# MEASURING SAMPLE QUALITY WITH DIFFUSIONS

By Jackson Gorham, Andrew B. Duncan, Sebastian J. Vollmer, and Lester Mackey

*Stanford University, University of Sussex, University of Warwick, and Microsoft Research*

Standard Markov chain Monte Carlo diagnostics, like effective sample size, are ineffective for biased sampling procedures that sacrifice asymptotic correctness for computational speed. Recent work addresses this issue for a class of strongly log-concave target distributions by constructing a computable discrepancy measure based on Stein's method that provably determines convergence to the target. We generalize this approach to cover any target with a fast-coupling Itô diffusion by bounding the derivatives of Stein equation solutions in terms of Markov process coupling times. As example applications, we develop computable and convergence-determining *diffusion Stein discrepancies* for log-concave, heavy-tailed, and multimodal targets and use these quality measures to select the hyperparameters of biased samplers, compare random and deterministic quadrature rules, and quantify bias-variance tradeoffs in approximate Markov chain Monte Carlo. Our explicit multivariate Stein factor bounds may be of independent interest.

**1. Introduction.** In Bayesian inference and maximum likelihood estimation [32], it is common to encounter complex target distributions with unknown normalizing constants and intractable expectations. Traditional Markov chain Monte Carlo (MCMC) methods [7] contend with such targets by generating consistent sample estimates of the intractable expectations. A recent alternative approach [see, e.g., 93, 1, 51] is to employ biased MCMC procedures that sacrifice consistency to improve the speed of sampling. The argument is compelling: the reduction in Monte Carlo variance from more rapid sampling can outweigh the bias incurred and yield more accurate estimates overall. However, the extra degree of freedom poses new challenges for selecting samplers and their tuning parameters, as traditional MCMC diagnostics, like effective sample size and asymptotic variance, pooled and within-chain variance measures, and mean and trace plots [7], do not detect sample bias.

*MSC 2010 subject classifications:* Primary 60J60; 62-04; 62E17; 60E15; 65C60; secondary 62-07; 65C05; 68T05

*Keywords and phrases:* Itô diffusion, Stein discrepancy, sample quality, multivariate Stein factors, Wasserstein decay, Stein's method, Markov chain Monte Carlo

Gorham and Mackey [35] addressed this issue for a class of strongly log-concave target distributions [62] by introducing a computable discrepancy based on Stein's method that provably determines convergence to the target and satisfies our quality measure design criteria detailed in Section 2. By relating Stein's method to Markov process coupling rates in Section 3, we generalize their approach to develop computable and convergence-determining *diffusion Stein discrepancies* for any target with a fast-coupling Itô diffusion. Our proofs produce uniform multivariate *Stein factor* bounds on the derivatives of Stein equation solutions that may be of independent interest. In Section 4, we provide examples of practically checkable sufficient conditions for fast coupling and illustrate the process of verifying these conditions for log-concave, heavy-tailed, and multimodal targets. Section 5 describes a practical algorithm for computing diffusion Stein discrepancies using a geometric spanner and linear programming. In Section 6, we deploy our discrepancies to select the hyperparameters of biased samplers, compare random and deterministic quadrature rules, and quantify bias-variance tradeoffs in approximate Markov chain Monte Carlo. A discussion of related and future work follows in Section 7, and all proofs are deferred to the appendices.

**Notation** For $r \in [1, \infty]$, let $\|\cdot\|_r$ denote the $\ell^r$ norm on $\mathbb{R}^d$. We will use $\|\cdot\|$ as a generic norm on $\mathbb{R}^d$ satisfying $\|\cdot\| \geq \|\cdot\|_2$ and define the associated dual norms, $\|v\|^* \triangleq \sup_{u\in\mathbb{R}^d:\|u\|=1} \langle u, v\rangle$ for vectors $v \in \mathbb{R}^d$ and $\|W\|^* \triangleq \sup_{u\in\mathbb{R}^d:\|u\|=1} \|Wu\|^*$ for matrices $W \in \mathbb{R}^{d\times d}$. Let $e_j$ be the $j$-th standard basis vector, $\nabla_j$ be the partial derivative $\frac{\partial}{\partial x_j}$, and $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ be the smallest and largest eigenvalues of a symmetric matrix. For any real vector $v$ and tensor $T$, let $\|v\|_{op} \triangleq \|v\|_2$ and $\|T\|_{op} \triangleq \sup_{\|u\|_2=1} \|T[u]\|_{op}$. For each sufficiently differentiable vector- or matrix-valued function $g$, we define the bound $M_0(g) \triangleq \sup_{x\in\mathbb{R}^d} \|g(x)\|_{op}$ and the $k$-th order Lipschitz coefficients

$$M_k(g) \triangleq \sup_{x,y\in\mathbb{R}^d, x\neq y} \frac{\|\nabla^{k-1}g(x) - \nabla^{k-1}g(y)\|_{op}}{\|x-y\|_2} \quad \text{for} \quad k \in \mathbb{N}$$

and $\nabla^0$ the identity operator. For each differentiable matrix-valued function $a$, we let $\langle\nabla, a(x)\rangle = \sum_j e_j \sum_k \nabla_k a_{jk}(x)$ represent the divergence operator applied to each row of $a$ and define the Lipschitz coefficients $F_k(a) \triangleq \sup_{x\in\mathbb{R}^d, \|v_1\|_2=1,\ldots,\|v_k\|_2=1} \|\nabla^k a(x)[v_1, \ldots, v_k]\|_F$ for $\|\cdot\|_F$ the Frobenius norm. Finally, when the domain and range of a function $f$ can be inferred from context, we write $f \in C^k$ to indicate that $f$ has $k$ continuous derivatives.

## 2. Measuring sample quality.

Consider a target probability distribution $P$ with finite mean, continuously differentiable density $p$, and support on all of $\mathbb{R}^d$. We will name the set of all such distributions $\mathcal{P}_1$. We assume that $p$ can be evaluated up to its normalizing constant but that exact expectations under $P$ are unattainable for most functions of interest. We will therefore use a *weighted sample*, represented as a discrete distribution $Q_n = \sum_{i=1}^n q(x_i)\delta_{x_i}$, to approximate intractable expectations $\mathbb{E}_P[h(Z)]$ with tractable sample estimates $\mathbb{E}_{Q_n}[h(X)] = \sum_{i=1}^n q(x_i)h(x_i)$. Here, the support of $Q_n$ is a collection of distinct sample points $x_1, \ldots, x_n \in \mathbb{R}^d$, and the weight $q(x_i)$ associated with each point is governed by a probability mass function $q$. We assume nothing about the process generating the sample points, so they may be the product of any random or deterministic mechanism.

Our ultimate goal is to develop a computable quality measure suitable for comparing any two samples approximating the same target distribution. More precisely, we seek to quantify how well $\mathbb{E}_{Q_n}$ approximates $\mathbb{E}_P$ in a manner that, at the very least, *(i)* indicates when a sample sequence is converging to $P$, *(ii)* identifies when a sample sequence is not converging to $P$, and *(iii)* is computationally tractable. A natural starting point is to consider the maximum error incurred by the sample approximation over a class of scalar test functions $\mathcal{H}$,

$$d_{\mathcal{H}}(Q_n, P) \triangleq \sup_{h\in\mathcal{H}} |\mathbb{E}_P[h(Z)] - \mathbb{E}_{Q_n}[h(X)]|. \tag{1}$$

When $\mathcal{H}$ is convergence determining, the measure (1) is an *integral probability metric* (IPM) [66], and $d_{\mathcal{H}}(Q_n, P)$ converges to zero only if the sample sequence $(Q_n)_{n\geq 1}$ converges in distribution to $P$.

While a variety of standard probability metrics are representable as IPMs [66], the intractability of integration under $P$ precludes us from computing most of these candidate quality measures. Recently, Gorham and Mackey [35] sidestepped this issue by constructing a class of test functions $h$ known a priori to have zero mean under $P$. Their resulting quality measure – the *Langevin graph Stein discrepancy* – satisfied our computability and convergence detection requirements (Desiderata *(i)* and *(iii)*) and detected sample sequence non-convergence (Desideratum *(ii)*) for strongly log concave targets with bounded third and fourth derivatives [62]. In the next section we will greatly extend the reach of the Stein discrepancy approach to measuring sample quality by introducing a diverse family of practical operators for generating mean zero functions under $P$ and establishing broad conditions under which the resulting Stein discrepancies detect non-convergence. We begin by reviewing the principles of Stein's method that underlie the Stein discrepancy.

**3. Stein's method.** In the early 1970s, Charles Stein [86] introduced a powerful three-step approach to upper-bounding a reference IPM $d_{\mathcal{H}}$:

1. First, identify an operator $\mathcal{T}$ that maps input functions[1] $g : \mathbb{R}^d \to \mathbb{R}^d$ in a domain $\mathcal{G}$ into mean-zero functions under $P$, i.e.,

$$\mathbb{E}_P[(\mathcal{T}g)(Z)] = 0 \quad \text{for all} \quad g \in \mathcal{G}.$$

   The operator $\mathcal{T}$ and its domain $\mathcal{G}$ define the *Stein discrepancy* [35],

$$\begin{aligned} \mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}) &\triangleq \sup_{g\in\mathcal{G}} |\mathbb{E}_{Q_n}[(\mathcal{T}g)(X)]| \\ &= \sup_{g\in\mathcal{G}} |\mathbb{E}_{Q_n}[(\mathcal{T}g)(X)] - \mathbb{E}_P[(\mathcal{T}g)(Z)]| = d_{\mathcal{T}\mathcal{G}}(Q_n, P), \end{aligned} \tag{2}$$

   a quality measure which takes the form of an integral probability metric while avoiding explicit integration under $P$.
2. Next, prove that, for each test function $h$ in the reference class $\mathcal{H}$, the *Stein equation*

$$h(x) - \mathbb{E}_P[h(Z)] = (\mathcal{T}g_h)(x) \tag{3}$$

   admits a solution $g_h \in \mathcal{G}$. This step ensures that the reference metric $d_{\mathcal{H}}$ lower bounds the Stein discrepancy (Desideratum *(ii)*) and, in practice, can be carried out simultaneously for large classes of target distributions.
3. Finally, use whatever means necessary to upper bound the Stein discrepancy and thereby establish convergence to zero under appropriate conditions (Desideratum *(i)*). Our general result, Proposition 8, suffices for this purpose.

While Stein's method is traditionally used as analytical tool to establish rates of distributional convergence, we aim, following [35], to develop the method into a practical computational tool for measuring the quality of a sample. We begin by assessing the convergence properties of a broad class of Stein operators derived from Itô diffusions. Our efforts will culminate in Section 5, where we show how to explicitly compute the Stein discrepancy (2) given any sample measure $Q_n$ and appropriate choices of $\mathcal{T}$ and $\mathcal{G}$.

3.1. *Identifying a Stein operator.* To identify an operator $\mathcal{T}$ that generates mean-zero functions under $P$, we will appeal to the elegant and widely applicable *generator method* construction of Barbour [4, 5] and Götze [37]. These authors note that if $(Z_t)_{t\geq 0}$ is a Feller process with invariant measure $P$, then the *infinitesimal generator* $\mathcal{A}$ of the process, defined pointwise by

$$(\mathcal{A}u)(x) = \lim_{t\to 0} (\mathbb{E}[u(Z_t) \mid Z_0 = x] - u(x))/t \tag{4}$$

[1]Real-valued $g$ are also common, but $\mathbb{R}^d$-valued $g$ are more convenient for our purposes.

satisfies $\mathbb{E}_P[(\mathcal{A}u)(Z)] = 0$ under very mild restrictions on $u$ and $\mathcal{A}$. Gorham and Mackey [35] developed a *Langevin Stein operator* based on the generator a specific Markov process – the Langevin diffusion described in (D1). Here, we will consider a broader class of continuous Markov processes known as *Itô diffusions.*

Definition 1 (Itô diffusion [71, Def. 7.1.1]). *A (time-homogeneous)* Itô diffusion *with starting point $x \in \mathbb{R}^d$, Lipschitz* drift coefficient $b : \mathbb{R}^d \to \mathbb{R}^d$, *and Lipschitz* diffusion coefficient $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times m}$ *is a stochastic process $(Z_{t,x})_{t\geq 0}$ solving the Itô stochastic differential equation*

$$dZ_{t,x} = b(Z_{t,x})\,dt + \sigma(Z_{t,x})\,dW_t \quad \textit{with} \quad Z_{0,x} = x \in \mathbb{R}^d, \tag{5}$$

*where $(W_t)_{t\geq 0}$ is an $m$-dimensional Wiener process.*

As the next theorem, distilled from [61, Thm. 2] and [74, Sec. 4.6], shows, it is straightforward to construct Itô diffusions with a given invariant measure $P$ (see also [49, 46]).

Theorem 2 ([61, Thm. 2] and [74, Sec. 4.6]). *Fix an Itô diffusion with $C^1$ drift and diffusion coefficients $b$ and $\sigma$, and define its* covariance coefficient $a(x) \triangleq \sigma(x)\sigma(x)^\top$. *$P \in \mathcal{P}_1$ is an invariant measure of this diffusion if and only if $b(x) = \frac{1}{2}\frac{1}{p(x)}\langle\nabla, p(x)a(x)\rangle + f(x)$ for a* non-reversible component *$f \in C^1$ satisfying $\langle\nabla, p(x)f(x)\rangle = 0$ for all $x \in \mathbb{R}^d$. If $f$ is $P$-integrable, then*

$$b(x) = \tfrac{1}{2}\tfrac{1}{p(x)}\langle\nabla, p(x)(a(x) + c(x))\rangle \tag{6}$$

*for $c$ a differentiable $P$-integrable skew-symmetric $d \times d$ matrix–valued function termed the* stream coefficient *[16, 53]. In this case, for all $u \in C^2 \cap \mathrm{dom}(\mathcal{A})$, the infinitesimal generator (4) of the diffusion takes the form*

$$(\mathcal{A}u)(x) = \tfrac{1}{2}\tfrac{1}{p(x)}\langle\nabla, p(x)(a(x) + c(x))\nabla u(x)\rangle.^{2} \tag{7}$$

Remarks. Theorem 2 does not require Lipschitz assumptions on $b$ or $\sigma$. An example of a non-reversible component which is not $P$-integrable is $f(x) = v/p(x)$ for any constant vector $v \in \mathbb{R}^d$. Prominent examples of $P$-targeted diffusions include

[2] We have chosen an atypical form for the infinitesimal generator in (7), as it will give rise to a first-order differential operator (8) with more desirable properties. One can check, for instance, that the first order operator $(\mathcal{T}g)(x) = 2\langle b(x), g(x)\rangle + \langle a(x), \nabla g(x)\rangle$ derived from the standard form of the generator, $(\mathcal{A}u)(x) = \langle b(x), \nabla u(x)\rangle + \frac{1}{2}\langle a(x), \nabla^2 u(x)\rangle$, fails to satisfy Proposition 3 whenever the non-reversible component $f(x) \not\equiv 0$.

(D1) the *(overdamped) Langevin diffusion* (also known as the *Brownian* or *Smoluchowski dynamics*) [74, Secs. 6.5 and 4.5], where $a \equiv I$ and $c \equiv 0$;
(D2) the *preconditioned Langevin diffusion* [88], where $c \equiv 0$ and $a \equiv \sigma\sigma^\top$ for a constant diffusion coefficient $\sigma \in \mathbb{R}^{d\times m}$ ;
(D3) the *Riemannian Langevin diffusion* [49, 82, 33], where $c \equiv 0$ and $a$ is not constant;
(D4) the *non-reversible preconditioned Langevin diffusion* [see, e.g., 61, 20, 79], where $a \equiv \sigma\sigma^\top$ for $\sigma \in \mathbb{R}^{d\times m}$ constant and $c$ not identically 0;
(D5) and the *second-order* or *underdamped Langevin diffusion* [43], where we target the joint distribution $P \otimes \mathcal{N}(0, I)$ on $\mathbb{R}^{2d}$ with

$$a \equiv 2\begin{pmatrix}0 & 0\\ 0 & I\end{pmatrix} \text{ and } c \equiv 2\begin{pmatrix}0 & -I\\ I & 0\end{pmatrix}.$$

We will present detailed examples making use of these diffusion classes in Sections 4 and 6.

Theorem 2 forms the basis for our Stein operator of choice, the *diffusion Stein operator* $\mathcal{T}$, defined by substituting $g$ for $\frac{1}{2}\nabla u$ in the generator (7):

$$(\mathcal{T}g)(x) = \tfrac{1}{p(x)}\langle \nabla, p(x)(a(x) + c(x))g(x)\rangle. \tag{8}$$

$\mathcal{T}$ is an appropriate choice for our setting as it depends on $P$ only through $\nabla \log p$ and is therefore computable even when the normalizing constant of $p$ is unavailable. One suitable domain for $\mathcal{T}$ is the *classical Stein set* [35] of 1-bounded functions with 1-bounded, 1-Lipschitz derivatives:

$$\mathcal{G}_{\|\cdot\|} \triangleq \left\{ g : \mathbb{R}^d \to \mathbb{R}^d \,\middle|\, \sup_{x\neq y\in\mathbb{R}^d} \max\left(\|g(x)\|^*, \|\nabla g(x)\|^*, \tfrac{\|\nabla g(x)-\nabla g(y)\|^*}{\|x-y\|}\right) \le 1 \right\}.$$

Indeed, our next proposition, proved in Section A, shows that, on this domain, the diffusion Stein operator generates mean-zero functions under $P$.

Proposition 3. *If $\mathcal{T}$ is the diffusion Stein operator (8) for $P \in \mathcal{P}_1$ with $a, c \in C^1$ and $a$, $c$, $b$ (6) $P$-integrable, then $\mathbb{E}_P[(\mathcal{T}g)(Z)] = 0$ for all $g \in \mathcal{G}_{\|\cdot\|}$.*

Together, $\mathcal{T}$ and $\mathcal{G}_{\|\cdot\|}$ give rise to the *classical diffusion Stein discrepancy* $\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|})$, our primary object of study in Sections 3.2 and 3.3.

3.2. *Lower bounding the diffusion Stein discrepancy.* To establish that the classical diffusion Stein discrepancy detects non-convergence (Desideratum *(ii)*), we will lower bound the discrepancy in terms of the $L^1$ *Wasserstein distance*, $d_{\mathcal{W}_{\|\cdot\|_2}}$, a standard reference IPM generated by

$$\mathcal{H} = \mathcal{W}_{\|\cdot\|_2} \triangleq \{h : \mathbb{R}^d \to \mathbb{R} \mid \sup_{x\neq y\in\mathbb{R}^d} |h(x) - h(y)| \le \|x - y\|_2\}.$$

The first step is to show that, for each $h \in \mathcal{W}_{\|\cdot\|_2}$, the solution $g_h$ to the Stein equation (3) with diffusion Stein operator (8) has low-order derivatives uniformly bounded by target-specific constants called *Stein factors*.

Explicit Langevin diffusion (D1) Stein factor bounds are readily available for a wide variety of univariate targets[3] (see, e.g., [87, 11, 12] for explicit bounds or [54] for a recent review). In contrast, in the multivariate setting, efforts to establish Stein factors have focused on Gaussian [5, 37, 78, 10, 65, 67, 29], Dirichlet [28], and strongly log-concave [62] targets with preconditioned Langevin (D2) operators. To extend the reach of the literature, we will derive multivariate Stein factors for targets with fast-coupling Itô diffusions. Our measure of coupling speed is the *Wasserstein decay rate*.

DEFINITION 4 (Wasserstein decay rate). *Let* $(P_t)_{t\geq 0}$ *be the* transition semigroup *of an Itô diffusion* $(Z_{t,x})_{t\geq 0}$ *defined via*

$$(P_t f)(x) \triangleq \mathbb{E}[f(Z_{t,x})] \quad \textit{for all measurable } f, \quad x \in \mathbb{R}^d, \quad \textit{and} \quad t \geq 0.$$

*For any non-increasing integrable function* $r : \mathbb{R}_{\geq 0} \to \mathbb{R}$*, we say that* $(P_t)_{t\geq 0}$ *has* Wasserstein decay rate $r$ *if*

$$d_{\mathcal{W}_{\|\cdot\|_2}}(\delta_x P_t,\, \delta_y P_t) \leq r(t)\, d_{\mathcal{W}_{\|\cdot\|_2}}(\delta_x, \delta_y) \quad \textit{for all } x, y \in \mathbb{R}^d \textit{ and } t \geq 0, \tag{9}$$

*where* $\delta_x P_t$ *denotes the distribution of* $Z_{t,x}$*.*

Our next result, proved in Section B, shows that the smoothness of a solution $g_h$ to a Stein equation is controlled by the rate of Wasserstein decay and hence by how quickly two diffusions with distinct starting points couple. The Stein factor bounds on the derivatives of $u_h$ and $g_h$ may be of independent interest for establishing rates of distributional convergence.

THEOREM 5 (Stein factors from Wasserstein decay). *Fix any Lipschitz* $h$*. If an Itô diffusion has invariant measure* $P \in \mathcal{P}_1$*, transition semigroup* $(P_t)_{t\geq 0}$*, Wasserstein decay rate* $r$*, and infinitesimal generator* $\mathcal{A}$ (4)*, then*

$$u_h \triangleq \int_0^\infty \mathbb{E}_P[h(Z)] - P_t h \, dt \tag{10}$$

*is twice continuously differentiable and satisfies*

$$M_1(u_h) \leq M_1(h) \int_0^\infty r(t)\, dt \qquad \textit{and} \qquad h - \mathbb{E}_P[h(Z)] = \mathcal{A} u_h.$$

*Hence,* $g_h \triangleq \frac{1}{2}\nabla u_h$ *solves the Stein equation* (3) *with diffusion Stein operator* (8) *whenever* $\mathcal{A}$ *has the form* (7)*. If the drift and diffusion coefficients*

[3]The Langevin operator recovers *Stein's density method operator [87]* when $d = 1$.

*$b$ and $\sigma$ have locally Lipschitz second derivatives and a right inverse $\sigma^{-1}(x)$ for each $x \in \mathbb{R}^d$, and $h \in C^2$ with bounded second derivatives, then*

$$M_2(u_h) \le M_1(h)(\beta_1 + \beta_2), \tag{11}$$

*where*

$$\beta_1 = r(0)^2(2^{3/2}M_0(\sigma^{-1}) + M_1(\sigma)M_0(\sigma^{-1}) + \tfrac{2}{3}\sqrt{\alpha}) \quad \textit{and}$$
$$\beta_2 = s_r r(\tfrac{1}{2})(\sqrt{2}e^{\gamma_2/2}M_0(\sigma^{-1}) + e^{\gamma_2/2}M_1(\sigma)M_0(\sigma^{-1}) + e^{\gamma_4/2}\sqrt{\alpha})$$

*for $\gamma_\rho \triangleq \rho M_1(b) + \frac{\rho^2-2\rho}{2}M_1(\sigma)^2 + \frac{\rho}{2}F_1(\sigma)^2$, $\alpha \triangleq \frac{M_2(b)^2}{2M_1(b)+4M_1(\sigma)^2} + 2F_2(\sigma)^2$.*

A first consequence of Theorem 5, proved in Section D, concerns Stein operators (8) with constant covariance and stream matrices $a$ and $c$. In this setting, fast Wasserstein decay implies that the diffusion Stein discrepancy converges to zero only if the Wasserstein distance does (Desideratum *(ii)*).

THEOREM 6 (Stein discrepancy lower bound: constant $a$ and $c$). *Consider an Itô diffusion with diffusion Stein operator $\mathcal{T}$ (8) for $P \in \mathcal{P}_1$, Wasserstein decay rate $r$, constant covariance and stream matrices $a$ and $c$, and Lipschitz drift $b(x) = \frac{1}{2}(a+c)\nabla \log p(x)$. If $s_r \triangleq \int_0^\infty r(t)\,dt$, then*

$$\begin{aligned} & d_{\mathcal{W}_{\|\cdot\|_2}}(Q_n, P) \\ & \le\ 3s_r \max\Big(\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|}),\ \sqrt[3]{\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|})\sqrt{2}\,\mathbb{E}[\|G\|_2]^2(2M_1(b) + \tfrac{1}{s_r})^2}\Big), \end{aligned} \tag{12}$$

*where $G \in \mathbb{R}^d$ is a standard normal vector and $M_1(b) \le \frac{1}{2}\|a+c\|_{op}M_2(\log p)$.*

Theorem 6 in fact provides an explicit upper bound on the Wasserstein distance in terms of the Stein discrepancy and the Wasserstein decay rate. Under additional smoothness assumptions on the coefficients, the explicit relationship between Stein discrepancy and Wasserstein distance can be improved and extended to diffusions with non-constant diffusion coefficient, as our next result, proved in Section E, shows.

THEOREM 7 (Stein discrepancy lower bound: non-constant $a$ and $c$). *Consider an Itô diffusion for $P \in \mathcal{P}_1$ with diffusion Stein operator $\mathcal{T}$ (8), Wasserstein decay rate $r$, and Lipschitz drift and diffusion coefficients $b$ (6) and $\sigma$ with locally Lipschitz second derivatives. If $s_r \triangleq \int_0^\infty r(t)\,dt$, then*

$$\begin{aligned} & d_{\mathcal{W}_{\|\cdot\|_2}}(Q_n, P) \\ & \le 2\max\Big(\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|})\max(s_r, \beta_1+\beta_2),\ \sqrt{\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|})\sqrt{2/\pi}(\beta_1+\beta_2)\zeta}\Big), \end{aligned}$$

*for $\beta_1$, $\beta_2$ defined in Theorem 5 and*

$$\zeta \triangleq \mathbb{E}[\|G\|_2](1 + 2M_1(b)M_0(g_h) + M_1^*(m)M_1(g_h))$$

*where $G \in \mathbb{R}^d$ is a standard normal vector, $m \triangleq a + c$, and $M_1^*(m) \triangleq \sup_{x\neq y} \|m(x) - m(y)\|_{op}^*/\|x - y\|_2$.*

*If, additionally, $\nabla^3 b$ and $\nabla^3 \sigma$ are locally Lipschitz, then, for all $\iota \in (0, 1)$,*

$$\begin{aligned} &d_{\mathcal{W}_{\|\cdot\|_2}}(Q_n, P) \\ &\le 2\max\Big(\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|}) \max(s_r, \beta_1 + \beta_2), \tfrac{\zeta}{\iota}\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|})^{\frac{1}{1+\iota}} (\tfrac{1+\iota s_r}{K\zeta/\sqrt{d}})^{\frac{1}{1+\iota}}\Big), \end{aligned}$$

*for $K > 0$ a constant depending only on $M_{1:3}(\sigma), M_{1:3}(b)$, $M_0(\sigma^{-1})$, and $r$.*

In Section 4, we will present practically checkable conditions implying fast Wasserstein decay and discuss both broad families and specific examples of diffusion-target pairings covered by this theory.

3.3. *Upper bounding the diffusion Stein discrepancy.* In upper bounding the Stein discrepancy, one classically aims to establish rates of convergence to $P$ for specific sequences $(Q_n)_{n=1}^\infty$. Since our interest is in explicitly computing Stein discrepancies for arbitrary sample sequences, our general upper bound in Proposition 8 serves principally to provide sufficient conditions under which the classical diffusion Stein discrepancy converges to zero.

Proposition 8 (Stein discrepancy upper bound). *Let $\mathcal{T}$ be the diffusion Stein operator (8) for $P \in \mathcal{P}_1$. If $m \triangleq a + c$ and $b$ (6) are $P$-integrable,*

$$\begin{aligned} \mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|}) &\le \inf_{X\sim Q_n, Z\sim P} (\mathbb{E}[2\|b(X) - b(Z)\| + \|m(X) - m(Z)\|] \\ &\qquad + \mathbb{E}[(2\|b(Z)\| + \|m(Z)\|)\min(\|X - Z\|, 2)]) \\ &\le \mathcal{W}_{s,\|\cdot\|}(Q_n, P)(2M_1^{\|\cdot\|}(b) + M_1^{\|\cdot\|}(m)) \\ &\quad + \mathcal{W}_{s,\|\cdot\|}(Q_n, P)^t\, 2^{1-t}\, \mathbb{E}[(2\|b(Z)\| + \|m(Z)\|)^{s/(s-t)}]^{(s-t)/s} \end{aligned}$$

*for any $s \ge 1$ and $t \in (0, 1]$, where $\mathcal{W}_{s,\|\cdot\|}(Q_n, P) \triangleq \inf_{X\sim Q_n, Z\sim P} \mathbb{E}[\|X - Z\|^s]^{1/s}$ represents the $L^s$ Wasserstein distance.*

This result, proved in Section F, complements the Wasserstein distance lower bounds of Section 3.2 and implies that, for Lipschitz and sufficiently integrable $m$ and $b$, the diffusion Stein discrepancy converges to zero whenever $Q_n$ converges to $P$ in Wasserstein distance.

3.4. *Extension to non-uniform Stein sets.* For any $c_1, c_2, c_3 > 0$, our analyses and algorithms readily accommodate the non-uniform Stein set

$$\mathcal{G}_{\|\cdot\|}^{c_{1:3}} \triangleq \left\{ g : \mathbb{R}^d \to \mathbb{R}^d \,\middle|\, \sup_{x \neq y \in \mathbb{R}^d} \max\left( \tfrac{\|g(x)\|^*}{c_1}, \tfrac{\|\nabla g(x)\|^*}{c_2}, \tfrac{\|\nabla g(x) - \nabla g(y)\|^*}{c_3 \|x-y\|} \right) \le 1 \right\}.$$

This added flexibility can be valuable when tight upper bounds on a reference IPM, like the Wasserstein distance, are available for a particular choice of Stein factors $(c_1, c_2, c_3)$. When such Stein factors are unknown or difficult to compute, we recommend the parameter-free classical Stein set and graph Stein set of the sequel as practical defaults, since the classical Stein discrepancy is strongly equivalent to any non-uniform Stein discrepancy:

PROPOSITION 9 (Equivalence of non-uniform Stein discrepancies). *For any $c_1, c_2, c_3 > 0$,*

$$\min(c_1, c_2, c_3)\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|}) \le \mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|}^{c_{1:3}}) \le \max(c_1, c_2, c_3)\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|}).$$

REMARK. The short proof follows exactly as in [35, Prop. 4].

**4. Sufficient conditions for Wasserstein decay.** Since the Stein discrepancy lower bounds of Section 3 depend on the Wasserstein decay (9) of the chosen diffusion, we next provide examples of practically checkable sufficient conditions for Wasserstein decay and illustrate the process of verifying these conditions for a variety of specific diffusion-target pairings.

4.1. *Uniform dissipativity.* It is well known [see, e.g., 8, Eq. 7] that the Langevin diffusion (D1) enjoys exponential Wasserstein decay whenever $\log p$ is $k$-strongly log concave, i.e., when the drift $b = \frac{1}{2}\nabla \log p$ satisfies $\langle b(x) - b(y), x - y\rangle \le -\frac{k}{2}\|x - y\|_2^2$ for $k > 0$. An analogous *uniform dissipativity* condition gives explicit exponential decay for a generic Itô diffusion:

THEOREM 10 (Wasserstein decay: uniform dissipativity). *Fix $k > 0$ and $G \succ 0$, and let $\|w\|_G^2 \triangleq \langle w, Gw\rangle$, for any vector or matrix $w \in \mathbb{R}^{d \times d'}$, $d' \ge 1$. An Itô diffusion with drift and diffusion coefficients $b$ and $\sigma$ satisfying*

$$2\langle b(x) - b(y), G(x - y)\rangle + \|\sigma(x) - \sigma(y)\|_G^2 \le -k\|x - y\|_G^2 \text{ for all } x, y \in \mathbb{R}^d$$

*has Wasserstein decay rate (9)* $r(t) = e^{-kt/2}\sqrt{\lambda_{\max}(G)/\lambda_{\min}(G)}$.

REMARK. Theorem 10 holds even when the drift $b$ is not Lipschitz.

Hence, if the drift $b$ of an Itô diffusion is $-k/2$-*one-sided Lipschitz*, i.e.,

$$2\langle b(x) - b(y), G(x-y)\rangle \le -k\|x-y\|_G^2 \quad \text{for all} \quad x, y \in \mathbb{R}^d \tag{13}$$

and some $G \succ 0$, and the diffusion coefficient $\sigma$ is $\sqrt{k'}$-Lipschitz, that is,

$$\|\sigma(x) - \sigma(y)\|_G^2 \le k'\|x-y\|_G^2 \quad \text{for all} \quad x, y \in \mathbb{R}^d,$$

then, whenever $k' < k$, the diffusion exhibits exponential Wasserstein decay. with rate $e^{-(k-k')t/2}\sqrt{\lambda_{\max}(G)/\lambda_{\min}(G)}$. The proof of Theorem 10 in Section G relies on a synchronous coupling of Itô diffusions and mimics [8, Sec. 1].

Example 1 (Bayesian logistic regression with Gaussian prior). A one-sided Lipschitz drift arises naturally in the setting of Bayesian logistic regression [31], a canonical model of binary outcomes $y \in \{-1, 1\}$ given measured covariates $v \in \mathbb{R}^d$. Consider the log density of a Bayesian logistic regression posterior based on a dataset of $L$ observations $(v_l, y_l)$ and a $\mathcal{N}(\mu, \Sigma)$ prior:

$$\log p(\beta) = \underbrace{-\tfrac{1}{2}\|\Sigma^{-1/2}(\beta-\mu)\|_2^2}_{\text{multivariate Gaussian prior}} \underbrace{-\textstyle\sum_{l=1}^L \log(1+\exp(-y_l\langle v_l, \beta\rangle))}_{\text{logistic regression likelihood}} + \text{const.}$$

Here, our inferential target is the unobserved parameter vector $\beta \in \mathbb{R}^d$. Since

$$-\Sigma^{-1} \succcurlyeq \nabla^2 \log p(\beta) = -\Sigma^{-1} - \textstyle\sum_{l=1}^L \frac{e^{y_l\langle v_l, \beta\rangle}}{(1+e^{y_l\langle v_l,\beta\rangle})^2} v_l v_l^\top \succcurlyeq -\Sigma^{-1} - \frac{1}{4}\sum_{l=1}^L v_l v_l^\top,$$

the $P$-targeted preconditioned Langevin diffusion (D2) drift $b(\beta) = \frac{1}{2}\Sigma\nabla \log p(\beta)$ satisfies (13) with $k = 1$ and $G = \Sigma^{-1}$ and $M_1(b) \le \frac{1}{2}\|I + \frac{1}{4}\Sigma\sum_{l=1}^L v_l v_l^\top\|_{op}$. Hence, the diffusion enjoys geometric Wasserstein decay (Theorem 10) and a Wasserstein lower bound on the Stein discrepancy (Theorem 6).

Example 2 (Bayesian Huber regression with Gaussian prior). Huber's least favorable distribution provides a robust error model for the regression of a continuous response $y \in \mathbb{R}$ onto a vector of measured covariates $v \in \mathbb{R}^d$ [44]. Given $L$ observations $(v_l, y_l)$ and a $\mathcal{N}(\mu, \Sigma)$ prior on an unknown parameter vector $\beta \in \mathbb{R}^d$, the Bayesian Huber regression log posterior takes the form

$$\log p(\beta) = \underbrace{-\tfrac{1}{2}\|\Sigma^{-1/2}(\beta-\mu)\|_2^2}_{\text{multivariate Gaussian prior}} \underbrace{-\textstyle\sum_{l=1}^L \rho_c(y_l - \langle v_l, \beta\rangle)}_{\text{Huber's least favorable likelihood}} + \text{const.}$$

where $\rho_c(r) \triangleq \frac{1}{2}r^2\mathbb{I}[|r| \le c] + c(|r| - \frac{1}{2}c)\mathbb{I}[|r| > c]$ for fixed $c > 0$. Since $\rho_c'(r) = \min(\max(r, -c), c)$ is 1-Lipschitz and convex, and the Hessian of the log prior is $-\Sigma^{-1}$, the $P$-targeted preconditioned Langevin diffusion (D2) drift $b(\beta) = \frac{1}{2}\Sigma\nabla \log p(\beta)$ satisfies (13) with $k = 1$ and $G = \Sigma^{-1}$ and $M_1(b) \le \frac{1}{2}\|I + \Sigma\sum_{l=1}^L v_l v_l^\top\|_{op}$. This is again sufficient for exponential Wasserstein decay and a Wasserstein lower bound on the Stein discrepancy.

4.2. *Distant dissipativity, constant $\sigma$.* When the diffusion coefficient $\sigma$ is constant with $a \triangleq \frac{1}{2}\sigma\sigma^\top$ invertible, Eberle [22] showed that a *distant dissipativity* condition is sufficient for exponential Wasserstein decay. Specifically, if we define a one-sided Lipschitz constant conditioned on a distance $r > 0$,

$$-\kappa(r) = \sup\{2(b(x) - b(y))^\top a^{-1}(x - y)/r^2 : (x - y)^\top a^{-1}(x - y) = r^2\},$$

then [22, Cor. 2] establishes exponential Wasserstein decay whenever $\kappa$ is continuous with $\liminf_{r\to\infty} \kappa(r) > 0$ and $\int_0^1 r\kappa(r)^- dr < \infty$. For a Lipschitz drift, this holds whenever $b$ is dissipative at large distances, that is, whenever, for some $k > 0$, we have $\kappa(r) \geq k$ for all $r$ sufficiently large [22, Lem. 1].

Example 3 (Gaussian mixture with common covariance). Consider an $m$-component mixture density $p(x) = \sum_{j=1}^m w_j \phi_j(x)$, where the component weights $w_j \geq 0$ sum to one and $\phi_j$ is the density of a $\mathcal{N}(\mu_j, \Sigma)$ distribution on $\mathbb{R}^d$. Fix any $x, y \in \mathbb{R}^d$. If $\|\Sigma^{-1/2}(x - y)\|_2 = r$, the $P$-targeted preconditioned Langevin diffusion (D2) with drift $b(z) = \frac{1}{2}a\nabla \log p(z)$ and $a = \Sigma$ satisfies

$$\begin{aligned} &2(b(x) - b(y))^\top a^{-1}(x - y) = (\nabla \log p(x) - \nabla \log p(y))^\top (x - y) \\ &= -r^2 + \langle \Sigma^{-1/2}(\mu(x) - \mu(y)), \Sigma^{-1/2}(x - y)\rangle \leq -r^2 + r\Delta, \end{aligned}$$

by Cauchy-Schwarz and Jensen's inequality, for $\Delta \triangleq \sup_{j,k} \|\Sigma^{-1/2}(\mu_j - \mu_k)\|_2$, $\mu(x) \triangleq \sum_{j=1}^m \pi_j(x)\mu_j$, and $\pi_j(x) \triangleq \frac{w_j\phi_j(x)}{p(x)}$. Moreover, by the mean value theorem, Cauchy-Schwarz, and Jensen's inequality, we have, for each $v \in \mathbb{R}^d$,

$$\begin{aligned} &2\langle \Sigma^{-1/2}(b(x) - b(y)), v\rangle = \langle \Sigma^{-1/2}(\nabla\mu(z) - I)(x - y), v\rangle \\ &= \langle (\Sigma^{-1/2}S(z)\Sigma^{-1/2} - I)\Sigma^{-1/2}(x - y), v\rangle \leq \|v\|_2 \|\Sigma^{-1/2}(x - y)\|_2\, L, \end{aligned}$$

for some $z \in \mathbb{R}^d$, $S(x) \triangleq \frac{1}{2}\sum_{j=1}^m \sum_{k=1}^m \pi_j(x)\pi_k(x)(\mu_j - \mu_k)(\mu_j - \mu_k)^\top$, and $L \triangleq \sup_{j,k} |1 - \|\Sigma^{-1/2}(\mu_j - \mu_k)\|_2^2/2|$. Hence, $b$ is Lipschitz, and $\kappa(r) \geq \frac{1}{2}$ when $r > 2\Delta$, so our diffusion enjoys exponential Wasserstein decay [22, Lem. 1] and a Stein discrepancy upper bound on the Wasserstein distance.

4.3. *Distant dissipativity, non-constant $\sigma$.* Using a combination of synchronous and reflection couplings, Wang [92, Thm. 2.6] showed that diffusions satisfying a distant dissipativity condition exhibit exponential Wasserstein decay, even when the diffusion coefficient $\sigma$ is non-constant. In Section H, we combine the coupling strategy of [92, Thm. 2.6] with the approach of [22] for diffusions with constant $\sigma$ to obtain the following explicit Wasserstein decay rate for distantly dissipative diffusions with bounded $\sigma^{-1}$.

THEOREM 11 (Wasserstein decay: distant dissipativity). *Let $(P_t)_{t\geq 0}$ be the transition semigroup of an Itô diffusion with drift and diffusion coefficients $b$ and $\sigma$. Define the truncated diffusion coefficient*

$$\sigma_0(x) = (\sigma(x)\sigma(x)^\top - \lambda_0^2 I)^{1/2} \quad \textit{for some} \quad \lambda_0 \in [0, 1/M_0(\sigma^{-1})]$$

*and the distance-conditional dissipativity function*

$$\kappa(r) = \inf\{ -2\alpha(\langle b(x) - b(y), x-y\rangle + \tfrac{1}{2}\|\sigma_0(x) - \sigma_0(y)\|_F^2 - \tfrac{1}{2}\|(\sigma_0(x) - \sigma_0(y))^\top (x-y)\|_2^2/r^2)/r^2 : \|x-y\|_2 = r\} \tag{14}$$

*for some*

$$m_0 \leq \inf_{x\neq y} \tfrac{\|(\sigma_0(x)-\sigma_0(y))^\top(x-y)\|_2}{\|x-y\|_2} \quad \textit{and} \quad \alpha \triangleq 1/(\lambda_0^2 + m_0^2/4).$$

*If the constants*

$$\begin{aligned} R_0 &= \inf\{R \geq 0 \,:\, \kappa(r) \geq 0,\ \forall\, r \geq R\}, \\ R_1 &= \inf\{R \geq R_0 \,:\, \kappa(r) R(R - R_0) \geq 8,\ \forall\, r \geq R\}, \end{aligned}$$

*satisfy $R_0 \leq R_1 < \infty$ then, for all $x, y \in \mathbb{R}^d$ and $t \geq 0$,*

$$d_{\mathcal{W}_{\|\cdot\|_2}}(\delta_x P_t, \delta_y P_t) \leq 2\varphi(R_0)^{-1} e^{-ct}\, d_{\mathcal{W}_{\|\cdot\|_2}}(\delta_x, \delta_y) \tag{15}$$

*where*

$$\varphi(r) = e^{-\frac{1}{4}\int_0^r s\kappa(s)^- \,ds} \quad \textit{and} \quad \tfrac{1}{c} \;=\; \alpha \int_0^{R_1}\int_0^s \exp(\tfrac{1}{4}\int_t^s u\kappa(u)^- \,du)\,dt\,ds.$$

REMARK. Theorem 11 holds even when the drift $b$ is not Lipschitz.

The Wasserstein decay rate (15) in Theorem 11 has a simple form when the diffusion is dissipative at large distances and $\kappa$ is bounded below. These rates follow exactly as in [22, Lem. 1].

COROLLARY 12. *Under the conditions of Theorem 11, suppose that*

$$\kappa(r) \geq -L \;\; \textit{for } r \leq R \quad \textit{and} \quad \kappa(r) \geq K \;\; \textit{for } r > R$$

*for $R, L \geq 0$ and $K > 0$. If $LR_0^2 \leq 8$ then*

$$\alpha^{-1}c^{-1} \;\leq\; \tfrac{e-1}{2}R^2 + e\sqrt{8K^{-1}}\,R + 4K^{-1} \;\leq\; \tfrac{3e}{2}\max(R^2, 8K^{-1}),$$

*and if $LR_0^2 \geq 8$ then*

$$\alpha^{-1}c^{-1} \;\leq\; 8\sqrt{2\pi}R^{-1}L^{-1/2}(L^{-1} + K^{-1})\exp\Big(\tfrac{LR^2}{8}\Big) + 32R^{-2}K^{-2}.$$

EXAMPLE 4 (Multivariate Student's t regression with pseudo-Huber prior). The multivariate Student's t distribution is also commonly employed as a robust error model for the linear regression of continuous responses $y \in \mathbb{R}^L$ onto measured covariates $V \in \mathbb{R}^{L\times d}$ [94, 55]. Under a pseudo-Huber prior [42], a Bayesian multivariate Student's t regression posterior takes the form

$$p(\beta) \propto \underbrace{\exp(\delta^2(1-\sqrt{1+\|\beta/\delta\|_2^2}))}_{\text{pseudo-Huber prior}} \underbrace{(1+\tfrac{1}{\nu}(y-V\beta)^\top\Sigma^{-1}(y-V\beta))^{-(\nu+L)/2}}_{\text{multivariate Student's t likelihood}}$$

for fixed $\delta, \nu > 0$ and $\Sigma \succ 0$. Introduce the shorthand $\psi_\lambda(r) \triangleq 2\sqrt{1+r^2/\delta^2} - \lambda^2$ for each $\lambda \in [0,\sqrt{2})$ and $\xi(\beta) \triangleq 1 + \frac{1}{\nu}(y-V\beta)^\top\Sigma^{-1}(y-V\beta)$. Since

$$\nabla \log p(\beta) = -2\beta/\psi_0(\|\beta\|_2) + (1+\tfrac{\nu}{L})V^\top\Sigma^{-1}(y-V\beta)/\xi(\beta)$$

is bounded, no $P$-targeted preconditioned Langevin diffusion (D2) will satisfy the distant dissipativity conditions of Section 4.2. However, we will show that whenever $V^\top V \succ 0$, the Riemannian Langevin diffusion (D3) with $\sigma(\beta) = \sqrt{\psi_0(\|\beta\|_2)}I \in \mathbb{R}^{d\times d}$, $a(\beta) = \frac{1}{2}\psi_0(\|\beta\|_2)I$, and $b(\beta) = a(\beta)\nabla\log p(\beta) + \langle\nabla, a(\beta)\rangle$ satisfies the Wasserstein decay preconditions of Corollary 12.

Indeed, fix any $\lambda_0 \in (0, 1/M_0(\sigma^{-1})) = (0,\sqrt{2})$. Since $M_1(\sqrt{\psi_\lambda}) \le \frac{1}{\delta\sqrt{2-\lambda^2}}$, $M_1(\psi_\lambda) \le \frac{2}{\delta}$, and $M_2(\psi_\lambda) \le \frac{2}{\delta^2}$, $\sigma_0$, $\sigma$, $a$, and $\nabla a$ are all Lipschitz. The drift $b$ is also Lipschitz, since $\nabla \log p$ and the product of $a(\beta)$ and

$$\begin{aligned}&\nabla^2 \log p(\beta) = -2I/\psi_0(\|\beta\|_2) + 8\beta\beta^\top/(\delta^2\psi_0^3(\|\beta\|_2))\\ &+ \left(1+\tfrac{\nu}{L}\right)(2V^\top\Sigma^{-1}(y-V\beta)(y-V\beta)^\top\Sigma^{-1}V/\xi^2(\beta) - V^\top\Sigma^{-1}V/\xi(\beta)).\end{aligned}$$

are bounded. Hence, $\kappa$ (14) is bounded below. Moreover, the the Hölder continuity of $x \mapsto \sqrt{x}$, Cauchy-Schwarz, and the triangle inequality imply

$$\begin{aligned}\kappa(r) &\ge \inf_{\|\beta-\beta'\|_2=r} \tfrac{2\alpha}{r^2}(\langle b(\beta')-b(\beta), \beta-\beta'\rangle - \tfrac{d-1}{2}|\sqrt{\psi_{\lambda_0}(\|\beta\|_2)} - \sqrt{\psi_{\lambda_0}(\|\beta'\|_2)}|^2)\\ &\ge 2\alpha - \tfrac{2\alpha}{r}(\tfrac{d-1}{\delta} + M_1(\psi_0) + \sup_\beta (1+\tfrac{\nu}{L})\psi_0(\|\beta\|_2)\|V^\top\Sigma^{-1}(y-V\beta)\|_2/\xi(\beta))\\ &\ge 2\alpha - \tfrac{2\alpha}{r}\left(\tfrac{d+1}{\delta} + \sup_s (1+\tfrac{\nu}{L})\tfrac{2(1+s/\delta)(\|V^\top\Sigma^{-1}y\|_2 + s\|V^\top\Sigma^{-1}V\|_{op})}{1+\frac{1}{\nu}\max(0, s/\|(V^\top\Sigma^{-1}V)^{-1}\|_{op} - \|\Sigma^{-1}y\|_2)^2}\right).\end{aligned}$$

Letting $\zeta$ represent the supremum in the final inequality, we see that $\kappa(r) \ge \alpha = 1/\lambda_0^2$ whenever $r \ge 2(\frac{d+1}{\delta} + \zeta)$. Hence, Corollary 12 delivers exponential Wasserstein decay. A Wasserstein lower bound on the Stein discrepancy now follows from Theorem 7, since $M_2(\sqrt{\psi_0}) \le \frac{1}{\sqrt{2}\delta^2}$, $M_3(\psi_0) \le \frac{96}{25\sqrt{5}\delta^3}$, and $a(\beta)\nabla^2\log p(\beta)$ is Lipschitz, and hence $M_2(\sigma)$ and $M_2(b)$ are bounded.

**5. Computing Stein discrepancies.** In this section, we introduce a computationally tractable Stein discrepancy that inherits the favorable convergence properties established in Sections 3 and 4. While we only explicitly discuss target distributions supported on all of $\mathbb{R}^d$, constrained domains of the form $(\alpha_1, \beta_1) \times \cdots \times (\alpha_d, \beta_d)$ where $-\infty \le \alpha_i < \beta_i \le \infty$ for all $1 \le i \le d$ can be handled by introducing boundary constraints as in [35, Section 4.4].

5.1. *Spanner Stein discrepancies.* For any sample $Q_n$, Stein operator $\mathcal{T}$, and Stein set $\mathcal{G}$, the Stein discrepancy $\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G})$ is recovered by solving an optimization problem over functions $g \in \mathcal{G}$. For example, if we write $m \triangleq a + c$ and $b(x) \triangleq \frac{1}{2}\frac{1}{p(x)}\langle \nabla, p(x)m(x)\rangle$, the classical diffusion Stein discrepancy is the value

$$\begin{aligned}\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|}) = \sup\nolimits_g &\textstyle\sum_{i=1}^n q(x_i)(2\langle b(x_i), g(x_i)\rangle + \langle m(x_i), \nabla g(x_i)\rangle) \\ &\text{s.t. } \max(\|g(x)\|^*, \|\nabla g(x)\|^*, \tfrac{\|\nabla g(x) - \nabla g(y)\|^*}{\|x-y\|}) \le 1, \forall x, y \in \mathbb{R}^d.\end{aligned}$$

For all Stein sets, the diffusion Stein discrepancy objective is linear in $g$ and only queries $g$ and $\nabla g$ at the $n$ sample points underlying $Q_n$. However, the classical Stein set $\mathcal{G}_{\|\cdot\|}$ constrains $g$ at all points in its domain, resulting in an infinite-dimensional optimization problem.[4]

To obtain a finite-dimensional problem that is both convergence-determining and straightforward to optimize, we will make use of the *graph Stein sets* of [35]. For a given graph $G = (V, E)$ with $V = \mathrm{supp}(Q_n)$, the graph Stein set,

$$\begin{aligned}\mathcal{G}_{\|\cdot\|, Q_n, G} = \Big\{ g : \max(&\|g(v)\|^*, \|\nabla g(v)\|^*, \tfrac{\|g(x)-g(y)\|^*}{\|x-y\|}, \tfrac{\|\nabla g(x) - \nabla g(y)\|^*}{\|x-y\|}) \le 1, \\ &\tfrac{\|g(x)-g(y)-\nabla g(x)(x-y)\|^*}{\frac{1}{2}\|x-y\|^2} \le 1, \tfrac{\|g(x)-g(y)-\nabla g(y)(x-y)\|^*}{\frac{1}{2}\|x-y\|^2} \le 1,\ \forall (x,y) \in E, v \in V \Big\},\end{aligned}$$

imposes boundedness constraints only at sample points and smoothness constraints only at pairs of sample points enumerated in the edge set $E$. The graph is termed a *t-spanner* [14, 75] if each edge $(x, y) \in E$ is assigned the weight $\|x - y\|$, and, for all $x' \neq y' \in V$, there exists a path between $x'$ and $y'$ in the graph with total path weight no greater than $t\|x' - y'\|$. Remarkably, for any linear Stein operator $\mathcal{T}$, a *spanner Stein discrepancy* $\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|, Q_n, G_t})$ based on a $t$-spanner $G_t$ is equivalent to the classical Stein discrepancy in the following strong sense, implying Desiderata *(i)* and *(ii)*.

[4] When $d = 1$, the problem reduces to a finite-dimensional convex quadratically constrained quadratic program with linear objective as in [35, Thm. 9].

PROPOSITION 13 (Equivalence of classical and spanner Stein discrepancies). *If $G_t = (\mathrm{supp}(Q_n), E)$ is a $t$-spanner for $t \geq 1$, then*

$$\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|}) \leq \mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|, Q_n, G_t}) \leq \kappa_d t^2 \, \mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|})$$

*where $\kappa_d$ is independent of $(Q_n, P, \mathcal{T}, G_t)$ and depends only on $d$ and $\|\cdot\|$.*

REMARK. The proof relies on the Whitney-Glaeser extension theorem [84, Thm. 1.4] of Glaeser [34] and follows exactly as in [35, Prop. 5 and 6].

When $d = 1$, a $t$-spanner with exactly $n-1$ edges is obtained in $O(n \log n)$ time for all $t \geq 1$ by introducing edges just between sample points that are adjacent in sorted order. More generally, if $\|\cdot\|$ is an $\ell^p$ norm, one can construct a 2-spanner with $O(\kappa_d' n)$ edges in $O(\kappa_d' n \log(n))$ expected time where $\kappa_d'$ is a constant that depends only on the norm $\|\cdot\|$ and the dimension $d$ [41]. Hence, a spanner Stein discrepancy can be computed by solving a finite-dimensional convex optimization problem with a linear objective, $O(n)$ variables, and $O(\kappa_d' n)$ convex constraints, making it an appealing choice for a computable quality measure (Desideratum *(iii)*).

5.2. *Decoupled linear programs.* Moreover, if we choose the norm $\|\cdot\| = \|\cdot\|_1$, the graph Stein discrepancy optimization problem decouples into $d$ independent linear programs (LPs) that can be solved in parallel using off-the-shelf solvers. Indeed, for any $G = (\mathrm{supp}(Q_n), E)$, $\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|_1, Q_n, G})$ equals

$$\sum_{j=1}^d \sup_{\psi_j \in \mathbb{R}^n, \Psi_j \in \mathbb{R}^{d \times n}} \sum_{i=1}^n q(x_i)(2b_j(x_i)\psi_{ji} + \sum_{k=1}^d m_{jk}(x_i)\Psi_{jki}) \tag{16}$$

$$\text{s.t. } \|\psi_j\|_\infty \leq 1,\ \|\Psi_j\|_\infty \leq 1,\ \text{and for all } i \neq l, (x_i, x_l) \in E$$

$$\max\left(\tfrac{|\psi_{ji} - \psi_{jl}|}{\|x_i - x_l\|_1}, \tfrac{\|\Psi_j(e_i - e_k)\|_\infty}{\|x_i - x_l\|_1}, \tfrac{|\psi_{ji} - \psi_{jl} - \langle \Psi_j e_i, x_i - x_l \rangle|}{\frac{1}{2}\|x_i - x_l\|_1^2}, \tfrac{|\psi_{ji} - \psi_{jl} - \langle \Psi_j e_i, x_l - x_i \rangle|}{\frac{1}{2}\|x_i - x_l\|_1^2}\right) \leq 1,$$

where $\psi_{ji}$ and $\Psi_{jki}$ represent the values $g_j(x_i)$ and $\nabla_k g_j(x_i)$ respectively. Therefore, our recommended quality measure is the 2-spanner diffusion Stein discrepancy with $\|\cdot\| = \|\cdot\|_1$. Its computation is summarized in Algorithm 1.

**6. Experiments.** In this section, we use our proposed quality measures to select hyperparameters for biased samplers, to quantify a bias-variance trade-off for approximate MCMC, and to compare deterministic and random quadrature rules. We solve all linear programs using Julia for Mathematical Programming [59] with the Gurobi 6.0.4 solver [72] and use the C++ greedy spanner implementation of Bouts et al. [6] to compute our 2-spanners. Our

**Algorithm 1** Spanner diffusion Stein discrepancy, $\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|_1, Q_n, G_2})$

**input:** sample $Q_n$, target score $\nabla \log p$, covariance coefficient $a$, stream coefficient $c$
$G_2 \leftarrow$ 2-spanner of $V = \mathrm{supp}(Q_n)$
**for** $j = 1$ **to** $d$ **do (in parallel)**
  $\tau_j \leftarrow$ Optimal value of $j$-th coordinate linear program (16) with graph $G_2$
**return** $\sum_{j=1}^{d} \tau_j$

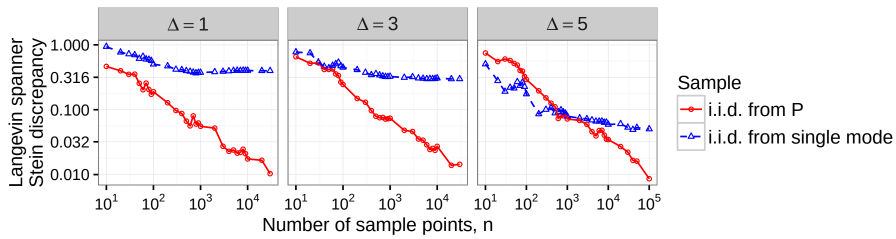
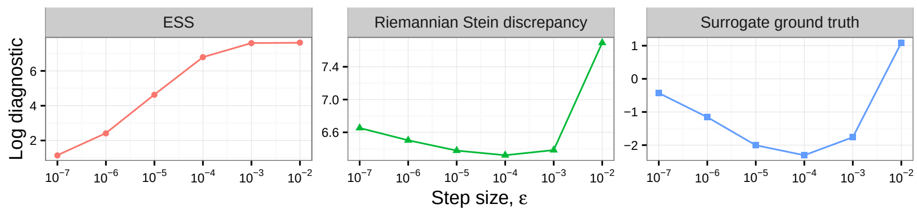



**Fig 1:** Stein discrepancy for normal mixture target $P$ with $\Delta$ mode separation (Section 6.1).

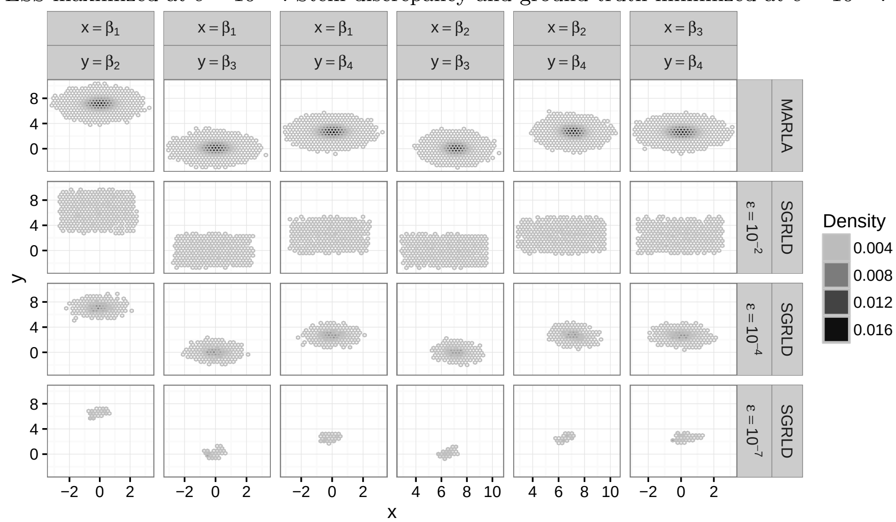
timings were obtained on a single core of an Intel Xeon CPU E5-2650 v2 @ 2.60GHz. Code reconstructing all experiments is available on the Julia package site `https://jgorham.github.io/SteinDiscrepancy.jl/`.

6.1. *A simple example.* We first present a simple example to illustrate several Stein discrepancy properties. For a Gaussian mixture target $P$ (Example 3) with $p(x) \propto e^{-\frac{1}{2}(x-\frac{\Delta}{2})^2} + e^{-\frac{1}{2}(x+\frac{\Delta}{2})^2}$ and $\Delta > 0$, we simulate one i.i.d. sequence of sample points from $P$ and a second i.i.d. sequence from $\mathcal{N}(-\frac{\Delta}{2}, 1)$, which represents only one component of $P$. For various mode separations $\Delta$, Figure 1 shows that the Langevin spanner Stein discrepancy (D1) applied to the first $n$ Gaussian mixture sample points decreases to zero at a $n^{-1/2}$ rate, while the discrepancy applied to the single mode sequence stays bounded away from zero. However, Figure 1 also indicates that larger sample sizes are needed to distinguish between the mixture and single mode sample sequences when $\Delta$ is large. This accords with our theory (see Example 3, Corollary 12, and Theorem 6), which implies that both the Langevin diffusion Wasserstein decay rate and the bound relating Stein to Wasserstein degrade as the mixture mode separation $\Delta$ increases.

6.2. *Selecting sampler hyperparameters.* Stochastic Gradient Riemannian Langevin Dynamics (SGRLD) [73] with a constant step size $\epsilon$ is an approximate MCMC procedure designed to accelerate posterior inference. Unlike asymptotically correct MCMC algorithms, SGRLD has a stationary distribution that deviates increasingly from its target $P$ as its step size $\epsilon$ grows. On the other hand, if $\epsilon$ is too small, SGRLD fails to explore the sample space sufficiently quickly. Hence, an appropriate setting of $\epsilon$ is paramount for accurate inference.




**(a)** Step size selection criteria and surrogate ground truth (median marginal Wasserstein). ESS maximized at $\epsilon = 10^{-2}$. Stein discrepancy and ground truth minimized at $\epsilon = 10^{-4}$.




**(b)** Bivariate hexbin plots. **Top row:** surrogate ground truth sample ($2 \times 10^8$ MARLA points). **Bottom 3 rows:** $2,000$ SGRLD sample points for various step sizes $\epsilon$.

**Fig 2:** Step size selection, stochastic gradient Riemannian Langevin dynamics (Section 6.2).

To demonstrate the value of diffusion Stein discrepancies for hyperparameter selection, we analyzed a biometric dataset of $L = 202$ athletes from the Australian Institute of Sport that was previously the focus of a heavy-tailed regression analysis [85]. In the notation of Example 4, we used SGRLD to conduct a Bayesian multivariate Student's $t$ regression ($\nu = 10$, $\Sigma = I$) of athlete lean body mass onto red blood count, white blood count, plasma ferritin concentration, and a constant regressor of value $1/\sqrt{L}$ with a pseudo-Huber prior ($\delta = 0.1$) on the unknown parameter vector $\beta \in \mathbb{R}^4$.

After standardizing the output variable and non-constant regressors and initializing each chain with an approximate posterior mode found by L-BFGS started at the origin, we ran SGRLD with minibatch size 30, metric $G(\beta) = 1/(2\sqrt{1 + \|\beta/\delta\|_2^2})I$, and a variety of step sizes $\epsilon$ to produce sample sequences of length $200,000$ thinned to length $2,000$. We then selected the step size that delivered the highest quality sample – either the maximum effective sample size (ESS, a popular MCMC mixing diagnostic based on asymptotic variance [7]) or the minimum Riemannian Langevin spanner Stein discrepancy with $a(\beta) = G^{-1}(\beta)$. The longest discrepancy computation consumed 6s for spanner construction and 65s to solve a coordinate optimization problem. As a surrogate measure of ground truth, we also generated a sample $Q^*$ of size $2 \times 10^8$ from the Metropolis-adjusted Riemannian Langevin Algorithm (MARLA) [33] with metric $G$ and compute the median bivariate marginal Wasserstein distance $d_{\mathcal{W}_{\|\cdot\|_1}}$ between each SGRLD sample and $Q^*$ thinned to $5,000$ points [39].

Figure 2a shows that ESS, which does not account for stationary distribution bias, selects the largest step size available, $\epsilon = 10^{-2}$. As seen in Figure 2b, this choice results in samples that are greatly overdispersed when compared with the ground truth MARLA sample $Q^*$. At the other extreme, the selection $\epsilon = 10^{-7}$ produces greatly underdispersed samples due to slow mixing. The Stein discrepancy chooses an intermediate value, $\epsilon = 10^{-4}$. The same value minimizes the surrogate ground truth Wasserstein measure and produces samples that most closely resemble the $Q^*$ in Figure 2b.

6.3. *Quantifying a bias-variance trade-off.* Approximate random walk Metropolis-Hastings (ARWMH) [51] with tolerance parameter $\epsilon$ is a biased MCMC procedure that accelerates posterior inference by approximating the standard MH correction. Qualitatively, a smaller setting of $\epsilon$ produces a more faithful approximation of the MH correction and less bias between the chain's stationary distribution and the target distribution of interest. A larger setting of $\epsilon$ leads to faster sampling and a more rapid reduction of Monte Carlo variance, as fewer datapoint likelihoods are computed per sampling step. We will quantify this bias-variance trade-off as a function of sampling time using the Langevin spanner Stein discrepancy.

In the notation of Example 2, we conduct a Bayesian Huber regression analysis ($c = 1$) of the log radon levels in $1,190$ Minnesota households [30] as a function of the log amount of uranium in the county, an indicator of whether the radon reading was performed in a basement, and an intercept term. A $\mathcal{N}(0, I)$ prior is placed on the coefficient vector $\beta$. We run ARWMH with minibatch size 5 and two settings of the tolerance threshold $\epsilon$ (0.1 and 0.2) for $10^7$ likelihood evaluations, discard the sample points from the first $10^5$ evaluations, and thin the remaining points to sequences of length $1,000$. Figure 3 displays the Langevin spanner Stein discrepancy applied to the first $n$ points in each sequence as a function of the likelihood evaluation count,

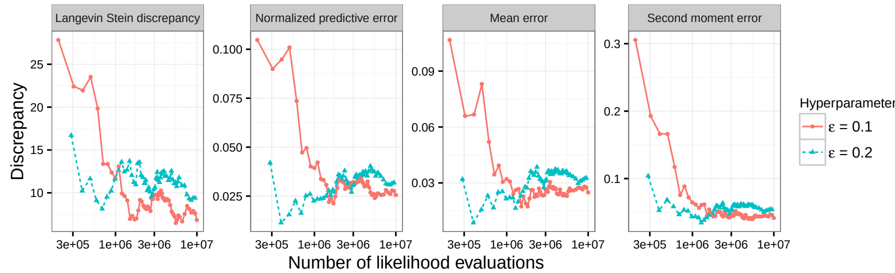



**Fig 3:** Bias-variance trade-off curves for approximate random walk MH (Section 6.3).

which serves as a proxy for sampling time. As expected, the higher tolerance sample ($\epsilon = 0.2$) is of higher Stein quality for a small computational budget but is eventually overtaken by the $\epsilon = 0.1$ sample with smaller asymptotic bias. The longest discrepancy computation consumed 0.8s for the spanner and 20.1s for a coordinate LP.

To provide external support for the Stein discrepancy quantification, we generate a Metropolis-adjusted Langevin chain [81] of length $10^8$ as a surrogate $Q^*$ for the target $P$ and display several measures of expectation error between $X \sim Q_n$ and $Z \sim Q^*$ in Figure 3: the normalized predictive error $\max_l |\mathbb{E}[\langle X - Z, v_l/\|v_l\|_\infty\rangle]|$ for $v_l$ the $l$-th datapoint covariate vector, the mean error $\frac{\max_j |\mathbb{E}[X_j - Z_j]|}{\max_j |\mathbb{E}_{Q^*}[Z_j]|}$, and the second moment error $\frac{\max_{j,k} |\mathbb{E}[X_j X_k - Z_j Z_k]|}{\max_{j,k} |\mathbb{E}_{Q^*}[Z_j Z_k]|}$. We see that the Stein discrepancy provides comparable results without the need for an additional surrogate chain.

6.4. *Comparing quadrature rules.* Stein discrepancies can also measure the quality of deterministic sample sequences designed to improve upon Monte Carlo sampling. For the Gaussian mixture target of Section 6.1, Figure 4 compares the median quality of 50 sample sequences generated from four quadrature rules recently studied in [52, Sec. 4.1]: i.i.d. sampling from $P$, Quasi-Monte Carlo (QMC) sampling using a deterministic quasirandom number generator, Frank-Wolfe (FW) kernel herding [13, 3], and fully-corrective Frank-Wolfe (FCFW) kernel herding [52]. The quality judgments of the Langevin spanner Stein discrepancy (D1) closely mimic those of the $L^1$ Wasserstein distance $d_{\mathcal{W}_{\|\cdot\|}}$, which is computable for simple univariate targets [90]. Each Stein discrepancy was computed in under 0.03s.

Under both diagnostics and as previously observed in other metrics [52], the i.i.d. samples are typically of lower median quality than their deterministic counterparts. More suprisingly and in contrast to past work focused on very smooth function classes [52], FCFW underperforms FW and QMC in our diagnostics for larger sample sizes. Apparently FCFW, which is heavily optimized for smooth function integration, has sacrificed approx-

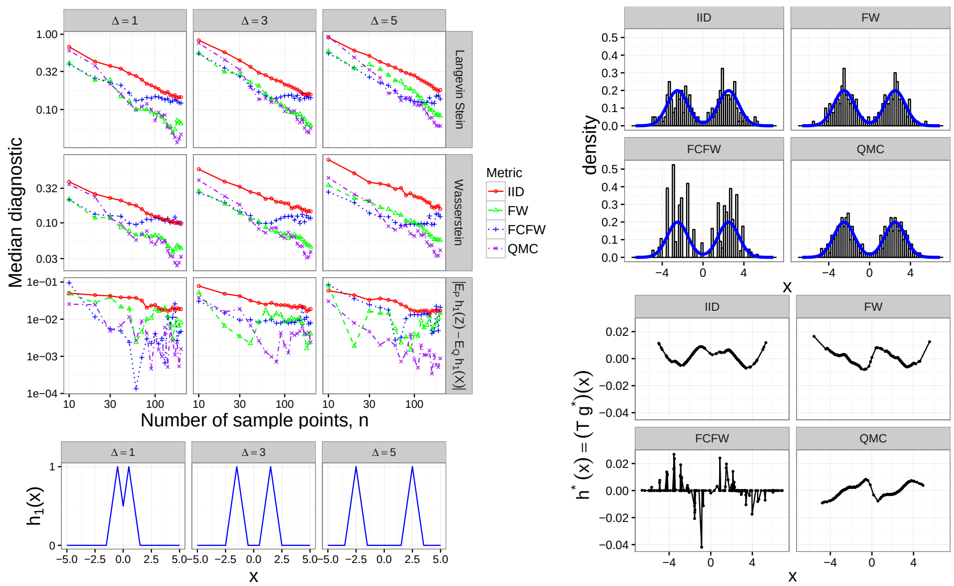



**Fig 4: Left:** Quadrature rule quality comparison for Gaussian mixture targets $P$ with mode separation $\Delta$ (Section 6.4). **Right:** (Top) Sample histograms with $p$ overlaid ($\Delta = 5$, $n = 200$). (Bottom) Optimal discriminating test functions $h^* = \mathcal{T}g^*$ from Stein program.

imation quality for less smooth test functions. For example, Figure 4 shows that QMC offers a better quadrature estimate than FCFW for $h_1(x) = \max\{0, 1 - \min_{j\in\{1,2\}} |x - \mu_j|\}$, a 1-Lipschitz approximation to the indicator of being within one standard deviation of a mode.

In addition to providing a sample quality score, the Stein discrepancy optimization problem produces an optimal Stein function $g^*$ and an associated test function $h^* = \mathcal{T}g^*$ that is mean zero under $P$ and best distinguishes the sample $Q_n$ from the target $P$. Figure 4 gives examples of these maximally discriminatve functions $h^*$ for a target mode separation of $\Delta = 5$ and length 200 sequences from each quadrature rule. We also display the associated sample histograms with overlaid target density. The optimal FCFW function reflects the jagged nature of the FCFW histogram.

**7. Connections and conclusions.** We developed quality measures suitable for comparing the fidelity of arbitrary "off-target" sample sequences by generating infinite collections of known target expectations.

*Alternative quality measures.* The score statistic of Fan et al. [25] and the Gibbs sampler convergence criteria of Zellner and Min [95] account for some sample biases but sacrifice differentiating power by exploiting only a finite number of known target expectations. For example, when $P = \mathcal{N}(0, 1)$, the score statistic [25] cannot differentiate two samples with the same means and variances. Maximum mean discrepancies (MMDs) over characteristic reproducing kernel Hilbert spaces [38] do detect arbitrary distributional biases but are only computable when the chosen kernel functions can be integrated under the target. In practice, one often approximates MMD using a sample from the target, but this requires a separate trustworthy sample from $P$.

While we have focused on the graph and classical Stein sets of [35], our diffusion Stein operators can also be paired with the reproducing kernel Hilbert space unit balls advocated in [69, 68, 15, 58, 36] to form tractable *kernel diffusion Stein discrepancies.* We have also restricted our attention to Stein operators arising from diffusion generators. These take the form $(\mathcal{T}g)(x) = \frac{1}{p(x)}\langle\nabla, p(x)m(x)g(x)\rangle$ with $m = a + c$ for $a(x)$ positive semidefinite and $c(x)$ skew-symmetric. More generally, if the matrix $m$ possesses eigenvalues having a negative real part, then the resulting operator need not correspond to a diffusion process. Such operators fall into the class of *pseudo-Fokker Planck* operators which have been studied in the context of quantum optics [80]. As noted in [18, 19] it is possible to obtain corresponding stochastic dynamics in an extended state space by introducing complex-valued noise terms; these operators may merit further study in future work.

*Alternative inferential tasks.* While our chief motivation is sample quality measurement, our work is also directly applicable to a variety of inferential tasks that currently rely on the Langevin operator introduced by [35, 70], including control variate design [70, 68], one sample hypothesis testing [15, 58], variational inference [57, 77], and importance sampling [56]. The Stein factor bounds of Theorem 5 can also be used, in the manner of [64, 47, 40], to characterize the error of numerical discretizations of diffusions. These works convert bounds on the solutions of Poisson equations – Stein factors – into central limit theorems for $\mathbb{E}_{Q_n}[h(X)] - \mathbb{E}_P[h(Z)]$, confidence intervals for $\mathbb{E}_P[h(Z)]$, and mean-squared error bounds for the estimate $\mathbb{E}_{Q_n}[h(X)]$. Teh et al. [89] and Vollmer et al. [91] extended these approaches to obtain error estimates for approximate discretizations of the Langevin diffusion on $\mathbb{R}^d$, while, independently of our work, Huggins and Zou [45] established error estimates for Itô diffusion approximations with biased drifts and constant diffusion coefficients. By Theorem 5, their results also hold for Itô diffusions with non-constant diffusion coefficients.

*Alternative targets.* Our exposition has focused on the Wasserstein distance $d_{\mathcal{W}_{\|\cdot\|}}$, which is only defined for distributions with finite means. A parallel development could be made for the Dudley metric [66] to target distributions with undefined mean. The work of Cerrai [9] also suggests that the Lipschitz condition on our drift and diffusion coefficients can be relaxed.

## APPENDIX A: PROOF OF PROPOSITION 3

Fix any $g \in \mathcal{G}_{\|\cdot\|}$. Since $g$ and $\nabla g$ are bounded and $b$, $a$, and $c$ are $P$-integrable, $\mathbb{E}_P[(\mathcal{T}g)(Z)]$ is finite. Define the ball $\mathcal{B}_r = \{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$ with $n_r(z)$ the outward facing unit normal vector for each $z$ on the boundary $\partial\mathcal{B}_r$. Since $z \mapsto p(z)(a(z)+c(z))g(z)$ is in $C^1$, we may apply the dominated convergence theorem and then the divergence theorem to obtain

$$\begin{aligned}\mathbb{E}_P[(\mathcal{T}g)(Z)] &= \lim_{r\to\infty} \int_{\mathcal{B}_r} \langle \nabla, p(z)(a(z)+c(z))g(z)\rangle \, dz \\ &= \lim_{r\to\infty} \int_{\partial\mathcal{B}_r} \langle n_r(z), (a(z)+c(z))g(z)p(z)\rangle \, dz.\end{aligned}$$

Let $f(r) = M_0(g) \int_{\partial\mathcal{B}_r} \|a(z)+c(z)\|_{op}\, p(z)\, dz$. Since $g$ and $n_r$ are bounded,

$$\int_{\partial\mathcal{B}_r} \langle n_r(z), (a(z)+c(z))g(z)p(z)\rangle \, dz \leq f(r).$$

The coarea formula [2] and the integrability of $a$ and $c$ further imply that

$$\int_0^\infty f(r)\, dr = \int_{\mathbb{R}^d} M_0(g)\|a(z)+c(z)\|_{op}\, p(z)\, dz < \infty.$$

Hence, $\liminf_{r\to\infty} f(r) = 0$, and therefore $\mathbb{E}_P[(\mathcal{T}g)(Z)] = 0$.

## APPENDIX B: PROOF OF THEOREM 5

Fix any $x \in \mathbb{R}^d$ and $h \in \mathcal{W}_{\|\cdot\|_2}$ with $\mathbb{E}_P[h(Z)] = 0$. Since the drift and diffusion coefficients are Lipschitz, [50, Thm. 3.4] guarantees that the diffusion $(Z_{t,x})_{t\geq 0}$ is well-defined. Using the shorthand $s_r \triangleq \int_0^\infty r(t)\, dt$, we will show that the posited function $u_h$ (10) exists and solves the *Poisson equation*

$$h = \mathcal{A}u_h \tag{17}$$

with infinitesimal generator $\mathcal{A}$, that $u_h$ is Lipschitz, that $u_h$ has a continuous Hessian, and that $u_h$ has a bounded Hessian under the additional smoothness assumptions.

*Existence of $u_h$ and solving the Poisson equation* (17). Consider the set $L \triangleq (1+\|x\|_2^2)C_0(\mathbb{R}^d) = \{(1+\|x\|_2^2)f : f \in C_0(\mathbb{R}^d)\}$, where $C_0(\mathbb{R}^d)$ is the set of continuous functions vanishing at infinity. Equipped with the norm $\|f\|_L = \sup_{x\in\mathbb{R}^d} |f(x)|/(1+\|x\|_2^2)$, the set $L$ is a Banach space [83]. As noted in [17], the space $L$ can also be characterized as the closure of the set of bounded continuous functions, $C_b(\mathbb{R}^d)$, in the set $\{f : \mathbb{R}^d \to \mathbb{R} : \|f\|_L < \infty\}$. To discuss the well-posedness of the Poisson equation (17), we first show that the transition semigroup of an Itô diffusion is strongly continuous on $L$.

PROPOSITION 14. *The transition semigroup $(P_t)_{t\geq 0}$ of an Itô diffusion with Lipschitz drift and diffusion coefficients is strongly continuous on $L$.*

PROOF. Fix any $f \in L$ and $x \in \mathbb{R}^d$. We first show that $(P_t f)(x)$ converges pointwise to $f(x)$ as $t \to 0^+$. Since the associated Itô process $(Z_{t,x})_{t\geq 0}$ is almost surely pathwise continuous [50, Thm. 3.4] and $f$ is continuous in a neighborhood of $x$, it follows that $f(Z_{t,x}) \to f(x)$ as $t \to 0^+$, almost surely. Moreover, [27, Sec. 5, Cor. 1.2] implies that

$$\mathbb{E}\big[\sup_{0\leq t\leq 1} |f(Z_{t,x})|\big] \leq \|f\|_L(1 + \mathbb{E}\big[\sup_{0\leq t\leq 1} \|Z_{t,x}\|_2^2\big]) \leq C\|f\|_L(1 + \|x\|_2^2),$$

for some $C > 0$ depending only on $M_1(b)$ and $M_1(\sigma)$. The dominated convergence theorem now yields the desired pointwise convergence.

To prove the strong continuity of $(P_t)_{t\geq 0}$, it suffices, by [23, Thm. I.5.8, p. 40], to verify that $(P_t)_{t\geq 0}$ is weakly continuous, i.e., that $l(P_t f) \to l(f)$, as $t \to 0^+$, for all elements $l$ of the dual space $L^*$. To this end, fix any $l \in L^*$. By the Riesz-Markov theorem for $L$ [17, Theorem 2.4], there exists a finite signed Radon measure $\mu$ such that

$$l(f) = \int_{\mathbb{R}^d} f(x)\mu(dx) \text{ and } \int_{\mathbb{R}^d}(1 + \|x\|_2^2)|\mu|(dx) = \|l\|_{L^*}, \tag{18}$$

for $\|\cdot\|_{L^*}$ the dual norm. By Jensen's inequality and [27, Sec. 5, Cor. 1.2],

$$\forall t, \|(P_t f)(x)\|_2 \leq \mathbb{E}[|f(Z_{t,x})|] \leq \|f\|_L \mathbb{E}\big[1 + \|Z_{t,x}\|_2^2\big] \leq C\|f\|_L(1 + \|x\|_2^2).$$

Since $1 + \|x\|_2^2$ is $|\mu|$-integrable by (18), dominated convergence gives

$$\lim_{t\to 0^+} l(P_t f) = \lim_{t\to 0^+} \int_{\mathbb{R}^d}(P_t f)(x)\mu(dx) = \int_{\mathbb{R}^d} f(x)\mu(dx) = l(f),$$

yielding the result. $\square$

Consider the infinitesimal generator $\mathcal{A}$ of the semigroup $(P_t)_{t\geq 0}$ on $L$ with

$$\mathrm{dom}(\mathcal{A}) = \big\{f \in L : \lim_{t\to 0^+} \tfrac{P_t f - f}{t} \text{ exists in the } \|\cdot\|_L \text{ norm}\big\}.$$

Since $P_t$ is strongly continuous on $L$ and $h \in L$, [24, Prop. 1.5] implies that

$$h - P_t h = -\mathcal{A}\int_0^t P_s h\, ds = \mathcal{A}u_{h,t} \quad \text{for} \quad u_{h,t} \triangleq -\int_0^t P_s h\, ds.$$

The stationarity of $P$ and the definitions of $d_{\mathcal{W}_{\|\cdot\|_2}}$ and $r$ imply that

$$\|P_t h\|_L \leq \sup_{x\in\mathbb{R}^d} \tfrac{\mathbb{E}_P[d_{\mathcal{W}_{\|\cdot\|_2}}(\delta_x P_t, \delta_Z P_t)]}{1+\|x\|_2^2} \leq r(t) \sup_{x\in R^d} \tfrac{\mathbb{E}_P[\|x-Z\|_2]}{1+\|x\|_2^2},$$

and hence $\|P_t h\|_L \to 0$ as $t \to \infty$, since $P$ has a finite mean, and $r(t) \to 0$ as $t \to \infty$ as $r$ is integrable and monotonic. Arguing similarly,

$$\|u_{h,t} - u_{h,t'}\|_L \leq \|\int_t^{t'} \mathbb{E}_P[d_{\mathcal{W}_{\|\cdot\|_2}}(\delta_x P_s, \delta_Z P_s)]ds\|_L \leq \sup_{x\in R^d} \tfrac{\mathbb{E}_P[\|x-Z\|_2]}{1+\|x\|_2^2} \int_t^{t'} r(s)\, ds.$$

Thus, it follows that $(u_{h,t})_{t>0}$ is a Cauchy sequence in $L$ with limit $u_h = \int_0^\infty P_s h\, ds \in L$. Thus, $(h - P_t h, u_{h,t}) \to (h, u_h)$ in the graph norm on $L \times L$, and since $\mathcal{A}$ is closed [24, Cor. 1.6], $u_h \in \mathrm{dom}(\mathcal{A})$ and $h = \mathcal{A} u_h$.

Remark. The choice of the Banach space is crucial for the argument above. As noted in [63] and contrary to the claim in [5], the semigroup $(P_t)_{t\geq 0}$ fails to be strongly continuous over the Banach space $\widetilde{L} \triangleq (1 + \|x\|_2^2)C_b(\mathbb{R}^d)$ when $(Z_{t,x})_{t\geq 0}$ is an Ornstein-Uhlenbeck process, i.e., a Langevin diffusion (D1) with a multivariate Gaussian invariant measure.

*Lipschitz continuity of $u_h$.* To demonstrate that $u_h$ is Lipschitz, we choose an arbitrary $v \in \mathbb{R}^d$, and apply the definition of the Wasserstein distance, the assumed decay rate, and the integrability of $r$ to obtain

$$\begin{aligned}&\|u_h(x+v) - u_h(x)\|_2 \leq \int_0^\infty \|\mathbb{E}[h(Z_{t,x}) - h(Z_{t,x+v})]\|_2\, dt \\ &\leq \int_0^\infty d_{\mathcal{W}_{\|\cdot\|}}(\delta_x P_t, \delta_{x+v} P_t)\, dt \leq d_{\mathcal{W}_{\|\cdot\|}}(\delta_x, \delta_{x+v})\, s_r = \|v\|_2\, s_r < \infty.\end{aligned}$$

*Continuity of $\nabla^2 u_h$.* Since $u_h \in \mathrm{dom}(\mathcal{A})$ is a continuous solution of the Poisson equation (17), and since the infinitesimal generator agrees with the characteristic operator of a diffusion when both are defined [71, p. 129], Thm. 5.9 of [21] implies that $u_h \in C^2$.

*Boundedness of $\nabla^2 u_h$.* Now, instantiate the additional smoothness assumptions on $b$, $\sigma$, and $h$, and assume that $M_0(\sigma^{-1}), F_2(\sigma), M_2(b) < \infty$, or else (11) is vacuous. Lemma 15, established in Section C, shows that the semigroup $P_t h$ admits a bounded continuous Hessian, which is integrable in $t$.

Lemma 15 (Semigroup Hessian estimate). *Suppose that the drift and diffusion coefficients $b$ and $\sigma$ of an Itô diffusion are Lipschitz with Lipschitz gradients and locally Lipschitz second derivatives. If the transition semigroup $(P_t)_{t\geq 0}$ has Wasserstein decay rate $r$, and $\sigma(x)$ has a right inverse $\sigma^{-1}(x)$ for each $x \in \mathbb{R}^d$, then, for all $t > 0$ and any $f \in C^2$ with bounded first and second derivatives, $P_t f$ is twice continuously differentiable with*

$$M_1(P_t f) \leq M_1(f) r(t) \quad \textit{and} \tag{19}$$

$$M_2(P_t f) \leq \inf_{t_0 \in (0,t]} M_1(f) r(t-t_0) \sqrt{\tfrac{2}{t_0}} r(t_0) e^{t_0 \gamma_2/2} M_0(\sigma^{-1}) + M_1(f) r(t-t_0) r(t_0/2) e^{t_0\gamma_2/2} M_1(\sigma) M_0(\sigma^{-1}) + M_1(f) r(t-t_0) \sqrt{t_0}\, r(t_0/2) e^{t_0\gamma_4/2} \sqrt{\alpha} \tag{20}$$

*for* $\gamma_\rho \triangleq \rho M_1(b) + \frac{\rho^2 - 2\rho}{2} M_1(\sigma)^2 + \frac{\rho}{2} F_1(\sigma)^2$, $\alpha \triangleq \frac{M_2(b)^2}{2M_1(b) + 4M_1(\sigma)^2} + 2F_2(\sigma)^2$.

The dominated convergence theorem now implies that the Hessian of $u_h$ is obtained by differentiating twice under the integral sign. The advertised bound (11) on $\nabla^2 u_h$ follows by replacing the infimum on the right-hand side of the semigroup bound (20) with the selection $t_0 = \min(t, 1)$ and integrating the result over $t$.

## APPENDIX C: PROOF OF LEMMA 15

Fix any $x \in \mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$ in $C^2$ with bounded first and second derivatives, and let $(Z_{t,x})_{t\geq 0}$ be an Itô diffusion solving the stochastic differential equation (5) with starting point $Z_{0,x} = x$, underlying Wiener process $(W_t)_{t\geq 0}$, and transition semigroup $(P_t)_{t\geq 0}$. Our proof is divided into five pieces establishing, for each $t > 0$, the Lipschitz continuity of $P_t f$, the Lipschitz continuity of $\nabla P_t f$, the continuity of $\nabla^2 P_t f$, an initial bound on $\nabla^2 P_t f$, and the infimal bound (20) on $\nabla^2 P_t f$.

*Lipschitz continuity of* $P_t f$. The semigroup gradient bound (19) follows from the Lipschitz continuity of $f$ and the definitions of the Wasserstein decay rate and the Wasserstein distance, as, for any $y \in \mathbb{R}^d$ and $t \geq 0$,

$$(P_t f)(x) - (P_t f)(y) = \mathbb{E}[f(Z_{t,x}) - f(Z_{t,y})] \leq M_1(f) d_{\mathcal{W}_{\|\cdot\|_2}}(\delta_x P_t, \delta_y P_t) \leq M_1(f) r(t)\, d_{\mathcal{W}_{\|\cdot\|_2}}(\delta_x, \delta_y) = M_1(f) r(t) \|x - y\|_2.$$

*Lipschitz continuity of* $\nabla P_t f$. Fix any $v, v' \in \mathbb{R}^d$. Under our smoothness assumptions on $b$ and $\sigma$, [76, Theorem V.40] implies that $(Z_{t,x})_{t\geq 0}$ is twice continuously differentiable in $x$. The first directional *derivative flow* $(V_{t,v})_{t\geq 0}$ solves the *first variation equation*,

$$dV_{t,v} = \nabla b(Z_{t,x}) V_{t,v}\, dt + \nabla\sigma(Z_{t,x}) V_{t,v}\, dW_t \quad \text{with} \quad V_{0,v} = v, \tag{21}$$

obtained by formally differentiating the equation (5) defining $(Z_{t,x})_{t\geq 0}$ with respect to $x$ in the direction $v$. The second directional derivative flow $(U_{t,v,v'})_{t\geq 0}$ solves the *second variation equation*,

$$dU_{t,v,v'} = (\nabla b(Z_{t,x}) U_{t,v,v'} + \nabla^2 b(Z_{t,x})[V_{t,v'}] V_{t,v})\, dt + (\nabla\sigma(Z_{t,x}) U_{t,v,v'} + \nabla^2\sigma(Z_{t,x})[V_{t,v'}] V_{t,v})\, dW_t \;\text{ with }\; U_{0,v,v'} = 0, \tag{22}$$

obtained by differentiating (21) with respect to $x$ in the direction $v'$.

Since $f$ has bounded first and second derivatives, the dominated convergence theorem implies that, for each $t \geq 0$, $P_t f$ is twice differentiable with

$$\langle \nabla (P_t f)(x), v\rangle = \mathbb{E}[\langle \nabla f(Z_{t,x}), V_{t,v}\rangle] \quad \text{and}$$
$$v'^{\top}\nabla^2 (P_t f)(x) v = \mathbb{E}\Big[V_{t,v'}^{\top}\nabla^2 f(Z_{t,x})V_{t,v} + \langle \nabla f(Z_{t,x}), U_{t,v,v'}\rangle\Big] \tag{23}$$

obtained by differentiating under the integral sign. Lemma 16, proved in Section C.1, justifies the exchanges of derivative and expectation by ensuring that the derivative flows have moments bounded uniformly in $x$.

LEMMA 16 (Derivative flow bounds). *Suppose that $(Z_{t,x})_{t\geq 0}$ is an Itô diffusion with starting point $Z_{0,x} = x \in \mathbb{R}^d$, driving Wiener process $(W_t)_{t\geq 0}$, and Lipschitz drift and diffusion coefficients $b$ and $\sigma$ with Lipschitz gradients and locally Lipschitz second derivatives. If $(V_{t,v})_{t\geq 0}$ and $(U_{t,v,v'})_{t\geq 0}$ respectively solve the stochastic differential equations (21) and (22) for $v, v' \in \mathbb{R}^d$, then, for any $\rho \geq 2$,*

$$\mathbb{E}[\|V_{t,v}\|_2^{\rho}] \leq \|v\|_2^{\rho}\, e^{t\gamma_\rho} \quad \textit{and} \tag{24}$$
$$\mathbb{E}\big[\|U_{t,v,v'}\|_2^2\big] \leq \alpha \|v\|_2^2 \|v'\|_2^2 t e^{t\gamma_4} \tag{25}$$

*for $\gamma_\rho \triangleq \rho M_1(b) + \frac{\rho^2-2\rho}{2} M_1(\sigma)^2 + \frac{\rho}{2} F_1(\sigma)^2$ and $\alpha \triangleq \frac{M_2(b)^2}{2M_1(b)+4M_1(\sigma)^2} + 2F_2(\sigma)^2$.*

Since $\nabla f$ and $\nabla^2 f$ are bounded, and $(V_{t,v})_{t\geq 0}, (V_{t,v'})_{t\geq 0}$, and $(U_{t,v,v'})_{t\geq 0}$ have second moments bounded uniformly in $x$ by Lemma 16, the Hessian formula (23) implies that $\nabla^2 P_t f$ is bounded and hence that $\nabla P_t f$ is Lipschitz continuous for each $t \geq 0$.

*Continuity of $\nabla^2 P_t f$.* Hereafter we assume that $M_0(\sigma^{-1}) < \infty$, as the semigroup Hessian bound (20) is otherwise vacuous.

The Lipschitz continuity of $f$ and the Itô diffusion moment bound of [50, Thm. 3.4, part 4] together imply that

$$\mathbb{E}\big[f(Z_{t,x})^2\big] \leq \mathbb{E}\big[(|f(x)| + \|Z_{t,x} - x\|_2 M_1(f))^2\big] < \infty$$

for all $t \geq 0$. Since $\sigma^{-1}$ is bounded, and $\nabla b$ and $\nabla \sigma$ are bounded and Lipschitz, [26, Prop. 3.2] gives the following Bismut-Elworthy-Li-type formula for the directional derivative of $P_t f$ for each $t > 0$:

$$\langle \nabla (P_t f)(x), v\rangle = \tfrac{1}{t}\mathbb{E}\Big[f(Z_{t,x}) \textstyle\int_0^t \langle \sigma^{-1}(Z_{s,x})V_{s,v},\, dW_s\rangle\Big],$$

By interchanging derivative and integral, the dominated convergence theorem now delivers the Hessian expression

$$\begin{aligned}
(26)\qquad & v'^\top \nabla^2(P_tf)(x)v = \mathbb{E}[J_{1,x} + J_{2,x} + J_{3,x}] \quad \text{for} \\
& J_{1,x} \triangleq \tfrac{1}{t}\langle \nabla(P_tf)(Z_{t,x}), V_{t,v'}\rangle \textstyle\int_0^t \langle \sigma^{-1}(Z_{s,x})V_{s,v},\, dW_s\rangle, \\
& J_{2,x} \triangleq \tfrac{1}{t} f(Z_{t,x}) \textstyle\int_0^t \langle \nabla\sigma^{-1}(Z_{s,x})[V_{s,v'}]V_{s,v},\, dW_s\rangle, \quad \text{and} \\
& J_{3,x} \triangleq \tfrac{1}{t} f(Z_{t,x}) \textstyle\int_0^t \langle \sigma^{-1}(Z_{s,x})U_{s,v,v'},\, dW_s\rangle,
\end{aligned}$$

for each $t > 0$, provided that $J_{1,x}, J_{2,x}$, and $J_{3,x}$ are continuous in $x$. The requisite continuity follows from the Lipschitz continuity of $\nabla P_t f$ and $f$, the boundedness of $\sigma^{-1}$, $\nabla\sigma$, and $\nabla^2\sigma$, and the controlled moment growth and Hölder continuity of $(Z_{t,x})_{t\geq 0}$, $(V_{t,v})_{t\geq 0}, (V_{t,v'})_{t\geq 0}$, and $(U_{t,v,v'})_{t\geq 0}$ as functions of $x$ [76, Theorem V.40]. The dominated convergence theorem further implies that $\nabla^2 P_t f$ is continuous for each $t > 0$.

*Initial bound on* $\nabla^2 P_t f$. Now, we fix any $t > 0$ and turn to bounding $\nabla^2 P_t f$ in terms of $M_1(f)$, by bounding the expectations of $J_{1,x}, J_{2,x}$, and $J_{3,x}$ of (26) in turn.

To control $\mathbb{E}[J_{1,x}]$, we apply Cauchy-Schwarz, the Itô isometry [27, Eqs. 7.1 and 7.2], the semigroup gradient bound (19), the derivative flow bound (24), and the fact $e^{s\gamma_2} \leq e^{t\gamma_2}$ for all $s \leq t$ to obtain

$$\begin{aligned}
\mathbb{E}[J_{1,x}] &\leq \tfrac{1}{t}\sqrt{\mathbb{E}\big[\langle \nabla(P_tf)(Z_{t,x}), V_{t,v'}\rangle^2\big]\mathbb{E}[(\textstyle\int_0^t \langle \sigma^{-1}(Z_{s,x})V_{s,v},\, dW_s\rangle)^2]} \\
&\leq \tfrac{1}{t} M_1(P_tf)\sqrt{\mathbb{E}\big[\|V_{t,v'}\|_2^2\big] \textstyle\int_0^t \mathbb{E}\big[\|\sigma^{-1}(Z_{s,x})V_{s,v}\|_2^2\big]\, ds} \\
&\leq \tfrac{1}{t} M_1(P_tf) M_0(\sigma^{-1})\sqrt{\mathbb{E}\big[\|V_{t,v'}\|_2^2\big] \textstyle\int_0^t \mathbb{E}\big[\|V_{s,v}\|_2^2\big]\, ds} \\
&\leq \tfrac{1}{t} M_1(f) r(t) M_0(\sigma^{-1}) \|v'\|_2 \|v\|_2 \sqrt{e^{t\gamma_2} \textstyle\int_0^t e^{s\gamma_2}\, ds} \\
&\leq \sqrt{\tfrac{1}{t}}\, r(t) e^{t\gamma_2} M_1(f) M_0(\sigma^{-1}) \|v'\|_2 \|v\|_2,
\end{aligned}$$

where we have adopted the definition of $\gamma_\rho$ given in Lemma 16.

To control $\mathbb{E}[J_{2,x}]$, we will first rewrite the unbounded quantity $f(Z_{t,x})$ in terms of more manageable semigroup gradients. To this end, we note that, since $P_{t-s}f \in C^2$ for all $s \in [0,t]$, we may apply Itô's formula [27, Thm. 7.1] to $(s,x) \mapsto P_{t-s}f(x)$ to obtain the identity

$$(27)\qquad f(Z_{t,x}) = (P_tf)(x) + \textstyle\int_0^t \langle \nabla(P_{t-s}f)(Z_{s,x}), \sigma(Z_{s,x})\, dW_s\rangle.$$

Now we may rewrite $\mathbb{E}[J_{2,x}]$ as

$$\begin{aligned}
\mathbb{E}[J_{2,x}] &= \tfrac{1}{t}\mathbb{E}\Big[(P_t f)(x) \int_0^t \langle \nabla\sigma^{-1}(Z_{s,x})[V_{s,v'}]V_{s,v},\, dW_s\rangle \\
&\qquad + \int_0^t \langle \nabla(P_{t-s}f)(Z_{s,x}), \sigma(Z_{s,x})\, dW_s\rangle \int_0^t \langle \nabla\sigma^{-1}(Z_{s,x})[V_{s,v'}]V_{s,v},\, dW_s\rangle\Big] \\
&= \tfrac{1}{t}\mathbb{E}\Big[\int_0^t \langle \nabla(P_{t-s}f)(Z_{s,x}), \sigma(Z_{s,x})\nabla\sigma^{-1}(Z_{s,x})[V_{s,v'}]V_{s,v}\rangle\, ds\Big] \\
&= -\tfrac{1}{t}\mathbb{E}\Big[\int_0^t \langle \nabla(P_{t-s}f)(Z_{s,x}), \nabla\sigma(Z_{s,x})[V_{s,v'}]\sigma^{-1}(Z_{s,x})V_{s,v}\rangle\, ds\Big],
\end{aligned}$$

where we have used Dynkin's formula [27, Eq. 7.11], the Itô isometry, and the chain rule,

$$\nabla\sigma^{-1}(x)[v] = -\sigma^{-1}(x)\nabla\sigma(x)[v]\sigma^{-1}(x). \tag{28}$$

Finally, we bound $\mathbb{E}[J_{2,x}]$ using Cauchy-Schwarz, the semigroup gradient bound (19), the derivative flow bound (24), and the fact that $s \mapsto r(t-s)e^{s\gamma_2}$ is increasing:

$$\begin{aligned}
\mathbb{E}[J_{2,x}] &\le \tfrac{1}{t}M_1(\sigma)M_0(\sigma^{-1}) \int_0^t M_1(P_{t-s}f)\mathbb{E}\big[\|V_{s,v'}\|_2\|V_{s,v}\|_2\big]\, ds \\
&\le \tfrac{1}{t}M_1(\sigma)M_0(\sigma^{-1}) \int_0^t M_1(P_{t-s}f)\sqrt{\mathbb{E}\big[\|V_{s,v'}\|_2^2\big]\mathbb{E}\big[\|V_{s,v}\|_2^2\big]}\, ds \\
&\le \tfrac{1}{t}M_1(\sigma)M_0(\sigma^{-1})M_1(f)\|v'\|_2\|v\|_2 \int_0^t r(t-s)e^{s\gamma_2}\, ds \\
&\le r(0)e^{t\gamma_2}M_1(\sigma)M_0(\sigma^{-1})M_1(f)\|v'\|_2\|v\|_2.
\end{aligned}$$

To control $\mathbb{E}[J_{3,x}]$, we again appeal to Dynkin's formula and the Itô isometry to obtain

$$\begin{aligned}
\mathbb{E}[J_{3,x}] &= \tfrac{1}{t}\mathbb{E}\Big[(P_t f)(x) \int_0^t \langle \sigma^{-1}(Z_{s,x})U_{s,v,v'},\, dW_s\rangle \\
&\qquad + \int_0^t \langle \nabla(P_{t-s}f)(Z_{s,x}), \sigma(Z_{s,x})\, dW_s\rangle \int_0^t \langle \sigma^{-1}(Z_{s,x})U_{s,v,v'},\, dW_s\rangle\Big] \\
&= \mathbb{E}\Big[\int_0^t \langle \nabla(P_{t-s}f)(Z_{s,x}), U_{s,v,v'}\rangle\, ds\Big],
\end{aligned}$$

and we bound this expression using Cauchy-Schwarz, Jensen's inequality, the semigroup gradient bound (19), the second derivative flow bound (25), and the fact that $s \mapsto r(t-s)e^{s\gamma_4}$ is increasing:

$$\begin{aligned}
\mathbb{E}[J_{3,x}] &\le \tfrac{1}{t}\int_0^t M_1(P_{t-s}f)\mathbb{E}[\|U_s\|_2]\, ds \le \tfrac{1}{t}\int_0^t M_1(P_{t-s}f)\sqrt{\mathbb{E}\big[\|U_s\|_2^2\big]}\, ds \\
&\le \tfrac{1}{t}M_1(f)\sqrt{\alpha}\|v'\|_2\|v\|_2 \int_0^t r(t-s)\sqrt{s}e^{s\gamma_4}\, ds \\
&\le \tfrac{2}{3}\sqrt{t}\, r(0)e^{t\gamma_4}M_1(f)\sqrt{\alpha}\|v'\|_2\|v\|_2,
\end{aligned}$$

where $\alpha$ is defined in Lemma 16. The advertised result (20) for $t_0 = t$ follows by summing the bounds developed for $\mathbb{E}[J_{1,x}], \mathbb{E}[J_{2,x}]$, and $\mathbb{E}[J_{3,x}]$.

*Infimal bound on $\nabla^2 P_t f$.* To obtain the infimum over $t_0 \in (0,t]$ in (20), we adapt an argument of [9, Prop. 1.5.1]. Specifically, fix any $t_0 \in (0,t]$. Our work thus far shows that $v'^\top \nabla^2 (P_{t_0}\tilde{f})(x)v \leq M_1(\tilde{f})\zeta(t_0)$ for a real-valued function $\zeta$ and $\tilde{f} \in C^2$ with bounded first and second derivatives. Since we now know that $P_{t-t_0} f \in C^2$ with bounded first and second derivatives, the Markov property of the diffusion and the first derivative bound (19) yield

$$\begin{aligned} v'^\top \nabla^2 (P_t f)(x) v &= v'^\top \nabla^2 (P_{t_0} P_{t-t_0} f)(x) v \\ &\leq M_1(P_{t-t_0} f)\zeta(t_0) \leq M_1(f) r(t-t_0)\zeta(t_0). \end{aligned}$$

**C.1. Proof of Lemma 16: Derivative flow bounds.** Fix any $\rho \geq 2$ and $v \in \mathbb{R}^d$. Since Dynkin's formula and Cauchy-Schwarz give

$$\begin{aligned} &\mathbb{E}[\|V_{s,v}\|_2^\rho] = \|v\|_2^\rho + \mathbb{E}\Big[\textstyle\int_0^t \rho\langle V_{s,v}\|V_{s,v}\|_2^{\rho-2}, \nabla b(Z_{s,x})V_{s,v}\rangle \\ &+ \tfrac{\rho}{2}\|V_{s,v}\|_2^{\rho-4}((\rho-2)\|V_{s,v}^\top \nabla\sigma(Z_{s,x})[V_{s,v}]\|_2^2 + \|V_{s,v}\|_2^2 \|\nabla\sigma(Z_{s,x})[V_{s,v}]\|_F^2)\, ds\Big] \\ &\leq \|v\|_2^\rho + \textstyle\int_0^t (\rho M_1(b) + \frac{\rho^2-2\rho}{2} M_1(\sigma)^2 + \frac{\rho}{2}F_1(\sigma)^2)\mathbb{E}[\|V_{s,v}\|_2^\rho]\, ds, \end{aligned}$$

the advertised result (24) follows from Grönwall's inequality.

Now fix any $v, v' \in \mathbb{R}^d$, and define $U_t \triangleq U_{t,v,v'}$. Dynkin's formula and multiple applications of Cauchy-Schwarz and Young's inequality give

$$\begin{aligned} \mathbb{E}\big[\|U_t\|_2^2\big] &= \mathbb{E}\Big[\textstyle\int_0^t 2\langle U_s, \nabla b(Z_{s,x})U_s + \nabla^2 b(Z_{s,x})[V_{s,v'}]V_{s,v}\rangle \\ &\qquad + \|\nabla\sigma(Z_{s,x})[U_s] + \nabla^2\sigma(Z_{s,x})[V_{s,v'}]V_{s,v}\|_F^2\, ds\Big] \\ &\leq \mathbb{E}\Big[\textstyle\int_0^t 2\|U_s\|_2^2 M_1(b) + 2\|U_s\|_2\|V_{s,v}\|_2\|V_{s,v'}\|_2 M_2(b) \\ &\qquad + 2\|\nabla\sigma(Z_{s,x})[U_s]\|_F^2 + 2\|\nabla^2\sigma(Z_{s,x})[V_{s,v'}]V_{s,v}\|_F^2\, ds\Big] \\ &\leq \textstyle\int_0^t (2M_1(b) + 2F_1(\sigma)^2 + \epsilon)\mathbb{E}\big[\|U_s\|_2^2\big] \\ &\qquad + (M_2(b)^2/\epsilon + 2F_2(\sigma)^2)\mathbb{E}\big[\|V_{s,v}\|_2^2\|V_{s,v'}\|_2^2\big]\, ds \end{aligned}$$

for any $\epsilon > 0$. Letting $\gamma_\rho = \rho M_1(b) + \frac{\rho^2-2\rho}{2}M_1(\sigma)^2 + \frac{\rho}{2}F_1(\sigma)^2$, we see that, by Cauchy-Schwarz and our derivative flow bound (24),

$$\begin{aligned} \textstyle\int_0^t \mathbb{E}\big[\|V_{s,v}\|_2^2\|V_{s,v'}\|_2^2\big]\, ds &\leq \textstyle\int_0^t \sqrt{\mathbb{E}\big[\|V_{s,v}\|_2^4\big]\mathbb{E}\big[\|V_{s,v'}\|_2^4\big]}\, ds \\ &\leq \textstyle\int_0^t \|v\|_2^2\|v'\|_2^2 e^{s\gamma_4}\, ds = \|v\|_2^2\|v'\|_2^2 \frac{e^{t\gamma_4}-1}{\gamma_4}. \end{aligned}$$

Hence, if we choose $\epsilon = \gamma_4 - (2M_1(b) + 2F_1(\sigma)^2)$ and define $\alpha = M_2(b)^2/\epsilon + 2F_2(\sigma)^2$ we may write

$$\mathbb{E}\big[\|U_t\|_2^2\big] \leq \alpha\|v\|_2^2\|v'\|_2^2 \tfrac{e^{t\gamma_4}-1}{\gamma_4} + \textstyle\int_0^t \gamma_4 \mathbb{E}\big[\|U_s\|_2^2\big]\, ds.$$

Gronwall's inequality now yields the result (25) via

$$\mathbb{E}\big[\|U_t\|_2^2\big] \le \alpha\|v\|_2^2\|v'\|_2^2\Big(\tfrac{e^{t\gamma_4}-1}{\gamma_4} + \int_0^t \tfrac{e^{s\gamma_4}-1}{\gamma_4}\gamma_4 e^{(t-s)\gamma_4}\,ds\Big) = \alpha\|v\|_2^2\|v'\|_2^2 t e^{t\gamma_4}.$$

## APPENDIX D: PROOF OF THEOREM 6

We first derive the result for $\|\cdot\| = \|\cdot\|_2$. Without loss of generality, assume $h \in \mathcal{W}_{\|\cdot\|_2}$ with $\mathbb{E}_P[h(Z)] = 0$. Our high-level strategy is to relate the Wasserstein distance to the Stein discrepancy via the Stein equation (3) with diffusion Stein operator $\mathcal{T}$ (8). Since the infinitesimal generator $\mathcal{A}$ (4) has the form (7) by Theorem 2, Theorem 5 implies that there exists a continuously differentiable solution $g_h$ to the the Stein equation $h(x) = (\mathcal{T}g_h)(x)$ satisfying $M_0(g_h) \le s_r M_1(h) \le s_r$. Since boundedness alone is insufficient to declare that $g_h$ falls into a scaled copy of the classical Stein set $\mathcal{G}_{\|\cdot\|}$, we will develop a smoothed version of the Stein solution with greater regularity.

Since $a$ and $c$ are constant, $b(x) = \frac{1}{2}(a+c)\nabla \log p(x)$. Fix any $s > 0$ and consider the convolution $g_{h,s}(x) \triangleq \mathbb{E}[g_h(x+sG)]$. If the smoothing level $s$ is small, the Lipschitz continuity of $h$ implies that that $(\mathcal{T}g_{h,s})(x)$ provides a close approximation to $h(x)$ for each $x \in \mathbb{R}^d$:

$$\begin{aligned}
(29)\quad h(x) &\le \mathbb{E}[h(x+sG)] + M_1(h)s\mathbb{E}[\|G\|_2] \\
&\le \mathbb{E}\Big[\tfrac{1}{p(x+sG)}\langle \nabla, p(x+sG)(a+c)g_h(x+sG)\rangle\Big] + s\mathbb{E}[\|G\|_2] \\
&\le 2\mathbb{E}[\langle b(x+sG), g_h(x+sG)\rangle] + \mathbb{E}[\langle a+c, \nabla g_h(x+sG)\rangle] + s\mathbb{E}[\|G\|_2] \\
&\le (\mathcal{T}g_{h,s})(x) + s\mathbb{E}[\|G\|_2](1 + 2M_1(b)M_0(g_h)).
\end{aligned}$$

Moreover, our next lemma, proved in Section D.1, shows that the smoothed Stein solution admits a bounded Lipschitz gradient $\nabla g_{h,s}(x) = \mathbb{E}[\nabla g_h(x+sG)]$.

LEMMA 17 (Smoothing by Gaussian convolution). *Let $G \in \mathbb{R}^d$ be a standard normal random vector, and fix $s > 0$. If $f : \mathbb{R}^d \to \mathbb{R}$ is bounded and measurable, and $f_s(x) \triangleq \mathbb{E}[f(x+sG)]$, then*

$$M_0(f_s) \le M_0(f), \quad M_1(f_s) \le \sqrt{\tfrac{2}{\pi}}\tfrac{M_0(f)}{s}, \quad \text{and} \quad M_2(f_s) \le \sqrt{2}\tfrac{M_0(f)}{s^2}.$$

*If, additionally, $f \in C^1$, then $\nabla f_s(x) = \mathbb{E}[\nabla f(x+sG)]$.*

Indeed, for each non-zero $w \in \mathbb{R}^d$, we may apply Lemma 17 to the function $f_w(x) \triangleq \langle w, g_h(x)\rangle/\|w\|_2$ with convolution $f_{w,s}(x) = \langle w, g_{h,s}(x)\rangle/\|w\|_2$ to obtain the bounds

$$M_0(g_{h,s}) = \sup_{w\neq 0} M_0(f_{w,s}) \le \sup_{w\neq 0} M_0(f_w) = M_0(g_h) \le s_r,$$

$M_1(g_{h,s}) = \sup_{w\neq 0} M_1(f_{w,s}) \leq \sup_{w\neq 0} \sqrt{\frac{2}{\pi}}\frac{M_1(f_w)}{s} = \sqrt{\frac{2}{\pi}}\frac{M_1(f_w)}{s} \leq \sqrt{\frac{2}{\pi}}\frac{s_r}{s}$, and
$M_2(g_{h,s}) = \sup_{w\neq 0} M_2(f_{w,s}) \leq \sup_{w\neq 0} \frac{\sqrt{2}M_2(f_w)}{s^2} = \frac{\sqrt{2}M_2(f_w)}{s^2} \leq \frac{\sqrt{2}\,s_r}{s^2}$.

Hence, since our choice of $h$ was arbitrary, and

$$\kappa_s \triangleq \max\Big(1, \tfrac{1}{s}\sqrt{\tfrac{2}{\pi}}, \tfrac{\sqrt{2}}{s^2}\Big) = \max\Big(1, \tfrac{\sqrt{2}}{s^2}\Big) \geq \tfrac{\max(M_0(g_{h,s}), M_1(g_{h,s}), M_2(g_{h,s}))}{s_r},$$

we may take expectation under $Q_n$ and supremum over $h$ in (29) to reach

$$\begin{aligned} d_{\mathcal{W}_{\|\cdot\|_2}}(\mu,\nu) &\leq \inf_{s>0} \mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|_2}) s_r \kappa_s + s\mathbb{E}[\|G\|_2](1 + 2M_1(b)s_r) \\ &\leq \max\big(\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|_2})s_r, \eta\big) + 2\eta \leq 3\max\big(\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|_2})s_r, \eta\big), \end{aligned}$$

where we define $\eta = \sqrt[3]{\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|_2})\sqrt{2}s_r\,\mathbb{E}[\|G\|_2]^2(1+2M_1(b)s_r)^2}$ and select $s = \sqrt[3]{\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|_2})2\sqrt{2}s_r/(\mathbb{E}[\|G\|_2](1+2M_1(b)s_r))}$ to produce the second inequality.

The generic norm result now follows from the assumed norm domination property $\|\cdot\| \geq \|\cdot\|_2$, which implies $\mathcal{G}_{\|\cdot\|_2} \subseteq \mathcal{G}_{\|\cdot\|}$.

**D.1. Proof of Lemma 17: Smoothing by Gaussian convolution.** The conclusion $M_0(f_s) \leq M_0(f)$ follows from Hölder's inequality. Now, fix any $x$ and non-zero $v_1, v_2 \in \mathbb{R}^d$. Since $f_s = f \star \phi_s$, where $\phi_s \in C^\infty$ is the density of $sG$ and $\star$ is the convolution operator, Leibniz's rule implies that

$$\begin{aligned} \langle v_1, \nabla f_s(x)\rangle &= \langle v_1, (f \star \nabla\phi_s)(x)\rangle = \tfrac{1}{s^2}\textstyle\int f(x-y)\langle v_1, y\rangle \phi_s(y)dy \\ &\leq \tfrac{M_0(f)}{s^2}\textstyle\int |\langle v_1, y\rangle|\phi_s(y)\,dy = \sqrt{\tfrac{2}{\pi}}\tfrac{M_0(f)}{s}\|v_1\|_2, \end{aligned}$$

as $\langle v_1, G\rangle/\|v_1\|_2$ has a standard normal distribution. Leibniz's rule also gives

$$\begin{aligned} \nabla^2 f_s(x)[v_1, v_2] &= (f \star \nabla^2\phi_s)(x)[v_1, v_2] \\ &\leq \tfrac{M_0(f)}{s^2}\textstyle\int_{\mathbb{R}^d}\big|\langle v_1, zz^\top v_2\rangle/s^2 - \langle v_1, v_2\rangle\big|\phi_s(z)\,dz \\ &\leq \tfrac{M_0(f)}{s^2}\sqrt{\textstyle\int_{\mathbb{R}^d}\big|\langle v_1, zz^\top v_2\rangle/s^2 - \langle v_1, v_2\rangle\big|^2\phi_s(z)\,dz} \\ &= \tfrac{M_0(f)}{s^2}\sqrt{\langle v_1, v_2\rangle^2 + \|v_1\|_2^2\|v_2\|_2^2} \leq \tfrac{\sqrt{2}M_0(f)}{s^2}\|v_1\|_2\|v_2\|_2, \end{aligned}$$

where the last equality follows by Isserlis' theorem. Finally, when $f \in C^1$, Leibniz's rule gives $\nabla f_s = \nabla f \star \phi_s$.

# APPENDIX E: PROOF OF THEOREM 7

We will derive each inequality for $\|\cdot\| = \|\cdot\|_2$; the generic norm results will then follow from the property $\|\cdot\| \geq \|\cdot\|_2$, which implies $\mathcal{G}_{\|\cdot\|_2} \subseteq \mathcal{G}_{\|\cdot\|}$.

Fix any $h \in \mathcal{H} = \{h : \mathbb{R}^d \to \mathbb{R} \mid h \in C^2, M_1(h) \leq 1, M_2(h) < \infty, M_3(h) < \infty\}$ with $\mathbb{E}_P[h(Z)] = 0$. We assume that $M_1(b)$, $M_2(b)$, $M_1(\sigma)$, $F_2(\sigma)$, $M_1^*(m)$, and $M_0(\sigma^{-1})$ are all finite, or else the results are vacuous. Our high-level strategy is to relate the Wasserstein distance to the Stein discrepancy via the Stein equation (3) with diffusion Stein operator $\mathcal{T}$ (8). By Theorem 5, we know that there exists a Lipschitz solution $g_h$ to the the Stein equation $h(x) = (\mathcal{T} g_h)(x)$ satisfying $M_0(g_h) \leq s_r M_1(h) \leq s_r$ and $M_1(g_h) \leq \beta M_1(h) \leq \beta$, for $\beta \triangleq \beta_1 + \beta_2$, where $\beta_1$ and $\beta_2$ are defined in Theorem 5. Since a Lipschitz gradient is also needed to declare that $g_h$ falls into a scaled copy of the classical Stein set $\mathcal{G}_{\|\cdot\|}$, we will develop a smoothed version of the Stein solution with greater regularity.

To this end, fix any $s > 0$ and consider the convolution $g_{h,s}(x) \triangleq \mathbb{E}[g_h(x + sG)]$. If the smoothing level $s$ is small, the Lipschitz continuity of $m$ and $h$ implies that $(\mathcal{T} g_{h,s})(x)$ closely approximates $h(x)$ for each $x \in \mathbb{R}^d$:

$$
\begin{aligned}
(30)\quad & h(x) \leq \mathbb{E}[h(x + sG)] + M_1(h) s \mathbb{E}[\|G\|_2] \\
& \leq 2\mathbb{E}[\langle b(x + sG), g_h(x + sG)\rangle + \langle m(x + sG), \nabla g_h(x + sG)\rangle] + s\mathbb{E}[\|G\|_2] \\
& \leq (\mathcal{T} g_{h,s})(x) + s\zeta.
\end{aligned}
$$

**E.1. Proof of the first inequality.** Moreover, by an argument mirroring that of Theorem 6, Lemma 17 shows that $g_{h,s}$ admits a Lipschitz gradient $\nabla g_{h,s}(x) = \mathbb{E}[\nabla g_h(x + sG)]$ and satisfies the derivative bounds

$$
\begin{aligned}
(31)\qquad & M_0(g_{h,s}) \leq M_0(g_h) \leq s_r, \\
& M_1(g_{h,s}) = M_0(\nabla g_{h,s}) \leq M_0(\nabla g_h) \leq \beta, \quad \text{and} \\
& M_2(g_{h,s}) = M_1(\nabla g_{h,s}) \leq \sqrt{\tfrac{2}{\pi}} \tfrac{M_0(\nabla g_h)}{s} \leq \sqrt{\tfrac{2}{\pi}} \tfrac{\beta}{s}.
\end{aligned}
$$

Hence, since $\mathcal{H}$ is dense in $\mathcal{W}_{\|\cdot\|_2}$, and we may take expectation under $Q_n$ and supremum over $h$ in (30) to reach

$$
\begin{aligned}
& d_{\mathcal{W}_{\|\cdot\|_2}}(\mu, \nu) \leq \inf_{s>0} \mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|_2}) \max\Big(s_r, \beta, \sqrt{\tfrac{2}{\pi}} \tfrac{\beta}{s}\Big) + s\zeta \\
& \leq \max(\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|_2}) \max(s_r, \beta), \eta) + \eta \leq 2\max(\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|_2}) \max(s_r, \beta), \eta),
\end{aligned}
$$

where $\eta \triangleq s^*\zeta$ for $s^* = \sqrt{\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_{\|\cdot\|_2}) \sqrt{2/\pi}\beta/\zeta}$.

**E.2. Proof of the second inequality.** Assume now that $\nabla^3 b$ and $\nabla^3\sigma$ are bounded and locally Lipschitz, and consider the following extension of the seminorm $M_k$ to positive $k \notin \mathbb{N}$,

$$M_k(g) \triangleq \sup_{x,y\in\mathbb{R}^d, x\neq y} \frac{\|\nabla^{\lceil k\rceil-1}g(x)-\nabla^{\lceil k\rceil-1}g(y)\|_{op}}{\|x-y\|_2^{\{k\}}},$$

where $\{k\} \triangleq k - \lfloor k\rfloor$ denotes the fractional part of $k$. Fix any $\iota \in (0,1)$. Lemma 17 and an auxiliary smoothing lemma (Lemma 19 in the supplement) imply that $M_2(g_{h,s}) = M_1(\nabla g_{h,s}) \leq \sqrt{d}\frac{M_{1-\iota}(\nabla g_h)}{s^\iota}$. This improved dependence on $s$ will allow us to establish a near-linear relationship between the Stein discrepancy and the Wasserstein distance. The following lemma, proved in Section E.3, provides a suitable bound on $M_{1-\iota}(\nabla g_h)$.

LEMMA 18 (Hölder continuity of $\nabla g_h$). *Under the assumptions of Theorem 7, if $\nabla^3 b$ and $\nabla^3\sigma$ are bounded and locally Lipschitz, then there exists a constant $K > 0$, depending only on $M_{1:3}(b), M_{1:3}(\sigma), M_0(\sigma^{-1})$, and $r$, such that for all $\iota \in (0,1)$, all $h \in C^3$ with bounded second and third derivatives, and $g_h$ the Stein solution defined in Theorem 5,*

$$M_{1-\iota}(\nabla g_h) \leq \tfrac{1}{K}\big(\tfrac{1}{\iota} + s_r\big)M_1(h).$$

REMARK. The constant can be in principle be traced through the proof.

Hence, $M_2(g_{h,s}) \leq C_\iota/s^\iota$ for $C_\iota \triangleq \frac{\sqrt{d}}{K}(\frac{1}{\iota} + s_r)$. Following the derivation in Section E.1 and choosing $s^* = \Big(\frac{\iota C_\iota \mathcal{S}(Q_n,\mathcal{T},\mathcal{G}_{\|\cdot\|_2})}{\zeta}\Big)^{\frac{1}{\iota+1}}$ and $\eta \triangleq \frac{\zeta}{\iota}s^*$, we obtain

$$\begin{aligned} &d_{\mathcal{W}_{\|\cdot\|_2}}(P,Q_n) \leq \inf_{s>0} \mathcal{S}(Q_n,\mathcal{T},\mathcal{G}_{\|\cdot\|_2}) \max(s_r,\beta,C_\iota s^{-\iota}) + s\zeta \\ &\leq \max(\mathcal{S}(Q_n,\mathcal{T},\mathcal{G}_{\|\cdot\|_2})\max(s_r,\beta),\eta) + \eta\iota \leq 2\max(\mathcal{S}(Q_n,\mathcal{T},\mathcal{G}_{\|\cdot\|_2})\max(s_r,\beta),\eta). \end{aligned}$$

**E.3. Proof of Lemma 18: Hölder continuity of $\nabla g_h$.** We assume $M_1(h) \leq 1$ and obtain the general result by scaling. Fix any $\iota \in (0,1)$. The integral representation (10) of $u_h = \frac{1}{2}\nabla g_h$, the dominated convergence theorem, and Jensen's inequality imply

$$M_{1-\iota}(\nabla g_h) = M_{1-\iota}\big(-\textstyle\int_0^\infty \frac{1}{2}\nabla^2 P_t h\, dt\big) \leq \frac{1}{2}\int_0^\infty M_{1-\iota}(\nabla^2 P_t h)\, dt.$$

When $t \leq t_0 \triangleq 1$, a seminorm interpolation lemma (Lemma 20 in the supplement), a semigroup third derivative estimate (Lemma 21 in the supplement), and the semigroup second derivative estimate of Lemma 15 imply

$$M_{1-\iota}(\nabla^2 P_t h) \leq 2^\iota M_0(\nabla^2 P_t h)^\iota M_1(\nabla^2 P_t h)^{1-\iota} \leq t^{\iota/2-1}/K_1$$

for some constant $K_1 > 0$ depending only on $M_{1:3}(b), M_{1:3}(\sigma), M_0(\sigma^{-1})$, and $r$. Thus $\int_0^1 M_{1-\iota}(\nabla^2 P_t h)\,dt \le \frac{2}{K_1\iota}$. For $t > t_0$, Lemmas 20, 21, and 15 and the integrability of $r$ yield

$$\int_1^\infty M_{1-\iota}(\nabla^2 P_t h)\,dt \le \frac{2}{K_2}\int_1^\infty r(t-1)\,dt = \frac{2}{K_2}s_r$$

for a constant $K_2 > 0$ again depending only on $M_{1:3}(b), M_{1:3}(\sigma), M_0(\sigma^{-1})$, and $r$. Combining these bounds and choosing $K = \min(K_1, K_2)$ finishes the proof. An explicit constant $K$ can be obtained by tracing constants through the proof of Lemma 21.

## APPENDIX F: PROOF OF PROPOSITION 8

Fix any $g \in \mathcal{G}_{\|\cdot\|}$. Since $\mathbb{E}_P[(\mathcal{T}g)(Z)] = 0$ by Proposition 3, we may write

$$\begin{aligned}
|\mathbb{E}_{Q_n}[(\mathcal{T}g)(X)]| &= |\mathbb{E}_{Q_n}[(\mathcal{T}g)(X)] - \mathbb{E}_P[(\mathcal{T}g)(Z)]| \\
&= |2\mathbb{E}[\langle b(X) - b(Z), g(X)\rangle + \langle b(Z), g(X) - g(Z)\rangle] \\
&\quad + \mathbb{E}[\langle m(X) - m(Z), \nabla g(X)\rangle + \langle m(Z), \nabla g(X) - \nabla g(Z)\rangle]|. \qquad (32)
\end{aligned}$$

for any coupling of $X$ and $Z$. We obtain the first advertised inequality by repeatedly applying the Fenchel-Young inequality for dual norms, invoking the boundedness and Lipschitz constraints on $g$ and $\nabla g$, and taking a supremum over $g \in \mathcal{G}_{\|\cdot\|}$. The second inequality follows by invoking Jensen's inequality, the fact $\min(x, y) \le x^t y^{1-t}$ for all $x, y \ge 0$, Hölder's inequality, and finally the definition of $\mathcal{W}_{s,\|\cdot\|}$.

## APPENDIX G: PROOF OF THEOREM 10

Fix any $x, y \in \mathbb{R}^d$, and define two Itô diffusions solving $dZ_{t,x} = b(Z_{t,x})\,dt + \sigma(Z_{t,x})\,dW_t$ with $Z_{0,x} = x$ and $dZ_{t,y} = b(Z_{t,y})\,dt + \sigma(Z_{t,y})\,dW_t$ with $Z_{0,y} = y$, for $(W_t)_{t\ge 0}$ a shared Wiener process. Applying Dynkin's formula to the function $f(t, x) = e^{kt}\|x\|_G^2$ for the difference process $Z_{t,x} - Z_{t,y}$ yields

$$\begin{aligned}
&\mathbb{E}[f(t, Z_{t,x} - Z_{t,y})] = \|x - y\|_G^2 + \mathbb{E}[\textstyle\int_0^t k e^{ks}\|Z_{s,x} - Z_{s,y}\|_G^2\,ds] \\
&\quad + \mathbb{E}[\textstyle\int_0^t e^{ks}\big(\|\sigma(Z_{s,x}) - \sigma(Z_{s,y})\|_G^2 + 2\langle b(Z_{s,x}) - b(Z_{s,y}), G(Z_{s,x} - Z_{s,y})\rangle\big)\,ds]
\end{aligned}$$

By the uniform dissipativity assumption, the right-hand side is at most $\|x - y\|_G^2 = d_{\mathcal{W}_{\|\cdot\|_G}}(\delta_x, \delta_y)^2$. For the transition semigroup $(P_t)_{t\ge 0}$.

$$\mathbb{E}[f(t, Z_{t,x} - Z_{t,y})] = e^{kt}\mathbb{E}\big[\|Z_{t,x} - Z_{t,y}\|_G^2\big] \ge d_{\mathcal{W}_{\|\cdot\|_G}}(\delta_x P_t, \delta_y P_t)^2,$$

by Cauchy-Schwarz. The result now follows from the fact that $\lambda_{\min}(G_1) \le \|z\|_G^2/\|z\|_2^2 \le \lambda_{\max}(G_1)$ for all $z \ne 0$.

## APPENDIX H: PROOF OF THEOREM 11

As in the proof of [92, Thm. 2.6], we fix two arbitrary starting points $x, y \in \mathbb{R}^d$ and define a pair of coupled Itô diffusions $(Z_{t,x})_{t\geq 0}$ and $(Z_{t,y})_{t\geq 0}$, each with associated marginal semigroup $(P_t)_{t\geq 0}$. Specifically, we set $Z_{0,x} = x$ and $Z_{0,y} = y$ and let $(Z_{t,x})_{t\geq 0}$ and $(Z_{t,y})_{t\geq 0}$ solve the equations

$$
\begin{aligned}
dZ_{t,x} &= b(Z_{t,x})\,dt + \sigma_0(Z_{t,x})\,dW_t' + \lambda_0\,dW_t''\\
dZ_{t,y} &= b(Z_{t,y})\,dt + \sigma_0(Z_{t,y})\,dW_t' + \lambda_0\big(I - 2\tfrac{Z_{t,x}-Z_{t,y}}{\|Z_{t,x}-Z_{t,y}\|_2}\tfrac{Z_{t,x}-Z_{t,y}^\top}{\|Z_{t,x}-Z_{t,y}\|_2}\big)\,dW_t'',
\end{aligned}
$$

where $(W_t')_{t\geq 0}$ is an $m$-dimensional Wiener process and $(W_t'')_{t\geq 0}$ is an independent $d$-dimensional Wiener process.

Following the argument of Eberle [22, Sec. 4], we define the difference process $Y_t = Z_{t,x} - Z_{t,y}$, its norm $r_t = \|Y_t\|_2$, and the one-dimensional Wiener process $W_t = \int_0^t \langle Y_s/r_s,\, dW_s''\rangle$, and apply the generalized Itô formula [48, Thm. 22.5] to obtain the stochastic differential equations

$$
\begin{aligned}
d\|Y_t\|_2^2 = {}&(2\langle Y_t, b(Z_{t,x}) - b(Z_{t,y})\rangle + \|\sigma_0(Z_{t,x}) - \sigma_0(Z_{t,y})\|_F^2 + 4\lambda_0^2)\,dt\\
&+ 2\langle Y_t, (\sigma_0(Z_{t,x}) - \sigma_0(Z_{t,y}))\,dW_t'\rangle + 4\lambda_0\|Y_t\|_2\,dW_t
\end{aligned}
$$

and

$$
\begin{aligned}
&df(r_t) = f'(r_t)/(r_t)\langle Y_t, (\sigma_0(Z_{t,x}) - \sigma_0(Z_{t,y}))\,dW_t'\rangle + 2\lambda_0 f'(r_t)\,dW_t\\
&+ (f''(r_t)(2\lambda_0^2 + \tfrac{1}{2}\|(\sigma_0(Z_{t,x}) - \sigma_0(Z_{t,y}))^\top Y_t\|_2^2/r_t^2) - \tfrac{1}{2\alpha}f'(r_t)\kappa(r_t)r_t)\,dt
\end{aligned}
$$

for any concave increasing $f : [0,\infty) \mapsto [0,\infty)$ with absolutely continuous derivative, $f(0) = 0$, and $f'(0) = 1$. Since the drift term in the latter equation is bounded above by

$$
\beta_t \triangleq (2/\alpha)(f''(r_t) - (1/4)f'(r_t)\kappa(r_t)r_t),
$$

the argument of [22, p. 15] shows that the results of [22, Thm. 1 and Cor. 2] hold for our choice of $\alpha$ and $\kappa$.

## ACKNOWLEDGMENTS

We thank Simon Lacoste-Julien for sharing his quadrature code, Martin Hairer for discussing interpolation inequalities, and Andreas Eberle for reading an earlier version of this manuscript. This material is based upon work supported by the grant EPSRC EP/N000188/1, the National Science Foundation DMS RTG Grant No. 1501767, the National Science Foundation Graduate Research Fellowship under Grant No. DGE-114747, and the Frederick E. Terman Fellowship.

## REFERENCES


[1] S. Ahn, A. Korattikara, and M. Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *Proc. 29th ICML*, ICML'12, 2012.

[2] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford University Press, 2000.

[3] F. Bach, S. Lacoste-Julien, and G. Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proc. 29th ICML*, ICML'12, 2012.

[4] A. D. Barbour. Stein's method and Poisson process convergence. *J. Appl. Probab.*, (Special Vol. 25A):175–184, 1988. ISSN 0021-9002. A celebration of applied probability.

[5] A. D. Barbour. Stein's method for diffusion approximations. *Probab. Theory Related Fields*, 84(3):297–322, 1990. ISSN 0178-8051.

[6] Q. W. Bouts, A. P. ten Brink, and K. Buchin. A framework for Computing the Greedy Spanner. In *Proc. of 30th SOCG*, pages 11:11–11:19, New York, NY, 2014. ACM.

[7] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

[8] P. Cattiaux and A. Guillin. Semi log-concave Markov diffusions. In *Séminaire de Probabilités XLVI*, volume 2123 of *Lecture Notes in Math.*, pages 231–292. Springer, Cham, 2014.

[9] S. Cerrai. *Second order PDE's in finite and infinite dimension: a probabilistic approach*, volume 1762. Springer Science & Business Media, 2001.

[10] S. Chatterjee and E. Meckes. Multivariate normal approximation using exchangeable pairs. *ALEA Lat. Am. J. Probab. Math. Stat.*, 4:257–283, 2008. ISSN 1980-0436.

[11] S. Chatterjee and Q. Shao. Nonnormal approximation by Stein's method of exchangeable pairs with application to the Curie-Weiss model. *Ann. Appl. Probab.*, 21 (2):464–483, 2011. ISSN 1050-5164.

[12] L. Chen, L. Goldstein, and Q. Shao. *Normal approximation by Stein's method*. Probability and its Applications. Springer, Heidelberg, 2011. ISBN 978-3-642-15006-7.

[13] Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. In *UAI*, 2010.

[14] P. Chew. There is a Planar Graph Almost As Good As the Complete Graph. In *Proc. 2nd SOCG*, pages 169–177, New York, NY, 1986. ACM.

[15] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *Proc. 33rd ICML*, ICML, 2016.

[16] C. Conca and M. Vanninathan. Periodic homogenization problems in incompressible fluid equations. *Handbook of Mathematical Fluid Dynamics*, 4:649–698, 2007.

[17] P. Doersek and J. Teichmann. A Semigroup Point Of View On Splitting Schemes For Stochastic (Partial) Differential Equations. *ArXiv e-prints*, Nov. 2010.

[18] P. D. Drummond and C. W. Gardiner. Generalised P-representations in quantum optics. *J. Phys. A*, 13(7):2353, 1980.

[19] P. D. Drummond and D. F. Walls. Quantum theory of optical bistability. I. Nonlinear polarisability model. *J. Phys. A*, 13(2):725, 1980.

[20] A. B. Duncan, T. Lelievre, and G. Pavliotis. Variance reduction using nonreversible Langevin samplers. *J. Stat. Phys.*, 163(3):457–491, 2016.

[21] E. Dynkin. *Markov Processes : Volume 1*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1965. ISBN 978-3-662-00033-5.

[22] A. Eberle. Reflection couplings and contraction rates for diffusions. *Probab. Theory Related Fields*, pages 1–36, 2015.

[23] K. Engel and R. Nagel. *One-parameter semigroups for linear evolution equations*, volume 194 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2000. ISBN 0-387-98463-1.

[24] S. N. Ethier and T. G. Kurtz. *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986. ISBN 0-471-08186-8.

[25] Y. Fan, S. P. Brooks, and A. Gelman. Output assessment for Monte Carlo simulations via the score statistic. *J. Comp. Graph. Stat.*, 15(1), 2006.

[26] E. Fournié, J. Lasry, J. Lebuchoux, P. Lions, and N. Touzi. Applications of Malliavin calculus to Monte Carlo methods in finance. *Fin. Stochastics*, 3(4):391–412, 1999. ISSN 0949-2984.

[27] A. Friedman. *Stochastic differential equations and applications. Vol. 1*. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Probability and Mathematical Statistics, Vol. 28.

[28] H. L. Gan, A. Röllin, and N. Ross. Dirichlet approximation of equilibrium distributions in Cannings models with mutation. *ArXiv e-prints*, Feb. 2016.

[29] R. E. Gaunt. Rates of convergence in normal approximation under moment conditions via new bounds on solutions of the stein equation. *J. Theoret. Probab.*, 29(1):231–247, 2016.

[30] A. Gelman. Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, 2012.

[31] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian data analysis*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL, third edition, 2014. ISBN 978-1-4398-4095-5.

[32] C. J. Geyer. Markov chain Monte Carlo maximum likelihood. *Computer Science and Statistics: Proc. 23rd Symp. Interface*, pages 156–163, 1991.

[33] M. Girolami and B. Calderhead. Riemann Manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B*, 73(2):123–214, 2011.

[34] G. Glaeser. Étude de quelques algèbres tayloriennes. *J. Analyse Math.*, 6:1–124; erratum, insert to 6 (1958), no. 2, 1958.

[35] J. Gorham and L. Mackey. Measuring sample quality with Stein's method. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Adv. NIPS 28*, pages 226–234. Curran Associates, Inc., 2015.

[36] J. Gorham and L. Mackey. Measuring sample quality with kernels. *arXiv:1703.01717*, Mar. 2017.

[37] F. Götze. On the rate of convergence in the multivariate CLT. *Ann. Probab.*, 19(2): 724–739, 1991.

[38] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In *Adv. NIPS 19*, pages 513–520, 2006.

[39] J. Gudmundsson, O. Klein, C. Knauer, and M. Smid. Small Manhattan Networks and Algorithmic Applications for the Earth Movers Distance. In *Proc. 23rd EuroCG*, pages 174–177, 2007.

[40] M. Hairer, A. Stuart, and S. Vollmer. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.*, 2014.

[41] S. Har-Peled and M. Mendel. Fast construction of nets in low-dimensional metrics and their applications. *SIAM J. Comput.*, 35(5):1148–1184, 2006.

[42] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[43] A. Horowitz. The second order Langevin equation and numerical simulations. *Nucl. Phys. B*, 280:510–522, 1987.

[44] P. Huber and E. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2009. ISBN 978-0-470-12990-6.

[45] J. H. Huggins and J. Zou. Quantifying the accuracy of approximate diffusions and Markov chains. *arXiv:1605.06420*, May 2016.

[46] C. Hwang, S. Hwang-Ma, and S. Sheu. Accelerating Gaussian diffusions. *Ann. Appl. Probab.*, pages 897–913, 1993.

[47] A. Joulin and Y. Ollivier. Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Probab.*, 38(6):2418–2442, 11 2010.

[48] O. Kallenberg. *Foundations of modern probability*. Probability and its Applications. Springer-Verlag, New York, second edition, 2002. ISBN 0-387-95313-2.

[49] J. Kent. Time-reversible diffusions. *Ann. Appl. Probab.*, pages 819–835, 1978.

[50] R. Khasminskii. *Stochastic stability of differential equations*, volume 66 of *Stochastic Modelling and Applied Probability*. Springer, Heidelberg, second edition, 2012. ISBN 978-3-642-23279-4. With contributions by G. N. Milstein and M. B. Nevelson.

[51] A. Korattikara, Y. Chen, and M. Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *Proc. of 31st ICML*, ICML'14, 2014.

[52] S. Lacoste-Julien, F. Lindsten, and F. Bach. Sequential kernel herding: Frank-Wolfe optimization for particle filtering. In *AISTATS*, 2015.

[53] C. Landim, S. Olla, and H. Yau. Convection–diffusion equation with space–time ergodic random flow. *Probability theory and related fields*, 112(2):203–220, 1998.

[54] C. Ley, G. Reinert, and Y. Swan. Stein's method for comparison of univariate distributions. *Probab. Surveys*, 14:1–52, 2017.

[55] C. Liu. Bayesian robust multivariate linear regression with incomplete data. *JASA*, 91(435):1219–1227, 1996. ISSN 01621459.

[56] Q. Liu and J. Lee. Black-box Importance Sampling. *arXiv:1610.05247*, Oct. 2016.

[57] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *arXiv:1608.04471*, 2016.

[58] Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proc. of 33rd ICML*, volume 48 of *ICML*, pages 276–284, 2016.

[59] M. Lubin and I. Dunning. Computing in operations research using Julia. *INFORMS Journal on Computing*, 27(2):238–248, 2015.

[60] A. Lunardi. An introduction to interpolation theory. 2007.

[61] Y. Ma, T. Chen, and E. Fox. A complete recipe for stochastic gradient MCMC. In *Adv. NIPS 28*, pages 2899–2907, 2015.

[62] L. Mackey and J. Gorham. Multivariate Stein factors for a class of strongly log-concave distributions. *Electron. Commun. Probab.*, 21:14 pp., 2016.

[63] L. Manca. *Kolmogorov operators in spaces of continuous functions and equations for measures*. PhD thesis, Scuola Normale Superiore di Pisa, 2008.

[64] J. Mattingly, A. Stuart, and M. Tretyakov. Convergence of numerical time-averaging and stationary measures via Poisson equations. *SIAM J. Numer. Anal.*, 48(2):552–577, 2010.

[65] E. Meckes. On Stein's method for multivariate normal approximation. In *High dimensional probability V: the Luminy volume*, volume 5 of *Inst. Math. Stat. Collect.*, pages 153–178. Inst. Math. Statist., Beachwood, OH, 2009.

[66] A. Müller. Integral probability metrics and their generating classes of functions. *Ann. Appl. Probab.*, 29(2):pp. 429–443, 1997.

[67] I. Nourdin, G. Peccati, and A. Réveillac. Multivariate normal approximation using Stein's method and Malliavin calculus. *Ann. Inst. Henri Poincaré Probab. Stat.*, 46 (1):45–58, 2010. ISSN 0246-0203.

[68] C. Oates and M. Girolami. Control functionals for Quasi-Monte Carlo integration. *arXiv:1501.03379*, 2015.
[69] C. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *arXiv:1410.2392*, Oct. 2014.
[70] C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages n/a–n/a, 2016. ISSN 1467-9868.
[71] B. Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 6 edition, 2013.
[72] G. Optimization. Gurobi optimizer reference manual, 2015. URL http://www.gurobi.com.
[73] S. Patterson and Y. Teh. Stochastic gradient Riemannian langevin dynamics on the probability simplex. In *Adv. NIPS 26*, pages 3102–3110, 2013.
[74] G. A. Pavliotis. *Stochastic processes and applications*, volume 60 of *Texts in Applied Mathematics*. Springer, New York, 2014. ISBN 978-1-4939-1322-0; 978-1-4939-1323-7. Diffusion processes, the Fokker-Planck and Langevin equations.
[75] D. Peleg and A. Schäffer. Graph spanners. *J. Graph Theory*, 13(1):99–116, 1989.
[76] P. Protter. *Stochastic integration and differential equations*, volume 21 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2005. ISBN 3-540-00313-4. Second edition. Version 2.1, Corrected third printing.
[77] R. Ranganath, D. Tran, J. Altosaar, and D. Blei. Operator variational inference. In *Advances in Neural Information Processing Systems*, pages 496–504, 2016.
[78] G. Reinert and A. Röllin. Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. *Ann. Probab.*, 37(6):2150–2173, 2009. ISSN 0091-1798.
[79] L. Rey-Bellet and K. Spiliopoulos. Irreversible Langevin samplers and variance reduction: a large deviation approach. *arXiv preprint arXiv:1404.0105*, 2014.
[80] H. Risken. *The Fokker-Planck Equation: Methods of Solutions and Applications*. Springer Series in Synergetics. Springer, 2nd ed. 1989. 3rd printing edition, Sept. 1996. ISBN 354061530X.
[81] G. Roberts and R. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. ISSN 1350-7265.
[82] G. O. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. 4(4):337–357, 2002.
[83] M. Röckner, Z. Sobol, et al. Kolmogorov equations in infinite dimensions: well-posedness and regularity of solutions, with applications to stochastic generalized burgers equations. *Ann. Probab.*, 34(2):663–727, 2006.
[84] P. Shvartsman. The Whitney extension problem and Lipschitz selections of set-valued mappings in jet-spaces. *Trans. Amer. Math. Soc.*, 360(10):5529–5550, 2008.
[85] G. Soffritti and G. Galimberti. Multivariate linear regression with non-normal errors: a solution based on mixture models. *Statistics and Computing*, 21(4):523–536, 2011.
[86] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. 6th Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pages 583–602. Univ. California Press, Berkeley, Calif., 1972.
[87] C. Stein, P. Diaconis, S. Holmes, and G. Reinert. Use of exchangeable pairs in the analysis of simulations. In *Stein's method: expository lectures and applications*, volume 46 of *IMS Lecture Notes Monogr. Ser.*, pages 1–26. Inst. Math. Statist., Beachwood, OH, 2004.
[88] A. Stuart, J. Voss, P. Wilberg, et al. Conditional path sampling of SDEs and the

Langevin MCMC method. *Commun. Math. Sci.*, 2(4):685–697, 2004.

[89] Y. Teh, A. Thiéry, and S. Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *arXiv:1409.0578*, 2014.

[90] S. Vallender. Calculation of the Wasserstein distance between probability distributions on the line. *Theory Probab. Appl.*, 18(4):784–786, 1974.

[91] S. Vollmer, K. Zygalakis, and Y. Teh. Exploration of the (non-)asymptotic bias and variance of stochastic gradient langevin dynamics. *J. Mach. Learn. Res.*, 17(159): 1–48, 2016.

[92] F. Wang. Exponential Contraction in Wasserstein Distances for Diffusion Semigroups with Negative Curvature. *ArXiv e-prints*, Mar. 2016.

[93] M. Welling and Y. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*, 2011.

[94] A. Zellner. Bayesian and Non-Bayesian analysis of the regression model with multivariate student-t error terms. *JASA*, 71(354):400–405, 1976.

[95] A. Zellner and C. Min. Gibbs sampler convergence criteria. *JASA*, 90(431):921–927, 1995.

## SUPPLEMENTARY APPENDIX I: SMOOTHING AND INTERPOLATION

We present in this section two essentially standard results on smoothing by convolution and seminorm interpolation [see, e.g., 60, Ex. 1.1.8] which support the proof of Theorem 7. Throughout, we let $G \in \mathbb{R}^d$ be a standard normal vector and $\phi \in C^\infty$ be its probability density. For any $s > 0$ and function $f : \mathbb{R}^d \to \mathbb{R}$ we define

$$f_s(x) \triangleq \mathbb{E}[f(x + sG)] = s^{-d} \int f(y)\phi\big(\tfrac{x-y}{s}\big)dy.$$

The first result bounds the Lipschitz constant of $f_s$ in terms of the Hölder continuity of $f$.

LEMMA 19 (Smoothing by convolution II). *Fix $\iota \in (0, 1)$ and consider any $f : \mathbb{R}^d \to \mathbb{R}$ with $M_{1-\iota}(f) < \infty$. For all $s > 0$,*

$$M_1(f_s) \le \mathbb{E}\big[\|G\|_2^{2-2\iota}\big]^{1/2} M_{1-\iota}(f)s^{-\iota}.$$

PROOF. Fix any $\|v\|_2 \le 1$ and $x \in \mathbb{R}^d$. Leibniz's rule implies that

$$\langle\nabla f_s(x), v\rangle = s^{-d-1} \int f(y)\langle\nabla\phi\big(\tfrac{x-y}{s}\big), v\rangle\, dy.$$

Because $s^{-d} \int \nabla\langle\phi(\frac{x-y}{s}), v\rangle\, dy = 0$ for any $v \in \mathbb{R}^d$, we also have

$$\begin{aligned}
|\langle\nabla f_s(x), v\rangle| = |s^{-d-1} \textstyle\int f(y)\langle\nabla\phi\big(\tfrac{x-y}{s}\big), v\rangle\, dy| &= |s^{-d-1} \textstyle\int [f(y) - f(x)]\langle\nabla\phi\big(\tfrac{x-y}{s}\big), v\rangle\, dy| \\
&= |s^{-d-1} \textstyle\int [f(x - z) - f(x)]\langle\nabla\phi\big(\tfrac{z}{s}\big), v\rangle\, dz| \\
&\le s^{-d-1} \textstyle\int M_{1-\iota}(f)\|z\|_2^{1-\iota}|\langle\nabla\phi\big(\tfrac{z}{s}\big), v\rangle|\, dz \\
&= M_{1-\iota}(f)s^{-\iota} \textstyle\int \|\omega\|_2^{1-\iota}|\langle\nabla\phi(\omega), v\rangle|\, d\omega,
\end{aligned}$$

where we have used substitutions $z \triangleq x - y$ and $\omega \triangleq z/s$. Finally, as $\nabla\phi(\omega) = -\omega\phi(\omega)$ for all $\omega \in \mathbb{R}^d$, we can use the spherical symmetry of the standard normal and Cauchy-Schwarz to yield

$$\begin{aligned}
\textstyle\int \|\omega\|_2^{1-\iota}|\langle\nabla\phi(\omega), v\rangle|\, d\omega = \mathbb{E}\big[\|G\|_2^{1-\iota}|\langle G, v\rangle|\big] &\le \mathbb{E}\big[\|G\|_2^{2-2\iota}\big]^{1/2}\mathbb{E}\big[|\langle G, v\rangle|^2\big]^{1/2} \\
&= \mathbb{E}\big[\|G\|_2^{2-2\iota}\big]^{1/2}\mathbb{E}\big[G_1^2\big]^{1/2} = \mathbb{E}\big[\|G\|_2^{2-2\iota}\big]^{1/2},
\end{aligned}$$

concluding the lemma. □

The second result provides interpolation bounds for the Hölder seminorm $M_k$ where $k \notin \mathbb{N}$.

LEMMA 20 (Seminorm interpolation). *Let $k > 0$ and $f \in C^{\lceil k\rceil}(\mathbb{R}^d)$. Then we have that*

$$M_k(f) \le 2^{1-\{k\}}\big(M_{\lceil k\rceil-1}(f)\big)^{1-\{k\}}\big(M_{\lceil k\rceil}(f)\big)^{\{k\}}.$$

PROOF. For $m \in \mathbb{N}$, let $V_m = \{(v_1, \ldots, v_m) : \|v_i\|_2 \le 1 \text{ for each } i \in \{1, \ldots, m\}\}$. Using the fundamental theorem of calculus we obtain

$$\begin{aligned}
&\sup_{V_{\lceil k\rceil-1}} \left|\nabla^{\lceil k\rceil-1} f(x)[v_1, v_2, \ldots, v_{\lceil k\rceil-1}] - \nabla^{\lceil k\rceil-1} f(y)[v_1, v_2, \ldots, v_{\lceil k\rceil-1}]\right| \\
&\qquad = \sup_{V_{\lceil k\rceil-1}} \left|\int_0^1 \nabla^{\lceil k\rceil} f(x + s(y-x))[v_1, v_2, \ldots, v_{\lceil k\rceil-1}, y-x] ds\right| \\
&\qquad \le \sup_{V_{\lceil k\rceil-1}} \left|\sup_z \nabla^{\lceil k\rceil} f(z)[v_1, v_2, \ldots, v_{\lceil k\rceil-1}, y-x]\right| \\
&\qquad \le \sup_z \|\nabla^{\lceil k\rceil} f(z)\|_{op} \|x-y\|_2.
\end{aligned}$$

An application of the triangle inequality gives rise to

$$\sup_{V_{\lceil k\rceil-1}} \left|\nabla^{\lceil k\rceil-1} f(x)[v_1, v_2, \ldots, v_{\lceil k\rceil-1}] - \nabla^{\lceil k\rceil-1} f(y)[v_1, v_2, \ldots, v_{\lceil k\rceil-1}]\right| \le 2 \sup_z \|\nabla^{\lceil k\rceil-1} f(z)\|_{op}.$$

There we obtain

$$\begin{aligned}
M_k(f) &= \sup_{x,y\in\mathbb{R}^d; x\neq y} \frac{\|\nabla^{\lceil k\rceil-1} f(x) - \nabla^{\lceil k\rceil-1} f(y)\|_{op}}{\|x-y\|_2^{\{k\}}} \\
&\le \sup_{x,y\in\mathbb{R}^d; x\neq y} \frac{2^{1-\{k\}} \left(\sup_z \|\nabla^{\lceil k\rceil} f(z)\|_{op}\right)^{\{k\}} \left(\sup_z \|\nabla^{\lceil k\rceil-1} f(z)\|_{op}\right)^{1-\{k\}} \|x-y\|_2^{\{k\}}}{\|x-y\|_2^{\{k\}}} \\
&\le 2^{1-\{k\}} \left(\sup_z \|\nabla^{\lceil k\rceil} f(z)\|_{op}\right)^{\{k\}} \left(\sup_z \|\nabla^{\lceil k\rceil-1} f(z)\|_{op}\right)^{1-\{k\}} \\
&\le 2^{1-\{k\}} \left(M_{\lceil k\rceil}(f)\right)^{\{k\}} \left(M_{\lceil k\rceil-1}(f)\right)^{1-\{k\}}
\end{aligned}$$

thus proving the statement. □

## SUPPLEMENTARY APPENDIX J: SEMIGROUP THIRD DERIVATIVE ESTIMATE

LEMMA 21 (Semigroup third derivative estimate). *Suppose that the drift and diffusion coefficients $b$ and $\sigma$ of an Itô diffusion have bounded, locally Lipschitz first, second, and third derivatives. If the transition semigroup $(P_t)_{t\ge0}$ has Wasserstein decay rate $r$, $\sigma(x)$ has a right inverse $\sigma^{-1}(x)$ for each $x \in \mathbb{R}^d$, and $M_0(\sigma^{-1}) < \infty$, then, for all $t > 0$ and any $f \in C^3$ with bounded second and third derivatives,*

$$M_3(P_t f) \le \inf_{t_0\in(0,t]} M_1(f) r(t-t_0) \frac{c}{t_0} e^{Ct_0} \tag{33}$$

*for constants $c, C$ depending only on $M_{1:3}(\sigma), M_{1:3}(b), M_0(\sigma^{-1})$, and $r$.*

PROOF. Our proof closely follows that of Lemma 15 in Section C, and we will only highlight the important differences. Throughout, $c$ and $C$ will represent arbitrary constants depending only on $M_{1:3}(\sigma), M_{1:3}(b), M_0(\sigma^{-1})$, and $r$ that may change from expression to expression.

Fix any $v, v', v''$ with unit Euclidean norms in $\mathbb{R}^d$ and, without loss of generality, fix any $f : \mathbb{R}^d \to \mathbb{R}$ in $C^3$ with $M_1(f) \le 1$ and bounded second and third derivatives. Let $(Z_{t,x})_{t\ge0}$ be an Itô diffusion solving the stochastic differential equation (5) with starting point $Z_{0,x} = x$, underlying Wiener process $(W_t)_{t\geq 0}$, and transition semigroup $(P_t)_{t\geq 0}$. Under our smoothness assumptions on $b$ and $\sigma$, [76, Theorem V.40] implies that $(Z_{t,x})_{t\geq 0}$ is thrice continuously differentiable in $x$ with third directional derivative flow $(Y_{t,v,v',v''})_{t\geq 0}$ solving the *third variation equation*,

$$\begin{aligned} dY_{t,v,v',v''} &= \nabla b(Z_{t,x})Y_{t,v,v',v''}\,dt + \nabla^2 b(Z_{t,x})[U_{t,v,v'}]V_{t,v''}\,dt \\ &+ \nabla^3 b(Z_{t,x})[V_{t,v}, V_{t,v'}, V_{t,v''}]\,dt + \nabla^2 b(Z_{t,x})[U_{t,v',v''}]V_{t,v}\,dt \\ &+ \nabla^2 b(Z_{t,x})[U_{t,v,v''}]V_{t,v'}\,dt \\ &+ \nabla\sigma(Z_{t,x})Y_{t,v,v',v''}\,dW_t + \nabla^2\sigma(Z_{t,x})[U_{t,v,v'}]V_{t,v''}\,dW_t \\ &+ \nabla^3\sigma(Z_{t,x})[V_{t,v}, V_{t,v'}, V_{t,v''}]\,dW_t + \nabla^2\sigma(Z_{t,x})[U_{t,v',v''}]V_{t,v}\,dW_t \\ &+ \nabla^2\sigma(Z_{t,x})[U_{t,v,v''}]V_{t,v'}\,dW_t \quad \text{with} \quad Y_{0,v,v',v''} = 0, \end{aligned} \tag{34}$$

obtained by differentiating (22) with respect to $x$ in the direction $v''$.

In a manner analogous to the derivation of (26) in proof of Lemma 15, we can derive an expression for the third derivative of the semi-group,

$$\begin{aligned} &\nabla^3(P_t f)(x)[v, v', v''] = \mathbb{E}\Big[\textstyle\sum_{i,j} J_{i,j,x}\Big] \quad \text{for} \\ &J_{1,1,x} \triangleq \tfrac{1}{t}\langle\nabla^2(P_t f)(Z_{t,x})V_{t,v''}, V_{t,v'}\rangle \textstyle\int_0^t \langle\sigma^{-1}(Z_{s,x})V_{s,v},\, dW_s\rangle, \\ &J_{1,2,x} \triangleq \tfrac{1}{t}\langle\nabla(P_t f)(Z_{t,x}), U_{t,v',v''}\rangle \textstyle\int_0^t \langle\sigma^{-1}(Z_{s,x})V_{s,v},\, dW_s\rangle, \\ &J_{1,3,x} \triangleq \tfrac{1}{t}\langle\nabla(P_t f)(Z_{t,x}), V_{t,v'}\rangle \textstyle\int_0^t \langle\nabla\sigma^{-1}(Z_{s,x})[V_{s,v''}]V_{s,v},\, dW_s\rangle, \\ &J_{2,1x} \triangleq \tfrac{1}{t}\langle\nabla f(Z_{t,x}), V_{t,v''}\rangle \textstyle\int_0^t \langle\nabla\sigma^{-1}(Z_{s,x})[V_{s,v'}]V_{s,v},\, dW_s\rangle, \\ &J_{2,2,x} \triangleq \tfrac{1}{t}f(Z_{t,x}) \textstyle\int_0^t \langle\nabla^2\sigma^{-1}(Z_{s,x})[V_{s,v''}][V_{s,v'}]V_{s,v},\, dW_s\rangle, \\ &J_{2,3,x} \triangleq \tfrac{1}{t}f(Z_{t,x}) \textstyle\int_0^t \langle\nabla\sigma^{-1}(Z_{s,x})[U_{s,v',v''}]V_{s,v},\, dW_s\rangle, \\ &J_{2,4,x} \triangleq \tfrac{1}{t}f(Z_{t,x}) \textstyle\int_0^t \langle\nabla\sigma^{-1}(Z_{s,x})[V_{s,v'}]U_{s,v,v''},\, dW_s\rangle, \\ &J_{3,1,x} \triangleq \tfrac{1}{t}\langle\nabla f(Z_{t,x}), V_{t,v''}\rangle \textstyle\int_0^t \langle\sigma^{-1}(Z_{s,x})U_{s,v,v'},\, dW_s\rangle, \\ &J_{3,2,x} \triangleq \tfrac{1}{t}f(Z_{t,x}) \textstyle\int_0^t \langle\nabla\sigma^{-1}(Z_{s,x})[V_{s,v''}]U_{s,v,v'},\, dW_s\rangle, \\ &J_{3,3,x} \triangleq \tfrac{1}{t}f(Z_{t,x}) \textstyle\int_0^t \langle\sigma^{-1}(Z_{s,x})Y_{s,v,v',v''},\, dW_s\rangle. \end{aligned} \tag{35}$$

We will bound each term $J_{i,j,x}$ in (35) in turn.

**J.1. The $J_{1,\cdot,x}$ terms.** We will show the calculation step by step for the first term. By Cauchy-Schwarz,

$$\begin{aligned} \mathbb{E}[J_{1,1,x}] &= \tfrac{1}{t}\mathbb{E}\Big[\langle\nabla^2(P_t f)(Z_{t,x})V_{t,v''}, V_{t,v'}\rangle \textstyle\int_0^t \langle\sigma^{-1}(Z_{s,x})V_{s,v},\, dW_s\rangle\Big] \\ &\leq \tfrac{1}{t}\sqrt{\mathbb{E}\big[\langle\nabla^2(P_t f)(Z_{t,x})V_{t,v''}, V_{t,v'}\rangle^2\big]\mathbb{E}[(\textstyle\int_0^t \langle\sigma^{-1}(Z_{s,x})V_{s,v},\, dW_s\rangle)^2]}. \end{aligned}$$

We use the derivative flow bounds of Lemma 16 to realize

$$\sqrt{\mathbb{E}\big[\|V_{t,v'}\|_2^2\|V_{t,v''}\|_2^2\big]} \le \sqrt[4]{\mathbb{E}\|V_{t,v'}\|_2^4\mathbb{E}\|V_{t,v''}\|_2^4} \le \|v'\|_2\|v''\|_2 e^{\frac{1}{2}t\gamma_4}.$$

Combining the above with the definition of $M_2$ and Cauchy-Schwarz yields

$$\begin{aligned}
\mathbb{E}[J_{1,1,x}] &\le \tfrac{1}{t}M_2(P_tf)\sqrt{\mathbb{E}\big[\|V_{t,v'}\|_2^2\|V_{t,v''}\|_2^2\big]\mathbb{E}\Big[(\textstyle\int_0^t \|\sigma^{-1}(Z_{s,x})V_{s,v}\,ds\|_2)^2\Big]}\\
&\le \tfrac{1}{t}M_2(P_tf)\|v'\|_2\|v''\|_2\|v\|_2 e^{\frac{1}{2}t\gamma_4}M_0(\sigma^{-1})\Big(\tfrac{1}{\gamma_2}(e^{\gamma_2 t}-1)\Big)^{\frac{1}{2}}\\
&\le \tfrac{1}{\sqrt{t}}M_2(P_tf)\|v'\|_2\|v''\|_2\|v\|_2 e^{\frac{1}{2}t(\gamma_4+\gamma_2)}M_0(\sigma^{-1}) \le \tfrac{c}{t}e^{Ct},
\end{aligned}$$

where we used $M_2(P_tf)$ is of order $\max(t^{-\frac{1}{2}}, e^{c_0 t})$ for some constant $c_0 > 0$ by Lemma 15 and the mean value theorem. Similar reasoning yields

$$\begin{aligned}
\mathbb{E}[J_{1,2,x}] &= \mathbb{E}\tfrac{1}{t}\langle\nabla(P_tf)(Z_{t,x}), U_{t,v',v''}\rangle \textstyle\int_0^t\langle\sigma^{-1}(Z_{s,x})V_{s,v},\, dW_s\rangle\\
&\le \tfrac{1}{t}M_1(f)r(t)\sqrt{t}e^{\frac{1}{4}t\gamma_4}M_0(\sigma^{-1})\Big(\tfrac{1}{\gamma_2}(e^{\gamma_2 t}-1)\Big)^{\frac{1}{2}} \le ce^{Ct}
\end{aligned}$$

and using equation (28)

$$\begin{aligned}
\mathbb{E}[J_{1,3,x}] &\le \tfrac{1}{t}M_1(f)e^{\frac{1}{2}\gamma_2 t}\|v'\|_2\sqrt{\mathbb{E}\textstyle\int_0^t M_1(\sigma^{-1})^2\|V_{s,v}\|_2^2\|V_{s,v''}\|_2^2\,ds}\\
&\le \tfrac{1}{t}M_1(f)e^{\frac{1}{2}\gamma_2 t}\|v'\|_2 M_1(\sigma^{-1})\|v\|_2\|v''\|_2\Big(\textstyle\int_0^t e^{\gamma_4 s}ds\Big)^{\frac{1}{2}}\\
&\le t^{-\frac{1}{2}}e^{\gamma_2 t/2}\|v'\|_2 M_0(\sigma^{-1})^2 M_1(\sigma)\|v\|_2\|v''\|_2 e^{\gamma_4 t/2} \le \tfrac{c}{\sqrt{t}}e^{Ct}.
\end{aligned}$$

**J.2. The $J_{2,\cdot,x}$ terms.** To tackle the $J_{2,\cdot,x}$ terms, we will rewrite the unbounded quantity $f(Z_{t,x})$ using (27) and its directional derivative using

$$\begin{aligned}
\langle\nabla_x f(Z_{t,x}), v\rangle = &\int_0^t \nabla^2(P_{t-s}f)(Z_{s,x})[\sigma(Z_{s,x})\,dW_s][v] + \int_0^t\langle\nabla(P_{t-s}f)(Z_{s,x}), v^t\nabla\sigma(Z_{s,x})\,dW_s\rangle\\
&+ \langle\nabla_x(P_tf)(Z_{s,x}), v\rangle
\end{aligned}\tag{36}$$

The first part of $J_{2,1,x}$ becomes

$$\begin{aligned}
\langle\nabla f(Z_{t,x}), V_{t,v''}\rangle = &\textstyle\int_0^t \nabla^2(P_{t-s}f)(Z_{s,x})[\sigma(Z_{s,x})\,dW_s][V_{t,v''}]\\
&+ \textstyle\int_0^t\langle\nabla(P_{t-s}f)(Z_{s,x}), \nabla\sigma(Z_{s,x})[V_{t,v''}]\,dW_s\rangle.
\end{aligned}\tag{37}$$

Using (27), we write $J_{2,1,x} = J_{2,1,1,x} + J_{2,1,2,x}$ where

$$\begin{aligned}
\mathbb{E}[J_{2,1,1,x}] &= \tfrac{1}{t}\mathbb{E}\textstyle\int_0^t \nabla^2(P_{t-s}f)(Z_{s,x})[\sigma(Z_{s,x})\,dW_s][V_{t,v''}]\int_0^t\langle\nabla\sigma^{-1}(Z_{s,x})[V_{s,v'}]V_{s,v},\, dW_s\rangle\\
&= \tfrac{1}{t}\mathbb{E}\textstyle\int_0^t \nabla^2(P_{t-s}f)(Z_{s,x})[\sigma(Z_{s,x})\nabla\sigma^{-1}(Z_{s,x})[V_{s,v'}]V_{s,v}][V_{t,v''}]ds.
\end{aligned}$$

Since

$$\left|\nabla^2(P_{t-s}f)(Z_{s,x})[w_1,w_2]\right| \leq \|w_1\|_2\|\nabla^2(P_{t-s}f)(Z_{s,x})w_2\|_2 \leq \|w_1\|_2\|w_2\|_2 M_2(P_{t-s}f),$$

by Cauchy-Schwarz, we use the chain rule identity (28) and Lemma 15 to conclude

$$\mathbb{E}[J_{2,1,1,x}] \leq M_1(\sigma)M_0(\sigma^{-1})\tfrac{1}{t}\int_0^t \mathbb{E}M_2(P_{t-s}f)\|V_{s,v}\|_2\|V_{s,v'}\|_2\|V_{s,v''}\|_2 ds \leq \tfrac{c}{\sqrt{t}}e^{Ct}.$$

The remaining term $\mathbb{E}[J_{2,1,2,x}]$ can be bounded similarly:

$$\begin{aligned}
&\mathbb{E}\tfrac{1}{t}\int_0^t\langle\nabla(P_{t-s}f)(Z_{s,x}),\nabla\sigma(Z_{s,x})[V_{t,v''}]\,dW_s\rangle\int_0^t\langle\nabla\sigma^{-1}(Z_{s,x})[V_{s,v'}]V_{s,v},\,dW_s\rangle\\
=&\tfrac{1}{t}\int_0^t\mathbb{E}\langle\nabla(P_{t-s}f)(Z_{s,x}),\nabla\sigma(Z_{s,x})[V_{t,v''}]\nabla\sigma^{-1}(Z_{s,x})[V_{s,v'}]V_{s,v}\rangle ds\\
\leq&M_1(\sigma)M_1(\sigma^{-1})\tfrac{1}{t}\int_0^t\mathbb{E}M_1(P_{t-s}f)\|V_{s,v}\|_2\|V_{s,v'}\|_2\|V_{s,v''}\|_2 ds\\
\leq&M_1(\sigma)^2M_0(\sigma^{-1})^2\tfrac{1}{t}\int_0^t\mathbb{E}M_1(P_{t-s}f)\|V_{s,v}\|_2\|V_{s,v'}\|_2\|V_{s,v''}\|_2 ds\leq ce^{Ct}.
\end{aligned}$$

We bound the next term

$$\begin{aligned}
\mathbb{E}J_{2,2,x}=&\ \mathbb{E}\tfrac{1}{t}f(Z_{t,x})\int_0^t\langle\nabla^2\sigma^{-1}(Z_{s,x})[V_{s,v''}][V_{s,v'}]V_{s,v},\,dW_s\rangle\\
&=\mathbb{E}\tfrac{1}{t}\int_0^t\langle\nabla(P_{t-s}f)(Z_{s,x}),\sigma(Z_{s,x})\,dW_s\rangle.\int_0^t\langle\nabla^2\sigma^{-1}(Z_{s,x})[V_{s,v''}][V_{s,v'}]V_{s,v},\,dW_s\rangle\\
&=\tfrac{1}{t}\mathbb{E}\int_0^t\langle\nabla(P_{t-s}f)(Z_{s,x}),\sigma(Z_{s,x})\nabla^2\sigma^{-1}(Z_{s,x})[V_{s,v''}][V_{s,v'}]V_{s,v}\rangle ds\\
&\leq\tfrac{1}{t}r(0)\big(2M_1(\sigma)^2M_0(\sigma^{-1})^2+M_1(\sigma)M_2(\sigma)M_0(\sigma^{-1})\big)\int_0^t\mathbb{E}\big[\|V_{s,v''}\|_2\|V_{s,v'}\|_2\|V_{s,v'}\|_2\big]ds.\\
&\leq r(0)\big(2M_1(\sigma)^2M_0(\sigma^{-1})^2+M_1(\sigma)M_2(\sigma)M_0(\sigma^{-1})\big)e^{\gamma_3 t}\|v''\|_2\|v'\|_2\|v\|_2\leq ce^{Ct},
\end{aligned}$$

where we used the chain rule expression

$$\begin{aligned}
\nabla^2\sigma^{-1}(x)[v][v'] = -\sigma(x)^{-1}\Big(&-\nabla\sigma(x)[v]\sigma(x)^{-1}\nabla\sigma[v'](x)\\
&-\nabla\sigma(x)[v']\sigma(x)^{-1}\nabla\sigma(x)[v]\\
&-\nabla\sigma(x)[v']\sigma(x)^{-1}\nabla^2\sigma(x)[v][v']\Big)\sigma(x)^{-1}
\end{aligned}$$

to rewrite $\sigma(Z_{s,x})\nabla^2\sigma^{-1}(Z_{s,x})$. The next term satisfies

$$\begin{aligned}
\mathbb{E}[J_{2,3,x}] &= \mathbb{E}\tfrac{1}{t}\int_0^t\langle\nabla(P_{t-s}f)(Z_{s,x}),\sigma(Z_{s,x})\,dW_s\rangle\int_0^t\langle\nabla\sigma^{-1}(Z_{s,x})[U_{s,v',v''}]V_{s,v},\,dW_s\rangle\\
&= \mathbb{E}\tfrac{1}{t}\int_0^t\langle\nabla(P_{t-s}f)(Z_{s,x}),\sigma(Z_{s,x})\nabla\sigma^{-1}(Z_{s,x})[U_{s,v',v''}]V_{s,v}\rangle ds\\
&\leq \tfrac{1}{t}M_0(\sigma^{-1})M_1(\sigma)r(x)\int_0^t\mathbb{E}\|U_{s,v',v''}\|\|V_{s,v}\|ds\\
&\leq M_0(\sigma^{-1})M_1(\sigma)r(x)\tfrac{1}{t}\int_0^t\|v\|\|v'\|\|v''\|(\alpha s e^{\gamma_4 s})^{\frac{1}{2}}e^{s\gamma_2/2}ds\leq ce^{Ct}.
\end{aligned}$$

The term $\mathbb{E}[J_{2,4,x}]$ can be bounded in the same way by swapping the roles of $v$ and $v'$.

**J.3. The $J_{3,\cdot}$ terms.** We again use the identity (37) and split

$$\mathbb{E}[J_{3,1,x}]=\mathbb{E}\Big[\tfrac{1}{t}\langle\nabla f(Z_{t,x}),V_{t,v''}\rangle\int_0^t\langle\sigma^{-1}(Z_{s,x})U_{s,v,v'},\,dW_s\rangle\Big],$$

into

$$\begin{aligned}\mathbb{E}[J_{3,1,1,x}]&=\tfrac{1}{t}\mathbb{E}\int_0^t\nabla^2(P_{t-s}f)(Z_{s,x})[\sigma(Z_{s,x})\sigma^{-1}(Z_{s,x})U_{s,v,v'}][V_{t,v''}]ds\\&\le\tfrac{1}{t}\int_0^t\mathbb{E}M_2(P_{t-s}f)\|U_{s,v,v'}\|_2\|V_{t,v''}\|_2ds\le\tfrac{c}{\sqrt{t}}e^{Ct}.\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[J_{3,1,2,x}]&=\tfrac{1}{t}\mathbb{E}\int_0^t\langle\nabla(P_{t-s}f)(Z_{s,x}),\nabla\sigma(Z_{s,x})[V_{t,v''}]\sigma^{-1}(Z_{s,x})U_{s,v,v'}\rangle ds\\&\le M_1(\sigma)M_0(\sigma^{-1})\tfrac{1}{t}\mathbb{E}\int_0^t M_1(P_{t-s}f)\|U_{s,v,v'}\|_2\|V_{t,v''}\|_2ds\le\tfrac{c}{\sqrt{t}}e^{Ct}.\end{aligned}$$

Notice that $J_{3,2}$ is the same as $J_{2,3}$ except for different labelling of the $v$'s. Now we consider the last term

$$J_{3,3,x}=\tfrac{1}{t}f(Z_{t,x})\int_0^t\langle\sigma^{-1}(Z_{s,x})Y_{s,v,v',v''},\,dW_s\rangle,$$

Using (27), we see that

$$\begin{aligned}\mathbb{E}[J_{3,3,x}]&=\tfrac{1}{t}\mathbb{E}\int_0^t\langle\nabla(P_{t-s}f)(Z_{s,x}),Y_{s,v,v',v''}\rangle ds\le\tfrac{1}{t}\int_0^t M_1(P_{t-s}f)\mathbb{E}\|Y_{s,v,v',v''}\|_2ds\\&\le\tfrac{1}{t}\int_0^t M_1(P_{t-s}f)\big(\mathbb{E}\|Y_{s,v,v',v''}\|_2^2\big)^{\frac{1}{2}}ds.\end{aligned}$$

This final expression is bounded by $ce^{Ct}$ provided that $\mathbb{E}\|Y_{s,v,v',v''}\|_2^2\le ce^{Cs}$. We will establish such a bound for the third directional derivative flow in the next section. By combining the resulting bounds on each $J_{i,j,x}$ term together, we can now establish the infimal bound and conclude the proof of Lemma 21. Adapting the argument of [9, Prop. 1.5.1] on result of Lemma 21 in Section J, we obtain

$$\begin{aligned}\big\|\nabla^3P_tf[v,v',v'']\big\|_{\mathrm{op}}&=\big\|\nabla^3P_{t_0}(P_{t-t_0}f)[v,v',v'']\big\|_{op}\\&\le M_1(P_{t-t_0}f)\frac{c}{t_0}e^{Ct_0}\le M_1(f)r(t-t_0)\frac{c}{t_0}e^{Ct_0}.\end{aligned}\tag{38}$$

**J.4. Third derivative flow bound.** Introduce the shorthand $(Y_t)_{t\ge0}$ for $(Y_{t,v,v',v''})_{t\ge0}$ solving the third variation equation (34). Dynkin's formula gives $\mathbb{E}\|Y_t\|_2^2=\int_0^t T_1+T_2\,ds$ for

$$\begin{aligned}T_1&\triangleq\mathbb{E}2\Big\langle Y_s,\nabla b(Z_{s,x})Y_s+\nabla^2b(Z_{s,x})[U_{s,v,v'}]V_{s,v''}+\nabla^3b(Z_{s,x})[V_{s,v},V_{s,v'},V_{s,v''}]\\&\quad+\nabla^2b(Z_{s,x})[U_{s,v',v''}]V_{s,v}+\nabla^2b(Z_{t,x})[U_{t,v,v''}]V_{t,v'}\Big\rangle\\T_2&\triangleq\mathbb{E}\Big\|\nabla\sigma(Z_{s,x})[Y_s]+\nabla^2\sigma(Z_{s,x})[U_{s,v,v'}]V_{s,v''}+\nabla^3\sigma(Z_{s,x})[V_{s,v},V_{s,v'},V_{s,v''}]\\&\quad+\nabla^2\sigma(Z_{s,x})[U_{s,v',v''}]V_{s,v}+\nabla^2\sigma(Z_{s,x})[U_{s,v,v''}]V_{s,v'}\Big\|_F^2.\end{aligned}$$

We have by Cauchy-Schwarz and Young's inequality

$$
\begin{aligned}
\tfrac{T_1}{2} &\leq \mathbb{E}\big(\|Y_s\|_2^2 M_1(b) + M_2(b)\|Y_s\|_2 \underbrace{\|U_{s,v,v'}\|_2\|V_{s,v''}\|_2}_{+2\text{ permutations}} + M_3(b)\|Y_s\|_2\|V_{s,v}\|_2\|V_{s,v'}\|_2\|V_{s,v''}\|_2\big) \\
&\leq \mathbb{E}\big(\|Y_s\|_2^2\big(M_1(b) + M_2^2(b) + M_3^2(b)\big) + \underbrace{M_2^2(b)\|U_{s,v,v'}\|_2^2\|V_{s,v''}\|_2^2}_{+2\text{ permutations}} \\
&\quad + M_3(b)^2\big(\|V_{s,v}\|_2\|V_{s,v'}\|_2\|V_{s,v''}\|_2\big)^2\big)
\end{aligned}
$$

and

$$
\begin{aligned}
\tfrac{T_2}{4} &\leq \mathbb{E}\|Y_2\|_2^2\big(M_1(\sigma)^2 + \|\nabla^2\sigma\|_{F_3}^2 + \|\nabla^3\sigma\|_{F_3}^2\big) \\
&\quad + \|\nabla^3\sigma\|_{F_3}^2\mathbb{E}(\|V_s\|\|V_s'\|\|V_s''\|)^2 + \|\nabla^2\sigma\|_{F_3}^2 \underbrace{\mathbb{E}\big[\|U_{s,v,v'}\|_2^2\|V_{s,v''}\|_2^2\big]}_{+2\text{ permutations}}.
\end{aligned}
$$

Provided that we establish a bound of $\mathbb{E}\|U_{s,v,v'}\|_2^4 \leq cte^{Ct}$, we have that overall

$$
\mathbb{E}\|Y_t\|_2^2 \leq \int_0^t c\mathbb{E}\|Y_s\|_2^2 ds + ce^{Ct}.
$$

We can conclude using Gronwall's inequality that

$$
\mathbb{E}\|Y_t\|_2^2 \leq ce^{Ct}. \tag{39}
$$

It remains to establish bounds on $\mathbb{E}\|U_{t,v,v'}\|_2^\rho$ for $\rho > 2$. Recall that the second derivative flow solves (22). Applying Ito's formula to $f(U_{t,v,v'}) = \|U_{t,v,v'}\|_2^\rho$, taking expectations, and introducing the shorthand $U_t = U_{t,v,v'}$, we obtain

$$
\begin{aligned}
\mathbb{E}[\|U_t\|_2^\rho] &= \|U_0\|_2^\rho + \mathbb{E}\Big[\int_0^t \rho\langle U_s\|U_s,\|_2^{\rho-2}, \nabla b(Z_{s,x})U_s + \nabla^2 b(Z_{s,x})[V_{s,v'}]V_{s,v})\rangle \\
&\quad + \tfrac{\rho}{2}\|U_s\|_2^{\rho-4}((\rho-2)\|U_s^\top\nabla\sigma(Z_{s,x})[U_{s,v}] + U_s^\top\nabla^2\sigma(Z_{s,x})[V_{s,v'}]V_{s,v}\|_2^2 \\
&\quad + \|U_{s,v}\|_2^2\|\nabla\sigma(Z_{s,x})[U_{s,}] + \nabla^2\sigma(Z_{s,x})[V_{s,v'}]V_{s,v}\|_F^2)\, ds\Big] \\
&\leq \|U_0\|_2^\rho + \int_0^t \rho M_1(b)\|U_s\|_2^\rho + \rho M_2(b)\|U_s\|_2^{\rho-1}\|V_s\|_2\|V_s'\|_2 \\
&\quad + \tfrac{\rho^2-\rho}{2}\Big(M_1(\sigma)^2\|U_s\|_2^\rho + M_2(\sigma)^2\|U_s\|_2^{\rho-2}\|V_s\|_2\|V_s'\|_2\Big) \\
&\quad + \tfrac{\rho}{2}\Big(F_1(\sigma)^2\|U_s\|_2^\rho + F_2(\sigma)^2\|U_s\|_2^{\rho-2}\|V_s\|_2\|V_s'\|_2\Big) ds \\
&\leq \|U_0\|_2^\rho + \int_0^t \mathbb{E}[\|U_s\|_2^\rho](\rho M_1(b) + (\rho-1)M_2(b) + M_1(\sigma)^2\tfrac{\rho^2-\rho}{2} + M_2(\sigma)^2\tfrac{(\rho-1)^2}{2} + F_2(\sigma)\tfrac{\rho-1}{2})ds \\
&\quad + \int_0^t\Big(M_2(b) + \tfrac{\rho-1}{2}M_2(\sigma)^2 + \tfrac{1}{2}\Big)\mathbb{E}[(\|V_s\|_2\|V_s'\|_2)^\rho]ds \\
&\leq \|U_0\|_2^\rho + \int_0^t \mathbb{E}[\|U_s\|_2^\rho](\rho M_1(b) + (\rho-1)M_2(b) + M_1(\sigma)^2\tfrac{\rho^2-\rho}{2} + M_2(\sigma)^2\tfrac{(\rho-1)^2}{2} + F_2(\sigma)\tfrac{\rho-1}{2})ds \\
&\quad + \int_0^t\Big(M_2(b) + \tfrac{\rho-1}{2}M_2(\sigma)^2 + \tfrac{1}{2}\Big)(\|v\|_2\|v'\|_2)^\rho e^{\gamma_{2\rho}s}ds
\end{aligned}
$$

where we use that, by Young's inequality,

$$\|U_s\|_2^{\rho-1}\|V_s\|_2\|V_s'\|_2 \leq \tfrac{\rho-1}{\rho}\|U_s\|_2^\rho + \tfrac{1}{\rho}\|V_s\|_2^\rho\|V_s'\|_2^\rho,$$

and similarly

$$\|U_s\|_2^{\rho-2}\|V_s\|_2\|V_s'\|_2 \leq \tfrac{\rho-2}{\rho}\|U_s\|_2^\rho + \tfrac{2}{\rho}\|V_s\|_2^{\rho/2}\|V_s'\|_2^{\rho/2}.$$

Following the arguments of Section C.1, Grönwall's inequality gives

$$\mathbb{E}[\|U_t\|_2^\rho] \leq \Big(M_2(b) + \tfrac{\rho-1}{2}M_2(\sigma)^2 + \tfrac{1}{2}\Big)(\|v\|_2\|v'\|_2)^\rho e^{\gamma_{2\rho}t}t\exp(\gamma_\rho t).$$

$\square$