tag{27}$$

□

Therefore, Noisy-SGD gives the same solution as GD but on a different potential ϕ_{α_∞} and from an effective initialization $\tilde{\beta}_0$, while SGD only changes α. For Noisy-SGD, we note that α_∞ decreases with σ (eq. 21): the more noise is added, the smaller is the effective α. If it did not change the effective initialization too, it would directly imply "better" implicit bias.

Implicit Bias We have observed that β_∞ is the solution to a perturbed minimization problem. We show now that the distance between β_∞ and the optimizer of ϕ_{α_∞} can be controlled by the perturbation σ:

Proposition 3. *Under the assumptions of theorem 2 and on the high probability set of convergence of (β_t) let $\beta^*_{\alpha_\infty} := \arg\min_{X\beta = Y} \phi_{\alpha_\infty}(\beta)$ and $r_\infty := \sigma \int_0^\infty \sqrt{\gamma L(\beta_s)}(1-P)dBs$. Since $\alpha_\infty > 0$, ϕ_{α_∞} is μ strongly convex for $\mu > 0$. Then we have:*

$$\frac{1}{\mu}\|r_\infty\|_2 \geq \|\beta^*_{\alpha_\infty} - \beta_\infty\|_2 \tag{28}$$

Proof. See Appendix B. □

We observe that although increasing σ implies a lower α_∞ than SGD, which should yield to sparser solutions, it also implies a new effective initialization $\tilde{\beta}_0$ that deviates the solution from the sparse $\beta^*_{\alpha_\infty}$. However, in Figure 3, we show that Noisy-SGD still leads to sparser solutions than SGD. This might be attributed to the fact that the gain from a smaller α is in this case more important than the perturbation controlled by $\frac{1}{\mu}\|r_\infty\| \propto \sigma$, as shown in Figure 4. We give more details in the experiment section 4.2.

Proofs are derived in Appendix B.1 with more general noise forms, with stability results for the implicit bias.

4 Experiments

4.1 Noisy SGD on ImageNet

Dataset and architecture To compare Noisy-SGD to DP-SGD, we adopt the exact same set-up as Sander et al. (2023). We train a Normalizer-Free ResNets (NF-ResNets) (Brock et al., 2021) with $d = 25$M parameters on the ImageNet-1K dataset (Deng et al., 2009; Russakovsky et al., 2014) with blurred faces, which contains 1.2 million images partitioned into 1000 categories. We use the timm (Wightman, 2019) library based on Pytorch (Paszke et al., 2019).

Optimization For SGD and Noisy-SGD, we use a constant learning rate of 0.5 and no momentum. We perform an exponential moving average of the weights (Tan and Le, 2019) with a decay parameter of 0.999 (similar to Sander et al. (2023)). We set the effective

noise σ/B constant to 6×10^{-4} for all Noisy-SGD experiments of Figure 1, which is 4 times greater than the noise used by Sander et al. (2023) for ImageNet.

Implicit bias and Noise level In Figure 2, we show that the additional Gaussian noise has a bigger l_2 norm than the gradients throughout training, for the same set-up as in Figure 1. It is an important observation because it shows that the additional Gaussian noise is not negligible, and thus that a similar implicit bias to the one of SGD persists even under strong Gaussian perturbation. Moreover, we notice that the training trajectories of SGD and Noisy-SGD do differ in Figure 1, especially at the beginning. However, the gradient norm $||\frac{1}{B}\sum_{i\in\mathcal{B}_k}\nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}_k)||$ increases during training, while the corresponding quantity in DP-SGD is bounded by 1 (*cf.* Equation 2 when $C = 1$), which makes the actual signal-to-noise ratio greater in Noisy-SGD compared to DP-SGD.

4.2 Diagonal Linear Network

We adopt the same set-up as Pesme et al. (2021) for sparse regression. We select parameters n = 40 and d = 100, and then create a sparse model $\beta^*_{l_0}$ with the constraint that its l_0 norm is equal to 5. We generate the features x_i from a normal distribution with mean 0 and identity covariance matrix $N(0, I)$, and compute the labels as $y_i = x_i^T\beta^*_{l_0}$. We always use the same step size of $\gamma = 1/(1.3||\bar{X}\bar{X}^T||_2)$. Notice that $||\beta_t - \beta^*_{l_0}||_2^2$ is the validation loss in the experiments.

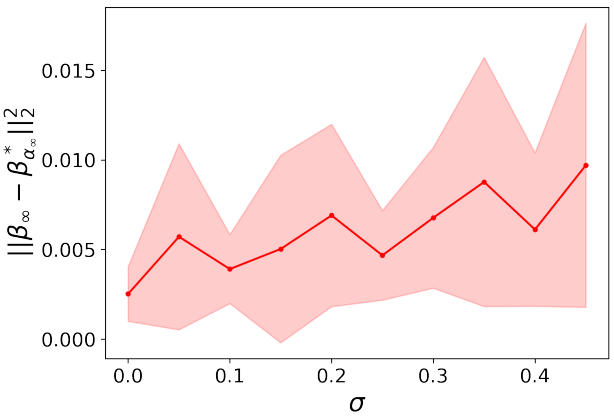

Figure 4: DLN: Distance between $\beta^*_{\alpha_\infty}$, the solution that minimizes ϕ_{α_∞} —obtained by GD from α_∞— and the one obtained by Noisy-SGD (see Proposition 3). Shaded areas represent one standard deviation over 10 runs. For small σ, the distance is smaller than the distance between the solutions of SGD and the sparse solution β_{l_0} (see Figure 3), explaining why the implicit bias persists and can be enhanced by Gaussian noise.

Noisy-SGD's improved implicit Bias In Figure 3, we show that noisy-SGD produces sparser solutions than the one obtaimed with SGD, as it is closer to the sparse interpolator β_{l_0}. This effect is primarily due to the impact of having a smaller effective value α which outweighs the impact of the perturbation governed by the new effective initialization $\tilde{\beta}_0$: $\frac{1}{\mu}\|r_\infty\| \propto \sigma$ (see Equation 3). In Appendix B.2, we also illustrate scenarios where the parameter σ becomes more prominent in comparison to α, and for which Noisy-SGD still induce a stronger implicit bias.

Impact of σ In Figure 4, we run Noisy-SGD for different values of σ, 10 times for each value, and show the averages and standard deviations. As expected from Proposition 3, the distance between $\beta^*_{\alpha_\infty}$, the actual minimizer under constraints of ϕ_{α_∞} and β_∞, the solution obtained from Noisy-SGD, increases with σ. We observe that the order of magnitude of this increase, which was hidden inside the constants of the Proposition, is reasonable when compared to the distance between the solution of SGD and the sparse interpolator, as depicted in Figure 3. It explains why using $\sigma = 0.5$ still helps when starting from $\alpha = 0.1$.

5 Conclusion

The challenge of achieving strong performance in large-batch DP-SGD training remains a critical obstacle in balancing privacy and utility. This performance bottleneck in DP-SGD is poised to become more pronounced over time as the theoretical solution to further bridging the performance gap between private and non-private training lies in the expansion of both training set size and batch size (Sander et al., 2023).

We have observed that Noisy-SGD exhibits a performance decline with the batch size, indicating that the gradient noise in SGD still plays a pivotal role, even with Gaussian perturbation. It puts in perspective a key argument supporting that the inherent geometry of SGD's noise is responsible for its bias. For DLNs, a seemingly minor addition of Gaussian noise disrupts a crucial KKT condition, necessitating a complex mathematical alternative. We've derived proofs for broader noise in Appendix B.1 and offer stability results that extend the amplification of the implicit bias. Overall, our study underscores two critical points:

- The specific gradient noise geometry in SGD is robust to Gaussian perturbations, and different geometries can lead to different implicit biases
- Further work could try to leverage methods developed to improve large-batch training in non-private settings (*e.g.*, employing the LARS opti-

mizer (You et al., 2017) for convolutional DNNs) to enhance the performance of DP-SGD, thus advancing the trade-off between privacy and utility.

6 Discussion

Our decision to study Noisy-SGD as a substitute for DP-SGD is primarily based on the observations depicted in figure 1. Notably, a similar implicit bias of DP-SGD is discernible even in the absence of clipping. We delve here into the implications of using Noisy-SGD as a stand-in DP-SGD. This includes considerations from both optimization and privacy perspectives.

An optimization perspective. Noisy-SGD only aligns with DP-SGD if the gradients are bounded, an assumption that is commonly done in convergence studies (Bassily et al., 2014). This is not the case for classical neural networks, nor for the DLN case that we studied. Without any assumption, the clipping operation can bias the (expected) direction of the update. However, assuming symmetricity of the gradients, which is a reasonable assumption (Chen et al., 2020), the drift maintains its direction after clipping.

If we take clipping into account for DLNs while assuming symetricity, the noise term can still be approximated by $\sigma_t \circ w_t dB_t$, where σ_t now depends on w_t. Indeed, if we denote $y_t := \langle \beta - \beta^*, x_{i_t} \rangle x_{i_t}$, we get $w_{t+1} = w_t - \gamma \mathbb{E}[\text{clip}(y_t \circ w_t)] - \gamma(\text{clip}(y_t \circ w_t) - \mathbb{E}[\text{clip}(y_t \circ w_t)])$. The mean zero term rewrites as $\alpha(y_t, w_t) y_t \circ w_t - \mathbb{E}[\alpha(y_t, w_t) y_t \circ w_t]$, with first order of the covariance$(1/n)\text{diag}(w_t)\alpha^2 \mathbb{E}[y_t y_t^T]\text{diag}(w_t)$. Our extension in appendix B.1 is a first approximation.

A privacy perspective. We deliberately focus on the sole effect of noise addition and its impact on the optimization procedure. Nevertheless, we do theoretically show how the noise level impacts our results, with additional experiments presented in Appendix B.2, and general forms in appendix B.1. However, vanishing noise would always lead to exploding privacy guarantees. We use this modelisation as a first approximation, as it changes the structure of the noise similarly to DP. Moreover, it is necessary for convergence as the learning is constant; studying a decreasing learning rate with fixed noise could have been an alternative, but we should have taken a different angle (unknown) to study the bias. In a similar vein, one could consider the clipping component more comprehensively. One potential approach could be to proportionally decrease the clipping value in tandem with the noise magnitude. This method would not only uphold the privacy guarantees but also ensure convergence

Acknowledgements

Our sincere thanks to Pierre Stock and Alexandre Sablayrolles for their initial guidance, Scott Pesme for his critical insights on the mathematical derivations of the DLN part, and Ilya Mironov for his constructive feedback on our draft.

References

Tom Sander, Pierre Stock, and Alexandre Sablayrolles. Tan without a burn: Scaling laws of dp-sgd. In *International Conference on Machine Learning*, pages 29937–29949. PMLR, 2023.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL `https://doi.org/10.1214/aoms/1177729586`.

M. Duflo. *Algorithmes stochastiques*. Mathématiques et Applications. Springer Berlin Heidelberg, 1996. ISBN 9783540606994. URL `https://books.google.co.uk/books?id=ffkzAAAACAAJ`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Samuel L. Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. *CoRR*, abs/2006.15081, 2020. URL `https://arxiv.org/abs/2006.15081`.

Dominic Masters and Carlo Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018.

Jeff Z. HaoChen, Colin Wei, Jason D. Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance, 2020.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium*, pages 267–284, 2019.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, page 265–284, 2006.

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning*, pages 12208–12218. PMLR, 2021.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L. Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale, 2022. URL `https://arxiv.org/abs/2204.13650`.

Yaodong Yu, Maziar Sanjabi, Yi Ma, Kamalika Chaudhuri, and Chuan Guo. Vip: A differentially private foundation model for computer vision. *arXiv preprint arXiv:2306.08842*, 2023.

Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1138–1156. IEEE, 2022.

Chuan Guo, Brian Karrer, Kamalika Chaudhuri, and Laurens van der Maaten. Bounding training data reconstruction in private (deep) learning. In *International Conference on Machine Learning*, pages 8056–8071. PMLR, 2022.

Chuan Guo, Alexandre Sablayrolles, and Maziar Sanjabi. Analyzing privacy leakage in machine learning via multiple hypothesis testing: A lesson from fano. In *International Conference on Machine Learning*, pages 11998–12011. PMLR, 2023.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization, 2011.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 464–473. IEEE, 2014.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017.

Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 521–532. IEEE, 2018.

Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. Large-scale differentially private bert, 2021. URL `https://arxiv.org/abs/2108.01624`.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners, 2021.

Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the Sampled Gaussian Mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

Pillaud-Vivien and Scott Pesm. Implicit bias of sgd, 2022. URL `https://francisbach.com/implicit-bias-sgd/`.

Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type. part i: Discrete time analysis, 2021.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear

networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2017.

S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Communications and Control Engineering. Springer London, 2012. ISBN 9781447132691. URL https://books.google.co.uk/books?id=kTX6oAEACAAJ.

Bernt Øksendal. *Stochastic Differential Equations*, pages 65–84. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003. ISBN 978-3-642-14394-6. doi: 10.1007/978-3-642-14394-6_5. URL https://doi.org/10.1007/978-3-642-14394-6_5.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20 (40):1–47, 2019. URL http://jmlr.org/papers/v20/17-526.html.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss, 2020.

Tomas Vaškevičius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery, 2019.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3635–3673. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/woodworth20a.html.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003. ISSN 0167-6377. doi: https://doi.org/10.1016/S0167-6377(02)00231-6. URL https://www.sciencedirect.com/science/article/pii/S0167637702002316.

Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 233–244. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/ali20a.html.

Andrew Brock, Soham De, Samuel L. Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization, 2021. URL https://arxiv.org/abs/2102.06171.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2014. URL https://arxiv.org/abs/1409.0575.

Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks, 2017.

Xiangyi Chen, Steven Z Wu, and Mingyi Hong. Understanding gradient clipping in private sgd: A geometric perspective. *Advances in Neural Information Processing Systems*, 33:13773–13782, 2020.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales, 2020.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [No]
   (c) New assets either in the supplemental material or as a URL, if applicable. [No]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Linear Least Square

We detail the proofs of the Propositions presented in Section 3.2 focusing on the continuous analysis of the implicit bias of SGD and Noisy-SGD in the Linear Least Square Setting. We start with the under parametrized case in Section A.1 where the number of parameters d is smaller than the number of training data points n. Subsequently, we delve into the over parameterized case in Section A.2.

A.1 Under Parameterized Case

SDE for SGD In situations where the model is underparameterized, the gradient noise encompasses the entire space of $\mathbb{R}^d$ and there is a specific value of $\epsilon > 0$ such that $||X\theta_t - Y||$ remains greater than ϵ for all time steps t, as documented in Ali et al. (2020). We thus consider the following SDE as the continuous approximation of SGD in the linear least square under parametrized setting. In this framework, the Euler discretization with a step size of γ corresponds to the implementation of SGD, incorporating the noise modelization previously described:

$$d\theta_t = -\bar{X}^T(\bar{X}\theta_t - Y)dt + \sqrt{\gamma}\epsilon\bar{X}^T dW_t \tag{29}$$

This process is a Orsnten-Uhlenbeck process that can be written as:

$$d\theta_t = \beta(\mu - \theta_t)dt + \Sigma dW_t \tag{30}$$

with $\mu = (X^TX)^{-1}X^TY = X^\dagger Y := \theta^{LS}$ for $X^\dagger$ the pseudo-inverse of X, $\beta = \bar{X}^T\bar{X}$ and $\Sigma = \sqrt{\gamma}\epsilon\bar{X}^T$.

The stationary distribution of this temporally homogeneous Markov Chain is:

$$\theta_\infty = \mathcal{N}(\theta^{LS}, w) \tag{31}$$

for w that verifies the following Lyapunov equation:

$$\beta w + w\beta^T = 2D = \gamma\epsilon^2\beta \tag{32}$$

For $D = \Sigma^T\Sigma/2 = \frac{\gamma\epsilon^2}{2}\beta$. The solution of this Lyapunov equation can be directly solved as:

$$w = \int_0^\infty e^{-t\beta}(-2D)e^{-t\beta}\,dt = \int_0^\infty e^{-t\beta}(-\gamma\epsilon^2\beta)e^{-t\beta}\,dt = -\gamma\epsilon^2\beta\int_0^\infty e^{-2t\beta}\,dt = \frac{\gamma\epsilon^2}{2}I_d \tag{33}$$

As $\int_0^\infty e^{-2t\beta}\,dt = \beta^{-1}/2$. The commutativity results from the fact that all matrices are sums of powers of β within the integrals.

Noisy-SDE Let us now consider adding Gaussian noise to the modelisation of the natural SGD noise:

$$d\theta_t = -\bar{X}^T(\bar{X}\theta_t - Y)dt + \sqrt{\gamma}\epsilon\bar{X}^T dW_t + \sigma d\tilde{W}_t \tag{34}$$

we get the same equations for $\Sigma = (\sqrt{\gamma}\epsilon\bar{X}^T|\sigma I_d) \in \mathbf{R}^{d\times(n+d)}$ and $W = (W^T|\tilde{W}^T)^T$.

So $D = \frac{\Sigma\Sigma^T}{2} = \frac{\frac{\gamma}{n}\epsilon^2 X^T X + \sigma^2 I_d}{2} = \frac{\gamma\epsilon^2}{2}\beta + \frac{\sigma^2}{2}I_d$, and in this case:

$$w = \int_0^\infty e^{-t\beta}(-\gamma\epsilon^2\beta)e^{-t\beta}\,dt - \frac{\sigma^2}{2}(-\frac{1}{2}\beta^{-1}) = \frac{\gamma\epsilon^2}{2}I_d + \frac{\sigma^2}{4}\beta^{-1} \tag{35}$$

With $\beta^{-1} = (\bar{X}^T\bar{X})^{-1}$. Finally:

$$\boxed{\theta_\infty \sim \mathcal{N}\left(\theta^{LS}, \frac{\gamma\epsilon^2}{2}I_d + \frac{\sigma^2}{4}(\bar{X}^T\bar{X})^{-1}\right)} \tag{36}$$

In this case, we observe that adding spherical noise adds a dependence on the data distribution to the variance term.

A.2 Over Parameterized Case

In the overparametrised regime, the noise vanishes at every global optimum $\theta^\star$, and is degenerate in the directions of $\mathrm{Ker}(X)$, giving the following continuous approximation (Pillaud-Vivien and Pesm, 2022; Ali et al., 2020):

$$d\theta_t = -\bar{X}^T(\bar{X}\theta_t - Y)dt + \sqrt{\gamma}\|X\theta_t - Y\|_2\bar{X}^T dB_t \tag{37}$$

In this case, θ_t converges almost surely to θ^{LS} for $\gamma > 0$ and $\theta_0 = 0$. Let us now consider the following noisy version of SGD, with an additional noise that has the same magnitude as the gradient noise:

$$\begin{aligned} d\beta_t = &-\bar{X}^T(\bar{X}\beta_t - Y)dt + \sqrt{\gamma}\|X\beta_t - Y\|_2\bar{X}^T dB_t \\ &+ \sqrt{\gamma}\|X\beta_t - Y\|_2\sigma d\tilde{B}_t \end{aligned}$$

We show that we can control the difference between the iterates of the first SDE and its noisy version: $\eta_t = \|\theta_t - \beta_t\|_2^2$. More precisely,

If $\gamma \leq 2/Tr(\bar{X}^T\bar{X})$ then

$$\boxed{\mathbf{E}(\eta_t) \leq \gamma d\sigma^2 \int_0^t L(\beta_t)ds}$$

Proof. We can write $d\beta_t = -\bar{X}^T(\bar{X}\beta_t - Y)dt + \sqrt{\gamma}\|X\beta_t - Y\|_2(\bar{X}^T|\sigma I_d)dB'_t$ for $B' = (\frac{B}{\tilde{B}}) \in \mathbb{R}^{n+d}$. Similarly to θ, for $\beta_0 = 0$ and $\gamma > 0$, β converges almost surely. So $\theta - \beta$ converges and finally η converges. Applying Itô's formula, we get:

$$\begin{aligned} d\eta_t = &[-2\|\bar{X}(\theta_t - \beta_t)\|^2 \\ &+ \gamma Tr(\bar{X}^T\bar{X})(\sqrt{R_n(\theta_t)} - \sqrt{R_n(\beta_t)})^2 \\ &+ d\sigma^2\gamma R_n(\beta_t)]dt + ...dW_t + ...d\tilde{W}_t \end{aligned}$$

Straightforwardly using the triangular inequality, we have:

$$(\sqrt{R_n(\theta_t)} - \sqrt{R_n(\beta_t)})^2 \leq ||\bar{X}(\theta_t - \beta_t)||_2^2 = ||\bar{X}\theta_t - Y - (\bar{X}\beta_t - Y)||_2^2 \leq (\sqrt{R_n(\theta_t)} + \sqrt{R_n(\beta_t)})^2 \tag{38}$$

Looking now at the expected value, the Brownian motions disappear and injecting Inequality 38:

$$\mathbb{E}(\eta_t) = \gamma d\sigma^2 \int_0^t R_n(\beta_s)ds - 2\int_0^t \|\bar{X}(\theta_s - \beta_s)\|_2^2 ds + \gamma Tr(\bar{X}^T\bar{X}) \int_0^t (\sqrt{R_n(\theta_s)} - \sqrt{R_n(\beta_s)})^2$$
$$\leq \gamma d\sigma^2 \int_0^t R_n(\beta_t)ds + (\gamma Tr(\bar{X}^T\bar{X}) - 2)\int_0^t \|\bar{X}(\theta_s - \beta_s)\|^2 ds$$

Finally, for $\gamma < 2/Tr(\bar{X}^T\bar{X})$,

$$\boxed{\mathbf{E}(\eta_t) \leq \gamma d\sigma^2 \int_0^t R_n(\beta_s)ds = \gamma d\sigma^2 \int_0^t L(\beta_s)ds}$$

and in particular,

$$\boxed{\mathbf{E}(\eta_\infty) \leq \gamma d\sigma^2 \int_0^\infty L(\beta_s)ds}$$

□

B Diagonal Linear Network

In this Section, we first derive in Section B.1 the mathematical proofs of the Propositions and Theorems presented in Section 3.3. We then show more empirical results on the sparsity of the solutions obtained with Noisy-SGD, when training a DLN.

B.1 Mathematical proofs

B.1.1 Modelisation

Recall that the weights $(w_{t,\pm})_{t\geq 0}$ are defined through the following SDE

$$dw_{t,\pm} = \pm[\bar{X}^T r(w_t)] \odot w_{t,\pm}dt + 2\sqrt{\gamma L(w_t)}w_{t,\pm} \odot \bar{X}^T dB_{t,\pm} + 2\sigma\sqrt{\gamma L(w_t)}w_{t,\pm} \odot d\tilde{B}_{t,\pm}$$

The Euler discretization of that SDE with step size γ is exactly

$$w_{t+1,\pm} = w_{t,\pm} - \gamma\nabla_{w_{t,\pm}} L(w_t) \pm 2\sqrt{\gamma L(w_t)}[\bar{X}^T\epsilon] \odot w_{t,\pm} \pm \gamma\sigma_t Z_{t,\pm} \odot w_{t,\pm},$$

Where $\epsilon \sim \mathcal{N}(0, \sqrt{\gamma}Id)$. As seen in (Pesme et al., 2021, Appendix A) the covariance of $2\sqrt{nL(w_t)}\frac{1}{\sqrt{\gamma}}\epsilon$ is equal to the one of ζ_{i_t}. Where

$$\zeta_{i_t} = -\left(\langle\beta - \beta^*, x_{i_t}\rangle e_{i_t} - \mathbb{E}_{i_t}[\langle\beta - \beta^*, x_{i_t}\rangle e_{i_t}]\right)$$

with i_t the random index chosen for the step of SGD. Thus the discretization scheme defines a Markov-Chain whose noise is the one found in equation (18) (the other term being independant from the rest).

B.1.2 Proof of Proposition 1

This section and the following one essentially adapt the proof of a similar result found in Pesme et al. (2021) in order to generalize it to more complex noise pattern. In particular extend it to a noise which does not depend on the loss and thus the geometry of the gradients. We consider throughout this section a more general setting. Let $A \in \mathbb{R}^{p\times d}$ which in the main text is either equal to $\bar{X}$ and thus $p = n$ or to $\left(\bar{X} \mid \sigma I_d\right)$ and $p = d + n$. And let a covariance noise matrix (σ_t) which is deterministic and valued in $\mathbb{R}^{p'\times d}$. Set $(w_{t,\pm})$ solutions of the following SDE with $w_{0,\pm} = \alpha$.

$$dw_{t,\pm} = -\nabla_{w_\pm} L(\beta_t)dt \pm 2\sqrt{\gamma L(\beta_t)}w_{t,\pm} \odot \left[A^T dB_t\right] \pm 2w_{t,\pm} \odot \left[\sigma_t^T d\tilde{B}_t\right]$$
$$= \pm\left(-\bar{X}^T r(\beta_t) \odot w_{t,\pm}dt + 2\sqrt{\gamma L(\beta_t)}w_{t,\pm} \odot \left[A^T dB_t\right] + 2w_{t,\pm} \odot \left[\sigma_t^T d\tilde{B}_t\right]\right)$$

where $r(w) = \bar{X}(w_+^2 - w_-^2 - \beta^*)$ and $(B_t)_t, (\tilde{B}_t)_t$ are two independant standard brownian motions respectively valued in $\mathbb{R}^p$ and $\mathbb{R}^{p'}$. This setting accounts for a decaying noise which has a geometry different to the one of the data as well a as a small noise which is solely of the magnitude of the parameters.

Lemma 1. *Consider the iterates (w_t) defined above (9) then $(\beta_t = w_{t,+}^2 - w_{t,-}^2)$ satisfies*

$$\beta_t = 2\alpha_t^2 sinh(2\eta_t + 2\delta_t)$$

where

$$\eta_t = -\int_0^t \bar{X}^T r(w_s)ds + 2\sqrt{\gamma}\int_0^t \sqrt{L(w_s)}A^T dB_s, \quad \delta_t = 2\int_0^t \sigma_s^T d\tilde{B}_s$$

and

$$\alpha_t = \alpha \odot \exp\left(-2\gamma diag(A^T A)\int_0^t L(\beta_s)ds - 2\int_0^t diag(\sigma_t^T\sigma_t)ds\right)$$

Proof. A direct application of Itô's lemma grants

$$w_{t,\pm} = w_{0,\pm} \odot \exp(\pm\left[-\int_0^t \bar{X}^T r(w_s)ds + 2\sqrt{\gamma}\int_0^t \sqrt{L(w_s)}A^T dB_s + 2\int_0^t \sigma_s^T d\tilde{B}_s\right])$$
$$\odot \exp(-2\gamma \mathrm{diag}(A^T A)\int_0^t L(\beta_s)ds - 2\int_0^t \mathrm{diag}(\sigma_t^T\sigma_t)ds)$$

Thus since $\beta_t = w_{t,+}^2 - w_{t,-}^2$ we have

$$\begin{aligned}\beta_t &= \alpha_t^2 \odot (\exp(+2\eta_t + 2\delta_t) - \exp(-2\eta_t - 2\delta_t)) \\ &= 2\alpha_t^2 \sinh(2\eta_t + 2\delta_t)\end{aligned}$$

□

The next result is the generalization of the result introducing the notion of mirror gradient descent with varying potential found in (Pesme et al., 2021, Proposition 1).

Proposition 4. *The flow $(\beta_t)_{t\geq 0}$ associated to $(w_{t,\pm})_{t\geq 0}$ follows a stochastic mirror gradient with varying potential defined by:*

$$d\nabla\phi_{\alpha_t}(\beta_t) = -\nabla L(\beta_t)dt + \sqrt{\gamma L(\beta_t)}A^T dB_t + \sigma_t^T d\tilde{B}_t$$

Proof. The expression of β_t from lemma 1 can be inversed in order to use ϕ_{α_t}. Indeed

$$\mathrm{arcsinh}\left(\frac{\beta_t}{2\alpha_t^2}\right) = 2\eta_t + 2\delta_t$$

which implies

$$d\mathrm{arcsinh}\left(\frac{\beta_t}{2\alpha_t^2}\right) = -2\bar{X}^T r(w_t) + 4\sqrt{\gamma L(\beta_t)}A^T dB_t + 4\sigma_t^T d\tilde{B}_t$$

Notice that $\nabla L(\beta_t) = \frac{1}{2}\bar{X}^T r(w_t)$ and $\nabla\phi_\alpha(\beta) = \frac{1}{4}\mathrm{arcsinh}(\frac{\beta}{2\alpha^2})$ we have the wanted result. □

B.1.3 Proof of Theorem 2

In order to prove the convergence of the integral of the loss we will introduce a perturbation of a Bregman divergence which will control the norm of the iterates and the integral of the loss. Moreover that process will be nicely controlled with high probability. Let β^* any interpolator, the process is the following

$$\begin{aligned}V_t = &-\phi_{\alpha_t}(\beta_t) + \langle\nabla\phi_{\alpha_t}(\beta_t), \beta_t - \beta^*\rangle \\ &+ \langle|\beta^*|, \gamma\mathrm{diag}(A^T A)\int_0^t L(\beta_s)ds + \int_0^t \mathrm{diag}(\sigma_s^T\sigma_s)ds\rangle\end{aligned}$$

We denote by D_t the Bregman divergence part.

Lemma 2. *For $t \geq 0$ we have*

$$\begin{aligned} dD_t = &-2L(\beta_t) + \gamma L(\beta_t)\langle \xi_t, diag(A^T A)\rangle dt + \langle \xi_t, diag(\sigma_t^T \sigma_t)\rangle dt \\ &+ \sqrt{\gamma L(\beta_t)}\langle A^T dB_t, \beta_t - \beta^*\rangle + \langle \sigma_t^T d\tilde{B}_t, \beta_t - \beta^*\rangle \end{aligned}$$

where $\xi_t = \sqrt{\beta_t^2 + 4\alpha_t^4} = w_{t,+}^2 + w_{t,-}^2$

Proof. By definition $D_t = -\phi_{\alpha_t}(\beta_t) + \langle \nabla \phi_{\alpha_t}(\beta_t), \beta_t - \beta^* \rangle$ thus a direct application of Itô's lemma grants

$$\begin{aligned} dD_t &= -\langle \nabla \phi_{\alpha_t}(\beta_t), d\beta_t \rangle - \langle \nabla_\alpha \phi(\alpha_t^2, \beta), d[\alpha_t^2] \rangle - \frac{1}{2}\mathrm{Tr}\left(\nabla^2 \phi_{\alpha_t}(\beta_t) d\langle \beta \rangle_t\right) \\ &+ \langle d\nabla \phi_{\alpha_t}(\beta_t), \beta_t - \beta^* \rangle + \langle \nabla \phi_{\alpha_t}(\beta_t), d\beta_t \rangle + \mathrm{Tr}\left(d\langle \nabla \phi_{\alpha_t}(\beta_t), \beta_t \rangle\right) \\ &= -\langle \nabla_\alpha \phi(\alpha_t^2, \beta), d[\alpha_t^2] \rangle - \frac{1}{2}\mathrm{Tr}\left(\nabla^2 \phi_{\alpha_t}(\beta_t) d\langle \beta \rangle_t\right) + \langle d\nabla \phi_{\alpha_t}(\beta_t), \beta_t - \beta^* \rangle \\ &+ \mathrm{Tr}\left(d\langle \nabla \phi_{\alpha_t}(\beta_t), \beta_t \rangle\right) \end{aligned}$$

We will compute one by one the four terms left. Note that todo so we will need to compute the quadratic variation of β_t. Since β_t can be decomposed in a drift part and a martingale part as follows

$$d\beta_t = \mathrm{drift}(\beta)_t dt + 4\sqrt{\gamma L(\beta_t)}\xi_t \circledcirc A^T dB_t + 4\xi_t \circledcirc \sigma_t^T d\tilde{B}_t$$

we obtain this formula for the quadratic variation

$$d\langle \beta \rangle_t = 16\gamma L(\beta_t)(A^T A) \circledcirc (\xi_t \xi_t^T) + 16(\sigma_t^T \sigma_t) \circledcirc (\xi_t \xi_t^T)$$

Since α_t has no martingale part the chain rule grants that the first term is equal to

$$\langle \nabla_\alpha \phi(\alpha_t^2, \beta), d[\alpha_t^2] \rangle = \gamma L(\beta_t)\langle \xi_t, \mathrm{diag}(A^T A)\rangle dt + \langle \xi_t, \mathrm{diag}(\sigma_t^T \sigma_t)\rangle dt$$

Since $\nabla^2 \phi_\alpha(\beta) = \frac{1}{4}\mathrm{diag}(\frac{1}{\xi})$ the second term is equal to

$$\frac{1}{2}\mathrm{Tr}\left(\nabla^2 \phi_{\alpha_t}(\beta_t) d\langle \beta \rangle_t\right) = 2\gamma L(\beta_t)\langle \xi_t, \mathrm{diag}(A^T A)\rangle dt + 2\langle \xi_t, \gamma \mathrm{diag}(\sigma_t^T \sigma_t)\rangle dt$$

Using proposition 4 the third term is equal to

$$\langle d\nabla \phi_{\alpha_t}(\beta_t), \beta_t - \beta^* \rangle = -\langle \nabla L(\beta_t), \beta_t - \beta^* \rangle + \sqrt{\gamma L(\beta_t)}\langle A^T dB_t, \beta_t - \beta^* \rangle + \langle \sigma_t^T d\tilde{B}_t, \beta_t - \beta^* \rangle$$

the fourth term

$$\mathrm{Tr}\left(d\langle \nabla \phi_{\alpha_t}(\beta_t), \beta_t \rangle\right) = 4\gamma L(\beta_t)\langle \xi_t, \mathrm{diag}(A^T A)\rangle dt + 4\langle \xi_t, \gamma \mathrm{diag}(\sigma_t^T \sigma_t)\rangle dt$$

Note that

$$\langle \nabla L(\beta_t), \beta_t - \beta^* \rangle = 2L(\beta_t)$$

Thus combined it grants

$$\begin{aligned} dD_t = &-2L(\beta_t) + \gamma L(\beta_t)\langle \xi_t, \mathrm{diag}(A^T A)\rangle dt + \langle \xi_t, \mathrm{diag}(\sigma_t^T \sigma_t)\rangle dt \\ &+ \sqrt{\gamma L(\beta_t)}\langle A^T dB_t, \beta_t - \beta^* \rangle + \langle \sigma_t^T d\tilde{B}_t, \beta_t - \beta^* \rangle \end{aligned}$$

□

The Bregman divergence is useful to control the iterations as the next lemma will state. Moreover under suitable assumptions on the noise σ_t the martingale part being close to its quadratic variation will enable us to control V_t itself and in turns the loss.

Lemma 3. *For $t \geq 0$ we have*

$$\|\xi_t\|_1 \leq 4V_t + \|\beta^*\|_1 \ln\left(\frac{\sqrt{2}\|\xi_t\|_1}{\min \alpha_i^2}\right)$$

Proof. A direct computation grants

$$\|\xi_t\|_1 = 4D_t + \langle \operatorname{arcsinh}\frac{\beta_t}{4\alpha_t^2}, \beta^* \rangle \tag{39}$$

Since $\operatorname{arcsinh}(x) \leq \ln(2(x+1))$ we have

$$\begin{aligned}
\langle \operatorname{arcsinh}\frac{\beta_t}{4\alpha_t^2}, \beta^* \rangle &\leq \sum_i |\beta_i^*| \ln\left(\frac{|\beta_{i,t}| + 2\alpha_{i,t}^2}{\alpha_{i,t}^2}\right) \\
&\leq \sum_i |\beta_i^*| \ln\left(\frac{|\beta_{i,t}| + 2\alpha_{i,t}^2}{\min \alpha_i^2}\right) + 4\langle |\beta^*|, \gamma \operatorname{diag}(A^T A)\int_0^t L(\beta_s)ds + \int_0^t \operatorname{diag}(\sigma_s^T \sigma_s)ds\rangle \\
&\leq \|\beta^*\|_1 \ln\left(\frac{\sqrt{2}\|\xi_t\|_1}{\min \alpha_i^2}\right) + 4\langle |\beta^*|, \gamma \operatorname{diag}(A^T A)\int_0^t L(\beta_s)ds + \int_0^t \operatorname{diag}(\sigma_s^T \sigma_s)ds\rangle
\end{aligned}$$

Because $\alpha_t^2 = \alpha^2 \circledcirc \exp\left(-4\gamma \operatorname{diag}(A^T A)\int_0^t L(\beta_s)ds - 4\int_0^t \operatorname{diag}(\sigma_s^T \sigma_s)ds\right)$ and $|\beta_{i,t}| + 2\alpha_{i,t}^2 \leq \sqrt{2}\|\xi_t\|_1$. Thus plugging that inequation in the first relationship (39) grants the wanted result. □

We now need to control V_t, however in order to only have to manage bounded variation processes we will use the fact that a Martingale is controlled by its quadratic variation

Lemma 4 (Howard et al. (2020), Corollary 11)**.** *For any locally square integrable process (S_t) with a.s. continuous trajectories and any $a, b > 0$*

$$\mathbb{P}(\exists t \geq 0, S_t \geq a + b\langle S \rangle_t) \leq \exp(-2ab)$$

We can use this lemma for the following two processes

$$\int_0^t \sqrt{\gamma L(\beta_s)}\langle A^T dB_s, \beta_s - \beta^* \rangle, \quad \int_0^t \langle \sigma_s^T d\tilde{B}_s, \beta_s - \beta^* \rangle$$

For $a, b > 0$ we have thanks to lemma 4 with probability at least $1 - 2\exp(-ab)$ for any $t \geq 0$

$$\begin{aligned}
&\left|\int_0^t \sqrt{\gamma L(\beta_s)}\langle A^T dB_s, \beta_s - \beta^* \rangle + \langle \sigma_s^T d\tilde{B}_s, \beta_s - \beta^* \rangle\right| \\
&\leq a + b\gamma\|A\|^2 \int_0^t L(\beta_s)(\|\beta_s\|^2 + \|\beta^*\|^2)ds + b\int_0^t \|\sigma_s\|^2(\|\beta_s\|^2 + \|\beta^*\|^2)ds
\end{aligned}$$

because the quadratic variation of both processes are upperbounded by the term on the right. We denote by $\mathcal{A}$ the set on which the inequality above is satsfied, we recall that this set has probability at least $1 - 2\exp(-ab)$.

Lemma 5. *On $\mathcal{A}$ for γ and $\int \|\sigma_s\|^2 ds$ small enough the iterates are bounded by a constant that only depends on $a, b, \|\beta^*\|, \alpha$.*

Proof. Since we are on the event $\mathcal{A}$ the bound on the martingale part is true and thus

$$V_t \leq V_0 + a - 2\int_0^t L(\beta_s)U_s ds + \int_0^t \langle |\beta^*|, \operatorname{diag}(\sigma_s^T \sigma_s)\rangle + b\|\sigma_s\|^2(\|\beta_s\|^2 + \|\beta^*\|^2)ds$$

Where $U_s = 1 - \frac{\gamma}{2}\left[\langle \operatorname{diag}(A^T A, \xi_s + |\beta^*|\rangle + 2b\|A\|^2(\|\beta_s\|^2 + \|\beta^*\|^2)ds)\right]$. Assume that $\int_0^\infty \|\sigma_s\|^2 ds \leq 1$, then we have

$$V_t \leq C - 2\int_0^t L(\beta_s)U_s ds + \int_0^t C'\|\sigma_s\|^2\|\beta_s\|^2 ds$$

Where C depends on $a, b, \|\beta^*\|_1$ and V_0 which depends on α also C' depends on b. In particular if $U_s \geq 0$ for $s \in [0, \tau]$ using lemma 3 we have for other constants C, C' and $t \in [0, \tau]$

$$\|\xi_t\|_1 \leq \|\beta^*\|_1 \ln\left(\frac{\sqrt{2}\|\xi_t\|_1}{\min \alpha_i^2}\right) + C + \int_0^t C'\|\sigma_s\|\|\beta_s\|^2 ds$$

Since $\|\beta_s\|^2 \leq \|\xi_s\|^2$ and by the sublinear growth of the log we get

$$\|\xi_t\|_2 \leq C + \int_0^t C' \|\sigma_s\|^2 \|\xi_s\|^2 ds$$

For another set of constants C, C'. We now use a bootstrap argument let $I = \{t \in [0,\tau] \mid \forall s \leq t, \|\xi_t\|_2 \leq 2C\}$. Assume that $\int_0^\infty \|\sigma_s\|^2 ds \leq 1/4CC'$ then $I = [0,\tau]$. Indeed $0 \in I$ and I is an interval. Let $t = \sup I$, first $t \in I$ and if $t = \tau$ there is nothing to prove. If $t < \tau$ then

$$\frac{\|\xi_t\|_2}{2C} \leq \frac{1}{2} + \int_0^t 2CC' \|\sigma_s\|^2 \left(\frac{\|\xi_s\|_2}{2C}\right)^2 ds \leq \frac{1}{2} + \int_0^t 2CC' \|\sigma_s^2\| ds < 1$$

And by continuity of ξ we have $t + \epsilon \in I$. Thus $I = [0,\tau]$ and for any $t \in [0,\tau]$ we have that ξ is bounded by a constant which only depends on the parameters of the problem. Now we shall see that for a nice choice of γ we have $\tau = +\infty$ on $\mathcal{A}$. Indeed for γ small enough we have $U_0 \geq 1/2$. Now let t the waiting time for U to reach $1/2$. Then by continuity of U for $s \leq t$ we have $U_s \geq 1/2 > 0$ thus the bound on the iterates found above is valid and for γ small enough we have that $U_t > 1/2$ which is a contradiction by continuity of U. Thus there is a threshold under which U is always greater than $1/2$. Thus the above bound on the iterates is valid for all times. □

Proposition 5. *On $\mathcal{A}$ for γ and $\int \|\sigma_s\|^2 ds$ small enough the integral of the loss converges.*

Proof. In the proof of the last lemma we have shown that for the right choice of γ and $\int \|\sigma_s\|^2 ds$, U_s is greater than $\frac{1}{2}$ on $\mathcal{A}$. Thus by boundedness of the iterates we have

$$\int_0^t L(\beta_s) ds \leq C - V_t + C' \int_0^t \|\sigma_s\|^2 ds$$

It remains to lower bound V in order to prove convergence of the integral of the loss. Or the computations in lemma 3 shows that

$$V_t \geq \frac{1}{4}\left(\|\xi_t\|_1 - \|\beta^*\|_1 \ln\left(\frac{\sqrt{2}\|\xi_t\|_1}{\min \alpha_i^2}\right)\right)$$

which is lower bounded because the iterates are bounded. Thus the integral converges. □

Finally we have all the ingredients to show that the iterates converge.

Proposition 6. *On $\mathcal{A}$ for γ and $\int \|\sigma_s\|^2 ds$ small enough the iterates converge.*

Proof. We have the following integral expression of D_t thanks to lemma 2

$$\begin{aligned} D_t = D_0 &+ \int_0^t -2L(\beta_s) + \gamma L(\beta_s)\langle \xi_s, \mathrm{diag}(A^T A)\rangle + \langle \xi_s, \mathrm{diag}(\sigma_s^T \sigma_s)\rangle ds \\ &+ \int_0^t \sqrt{\gamma L(\beta_s)}\langle \bar{X}^T dB_s, \beta_s - \beta^*\rangle + \langle \sigma_s d\tilde{B}_s, \beta_s - \beta^*\rangle ds \end{aligned}$$

Note that D_t converges because the bouded variation part converges due to the convergence of the integral of the loss and the boundedness of the iterates. The martingale part converges because the quadratic variation is converging. Recall that

$$D_t = -\phi_{\alpha_t} + \langle \nabla \phi_{\alpha_t}(\beta_t), \beta_t - \beta^*\rangle$$

We have that D converges for any choice of interpolator β^*.
The integral of the loss is convergent thus up to extraction $L(\beta_t)$ converges to 0. Since β_t is bounded a second extraction ensures the converges of the iterates. Thus there is a subsequence $\beta_{\psi(t)}$ such that it converges to β_∞ and $L(\beta_\infty) = 0$ thus it is an interpolator. And we have

$$\begin{aligned} \phi_{\alpha_\infty}(\beta_\infty) - D_t &= \phi_{\alpha_\infty}(\beta_\infty) - \phi_{\alpha_t}(\beta_t) - \langle \nabla \phi_{\alpha_t}(\beta_t), \beta_t - \beta_\infty\rangle \\ &\geq \phi_{\alpha_t}(\beta_\infty) - \phi_{\alpha_t}(\beta_t) - \langle \nabla \phi_{\alpha_t}(\beta_t), \beta_t - \beta_\infty\rangle \\ &= D_{\phi_{\alpha_t}}(\beta_\infty, \beta) \\ &\geq 0 \end{aligned}$$

Because $\alpha \mapsto \phi_\alpha(\beta)$ is decreasing and $\alpha_t \geq \alpha_\infty$. Note that $\phi_{\alpha_\infty}(\beta_\infty) - D_{\psi(t)} \to 0$ by convergence of $\beta_{\psi(t)}$ to β_∞. Thus by convergence of D we have that $\phi_{\alpha_\infty}(\beta_\infty) - D_t \to 0$ and in turn $D_{\phi_{\alpha_t}}(\beta_\infty, \beta) \to 0$. Recall that

$$\nabla^2 \phi_\alpha(\beta) = \frac{1}{4}\text{diag}\left(\frac{1}{\sqrt{\beta^2 + 4\alpha^2}}\right) \geq \frac{1}{4\sqrt{\|\beta\|^2 + 4\max \alpha_i^2}} Id$$

thus ϕ_{α_t} is strongly convex of with a parameter that does not depend on t by decreasingness of α_t and boundedness of β_t. Thus we have that $D_{\phi_{\alpha_t}}(\beta_\infty, \beta) \geq \mu/2\|\beta_t - \beta_\infty\|^2$ which concludes the convergence. □

B.1.4 Proof of Proposition 3

In the last section we have shown that the iterates converge toward an interpolator. However that interpolator is not a minimizer of the potential ϕ_{α_∞} as shown in proposition 2. However by strong convexity of ϕ_α we have a control of the distance between β_∞ and the minimizer of ϕ_{α_∞} which we denote by $\beta^*_{\alpha_\infty}$

Proof. Let $\phi(\beta) = \phi_{\alpha_\infty}(\beta) + \iota_{X\beta=Y}$ where ι is the indicator function equal to 0 if the condition is satisfied and $+\infty$ otherwise. Note that ϕ is still strongly convex. Indeed, we can show that $\phi - \mu\|.\|^2$ is convex: let $\beta, \nu \in \mathbb{R}^d$, $\lambda \in (0,1)$. $C = \{\beta \text{ s.t. } X\beta = Y\}$ is a convex set as an affine subspace of $\mathbb{R}^d$. Therefore:

$$\lambda\beta + (1-\lambda)\nu \notin C \implies \beta \notin C \text{ or } \nu \notin C \tag{40}$$

$$\iota_{X(\lambda\beta+(1-\lambda)\nu)=Y} = \infty \implies \iota_{X\beta=Y} = \infty \text{ or } \iota_{X\nu=Y} = \infty \tag{41}$$

We also have that $\phi_{\alpha_\infty} - \mu\|.\|^2$ is convex on a bounded set which contains β_∞ and $\beta^*_{\alpha_\infty}$. Finally $\phi - \mu\|.\|^2$ is convex, i.e. ϕ is $\mu-$strongly convex on a bounded set.

We have β_∞ solution of

$$\min \phi(\beta) - \langle r_\infty, \beta\rangle \tag{42}$$

And $\beta^*_{\alpha_\infty}$ is the solution to

$$\min \phi(\beta) \tag{43}$$

Since ϕ is strongly convex we have that $f = \phi - \mu\|\cdot - \beta^*\|^2$ is convex. Notice that $0 \in \partial f(\beta_\infty) - r_\infty$ thus

$$\phi(\beta^*) - \langle r_\infty, \beta_0\rangle - \mu\|\beta^* - \beta_\infty\|^2 \geq \phi(\beta_\infty) - \langle r_\infty, \beta_\infty\rangle \tag{44}$$

We thus observe that:

$$\phi(\beta^*) - \langle r_\infty, \beta^*\rangle \geq \phi(\beta_\infty) - \langle r_\infty, \beta_\infty\rangle + \mu\|\beta^* - \beta_\infty\|^2 \tag{45}$$

Thus by optimality of β^* ($\phi(\beta^*) - \phi(\beta_\infty) \leq 0$)

$$\begin{aligned}\langle r_\infty, \beta_\infty - \beta^*\rangle &\geq \phi(\beta^*) - \langle r_\infty, \beta^*\rangle - (\phi(\beta_\infty) - \langle r_\infty, \beta_\infty\rangle) \\ &\geq \mu\|\beta^* - \beta_\infty\|^2\end{aligned} \tag{46}$$

Finally using Cauchy-Schwarz inequality we have

$$\boxed{\|r_\infty\| \geq \mu\|\beta^* - \beta_\infty\|} \tag{47}$$

□

B.2 Additional Numerical Experiments

Reminder We adopt the same set-up as Pesme et al. (2021) for sparse regression. We select parameters n = 40 and d = 100, and then create a sparse model $\beta^*_{l_0}$ with the constraint that its l_0 norm is equal to 5. We generate the features x_i from a normal distribution with mean 0 and identity covariance matrix $N(0, I)$, and compute the labels as $y_i = x_i^T \beta^*_{l_0}$. We always use the same step size of $\gamma = 1/(1.3||\bar{X}^T\bar{X}||_2)$. Notice that $||\beta_t - \beta^*_{l_0}||_2^2$ is the validation loss in the experiments.

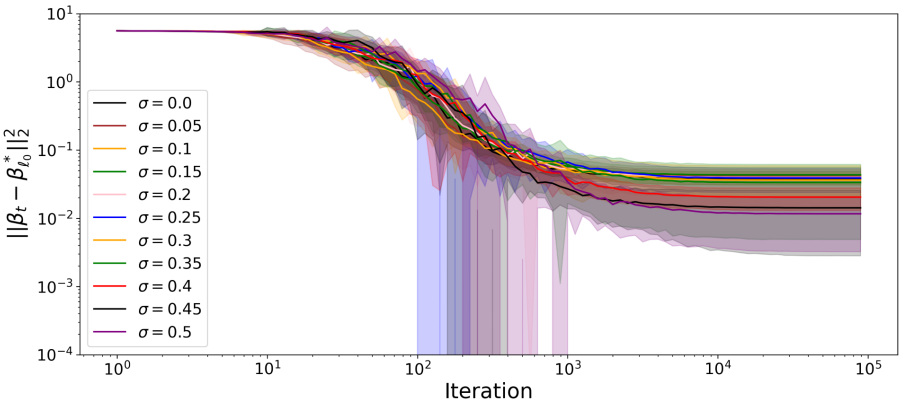

Figure 5: Diagonal Linear Network from $\alpha = 0.1$: Implicit Bias of SGD and Noisy SGD different values of σ in Equation 18, **from** $\alpha = 0.1$. Shaded areas represent one standard variation over 5 runs, and plain lines represent the average values. Starting from this initialization, the bigger σ is, the closer the solution obtained with Noisy-SGD is to the sparse solution $\beta^*_{l_0}$.

Introduction As outlined in Section 3.3 and prooved in Section B.1 in a more general setting, our investigation reveals that Noisy-SGD yields an identical solution to GD, albeit (1) operating on an altered potential ϕ_{α_∞}, and (2) starting from an effective non-zero initialization denoted as $\tilde{\beta}_0$. In contrast, the sole parameter affected by SGD is α. In the context of Noisy-SGD, it is noteworthy that α_∞ diminishes with increasing σ. This signifies that as more noise is introduced, the effective α becomes smaller. If this did not influence the effective initialization, it would directly imply an "improved" implicit bias, in the case of the sparse regression under study.

Trade off To be more precise, if the effective initialization $\tilde{\beta}_0$ remained unaltered, the noisy process would ultimately converge to $\beta^\star_{\alpha_\infty}$, which is the solution that minimizes ϕ_{α_∞}. We already know that this solution is sparser than the one obtained via SGD because α_∞ is smaller (as referenced in Equation 21). As demonstrated in Proposition 3, the presence of the new effective initialization $\tilde{\beta}_0$ does introduce some perturbation to the solution. However, the degree of its influence directly depends on the level of noise introduced. This ensures that β_∞, which is the solution obtained through Noisy-SGD, remains in close proximity to $\beta^\star_{\alpha_\infty}$. This scenario implies the existence of a tradeoff. Increasing the value of σ will:

1. Augment the sparsity of $\beta^\star_{\alpha_\infty}$ as $\beta^\star_{\alpha_\infty}$ gets closer to the (sparse) ground truth $\beta^\star_{l_0}$
2. Expand the gap between β_∞ and $\beta^\star_{\alpha_\infty}$, which could potentially diminish the impact of the first outcome if β_∞ gets far from $\beta^\star_{l_0}$

In summary, if the influence of **1.** surpasses the effect of **2.**, Noisy-SGD would enhance the implicit bias. Conversely, if **2.** outweighs **1.**, this enhancement may not necessarily occur. To empirically quantify the trade-off, we measure the gap between $\beta^\star_{\alpha_\infty}$ and β_∞ across varying values of α and σ. Our approach involves (a) obtaining β_∞ through Noisy-SGD, (b) approximating α_∞ numerically based on the loss logs, and (c) subsequently executing GD with $\alpha = \alpha_\infty$ to obtain $\beta^\star_{\alpha_\infty}$. In Figure 4, our findings underscore that, for $\alpha = 0.1$, the distance between $\beta^\star_{\alpha_\infty}$ and β_∞ seem to indeed increase with growing values of σ, validating the second point (**2.**). Moreover, as revealed in Figure 3, when using $\alpha = 0.1$, Noisy-SGD consistently converges to sparser solutions compared to standard SGD. This trend remains consistent across various noise levels, as illustrated in Figure 5. Specifically, as we elevate the magnitude of σ, the solutions become progressively closer to the sparse interpolator (ground truth). In this context, the influence of the new initialization $\tilde{\beta}_0$ appears to be of secondary importance, with therefore point **1.** consistently holding more significance than point **2.**

Nonetheless, Figure 4 reveals that the magnitude of the gap between $\beta^\star_{\alpha_\infty}$ and β_∞ is smaller or comparable to the gap between β_∞ and $\beta^\star_{l_0}$ in Figure 3. This observation implies that Noisy-SGD may be advantageous only in this specific scenario, as the impact of **2.** is small. However, a question arises: if the initial α were even smaller, would **1.** still dominate **2.**?

Figure 6 provides insight into this inquiry. Even with an initial α reduced by a factor of ten, we observe that the

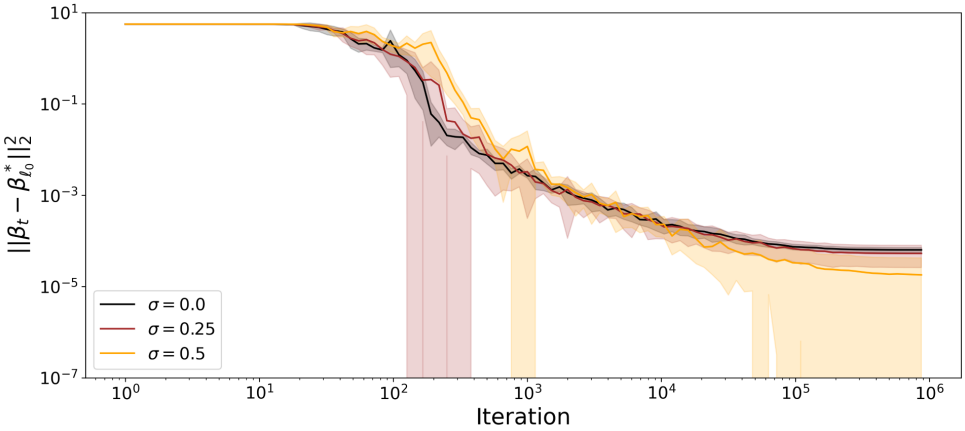

Figure 6: Diagonal Linear Network from $\alpha = 0.01$: Implicit Bias of SGD and Noisy SGD different values of σ in Equation 18, **from small initialization** $\alpha = 0.01$. Shaded areas represent one standard variation over 5 runs, and plain lines represent the average values. Starting from this initialization, the bigger σ is, the closer the solution obtained with Noisy-SGD is to the sparse solution $\beta^*_{l_0}$.

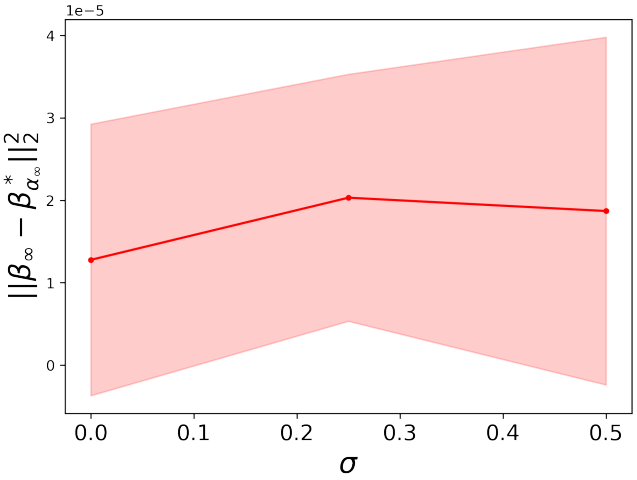

Figure 7: DLN from $\alpha = 0.1$: Distance between $\beta^*_{\alpha_\infty}$, the solution that minimizes ϕ_{α_∞} —obtained by GD from α_∞— and β_∞, the one obtained by Noisy-SGD **from** $\alpha = 0.01$ (see Proposition 3). Shaded areas represent one standard deviation over 5 runs. Similar to the case of $\alpha = 0.1$ (see Figure 4), the distance is of the same order of magnitude than the distance between β_∞ and the sparse solution β_{l_0} (see Figure 6). It explains why why the implicit bias is enhanced by Gaussian noise in this case.

introduction of noise continues to exert a positive influence on the sparsity of the solution. Furthermore, when examining the gap between $\beta^\star_{\alpha_\infty}$ and β_∞ in Figure 7, we find that, similarly to the case where $\alpha = 0.1$, it remains of a comparable order of magnitude to the gap between β_∞ and $\beta^\star_{l_0}$ observed in Figure 6. This observation helps to elucidate why the advantages of Noisy-SGD persist under these conditions: the distance between $\beta^\star_{\alpha_\infty}$ and β_∞ empirically also depends (inversely) on α, which mitigates the impact of **2.** over **1.**

Conclusion: The effect of Noisy-SGD on the solution's sparsity, as discussed in Section 3.3, remains significant even when initiated with smaller values of α, which corresponds to cases where the solutions obtained from GD and SGD already exhibit some enhanced degree of sparsity. Therefore, the practical implications of the theoretical trade off appears to be modest, and the reinforcement of the implicit bias from Noisy-SGD prevails: the introduction of Gaussian noise enhances or at least maintains the implicit bias of SGD in a robust way.