

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,

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

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

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

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

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

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

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

$$= \nabla \log p_t(x). \quad (19)$$

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} . \square

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

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

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

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

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

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

Once we note that,

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

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

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

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

where

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

and $\hat{s}^*(x, t) = \nabla \hat{p}_t(x)$. 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 $\hat{\ell}_{\text{dsm}}(s; S, \tau)$ is identical, up to a constant, to the objective*

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

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 $t \in \text{supp}(\tau)$.

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 $s^*(x, t) = \nabla \log p_t(x)$ satisfies $\hat{s}^* \in L^0(\mathbb{R}^d \times [0, T]; \mathbb{R}^d)$ as well as,

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

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

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 $\tau_A = \inf_{p \in A} d(p, \text{Med}(A))$, where we define the set,*

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

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*

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

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 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 $\hat{\ell}_{\text{dsm}}(s; S)$ in place of $\hat{\ell}_{\text{dsm}}(s; S, \tau)$.

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

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

Furthermore, it holds that

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

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

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

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

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

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

$$\left| \mathbb{E}[\ell_{\text{dsm}}(\hat{s}; \tau)]^{1/2} - \mathbb{E}[\hat{\ell}_{\text{dsm}}(\hat{s}; S)]^{1/2} \right| \quad (26)$$

$$\begin{aligned} &\leq \mathbb{E}\left[\inf_{(\hat{s}, \hat{s}^i) \in \Gamma_i} \frac{1}{N} \sum_{i=1}^N \int \mathbb{E}[\|\hat{s}(X_t, t) - \hat{s}^i(X_t, t)\|^2 | X_0 = x_i, S] \tau(dt)\right]^{1/2} \\ &\leq \varepsilon_{\text{stab}}, \end{aligned} \quad (27)$$

completing the proof of the bound in (24).

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

$$\begin{aligned}\mathbb{E}[\ell_{\text{dsm}}(\hat{s}; \tau)] &= \mathbb{E}[\hat{\ell}_{\text{dsm}}(\hat{s}; S)] + \mathbb{E}[\ell_{\text{dsm}}(\hat{s}; \tau) - \hat{\ell}_{\text{dsm}}(\hat{s}; S)] + \mathbb{E}[\hat{\ell}_{\text{dsm}}(\nabla \log \hat{p}_t; S)] \\ &\quad - \ell_{\text{dsm}}(\nabla \log p_t; \tau).\end{aligned}\quad (28)$$

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

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

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

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

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

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 $\hat{\ell}_{\text{dsm}}(\cdot; S, \tau)$ 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, $\hat{s} = A_{\text{erm}}(S)$ and $\hat{s}^i = A_{\text{erm}}(S)$ are close in L^2 space, averaged over the full dataset. The first half of this proof utilises the fact that $\hat{\ell}_{\text{dsm}}$ 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 $\varepsilon_{\text{stab}}$, then for any $i \in [N]$, we obtain,

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

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

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

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

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

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

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

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

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

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

where we recall that $\varepsilon_{\text{stab}}$ 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 $\mathbb{E}[\hat{\ell}_{\text{dsm}}(\hat{s}^i; \{\tilde{x}\})] = \mathbb{E}[\hat{\ell}_{\text{dsm}}(\hat{s}; S)]$ and $\mathbb{E}[\hat{\ell}_{\text{dsm}}(\hat{s}^i; \{x_i\})] =$

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

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

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

$$\begin{aligned} \mathbb{E}[\hat{\ell}_{\text{dsm}}(\hat{s})] &= \mathbb{E}[\hat{\ell}_{\text{sm}}(\hat{s})] + \mathbb{E}[\hat{C}_{\text{sm}}] \\ &\leq \mathbb{E}[\hat{\ell}_{\text{sm}}(\hat{s})] + C_{\text{sm}}, \end{aligned}$$

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

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

□

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 $L^2(p_{t|0}(\cdot|\hat{x}))$ 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 $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$, 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 $i \in [N], r \in (0, \tau_{\text{reach}}]$ and any decreasing function $\phi : (0, \infty) \rightarrow \mathbb{R}_+$, we have the bound*

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

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 $\mu = \nu_{\text{data}}(B_r(x_i))$. To utilise concentration of the empirical process, we first rewrite the probability

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

Therefore, by Chernoff's inequality we obtain

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

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,

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

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

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 $\epsilon := \inf \text{supp}(\tau) \in (0, \tau_{\text{reach}}^2)$, then for any $c \in (0, 1)$ and sufficiently large N , the score matching algorithm A_{erm} is score stable with,

$$\varepsilon_{\text{stab}}^2 \lesssim C(CC_{\text{sm}}N^{-2} + \mathbb{E}[\hat{\ell}_{\text{sm}}(\hat{s})])^c, \quad C = \frac{D_{\mathcal{H}}^2}{\sigma_\epsilon^4} \vee \frac{1}{c_\nu \sigma_\epsilon^{d^*}}.$$

Proof. We use the shorthand $\hat{\ell}_{\text{sm}}(s) = \hat{\ell}_{\text{sm}}(s; S, \tau)$, $\hat{\ell}_{\text{dsm}}(s) = \hat{\ell}_{\text{dsm}}(s; S, \tau)$, $\ell_{\text{sm}}(s) = \ell_{\text{sm}}(s; \tau)$ 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $\sigma_\epsilon \leq \tau_{\text{reach}}$ we obtain,

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

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

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

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

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

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

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

Optimising γ leads to the bound,

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

Optimising p , we obtain

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

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. \square

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 Potapchik et al. (2024), they consider the following discretisation scheme, based on the scheme of (Benton et al., 2024):

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

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 Potapchik et al. (2024).

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

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

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 $\hat{s} = A_{\text{erm}}(S)$. We begin with Lemma 22 which provides the bound,

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

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,

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

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 $\sigma_{\epsilon(1+\kappa)}^2 \geq \epsilon(1 + \kappa)$ and thus we obtain,

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

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

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 $\kappa \rightarrow 0^+$. 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 $\lim_{\kappa \rightarrow 0^+} B_{\kappa} < \infty$.

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

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

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

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

Combining these leads to the bound in the statement. \square

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

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

Proof. We begin with the bound,

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

By comparison, this leads to the bound

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

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,

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

□

We are now ready to prove Proposition [11](#)

Proposition [11](#). 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 [\(14\)](#). 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,

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

where $R^2 = \mathbb{E}[\|\theta_0\|^2]$, $v = (\bar{M} B_\ell C_\tau^{1/2} + \bar{L}^2 - \lambda) \vee 0$ 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 :

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

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,

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

We control the expectation of this by first noting that,

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

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

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

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,

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

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 $k_0 \leq k$,

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

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

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

From this we obtain the following:

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

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

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

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

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

Finally, we obtain score stability using the fact that

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

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,

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

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

for some $b, \tilde{b} : \mathbb{R}^d \rightarrow \mathbb{R}^d$, $\sigma, \tilde{\sigma} : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ 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,

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

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

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

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

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

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

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

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, $\|b(x) - b(y)\|_{G^+} \leq L_b \|x - y\|_{G^+}$ and $\|\sigma(x) - \sigma(y)\|_{op, G^+} \leq L_\sigma \|x - y\|_{G^+}$ for all $x, y \in \mathbb{R}^n$ and for some $L_b, L_\sigma \geq 0$.

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 $\|b(x) - \tilde{b}(x)\|_{G^+} \leq \tilde{B}_b, \|\sigma(x) - \tilde{\sigma}(x)\|_{op, G^+} \leq \tilde{B}_\sigma$ for all $x \in \mathbb{R}^n$ and for some $\tilde{B}_b, \tilde{B}_\sigma \geq 0$.

We define the objects,

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

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,

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

where $a = 6L_b r_1 / c_0$, $r_1 = 4(1 + \eta_0 L_b)B/\lambda$, $r_2 = r_1 + \sqrt{\eta_0}$ 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:

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

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 $\eta \leq \eta_0$ and $x, y \in \mathbb{R}^d$, it holds that

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

where

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

$$\eta_0 := \min \left\{ \frac{\lambda}{4L^2}, \frac{16}{\lambda}, \frac{1}{2L}, \frac{2c_0 \log(3/2)\lambda^2}{432L^2B^2}, \frac{4B^2}{\lambda^2}, \frac{c_0^2(\log(2))^2\lambda^2}{2304L^2B^2} \right\},$$

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 $z \sim \mathcal{N}(0, 1)$,

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

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

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 $\eta \leq 4\tilde{m}^2$, we have

$$\mathbb{E}[(|t' - s'| - |t - s|)^2 \mathbb{1}_{|t' - s'| \in I_{|t-s|}}] \geq \frac{1}{2}c_0 \min(\sqrt{\eta}, |t - s|)\sqrt{\eta},$$

$$\text{where } I_r = \begin{cases} (0, r + \sqrt{\eta}), & \text{if } r \leq \sqrt{\eta}, \\ (r - \sqrt{\eta}, r), & \text{otherwise,} \end{cases}$$

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,

$$\begin{aligned} \langle e, G^+ X' \rangle &= \langle e, G^+ \hat{x} \rangle + \sqrt{h} \langle (G^{1/2})^+ e, Z \rangle, \\ \langle e, G^+ Y'_s \rangle &= \langle e, G^+ \hat{y} \rangle + \sqrt{h} \langle (G^{1/2})^+ e, Z \rangle. \end{aligned}$$

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,

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

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

$$t' = \langle e, G^+ X' \rangle, \quad s' = \langle e, G^+ Y_r' \rangle \quad (42)$$

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

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 $\eta \leq 4m^2$, we have the following:*

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

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

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

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 $\langle e_i, G^+(X' - Y_r') \rangle = 0$. 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. \square

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,

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

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

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

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

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

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

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

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,

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

The second term is then bounded by,

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

and the first term is bounded by,

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

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 $2\eta^{1/2} L_\sigma \bar{z}_1 \leq 1$, $\eta^{1/2} L_\sigma (\bar{z}_1^2 + \bar{z}_2^2) \leq \bar{z}_1$, 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:

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

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

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 $r \geq r_2$,*

$$f\left(\left(1 - \frac{\eta K}{2}\right)r\right) - f(r) \leq -\eta c f(r).$$

Using the fact that f is increasing, it follows from lemmas [30](#) and [31](#) that,

$$f(\hat{r}) \leq f\left(\left(1 - \frac{\eta K}{2}\right)r\right) \leq (1 - \eta c)f(r).$$

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,

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

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

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

Since we have the derivative $f'(r) = e^{-ar} = f'(\hat{r})e^{-a(r-\hat{r})} \leq f'(\hat{r})e^{a\eta Lr_1}$, it holds that

$$f(\hat{r}) - f(r) \leq f'(\hat{r})e^{a\eta Lr_1}\eta L\hat{r}, \tag{51}$$

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,

$$\mathbb{E}[f(R')|Z'] - f(r) \leq f'(\hat{r})e^{a\eta Lr_1}\eta L\hat{r} + \frac{1}{4} \sup_{\theta \in I_{\hat{r}}} f''(\theta) c_0 \min(\sqrt{\eta}, \hat{r}) \sqrt{\eta}.$$

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 $\hat{r} \in [0, r_1]$,*

$$f'(\hat{r})e^{a\eta Lr_1}\eta L\hat{r} + \frac{1}{4} c_0 \min(\sqrt{\eta}, \hat{r}) \sqrt{\eta} \sup_{I_{\hat{r}}} f''(\hat{r}) \leq -chf(\hat{r}).$$

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,

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

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,

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

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 $\bar{z}_1 = \sqrt{2\lambda_{\text{gap}}(G)^{-1}(n-1)}$,

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

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

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

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

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

When $r \leq r_2$, we have

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

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,

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

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

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

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

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

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

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

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 $\hat{\theta}_k$ 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,

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

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 I3, we apply the same argument used in the proof of Proposition I1 to obtain

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

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

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

From this, we deduce,

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

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

$$\begin{aligned} \|v_{i,j}(\theta) - v_{i,j}(\theta')\|_{G^+} &\leq 2 \|s_{\theta}(X(i,j), t(i,j)) - s_{\theta'}(X(i,j), t(i,j))\|_{G^+} + \|\nabla_{\theta} s_{\theta}(X(i,j), t(i,j))\|_{op, G^+} \\ &\quad + 2 \|s_{\theta}(X(i,j), t(i,j)) - \nabla \log p_{t(i,j)|0}(X(i,j)|x_i)\|_{G^+} + \|\nabla_{\theta} s_{\theta}(X(i,j), t(i,j)) \\ &\quad - \nabla_{\theta} s_{\theta'}(X(i,j), t(i,j))\|_{op, G^+} \\ &\leq 2L(X(i,j), t(i,j))^2 \|\theta - \theta'\|_{G^+} + \frac{2c^{1/2} \mu_t}{\sigma_t^2} M(X(i,j), t(i,j)) \|\theta - \theta'\|_{G^+}. \end{aligned}$$

Computing the variance of this leads to the bound,

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

Next, we use a similar argument to the proof of Proposition I1 to obtain

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

$$\begin{aligned}
\|\sigma(\theta) - \tilde{\sigma}(\theta')\|_{op, G^+} &\leq \sqrt{\eta} \left(\frac{1}{PN_B^2} \sum_{i \in B} \text{Var} \left(\|\text{Clip}_C(v_{i,j}(\theta)) - \text{Clip}_C(v_{i,j}(\theta'))\|_{G^+} \middle| \theta, B, S \right) \right)^{1/2} \\
&\leq \sqrt{\frac{\eta}{PN_B^2}} \text{Var} \left(\|\text{Clip}_C(v_{N,j}(\theta)) - \text{Clip}_C(v_{N,j}(\theta'))\|_{G^+} \middle| \theta, B, S \right)^{1/2} \mathbb{1}_{N \in B}
\end{aligned}$$

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

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

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,

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

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

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

Thus, by comparison, we obtain the bound,

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

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

$$\begin{aligned}
f(r) &\geq \left(\left(\frac{1}{a} (1 - e^{-ar_2}) \right) \wedge \left(\frac{1}{2r_2} e^{-ar_2} \right) \right) r^2 \\
&\geq \frac{1}{2r_2} e^{-ar_2} \left(\left(\frac{2r_2}{a} (e^{ar_2} - 1) \right) \wedge 1 \right) r^2 \\
&\geq \frac{1}{2r_2} e^{-ar_2} ((2r_2^2) \wedge 1) r^2.
\end{aligned}$$

Thus, using assumption 13, it follows that

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

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

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

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

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