IMPLICIT REGULARISATION IN DIFFUSION MODELS: AN ALGORITHM-DEPENDENT GENERALISATION ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

The success of denoising diffusion models raises important questions regarding their generalisation behaviour, particularly in high-dimensional settings. Notably, it has been shown that when training and sampling are performed perfectly, these models memorise training data—implying that some form of regularisation is essential for generalisation. Existing theoretical analyses primarily rely on algorithm-independent techniques such as uniform convergence, heavily utilising model structure to obtain generalisation bounds. In this work, we instead leverage the algorithmic aspects that promote generalisation in diffusion models, developing a general theory of algorithm-dependent generalisation for this setting. Borrowing from the framework of algorithmic stability, we introduce the notion of score stability, which quantifies the sensitivity of score-matching algorithms to dataset perturbations. We derive generalisation bounds in terms of score stability, and apply our framework to several fundamental learning settings, identifying sources of regularisation. In particular, we consider denoising score matching with early stopping (denoising regularisation), sampler-wide coarse discretisation (sampler regularisation) and optimising with SGD (optimisation regularisation). By grounding our analysis in algorithmic properties rather than model structure, we identify multiple sources of implicit regularisation unique to diffusion models that have so far been overlooked in the literature.

1 INTRODUCTION

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) are a class of generative models that have achieved state-of-the-art performance across image, audio, video, and protein synthesis tasks (Rombach et al., 2022; Saharia et al., 2022; Ramesh et al., 2022; Watson et al., 2023; Esser et al., 2024). Their ability to generate high-quality samples from complex, high-dimensional distributions with limited data motivates the need for a theoretical understanding of the mechanisms underpinning their strong generalisation capabilities.

The goal of diffusion models is to generate new synthetic samples from a data distribution ν_{data} using a finite set of N data points $\{x_i\}_{i=1}^N$. Central to the methodology is a unique approach to generating data, formulating it as the iterative transformation of noise into data, or equivalently, the reversal of a diffusion process (Song et al., 2021). This diffusion process, called the *forward process*, is defined by the stochastic differential equation (SDE),

$$dX_t = -\alpha X_t\, dt + \sqrt{2}\, dW_t, \qquad X_0 \sim \nu_{\mathrm{data}}, \qquad t \in [0, T], \tag{1}$$

for some $\alpha \geq 0$, where W_t denotes the Brownian motion in $\mathbb{R}^d$ and $T > 0$ is the terminal time. It can then be shown that the time-reversal of this process, $Y_t := X_{T-t}$ admits a weak formulation as a solution to the SDE,

$$dY_t = \alpha Y_t dt + 2\nabla \log p_{T-t}(Y_t) dt + \sqrt{2} dW_t, \qquad Y_0 \sim p_T, \qquad t \in [0, T), \tag{2}$$

where p_t denotes the marginal density of X_t (Haussmann & Pardoux, 1986). Therefore, simulating samples from $\nu_{\mathrm{data}} = p_0$ can be achieved by solving the diffusion process in (2), which requires

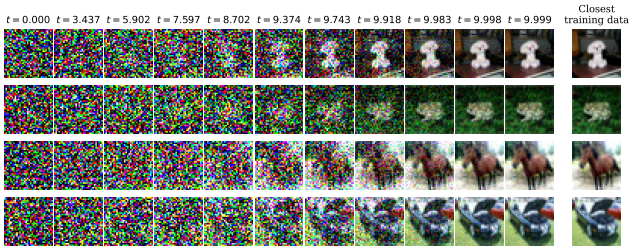

Figure 1: Samples generated using the empirical score function on CIFAR-10 compared to the closest image in the dataset, illustrating memorisation of the training data.

an approximation of the score function, $\nabla \log p_t$. This is achieved by fitting a time-dependent deep neural network to minimise a weighted L^2-distance called the *(population) score matching loss*:

$$\ell_{\mathrm{sm}}(s;\tau) := \int \mathbb{E}_{X_t}[\|s(X_t,t) - \nabla \log p_t(X_t)\|^2]\,\tau(dt), \tag{3}$$

where τ is a probability measure over $(0,T]$ that determines the weighting of the timepoints. Since $\nabla \log p_t$ is unknown, typically the (population) *denoising score matching loss* ℓ_{dsm}, which differs from $\ell_{\mathrm{sm}}(s)$ only by a constant, is used instead and is then approximated using the dataset, forming the *empirical denoising score matching loss* $\hat{\ell}_{\mathrm{dsm}}$ (see equations (6) and (8)). The score network $s(x,t)$ is trained on this objective using standard stochastic optimisation methods relying on mini-batching. Once an approximation is obtained, samples are generated by numerically solving the reverse-time SDE, (2). Both *score matching* and *sampling* introduce distinct challenges and design choices that impact the quality of model output (Karras et al., 2022).

Score matching presents a key difference from standard supervised learning. In the space of all L^2 score functions, the empirical objective $\hat{\ell}_{\mathrm{dsm}}$ possesses a unique minimiser—the empirical score function—as a result of the integration over $X_t|X_0$ (see Lemma 1). This contrasts with traditional supervised learning, where the empirical risk minimisation problem can have infinitely many solutions (e.g., in overparameterised regression) and often requires regularisation to be well-posed. As shown in Figure 1, sampling with this empirical score leads to exact recovery of the training data (Pidstrigach, 2022). This behaviour is distinct from 'benign overfitting', a phenomenon from the deep learning literature where interpolating the data does not necessarily prevent generalisation (Bartlett et al., 2021; Zhang et al., 2021). This divergence suggests that existing theory may be insufficient to explain the success of diffusion models, highlighting the need for new frameworks tailored to this setting.

Recently, there has been a drive towards developing theory for better understanding the unique structure of diffusion models. The most developed subset of this work focuses on connecting sample quality to score matching by deriving upper bounds on distribution error (e.g. KL divergence, total variation, or Wasserstein distance) between model samples and the data distribution, controlling it by the population score matching loss (De Bortoli et al., 2021; De Bortoli, 2022; Lee et al., 2022; Chen et al., 2023; Benton et al., 2024; Potaptchik et al., 2024). These results, often referred to as *convergence bounds*, typically take the form,

$$\text{Distribution error} \lesssim \ell_{\mathrm{sm}}(s) + \Delta,$$

where Δ is the discretisation error of the sampling scheme, which can be made small with sufficiently fine discretisation. However, since ℓ_{sm} is not computable, these bounds say little about performance under empirical guarantees—that is, their *generalisation* properties. One line of work, initiated by Oko et al. (2023) and extended in (Azangulov et al., 2024; Tang & Yang, 2024), applies classical uniform convergence theory to bound the generalisation gap from the decomposition,

$$\ell_{\mathrm{sm}}(s) = \hat{\ell}_{\mathrm{sm}}(s) + \underbrace{\ell_{\mathrm{sm}}(s) - \hat{\ell}_{\mathrm{sm}}(s)}_{\text{generalisation gap}}, \tag{4}$$

where $\hat{\ell}_{\mathrm{sm}}$ denotes the empirical counterpart to ℓ_{sm}. These results rely on covering number bounds for specific classes of neural networks and, while informative, they are limited to carefully chosen model classes and do not account for algorithmic properties. An alternative approach by De Bortoli

(2022) uses a decomposition of the Wasserstein distance that leverages convergence properties of the empirical measure. Though more model-agnostic, this method overlooks how diffusion models generate *novel* data. Both lines of work are fundamentally algorithm-independent, in that they lack any utilisation of the algorithmic aspects that uniquely define diffusion models. Recent efforts aim to incorporate algorithmic effects by restricting the problem. For instance, Shah et al. (2023); Chen et al. (2024) consider Gaussian mixture targets, while Li et al. (2023); Yang (2022) study random feature models. These settings allow for finer analysis of the role of the score matching algorithm, but remain limited in scope, leaving open the challenge of developing a more general algorithm-dependent theory of generalisation in diffusion models.

As noted earlier, if the empirical score matching loss was completely minimised and sampling was performed perfectly, the diffusion model would simply return training data, failing to generalise. Therefore, the observed success of diffusion models in producing novel data implies that, in practice, they either avoid completely minimising $\hat{\ell}_{\mathrm{sm}}$ or must avoid perfectly sampling. This suggests that (implicit) regularisation in the score matching or sampling algorithm is crucial for generalisation, making algorithmic considerations essential for understanding diffusion models.

1.1 Our contributions

We introduce score stability, a general, algorithm-dependent framework for analysing diffusion model generalisation based on the classical approach of algorithmic stability. This framework quantifies an algorithm's dependence on individual training examples, from which we derive expected generalisation gap bounds for score matching losses. Using the score stability framework, we then analyse several examples of score matching algorithms, identifying three distinct sources of implicit regularisation in diffusion model training and sampling: noising, sampler, and optimisation-induced regularisation.

Denoising regularisation. To begin with, we consider the empirical risk minimisation algorithm (ERM) that minimises $\hat{\ell}_{\mathrm{dsm}}$ over a hypothesis class $\mathcal{H}$. Through a score stability analysis, we reveal a regularisation source within this objective when early stopping of the forward process is used—a standard practice in the diffusion model literature. Utilising properties of the noising forward process, we obtain generalisation gap bounds with near-linear rate, $\epsilon^{-d^*/4}(\epsilon^{-d^*/2}N^{-2} + \min_{\mathcal{H}} \hat{\ell}_{\mathrm{sm}})^{c/2}$ for any $c < 1$, where $\epsilon > 0$ is the early stopping time and d^* is the dimension of the data support.

Sampler regularisation. We then apply this analysis to discrete-time sampling algorithms, deriving statistical guarantees for the expected KL divergence between the true data distribution and samples generated by the diffusion model. The bound we derive is formed of two stages: we obtain generic rates $\epsilon^{-1/2}(\epsilon^{-d^*/2}N^{-2} + \min_{\mathcal{H}} \hat{\ell}_{\mathrm{sm}})^{c/d^*}$ but when N^{-2} and $\min_{\mathcal{H}} \hat{\ell}_{\mathrm{sm}}$ are sufficiently small relative to ϵ, we obtain bounds with rates $\epsilon^{-d^*/4}(\epsilon^{-d^*/2}N^{-2} + \min_{\mathcal{H}} \hat{\ell}_{\mathrm{sm}})^{c/2}$ that are faster in N and $\min_{\mathcal{H}} \hat{\ell}_{\mathrm{sm}}$. To derive this bound, we utilise regularisation brought about by the coarseness of the discretisation. We find that by increasing discretisation coarseness, we can improve the generalisation gap bound at the expense of worsening the discretisation error term.

Optimisation regularisation. Finally, we consider the role of the optimisation scheme, analysing stochastic gradient descent (SGD) with gradient clipping and weight decay. On the model class, we assume only structural assumptions typical in the optimisation literature, including non-global Lipschitz and smoothness assumptions. While this initially yields bounds that grow with the number of iterations, we more closely inspect the impact of the high-variance gradient estimator used in diffusion training. We show this gradient noise induces a contractive behaviour in the training dynamics, which we harness to obtain stability bounds that do not grow with the number of iterations (Proposition 14), showing that the noisy dynamics enable tighter generalisation guarantees.

2 Background

Suppose that the data distribution ν_{data} is on $\mathbb{R}^d$ and we are provided a finite dataset of samples $S = \{x_1, ..., x_N\}$ which we assume are sampled independently and identically (i.i.d.) from ν_{data}. As discussed in the introduction, diffusion models are formed of two distinct stages. The first stage, *score matching*, consists of learning an approximation to the score function $\nabla \log p_t$ using the dataset S. In this work, we take a *score function* to be any function belonging to the set $L^0(\mathbb{R}^d \times [0, T]; \mathbb{R}^d)$, the

set of Borel measurable functions of the form $\mathbb{R}^d \times [0,T] \to \mathbb{R}^d$. Then, a *score matching algorithm* is taken to be any mapping of the form $A_{\text{sm}} : (\cup_{N=1}^{\infty}(\mathbb{R}^d)^{\otimes N}) \times \Omega \to \mathcal{H}$ where $\mathcal{H}$ is a measurable subset of $L^0(\mathbb{R}^d \times [0,T]; \mathbb{R}^d)$. Here, Ω is the event space belonging to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The second stage of diffusion models, *sampling*, consists of generating samples with the learned score function. We take a *sampling algorithm* to be a mapping of the form, $A_{\text{samp}} : \mathcal{H} \to \mathcal{P}(\mathbb{R}^d)$ where $\mathcal{P}(\mathbb{R}^d)$ denotes the set of Borel measures on $\mathbb{R}^d$. Typically, sampling is performed using an approximation to the reverse process given in (2), replacing $\nabla \log p_t$ with the learned score function $s(\cdot, t)$ and replacing its initial distribution, p_T with a data-independent prior, $p_{\text{prior}} = \mathcal{N}(\mathbf{0}, \sigma^2_{\text{prior}} I)$. In the case of $\alpha > 0$, we choose $\sigma^2_{\text{prior}} = \alpha^{-1}$ so that the prior coincides with the stationary distribution of the forward process, and when $\alpha = 0$ we simply set $\sigma^2_{\text{prior}} = 2T$. With this, we arrive at the SDE,

$$d\hat{Y}_t = \alpha \hat{Y}_t dt + 2s(\hat{Y}_t, T-t)dt + \sqrt{2}dW_t, \quad \hat{Y}_0 \sim p_{\text{prior}}. \tag{5}$$

Thus, a sample is generated by sampling from $\hat{Y}_T$, or more commonly, the process is terminated early, sampling from $\hat{Y}_{T-\epsilon}$ for some small $\epsilon > 0$. Therefore, diffusion models are density estimation algorithms formed from the composition $A_{\text{samp}} \circ A_{\text{sm}}$.

Denoising score matching and overfitting. As stated in the introduction, computing ℓ_{sm} requires access to the population score function, $\nabla \log p_t$. So instead, the *(population) denoising score matching loss* is used in its place:

$$\ell_{\text{dsm}}(s; \tau) := \mathbb{E}_{X_0 \sim \nu}\left[\int \mathbb{E}_{X_t|X_0}[\|s(X_t, t) - \nabla \log p_{t|0}(X_t|X_0)\|^2 | X_0] \, \tau(dt) \right], \tag{6}$$

which differs from $\ell_{\text{sm}}(s)$ only by a constant C_{sm}, (see Lemma 16) whilst being easier to approximate without access to $\nabla \log p_t$ (Hyvärinen, 2005). Since $p_{t|0}$ is a Gaussian kernel, its score is given by,

$$\nabla_y \log p_{t|0}(y|x) = \frac{\mu_t x - y}{\sigma_t^2}, \qquad \mu_t = e^{-\alpha t}, \qquad \sigma_t^2 = \alpha^{-1}(1 - \mu_t^2). \tag{7}$$

In practice, the objective in (6) is further approximated via Monte Carlo estimation using the dataset which leads to the *empirical denoising score matching loss*,

$$\hat{\ell}_{\text{dsm}}(s; S, \tau) := \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \displaystyle\int \mathbb{E}_{X_t|X_0}[\|s(X_t, t) - \nabla \log p_{t|0}(X_t|x_i)\|^2 | X_0 = x_i] \, \tau(dt). \tag{8}$$

In the following lemma, we highlight the important property that this can equivalently be defined as the denoising score matching objective for the process $\hat{X}_t$ which evolves as in (1) but with the initial distribution given by the empirical distribution, $\hat{X}_0 \sim \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}(dx)$.

Lemma 1. *The objective $\hat{\ell}_{\text{dsm}}(s; S, \tau)$ is identical, up to a constant, to the objective*

$$\hat{\ell}_{\text{sm}}(s; S, \tau) := \int \mathbb{E}[\|s(\hat{X}_t, t) - \nabla \log \hat{p}_t(\hat{X}_t)\|^2 | S] \tau(dt), \tag{9}$$

where $\hat{p}_t$ is the marginal density of $\hat{X}_t$. Therefore, any minimiser of $\hat{\ell}_{\text{dsm}}(\cdot; S, \tau)$ on $L^0(\mathbb{R}^d \times [0,T]; \mathbb{R}^d)$ is identical to $\nabla \log \hat{p}_t$ a.e. for any $t \in \text{supp}(\tau)$.

See Appendix A.2 for the proof. This lemma shows that, unlike in traditional supervised learning problems, the empirical objective here admits a single unique minimiser, the *empirical score function*, $\nabla \log \hat{p}_t$. The nature of this score function and the samples it generates has been the focus of several recent studies, notably (Pidstrigach, 2022) which shows that with perfect sampling, any score function sufficiently close to $\nabla \log \hat{p}_t$ recovers the training data.

Other notation. When the score matching algorithm A_{sm} is random, we use $A_{\text{sm}}(S)$ as shorthand for the random score function $(x, t, \omega) \mapsto A_{\text{sm}}(S, \omega)(x, t)$. Given two random score functions s, s', we let $\Gamma(s, s')$ denote the set of all couplings of these random functions (Appendix A.1 for details).

3 Score stability and generalisation

Algorithmic stability is a classical technique in learning theory used to understand the generalisation properties of a variety of important learning algorithms (Kearns & Ron, 1999; Devroye & Wagner,

1979; Bousquet & Elisseeff, 2002; Hardt et al., 2016). While there are various formulations, they all share the common aim of connecting properties of a learning algorithm to its robustness under changes in the dataset. Its use has primarily been focused around regression and classification problems—in this section, we propose a notion of stability that applies specifically to diffusion models.

We introduce the notion of *score stability* which quantifies how sensitive a score matching algorithm A_{sm} is to individual changes in the dataset. We do this by defining the adjacent dataset $S^i := \{x_1, ..., x_{i-1}, \tilde{x}, x_{i+1}, ..., x_N\}$ where $\tilde{x} \sim \nu_{\mathrm{data}}$, independent from S, and then measuring the similarity between the score functions $\hat{s} = A_{\mathrm{sm}}(S)$ and $\hat{s}^i = A_{\mathrm{sm}}(S^i)$.

Definition 2. *A score matching algorithm A_{sm} is score stable with constant $\varepsilon_{stab} > 0$ if for any $i \in [N]$ it holds that,*

$$\mathbb{E}_{S,\tilde{x}}\left[\inf_{(\hat{s},\hat{s}^i)\in\Gamma_i} \int \mathbb{E}[\|\hat{s}(X_t,t) - \hat{s}^i(X_t,t)\|^2 | X_0 = \tilde{x}, S, \tilde{x}]\, \tau(dt)\right] \leq \varepsilon_{stab}^2,$$

where $\Gamma_i = \Gamma(A_{\mathrm{sm}}(S), A_{\mathrm{sm}}(S^i))$.

Since A_{sm} may be random, we define score stability in terms of the best-case coupling of the random score functions $\hat{s}$, $\hat{s}^i$. We recall that $\Gamma(\cdot,\cdot)$ denotes the set of couplings between two random score functions, and when it is not random, it is given by the singleton $\Gamma_i = \{(A_{\mathrm{sm}}(S), A_{\mathrm{sm}}(S^i))\}$. In the following theorem, we connect score stability to generalisation by controlling the expected generalisation gap by the score stability constant.

Theorem 3. *Suppose that the score matching algorithm A_{sm} is score stable with constant ε_{stab}. Then, with $\hat{s} = A_{\mathrm{sm}}(S)$, it holds that*

$$\left|\mathbb{E}\big[\ell_{\mathrm{dsm}}(\hat{s};\tau)\big]^{1/2} - \mathbb{E}\big[\hat{\ell}_{\mathrm{dsm}}(\hat{s};S,\tau)\big]^{1/2}\right| \leq \varepsilon_{stab}. \tag{10}$$

Furthermore, it holds that

$$\mathbb{E}\big[\ell_{\mathrm{sm}}(\hat{s};\tau)\big] - \mathbb{E}\big[\hat{\ell}_{\mathrm{sm}}(\hat{s};S,\tau)\big] \leq 2\,\varepsilon_{stab}\,\mathbb{E}\big[\hat{\ell}_{\mathrm{dsm}}(\hat{s};S,\tau)\big]^{1/2} + \varepsilon_{stab}^2. \tag{11}$$

With Theorem 3, we obtain that the generalisation gap for both the denoising score matching loss and the score matching loss decays at the same rate as score stability. We can further simplify the bound for the score matching loss using the fact that $\hat{\ell}_{\mathrm{dsm}}$ and $\hat{\ell}_{\mathrm{sm}}$ are identical up to a constant, to obtain,

$$\mathbb{E}\big[\ell_{\mathrm{sm}}(\hat{s};\tau)\big] \lesssim \mathbb{E}\big[\hat{\ell}_{\mathrm{sm}}(\hat{s};\tau)\big] + \varepsilon_{\mathrm{stab}}\, C_{\mathrm{sm}}^{1/2} + \varepsilon_{\mathrm{stab}}^2.$$

One should expect that if the score matching algorithm is effective, both $\hat{s}$ and $\hat{s}^i$ converge to the ground truth as N grows, and thus $\varepsilon_{\mathrm{stab}}$ should decrease to 0. Ascertaining the rate at which N decreases requires an analysis of the algorithm at hand, hence the categorisation of algorithmic stability as an algorithm-dependent approach. This contrasts with uniform learning, which utilises control over the hypothesis class, providing a worst-case bound that is independent from the algorithm.

In the following sections, we apply the framework of score stability to some common learning settings for diffusion models. We derive estimates of the score stability constant for these algorithms and identify features that promote generalisation.

4 Empirical score matching and implicit regularisation

We begin our examples by considering the score matching algorithm that minimises the empirical denoising score matching loss. Given a hypothesis class $\mathcal{H} \subseteq L^0(\mathbb{R}^d \times [0,T]; \mathbb{R}^d)$, we define this algorithm by,

$$A_{\mathrm{erm}}(S) = \mathrm{argmin}_{s\in\mathcal{H}} \hat{\ell}_{\mathrm{dsm}}(s;S,\tau).$$

While this algorithm is not often used in practice, it is the natural analogue to empirical risk minimisation from traditional supervised learning and thus serves as a canonical example. We consider the setting of the manifold hypothesis where the data distribution is supported on a submanifold of $\mathbb{R}^d$.

Assumption 4. *Suppose that ν_{data} is supported on a smooth submanifold of $\mathbb{R}^d$ that has dimension d^* and reach $\tau_{reach} > 0$. Furthermore, its density on the submanifold, p_ν, satisfies $c_\nu := \inf p_\nu > 0$.*

The reach describes the maximum distance where the projection to the manifold is uniquely defined and therefore, it quantifies the maximum curvature of the manifold. We refer to Appendix A.3 for the full definition. Several recent works have considered the assumption that ν_{data} lies on a submanifold of $\mathbb{R}^d$. These works argue that d^* can often be far smaller than d and so dependence with respect to d^* over d is favourable (De Bortoli, 2022; Pidstrigach, 2022; Loaiza-Ganem et al., 2024; Potaptchik et al., 2024; Huang et al., 2024). The assumption that the density is bounded from below has also appeared in several of these works (Potaptchik et al., 2024; Huang et al., 2024). We also make the following assumption about the class of score networks.

Assumption 5. *Suppose there exists $D_{\mathcal{H}} \geq 0$ such that for any $s, s' \in \mathcal{H}$, it holds that*

$$\|s(\cdot,t) - s'(\cdot,t)\|_{L^\infty} \leq D_{\mathcal{H}}/\sigma_t^2, \qquad \textit{for all } t \in \mathrm{supp}(\tau).$$

Under these assumptions, we obtain the following estimate for the stability constant.

Proposition 6. *Suppose that assumptions 4 and 5 hold and that $\epsilon := \inf \mathrm{supp}(\tau) \in (0, \tau^2_{reach})$, then for any $c \in (0,1)$ and sufficiently large N, the score matching algorithm A_{erm} is score stable with,*

$$\varepsilon^2_{stab} \lesssim C\big(CC_{\mathrm{sm}}N^{-2} + \mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})]\big)^c, \qquad C = \tfrac{D^2_{\mathcal{H}}}{\sigma^4_\epsilon} \vee \tfrac{1}{c_\nu \sigma^{d^*}_\epsilon}.$$

An interesting feature of Proposition 6 is that generalisation bounds under only basic assumptions about the structure of the hypothesis class *without any additional regularisation*. This contrasts with algorithmic stability in the setting of traditional supervised learning, where empirical risk minimisation is stable only when restricting the hypothesis class or with the use of explicit regularisation (Zhang et al., 2021; Bousquet & Elisseeff, 2002). Here, we show that the denoising score matching loss possesses the unique property that it is stable without the need for additional regularisation, suggesting that the denoising score matching loss possesses a form of *implicit regularisation.*

When $d^* > 4$, for ϵ sufficiently small, we have that $C = \mathcal{O}(c_\nu^{-1}\epsilon^{-d^*/2}), C_{\mathrm{sm}} = \mathcal{O}(d^*\epsilon^{-1})$. Since the bound only depends on d^* and not d, this suggests that diffusion models are automatically manifold-adaptive. The bound also heavily depends on ϵ, with it being smaller for larger ϵ and growing exponentially fast as ϵ approaches zero, indicating that the natural regularisation present in the score matching objective is more prevalent at larger noise scales. The requirement to have $\epsilon > 0$ is closely related to the technique of early stopping which is frequently used in the diffusion model literature (Song & Kingma, 2021; Karras et al., 2022). This is where the backwards process $\hat{Y}_t$ is terminated early by some small amount of time to avoid irregularity issues of the score function when close to convergence. Other theoretical works have also identified the importance of early stopping in the generalisation properties of diffusion models (Oko et al., 2023; Azangulov et al., 2024).

Proof summary We now provide a brief summary of the proof of Proposition 6. The first step of the proof technique utilises a fundamental property of the empirical denoising score matching objective, $\hat{\ell}_{\mathrm{dsm}}(s; S, \tau)$: that it is strongly convex in s in a data-dependent weighted L^2-space. Strong convexity is often used in algorithmic stability analyses, especially in deriving stability bounds for linear models—here we borrow a similar approach, but we analyse the stability of the algorithm in function space. With this, we arrive at the following inequality (see Lemma 19):

$$\int \mathbb{E}[\|\hat{s}(\hat{X}_t, t) - \hat{s}^i(\hat{X}_t, t)\|^2]\tau(dt) \lesssim \mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})] + \tfrac{\varepsilon_{\mathrm{stab}}}{N}(C^{1/2}_{\mathrm{sm}} + \varepsilon_{\mathrm{stab}}), \tag{12}$$

where $\varepsilon_{\mathrm{stab}}$ is the (yet-to-be bounded) score stability constant of A_{erm}.

The second step of the proof technique utilises a characteristic property of the heat kernel—that it smooths out functions. In particular, we utilise the celebrated Harnack inequality of Wang (1997) that captures this property by showing that for any positive measurable $\phi : \mathbb{R}^d \to \mathbb{R}_+$, $x, y \in \mathbb{R}^d$, it holds that

$$\mathbb{E}[\phi(X_t)|X_0 = x] \leq \mathbb{E}[\phi(X_t)^p|X_0 = y]^{1/p} \exp\big(\tfrac{\mu_t^2\|x-y\|^2}{2(p-1)\sigma_t^2}\big),$$

for any $t > 0, p > 1$. Utilising this bound, we convert the upper bound in (12) to a bound on the stability constant. The full proof can be found in Appendix C.

5 Stochastic sampling and score stability

In practice, the backwards process in (5) cannot be sampled exactly, so we instead rely on approximations based on numerical integration schemes. In this section, we investigate how algorithmic

stability interacts with these sampling schemes. We consider the Euler-Maruyama-type sampling scheme proposed in (Benton et al., 2024; Potaptchik et al., 2024) which discretises at the timesteps $(t_k)_{k=0}^K$, where $t_k = T - (1+\kappa)^{\frac{T-1}{\kappa}-k}$ for large k. The quantity $\kappa > 0$ is chosen freely and we choose $K \sim \log(\epsilon^{-1})/\log(1+\kappa)$ so that $t_K \approx \epsilon$ (see Appendix D for details). By sampling its terminating iterate $\hat{y}_K$, we obtain a *sampling algorithm*, A_{em}, that maps a score function s to the distribution $\mathrm{law}(\hat{y}_K)$, which approximates the distribution $\mathrm{law}(\hat{Y}_\epsilon)$.

In the previous section, we identified that early stopping of the backwards process benefits generalisation. In the present section, we will consider how coarseness of the discretisation scheme produces similar benefits. It is often the case that the score function is trained only at those time steps considered by the sampler, i.e. using the time-weighting, $\hat{\tau}_\kappa(dt) = \frac{1}{K}\sum_{k=0}^{K-1} \delta_{T-t_k}(dt)$ (Ho et al., 2020). As a result, the effective stopping time of the algorithm can be much larger than the early stopping time, ϵ. In the following proposition, we demonstrate how this benefits generalisation.

Proposition 7. *Consider the setting of Proposition 6 with $\alpha = 1$ and set $\tau = \hat{\tau}_\kappa$, then for sufficiently large N, $\kappa \le \epsilon^{-1}/4$ and any $c \in (0,1)$, we have that for $q_K = A_{\mathrm{em}} \circ A_{\mathrm{erm}}(S)$,*

$$\mathbb{E}[D(p_\epsilon \| q_K)] \lesssim \mathbb{E}[\hat{\ell}^\star_{\mathrm{sm},\kappa}] + B_\kappa^{\frac{1}{2}}(1+\kappa)^{-d^*} + \tfrac{B_\kappa}{C_{\mathrm{sm}}}(1+\kappa)^{-2d^*} + \kappa(1+\kappa)d^*\log(\epsilon^{-1})^2 + de^{-2T},$$

where $B_\kappa = \frac{C_{\mathrm{sm}}}{c_\nu}(\frac{C_{\mathrm{sm}}}{c_\nu}N^{-2} + \mathbb{E}[\hat{\ell}^\star_{\mathrm{sm},\kappa}])^c \epsilon^{-d^*}$, $\hat{\ell}^\star_{\mathrm{sm},\kappa} := \inf_{\mathcal{H}} \hat{\ell}_{\mathrm{sm}}(h; S, \hat{\tau}_\kappa)$.

The second and third terms of the bound in Proposition 7 are due to the score stability of the ERM algorithm and decay as κ increases. The fourth term of the bound captures the discretisation error and therefore increases with κ. What this result captures is that there is a trade-off between sampler accuracy and generalisation that is managed by the discretisation of the diffusion model. In the following corollary, this trade-off is optimised.

Corollary 8. *Consider the setting of Proposition 7, then for any $c \in (0,1)$ and sufficiently small ϵ, there exists $\kappa > 0$ such that with $q_K = A_{\mathrm{em}} \circ A_{\mathrm{erm}}(S)$*

$$\mathbb{E}[D(p_\epsilon \| q_K)] \lesssim \begin{cases} B_\kappa^{\frac{1}{2}} + C_{\mathrm{sm}}^{-1} B_\kappa, & \text{if } B_\kappa \le \log(\epsilon^{-1})^2, \\ \log(\epsilon^{-1}) B_\kappa^{\frac{1}{2(d^*+1)}} + (C_{\mathrm{sm}}^{-1} + d^*)\log(\epsilon^{-1})^2 B_\kappa^{\frac{1}{d^*+1}} + de^{-2T}, & \text{otherwise.} \end{cases}$$

The primary strength of this result over (Oko et al., 2023; Azangulov et al., 2024) is that we assume little about the hypothesis class. Their results require carefully constrained network architectures and a specific early stopping time to control complexity. In contrast, our result holds for any sufficiently small early stopping time, relying instead on a carefully chosen discretisation scheme, which is usually tuned in practice (Karras et al., 2022; Williams et al., 2024). The main drawback is that our general approach does not exploit the model class to adapt to smoothness properties of the underlying measure, which we leave for future work.

6 Stochastic optimisation and implicit regularisation

To learn the score function, it is common to choose it from a parametric hypothesis class $\{s_\theta : \theta \in \mathbb{R}^n\}$ (e.g. a deep neural network) by minimising $\hat{\ell}_{\mathrm{dsm}}$ via stochastic optimisation (Karras et al., 2024). In this section, we consider the score stability of this setting, focusing on stochastic gradient descent (SGD) with gradient clipping and weight decay. We consider the standard gradient estimator: given the mini-batch $(x'_i)_{i=1}^{N_B}$ of size $N_B \ll N$ we define the random estimator,

$$G(\theta, (x'_i)_{i=1}^{N_B}) = \tfrac{1}{N_B P}\sum_{i=1}^{N_B}\sum_{j=1}^{P} w_{t_{i,j}} \nabla_\theta \| s_\theta(X_{i,j}, t_{i,j}) - \nabla \log p_{t_{i,j}|0}(X_{i,j}|x)\|^2, \tag{13}$$

where we define the random variables $X_{i,j} = \mu_{t_{i,j}} x'_i + \sigma_{t_{i,j}}\xi_{i,j}$, $t_{i,j} \sim w_t^{-1}\tau(dt)$, $\xi_{i,j} \sim N(0, I_d)$. The additional variance introduced by the random variables $\xi_{i,j}$ and $t_{i,j}$ leads to a gradient estimator with significantly higher variance than in standard supervised learning. This presents several challenges during training, and various strategies have been proposed to mitigate this issue (Karras et al., 2024; Song & Kingma, 2021). For example, the weighting function $w : [0,T] \to \mathbb{R}_+$ can be tuned to reduce variance (Karras et al., 2022) or the number of resamples $P \in \mathbb{N}$ can be increased. We consider the following iterative scheme, defined for a given weight decay constant $\lambda > 0$ and clipping value $C > 0$:

$$\theta_{k+1} = (1 - \eta_k\lambda)\theta_k - \eta_k \,\mathrm{Clip}_C(G_k(\theta_k, (x_i)_{i \in B_k})), \tag{14}$$

where $\eta_k > 0$ and $B_k \subset [N]$ are the learning rates and mini-batch indices for each iteration $k \in \mathbb{N}$ and we define the clipping operator $\mathrm{Clip}_C(v) = (1 \wedge (C\|v\|^{-1}))v$. Both gradient clipping and weight decay are widely used in diffusion model training and are typically motivated by their stabilising effect on optimisation, minimising the impact of the high variance of the gradient estimator (Song et al., 2021; Ho et al., 2020). Throughout this section, we take the mini-batch B_k to be i.i.d. and uniformly sampled from $[N]$ without replacement. For the sake of simplicity, we suppose that the iterative scheme is terminated after $K \in \mathbb{N}$ iterations, where K is fixed and independent of the data.

6.1 Stability of SGD with weight decay and clipping

In our analysis, we avoid restricting the score network to a specific parametric class and instead make structural assumptions based on its smoothness properties. We recall that a function is Lipschitz with constant $L \geq 0$ if it is differentiable and its directional derivatives are uniformly bounded by L.

Assumption 9 (Smoothness of the score network)**.** *There exists $L : \mathbb{R}^d \times (0,T] \to \mathbb{R}_+$ and $M : \mathbb{R}^d \times (0,T] \to \mathbb{R}_+$ such that for almost all $x \in \mathbb{R}^d, t \in (0,T]$, $s_\theta(x,t)$ is Lipschitz and smooth (gradient-Lipschitz) in $\theta \in \mathbb{R}^n$ with constants $L(x,t)$ and $M(x,t)$, respectively. Furthermore, there exists constants $\overline{L}, \overline{M} \geq 0$ such that for any $x \in \mathrm{supp}(\nu_{\mathrm{data}})$,*

$$\int \mathbb{E}[L(X_t,t)^2 | X_0 = x]\, \tau(dt) \leq \overline{L}^2, \quad \int \mathbb{E}[M(X_t,t)^2 | X_0 = x]\, \tau(dt) \leq \overline{M}^2.$$

The use of Lipschitz and smoothness assumptions is commonplace in the analysis of optimisation schemes (Nesterov, 2018; Hardt et al., 2016). However, the assumption differs slightly from the usual in that we only require these properties to hold *almost everywhere* with respect to the input distribution and we allow the Lipschitz and smoothness constants to vary with the input, provided their square averages remain bounded. This relaxation enables us to accommodate common models that would otherwise violate global smoothness assumptions, such as ReLU networks.

Assumption 10. *Suppose there exists $B_\ell > 0$ such that for any $\theta \in \mathbb{R}^n$, it holds that*

$$\hat{\ell}_{\mathrm{dsm}}(s_\theta; \{x\}, \delta_t) \leq B_\ell^2 / \sigma_t^4, \qquad \text{for each } x \in \mathrm{supp}(\nu_{\mathrm{data}}), t \in \mathrm{supp}(\tau). \tag{15}$$

This property requires that the supported score functions are made of denoising functions that are concentrated on a compact set. To highlight that this can be achieved quite easily, we note that with the naive estimate $s(x,t) = -x/\sigma_t^2$, (15) is satisfied with $B_\ell^2 = \mathbb{E}[\|X_0\|^2]$.

In the following proposition we demonstrate score stability bounds in the case that the step size is decaying with a rate of $1/k$.

Proposition 11. *Consider the score matching algorithm $A_{\mathrm{sm}} : S \mapsto s_{\theta_K}$ for some fixed $K \in \mathbb{N}$ where $(\theta_k)_k$ is as given in* (14)*. Suppose that assumptions 9 and 10 hold and $\eta_k \leq \bar{\eta}/k$ for all $k < K$, for some $\bar{\eta} \in (0, \lambda^{-1})$. Then, we obtain that A_{sm} is score stable with constant,*

$$\varepsilon_{stab}^2 \lesssim \left(\frac{C}{\lambda} \vee R\right)^{1+\frac{\bar{\eta}\upsilon}{\bar{\eta}\upsilon+1}} \frac{\overline{L}^2}{(\bar{\eta}\upsilon) \vee 1} \left(\frac{C}{\bar{\eta}}\right)^{\frac{1}{\bar{\eta}\upsilon+1}} \frac{N_B K^{\frac{\bar{\eta}\upsilon}{\bar{\eta}\upsilon+1}}}{N},$$

where $R^2 = \mathbb{E}[\|\theta_0\|^2]$, $\upsilon = (\overline{M} B_\ell C_\tau^{1/2} + \overline{L}^2 - \lambda) \vee 0$ and $C_\tau = \int \sigma_t^{-4} \tau(dt)$.

Since the score matching algorithm is random, to control the stability constant we construct a coupling of the random score functions $A_{\mathrm{sm}}(S)$ and $A_{\mathrm{sm}}(S^i)$ through a coupling of the optimisation trajectories associated with training on S versus S^i.

6.2 Utilising noise in the gradient estimator

The primary drawback of Proposition 11 is that the bound grows with the number of iterations. This is particularly problematic since diffusion models often require numerous steps due to the high-variance gradient estimator. In this section, we improve this dependence by explicitly leveraging the noise in the gradient estimator. The idea that stochasticity in optimisation can act as a form of implicit regularisation has motivated the development of numerous learning algorithms and theoretical works in recent years (Srivastava et al., 2014; Bishop, 1995; Mou et al., 2018; Pensia et al., 2018). Here, we investigate how the noise intrinsic to the gradient estimator for $\hat{\ell}_{\mathrm{dsm}}$ can play a similar role in promoting generalisation in diffusion models.

To incorporate the effects of the gradient noise, we consider a simplified model in which the noise from the stochastic gradient estimator is approximated with a second-order Gaussian approximation:

$$\theta_{k+1} = (1-\eta\lambda)\theta_k - \eta\mathbb{E}\big[\mathrm{Clip}_C(G_k)\big|\theta_k, B_k, S\big] + \eta\,\mathrm{Cov}\left(\mathrm{Clip}_C(G_k)\big|\theta_k, B_k, S\right)^{1/2}\xi_k, \quad (16)$$

where $\xi_k \in \mathbb{R}^d$ is a standard Gaussian and we use $G_k := G_k(\theta_k, B_k)$. This approximation can be justified by observing that the inner summation in (13) is over conditionally i.i.d. variables, once conditioned on θ, B and S. Therefore, the gradient estimator G becomes approximately Gaussian as P grows large. For this analysis, we assume the following lower bound on the gradient noise.

Assumption 12. *There exists a positive semi-definite matrix $\overline{\Sigma} \in \mathbb{R}^{n\times n}$ such that for any $x \in \mathrm{supp}(\nu)$ and $\theta \in \mathbb{R}^n$,*

$$\mathrm{Cov}_{t\sim\tau, X_t|X_0}\left(\mathrm{Clip}_C(\nabla_\theta\|s_\theta(X_t,t)-x\|^2)\big|X_0 = x\right) \succcurlyeq \overline{\Sigma}.$$

Furthermore, the eigenvalues of $\overline{\Sigma}$, $(\lambda_i)_{i=1}^n$, possess the spectral gap $\lambda_{gap} := \min_{\lambda_i\neq 0}\lambda_i > 0$.

We use the matrix $\overline{\Sigma}$ to dictate the geometry on which we perform our analysis. In particular, we consider the weighted norm $\|v\|_{\overline{\Sigma}^+} := v^T\overline{\Sigma}^+ v$ where $\overline{\Sigma}^+$ is the pseudoinverse matrix.

Assumption 13. *For almost all $x \in \mathbb{R}^d, t \in (0,T]$, $s_\theta(x,t)$ is Lipschitz and smooth (gradient-Lipschitz) in $\theta \in \mathbb{R}^n$ with respect to the seminorm $\|\cdot\|_{\overline{\Sigma}^+}$ and with constants $L(x,t)$ and $M(x,t)$, respectively. Furthermore, there exists constants $\overline{L}, \overline{M} \geq 0$ such that for any $x \in \mathrm{supp}(\nu_{\mathrm{data}})$,*

$$\int\mathbb{E}[L(X_t,t)^4|X_0 = x]\,\tau(dt) \leq \overline{L}^4, \quad \int\mathbb{E}[M(X_t,t)^4|X_0 = x]\,\tau(dt) \leq \overline{M}^4.$$

By requiring that the Lipschitz and smoothness properties hold with respect to $\|\cdot\|_{\overline{\Sigma}^+}$, we effectively require that the gradient estimator adds noise in all directions aside from those that do not change the function (e.g. along symmetries in the parameter space). With this, we arrive at our time-convergent score stability bound for SGD.

Proposition 14. *Consider the score matching algorithm $A_{\mathrm{sm}} : S \mapsto s_{\theta_K}$ for some fixed $K \in \mathbb{N}$ where $(\theta_k)_k$ is as given in (16). Suppose that assumptions 10, 12 and 13 hold, then there exists some $\bar{\eta} > 0$ such that, if $\sup_p \eta_p \leq \bar{\eta}$, we obtain that A_{sm} is score stable with constant*

$$\varepsilon^2_{stab} \lesssim \frac{\overline{L}^2C^2(P+n)}{\lambda_{gap}N}\min\left\{\frac{\eta_{\min}\lambda_{gap}\lambda^2}{PN_BC}\sum_{k=0}^{K-1}\eta_k, \exp\left(\tilde{c}\frac{PN_BC}{\eta_{\min}\lambda_{gap}\lambda^2}\right)\right\},$$

where $\tilde{c} \lesssim (\overline{M}_4B_\ell C_\tau^{1/2} + \overline{L}_4^2)(PN_B\lambda_{gap})^{-1/2} \vee 1$, $\eta_{\min} = \min_k \eta_k$.

In this bound, we recover the $\frac{1}{\sqrt{N}}$ score stability bounds from Proposition 11 while also introducing the property that the bound does not grow endlessly with the number of iterations. This property is obtained using the noise in the gradient estimator and is not possible without additional noise. Through this analysis, we identify the generalisation benefit of a property unique to diffusion models and how they interact with SGD.

7 Conclusion and Future Work

In this paper, we propose a general algorithm-dependent framework for analysing the generalisation capabilities of diffusion models. We introduce *score stability*, which quantifies an algorithm's sensitivity to the dataset, and use it to derive expected generalisation gap bounds. Applying this framework to several common algorithms, we derive closed-form bounds and identify several previously overlooked sources of implicit regularisation in diffusion models. First, our analysis of empirical risk minimisation finds that the denoising score matching objective provides inherent stability guarantees without further regularisation (denoising regularisation). We then analyse how score stability interacts with discrete-time samplers, identifying that coarse discretisation can improve generalisation guarantees (sampler regularisation). Finally, we consider stochastic optimisation schemes for score matching, obtaining stability guarantees (optimisation regularisation).

This work opens several avenues for future research. Key directions include developing high-probability bounds, developing bounds on privacy and memorisation, tightening our analysis by incorporating data or model properties, like smoothness, and extending the framework to compare different sampling algorithms, such as the probability flow ODE.

REFERENCES

Eddie Aamari. *Convergence Rates for Geometric Inference*. PhD thesis, Université Paris-Saclay, September 2017.

Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. *Electronic Journal of Statistics*, 13(1):1359–1399, 2019.

Amit Attia and Tomer Koren. Uniform stability for first-order empirical risk minimization. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 3313–3332. PMLR, 2022.

Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv:2409.18804*, 2024.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer International Publishing, 2014.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *International Conference on Learning Representations*, 2024.

Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural Computation*, 7 (1):108–116, 1995.

Olivier Bousquet and Andre Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 745–754. PMLR, 2018.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023.

Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general Gaussian mixtures with efficient score matching. *arXiv preprint arXiv:2404.18893*, 2024.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, 2021.

L Devroye and T Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, September 1979.

Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability Theory and Related Fields*, 166(3):851–886, December 2016.

Andreas Eberle and Mateusz B Majka. Quantitative contraction rates for Markov chains on general state spaces. *Electronic Journal of Probability*, 24:1–36, January 2019.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 1225–1234, 2016.

U G Haussmann and E Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(4): 1188–1205, 1986.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Zhihan Huang, Yuting Wei, and Yuxin Chen. Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *arXiv [cs.LG]*, October 2024.

Aapo Hyvärinen. Estimation of Non-Normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In S. Koyejo and S. Mohamed and A. Agarwal and D. Belgrave and K. Cho and A. Oh (ed.), *Advances in Neural Information Processing Systems*, 2022.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24174–24184, June 2024.

M Kearns and D Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.

Puheng Li, Zhong Li, Huishuai Zhang, and Jiang Bian. On the generalization properties of diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Gabriel Loaiza-Ganem, Brendan Leigh Ross, Rasa Hosseinzadeh, Anthony L Caterini, and Jesse C Cresswell. Deep generative models through the lens of the manifold hypothesis: A survey and new connections. *arXiv [cs.LG]*, April 2024.

Mateusz B Majka, Aleksandar Mijatovic, and Lukasz Szpruch. Nonasymptotic bounds for sampling algorithms without log-concavity. *Ann. Appl. Probab.*, 30(4):1534–1581, August 2020.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In Bubeck, Sébastien and Perchet, Vianney and Rigollet, Philippe (ed.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 605–638. PMLR, 2018.

Yurii Nesterov. *Lectures on convex optimization*. Springer optimization and its applications. Springer International Publishing, Cham, Switzerland, 2 edition, December 2018.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, 2023.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 546–550, June 2018.

Jakiw Pidstrigach. Score-based generative models detect manifolds. In *Advances in Neural Information Processing Systems*, 2022.

Peter Potaptchik, Iskander Azangulov, and George Deligiannidis. Linear convergence of diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2410.09046*, 2024.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022.

Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of Gaussians using the DDPM objective. In *Advances in Neural Information Processing Systems*, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101. 03288*, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based generative modeling through stochastic differential equations. In *9International Conference on Learning Representations*, 2021.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Rong Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Simon Vary, David Martínez-Rubio, and Patrick Rebeschini. Black-box uniform stability for non-euclidean empirical risk minimization. *arXiv [cs.LG]*, December 2024.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Feng-Yu Wang. Logarithmic sobolev inequalities on noncompact riemannian manifolds. *Probability Theory and Related Fields*, 109(3):417–424, November 1997.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Christopher Williams, Andrew Campbell, Arnaud Doucet, and Saifuddin Syed. Score-optimal diffusion schedules. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Hongkang Yang. A mathematical framework for learning probability distributions. *arXiv preprint arXiv:2212.11481*, 2022.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Google Brain, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

A FURTHER BACKGROUND

We begin with some further details on notation and lemmas used throughout this work and provide proofs for the lemmas in Section 2.

A.1 RANDOM SCORE MATCHING ALGORITHMS

We begin with some additional details on how random score matching algorithms are defined in this work. Recalling the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we define the set of random score functions,

$$\mathcal{S} := \Big\{ s : \mathbb{R}^d \times [0,T] \times \Omega : s(\cdot,\cdot,\omega) \in L^0(\mathbb{R}^d \times [0,T]; \mathbb{R}^d) \Big\}.$$

For any random score matching algorithm $A_{\mathrm{sm}} : (\cup_{N=1}^{\infty} (\mathbb{R}^d)^{\otimes N}) \times \Omega \to L^0(\mathbb{R}^d \times [0,T]; \mathbb{R}^d)$, we use $A_{\mathrm{sm}}(S)$ as shorthand for the random score function $(\omega, x, t) \mapsto A_{\mathrm{sm}}(S, \omega)(x, t)$ belonging to $\mathcal{S}$.

Given two random score functions s, s', let $\Gamma(s, s')$ denote the set of all couplings of these functions which we define as,

$$\Gamma(s, s') := \Big\{ (\tilde{s}, \tilde{s}') \in \mathcal{S} \times \mathcal{S} : \tilde{s} \simeq s, \tilde{s}' \simeq s' \Big\},$$

where $\tilde{s} \simeq s$ denotes the fact that for any bounded measurable test function $\phi : L^0(\mathbb{R}^d \times [0,T]; \mathbb{R}^d) \to \mathbb{R}$, it holds that,

$$\int \phi(s(\cdot,\cdot,\omega)) d\mathbb{P} = \int \phi(\tilde{s}(\cdot,\cdot,\omega)) d\mathbb{P}.$$

A.2 PRELIMINARY LEMMAS

For the score matching loss bound, we begin with the fact that the score matching loss is equivalent to the denoising score matching loss up to an added constant Song et al. (2021); Hyvärinen (2005).

Lemma 15. *For any $t > 0$, $y \in \mathbb{R}^d$, we have*

$$\nabla \log p_t(y) = \frac{\mu_t \mathbb{E}[X_0 | X_t = y] - y}{\sigma_t^2}, \qquad \nabla \log \hat{p}_t(y) = \frac{\mu_t \mathbb{E}[\hat{X}_0 | \hat{X}_t = y, S] - y}{\sigma_t^2}. \tag{17}$$

Proof. We begin by showing that the conditional score is an unbiased estimate of $\nabla \log p_t$. For any $x \in \mathbb{R}^d, t > 0$, we have

$$\begin{aligned}
\mathbb{E}[\nabla \log p_{t|0}(X_t | X_0) | X_t = x] &= \int \nabla_x \log p_{t|0}(x|y)\, p_{0|t}(y|x) dy \\
&= \int \nabla \log p_{t|0}(x|y)\, \frac{p_{t|0}(x|y) p_0(y)}{p_t(x)} dy \\
&= \int \nabla p_{t|0}(x|y)\, \frac{p_0(y)}{p_t(x)} dy.
\end{aligned}$$

Therefore, using the exchangeability of gradients and integrals (note that $p_{t|0}$ is C^∞), we arrive at

$$\mathbb{E}[\nabla \log p_{t|0}(X_t | X_0) | X_t = x] = \frac{\nabla p_t(x)}{p_t(x)} \tag{18}$$

$$= \nabla \log p_t(x). \tag{19}$$

Alternatively, using (7), we obtain that the left-hand side takes the form,

$$\mathbb{E}[\nabla \log p_{t|0}(X_t | X_0) | X_t = x] = \frac{\mu_t \mathbb{E}[X_0 | X_t = x] - x}{\sigma_t^2},$$

completing the proof of the first equality in (17). For the second equality, concerning that empirical score function, the proof follows similarly once the empirical measure $\frac{1}{N} \sum_{i=1}^N \delta_{x_i}$ is considered in place of ν_{data}. □

Lemma 16. *For any integrable score function s, it holds that*

$$\ell_{\mathrm{dsm}}(s;\tau) = \ell_{\mathrm{sm}}(s;\tau) + C_{\mathrm{sm}},$$

where, given $s^\star(x,t) := \nabla \log p_t(x)$, we define

$$C_{\mathrm{sm}} := \int \frac{\mu_t^2}{\sigma_t^4}\mathbb{E}[\mathrm{Tr}\,\mathrm{Cov}(X_0|X_t)]\tau(dt) = \ell_{\mathrm{dsm}}(s^\star;\tau). \tag{20}$$

Proof. Let s be any score function. Using the equality in (19), we obtain the following bias-variance decomposition of $\ell_{\mathrm{dsm}}(s;\tau)$:

$$\begin{aligned}
&\ell_{\mathrm{dsm}}(s;\tau)\\
&= \int \mathbb{E}\Big[\|s(X_t,t) - \nabla \log p_{t|0}(X_t|X_0)\|^2\Big]\tau(dt)\\
&= \int \mathbb{E}\Big[\|s(X_t,t) - \nabla \log p_t(X_t)\|^2\Big]\tau(dt) + \int \mathbb{E}\Big[\|\nabla \log p_{t|0}(X_t|X_0) - \nabla \log p_t(X_t)\|^2\Big]\tau(dt)\\
&= \ell_{\mathrm{sm}}(s;\tau) + \int \mathbb{E}\Big[\,\mathrm{Tr}\,\mathrm{Cov}\left(\nabla \log p_{t|0}(X_t|X_0)\Big|X_t\right)\Big]\tau(dt).
\end{aligned}$$

Once we note that,

$$\begin{aligned}
\mathrm{Tr}\,\mathrm{Cov}\left(\nabla \log p_{t|0}(X_t|X_0)\Big|X_t\right) &= \mathrm{Tr}\,\mathrm{Cov}\left(\frac{\mu_t X_0 - x}{\sigma_t^2}\Big|X_t\right)\\
&= \frac{\mu_t^2}{\sigma_t^4}\,\mathrm{Tr}\,\mathrm{Cov}(X_0|X_t),
\end{aligned}$$

we obtain the bound $\ell_{\mathrm{dsm}}(s;\tau) = \ell_{\mathrm{sm}}(s;\tau) + C_{\mathrm{sm}}$ from the statement. To derive the equality $C_{\mathrm{sm}} = \ell_{\mathrm{dsm}}(s^\star;\tau)$, we use that $\ell_{\mathrm{sm}}(s^\star;\tau) = 0$ and so we obtain $\ell_{\mathrm{dsm}}(s^\star;\tau) = 0 + C_{\mathrm{sm}}$. □

Similarly, there is an equivalence between the empirical forms of the denoising score matching loss and the score matching loss,

$$\hat{\ell}_{\mathrm{dsm}}(s;S,\tau) = \hat{\ell}_{\mathrm{sm}}(s;S,\tau) + \hat{C}_{\mathrm{sm}}, \tag{21}$$

where

$$\hat{C}_{\mathrm{sm}} := \int \frac{\mu_t^2}{\sigma_t^4}\mathbb{E}[\mathrm{Tr}\,\mathrm{Cov}(\hat{X}_0|\hat{X}_t,S)|S]\tau(dt) = \hat{\ell}_{\mathrm{dsm}}(\hat{s}^\star;S,\tau), \tag{22}$$

and $\hat{s}^\star(x,t) = \nabla \hat{p}_t(x)$. This follows immediately from the above proof once the empirical measure $\frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ is considered in place of ν_{data}. This effectively completes the proof of Lemma 1 in Section 2.

Lemma 1. *The objective $\hat{\ell}_{\mathrm{dsm}}(s;S,\tau)$ is identical, up to a constant, to the objective*

$$\hat{\ell}_{\mathrm{sm}}(s;S,\tau) := \int \mathbb{E}[\|s(\hat{X}_t,t) - \nabla \log \hat{p}_t(\hat{X}_t)\|^2|S]\tau(dt), \tag{23}$$

where $\hat{p}_t$ is the marginal density of $\hat{X}_t$. Therefore, any minimiser of $\hat{\ell}_{\mathrm{dsm}}(\cdot;S,\tau)$ on $L^0(\mathbb{R}^d \times [0,T];\mathbb{R}^d)$ is identical to $\nabla \log \hat{p}_t$ a.e. for any $t \in \mathrm{supp}(\tau)$.

Proof. The proof follows nearly immediately from (21). Since $p_{t|0}$ is C^∞, $\nabla \log p_{t|0}$ is measurable and thus its empirical average $\nabla \log \hat{p}_t$ must be also. Therefore, the score function $s^\star(x,t) = \nabla \log \hat{p}_t(x)$ satisfies $\hat{s}^\star \in L^0(\mathbb{R}^d \times [0,T];\mathbb{R}^d)$ as well as,

$$\hat{\ell}_{\mathrm{sm}}(\hat{s}^\star;S,\tau) = 0.$$

Now let $s \in L^0(\mathbb{R}^d \times [0,T];\mathbb{R}^d)$ be any minimiser of $\hat{\ell}_{\mathrm{dsm}}(\cdot;S,\tau)$. Through the equivalence of $\hat{\ell}_{\mathrm{dsm}}$ and $\hat{\ell}_{\mathrm{sm}}$ up to a constant, it follows that s must also be a minimiser of $\hat{\ell}_{\mathrm{sm}}(\cdot;S,\tau)$ and, due to the existence of $\hat{s}^\star$, must satisfy $\hat{\ell}_{\mathrm{sm}}(s;S,\tau) = 0$ also. Letting $t \in \mathrm{supp}(t)$, we note that since $t > 0$, we must have that $p_{t|0}$ has full support and thus, $s(\cdot,t) = s^\star(\cdot,t)$ almost everywhere. □

A.3 MANIFOLDS

We also introduce some basic properties of smooth manifolds, primarily referencing Aamari et al. (2019). We define the manifold reach and include a known property of this quantity.

Definition 17. *The reach of a set $A \subset \mathbb{R}^d$, is defined by $\tau_A = \inf_{p \in A} d(p, Med(A))$, where we define the set,*

$$Med(A) = \Big\{ z \in \mathbb{R}^d : \exists p, q \in A \text{ s.t. } p \neq q, \|p - z\| = \|q - z\| \Big\}.$$

Lemma 18. *Suppose that the measure μ is supported on a manifold M with reach $\tau_M > 0$ and dimension d^*. Then, for any $r \leq \tau_M$, we have*

$$\mu(B_r(x)) \geq \Big| \inf_{B_r(x)} p_\mu \Big| \, r^{d^*},$$

where p_μ denotes the density of μ with respect to the volume measure on M.

For the proof of this lemma, we refer to the proof of Proposition 4.3 in Aamari et al. (2019) or Lemma III.23 in Aamari (2017).

B PROOFS FOR THE GENERALISATION GAP BOUNDS

We now provide provide the proof of theorem 3 that bound the generalisation gap under score stability guarantees. For the sake of brevity, throughout this section we suppress the notation for the time weighting, for example, using the shorthand $\hat{\ell}_{\mathrm{sm}}(s; S)$ in place of $\hat{\ell}_{\mathrm{sm}}(s; S, \tau)$.

Theorem 3. *Suppose that the score matching algorithm A_{sm} is score stable with constant ε_{stab}. Then, with $\hat{s} = A_{\mathrm{sm}}(S)$, it holds that*

$$\big| \mathbb{E}\big[\ell_{\mathrm{dsm}}(\hat{s}; \tau)\big]^{1/2} - \mathbb{E}\big[\hat{\ell}_{\mathrm{dsm}}(\hat{s}; S, \tau)\big]^{1/2} \big| \leq \varepsilon_{stab}. \tag{24}$$

Furthermore, it holds that

$$\mathbb{E}\big[\ell_{\mathrm{sm}}(\hat{s}; \tau)\big] - \mathbb{E}\big[\hat{\ell}_{\mathrm{sm}}(\hat{s}; S, \tau)\big] \leq 2\, \varepsilon_{stab}\, \mathbb{E}\big[\hat{\ell}_{\mathrm{dsm}}(\hat{s}; S, \tau)\big]^{1/2} + \varepsilon_{stab}^2. \tag{25}$$

Proof. Setting $\hat{s} = A_{\mathrm{sm}}(S)$ and $\hat{s}^i = A_{\mathrm{sm}}(S^i)$, we use the property that $(\hat{s}, \tilde{x})$ and $(\hat{s}^i, x_i)$ are distributed identically to obtain that,

$$\begin{aligned}
\mathbb{E}[\ell_{\mathrm{dsm}}(\hat{s}; \tau)] &= \mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s}; \{\tilde{x}\})] \\
&= \mathbb{E}\Big[\frac{1}{N} \sum_{i=1}^{N} \hat{\ell}_{\mathrm{dsm}}(\hat{s}^i; \{x_i\})\Big] \\
&= \mathbb{E}\Big[\frac{1}{N} \sum_{i=1}^{N} \int \mathbb{E}_{X_t}[\|\hat{s}^i(X_t, t, \omega) - \nabla \log p_{t|0}(X_t|x_i)\|^2 | X_0 = x_i, S]\, \tau(dt)\Big].
\end{aligned}$$

Therefore, it follows from the triangle inequality in L^2-norm that

$$\Big| \mathbb{E}[\ell_{\mathrm{dsm}}(\hat{s}; \tau)]^{1/2} - \mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s}; S)]^{1/2} \Big| \leq \mathbb{E}\Big[\frac{1}{N} \sum_{i=1}^{N} \int \mathbb{E}[\|\hat{s}(X_t, t) - \hat{s}^i(X_t, t)\|^2 | X_0 = x_i, S]\, \tau(dt)\Big]^{1/2}$$

Note that if the algorithm A_{sm} is stochastic, the right-hand side would hold regardless of how $\hat{s}|S, \tilde{x}$ and $\hat{s}^i|S, \tilde{x}$ were coupled. Therefore the most efficient coupling can be chosen, leading to the bound,

$$\begin{aligned}
&\Big| \mathbb{E}[\ell_{\mathrm{dsm}}(\hat{s}; \tau)]^{1/2} - \mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s}; S)]^{1/2} \Big| \qquad (26) \\
&\qquad \leq \mathbb{E}\Big[\inf_{(\hat{s}, \hat{s}^i) \in \Gamma_i} \frac{1}{N} \sum_{i=1}^{N} \int \mathbb{E}[\|\hat{s}(X_t, t) - \hat{s}^i(X_t, t)\|^2 | X_0 = x_i, S]\, \tau(dt)\Big]^{1/2} \\
&\qquad \leq \varepsilon_{\mathrm{stab}}, \qquad (27)
\end{aligned}$$

completing the proof of the bound in (24).

To obtain the bound in (25), we use Lemma 16 to derive

$$\mathbb{E}[\ell_{\mathrm{sm}}(\hat{s};\tau)] = \mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s};S)] + \mathbb{E}[\ell_{\mathrm{dsm}}(\hat{s};\tau) - \hat{\ell}_{\mathrm{dsm}}(\hat{s};S)] + \mathbb{E}\big[\hat{\ell}_{\mathrm{dsm}}(\nabla \log \hat{p}_t;S)\big] - \ell_{\mathrm{dsm}}(\nabla \log p_t;\tau). \tag{28}$$

Since $\hat{\ell}_{\mathrm{dsm}}(\cdot;S)$ is a unbiased estimator of $\ell_{\mathrm{sm}}(\cdot;\tau)$, we have that

$$\ell_{\mathrm{dsm}}(\nabla \log p_t;\tau) = \mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\nabla \log p_t;S)] \geq \mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\nabla \log \hat{p}_t;S)], \tag{29}$$

where the inequality follows from the fact that $\nabla \log \hat{p}_t$ minimises $\hat{\ell}_{\mathrm{dsm}}$. Furthermore, using (27), we deduce the bound,

$$\begin{aligned}
&|\mathbb{E}[\ell_{\mathrm{dsm}}(\hat{s};\tau) - \hat{\ell}_{\mathrm{dsm}}(\hat{s};S)]| \\
&\quad = \Big(\mathbb{E}[\ell_{\mathrm{dsm}}(\hat{s};\tau)]^{1/2} + \mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s};S)]^{1/2}\Big)\Big|\mathbb{E}[\ell_{\mathrm{dsm}}(\hat{s};S)]^{1/2} - \mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s};S)]^{1/2}\Big| \\
&\quad \leq \Big(2\mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s};S)]^{1/2} + \varepsilon_{\text{stab}}\Big)\varepsilon_{\text{stab}} \\
&\quad = 2\varepsilon_{\text{stab}}\mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s};S)]^{1/2} + \varepsilon_{\text{stab}}^2.
\end{aligned} \tag{30}$$

Thus, substituting (29) and (30) in to (28) recovers the bound in (25) in the statement. □

We obtain upper bounds relying on the fact that the constant separating the score matching loss from the denoising score matching loss is larger on average in the empirical case. One could obtain lower bounds through our techniques but this would require an analysis of the rate of convergence of this constant which is beyond the scope of this paper.

C Proofs for stability of empirical denoising score matching

In this section, we provide the proof for Theorem 3, where the algorithm that minimises $\hat{\ell}_{\mathrm{dsm}}(\cdot;S,\tau)$ over some class of score functions $\mathcal{H}$ is shown to be score stable.

C.1 On-average stability of the ERM algorithm

We begin with an important lemma that shows that under minimal assumptions, $\hat{s} = A_{\mathrm{erm}}(S)$ and $\hat{s}^i = A_{\mathrm{erm}}(S)$ are close in L^2 space, averaged over the full dataset. The first half of this proof utilises the fact that $\hat{\ell}_{\mathrm{dsm}}$ is 1-strongly convex in a weighted L^2 space, exploiting a well-known relationship between strong-convexity and algorithmic stability (e.g. see (Bousquet & Elisseeff, 2002; Charles & Papailiopoulos, 2018; Vary et al., 2024; Attia & Koren, 2022)).

Lemma 19. *Suppose that A_{erm} is score stable with constant ε_{stab}, then for any $i \in [N]$, we obtain,*

$$\mathbb{E}\bigg[\int\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2 \hat{p}_t(dy)\,\tau(dt)\bigg] \leq 8\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})] + \frac{8}{N}\varepsilon_{stab}(C_{\mathrm{sm}}^{1/2} + \varepsilon_{stab}) \tag{31}$$

where $\hat{s} = A_{\mathrm{erm}}(S), \hat{s}^i = A_{\mathrm{erm}}(S)$.

Proof. Choose $i \in [N]$ and let $\hat{s} = A_{\mathrm{erm}}(S), \hat{s}^i = A_{\mathrm{erm}}(S^i)$ so that $\hat{s} \in \operatorname{argmin}_{\mathcal{H}} \hat{\ell}_{\mathrm{dsm}}(\cdot;S,\tau), \hat{s}^i \in \operatorname{argmin}_{\mathcal{H}} \hat{\ell}_{\mathrm{dsm}}(\cdot;S^i,\tau)$. The proof begins with the following simple expression, that holds for all $j \in [N]$:

$$\begin{aligned}
&2\int \big\langle \hat{s}^i(y,t) - \hat{s}(y,t), \hat{s}^i - \nabla \log p_{t|0}(y|x_j)\big\rangle\, p_{t|0}(dy|x_j) \\
&\quad = \int \|\hat{s}^i(y,t) - \nabla \log p_{t|0}(y|x_j)\|^2\, p_{t|0}(dy|x_j) - \int \|\hat{s}(y,t) - \nabla \log p_{t|0}(y|x_j)\|^2\, p_{t|0}(dy|x_j) \\
&\qquad + \int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2\, p_{t|0}(dy|x_j).
\end{aligned}$$

By averaging over $j \in [N]$ and integrating with respect to $\tau(dt)$, we arrive at the upper bound,

$$\begin{aligned}
&\frac{2}{N}\sum_{j\in[N]}\int\int \left\langle \hat{s}^i(y,t) - \hat{s}(y,t), \hat{s}^i - \nabla \log p_{t|0}(y|x_j)\right\rangle p_{t|0}(dy|x_j)\,\tau(dt) \\
&\quad = \hat{\ell}_{\mathrm{dsm}}(\hat{s}^i; S, \tau) - \hat{\ell}_{\mathrm{dsm}}(\hat{s}; S, \tau) + \int\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2\, \hat{p}_t(dy)\,\tau(dt) \\
&\quad \geq \int\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2\, \hat{p}_t(dy)\,\tau(dt), \qquad (32)
\end{aligned}$$

where the inequality follows from the fact that $\hat{\ell}_{\mathrm{dsm}}(\hat{s}; S, \tau) \leq \hat{\ell}_{\mathrm{dsm}}(s; S, \tau)$ for any score function $s \in \mathcal{H}$. Additionally, the left-hand side is upper bounded using the Cauchy-Schwarz inequality to obtain,

$$\begin{aligned}
&\frac{2}{N}\sum_{x\in S}\int\int \left\langle \hat{s}^i(y,t) - \hat{s}(y,t), \hat{s}^i - \nabla \log p_{t|0}(y|x)\right\rangle p_{t|0}(dy|x)\,\tau(dt) \\
&\quad = \frac{2}{N}\sum_{x\in S^i}\int\int \left\langle \hat{s}^i(y,t) - \hat{s}(y,t), \hat{s}^i(y,t) - \nabla \log p_{t|0}(y|x)\right\rangle p_{t|0}(dy|x)\,\tau(dt) \\
&\qquad + \frac{2}{N}\int\int \left\langle \hat{s}^i(y,t) - \hat{s}(y,t), \hat{s}^i(y,t) - \nabla \log p_{t|0}(y|x_i)\right\rangle p_{t|0}(dy|x_i)\,\tau(dt) \\
&\qquad - \frac{2}{N}\int\int \left\langle \hat{s}^i(y,t) - \hat{s}(y,t), \hat{s}^i(y,t) - \nabla \log p_{t|0}(y|\tilde{x})\right\rangle p_{t|0}(dy|\tilde{x})\,\tau(dt) \\
&\quad \leq 2\hat{\ell}_{\mathrm{sm}}(\hat{s}^i; S^i, \tau)^{1/2}\bigg(\int\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2\, \hat{p}^i_t(dy)\,\tau(dt)\bigg)^{1/2} \\
&\qquad + \frac{2}{N}\hat{\ell}_{\mathrm{dsm}}(\hat{s}^i; \{x_i\}, \tau)^{1/2}\bigg(\int\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2\, p_{t|0}(dy|x_i)\,\tau(dt)\bigg)^{1/2} \\
&\qquad + \frac{2}{N}\hat{\ell}_{\mathrm{dsm}}(\hat{s}^i; \{\tilde{x}\}, \tau)^{1/2}\bigg(\int\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2\, p_{t|0}(dy|\tilde{x})\,\tau(dt)\bigg)^{1/2}, \qquad (33)
\end{aligned}$$

where $\hat{p}^i_t(dy) = \frac{1}{N}\sum_{x\in S^i} p_{t|0}(dy|x)$. Combining the expressions in (32) and (33) and taking the expectation, we derive the bound,

$$\begin{aligned}
&\mathbb{E}\bigg[\int\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2\, \hat{p}_t(dy)\,\tau(dt)\bigg] \\
&\quad \leq 2\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s}^i; S^i, \tau)]^{1/2}\mathbb{E}\bigg[\int\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2\, \hat{p}^i_t(dy)\,\tau(dt)\bigg]^{1/2} \\
&\qquad + \frac{2}{N}\mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s}^i; \{x_i\}, \tau)]^{1/2}\mathbb{E}\bigg[\int\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2\, p_{t|0}(dy|x_i)\,\tau(dt)\bigg]^{1/2} \\
&\qquad + \frac{2}{N}\mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s}^i; \{\tilde{x}\}, \tau)]^{1/2}\mathbb{E}\bigg[\int\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2\, p_{t|0}(dy|\tilde{x})\,\tau(dt)\bigg]^{1/2} \\
&\quad \leq 2\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s}; S, \tau)]^{1/2}\mathbb{E}\bigg[\int\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2\, \hat{p}_t(dy)\,\tau(dt)\bigg]^{1/2} \\
&\qquad + \frac{2}{N}\varepsilon_{\mathrm{stab}}\Big(\mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s}; S, \tau)]^{1/2} + \mathbb{E}[\ell_{\mathrm{dsm}}(\hat{s},)]^{1/2}\Big),
\end{aligned}$$

where we recall that $\varepsilon_{\mathrm{stab}}$ is the stability constant for A_{erm}. Here, we have used the fact that $(\hat{s}, S)$ has the same law as $(\hat{s}^i, S^i)$ and also $\mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s}^i; \{\tilde{x}\})] = \mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s}; S)]$ and $\mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s}^i; \{x_i\})] =$

$\mathbb{E}[\ell_{\mathrm{dsm}}(\hat{s})]$. By solving the quadratic equation, we deduce that the above inequality implies that,

$$\mathbb{E}\Big[\int\int \|\hat{s}^i(y,t)-\hat{s}(y,t)\|^2\,\hat{p}_t(dy)\,\tau(dt)\Big]$$
$$\leq \left(\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s};S,\tau)]^{\frac{1}{2}}+\sqrt{\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s};S,\tau)]+\tfrac{2}{N}\varepsilon_{\mathrm{stab}}(\mathbb{E}[\ell_{\mathrm{dsm}}(\hat{s};\tau)]^{\frac{1}{2}}+\mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s};S,\tau)]^{1/2})}\right)^2$$
$$\leq 4\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s};S,\tau)]+\frac{4}{N}\varepsilon_{\mathrm{stab}}(\mathbb{E}[\ell_{\mathrm{dsm}}(\hat{s};\tau)]^{\frac{1}{2}}+\mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s};S,\tau)]^{\frac{1}{2}}).$$

We simplify the above expression further using Theorem 3. Using the stability assumption, it follows from (24) that $\mathbb{E}[\ell_{\mathrm{dsm}}(\hat{s})]^{1/2}\leq\mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s})]^{1/2}+\varepsilon$. Furthermore, from Lemma 16, we have

$$\mathbb{E}[\hat{\ell}_{\mathrm{dsm}}(\hat{s})]=\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})]+\mathbb{E}[\hat{C}_{\mathrm{sm}}]$$
$$\leq \mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})]+C_{\mathrm{sm}},$$

where we recall the definitions of $\hat{C}_{\mathrm{sm}}$ and C_{sm} from (22) and (20) and recall that $\mathbb{E}[\hat{C}_{\mathrm{sm}}]\leq C_{\mathrm{sm}}$ from (29). Thus, from Young's inequality, we obtain the bound

$$\mathbb{E}\Big[\int\int \|\hat{s}^i(y,t)-\hat{s}(y,t)\|^2\,\hat{p}_t(dy)\,\tau(dt)\Big]$$
$$\leq 4\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})]+\frac{4}{N}\varepsilon_{\mathrm{stab}}(2\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})]^{1/2}+2C_{\mathrm{sm}}^{1/2}+\varepsilon_{\mathrm{stab}})$$
$$\leq 8\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})]+\frac{4}{N}\varepsilon_{\mathrm{stab}}(\varepsilon_{\mathrm{stab}}/N+2C_{\mathrm{sm}}^{1/2}+\varepsilon_{\mathrm{stab}})$$
$$\leq 8\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})]+\frac{8}{N}\varepsilon_{\mathrm{stab}}(C_{\mathrm{sm}}^{1/2}+\varepsilon_{\mathrm{stab}}).$$

□

C.2 Proof of Proposition 6

To obtain the stability bound in Proposition 6, we convert the result in Lemma 19, which is a bound in $L^2(\hat{p}_t)$, to a bound in $L^2(p_{t|0}(\cdot|\tilde{x})$ which is required of score stability. For this, we rely on two further lemmas, the first of which is a fundamental property of the Ornstein-Uhlenbeck process, captured by the Harnack inequality of Wang (1997) (see Theorem 5.6.1 Bakry et al. (2014)).

Lemma 20 (Wang's Harnack inequality)**.** *For each positive measurable function $\phi:\mathbb{R}^d\to\mathbb{R}$, every $t>0, p>1$ and every $x,y\in\mathbb{R}^d$, it holds that*

$$\mathbb{E}[\phi(X_t)|X_0=x]\leq\mathbb{E}[\phi(X_t)^p|X_0=y]^{1/p}\exp\left(\frac{\mu_t^2\|x-y\|^2}{2(p-1)\sigma_t^2}\right).$$

This result describes the stability of the diffusion semigroup under changes in initial position and shows that as t grows, the distribution of X_t depends less on X_0. The second lemma, for which we provide a proof, controls the empirical measure,

$$\hat{\nu}(dx)=\frac{1}{N}\sum_{i=1}^N\delta_{x_i}(dx),$$

on balls around training examples.

Lemma 21. *Suppose that Assumption 4 is satisfied, then for any $i\in[N], r\in(0,\tau_{reach}]$ and any decreasing function $\phi:(0,\infty)\to\mathbb{R}_+$, we have the bound*

$$\mathbb{E}\Big[\phi\Big(\hat{\nu}(B_r(x_i))\Big)\Big]\leq\phi(N^{-1})\exp(-c_\nu N^2 r^{d^*})+\phi(c_\nu r^{d^*}/2),$$

whenever $N\geq 4c_\nu^{-1}r^{-d^*}$, where $c_\nu=\inf p_\nu$.

Proof. We rewrite the object $\hat{\nu}(B_r(x_i))$ as an empirical average of Bernoulli random variables

$$\hat{\nu}(B_r(x_i)) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{x_j \in B_r(x_i)} = \frac{1}{N} + \frac{1}{N}\sum_{j\neq i} \mathbb{1}_{x_j \in B_r(x_i)}.$$

When conditioned on x_i, the random variables $(\mathbb{1}_{x_j \in B_r(x_i)})_{j\neq i}$ are independently and identically distributed Bernoulli random variable with probability $\mu = \nu_{\mathrm{data}}(B_r(x_i))$. To utilise concentration of the empirical process, we first rewrite the probability

$$\mathbb{P}\Big(\hat{\nu}(B_r(x_i)) \leq \mu/2 \Big| x_i\Big) \leq \mathbb{P}\Big(S_{N-1} \leq \frac{N\mu}{2} - 1 \Big| x_i\Big), \qquad S_{N-1} = \sum_{j\neq i} \mathbb{1}_{x_j \in B_r(x_i)}.$$

Therefore, by Chernoff's inequality we obtain

$$\begin{aligned}\mathbb{P}\Big(\hat{\nu}(B_r(x_i)) \leq \mu/2 \Big| x_i\Big) &\leq \exp\Big(-\mu^{-1}(N\mu/2 - 1)^2\Big) \\ &\leq \exp(-N^2\mu/16),\end{aligned}$$

where the last bound holds when $N \geq 4\mu^{-1}$. Therefore, using the above bound as well as the trivial bound $\hat{\nu}(B_r(x_i)) \geq N^{-1}$ we apply the law of total expectation to obtain,

$$\begin{aligned}\mathbb{E}\Big[\phi\Big(\hat{\nu}(B_r(x_i))\Big)\Big| x_i\Big] &= \mathbb{E}\Big[\phi\Big(\hat{\nu}(B_r(x_i))\Big)\Big| \hat{\nu}(B_r(x_i)) > \mu/2\Big] + \mathbb{P}\Big(\hat{\nu}(B_r(x_i)) \leq \mu/2 \Big| x_i\Big)\phi(N^{-1}) \\ &\leq \phi(\mu/2) + \exp(-N^2\mu/16)\phi(N^{-1}).\end{aligned}$$

To control μ, we use Lemma 18 which asserts that $\mu \geq c_\nu r^{d^*}$. □

This now brings us to the proof of the proposition, which we first restate.

Proposition 6. *Suppose that assumptions 4 and 5 hold and that $\epsilon := \inf \mathrm{supp}(\tau) \in (0, \tau^2_{reach})$, then for any $c \in (0,1)$ and sufficiently large N, the score matching algorithm A_{erm} is score stable with,*

$$\varepsilon^2_{stab} \lesssim C\big(CC_{\mathrm{sm}}N^{-2} + \mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})]\big)^c, \qquad C = \tfrac{D^2_{\mathcal{M}}}{\sigma^4_\epsilon} \vee \tfrac{1}{c_\nu \sigma^{d^*}_\epsilon}.$$

Proof. We use the shorthand $\hat{\ell}_{\mathrm{sm}}(s) = \hat{\ell}_{\mathrm{sm}}(s; S, \tau), \hat{\ell}_{\mathrm{dsm}}(s) = \hat{\ell}_{\mathrm{dsm}}(s; S, \tau), \ell_{\mathrm{sm}}(s) = \ell_{\mathrm{sm}}(s; \tau)$ for the sake of brevity. We start from Lemma 19 which provides a bound on the difference between $\hat{s}^i$ and $\hat{s}$ in $L^2(\hat{p}_t)$ and use it to develop a bound in $L^2(\hat{p}_{t|0}(\cdot|\tilde{x}))$, as required by score stability. In particular, we define the quantity

$$\varepsilon^2 = \mathbb{E}\bigg[\int\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2 \, p_{t|0}(dy|x_i)\, \tau(dt)\bigg],$$

so that, by the symmetric of the algorithm, A_{erm} is score stable with constant ε (we have that $\varepsilon < \infty$ from Assumption 5. Therefore, from Lemma 19, we have

$$\mathbb{E}\bigg[\int\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2 \, \hat{p}_t(dy)\, \tau(dt)\bigg] \leq 8\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})] + \frac{8}{N}\varepsilon(C^{1/2}_{\mathrm{sm}} + \varepsilon).$$

We proceed using Lemma 20 with $\phi(y) = \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2$ to obtain that for any $j \in [N]$, $p > 1$,

$$\begin{aligned}&\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2 \, p_{t|0}(dy|x_i) \\ &\qquad \leq \bigg(\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^{2p} \, p_{t|0}(dy|x_j)\bigg)^{1/p} \exp\left(\frac{\mu_t^2\|x_i - x_j\|^2}{2(p-1)\sigma_t^2}\right)\end{aligned}$$

Given any subset of the dataset $B \subset S$ with $x_i \in B$ we can average over the above bound to obtain,

$$\begin{aligned}
&\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2 \, p_{t|0}(dy|x_i) \\
&\qquad \le \frac{1}{|B|} \sum_{x \in B} \left(\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^{2p} \, p_{t|0}(dy|x) \right)^{1/p} \exp\left(\frac{\mu_t^2 \operatorname{diam}(B)^2}{2(p-1)\sigma_t^2} \right) \\
&\qquad \le \left(\frac{1}{|B|} \sum_{x \in B} \int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^{2p} \, p_{t|0}(dy|x) \right)^{1/p} \exp\left(\frac{\mu_t^2 \operatorname{diam}(B)^2}{2(p-1)\sigma_t^2} \right) \\
&\qquad \le \hat{\nu}(B)^{-1/p} \left(\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^{2p} \, \hat{p}_t(dy) \right)^{1/p} \exp\left(\frac{\mu_t^2 \operatorname{diam}(B)^2}{2(p-1)\sigma_t^2} \right) \\
&\qquad \le (D_{\mathcal{H}}/\sigma_t^2)^{2(1-1/p)} \hat{\nu}(B)^{-1/p} \left(\int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2 \, \hat{p}_t(dy) \right)^{\frac{1}{p}} \exp\left(\tfrac{\mu_t^2 \operatorname{diam}(B)^2}{2(p-1)\sigma_t^2} \right),
\end{aligned}$$

where in the final inequality we use the L^∞ bound in Assumption 5. Integrating with respect to τ and taking the expectation, we obtain,

$$\begin{aligned}
\varepsilon^2 &\le (D_{\mathcal{H}}/\sigma_\epsilon^2)^{2/q} \mathbb{E}\Big[\hat{\nu}(B)^{-1/p}\Big] \mathbb{E}\left[\int \int \|\hat{s}^i(y,t) - \hat{s}(y,t)\|^2 \, \hat{p}_t(dy)\tau(dt) \right] \exp\left(\frac{\mu_\epsilon^2 \operatorname{diam}(B)^2}{2(p-1)\sigma_\epsilon^2} \right) \\
&\le (D_{\mathcal{H}}/\sigma_\epsilon^2)^{2/q} \mathbb{E}\Big[\hat{\nu}(B)^{-q/p}\Big]^{1/q} \left(8\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})] + \frac{8}{N}\varepsilon(C_{\mathrm{sm}}^{1/2} + \varepsilon) \right)^{1/p} \exp\left(\frac{\mu_\epsilon^2 \operatorname{diam}(B)^2}{2(p-1)\sigma_\epsilon^2} \right),
\end{aligned}$$

where we define $q := (1 - 1/p)^{-1}$. Using Young's inequality, it follows that for any $\lambda > 0$,

$$\varepsilon^2 \le \frac{D_{\mathcal{H}}^2}{\sigma_\epsilon^4 \lambda^q q} \mathbb{E}\Big[\hat{\nu}(B)^{-q/p}\Big] \exp\left(\frac{q\mu_\epsilon^2 \operatorname{diam}(B)^2}{2(p-1)\sigma_\epsilon^2} \right) + \frac{\lambda^p}{p} \left(8\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})] + \frac{8}{N}\varepsilon(C_{\mathrm{sm}}^{1/2} + \varepsilon) \right).$$

Setting $\kappa := 8\lambda^p / pN$, we can rearrange this to obtain the quadratic inequality,

$$(1-\kappa)\varepsilon^2 - C_{\mathrm{sm}}^{1/2}\kappa\varepsilon \le \left(\frac{8}{Np\kappa} \right)^{q/p} \frac{D_{\mathcal{H}}^2}{\sigma_\epsilon^4 q} \mathbb{E}\Big[\hat{\nu}(B)^{-q/p}\Big] \exp\left(\frac{q\mu_\epsilon^2 \operatorname{diam}(B)^2}{2(p-1)\sigma_\epsilon^2} \right) + N\kappa\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})].$$

Requiring that $\kappa \le 1/2$, we solve the quadratic to obtain the inequality,

$$\frac{\varepsilon^2}{4} \le C_{\mathrm{sm}}\kappa^2 + \left(\frac{8}{Np\kappa} \right)^{q/p} \frac{D_{\mathcal{H}}^2}{\sigma_\epsilon^4 q} \mathbb{E}\Big[\hat{\nu}(B)^{-q/p}\Big] \exp\left(\frac{q\mu_\epsilon^2 \operatorname{diam}(B)^2}{2(p-1)\sigma_\epsilon^2} \right) + N\kappa\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})]. \tag{34}$$

Next, we optimise B by setting $B = B_{\sigma_\epsilon}(x_i) \cap S$. We apply Lemma 21 with $\phi(r) = r^{-q/p}$ to obtain that whenever $\sigma_\epsilon \le \tau_{\text{reach}}$ we obtain,

$$\begin{aligned}
\mathbb{E}\Big[\hat{\nu}(B)^{-q/p}\Big] &\le N^{q/p} \exp(-c_\nu N^2 r^{d^*}) + \left(\frac{2}{c_\nu r^{d^*}} \right)^{q/p} \\
&\le 2\left(\frac{2}{c_\nu \sigma_\epsilon^{d^*}} \right)^{q/p},
\end{aligned}$$

where the second inequality holds whenever $N \ge q/2p$. Returning to (34), it follows from the above that

$$\frac{\varepsilon^2}{4} \le C_{\mathrm{sm}}\kappa^2 + \left(\frac{16}{Npc_\nu\sigma_\epsilon^{d^*}\kappa} \right)^{q/p} \frac{2D_{\mathcal{H}}^2}{\sigma_\epsilon^4 q} \exp\left(\frac{2q}{p-1} \right) + N\kappa\mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})]. \tag{35}$$

We now choose κ by optimising the second two terms of this bound, by which we arrive at the choice

$$\kappa^{q/p+1} = \frac{2D_{\mathcal{H}}^2}{\sigma_\epsilon^4 pN\gamma} \exp\left(\frac{2q}{p-1} \right) \left(\frac{16}{Npc_\nu\sigma_\epsilon^{d^*}} \right)^{q/p},$$

for some $\gamma > 0$. Substituting this in to (35), we arrive at the bound

$$\begin{aligned}
\frac{\varepsilon^2}{4} \le{}& C_{\mathrm{sm}}(Np)^{-2} \left(\frac{2D_{\mathcal{H}}^2}{\sigma_\epsilon^4} \right)^{2/q} \exp\left(\frac{4}{p-1} \right) \left(\frac{16}{c_\nu\sigma_\epsilon^{d^*}} \right)^{2/p} \gamma^{-1/q} \\
&+ \left(\frac{2D_{\mathcal{H}}^2}{\sigma_\epsilon^4} \right)^{1/q} \exp\left(\frac{2}{p-1} \right) \left(\frac{16}{c_\nu\sigma_\epsilon^{d^*}} \right)^{1/p} \left(\frac{\gamma^{1/p}}{q} + \frac{1}{p\gamma^{1/q}} \mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})] \right).
\end{aligned}$$

Optimising γ leads to the bound,

$$\begin{aligned}\frac{\varepsilon^2}{4} &\leq \left(\frac{2D_{\mathcal{H}}^2}{\sigma_\epsilon^4}\right)^{\frac{1}{q}} \exp\left(\frac{2}{p-1}\right)\left(\frac{16}{c_\nu \sigma_\epsilon^{d^*}}\right)^{\frac{1}{p}}\left(C_{\mathrm{sm}}N^{-2}\left(\frac{2D_{\mathcal{H}}^2}{\sigma_\epsilon^4}\right)^{\frac{1}{q}} \exp\left(\frac{2}{p-1}\right)\left(\frac{16}{c_\nu \sigma_\epsilon^{d^*}}\right)^{\frac{1}{p}}\right.\\ &\qquad \left. + \mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})]\right)^{\frac{1}{p}}\\ &\leq \left(\frac{2D_{\mathcal{H}}^2}{\sigma_\epsilon^4} \vee \frac{16}{c_\nu \sigma_\epsilon^{d^*}}\right)\exp\left(\frac{4}{p-1}\right)\left(\left(\frac{2D_{\mathcal{H}}^2}{\sigma_\epsilon^4} \vee \frac{16}{c_\nu \sigma_\epsilon^{d^*}}\right)C_{\mathrm{sm}}N^{-2} + \mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})]\right)^{1/p}\end{aligned}$$

Optimising p, we obtain

$$\frac{\varepsilon^2}{4} \lesssim \left(\tfrac{2D_{\mathcal{H}}^2}{\sigma_\epsilon^4} \vee \tfrac{16}{c_\nu \sigma_\epsilon^{d^*}}\right)\exp\left(\frac{5}{2\sqrt{2}}\log(\alpha^{-1})^{1/2} - 2\right)\alpha, \quad \alpha = \left(\tfrac{2D_{\mathcal{H}}^2}{\sigma_\epsilon^4} \vee \tfrac{16}{c_\nu \sigma_\epsilon^{d^*}}\right)C_{\mathrm{sm}}N^{-2} + \mathbb{E}[\hat{\ell}_{\mathrm{sm}}(\hat{s})],$$

from which the bound in the statement follows. To obtain that $\kappa \leq 1/2$ and $N \geq q/2q$, it is sufficient to require that N is sufficiently large. □

D Proofs for sampling and score stability

In this section, we provide details for the discretisation scheme considered in Section 5 and give the proof for Proposition 7 and Corollary 8. In the work of Potaptchik et al. (2024), they consider the following discretisation scheme, based on the scheme of (Benton et al., 2024):

$$\hat{y}_{k+1} = \mu_{t_{k+1}-t_k}^{-1}\hat{y}_k + \frac{\sigma_{t_{k+1}-t_k}^2}{\mu_{t_{k+1}-t_k}} s(\hat{y}_k, T - t_k) + \sigma_{t_{k+1}-t_k}\frac{\sigma_{T-t_{k+1}}}{\sigma_{T-t_k}}\zeta_k, \qquad k \in \{0, ..., K-1\},$$

where $\zeta_k \sim N(0, I_d)$ and we recall that the timesteps $(t_k)_{k=0}^K$ are given by,

$$t_k = \begin{cases} \kappa k, & \text{if } k < \frac{T-1}{\kappa}, \\ T - (1+\kappa)^{\frac{T-1}{\kappa} - k}, & \text{if } \frac{T-1}{\kappa} \leq k \leq K, \end{cases}$$

where $L = \frac{T-1}{\kappa} > 0$, $K = \lfloor L + \log(\epsilon^{-1})/\log(1+\kappa)\rfloor$ and $\kappa > 0, T \geq 1$ is chosen freely. We recall the following result from Potaptchik et al. (2024).

Lemma 22. *Suppose that $\alpha = 1$ and Assumption 4 holds with $\operatorname{diam}\operatorname{supp}(\nu_{\mathrm{data}}) \leq 1$. Then, it holds that,*

$$\begin{aligned} D(p_\epsilon \| A_{\mathrm{em}}(s)) &\lesssim \ell_{\mathrm{sm}}(s; \hat{\tau}) + D(p_T \| p_\infty) + \Delta_{\kappa,K}, \\ \Delta_{\kappa,K} &= \kappa + d^*\kappa^2(K - L)(\log(\epsilon^{-1}) + \sup|\log(p_\nu)|), \end{aligned}$$

where we define the measure,

$$\hat{\tau}(dt) = \frac{1}{K}\sum_{k=0}^{K-1}\delta_{T-t_k}(dt).$$

D.1 Coarse discretisation and regularisation

Fix $\epsilon > 0$ and suppose that κ is such that $\log(\epsilon^{-1})/\log(1+\kappa)$ is an integer. Set $K = L + \log(\epsilon^{-1})/\log(1+\kappa)$ so that, according to the discretisation scheme,

$$t_K = T - (1+\kappa)^{-\log(\epsilon^{-1})/\log(1+\kappa)} = T - \epsilon.$$

Proof of Proposition 7. Let $\hat{s} = A_{\mathrm{erm}}(S)$. We begin with Lemma 22, which provides the bound,

$$\mathbb{E}[D(p_\epsilon \| A_{\mathrm{em}}(\hat{s}))] \lesssim \mathbb{E}[\ell_{\mathrm{sm}}(\hat{s}; S, \hat{\tau})] + D(p_T \| p_\infty) + \Delta_{\kappa,K}.$$

For ϵ sufficiently small we have the bound,

$$\begin{aligned}\Delta_{\kappa,K} &= \kappa + d^*\kappa^2\frac{\log(\epsilon^{-1})}{\log(1+\kappa)}(\log(\epsilon^{-1}) + \sup|\log(p_\nu)|) \\ &\lesssim \kappa(1+\kappa)d^*\log(\epsilon^{-1})^2.\end{aligned}$$

Using Theorem 3, we obtain that if the algorithm is $\varepsilon_{\text{stab}}$-score stable, we have

$$\begin{aligned}\mathbb{E}[\ell_{\text{sm}}(\hat{s};\hat{\tau})] &\lesssim \mathbb{E}[\hat{\ell}_{\text{sm}}(\hat{s};S,\hat{\tau})] + \varepsilon_{\text{stab}}\mathbb{E}[\hat{\ell}_{\text{dsm}}(\hat{s};S,\hat{\tau})]^{1/2} + \varepsilon_{\text{stab}}^2 \\ &\lesssim \mathbb{E}[\hat{\ell}_{\text{sm}}(\hat{s};\hat{\tau})] + \varepsilon_{\text{stab}} C_{\text{sm}}^{1/2} + \varepsilon_{\text{stab}}^2\end{aligned}$$

Using Proposition 6 we obtain that with $\tau = \hat{\tau}$, A_{erm} is score stable, with constant,

$$\begin{aligned}\varepsilon_{\text{stab}}^2 &\lesssim C\Big(CC_{\text{sm}}N^{-2} + \mathbb{E}[\hat{\ell}_{\text{sm}}(\hat{s})]\Big)^c \\ &\lesssim c_\nu^{-1}\sigma_{T-t_{K-1}}^{-d^*}\Big(c_\nu^{-1}\sigma_{T-t_{K-1}}^{-d^*}C_{\text{sm}}N^{-2} + \mathbb{E}[\hat{\ell}_{\text{sm}}(\hat{s})]\Big)^c.\end{aligned}$$

Now by definition, we have that

$$T - t_{K-1} = (1+\kappa)^{L-K+1} = \epsilon(1+\kappa),$$

so if we take ϵ, κ sufficiently small so that $\epsilon(1+\kappa) \leq \frac{1}{2}$, we also have $\sigma^2_{\epsilon(1+\kappa)} \geq \epsilon(1+\kappa)$ and thus we obtain,

$$\begin{aligned}\varepsilon_{\text{stab}}^2 &\lesssim c_\nu^{-1}\epsilon^{-d^*/2}(1+\kappa)^{-d^*/2}\Big(c_\nu^{-1}\epsilon^{-d^*/2}(1+\kappa)^{-d^*/2}C_{\text{sm}}N^{-2} + \mathbb{E}[\hat{\ell}_{\text{sm}}(\hat{s})]\Big)^c. \\ &\lesssim c_\nu^{-1}\epsilon^{-d^*}(1+\kappa)^{-d^*}\Big(c_\nu^{-1}C_{\text{sm}}N^{-2} + \mathbb{E}[\hat{\ell}_{\text{sm}}(\hat{s})]\Big)^c,\end{aligned}$$

where in the last inequality, we use that $\epsilon(1+\kappa) \leq 1/2$. □

We now proceed by proving Corollary 8 in which the bound in Proposition 7 is optimised.

Proof of Corollary 8. Let $\tilde{\tau}_\epsilon$ denote the weak limit of the measure τ_κ as $\kappa \to 0^+$. Since $\text{supp}(\tilde{\tau}_\epsilon) \subseteq [\epsilon, T]$ and $\epsilon > 0$, we know that $\inf_{\mathcal{H}} \hat{\ell}_{\text{sm}}(\cdot; S, \tilde{\tau}_\epsilon) < \infty$. From this, we deduce that $\lim_{\kappa\to 0^+} B_\kappa < \infty$. With this there exists $\kappa^* \geq 1$ which is the smallest quantity satisfying,

$$(1+\kappa^*)^{2d^*+2} = \frac{B_{\kappa^*}}{\log(\epsilon^{-1})^2} \vee 1.$$

In the case that $B_{\kappa^*} > \log(\epsilon^{-1})$, we have that

$$\begin{aligned}&B_{\kappa^*}^{1/2}(1+\kappa^*)^{-d^*} + \frac{B_{\kappa^*}}{C_{\text{sm}}}(1+\kappa^*)^{-2d^*} + \kappa^*(1+\kappa^*)d^*\log(\epsilon^{-1})^2 \\ &\quad= B_{\kappa^*}^{\frac{1}{2(d^*+1)}}\log(\epsilon^{-1})^{\frac{d^*}{d^*+1}} + (C_{\text{sm}}^{-1} + d^*)B_{\kappa^*}^{\frac{1}{d^*+1}} \\ &\quad\leq B_{\kappa^*}^{\frac{1}{2(d^*+1)}}\log(\epsilon^{-1}) + (C_{\text{sm}}^{-1} + d^*)B_{\kappa^*}^{\frac{1}{d^*+1}}\log(\epsilon^{-1})^2.\end{aligned}$$

Plus, if $B_{\kappa^*} \leq \log(\epsilon^{-1})$ and therefore $\kappa^* = 1$, then there exists κ such that,

$$B_\kappa^{1/2}(1+\kappa)^{-d^*} + \frac{B_\kappa}{C_{\text{sm}}}(1+\kappa)^{-2d^*} + \kappa(1+\kappa)d^*\log(\epsilon^{-1})^2 \lesssim B_\kappa^{1/2} + \frac{B_\kappa}{C_{\text{sm}}} + de^{-T}.$$

Combining these leads to the bound in the statement. □

E PROOFS FOR STABILITY OF SGD

In this section, we analyse the stochastic optimisation scheme in (14), deriving the score stability bounds given in Proposition 11. We begin with a basic lemma that follows from weight decay and gradient clipping.

Lemma 23. *Suppose that $\eta_k < \lambda^{-1}$ for all $k \in \mathbb{N}$, then for any $K \in \mathbb{N}$, it holds that*

$$\|\theta_K\| \leq \frac{Ce}{\lambda} \vee \|\theta_0\|.$$

Proof. We begin with the bound,

$$\begin{aligned}\|\theta_{k+1}\| &\leq (1-\eta_k\lambda)\|\theta_k\| + \eta_k\|\mathrm{Clip}_C(G_k(\theta_k, \{x_i\}_{i\in B_k}))\| \\ &\leq (1-\eta_k\lambda)\|\theta_k\| + \eta_k C.\end{aligned}$$

By comparison, this leads to the bound

$$\begin{aligned}\|\theta_k\| &\leq C\sum_{k=0}^{K-1}\eta_k\prod_{i=k+1}^{K-1}(1-\eta_k\lambda) + \prod_{k=0}^{K-1}(1-\eta_k\lambda)\|\theta_0\| \\ &\leq C\sum_{k=0}^{K-1}\eta_k\exp\Big(\lambda\sum_{i=0}^{k}\eta_k\Big) + \exp\Big(-\lambda\sum_{i=0}^{K-1}\eta_k\Big)\|\theta_0\| \\ &\leq C\exp\Big(-\lambda\sum_{i=0}^{K-1}\eta_k + \lambda\max_k\eta_k\Big)\sum_{k=0}^{K-1}\eta_k\exp\Big(\lambda\sum_{i=0}^{k-1}\eta_k\Big) + \exp\Big(-\lambda\sum_{i=0}^{K-1}\eta_k\Big)\|\theta_0\|\end{aligned}$$

Since the sum forms a left Riemann sum, approximating an integral of an increasing function, we can upper bound it by the integral over $\exp(\lambda t)$. Furthermore, we have that $\lambda\max_k\eta_k \leq 1$, which leads to the bound,

$$\begin{aligned}\|\theta_k\| &\leq Ce\exp\Big(-\lambda\sum_{i=0}^{K-1}\eta_k\Big)\int_0^{\sum_{k=0}^{K-1}\eta_k}\exp(\lambda t)dt + \exp\Big(-\lambda\sum_{i=0}^{K-1}\eta_k\Big)\|\theta_0\| \\ &\leq \frac{Ce}{\lambda}\Big(1-\exp\Big(-\lambda\sum_{k=0}^{K-1}\eta_k\Big)\Big) + \exp\Big(-\lambda\sum_{k=0}^{K-1}\eta_k\Big)\|\theta_0\| \\ &\leq \frac{Ce}{\lambda}\vee\|\theta_0\|.\end{aligned}$$

□

We are now ready to prove Proposition 11.

Proposition 11. *Consider the score matching algorithm $A_{\mathrm{sm}} : S \mapsto s_{\theta_K}$ for some fixed $K \in \mathbb{N}$ where $(\theta_k)_k$ is as given in* (14)*. Suppose that assumptions 9 and 10 hold and $\eta_k \leq \bar{\eta}/k$ for all $k < K$, for some $\bar{\eta} \in (0, \lambda^{-1})$. Then, we obtain that A_{sm} is score stable with constant,*

$$\varepsilon_{stab}^2 \lesssim \left(\frac{C}{\lambda}\vee R\right)^{1+\frac{\bar{\eta}\upsilon}{\bar{\eta}\upsilon+1}}\frac{\overline{L}^2}{(\bar{\eta}\upsilon)\vee 1}\left(\frac{C}{\bar{\eta}}\right)^{\frac{1}{\bar{\eta}\upsilon+1}}\frac{N_B K^{\frac{\bar{\eta}\upsilon}{\bar{\eta}\upsilon+1}}}{N},$$

where $R^2 = \mathbb{E}[\|\theta_0\|^2], \upsilon = (\overline{M}B_\ell C_\tau^{1/2} + \overline{L}^2 - \lambda)\vee 0$ and $C_\tau = \int\sigma_t^{-4}\tau(dt)$.

Proof. Since the stochastic mini-batch scheme, and therefore the resulting score matching algorithm, is symmetric to dataset permutations, we consider stability under changes in the N^{th} entry of the dataset, without loss of generality. Let θ_k be the process given in (14), using the dataset S and let $\tilde{\theta}_k$ be the same process using S^N instead of S:

$$\tilde{\theta}_{k+1} = (1-\eta\lambda)\tilde{\theta}_k - \eta_k\,\mathrm{Clip}_C(G_k(\tilde{\theta}_p, \{\tilde{x}_i\}_{i\in B_k})), \qquad \tilde{\theta}_0 = \theta_0,$$

where $\tilde{x}_i = x_i$ for $i \neq N$, $\tilde{x}_N = \tilde{x}$. By having the processes share the same mini-batch indices B_k and gradient approximation G_k (i.e. sharing the same random time variables $t_{i,j}$ and noise $\xi_{i,j}$), we couple the processes θ_k and $\tilde{\theta}_k$.

We proceed by first controlling the stability of the gradient estimator, computing the bound,

$$
\begin{aligned}
&\|G_k(\theta_k,(x_i)_{i\in B_k}) - G_k(\tilde{\theta}_k,(x_i)_{i\in B_k})\| \\
&\quad\le \frac{1}{N_B P}\sum_{i\in B_k}\sum_{j=1}^{P} w_{t_{i,j}}\|\nabla s_{\theta_k}(X_{i,j},t_{i,j})^T(s_{\theta_k}(X_{i,j},t_{i,j}) - \nabla \log p_{t_{i,j}|0}(X_{t_{i,j}}|x_i)) \\
&\qquad - \nabla s_{\tilde{\theta}_k}(X_{t_{i,j}},t_{i,j})^T(s_{\tilde{\theta}_k}(X_{t_{i,j}},t_{i,j}) - \nabla \log p_{t_{i,j}|0}(X_{t_{i,j}}|x_i)))\| \\
&\quad\le \frac{1}{N_P P}\sum_{i\in B_k}\sum_{j=1}^{P} w_{t_{i,j}}\Big(\|\nabla s_{\theta_k}(X_{t_{i,j}},t_{i,j}) - \nabla s_{\tilde{\theta}_k}(X_{t_{i,j}},t_{i,j})\|\|s_{\theta_k}(X_{t_{i,j}},t_{i,j}) \\
&\qquad - \nabla \log p_{t_{i,j}|0}(X_{t_{i,j}}|x_i)\| + \|\nabla s_{\tilde{\theta}_k}(X_{t_{i,j}},t_{i,j})\|\|s_{\theta_k}(X_{t_{i,j}},t_{i,j}) - s_{\tilde{\theta}_k}(X_{t_{i,j}},t_{i,j})\|\Big) \\
&\quad\le \frac{1}{N_P P}\sum_{i\in B_k}\sum_{j=1}^{P} w_{t_{i,j}}\Big(M(X_{t_{i,j}},t_{i,j})\|s_{\theta_k}(X_{t_{i,j}},t_{i,j}) - \nabla \log p_{t_{i,j}|0}(X_{t_{i,j}}|x_j)\| \\
&\qquad + L(X_{t_{i,j}},t_{i,j})^2\Big)\|\theta_k - \tilde{\theta}_k\|.
\end{aligned}
$$

We control the expectation of this by first noting that,

$$
\begin{aligned}
&\mathbb{E}\Big[w_{t_{i,j}}\Big(M(X_{t_{i,j}},t_{i,j})\|s_{\theta_k}(X_{t_{i,j}},t_{i,j}) - \nabla \log p_{t_{i,j}|0}(X_{t_{i,j}}|x_j)\| + L(X_{t_{i,j}},t_{i,j})^2\Big)\Big|\theta_k,\tilde{\theta}_k,S,\tilde{x}\Big] \\
&\quad\le \left(\int \mathbb{E}[M(X_t,t)^2|X_0=x_i]\tau(dt)\right)^{1/2}\left(\int \hat{\ell}_{\mathrm{dsm}}(s_{\theta_k};\{x_i\},\delta_t)\tau(dt)\right)^{1/2} \\
&\qquad + \int \mathbb{E}[L(X_t,t)^2|X_0=x_i]\tau(dt) \\
&\quad\le \overline{M}B_\ell C_\tau^{1/2} + \overline{L}^2,
\end{aligned}
$$

where we define the quantity $C_\tau := \int \sigma_t^{-4}\tau(dt)$. From this, it follows that

$$
\mathbb{E}\Big[\|G_k(\theta_k,(x_i)_{i\in B_k}) - G_k(\tilde{\theta}_k,(x_i)_{i\in B_k})\|\Big|\theta_k,\tilde{\theta}_k,S,\tilde{x}\Big] \le \left(\overline{M}B_\ell C_\tau^{1/2} + \overline{L}^2\right)\|\theta_k - \tilde{\theta}_k\|.
$$

Furthermore, we can control the difference between $G_k(\tilde{\theta}_k,(x_i)_{i\in B_k})$ and $G_k(\tilde{\theta}_k,(\tilde{x}_i)_{i\in B_k})$, using the fact that they are identical whenever $N \notin B_k$. Thus, obtaining,

$$
\begin{aligned}
&\mathbb{E}\Big[\|\mathrm{Clip}_C(G(\theta_k,(x_i)_{i\in B_k})) - \mathrm{Clip}_C(G(\tilde{\theta}_k,(\tilde{x}_i)_{i\in B_k}))\|\Big|\theta_k,\tilde{\theta}_k,S,\tilde{x}\Big] \\
&\qquad\le \mathbb{E}\Big[\|G(\theta_k,(x_i)_{i\in B_k}) - G(\tilde{\theta}_k,(x_i)_{i\in B_k})\|\Big|\theta_k,\tilde{\theta}_k,S,\tilde{x}\Big] \\
&\qquad\quad + \mathbb{E}\Big[\|\mathrm{Clip}_C(G(\tilde{\theta}_k,(x_i)_{i\in B_k})) - \mathrm{Clip}_C(G(\tilde{\theta}_k,(\tilde{x}_i)_{i\in B_k}))\|\Big|\theta_k,\tilde{\theta}_k,S,\tilde{x}\Big] \\
&\qquad\le \left(\overline{M}B_\ell C_\tau^{1/2} + \overline{L}^2\right)\|\theta_k - \tilde{\theta}_k\| + 2C\frac{N_B}{N},
\end{aligned}
$$

where we have used the fact that $\mathbb{P}(N \in B_k) = \frac{N_B}{N}$. Thus, using (14), we obtain that for any $k_0 \le k$,

$$
\begin{aligned}
&\mathbb{E}\Big[\|\theta_{k+1} - \tilde{\theta}_{k+1}\|\Big|\theta_{k_0},\tilde{\theta}_{k_0},S,\tilde{x}\Big] \\
&\qquad\le \Big(1 + \eta_k\Big(\overline{M}B_\ell C_\tau^{1/2} + \overline{L}^2 - \lambda\Big)\Big)\,\mathbb{E}\Big[\|\theta_k - \tilde{\theta}_k\|\Big|\theta_{k_0},\tilde{\theta}_{k_0},S,\tilde{x}\Big] + 2\eta_k C\frac{N_B}{N}. \\
&\qquad\le (1+\eta_k\upsilon)\,\mathbb{E}\Big[\|\theta_k - \tilde{\theta}_k\|\Big|\theta_{k_0},\tilde{\theta}_{k_0},S,\tilde{x}\Big] + 2\eta_k C\frac{N_B}{N},
\end{aligned}
$$

where $\upsilon = \overline{M}B_\ell C_\tau^{1/2} + \overline{L}^2 - \lambda$. Thus, by comparison, we obtain,

$$
\mathbb{E}\Big[\|\theta_K - \tilde{\theta}_K\|\Big|\theta_{k_0},\tilde{\theta}_{k_0},S,\tilde{x}\Big] \le \sum_{i=k_0}^{K-1} 2\eta_i C\frac{N_B}{N}\prod_{j=i+1}^{K-1}(1+\eta_j\upsilon) + \|\theta_{k_0} - \tilde{\theta}_{k_0}\|\prod_{j=k_0}^{K-1}(1+\eta_j\upsilon).
$$

From this we obtain the following:

$$\begin{aligned}\mathbb{E}[\|\theta_K - \tilde{\theta}_K\| | \theta_{k_0} = \tilde{\theta}_{k_0}, S, \tilde{x}] &\leq 2C\frac{N_B}{N}\sum_{i=k_0}^{K-1}\eta_i \exp\left(\sum_{j=i+1}^{K-1}\eta_j \upsilon\right) \\ &\leq \frac{2CN_B\bar{\eta}}{N}\sum_{i=k_0}^{K-1}\frac{1}{i}\left(\frac{K}{i}\right)^{\bar{\eta}\upsilon} \\ &\lesssim \frac{CN_B}{N\upsilon}\left(\frac{K}{k_0}\right)^{\bar{\eta}\upsilon},\end{aligned}$$

where we use the fact that $\sum_{j=i+1}^{K-1}\frac{1}{j} \leq \log(K) - \log(i)$. By the law of total probability, we have

$$\begin{aligned}&\mathbb{E}[\|\theta_K - \tilde{\theta}_K\| | \theta_0] \\ &\quad= \mathbb{E}[\|\theta_K - \tilde{\theta}_K\| | \theta_{k_0} = \tilde{\theta}_{k_0}]\mathbb{P}(\theta_{k_0} = \tilde{\theta}_{k_0}|\theta_0) + \mathbb{E}[\|\theta_K - \tilde{\theta}_K\| | \theta_{k_0} \neq \tilde{\theta}_{k_0}, \theta_0]\mathbb{P}(\theta_{k_0} \neq \tilde{\theta}_{k_0}|\theta_0) \\ &\quad\lesssim \frac{CN_B}{N\upsilon}\left(\frac{K}{k_0}\right)^{\bar{\eta}\upsilon} + \left(\frac{Ce}{\lambda} \vee \|\theta_0\|\right)\frac{k_0 N_B}{N},\end{aligned}$$

where in the second inequality, we use Lemma 23. Thus, optimising k_0 leads to the bound,

$$\mathbb{E}[\|\theta_K - \tilde{\theta}_K\| | \theta_0] \lesssim \left(\frac{C}{c}\right)^{\frac{1}{\upsilon+1}}(1 + 1/c\upsilon)\left(\frac{Ce}{\lambda} \vee \|\theta_0\|\right)^{\frac{c\upsilon}{c\upsilon+1}}\frac{N_B}{N}K^{\frac{c\upsilon}{c\upsilon+1}}.$$

Finally, we obtain score stability using the fact that

$$\begin{aligned}\int \mathbb{E}[\|s_{\theta_K}(X_t, t) - s_{\tilde{\theta}_K}(X_t, t)\|^2 | X_0 = \tilde{x}, S]\tau(dt) &\\ &\leq \mathbb{E}\Big[\bar{L}^2\|\theta_K - \tilde{\theta}_K\|^2\Big] \\ &\leq 2\mathbb{E}\Big[\bar{L}^2\left(\frac{Ce}{\lambda} \vee \|\theta_0\|\right)\|\theta_K - \tilde{\theta}_K\|\Big] \\ &\lesssim \bar{L}^2\left(\frac{Ce}{\lambda} \vee R\right)^{1+\frac{c\upsilon}{c\upsilon+1}}\left(\frac{C}{c}\right)^{\frac{1}{c\upsilon+1}}(1 + 1/c\upsilon)\frac{N_B}{N}K^{\frac{c\upsilon}{c\upsilon+1}},\end{aligned}$$

where $R^2 = \mathbb{E}\|\theta_0\|^2$. □

F WASSERSTEIN CONTRACTIONS

In this section, we derive the Wasserstein contraction result used in the proof of Proposition 14. We begin with the more abstract problem of deriving Wasserstein contractions for a discrete time diffusion process with anisotropic non-constant volatility. We consider stochastic processes given by the discrete-time update,

$$x_{k+1} = (1 - \eta\lambda)x_k + \eta b(x_k) + \sqrt{\eta}\sigma(x_k)\xi_k, \tag{36}$$

$$y_{k+1} = (1 - \eta\lambda)y_k + \eta\tilde{b}(y_k) + \sqrt{\eta}\tilde{\sigma}(y_k)\xi_k, \tag{37}$$

for some $b, \tilde{b} : \mathbb{R}^d \to \mathbb{R}^d, \sigma, \tilde{\sigma} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ where $\xi_k \sim N(0, I_d)$, and we show that the laws of x_k and y_k contract in Wasserstein distance. We borrow the strategy developed by Eberle (2016) and extended in (Eberle & Majka, 2019; Majka et al., 2020), constructing a coupling and a metric for which exponential contractions of the coupling can be obtained. However, these works are restricted to the setting of isotropic noise with constant volatility (i.e. $\sigma(x) = cI_d$) and so some careful modification to the strategy is required. In particular, we analyse this process with respect to the seminorm $\|\cdot\|_{G^+}$ given by $\|x\|_{G^+}^2 = x^T G^+ x$, where G^+ denotes the Moore-Penrose pseudoinverse of the matrix G. Furthermore, we allow for x_k and y_k to have different bias and volatility terms and so controlling for this will also require some modifications to the proof technique.

To define our coupling we first suppose that there exists a symmetric positive semi-definite matrix $G \in \mathbb{R}^{d\times d}$ such that $\sigma(x), \tilde{\sigma}(y) \succcurlyeq G^{1/2}$ for all $x \in \mathbb{R}^d$, and to couple an update from the above process starting at $x, y \in \mathbb{R}^d$, we first define the update,

$$\tilde{x} = (1-\eta\lambda)x + \eta b(x), \qquad \tilde{y} = (1-\eta\lambda)y + \eta\tilde{b}(y),$$
$$\hat{x} = \tilde{x} + \sqrt{\eta}(\sigma(x) - G^{1/2})Z', \qquad \hat{y} = \tilde{y} + \sqrt{\eta}(\tilde{\sigma}(y) - G^{1/2})Z',$$

where $Z' \sim N(0, I_d)$. We then define the *synchronous coupled* processes,

$$X' = \hat{x} + \sqrt{\eta}G^{1/2}Z$$
$$Y'_s = \hat{y} + \sqrt{\eta}G^{1/2}Z,$$

with $Z \sim N(0, I_d)$. We also consider the reflection coupling,

$$Y'_r = \hat{y} + \sqrt{\eta}G^{1/2}\Big(I - 2(G^{1/2})^+ee^T(G^{1/2})^+\Big)Z, \qquad \text{with } e = (\hat{x}-\hat{y})/\|\hat{x}-\hat{y}\|_{G^+} \tag{38}$$

which has the noise act in the mirrored direction. We combine these couplings to arrive at the final coupling (X', Y'):

$$Y' = \begin{cases} X', & \text{if } \zeta \le \phi_{\hat{y},\eta G}(X')/\phi_{\hat{x},\eta G}(X'), |\langle e, Z\rangle|^2 < m^2/\eta \text{ and } \hat{r} \le r_1 \\ Y'_r, & \text{if } \zeta > \phi_{\hat{y},\eta G}(X')/\phi_{\hat{x},\eta G}(X'),\ |\langle e, Z\rangle|^2 < m^2/\eta \text{ and } \hat{r} \le r_1 \\ Y'_s, & \text{otherwise,} \end{cases} \tag{39}$$

for some fixed $m > 0$.

We assume the following regularity properties.

Assumption 24. *Suppose that b is bounded, satisfying $B := \sup_{x\in\mathbb{R}^n} \|b(x)\|_{G^+} < \infty$ and we have the Lipschitz property, $\|b(x) - b(y)\|_{G^+} \le L_b\|x-y\|_{G^+}$ and $\|\sigma(x) - \sigma(y)\|_{op,G^+} \le L_\sigma\|x-y\|_{G^+}$ for all $x, y \in \mathbb{R}^n$ and for some $L_b, L_\sigma \ge 0$.*

We also allow for $b \neq \tilde{b}$ and $\sigma \neq \tilde{\sigma}$, making the following assumption.

Assumption 25. *Suppose that $b, \tilde{b}$ satisfy $\|b(x) - \tilde{b}(x)\|_{G^+} \le \tilde{B}_b, \|\sigma(x) - \tilde{\sigma}(x)\|_{op,G^+} \le \tilde{B}_\sigma$ for all $x \in \mathbb{R}^n$ and for some $\tilde{B}_b, \tilde{B}_\sigma \ge 0$.*

We define the objects,

$$R = \|x-y\|_{G^+}, \qquad \tilde{r} = \|\tilde{x}-\tilde{y}\|_{G^+}, \qquad \hat{r} = \|\hat{x}-\hat{y}\|_{G^+}, \qquad R' = \|X'-Y'\|_{G^+}.$$

We wish to show that R' contracts in expectation, i.e. it is less than R on average. We modify the metric to guarantee this is possible. We define the function,

$$f(r) = \begin{cases} \frac{1}{a}(1-e^{-ar}), & \text{if } r \le r_2, \\ \frac{1}{a}(1-e^{-ar_2}) + \frac{1}{2r_2}e^{-ar_2}(r^2 - r_2^2), & \text{otherwise,} \end{cases}$$

where $a = 6L_br_1/c_0$, $r_1 = 4(1+\eta_0L_b)B/\lambda$, $r_2 = r_1 + \sqrt{\eta_0}$ and c_0, η_0 are defined below. The coupling and the strategy for proving contractions is closely based on an analysis in Majka et al. (2020) and for the sake of comparison, we rely on similar notation. We will also heavily borrow properties of the function f that are proven in this work.

By allowing σ to be non-constant, we run in to additional complications that are controlled by making the following assumption about the scale of L_σ.

Assumption 26. *Suppose that the following three inequalities hold:*

$$n-1, (\lambda^2/16L_\sigma^2 - 1)^2 \ge 32\log\left(\frac{8L_\sigma(6\vee(4a))\kappa_0^{1/2}}{\sqrt{\eta}(1-e^{-ar_2})c}\right), \qquad L_\sigma^2 \le \lambda/8n,$$

for some universal constant κ_0.

Under these assumptions, we obtain exponential contractions.

Proposition 27. *Suppose that assumptions 24, 25 and 26 hold and $m = \sqrt{\eta_0}/2$, then for any $\eta \le \eta_0$ and $x, y \in \mathbb{R}^d$, it holds that*

$$\mathbb{E}[f(R')] \le (1 - \eta c/4) f(r) + \frac{3}{2r_2} e^{-ar_2}(\eta^2 \tilde{B}^2 + \eta n \tilde{B}_\sigma^2),$$

where

$$c := \min\left\{ e^{-ar_2}\frac{\lambda}{16}, \frac{\frac{1}{2}e^{-ar_2}r_2}{\frac{1}{a}(1-e^{-ar_2})}\frac{\lambda}{16}, \frac{9L^2 r_1^2}{2c_0} e^{-6Lr_1^2/c_0}, \frac{3Lr_1}{16\sqrt{\eta_0}} \right\},$$
$$\eta_0 := \min\left\{ \frac{\lambda}{4L^2}, \frac{16}{\lambda}, \frac{1}{2L}, \frac{2c_0 \log(3/2)\lambda^2}{432 L^2 B^2}, \frac{4B^2}{\lambda^2}, \frac{c_0^2 (\log(2))^2 \lambda^2}{2304 L^2 B^2} \right\},$$

for some universal c_0 and $L = 2(L_b - \lambda)_+ + 4\eta^{-1/2} L_\sigma \sqrt{2(n-1)}$.

F.1 The coupling

Before we provide the proof of Proposition 27, we provide an explanation of how the coupling is arrived at. We begin by discussing the one-dimensional coupling of the Gaussian distribution that the construction is ultimately based on. Consider the following coupling of $\mathcal{N}(t, \eta)$ and $\mathcal{N}(s, \eta)$ for $t, s \in \mathbb{R}$: with $z \sim \mathcal{N}(0, 1)$,

$$t' = t + \sqrt{\eta} z, \tag{40}$$

$$s' = \begin{cases} t', & \text{if } \zeta \le \phi_{s,\eta}(t')/\phi_{t,\eta}(t'), |\sqrt{\eta} z| < \tilde{m}, \text{ and } |t - s| \le r_1, \\ s - \sqrt{\eta} z, & \text{if } \zeta > \phi_{s,\eta}(t')/\phi_{t,\eta}(t'), |\sqrt{\eta} z| < \tilde{m}, \text{ and } |t - s| \le r_1, \\ s + \sqrt{\eta} z, & \text{otherwise.} \end{cases} \tag{41}$$

This coupling has the following property given in lemmas 3.1 and 3.2 of Majka et al. (2020).

Lemma 28. *For the coupling defined in* (40) *and* (41)*, we have*

$$\mathbb{E}[|t' - s'|] = |t - s|,$$

and if $\eta \le 4\tilde{m}^2$, we have

$$\mathbb{E}\Big[(|t' - s'| - |t - s|)^2 \mathbb{1}_{|t'-s'| \in I_{|t-s|}}\Big] \ge \frac{1}{2} c_0 \min(\sqrt{\eta}, |t - s|)\sqrt{\eta},$$
$$\text{where} \qquad I_r = \begin{cases} (0, r + \sqrt{\eta}), & \text{if } r \le \sqrt{\eta}, \\ (r - \sqrt{\eta}, r), & \text{otherwise,} \end{cases}$$

for some universal constant $c_0 > 0$.

Thus, through the second bound, we have control of the probability that $|t' - s'|$ contracts below $|t - s|$. The coupling proposed in (39) is a multivariate analogue of this that also accounts for the diffusion coefficient $G^{1/2}$. Let the vector $e \in \mathbb{R}^d$ be as defined in (38), then we obtain that,

$$\langle e, G^+ X' \rangle = \langle e, G^+ \hat{x} \rangle + \sqrt{h} \langle (G^{1/2})^+ e, Z \rangle,$$
$$\langle e, G^+ Y_s' \rangle = \langle e, G^+ \hat{y} \rangle + \sqrt{h} \langle (G^{1/2})^+ e, Z \rangle.$$

Therefore, $\langle e, G^+ X' \rangle, \langle e, G^+ Y_s' \rangle$ are a synchronous coupling of $\mathcal{N}(\langle e, G^+ \hat{x} \rangle, h)$ and $\mathcal{N}(\langle e, G^+ \hat{y} \rangle, h)$. Furthermore, we have

$$\begin{aligned} \langle e, G^+ Y_r' \rangle &= \langle e, G^+ \hat{y} \rangle + \sqrt{h} \langle (G^{1/2})^+ e, (I - 2(G^{1/2})^+ e e^T (G^{1/2})^+) Z \rangle \\ &= \langle e, G^+ \hat{y} \rangle + \sqrt{h} \langle (G^{1/2})^+ e, Z \rangle - 2\sqrt{h} \langle e, G^+ e \rangle \langle (G^{1/2})^+ e, Z \rangle \\ &= \langle e, G^+ \hat{y} \rangle - \sqrt{h} \langle (G^{1/2})^+ e, Z \rangle, \end{aligned}$$

and so $\langle e, G^+X'\rangle, \langle e, G^+Y_r'\rangle$ is the one-dimensional reflection coupling. Finally we obtain,

$$\begin{aligned}
\frac{\phi_{\hat{y},\eta G}(X')}{\phi_{\hat{x},\eta G}(X')} &= \frac{\phi_{(G^{1/2})^+(\hat{y}-\hat{x}),\eta(G^{1/2})+G^{1/2}}(\sqrt{\eta}Z)}{\phi_{\mathbf{0},\eta(G^{1/2})^+G^{1/2}}(\sqrt{\eta}Z)} \\
&= \exp\left(-\frac{1}{2\eta}\|\sqrt{\eta}Z-(G^{1/2})^+(\hat{y}-\hat{x})\|^2_{(G^{1/2})+G^{1/2}} + \frac{1}{2\eta}\|\sqrt{\eta}Z\|^2_{(G^{1/2})+G^{1/2}}\right) \\
&= \exp\left(-\frac{1}{2\eta}\|\hat{y}-\hat{x}\|^2_{G^+} + \frac{1}{\eta}\sqrt{\eta}\langle(G^{1/2})^+(\hat{y}-\hat{x}), Z\rangle\right) \\
&= \exp\left(-\frac{1}{2\eta}|\langle e, G^+(\hat{y}-\hat{x})\rangle|^2 + \frac{1}{\eta}\sqrt{\eta}\langle e, G^+(\hat{y}-\hat{x})\rangle\langle(G^{1/2})^+e, Z\rangle\right) \\
&= \exp\left(-\frac{1}{2\eta}(\sqrt{\eta}\langle e, G^+Z\rangle - \langle e, G^+(\hat{y}-\hat{x})\rangle)^2 + \frac{|\sqrt{\eta}\langle(G^{1/2})^+e, Z\rangle|^2}{2\eta}\right) \\
&= \frac{\phi_{\langle e,G^+(\hat{y}-\hat{x})\rangle,\eta}(\sqrt{\eta}\langle(G^{1/2})^+e, Z\rangle)}{\phi_{0,\eta}(\sqrt{\eta}\langle(G^{1/2})^+e, Z\rangle)} \\
&= \frac{\phi_{\langle e,G^+\hat{y}\rangle,\eta}(\langle e, G^+X'\rangle)}{\phi_{\langle e,G^+\hat{x}\rangle,\eta}(\langle e, G^+X'\rangle)}.
\end{aligned}$$

From this, we deduce that $\langle e, G^+X'\rangle, \langle e, G^+Y'\rangle$ are coupled as in (40), (41). The equivalence follows by setting

$$t' = \langle e, G^+X'\rangle, \qquad s' = \langle e, G^+Y'\rangle \tag{42}$$

$$t = \langle e, G^+\hat{x}\rangle, \qquad s = \langle e, G^+\hat{y}\rangle, \qquad z = \langle(G^{1/2})^+e, Z\rangle. \tag{43}$$

Through this equivalence, we can extend the previous lemma to obtain the following result about the high dimensional coupling.

Lemma 29. *For the coupling defined in* (39)*, we obtain that for $\eta \le 4m^2$, we have the following:*

$$\mathbb{E}[R'] = \hat{r}, \qquad \mathbb{E}\Big[(R'-\hat{r})^2\mathbb{1}_{R'\in I_{\hat{r}}}\Big] \ge \frac{1}{2}c_0\min(\sqrt{\eta}, \hat{r})\sqrt{\eta},$$

where c_0 and I_r is as in Lemma 28.

Proof. Let $\{e_i\}_{i=1}^n$ be a basis of $\mathbb{R}^n$ with respect to the inner product $\langle\cdot,\cdot\rangle_{G^+}$ with $e_1 = e$. Then, we have that

$$\begin{aligned}
(R')^2 &= \sum_{i=1}^n |\langle e_i, G^+(X'-Y')\rangle|^2 \\
&= |t'-s'|^2 + \sum_{i=2}^n |\langle e_i, G^+(X'-Y')\rangle|^2,
\end{aligned} \tag{44}$$

where t', s' are as defined in (42). For any $i \neq 1$, we can use that $e_i \perp e$, to obtain that

$$\begin{aligned}
\langle e_i, G^+Y_r'\rangle &= \langle e_i, G^+\hat{y}\rangle + \sqrt{h}\langle e_i, e\rangle + 2\sqrt{h}\langle e_i, Z\rangle \\
&= \langle e_i, G^+\hat{y}\rangle + 2\sqrt{h}z.
\end{aligned}$$

From this, we obtain that,

$$\langle e_i, G^+(X'-Y_r')\rangle = \langle e_i, G^+(\hat{x}-\hat{y})\rangle = 0.$$

This also holds for the synchronous coupling and hence we obtain $\langle e_i, G^+(X'-Y')\rangle = 0$. Combined with (44), we obtain that $R' = |t'-s'|$. Similarly it can be shown that $\hat{r} = |t-s|$ and thus, from Lemma 28, the statement of the lemma follows. □

F.2 Proof of Proposition 27

We begin by considering the setting where Z' is truncated Gaussian noise and that $b = \tilde{b}$, $\sigma = \tilde{\sigma}$. We will then extend this to the more general setting in Section F.2.3. We begin by decomposing $\xi \sim N(0, I_d)$ in to directions parallel and perpendicular to the radial vector,

$$\xi_1 = vv^T\xi, \qquad \xi_2 = (I - vv^T)\xi, \qquad v = \frac{\tilde{x} - \tilde{y}}{\|\tilde{x} - \tilde{y}\|_{G^+}}.$$

We then clip each direction according to constants $\bar{z}_1, \bar{z}_2 > 0$ and add them together:

$$Z' = (1 \wedge \bar{z}_1\|\xi_1\|_{G^+}^{-1})\xi_1 + (1 \wedge \bar{z}_2\|\xi_2\|_{G^+}^{-1})\xi_2. \tag{45}$$

To prove that the process is contractive, we consider two cases based on the initial distance r.

F.2.1 The case of $r \geq r_1$

When r is large, we can rely on contractive properties following from the weight decay. For this, we obtain the following.

Lemma 30. *Suppose that Assumption 24 holds and that $4\bar{z}_1 \leq \lambda L_\sigma^{-1}\sqrt{\eta}, 2\bar{z}_2 \leq \sqrt{\lambda}L_\sigma, \eta \leq \lambda^{-1}$. Then whenever $r \geq 4B/\lambda$, we have*

$$\hat{r} \leq \left(1 - \frac{\eta\lambda}{8}\right)r, \tag{46}$$

and when $r < 4B/\lambda$,

$$\hat{r} \leq (1 + \eta L)r, \tag{47}$$

where $L = 2(L_b - \lambda)_+ + 4\eta^{-1/2}L_\sigma\bar{z}_1$.

Proof. From the triangle inequality and the Lipschitz property of b, we obtain

$$\begin{aligned}\tilde{r} &\leq (1 - \eta\lambda)\|x - y\|_{G^+} + \eta\|b(x) - b(y)\|_{G^+} \\ &\leq (1 + \eta(L_b - \lambda)_+)r.\end{aligned}$$

Alternatively, we can use the fact that $\|b\|_{G^+} \leq B$ to obtain, $\tilde{r} \leq (1 - \eta\lambda)r + 2\eta B$. In particular, if $r \geq 4B/\lambda$, we obtain $\tilde{r} \leq (1 - \eta\lambda/2)r$. Next we bound $\hat{r}$ using the decomposition,

$$\begin{aligned}\hat{r}^2 &= \|\tilde{x} - \tilde{y} + \sqrt{\eta}(\sigma(x) - \sigma(y))Z'\|_{G^+}^2 \\ &= \|\tilde{x} - \tilde{y} + \sqrt{\eta}(\sigma(x) - \sigma(y))(1 \wedge \bar{z}_1\|\xi_1\|_{G^+}^{-1})\xi_1\|_{G^+}^2 + \|\sqrt{\eta}(\sigma(x) - \sigma(y))(1 \wedge \bar{z}_2\|\xi_2\|_{G^+}^{-1})\xi_2\|_{G^+}^2 \\ &\leq \|\tilde{x} - \tilde{y} + \sqrt{\eta}(\sigma(x) - \sigma(y))(1 \wedge \bar{z}_1\|\xi_1\|_{G^+}^{-1})\xi_1\|_{G^+}^2 \end{aligned} \tag{48}$$

The second term is then bounded by,

$$\begin{aligned}\|\sqrt{\eta}(\sigma(x) - \sigma(y))(1 \wedge \bar{z}_2\|\xi_2\|_{G^+}^{-1})\xi_2\|_{G^+}^2 &\leq \eta\|\sigma(x) - \sigma(y)\|_{op,G^+}(1 \wedge \bar{z}_2\|\xi_2\|_{G^+}^{-1})^2\|\xi_2\|_{G^+}^2 \\ &\leq \eta L_\sigma^2\bar{z}_2^2 r^2,\end{aligned} \tag{49}$$

and the first term is bounded by,

$$\begin{aligned}\|\tilde{x} - \tilde{y} + \sqrt{\eta}(\sigma(x) - \sigma(y))(1 \wedge \bar{z}_1\|\xi_1\|_{G^+}^{-1})\xi_1\|_{G^+}^2 &\leq \tilde{r}^2 + \eta L_\sigma^2\bar{z}_1^2 r^2 + 2\sqrt{\eta}L_\sigma\langle v, G^+\xi_1\rangle\tilde{r}^2 \\ &\leq (1 + 2\sqrt{\eta}L_\sigma\bar{z}_1)\tilde{r}^2 + \eta L_\sigma^2\bar{z}_1^2 r^2.\end{aligned} \tag{50}$$

We substitute (49) and (50) in to (48) to obtain

$$\begin{aligned}\hat{r}^2 &\leq (1 + 2\sqrt{\eta}L_\sigma\bar{z}_1)\tilde{r}^2 + \eta L_\sigma^2(\bar{z}_1^2 + \bar{z}_2^2)r^2 \\ &\leq (1 + \eta(L_b - \lambda)_+)^2(1 + 2\sqrt{\eta}L_\sigma\bar{z}_1)r^2 + \eta L_\sigma^2(\bar{z}_1^2 + \bar{z}_2^2)r^2 \\ &\leq (1 + \eta(L_b - \lambda)_+ + 2\eta^{3/2}(L_b - \lambda)_+L_\sigma\bar{z}_1 + 2\sqrt{\eta}L_\sigma\bar{z}_1 + \eta L_\sigma^2(\bar{z}_1^2 + \bar{z}_2^2))r^2 \\ &\leq (1 + 2\eta(L_b - \lambda)_+ + 4\sqrt{\eta}L_\sigma\bar{z}_1)r^2,\end{aligned}$$

where we have used that $2\eta^{1/2}L_\sigma\bar{z}_1 \leq 1, \eta^{1/2}L_\sigma(z_1^2 + z_2^2) \leq \bar{z}_1$, producing the bound in (47). In the case that $r \leq 4B/\lambda$, we can use the fact that $2\eta^{1/2}L_\sigma\bar{z}_1 \leq \eta\lambda/2$ and $L_\sigma^2(\bar{z}_1^2 + \bar{z}_2^2) \leq \lambda/4$ to refine this bound:

$$\begin{aligned}\hat{r}^2 &\leq (1 - \eta\lambda/2)^2(1 + 2\sqrt{\eta}L_\sigma\bar{z}_1)r^2 + \eta L_\sigma^2(z_1^2 + z_2^2)r^2 \\ &\leq (1 - \eta\lambda/2)(1 - \eta^2\lambda^2/4)^2r^2 + \eta L_\sigma^2(z_1^2 + z_2^2)r^2 \\ &\leq (1 - \eta\lambda/4)r^2.\end{aligned}$$

Using the fact that $(1 - \eta\lambda/4)^{1/2} \leq 1 - \eta\lambda/8$, we obtain the bound in (46). □

We will also need a property of f given in Majka et al. (2020).

Lemma 31. *The function f satisfies the property that for all $r \geq r_2$,*

$$f\Big(\Big(1 - \tfrac{\eta K}{2}\Big)r\Big) - f(r) \leq -\eta c f(r).$$

Using the fact that f is increasing, it follows from lemmas 30 and 31 that,

$$f(\hat{r}) \leq f\Big(\Big(1 - \tfrac{\eta K}{2}\Big)r\Big) \leq (1 - \eta c) f(r).$$

Thus, to obtain contractions of $\mathbb{E}[f(R')]$ when $r \geq r_1$, it is sufficient to show that $\mathbb{E}[f(R')|Z'] \leq f(\hat{r})$. Note that when $\hat{r} \geq r_1$ or $\|\sqrt{\eta}Z\| \geq m$, the synchronous coupling is used and so $R' = \hat{r}$. Furthermore, if $\hat{r} < r_1$ and $\|\sqrt{\eta}Z\| < m$, we have that $R' \leq r_2$ and thus, using the concavity of f, we deduce that

$$\mathbb{E}[f(R')|Z'] - f(\hat{r}) \leq f'(\hat{r})(\mathbb{E}[R'|Z'] - \hat{r}) = 0.$$

Thus, we have shown that whenever $r \geq r_1$, $\mathbb{E}[f(R')|Z'] \leq f(\hat{r})$.

F.2.2 The case of $r < r_1$

When r is small we no longer have contractions due to weight decay and must instead rely on properties of the coupling and function. From Taylor's theorem we have the following:

$$f(R') - f(\hat{r}) = f'(\hat{r})(R' - \hat{r}) + \frac{1}{2}\sup_{\theta} f''(\theta)(R' - \hat{r})^2.$$

where the supremum is between all $\theta \geq 0$ between R' and $\hat{r}$. We note that in the present setting, $\hat{r} \leq r_1$ also (this follows from Lemma 30) and furthermore $R' - \hat{r} \leq 2m \leq r_2$. Therefore, we can use that f is concave between R' and $\hat{r}$ and so f'' is negative. Using this fact, as well as the fact that $\mathbb{E}[R'|Z'] = \hat{r}$, we obtain the bound,

$$\begin{aligned}
\mathbb{E}[f(R')|Z'] - f(\hat{r}) &\leq \frac{1}{2}\mathbb{E}\Big[\sup_{\theta} f''(\theta)(R' - \hat{r})^2 \mathbb{1}_{R' \in I_{\hat{r}}}\Big] \\
&\leq \frac{1}{2}\sup_{\theta \in I_{\hat{r}}} f''(\theta)\mathbb{E}\Big[(R' - \hat{r})^2 \mathbb{1}_{R' \in I_{\hat{r}}}\Big|Z'\Big] \\
&\leq \frac{1}{4}\sup_{\theta \in I_{\hat{r}}} f''(\theta) c_0 \min(\sqrt{\eta}, \hat{r})\sqrt{\eta}.
\end{aligned}$$

Furthermore, we analyse the contractions between $\hat{r}$ and r using the fact that the function is concave between these values, obtaining,

$$f(\hat{r}) - f(r) \leq f'(r)(\hat{r} - r) \leq f'(r)\eta L r.$$

Since we have the derivative $f'(r) = e^{-ar} = f'(\hat{r})e^{-a(r-\hat{r})} \leq f'(\hat{r})e^{a\eta L r_1}$, it holds that

$$f(\hat{r}) - f(r) \leq f'(\hat{r})e^{a\eta L r_1}\eta L \hat{r}, \tag{51}$$

where we have used that $f(\hat{r}) - f(r) \leq 0$ holds trivially whenever $r \geq \hat{r}$. Putting these together, we obtain the bound,

$$\mathbb{E}[f(R')|Z'] - f(r) \leq f'(\hat{r})e^{a\eta L r_1}\eta L\hat{r} + \frac{1}{4}\sup_{\theta \in I_{\hat{r}}} f''(\theta) c_0 \min(\sqrt{\eta}, \hat{r})\sqrt{\eta}.$$

To complete the analysis of this case, we borrow a result from Majka et al. (2020).

Lemma 32. *The function f, satisfies the property that for all $\hat{r} \in [0, r_1]$,*

$$f'(\hat{r})e^{a\eta L r_1}\eta L\hat{r} + \frac{1}{4}c_0 \min(\sqrt{\eta}, \hat{r})\sqrt{\eta}\sup_{I_{\hat{r}}} f''(\hat{r}) \leq -ch f(\hat{r}).$$

Between this section and the previous, we have shown that for any $x, y \in \mathbb{R}^n$,

$$\mathbb{E}[f(R')] \leq (1 - \eta c/2) f(r),$$

in the setting where Z' is the truncated Gaussian defined in (45) and $b = \tilde{b}, \sigma = \tilde{\sigma}$.

F.2.3 FULL NOISE AND INACCURATE DRIFT

We now consider the more general case of $b \neq \tilde{b}$, $\sigma \neq \tilde{\sigma}$ necessarily and also set $Z' = \xi$, so that it is Gaussian distributed. We do this by borrowing the contraction analysis above. We use the notation $R'' = \|X' - Y'\|_{G^+}$ to not confuse it with R' used above. We obtain,

$$\begin{aligned}
R'' &\leq R' + \eta\|b(y) - \tilde{b}(y)\|_{G^+} + \sqrt{\eta}\|(\sigma(y) - \tilde{\sigma}(y))\xi\|_{G^+} + \sqrt{\eta}\|(\sigma(x) - \sigma(y))(Z' - \xi_1 - \xi_2)\|_{G^+} \\
&\leq R' + \eta\tilde{B} + \sqrt{\eta}\tilde{B}_\sigma\|\xi\|_{G^+} \\
&\quad + \sqrt{\eta}\|\sigma(x) - \sigma(y)\|_{op,G^+}\|Z' - (1 \wedge \bar{z}_1\|\xi_1\|_{G^+}^{-1})\xi_1 - (1 \wedge \bar{z}_2\|\xi_2\|^{-1})\xi_2\|_{G^+} \\
&\leq R' + \eta\tilde{B} + \sqrt{\eta}\tilde{B}_\sigma\|\xi\|_{G^+} + \sqrt{\eta}L_\sigma r(\|\xi_1\|_{G^+}\mathbb{1}_{\|\xi_1\|_{G^+} \geq \bar{z}_1} + \|\xi_2\|_{G^+}\mathbb{1}_{\|\xi_2\|_{G^+} \geq \bar{z}_2}).
\end{aligned}$$

We use the following stability bound for the function f given in the proof of Theorem 2.5 in Majka et al. (2020).

Lemma 33. *For any $t, s \geq 0$, we have*

$$f(t) - f(s) \leq (r_2^{-1}e^{-ar_2}(t \vee s) + 1)|t - s|.$$

Thus, the difference between $f(R'')$ and $f(R')$ is given by,

$$\begin{aligned}
&f(R'') - f(R') \\
&\leq f(R' + \eta\tilde{B} + \sqrt{\eta}\tilde{B}_\sigma\|\xi\|_{G^+} + \sqrt{\eta}L_\sigma r(\|\xi_1\|_{G^+}\mathbb{1}_{\|\xi_1\|_{G^+} \geq \bar{z}_1} + \|\xi_2\|_{G^+}\mathbb{1}_{\|\xi_2\|_{G^+} \geq \bar{z}_2})) - f(R') \\
&\leq (r_2^{-1}e^{-ar_2}(R' + \eta\tilde{B} + \sqrt{\eta}\tilde{B}_\sigma\|\xi\|_{G^+} + \sqrt{\eta}L_\sigma r(\|\xi_1\|_{G^+}\mathbb{1}_{\|\xi_1\|_{G^+} \geq \bar{z}_1} + \|\xi_2\|_{G^+}\mathbb{1}_{\|\xi_2\|_{G^+} \geq \bar{z}_2})) \\
&\quad + 1)(\eta\tilde{B} + \sqrt{\eta}\tilde{B}_\sigma\|\xi\|_{G^+} + \sqrt{\eta}L_\sigma r(\|\xi_1\|_{G^+}\mathbb{1}_{\|\xi_1\|_{G^+} \geq \bar{z}_1} + \|\xi_2\|_{G^+}\mathbb{1}_{\|\xi_2\|_{G^+} \geq \bar{z}_2})). \qquad (52)
\end{aligned}$$

We now control the expected value of this. Using concentration of the χ^2 distribution (see Example 2.11 of Wainwright (2019)), we obtain that for any $\bar{z}_1 = \sqrt{2\lambda_{\text{gap}}(G)^{-1}(n-1)}$,

$$\begin{aligned}
&\mathbb{E}[\|\xi_2\|_{G^+}^2\mathbb{1}_{\|\xi_2\|_{G^+} \geq \bar{z}_2}] \\
&\quad \leq \lambda_{\text{gap}}(G)^{-1}\mathbb{E}[\|\xi_2\|^2\mathbb{1}_{\|\xi_2\| \geq \lambda_{\text{gap}}(G)^{1/2}\bar{z}_2}] \\
&\quad \leq \lambda_{\text{gap}}(G)^{-1}\int_{\lambda_{\text{gap}}(G)\bar{z}_2^2}^{\infty}\mathbb{P}(\|\xi_2\|^2 \geq r)\,dr \\
&\qquad + \lambda_{\text{gap}}(G)^{-1}\int_0^{\lambda_{\text{gap}}(G)\bar{z}_2^2}\mathbb{P}(\|\xi_2\| \geq \bar{z}_2)\,dr \\
&\quad \leq \lambda_{\text{gap}}(G)^{-1}\int_{\lambda_{\text{gap}}(G)\bar{z}_2^2}^{\infty}\exp\Big(-\frac{(r-(n-1))^2}{8n}\Big)dr \\
&\qquad + \lambda_{\text{gap}}(G)^{-1}\exp\Big(-\frac{(\lambda_{\text{gap}}(G)\bar{z}_2^2 - (n-1))^2}{8n}\Big)\bar{z}^2 \\
&\quad \leq \lambda_{\text{gap}}(G)^{-1}(\sqrt{8(n-1)\pi} + \lambda_{\text{gap}}(G)\bar{z}_2^2)\exp\Big(-\frac{(\lambda_{\text{gap}}(G)\bar{z}_2^2 - (n-1))^2}{8(n-1)}\Big) \\
&\quad \leq \lambda_{\text{gap}}(G)^{-1}\Big(\sqrt{8(n-1)\pi}\exp(-(n-1)/16) \\
&\qquad + \lambda_{\text{gap}}(G)\bar{z}_2^2\exp(-\bar{z}_2^4/64)\Big)\exp\Big(-\frac{(\lambda_{\text{gap}}(G)\bar{z}_2^2 - (n-1))^2}{16(n-1)}\Big) \\
&\quad \leq \kappa_0\lambda_{\text{gap}}(G)^{-1}\exp\Big(-\frac{(\lambda_{\text{gap}}(G)\bar{z}_2^2 - (n-1))^2}{16(n-1)}\Big) \\
&\quad \leq \kappa_0\lambda_{\text{gap}}(G)^{-1}\exp\Big(-\frac{n-1}{16}\Big),
\end{aligned}$$

for some universal constant $\kappa_0 \geq 1$ (independent of n and $\bar{z}$). Similarly, we have

$$\mathbb{E}[\|\xi_1\|_{G^+}^2\mathbb{1}_{\|\xi_1\|_{G^+} \geq \bar{z}_1}] \leq \kappa_0\lambda_{\text{gap}}(G)^{-1}\exp\Big(-\frac{(\lambda_{\text{gap}}(G)\bar{z}_1^2 - 1)^2}{16}\Big),$$

for any $\bar{z}_1 \geq \lambda_{\text{gap}}(G)^{-1/2}$. Therefore, we choose $\bar{z}_1 = \frac{\lambda}{4} L_\sigma^{-1}\sqrt{\eta}$We now return to (52) using these bounds as well as the fact that $\mathbb{E}[R'|Z'] = \hat{r}$. Defining the quantity,

$$A := \kappa_0^{1/2}\lambda_{\text{gap}}(G)^{-1/2}\exp(-(n-1)/32) + \kappa_0^{1/2}\lambda_{\text{gap}}(G)^{-1/2}\exp(-(\lambda_{\text{gap}}(G)\bar{z}_1^2 - 1)^2/32),$$

we obtain that for $\eta \leq \min\{\tilde{B}/2, d\tilde{B}_\sigma^2/4, 1/2L, 1/2L_\sigma^2 A^2\}$

$$\begin{aligned}
&\mathbb{E}[f(R'') - f(R')] \\
&\quad = (r_2^{-1}e^{-ar_2}(\mathbb{E}[\hat{r}^2]^{1/2} + \eta\tilde{B} + \sqrt{\eta d}\tilde{B}_\sigma + \sqrt{\eta}L_\sigma r A) + 1)(\eta\tilde{B} + \sqrt{\eta d}\tilde{B}_\sigma + \sqrt{\eta}L_\sigma r A) \\
&\quad \leq (r_2^{-1}e^{-ar_2}(1 + \eta L + \sqrt{\eta}L_\sigma A)r + 1)\sqrt{\eta}L_\sigma r A + \frac{1}{r_2}e^{-ar_2}(\eta^2\tilde{B}^2 + \eta d\tilde{B}_\sigma^2) \\
&\qquad + (r_2^{-1}e^{-ar_2}(1 + \eta L + \sqrt{\eta}L_\sigma A)r + 1)(\eta\tilde{B} + \sqrt{\eta d}\tilde{B}_\sigma) \\
&\qquad + r_2^{-1}e^{-ar_2}(\eta\tilde{B} + \sqrt{\eta d}\tilde{B}_\sigma)\sqrt{\eta}L_\sigma r^2 A \\
&\quad \leq (4r_2^{-1}e^{-ar_2}r + 1)\sqrt{\eta}L_\sigma r A + \frac{3}{2r_2}e^{-ar_2}(\eta^2\tilde{B}^2 + \eta d\tilde{B}_\sigma^2).
\end{aligned}$$

When $r \leq r_2$, we have

$$\begin{aligned}
&(r_2^{-1}e^{-ar_2}(1 + \eta L + 3\sqrt{\eta}L_\sigma A)r + 1)r \\
&\qquad \leq (e^{-ar_2}(1 + \eta L + \sqrt{\eta}L_\sigma A) + 1)r \\
&\qquad \leq (e^{-ar_2}(1 + \eta L + \sqrt{\eta}L_\sigma A) + 1)(a^{-1}(1 - e^{-ar_2}))^{-1}a^{-1}(1 - e^{-ar}) \\
&\qquad \leq 4(a^{-1}(1 - e^{-ar_2}))^{-1}f(r),
\end{aligned}$$

where in the final line, we used $\eta \leq L^{-1}$ and $\sqrt{\eta}L_\sigma\kappa_0^{1/2} \leq 1$. When $r > r_2$, we have

$$\begin{aligned}
(r_2^{-1}e^{-ar_2}(1 + \eta L + \sqrt{\eta}L_\sigma A)r + 1)r &\leq (e^{-ar_2}(1 + \eta L + \sqrt{\eta}L_\sigma A) + 1)r_2^{-1}r^2 \\
&\leq 2(2 + e^{ar_2})f(r).
\end{aligned}$$

Thus, we obtain,

$$\begin{aligned}
\mathbb{E}[f(R'')] &\leq \mathbb{E}[f(R')] + \sqrt{\eta}L_\sigma A\tfrac{6\vee(4a)}{1-e^{-ar_2}}f(r) + \frac{1}{2r_2}e^{-ar_2}(\eta^2\tilde{B}^2 + \eta d\tilde{B}_\sigma^2) \\
&\leq (1 - \eta c/2 + \sqrt{\eta}L_\sigma A\tfrac{6\vee(4a)}{1-e^{-ar_2}})f(r) + \frac{3}{2r_2}e^{-ar_2}(\eta^2\tilde{B}^2 + \eta d\tilde{B}_\sigma^2),
\end{aligned}$$

where we used $AL_\sigma\frac{6\vee(4a)}{1-e^{-ar_2}} \leq \sqrt{\eta}c/4$.

G Proofs for the stability of the noisy gradient estimator

Using the Wasserstein contraction obtained in the previous section, we will now prove Proposition 14.

Proposition 14. *Consider the score matching algorithm $A_{\text{sm}} : S \mapsto s_{\theta_K}$ for some fixed $K \in \mathbb{N}$ where $(\theta_k)_k$ is as given in* (16)*. Suppose that assumptions 10, 12 and 13 hold, then there exists some $\bar{\eta} > 0$ such that, if $\sup_p \eta_p \leq \bar{\eta}$, we obtain that A_{sm} is score stable with constant*

$$\varepsilon_{stab}^2 \lesssim \frac{\overline{L}^2 C^2(P+n)}{\lambda_{gap}N}\min\left\{\frac{\eta_{\min}\lambda_{gap}\lambda^2}{PN_BC}\sum_{k=0}^{K-1}\eta_k, \exp\left(\tilde{c}\frac{PN_BC}{\eta_{\min}\lambda_{gap}\lambda^2}\right)\right\},$$

where $\tilde{c} \lesssim (\overline{M}_4 B_\ell C_\tau^{1/2} + \overline{L}_4^2)(PN_B\lambda_{gap})^{-1/2} \vee 1$, $\eta_{\min} = \min_k \eta_k$.

The proof of Proposition follows from an application of Proposition 27 to the process in (16). Similar to the proof of Proposition 11, we obtain stability estimates by analysing the trajectories θ_k and $\tilde{\theta}_k$ trained on S and S^N with coupled minibatch indices. In particular, given a set of minibatch indices $B \subset [N]$ with $|B| = N_B$, if we set

$$\begin{aligned}
b(\theta) &:= \mathbb{E}\Big[\text{Clip}_C(G(\theta, (x_i)_{i\in B}))\Big|\theta, B, S\Big], \qquad \tilde{b}(\theta) := \mathbb{E}\Big[\text{Clip}_C(G(\theta, (\tilde{x}_i)_{i\in B}))\Big|\theta, B, S^N\Big] \\
\sigma(\theta) &:= \sqrt{\eta}\,\Sigma_S(\theta, B)^{1/2}, \qquad \tilde{\sigma}(\theta) := \sqrt{\eta}\,\Sigma_{S^N}(\theta, B)^{1/2},
\end{aligned}$$

where we use $(\tilde{x}_i)_{i=1}^N$ to denote the dataset S^N (i.e. $\tilde{x}_i = x_i$ for all $i \neq N$ and $\tilde{x}_N = \tilde{x}$), then the trajectories θ_k and $\tilde{\theta}_k$ are updated as in (36), (37). Using the shorthand, $v_{i,j}(\theta) = w_{t_{(i,j)}} \nabla_\theta \|s_\theta(X_{(i,j)}, t_{(i,j)}) - \nabla \log p_{t_{(i,j)}|0}(X_{(i,j)}|x_i)\|^2$, we obtain the bound,

$$\begin{aligned}\Sigma_S(\theta, B) &\succcurlyeq \mathrm{Cov}\left(\frac{1}{PN_B}\sum_{i\in B}\sum_{j=1}^{P}\mathrm{Clip}_C(v_{i,j}(\theta))\middle|\theta, B, S\right)\\ &\succcurlyeq \frac{1}{P}\frac{1}{N_B^2}\sum_{i\in B}\mathrm{Cov}\left(\mathrm{Clip}_C(v_{i,j}(\theta))\middle|\theta, B, S\right)\\ &\succcurlyeq \frac{1}{PN_B}\bar{\Sigma}.\end{aligned}$$

Therefore, we have $\sigma(\theta) \succcurlyeq \sqrt{\eta/PN_B}\,\bar{\Sigma}^{1/2} =: G^{1/2}$, and similarly, $\tilde{\sigma}(\theta) \succcurlyeq G^{1/2}$. The weighted norm $\|\cdot\|_{G^+}$ satisfies the property,

$$\|\theta\|_{G^+} \leq \lambda_{\max}(G^+)^{1/2}\|\theta\| \leq \sqrt{\frac{PN_B}{\eta\lambda_{\text{gap}}}}\|\theta\|$$

Therefore, due to the gradient clipping, we have $\|b(\theta)\|_{G^+} \leq \sqrt{PN_B/\eta\lambda_{\text{gap}}}C =: B$. Furthermore, by Assumption 13, we apply the same argument used in the proof of Proposition 11 to obtain

$$\|b(\theta) - b(\theta')\|_{G^+} \leq (\overline{M}_4 B_\ell C_\tau^{1/2} + \overline{L}_4^2)\|\theta - \theta'\|_{G^+}$$

so $L_b = \overline{M}_4 B_\ell C_\tau^{1/2} + \overline{L}_4^2$. To obtain the Lipschitz constant for the volatility matrix, we first obtain,

$$\begin{aligned}\sigma(\theta) - \sigma(\theta') \preccurlyeq \sqrt{\eta}\,\mathrm{Cov}\Bigg(\frac{1}{PN_B}\sum_{i\in B}\sum_{j=1}^{P}\Big((1 \vee (C\|v_{i,j}(\theta)\|^{-1}))v_{i,j}(\theta)\\ - (1 \vee (C\|v_{i,j}(\theta')\|^{-1}))v_{i,j}(\theta')\Big)\Bigg|\theta, B, S\Bigg)^{1/2}.\end{aligned}$$

From this, we deduce,

$$\begin{aligned}\|\sigma(\theta) - \sigma(\theta')\|_{op,G^+} &\leq \sqrt{\eta}\sup_{\|v\|_{G^+}=1}\mathrm{Var}\Big(\Big\langle G^+ v, \frac{1}{PN_B}\sum_{i\in B}\sum_{j=1}^{P}\Big((1 \vee (C\|v_{i,j}(\theta)\|^{-1}))v_{i,j}(\theta)\\ &\qquad - (1 \vee (C\|v_{i,j}(\theta')\|^{-1}))v_{i,j}(\theta')\Big)\Big\rangle\Big|\theta, B, S\Big)^{1/2}\\ &\leq \sqrt{\eta}\Bigg(\frac{1}{PN_B^2}\sum_{i\in B}\mathrm{Var}\left(\|v_{i,j}(\theta) - v_{i,j}(\theta')\|_{G^+}\middle|\theta, B, S\right)\Bigg)^{1/2}.\end{aligned}$$

To control this further, we use the Lipschitz assumption on to show that υ is Lipschitz also:

$$\begin{aligned}&\|v_{i,j}(\theta) - v_{i,j}(\theta')\|_{G^+}\\ &\quad\leq 2\|s_\theta(X_{(i,j)}, t_{(i,j)}) - s_{\theta'}(X_{(i,j)}, t_{(i,j)})\|_{G^+}\|\nabla_\theta s_\theta(X_{(i,j)}, t_{(i,j)})\|_{op,G^+}\\ &\qquad + 2\|s_\theta(X_{(i,j)}, t_{(i,j)}) - \nabla \log p_{t_{(i,j)}|0}(X_{(i,j)}|x_i)\|_{G^+}\|\nabla_\theta s_\theta(X_{(i,j)}, t_{(i,j)})\\ &\qquad - \nabla_\theta s_{\theta'}(X_{(i,j)}, t_{(i,j)})\|_{op,G^+}\\ &\quad\leq 2L(X_{(i,j)}, t_{(i,j)})^2\|\theta - \theta'\|_{G^+} + \frac{2c^{1/2}\mu_t}{\sigma_t^2}M(X_{(i,j)}, t_{(i,j)})\|\theta - \theta'\|_{G^+}.\end{aligned}$$

Computing the variance of this leads to the bound,

$$\|\sigma(\theta) - \sigma(\theta')\|_{op,G^+} \leq 2\sqrt{\frac{\eta}{PN_B\lambda_{\text{gap}}}}(\overline{M}_4 B_\ell C_\tau^{1/2} + \overline{L}_4^2)\|\theta - \theta'\|_{G^+} =: L_\sigma\|\theta - \theta'\|_{G^+}.$$

Next, we use a similar argument to the proof of Proposition 11 to obtain

$$\|b(\theta) - \tilde{b}(\theta)\|_{G^+} \leq \sqrt{\frac{PN_B}{\eta\lambda_{\text{gap}}}}\|b(\theta) - \tilde{b}(\theta)\| \leq \sqrt{\frac{PN_B}{\eta\lambda_{\text{gap}}}}\frac{2C}{N_B}\mathbb{1}_{N\in B} =: \tilde{B}_b.$$

$$\|\sigma(\theta)-\tilde{\sigma}(\theta')\|_{op,G^+} \le \sqrt{\eta}\left(\frac{1}{PN_B^2}\sum_{i\in B}\mathrm{Var}\left(\|\mathrm{Clip}_C(v_{i,j}(\theta))-\mathrm{Clip}_C(v_{i,j}(\theta'))\|_{G^+}\Big|\theta,B,S\right)\right)^{1/2}$$
$$\le \sqrt{\frac{\eta}{PN_B^2}}\,\mathrm{Var}\left(\|\mathrm{Clip}_C(v_{N,j}(\theta))-\mathrm{Clip}_C(v_{N,j}(\theta'))\|_{G^+}\Big|\theta,B,S\right)^{1/2}\mathbb{1}_{N\in B}$$

Since $0 \le \|\mathrm{Clip}_C(v_{i,N}(\theta))-\mathrm{Clip}_C(\tilde{v}_{i,N}(\theta'))\|_{G^+} \le 2C\sqrt{\frac{PN_B}{\eta\lambda_{\text{gap}}}}$

$$\|\sigma(\theta)-\tilde{\sigma}(\theta')\|_{op,G^+} \le \sqrt{\frac{\eta}{PN_B^2}}\sqrt{\frac{PN_B}{\eta\lambda_{\text{gap}}}}C\mathbb{1}_{N\in B}$$
$$\le \sqrt{\frac{1}{N_B\lambda_{\text{gap}}}}C\mathbb{1}_{N\in B}$$
$$=: \tilde{B}_\sigma.$$

Therefore we have satisfied all assumptions of Proposition 27 aside from Assumption 26. To satisfy this assumption we use that $L_\sigma \sim \sqrt{\eta/P}$ and so if η is sufficiently small, or P is sufficiently large, this assumption is satisfied once n is sufficiently large also.

Using Proposition 27, we obtain the contraction,

$$\mathbb{E}[d(\theta_{k+1},\tilde{\theta}_{k+1})|\theta_k,\tilde{\theta}_k,B_k] \le (1-\eta c/2)d(\theta_k,\tilde{\theta}_k)+\frac{3}{2r_2}e^{-ar_2}\left(\eta\tfrac{4PC^2}{\lambda_{\text{gap}}N_B}\mathbb{1}_{N\in B}+\tfrac{\eta n}{N_B\lambda_{\text{gap}}}C^2\mathbb{1}_{N\in B}\right).$$

Using the fact that $\mathbb{P}(N\in B_k)=N_B/N$, we obtain,

$$\mathbb{E}[d(\theta_{k+1},\tilde{\theta}_{k+1})] \le (1-\eta c/2)\mathbb{E}[d(\theta_k,\tilde{\theta}_k)]+\eta\frac{3}{2r_2}e^{-ar_2}\left(\frac{4PC^2}{\lambda_{\text{gap}}}+\frac{n}{\lambda_{\text{gap}}}C^2\right)\frac{1}{N}.$$

Thus, by comparison, we obtain the bound,

$$\mathbb{E}[d(\theta_K,\tilde{\theta}_K)] \le \frac{3}{2r_2}e^{-ar_2}\left(\frac{4PC^2}{\lambda_{\text{gap}}}+\frac{n}{\lambda_{\text{gap}}}C^2\right)\frac{1}{N}\eta\sum_{k=0}^{K-1}(1-\eta c/2)^k$$
$$= \frac{3}{2r_2}e^{-ar_2}\left(\frac{4PC^2}{\lambda_{\text{gap}}}+\frac{n}{\lambda_{\text{gap}}}C^2\right)\frac{1-(1-\eta c/2)^K}{Nc/2}$$
$$\le \frac{3}{2r_2}e^{-ar_2}(4P+n)\frac{C^2}{\lambda_{\text{gap}}N}(\eta K\wedge 2/c).$$

By the definition of $f(r)$, we have that it dominates r^2 up to a multiplicative constant:

$$f(r) \ge \left(\left(\frac{1}{a}(1-e^{-ar_2})\right)\wedge\left(\frac{1}{2r_2}e^{-ar_2}\right)\right)r^2$$
$$\ge \frac{1}{2r_2}e^{-ar_2}\left(\left(\frac{2r_2}{a}(e^{ar_2}-1)\right)\wedge 1\right)r^2$$
$$\ge \frac{1}{2r_2}e^{-ar_2}((2r_2^2)\wedge 1)r^2.$$

Thus, using assumption 13, it follows that

$$\int \mathbb{E}[\|s_{\theta_K}(X_t,t)-s_{\tilde{\theta}_K}(X_t,t)\|^2|X_0=\tilde{x},S]\tau(dt)$$
$$\le \bar{L}^2\mathbb{E}\left[\|\theta_K-\tilde{\theta}_K\|_{G^+}^2\right]$$
$$\le \bar{L}^2\left(\frac{1}{2r_2}e^{-ar_2}((2r_2^2)\wedge 1)\right)^{-1}\mathbb{E}\left[d(\theta_K,\tilde{\theta}_K)\right]$$
$$\le 3\bar{L}^2((2r_2^2)^{-1}\vee 1)(4P+n)\frac{C^2}{\lambda_{\text{gap}}N}(\eta K\wedge 2/c).$$

We then use the fact that when η is sufficiently small, we obtain the estimate $\eta_0 \gtrsim \lambda^{-1}$ and therefore,

$$r_1^2, r_2^2 \geq \frac{PN_B C}{\eta \lambda_{\text{gap}} \lambda^2}, \qquad L \lesssim (\overline{M}_4 B_\ell C_\tau^{1/2} + \overline{L}_4^2)(PN_B \lambda_{\text{gap}})^{-1/2} \vee 1$$

and since L and r_1 explode as $\eta \to 0^+$, we also have,

$$\begin{aligned} r_2^2 c &\gtrsim L^2 r_1^4 \exp(-6Lr_1^2/c_0) \\ &\gtrsim \exp(-6Lr_1^2/c_0). \end{aligned}$$