Neural Stochastic Differential Equations on Compact State-Spaces

Yue-Jane Liu [*1] **Malinda Lu** [*1] **Matthew K. Nock** [2] **Yaniv Yacoby** [1]

Abstract

Many modern probabilistic models rely on SDEs, but their adoption is hampered by instability, poor inductive bias outside bounded domains, and reliance on restrictive dynamics or training tricks. While recent work constrains SDEs to compact spaces using reflected dynamics, these approaches lack continuous dynamics and efficient high-order solvers, limiting interpretability and applicability. We propose a novel class of neural SDEs on compact polyhedral spaces with continuous dynamics, amenable to higher-order solvers, and with favorable inductive bias.

1. Introduction and Related Work

Stochastic differential equations (SDEs) have enabled the development of a large class of expressive probabilistic models, including models for continuous-time time series (e.g. Archambeau et al. (2007); Li et al. (2020)), diffusion models (Song et al., 2021), infinitely deep models (e.g. Xu et al. (2022)), and more. SDEs take the form, $dz_t = h(t, z_t) \cdot dt + g(t, z_t) \cdot dB_t$, wherein the change in state, dz_t, is modeled as a sum of deterministic and stochastic components. The "drift," h, represents a deterministic change to the state, and the "diffusion," g, represents the magnitude of stochasticity affecting the state. The stochasticity is given by the time-derivative Brownian motion, B_t.

While expressive, SDEs are notoriously unstable, especially when their dynamics, h and g, are non-linear. Their adoption so far often hinges on simplified dynamics (e.g. Ansari et al. (2023); Oh et al. (2024)) and training tricks (like KL-annealing, e.g. Li et al. (2020)). These approaches broadly have two shortcomings: (1) simplified dynamics and low-order solvers prevent scientific interpretation of the learned dynamics, and (2) the necessity of training tricks reduces their practicality.

Recent work on diffusion models observed that parameterizing SDEs on compact/bounded state spaces can improve their performance (Saharia et al., 2022; Lou & Ermon, 2023; Fishman et al., 2023a; Christopher et al., 2024). This is because many data sets (e.g. images) lie on compact Euclidean subspaces. In early-stage training, SDE trajectories often leave the compact region, requiring a large number of gradient steps just to return to it (while not necessarily fitting the data well), causing optimization to get stuck in poor local optima. In late-stage training, small perturbations to the dynamics may, again, yield trajectories that lie outside the region. Dynamics that respect the natural (compact) region of the data encode a stronger inductive bias towards admissible models, improving the stability of training and test-set generalization.

Of these recent works, SDEs with reflected (or clipped) dynamics are most promising because they apply to *all* SDE-based models. Reflected SDEs (RSDEs) augment the original SDE equation with a second process as follows:

$$dz_t = h(t, z_t) \cdot dt + g(t, z_t) \cdot dB_t + v(t, z_t) \cdot dC_t, \qquad dC_t = \mathbb{I}(z_t \in \partial K) \cdot dt,$$

where $\mathbb{I}(\cdot)$ is an indicator function. Here, dC_t "flips on" when z_t hits the boundary of the space, ∂K, allowing v to neutralize the outward-pointing component of the forward step (Pilipenko, 2014). Thus, RSDEs behave like SDEs on the interior of the space, but are reflected inwards at the boundary.

Despite their rich theory, RSDEs have two shortcomings. First, while RSDE trajectories are continuous (Skorokhod, 1961), their dynamics are not. As such, they may not faithfully describe many natural phenomena in physics, biology, engineering, and medicine (e.g. d'Onofrio (2013); Rohanizadegan et al. (2020)). Second, RSDEs lack efficient, high-order solvers

[*]Equal contribution [1]Department of Computer Science, Wellesley College, Wellesley MA, USA [2]Department of Psychology, Harvard University, Cambridge MA, USA. Correspondence to: Yaniv Yacoby <yy109@wellesley.edu>.

Accepted at Methods and Opportunities at Small Scale (MOSS), ICML 2025, Vancouver, Canada.

(e.g. Ding & Zhang (2008); Fishman et al. (2023b)). These challenges bar their use when model interpretation is important, for example, when modeling irregularly sampled healthcare time series.

Contributions. In this paper, we propose a novel class of expressive (neural) SDEs on compact polyhedral spaces using insights from stochastic viability theory (Aubin, 1991). Polyhedra represent a large class of useful spaces, including bounded rectangular spaces and simplexes, both of which are useful for a variety of temporal natural phenomena (e.g. Cresson et al. (2016)) as well as diffusion models (e.g. Lou & Ermon (2023)). Our contributions are: **(1)** We explain why chain-rule based approaches to SDEs on compact state-spaces struggle theoretically and empirically (Section 2). **(2)** We prove constraints on the drift/diffusion that ensure both stationary and non-stationary SDEs have an inductive bias for compact state spaces (Section 3). **(3)** We propose a neural SDE parameterization that provably satisfies these constraints. Since our parameterization is continuous, it captures a different class of natural phenomena than RSDEs and allows us to use higher-order solvers (Section 4). **(4)** Finally, we empirically demonstrate our parameterization has favorable inductive bias than baselines (Section 5).

Impact. We anticipate that our proposed method will benefit a wide range of models based on stochastic differential equations (SDEs). Importantly, our approach is fully compatible with established inference techniques (e.g. Archambeau et al. (2007); Kidger et al. (2021); Issa et al. (2023); Zhang et al. (2025)), as well as with standard training heuristics for mitigating numerical instability. While broadly applicable, our primary motivation for this work is to advance our understanding of suicide and related behavior. In particular, we work with data from intensive longitudinal studies, consisting of patient self-reports of affect and suicidal ideation, to gain a deeper understanding of how suicide risk develops over time. This data is irregularly sampled, partially observed, and lies a compact, rectangular state-space.

2. Challenges with Transforming SDEs on Euclidean State Spaces to Compact State Spaces

Notation. Consider the following Ito SDE:

$$dz_t = h(t, z_t) \cdot dt + (\text{diag} \circ g)(t, z_t) \cdot dB_t. \tag{1}$$

Here, $t \geq 0$ is time, $z_t \in K$ is the SDE's solution, which lies on a compact subset of Euclidean space, $K \subset \mathbb{R}^{D_z}$. Next, $h : \mathbb{R}_{\geq 0} \times K \to \mathbb{R}^{D_z}$ and $g : \mathbb{R}_{\geq 0} \times K \to \mathbb{R}^{D_z}_{\geq 0}$ are the drift and diffusion, respectively. We overload $\text{diag}(\cdot)$ to transform a vector into a diagonal matrix or to extract the diagonal vector from a matrix. Finally, we use e_d for the dth standard basis vector, ∇ for the Jacobian, $\langle \cdot, \cdot \rangle$ for inner products, and z_t^d, h^d, and g^d for the dth dimension of z_t, h and g, respectively.

Goal. Our goal is to find an expressive, *continuous* parameterization of h and g so that the SDE in Eq. 1 is *viable*:

Definition 2.1 (Milian (1995))**.** A stochastic process, z_t, is *viable* in K if, for every $t \in [0, \infty)$, $\mathbb{P}(z_t \in K) = 1$.

Transforming SDE solutions on $\mathbb{R}^{D_z}$ to solutions on K. The simplest way to ensure z_t lies on a compact space K is to derive a closed-form SDE for $f(z_t)$, where $f : \mathbb{R}^{D_z} \to K$. This is achieved with Ito's lemma for Ito SDEs and with the standard chain-rule for Stratonovich SDEs. While simple, however, this approach presents several challenges.

Challenge 1: Theory. There does not exist a continuous, surjective map, f, from an *open* set, $\mathbb{R}^{D_z}$, to a *closed* set, K. Practically, we may be willing to overlook this, mapping $\mathbb{R}^{D_z}$ to the *interior* of K instead. This is standard practice with simpler models, like classification models, using a sigmoid/softmax transform. But even with these simpler models, we may run into undesirable behavior. For example, given separable classes, the maximum likelihood estimator (MLE) for logistic regression is unbounded and the parameters tend towards infinity (Santner & Duffy, 1986), presenting challenges for model interpretation. This is because when the classes are separable, the MLE can continue to increase by pushing the classes away from each other, maximizing their probability under the sigmoid. Similar challenges may arise with SDEs, for example, when using them to parameterize a time-varying Bernoulli probability for forecasting.

Challenge 2: Numerical Stability. If we're willing to overlook Challenge 1, we find ourselves with numerically unstable dynamics. To see this, suppose we have a 1D SDE, $y_t \in \mathbb{R}$, with drift and diffusion $\tilde{h}$ and $\tilde{g}$, and our goal is to model an SDE $z_t \in (0, 1)^{D_z}$ by transforming y_t into z_t; that is, $z_t = f(y_t)$, where f is a sigmoid. We begin with Ito's lemma:

$$dz_t = \left[\tilde{h}(t, y_t) \cdot \frac{\partial f(y_t)}{\partial y_t} + \frac{1}{2} \cdot \tilde{g}(t, y_t) \cdot \frac{\partial^2 f(y_t)}{\partial^2 y_t}\right] \cdot dt + \left[\frac{1}{2} \cdot \tilde{g}(t, y_t) \cdot \frac{\partial f(y_t)}{\partial y_t}\right] \cdot dB_t.$$

Since we'd like to model dz_t as a function of z_t, we plug in $y_t = f^{-1}(z_t) = \text{sigmoid}^{-1}(z_t)$:

$$dz_t = \left[\tilde{h}(t, f^{-1}(z_t)) \cdot (z_t - z_t^2) + \tilde{g}(t, f^{-1}(z_t)) \cdot (2z_t^3 - 3z_t^2 + z_t)/2\right] \cdot dt + \left[\tilde{g}(t, f^{-1}(z_t)) \cdot (z_t - z_t^2)\right] \cdot dB_t. \quad (2)$$

Here, f^{-1} is unbounded and may lead to unbounded dynamics, leading to numerical instability. Moreover, proofs of SDE existence and uniqueness often require linearly bounded drift and diffusion (e.g. Theorem 5.2.1, Oksendal (2013)).

Challenge 3: Inductive Bias. If we're willing to overlook Challenge 1, we can overcome Challenge 2 as follows. We observe that arbitrarily expressive $\tilde{h}$ and $\tilde{g}$ can "undo" f^{-1} by internally composing it with f. Thus, we can define $h(t, z_t) = \tilde{h}(t, f^{-1}(z_t))$ and $g(t, z_t) = \tilde{g}(t, f^{-1}(z_t))$ and parameterize h and g directly, e.g. via neural networks (NNs):

$$dz_t = \left[h(t, z_t) \cdot (z_t - z_t^2) + g(t, z_t) \cdot (2z_t^3 - 3z_t^2 + z_t)/2\right] \cdot dt + \left[g(t, z_t) \cdot (z_t - z_t^2)\right] \cdot dB_t. \quad (3)$$

So long as h and g are bounded, this SDE has bounded dynamics, overcoming Challenge 2. But, this brings about yet another challenge: the inductive bias of this equation is suitable for few phenomena. As we show empirically in Section 5, samples from Eq. 3 tend to "stick" to the boundaries of the state space. This is undesirable, for example, in intensive longitudinal of mental health, when trajectories tend to oscillate rapidly (e.g. Wang et al. (2024)). This is because both $z_t - z_t^2$ and $2z_t^3 - 3z_t^2 + z_t$ vanish at the boundaries, and so does dz_t. We may hope to learn h and g that can compensate for this behavior by growing rapidly near the boundaries, but empirically, this is unlikely. This is because h and g would need to grow super-cubically near the boundaries, and NNs can struggle to learn polynomials (Cardell et al., 1994).

3. Constraints on Dynamics for SDEs on Compact Polyhedral State Spaces

Motivated by the challenges from Section 2, we prove necessary constraints on the drift/diffusion to ensure that both stationary and non-stationary SDEs have an inductive bias for compact polyhedral state spaces. Our approach leverages insights from stochastic viability theory (Aubin, 1991). To begin, we define polyhedral subspaces as follows:

Definition 3.1. Let $u, v \in \mathbb{R}^{D_z}$ and $\mathcal{H}(u, v) = \{z \in \mathbb{R}^{D_z} : \langle z - u, v\rangle \geq 0\}$ denote a closed half-space. A set $K \subset \mathbb{R}^{D_z}$ is a *polyhedron* if it is a finite intersection of closed half-spaces: $K = \bigcap_{s\in\{1,\cdots,S\}} \mathcal{H}(u_s, v_s)$.

In Theorem 3.2, Milian (1995) shows that, with linearly-bounded, Lipschitz continuous drift and diffusion, an Ito SDE is viable in a polyhedral subspace, K, if and only if (a) the drift pushes the trajectory towards the interior of K when $z_t \in \partial K$, and (b) the diffusion vanishes when $z_t \in \partial K$. Thus, at the boundary, the trajectory is deterministically pushed inwards. While Theorem 3.2 also holds for non-compact polyhedra, we focus on compact polyhedra from here on. In Appendix A, we extend this result to Stratonovich SDEs on compact polyhedra.

Theorem 3.2 (Milian (1995))**.** ***Suppose*** *that the drift and diffusion, $h(t, z_t)$ and $g(t, z_t)$, of an Ito SDE, defined for $t \geq 0$ and $z_t \in \mathbb{R}^{D_z}$, satisfy three conditions:* ***(i)*** *For each $T > 0$, there exists $C_T > 0$ such that for all $z_t \in K$ and $t \in [0, T]$, $\|h(t, z_t)\|^2 + \|g(t, z_t)\|^2 \leq C_T \cdot (1 + \|z_t\|^2)$.* ***(ii)*** *For all $T > 0$, $z_t, z_t' \in K$, and $t \in [0, T]$, $\|h(t, z_t) - h(t, z_t')\| + \|g(t, z_t) - g(t, z_t')\| \leq C_T \cdot \|z_t - z_t'\|$.* ***(iii)*** *For each $z_t \in K$, $h(t, z_t)$ and $g(t, z_t)$ are continuous.* ***Then*** *z_t is viable in K if and only if: for all $s \in [1, \ldots, S]$ and $z_t \in K$ such that when $\langle z_t - u_s, v_s\rangle = 0$, we have* ***(a)*** *$\langle h(t, z_t), v_s\rangle \geq 0$ and* ***(b)*** *$\langle g(t, z_t) \odot e_d, v_s\rangle = 0$ for $t \geq 0$ and $d \in [1, \ldots, D_z]$.*

We now extend this theorem to stationary SDEs on compact polyhedra by selecting a diffusion, g, that satisfies (i)-(iii) and (b) from Theorem 3.2, deriving a closed-form equation for the drift h as a function of g that ensures stationarity, and proving that h also satisfies all conditions from Theorem 3.2 (see proof in Appendix B).

Theorem 3.3. *Let K be a compact polyhedron, and $h(z_t)$ and $g(z_t)$ be the drift and diffusion, respectively, of an autonomous Ito SDE, defined for $t \geq 0$ and $z_t \in \mathbb{R}^{D_z}$.* ***Suppose*** *that for all $T > 0$, $z_t, z_t' \in K$, and $t \in [0, T]$, there exists $C_T > 0$ such that:* ***(i)*** *$\|g(z_t)\|^2 \leq C_T \cdot (1 + \|z_t\|^2)$.* ***(ii)*** *$\|g(z_t) - g(z_t')\| \leq C_T \cdot \|z_t - z_t'\|$ and $\|\text{diag}(\nabla_{z_t} g(z_t)) - \text{diag}(\nabla_{z_t'} g(z_t'))\| \leq C_T \cdot \|z_t - z_t'\|$.* ***(iii)*** *The unnormalized time-marginal, $\tilde{p}(t, z_t)$ satisfies $\|\log \tilde{p}(t, z_t) - \log \tilde{p}(t, z_t')\| \leq C_T \cdot \|z_t - z_t'\|$.* ***(iv)*** *$g(z_t)$ and $\tilde{p}(t, z_t')$ are differentiable with continuous partials.* ***Then*** *z_t is a solution to a stationary SDE with time-marginal, $p(t, z_t) = p(z_t)$, viable in K if:* ***(a)*** *$h(z_t) = \frac{1}{2} \cdot \text{diag}\left(\nabla_{z_t}[g(z_t)^2]\right) + \frac{1}{2} \cdot g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t)$.* ***(b)*** *For all $s \in [1, \ldots, S]$ and $z_t \in K$ such that when $\langle z_t - u_s, v_s\rangle = 0$, we have $\langle g(z_t) \odot e_d, v_s\rangle = 0$ for $t \geq 0$ and $d \in [1, \ldots, D_z]$.*

The parameterization of $h(z_t)$ in Theorem 3.3 is easily implemented in auto-differentiation frameworks with $g(z_t)$ and $\log \tilde{p}(z_t)$ ("score function") as NNs. We note that assumptions (i)-(iv) are easily satisfiable—see discussion in Appendix D. We will next describe a novel parameterization to satisfy conditions (a) and (b) from Theorem 3.2 and (b) from Theorem 3.3.

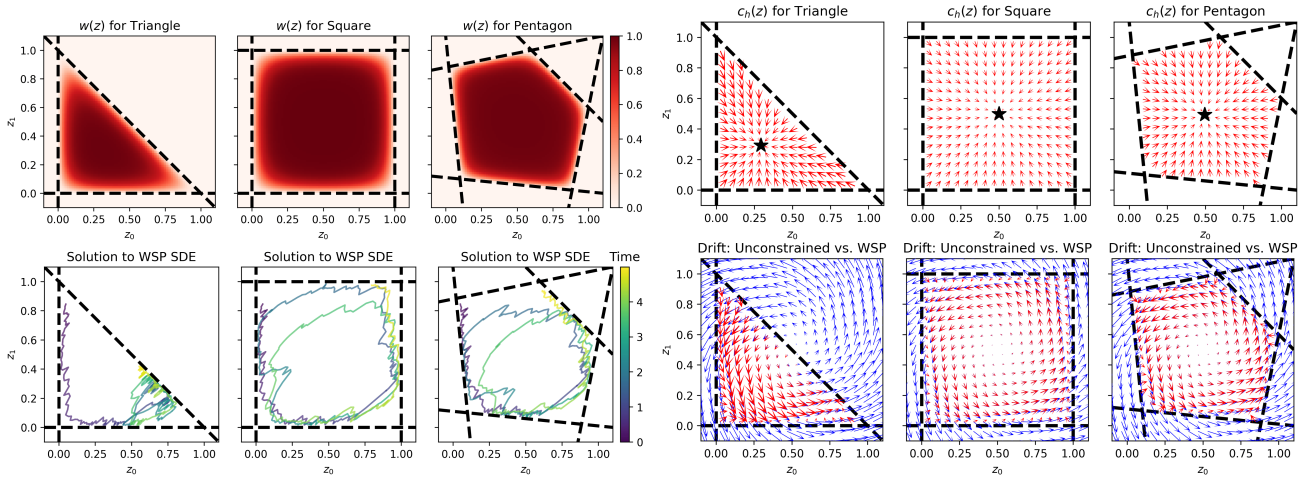

Figure 1. **Intuition behind WSP for different polyhedra.** Top left: $w(z)$ from Eq. 4, approaching 0 at the boundaries and 1 in the interior. Top right: $c_h(z)$ from Eq. 5, pointing towards the Chebyshev center ★. Bottom left: solutions to WSP SDE, successfully remaining in K. Bottom right: some unconstrained drift $\tilde{h}$ vs. WSP drift h (Eq. 5) matching in the interior of K, but differing near the bounds.

4. Parameterization of Expressive SDEs on Compact Polyhedral State Spaces

We propose a flexible, constraint-satisfying parameterization of drifts and diffusions from Theorems 3.2 and 3.3 that transforms any unconstrained SDE dynamics, whether NN-based or expert given, to lie on a target compact polyhedron.

Weighted Sums Parameterization (WSP). We observe that we can satisfy both constraints on the drift and diffusion using the same mechanism: $\text{WSP}(f, c, t, z) = w(z) \cdot f(t, z) + (1 - w(z)) \cdot c(z)$, using a different choice of $c(z)$ for each. Here, $f(\cdot)$ is the original, unconstrained dynamics, given by domain experts or by some flexible function class, like NNs; $c(\cdot)$ is a simple function that satisfies the constraints; finally, $w(z) \in [0, 1]$ weighs the of sum $f(\cdot)$ and $c(\cdot)$, approaching 0 at ∂K to favor $c(\cdot)$, and approaching 1 at the interior of K to favor $f(\cdot)$. Of many possible choices, we define $w(z)$ as:

$$w(z) = \tanh\left(\beta \cdot \prod_s \frac{e^{-d(u_s, v_s, z)}}{\sum_{s'} e^{-d(u_{s'}, v_{s'}, z)}} \cdot \tanh\left(\alpha \cdot d(u_s, v_s, z)\right)\right), \qquad d(u, v, z) = \frac{\langle z - u, v \rangle}{\|v\|}. \tag{4}$$

Here, $d(u, v, z) \geq 0$ is the shortest distance from z to the boundary of $\mathcal{H}(u, v)$, and $\alpha, \beta > 0$ are constants that determine how quickly to transition between $f(z)$ and $c(z)$. They can be learned jointly with the model parameters. The intuition behind Eq. 4 is that $w(z)$ should approach 1 as z approaches the *closest* of the S boundaries. As such, we take a convex combination of distances from z to each boundary, weighted by a softmin; this, in a sense, "selects" the closest distance. By taking a product of these weighted distances, we obtain a function that is 0 at all boundaries and positive elsewhere, using $\tanh$ to ensure $w(z) \in [0, 1]$. Fig. 1 (top-left) visualizes $w(z)$ for different polyhedra.

WSP-based SDEs. Given any unconstrained drift and diffusion, $\tilde{h} : \mathbb{R}_{\geq 0} \times K \to \mathbb{R}^{D_z}$ and $\tilde{g} : \mathbb{R}_{\geq 0} \times K \to \mathbb{R}^{D_z}_{\geq 0}$, we use WSP to obtain new drift and diffusion, h ans g, that satisfy (a)-(b) from Theorem 3.2:

$$\begin{aligned} h(t, z_t) &= \text{WSP}(\tilde{h}, c_h, t, z_t), \qquad c_h(z_t) = \gamma \cdot \frac{z^* - z_t}{\|z^* - z_t\| + \epsilon}, \quad \text{where} \quad z^* = \text{argmin}_{\bar{z}} \text{max}_{z \in K} \|z - \bar{z}\|^2, \\ g(t, z_t) &= \text{WSP}(\tilde{g}, c_g, t, z_t), \qquad c_g(z_t) = 0. \end{aligned} \tag{5}$$

Here, z^* is the Chebyshev center of K, easily computed via linear programming once per polyhedron (Boyd, 2004). $c_h(z_t)$ is therefore a push towards the center of K, with magnitude controlled by $\gamma > 0$ and $\epsilon > 0$, learned jointly with the other model parameters. We note that there are many possible choices of $c_h(z_t)$; we selected this one for its simplicity. Fig. 1 visualizes $c_h(z)$ and h for different polyhedra, showing WSP SDEs remain viable in K.

Next, we prove that, under Lipschitz continuity and linear boundedness of $\tilde{h}$ and $\tilde{g}$, WSP dynamics satisfy conditions of Theorem 3.2 (implying the diffusion g also satisfies conditions for Theorem 3.3)—see proof in Appendix C.

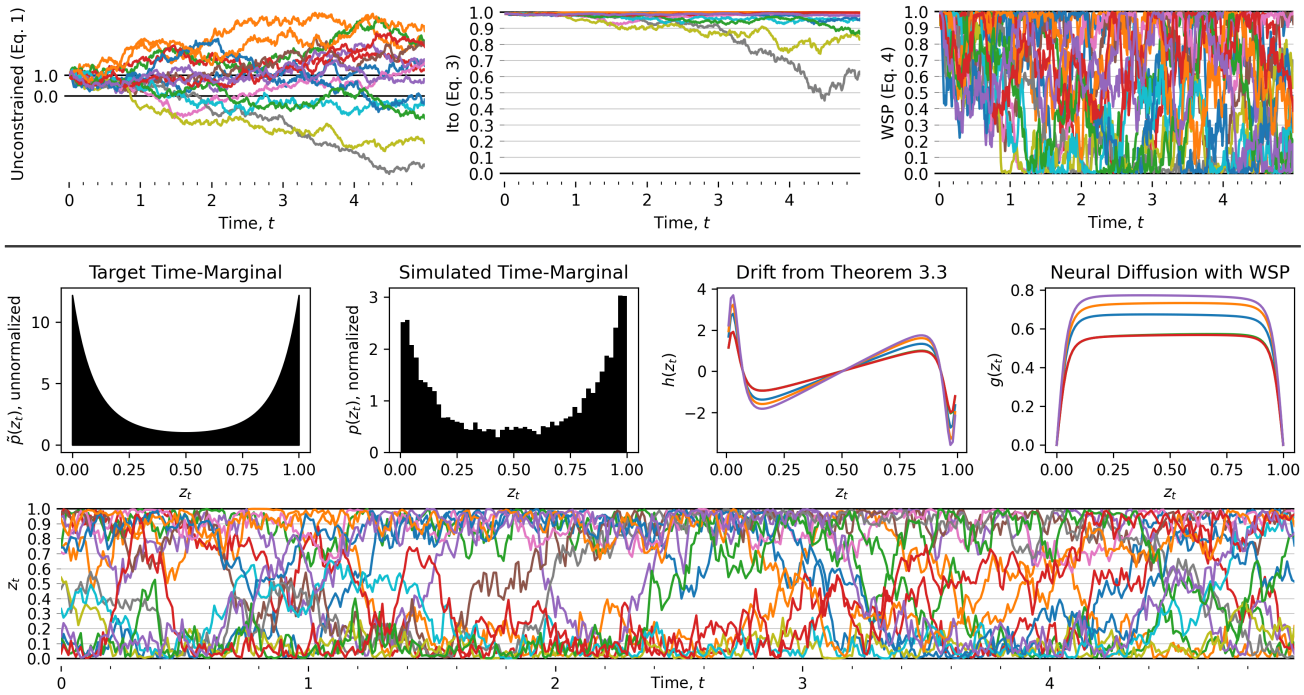

Figure 2. **Top: WSP exhibits better inductive bias than baselines.** Left: unconstrained SDE (Eq. 1) with NN quickly leaves $K = [0, 1]$. Middle: SDE transformed via sigmoid (Eqs. 2 and 3) sticks to the boundary. Right: SDE with WSP (Eq. 5) successfully remains in K. **Bottom: Stationary WSP exhibits favorable inductive bias.** Given a target time-marginal and WSP diffusion, drift derived from Theorem 3.3 yields an SDE viable in K with target stationary distribution.

Theorem 4.1. *Let K be a compact polyhedron. **Suppose** that $\tilde{h}$ and $\tilde{g}$, defined above, satisfy three conditions: **(i)** For each $T > 0$, there exists $C_T > 0$ such that for all $z_t \in K$ and $t \in [0, T]$, $\|\tilde{h}(t, z_t)\|^2 + \|\tilde{g}(t, z_t)\|^2 \leq C_T \cdot (1 + \|z_t\|^2)$. **(ii)** For all $T > 0$, $z_t, z'_t \in K$, and $t \in [0, T]$, $\|\tilde{h}(t, z_t) - \tilde{h}(t, z'_t)\| + \|\tilde{g}(t, z_t) - \tilde{g}(t, z'_t)\| \leq C_T \cdot \|z_t - z'_t\|$. **(iii)** For each $z_t \in K$, $\tilde{h}(t, z_t)$ and $\tilde{g}(t, z_t)$ are continuous. **Then** the solution z_t to the SDE with drift and diffusion, $h(t, z_t)$ and $g(t, z_t)$, defined in Eq. 5, is viable in K.*

5. Experiments, Results, and Future Work

Experiments: comparing the inductive biases of WSP vs. baselines. As discussed in Section 1, initialization plays a crucial role in the success of expressive SDE-based models. Thus, to empirically compare the inductive bias of WSP (Eq. 5) against baselines (Eqs. 1–3), we solve SDEs given by NN drifts and diffusions with randomly sampled weights, and with z_0 near the boundary to compare their inductive biases. While simple, *these preliminary experiments already show a stark improvement in inductive bias of WSP compared to baselines.* Details on setup in Appendix E.

Results: WSP has favorable inductive bias relative to baselines. Fig. 2 (top) shows WSP ensures SDE samples are viable in $K = [0, 1]$. In contrast, unconstrained SDEs quickly leave K, and SDEs based on Ito's lemma stick to the boundary (as explained in Section 2), limiting their expressivity. In Appendix F.1, we demonstrate WSP's favorable inductive bias also when Brownian motion is replaced with a smooth, pathwise expansion, which allows us to replace the SDE solver with a fast, adaptive ODE solver. Finally, in Fig. 2 (bottom), we show that the stationary SDE from Theorem 3.3 boasts the same favorable inductive bias, allowing us to match any target time-marginal. For additional results, see Appendix F.2.

Future Work. While our specific parameterization of WSP (i.e. choice of $w(z)$ and $c_h(z)$) shows favorable inductive bias over baselines, we do not yet know how amenable it is to gradient-based optimization, for example when used in a latent SDE or diffusion model. Thus, in future work, we will conduct empirical comparisons of WSP with baselines to determine whether it enables better model fit and more stable training dynamics.

Acknowledgments

The authors are grateful for funding from NIMH (U01MH116928) and from the Fuss Family Research Fund and the Chet and Will Griswold Suicide Prevention Fund. The authors are grateful to the Wellesley College Science Center Summer Research Program for supporting YL in the summer of 2025.

References

Anil, C., Lucas, J., and Grosse, R. Sorting out lipschitz function approximation. In *International conference on machine learning*, pp. 291–301. PMLR, 2019.

Ansari, A. F., Heng, A., Lim, A., and Soh, H. Neural continuous-discrete state space models for irregularly-sampled time series. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 926–951. PMLR, 23–29 Jul 2023.

Archambeau, C., Opper, M., Shen, Y., Cornford, D., and Shawe-Taylor, J. Variational inference for diffusion processes. *Advances in neural information processing systems*, 20, 2007.

Aubin, J.-P. *Viability Theory*. Modern Birkhäuser Classics. Birkhauser Boston, Secaucus, NJ, January 1991.

Barron, J. T. Continuously differentiable exponential linear units. *arXiv preprint arXiv:1704.07483*, 2017.

Boyd, S. Convex optimization. *Cambridge UP*, 2004.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., et al. Jax: composable transformations of python+ numpy programs. 2018.

Cai, G. and Lin, Y. Generation of non-gaussian stationary stochastic processes. *Physical Review E*, 54(1):299, 1996.

Cardell, N. S., Joerding, W., and Li, Y. Why some feedforward networks cannot learn some polynomials. *Neural computation*, 6(4):761–766, 1994.

Christopher, J. K., Baek, S., and Fioretto, N. Constrained synthesis with projected diffusion models. *Advances in Neural Information Processing Systems*, 37:89307–89333, 2024.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

Cresson, J., Puig, B., and Sonner, S. Stochastic models in biology and the invariance problem. *Discrete & Continuous Dynamical Systems-Series B*, 21(7), 2016.

Ding, D. and Zhang, Y. Y. Numerical solutions for reflected stochastic differential equations, 2008.

d'Onofrio, A. *Bounded noises in physics, biology, and engineering*. Springer, 2013.

Elfwing, S., Uchibe, E., and Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

Esmayli, B. Prove that the product of two lipschitz functions is locally lipschitz. Mathematics Stack Exchange, 2017. URL https://math.stackexchange.com/q/2171798. URL:https://math.stackexchange.com/q/2171798 (version: 2021-10-12).

Fishman, N., Klarner, L., Bortoli, V. D., Mathieu, E., and Hutchinson, M. J. Diffusion models for constrained domains. *Transactions on Machine Learning Research*, 2023a. ISSN 2835-8856.

Fishman, N., Klarner, L., Mathieu, E., Hutchinson, M., and De Bortoli, V. Metropolis sampling for constrained diffusion models. *Advances in Neural Information Processing Systems*, 36:62296–62331, 2023b.

Ghosh, S., Birrell, P. J., and De Angelis, D. Differentiable bayesian inference of sde parameters using a pathwise series expansion of brownian motion. In *International Conference on Artificial Intelligence and Statistics*, pp. 10982–10998. PMLR, 2022.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Gyawali, P., Ghimire, S., and Wang, L. Enhancing mixup-based semi-supervised learningwith explicit lipschitz regularization. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 1046–1051. IEEE, 2020.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Hurault, S., Leclaire, A., and Papadakis, N. Gradient step denoiser for convergent plug-and-play. In *International Conference on Learning Representations (ICLR'22)*, 2022.

Issa, Z., Horvath, B., Lemercier, M., and Salvi, C. Non-adversarial training of neural sdes with signature kernel scores. *Advances in Neural Information Processing Systems*, 36:11102–11126, 2023.

Kidger, P. *On Neural Differential Equations*. PhD thesis, University of Oxford, 2021.

Kidger, P., Foster, J., Li, X., and Lyons, T. J. Neural sdes as infinite-dimensional gans. In *International conference on machine learning*, pp. 5453–5463. PMLR, 2021.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017.

Li, X., Wong, T.-K. L., Chen, R. T., and Duvenaud, D. Scalable gradients for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, pp. 3870–3882. PMLR, 2020.

Liu, H.-T. D., Williams, F., Jacobson, A., Fidler, S., and Litany, O. Learning smooth neural functions via lipschitz regularization. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–13, 2022.

Lou, A. and Ermon, S. Reflected diffusion models. In *International Conference on Machine Learning*, pp. 22675–22701. PMLR, 2023.

Ma, Y.-A., Chen, T., and Fox, E. A complete recipe for stochastic gradient mcmc. *Advances in neural information processing systems*, 28, 2015.

Milian, A. Stochastic viability and a comparison theorem. In *Colloquium Mathematicum*, volume 68, pp. 297–316. Polska Akademia Nauk. Instytut Matematyczny PAN, 1995.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Oh, Y., Lim, D., and Kim, S. Stable neural stochastic differential equations in analyzing irregular time series data. In *The Twelfth International Conference on Learning Representations*, 2024.

Oksendal, B. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

Ormandy, C. Linking sampling and stochastic differential equations, 2019. URL https://chrisorm.github.io/SDE-S.html.

Phan, D., Pradhan, N., and Jankowiak, M. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.

Pilipenko, A. *An introduction to stochastic differential equations with reflection*, volume 1. Universitätsverlag Potsdam, 2014.

Rohanizadegan, Y., Sonner, S., and Eberl, H. J. Discrete attachment to a cellulolytic biofilm modeled by an itô stochastic differential equation. 2020.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Santner, T. J. and Duffy, D. E. A note on a. albert and ja anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 73(3):755–758, 1986.

Särkkä, S. and Solin, A. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019.

Skorokhod, A. V. Stochastic equations for diffusion processes in a bounded region. *Theory of Probability & Its Applications*, 6(3):264–274, 1961.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Wang, S. B., Van Genugten, R. D., Yacoby, Y., Pan, W., Bentley, K. H., Bird, S. A., Buonopane, R. J., Christie, A., Daniel, M., DeMarco, D., et al. Building personalized machine learning models using real-time monitoring data to predict idiographic suicidal thoughts. *Nature Mental Health*, pp. 1–10, 2024.

Wong, E. and Zakai, M. On the convergence of ordinary integrals to stochastic integrals. *The Annals of Mathematical Statistics*, 36(5):1560–1564, 1965.

Xu, W., Chen, R. T., Li, X., and Duvenaud, D. Infinitely deep bayesian neural networks with stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, pp. 721–738. PMLR, 2022.

Zhang, J., Viktorov, J., Jung, D., and Pitler, E. Efficient training of neural stochastic differential equations by matching finite dimensional distributions. In *The Thirteenth International Conference on Learning Representations*, 2025.

A. Extending Theorem 3.2 to Stratonovich SDEs on Compact Polyhedra

Corollary A.1. ***Suppose*** *that the drift and diffusion, $h(t, z_t)$ and $g(t, z_t)$, of a Stratonovich SDE, defined for $t \geq 0$ and $z_t \in \mathbb{R}^{D_z}$, satisfy conditions (i)-(iii) from Theorem 3.2. Suppose further that for all $T > 0$, $z_t, z'_t \in K$, and $t \in [0, T]$, $\|\text{diag}(\nabla_{z_t} g(t, z_t)) - \text{diag}(\nabla_{z'_t} g(t, z'_t))\| \leq C_T \cdot \|z_t - z'_t\|$.* ***Then*** *$z_t$ is viable in compact polyhedron K if and only if (a)-(b) from Theorem 3.2 hold.*

Proof. Given the Stratonovich interpretation of the SDE in Eq. 1,

$$dz_t = h(t, z_t) \cdot dt + (\text{diag} \circ g)(t, z_t) \circ dB_t, \tag{6}$$

we can write the equivalent Ito SDE as follows:

$$dz_t = \hat{h}(t, z_t) \cdot dt + (\text{diag} \circ g)(t, z_t) \cdot dB_t, \tag{7}$$

where,

$$\hat{h}(t, z_t) = h(t, z_t) + \frac{1}{2} \cdot \text{diag}(\nabla_{z_t} g(t, z_t)) \odot g(t, z_t). \tag{8}$$

Since in Eq. 7, the diffusion is unchanged, we only need to show that when h satisfies (i)-(iii) and (a), so does $\hat{h}$.

Proof that $\tilde{h}$ satisfies (i) from Theorem 3.2. We prove that for each $T > 0$, there exists $C_T > 0$ such that for all $z_t \in K$ and $t \in [0, T]$, $\|\tilde{h}(z_t)\|^2 \leq C_T \cdot (1 + \|z_t\|^2)$. We do this as follows:

$$\|\tilde{h}(z_t)\| = \left\| h(t, z_t) + \frac{1}{2} \cdot \text{diag}(\nabla_{z_t} g(t, z_t)) \odot g(t, z_t) \right\| \tag{9}$$

$$\leq \|h(t, z_t)\| + \frac{1}{2} \cdot \|\text{diag}(\nabla_{z_t} g(t, z_t)) \odot g(t, z_t)\| \tag{10}$$

$$\leq \|h(t, z_t)\| + \frac{1}{2} \cdot \|\text{diag}(\nabla_{z_t} g(t, z_t))\| \cdot \|g(t, z_t)\| \tag{11}$$

$$\leq \underbrace{\sqrt{C'_T \cdot (1 + \|z_t\|^2)}}_{\text{bounded via condition (i)}} + \underbrace{\frac{1}{2} \cdot \|\text{diag}(\nabla_{z_t} g(t, z_t))\|}_{\text{bounded by const. via condition (ii)}} \cdot \underbrace{\sqrt{C'_T \cdot (1 + \|z_t\|^2)}}_{\text{bounded via condition (i)}} \tag{12}$$

The above line can be written in the form $(1 + B) \cdot \sqrt{C'_T \cdot (1 + \|z_t\|^2)}$, which, when squared, gives us an inequality of the form, $\|\tilde{h}(z_t)\|^2 \leq C_T \cdot (1 + \|z_t\|^2)$.

Proof that $\tilde{h}$ satisfies (ii) from Theorem 3.2. Here, we prove that for all $T > 0$, $z_t, z'_t \in K$, and $t \in [0, T]$, $\|\tilde{h}(z_t) - \tilde{h}(z'_t)\| \leq C_T \cdot \|z_t - z'_t\|$:

$$\|\tilde{h}(z_t) - \tilde{h}(z'_t)\| = \left\| h(t, z_t) + \frac{1}{2} \cdot \text{diag}(\nabla_{z_t} g(t, z_t)) \odot g(t, z_t) - h(t, z'_t) - \frac{1}{2} \cdot \text{diag}(\nabla_{z'_t} g(t, z'_t)) \odot g(t, z'_t) \right\| \tag{13}$$

$$= \left\| h(t, z_t) - h(t', z'_t) + \frac{1}{2} \cdot \text{diag}(\nabla_{z_t} g(t, z_t)) \odot g(t, z_t) - \frac{1}{2} \cdot \text{diag}(\nabla_{z'_t} g(t, z'_t)) \odot g(t, z'_t) \right\| \tag{14}$$

$$\leq \|h(t, z_t) - h(t', z'_t)\| + \left\| \frac{1}{2} \cdot \text{diag}(\nabla_{z_t} g(t, z_t)) \odot g(t, z_t) - \frac{1}{2} \cdot \text{diag}(\nabla_{z'_t} g(t, z'_t)) \odot g(t, z'_t) \right\| \tag{15}$$

Using the trick by Esmayli (2017), we have

$$\|\tilde{h}(z_t) - \tilde{h}(z'_t)\| \leq \underbrace{\|h(t, z_t) - h(t', z'_t)\|}_{\leq C'_T \cdot \|z_t - z'_t\|} \tag{16}$$

$$+ \frac{1}{2} \cdot \underbrace{\left\|\text{diag}(\nabla_{z_t} g(t, z_t)) - \text{diag}(\nabla_{z'_t} g(t, z'_t))\right\|}_{\leq C'_T \cdot \|z_t - z'_t\|} \cdot \|g(t, z_t)\|$$

$$+ \frac{1}{2} \cdot \left\|\text{diag}(\nabla_{z'_t} g(t, z'_t))\right\| \cdot \underbrace{\|g(t, z_t) - g(t, z'_t)\|}_{\leq C'_T \cdot \|z_t - z'_t\|}$$

Finally, since both $g(t, z_t)$ and $\text{diag}(\nabla_{z'_t} g(t, z'_t))$ are Lipschitz on a bounded domain, they can be bounded by a constant.

Proof that $\tilde{h}$ satisfies (iii) from Theorem 3.2. Since $\tilde{h}$ is comprised of addition and scaling operations on continuous functions, it is also continuous.

Proof that $\tilde{h}$ satisfies (a) from Theorem 3.2. When $\langle z_t - u_s, v_s \rangle = 0$, we show (a) holds for $\tilde{h}$ as follows:

$$\langle \hat{h}(t, z_t), v_s \rangle = \underbrace{\langle h(t, z_t), v_s \rangle}_{\geq 0} + \frac{1}{2} \cdot \sum_{d=1}^{D_z} \underbrace{\frac{\partial g^d(t, z_t)}{\partial z_t^d}}_{\text{bounded}} \cdot \underbrace{g^d(t, z_t) \cdot v_s^d}_{=0} \geq 0. \tag{17}$$

The first term is non-negative. The second term is 0 since $\langle g(t, z_t) \odot e_d, v_s \rangle = 0$ when condition (b) holds for g, and when 0 is multiplied by the partial (bounded thanks to condition (ii) for g), we get 0. Thus, $\langle \hat{h}(t, z_t), v_s \rangle \geq 0$.

□

B. Proof of Theorem 3.3

Proof. To prove Theorem 3.3, we will show that the form for the drift listed in condition (a) results in a stationary SDE. Then, we will prove that this stationary SDE satisfies all conditions from Theorem 3.2, implying it is viable in K.

We find h by drawing inspiration from the derivation Cai & Lin (1996), which sets the Fokker-Planck-Kolmogorov (FPK) equation to 0 to obtain stationarity and then solves for the dynamics. In contrast to Cai & Lin (1996), instead of solving for the diffusion, which, in general, requires us to compute an intractable integral with no closed-form, we solve for the drift. This part of the proof is similar in spirit to the derivation of the stationary SDE used for stochastic gradient MCMC (Ma et al., 2015; Ormandy, 2019).

We begin by setting FPK equation equal to 0 to obtain stationarity for general SDEs of the form, $dz_t = h(z_t) \cdot dt + g(z_t) \cdot dB_t$, where $g(z_t) \in \mathbb{R}^{D_z \times D_z}$ is a full matrix. We will then adapt it to our case. As such, we denote $G(z_t) = g(z_t) \cdot \Sigma \cdot g(z_t)^\intercal$, where Σ is the covariance of the Brownian motion.

$$0 = \frac{\partial}{\partial t} p(t, z_t) = \frac{\partial}{\partial t} p(z_t) \quad \text{(simplified notation)} \tag{18}$$

$$= -\sum_{d=1}^{D_z} \frac{\partial}{\partial z_t^d} \left[h^d(z_t) \cdot p(z_t) \right] + \frac{1}{2} \sum_{d=1}^{D_z} \sum_{d'=1}^{D_z} \frac{\partial^2}{\partial z_t^d \partial z_t^{d'}} \left[G^{d,d'}(z_t) \cdot p(z_t) \right] \tag{19}$$

$$= \sum_{d=1}^{D_z} \frac{\partial}{\partial z_t^d} \left(-h^d(z_t) \cdot p(z_t) + \frac{1}{2} \sum_{d'=1}^{D_z} \frac{\partial}{\partial z_t^{d'}} \left[G^{d,d'}(z_t) \cdot p(z_t) \right] \right) \tag{20}$$

For this theorem, we only concern ourselves with identity covariance Brownian motion and diagonal diffusion; that is, we set $\Sigma = I$ and $g(z_t) \in \mathbb{R}^{D_z}$ to be a vector. As such, the above equation simplifies to:

$$0 = \sum_{d=1}^{D_z} \frac{\partial}{\partial z_t^d} \left(-h^d(z_t) \cdot p(z_t) + \frac{1}{2} \cdot \frac{\partial}{\partial z_t^{d'}} \left[g^d(z_t)^2 \cdot p(z_t) \right] \right) \tag{21}$$

One way to solve this equation for the drift is to ensure that for every $d \in [1, D_z]$.

$$h^d(z_t) \cdot p(z_t) = \frac{1}{2} \cdot \frac{\partial}{\partial z_t^{d'}} \left[g^d(z_t)^2 \cdot p(z_t) \right]. \tag{22}$$

We can then solve for the drift:

$$h^d(z_t) = \frac{1}{2 \cdot p(z_t)} \frac{\partial}{\partial z_t^d} \left[g^d(z_t)^2 \cdot p(z_t) \right] \tag{23}$$

$$= \frac{1}{2 \cdot \tilde{p}(z_t)/A} \cdot \frac{\partial}{\partial z_t^d} \left[g^d(z_t)^2 \cdot \tilde{p}(z_t)/A \right] \quad (A \text{ is the normalizing const.}) \tag{24}$$

$$= \frac{1}{2 \cdot \tilde{p}(z_t)} \cdot \frac{\partial}{\partial z_t^d} \left[g^d(z_t)^2 \cdot \tilde{p}(z_t) \right] \tag{25}$$

$$= g^d(z_t) \cdot \frac{\partial}{\partial z_t^d} g^d(z_t) + \frac{1}{2} \cdot g^d(z_t)^2 \cdot \frac{\frac{\partial}{\partial z_t^d} \tilde{p}(z_t)}{\tilde{p}(z_t)} \tag{26}$$

$$= g^d(z_t) \cdot \frac{\partial}{\partial z_t^d} g^d(z_t) + \frac{1}{2} \cdot g^d(z_t)^2 \cdot \frac{\partial}{\partial z_t^d} \log \tilde{p}(z_t) \quad \text{(log derivative trick)} \tag{27}$$

$$= \frac{1}{2} \frac{\partial}{\partial z_t^d} [g^d(z_t)^2] + \frac{1}{2} \cdot g^d(z_t)^2 \cdot \frac{\partial}{\partial z_t^d} \log \tilde{p}(z_t) \tag{28}$$

This gives us:

$$h(z_t) = \frac{1}{2} \cdot \operatorname{diag}\left(\nabla_{z_t} [g(z_t)^2] \right) + \frac{1}{2} \cdot g(z_t)^2 \odot \underbrace{\nabla_{z_t} \log \tilde{p}(z_t)}_{\text{score function}}. \tag{29}$$

Now that we have derived a closed-form equation for the drift, we will prove it satisfies all conditions of Theorem 3.2, thereby proving Theorem 3.3.

Proof that h satisfies (i) from Theorem 3.2. Here, we prove that for each $T > 0$, there exists $C_T > 0$ such that for all $z_t \in K$ and $t \in [0, T]$, $\|h(z_t)\|^2 \leq C_T \cdot (1 + \|z_t\|^2)$. We do this as follows:

$$\|h(z_t)\| = \frac{1}{2} \cdot \left\| \operatorname{diag}\left(\nabla_{z_t} [g(z_t)^2] \right) + g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t) \right\| \tag{30}$$

$$\leq \frac{1}{2} \cdot \left\| \operatorname{diag}\left(\nabla_{z_t} [g(z_t)^2] \right) \right\| + \frac{1}{2} \left\| g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t) \right\| \tag{31}$$

$$\leq \frac{1}{2} \cdot \underbrace{\left\| \operatorname{diag}\left(\nabla_{z_t} [g(z_t)^2] \right) \right\|}_{②} + \frac{1}{2} \underbrace{\|g(z_t)\|^2}_{①} \cdot \underbrace{\|\nabla_{z_t} \log \tilde{p}(z_t)\|}_{\text{bounded by const. via (ii)}} \tag{32}$$

Since K is compact and $g(z_t)$ is continuous and linearly bounded (i.e. $\|g(z_t)\|^2 \leq C_T(1 + \|z_t\|^2)$), we know that there exists some maximal value, M, that bounds g on K:

$$① = \|g(z_t)\| \leq \max_{z_t \in K} \|g(z_t)\| = M. \tag{33}$$

Next, we bound the gradient of g using condition (ii):

$$
\begin{aligned}
② &= \left\|\operatorname{diag}\left(\nabla_{z_t}[g(z_t)^2]\right)\right\| && (34)\\
&= \left\| \lim_{\epsilon\to 0} \begin{bmatrix} \frac{g^1(z_t+\epsilon\cdot e_1)^2 - g^1(z_t)^2}{\epsilon} \\ \vdots \\ \frac{g^D(z_t+\epsilon\cdot e_D)^2 - g^D(z_t)^2}{\epsilon} \end{bmatrix} \right\| && (35)\\
&= \lim_{\epsilon\to 0} \left\| \begin{bmatrix} \frac{g^1(z_t+\epsilon\cdot e_1)^2 - g^1(z_t)^2}{\epsilon} \\ \vdots \\ \frac{g^D(z_t+\epsilon\cdot e_D)^2 - g^D(z_t)^2}{\epsilon} \end{bmatrix} \right\| \quad \text{(by continuity of the norm)} && (36)\\
&\leq \sum_{d=1}^{D} \lim_{\epsilon\to 0} \left\| \frac{g^d(z_t+\epsilon\cdot e_d)^2 - g^d(z_t)^2}{\epsilon} \right\| \quad \text{(since the } \ell_2\text{-norm is upper bounded by the } \ell_1\text{-norm)} && (37)\\
&= \sum_{d=1}^{D} \lim_{\epsilon\to 0} \frac{1}{\epsilon} \cdot \left\| g^d(z_t+\epsilon\cdot e_d)^2 - g^d(z_t)^2 \right\| && (38)\\
&\leq \sum_{d=1}^{D} \lim_{\epsilon\to 0} \frac{1}{\epsilon} \cdot \left\| g(z_t+\epsilon\cdot e_d)^2 - g(z_t)^2 \right\| && (39)\\
&= \sum_{d=1}^{D} \lim_{\epsilon\to 0} \frac{1}{\epsilon} \cdot \left\| (g(z_t+\epsilon\cdot e_d) - g(z_t)) \odot g(z_t+\epsilon\cdot e_d) - g(z_t) \odot (g(z_t) - g(z_t+\epsilon\cdot e_d)) \right\| && (40)\\
&\leq \sum_{d=1}^{D} \lim_{\epsilon\to 0} \frac{1}{\epsilon} \cdot \left\| (g(z_t+\epsilon\cdot e_d) - g(z_t)) \odot g(z_t+\epsilon\cdot e_d) \right\| + \frac{1}{\epsilon} \cdot \left\| g(z_t) \odot (g(z_t) - g(z_t+\epsilon\cdot e_d)) \right\| && (41)\\
&\leq \sum_{d=1}^{D} \lim_{\epsilon\to 0} \frac{1}{\epsilon} \cdot \left\| g(z_t+\epsilon\cdot e_d) - g(z_t) \right\| \cdot \left\| g(z_t+\epsilon\cdot e_d) \right\| + \frac{1}{\epsilon} \cdot \left\| g(z_t) \right\| \cdot \left\| g(z_t) - g(z_t+\epsilon\cdot e_d) \right\| && (42)\\
&\leq \sum_{d=1}^{D} \lim_{\epsilon\to 0} \frac{M}{\epsilon} \cdot \left\| g(z_t+\epsilon\cdot e_d) - g(z_t) \right\| + \frac{M}{\epsilon} \cdot \left\| g(z_t) - g(z_t+\epsilon\cdot e_d) \right\| && (43)\\
&\leq \sum_{d=1}^{D} \lim_{\epsilon\to 0} \frac{2\cdot M}{\epsilon} \cdot C_T \cdot \left\| \epsilon\cdot e_d \right\| && (44)\\
&= 2\cdot D\cdot M\cdot C_T \cdot \|e_d\| && (45)
\end{aligned}
$$

where, in Eq. 40, we use the trick from Esmayli (2017). Putting all of this together, we have that $\|h(z_t)\|$ is bounded by some constant, for which we can always find a new constant C_T to further bound it: $\|h(z_t)\|^2 \leq C_T \cdot (1 + \|z_t\|^2)$.

Proof that h satisfies (ii) from Theorem 3.2. Here, we prove that for all $T > 0$, $z_t, z_t' \in K$, and $t \in [0, T]$, $\|h(z_t) - h(z_t')\| \leq C_T \cdot \|z_t - z_t'\|$.

We begin as follows:

$$
\begin{aligned}
\|h(z_t) - h(z_t')\| \leq \frac{1}{2} \cdot \underbrace{\left\|\operatorname{diag}\left(\nabla_{z_t}[g(z_t)^2]\right) - \operatorname{diag}\left(\nabla_{z_t'}[g(z_t')^2]\right)\right\|}_{③} \\
+ \frac{1}{2} \cdot \underbrace{\left\| g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t) - g(z_t')^2 \odot \nabla_{z_t'} \log \tilde{p}(z_t') \right\|}_{④}.
\end{aligned} \tag{46}
$$

Using the trick by Esmayli (2017) again, we bound ③ as follows:

$$
\begin{aligned}
③ &= \left\|\operatorname{diag}\left(\nabla_{z_t'}[g(z_t')^2]\right) - \operatorname{diag}\left(\nabla_{z_t}[g(z_t)^2]\right)\right\| && (47)\\
&= \left\|g(z_t') \odot \operatorname{diag}\left(\nabla_{z_t'} g(z_t')\right) - g(z_t) \odot \operatorname{diag}\left(\nabla_{z_t} g(z_t)\right)\right\| && (48)\\
&= \left\|(g(z_t') - g(z_t)) \odot \operatorname{diag}\left(\nabla_{z_t'} g(z_t')\right) - g(z_t) \odot \left(\operatorname{diag}\left(\nabla_{z_t} g(z_t)\right) - \operatorname{diag}\left(\nabla_{z_t'} g(z_t')\right)\right)\right\| && (49)\\
&\leq \left\|(g(z_t') - g(z_t)) \odot \operatorname{diag}\left(\nabla_{z_t'} g(z_t')\right)\right\| + \left\|g(z_t) \odot \left(\operatorname{diag}\left(\nabla_{z_t} g(z_t)\right) - \operatorname{diag}\left(\nabla_{z_t'} g(z_t')\right)\right)\right\| && (50)\\
&\leq \|g(z_t') - g(z_t)\| \cdot \left\|\operatorname{diag}\left(\nabla_{z_t'} g(z_t')\right)\right\| + \|g(z_t)\| \cdot \left\|\operatorname{diag}\left(\nabla_{z_t} g(z_t)\right) - \operatorname{diag}\left(\nabla_{z_t'} g(z_t')\right)\right\| && (51)\\
&\leq C_T \cdot \|z_t - z_t'\| \cdot \underbrace{\left\|\operatorname{diag}\left(\nabla_{z_t'} g(z_t')\right)\right\|}_{\text{bounded by const. via (ii)}} + M \cdot \underbrace{\left\|\operatorname{diag}\left(\nabla_{z_t} g(z_t)\right) - \operatorname{diag}\left(\nabla_{z_t'} g(z_t')\right)\right\|}_{\leq C_T \cdot \|z_t - z_t'\| \text{ via (ii)}} && (52)
\end{aligned}
$$

We similarly bound ④ as follows:

$$
\begin{aligned}
④ &= \left\|\operatorname{diag}\left(\nabla_{z_t}[g(z_t)^2]\right) - \operatorname{diag}\left(\nabla_{z_t'}[g(z_t')^2]\right)\right\| && (53)\\
&= \left\|(g(z_t)^2 - g(z_t')^2) \odot \nabla_{z_t} \log \tilde{p}(z_t) - g(z_t')^2 \odot \left(\nabla_{z_t'} \log \tilde{p}(z_t') - \nabla_{z_t} \log \tilde{p}(z_t)\right)\right\| && (54)\\
&\leq \left\|(g(z_t)^2 - g(z_t')^2) \odot \nabla_{z_t} \log \tilde{p}(z_t)\right\| + \left\|g(z_t')^2 \odot \left(\nabla_{z_t'} \log \tilde{p}(z_t') - \nabla_{z_t} \log \tilde{p}(z_t)\right)\right\| && (55)\\
&\leq \underbrace{\left\|g(z_t)^2 - g(z_t')^2\right\|}_{⑤} \cdot \underbrace{\|\nabla_{z_t} \log \tilde{p}(z_t)\|}_{\text{bounded by const. via (iii)}} + \underbrace{\left\|g(z_t')^2\right\|}_{\leq M^2} \cdot \underbrace{\left\|\nabla_{z_t'} \log \tilde{p}(z_t') - \nabla_{z_t} \log \tilde{p}(z_t)\right\|}_{\leq C_T \cdot \|z_t - z_t'\| \text{ via (iii)}} && (56)
\end{aligned}
$$

Finally, This leaves us to bound terms ⑤ by a function of the form, $C_T \cdot \|z_t - z_t'\|$:

$$
\begin{aligned}
⑤ &= \left\|g(z_t)^2 - g(z_t')^2\right\| && (57)\\
&= \|(g(z_t) - g(z_t')) \odot g(z_t) - g(z_t') \odot (g(z_t') - g(z_t))\| && (58)\\
&\leq \|(g(z_t) - g(z_t')) \odot g(z_t)\| + \|g(z_t') \odot (g(z_t') - g(z_t))\| && (59)\\
&\leq \|g(z_t) - g(z_t')\| \cdot \|g(z_t)\| + \|g(z_t')\| \cdot \|g(z_t') - g(z_t)\| && (60)\\
&\leq 2 \cdot M \cdot C_T \cdot \|z_t - z_t'\| && (61)
\end{aligned}
$$

Proof that h satisfies (iii) from Theorem 3.2. Here, we prove that for each $z_t \in K$, $h(z_t)$, defined for $t \geq 0$, is continuous.

Since continuous functions are closed under all operations used to define $h(z_t)$, and since $h(z_t)$ is defined in terms of other continuous functions, it is also continuous.

Proof that h satisfies (a) from Theorem 3.2. Here, we prove that, for all $s \in [1, \ldots, S]$ and $z_t \in K$ such that when $\langle z_t - u_s, v_s \rangle = 0$, we have $\langle h(z_t), v_s \rangle \geq 0$. We start as follows:

$$
\begin{aligned}
\langle h(z_t), v_s \rangle &= \frac{1}{2} \cdot \langle \operatorname{diag}\left(\nabla_{z_t}[g(z_t)^2]\right), v_s \rangle + \frac{1}{2} \cdot \langle g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t), v_s \rangle && (62)\\
&= \frac{1}{2} \cdot \langle \operatorname{diag}\left(\nabla_{z_t}[g(z_t)^2]\right), v_s \rangle + \frac{1}{2} \cdot v_s^\intercal \cdot \left(g(z_t)^2 \odot \nabla_{z_t} \log \tilde{p}(z_t)\right) && (63)\\
&= \frac{1}{2} \cdot \langle \operatorname{diag}\left(\nabla_{z_t}[g(z_t)^2]\right), v_s \rangle + \frac{1}{2} \cdot v_s^\intercal \cdot \left(\sum_{d=1}^{D_z} (g(z_t) \odot e_d) \cdot g^d(z_t) \cdot \nabla_{z_t^d} \log \tilde{p}(z_t)\right) && (64)\\
&= \frac{1}{2} \cdot \langle \operatorname{diag}\left(\nabla_{z_t}[g(z_t)^2]\right), v_s \rangle + \frac{1}{2} \cdot \left(\sum_{d=1}^{D_z} v_s^\intercal \cdot (g(z_t) \odot e_d) \cdot g^d(z_t) \cdot \nabla_{z_t^d} \log \tilde{p}(z_t)\right) && (65)\\
&= \frac{1}{2} \cdot \langle \operatorname{diag}\left(\nabla_{z_t}[g(z_t)^2]\right), v_s \rangle \quad (\text{since } \langle g(z_t) \odot e_d, v_s \rangle = 0) && (66)\\
&\geq 0 && (67)
\end{aligned}
$$

We arrive at the last line because (1) when $\langle z_t - u_s, v_s \rangle = 0$, z_t is on the boundary of K, and (2) in the interior of K, $g(z_t)$ is non-negative. As such, $\operatorname{diag}\left(\nabla_{z_t}[g(z_t)^2]\right)$ does not point towards the exterior of K, giving us that $\langle h(z_t), v_s \rangle \geq 0$.

□

C. Proof of Theorem 4.1

Proof. To prove Theorem 4.1, we will show that $h(t, z_t)$ and $g(t, z_t)$, defined in Eq. 5, satisfy (i)-(iii) and (a)-(b) in Theorem 3.2.

Proof that WSP satisfies (i) from Theorem 3.2. Here, we prove that for each $T > 0$, there exists $C_T > 0$ such that for all $z_t \in K$ and $t \in [0, T]$, $\|h(t, z_t)\|^2 + \|g(t, z_t)\|^2 \leq C_T \cdot (1 + \|z_t\|^2)$.

First, we show that $w(z_t)$, defined in Eq. 4, lies in $[0, 1]$. Since $\alpha > 0$, and for any $z_t \in K$, $d(u_s, v_s, z_t) \geq 0$ (since distances are non-negative), we have,

$$0 \leq \tanh\left(\alpha \cdot d(u_s, v_s, z_t)\right) \leq 1. \tag{68}$$

Next, since $\beta > 0$, we know that:

$$0 \leq \beta \cdot \prod_s \underbrace{\frac{e^{-d(u_s, v_s, z_t)}}{\sum_{s'} e^{-d(u_{s'}, v_{s'}, z_t)}}}_{\in [0,1]} \cdot \underbrace{\tanh\left(\alpha \cdot d(u_s, v_s, z_t)\right)}_{\in [0,1]} \tag{69}$$

This then gives us,

$$0 \leq \underbrace{\tanh\left(\beta \cdot \prod_s \frac{e^{-d(u_s, v_s, z_t)}}{\sum_{s'} e^{-d(u_{s'}, v_{s'}, z_t)}} \cdot \tanh\left(\alpha \cdot d(u_s, v_s, z_t)\right)\right)}_{\geq 0} \leq 1, \tag{70}$$

thereby showing that $w(z_t) \in [0, 1]$. Using this, we go on to show that h and g satisfy condition (i) from Theorem 3.2.

$$\begin{aligned}
\|h(t, z_t)\| &= \|\text{WSP}(\tilde{h}, c_h, t, z_t)\| && (71)\\
&= \|w(z_t) \cdot \tilde{h}(t, z_t) + (1 - w(z_t)) \cdot c_h(z_t)\| && (72)\\
&\leq \|w(z_t) \cdot \tilde{h}(t, z_t)\| + \|(1 - w(z_t)) \cdot c_h(z_t)\| && (73)\\
&\leq \|1 \cdot \tilde{h}(t, z_t)\| + \|(1 - 0) \cdot c_h(z_t)\| && (74)\\
&= \|\tilde{h}(t, z_t)\| + \|c_h(z_t)\| && (75)\\
&= \|\tilde{h}(t, z_t)\| + \left\|\gamma \cdot \frac{z^* - z_t}{\|z^* - z_t\| + \epsilon}\right\| && (76)\\
&= \|\tilde{h}(t, z_t)\| + \gamma \cdot \underbrace{\left\|\frac{z^* - z_t}{\|z^* - z_t\| + \epsilon}\right\|}_{<1} && (77)\\
&\leq \|\tilde{h}(t, z_t)\| + \gamma && (78)\\
&\leq \sqrt{C'_T \cdot (1 + \|z_t\|^2)} + \gamma && (79)
\end{aligned}$$

Thus,

$$\begin{aligned}
\|h(t, z_t)\|^2 &\leq \left(\sqrt{C'_T \cdot (1 + \|z_t\|^2)} + \gamma\right)^2 && (80)\\
&= C'_T \cdot (1 + \|z_t\|^2) + \gamma^2 + 2 \cdot \gamma \cdot \sqrt{C'_T \cdot (1 + \|z_t\|^2)} && (81)\\
&\leq C_T \cdot (1 + \|z_t\|^2) && (82)
\end{aligned}$$

for some $C_T > 0$.

Similarly for $\|g(t,z_t)\|^2$, we have

$$\begin{aligned}
\|g(t,z_t)\| &= \|\mathrm{WSP}(\tilde{g}, c_g, t, z_t)\| && (83)\\
&= \|w(z_t)\cdot\tilde{g}(t,z_t) + (1-w(z_t))\cdot c_g(z_t)\| && (84)\\
&\leq \|w(z_t)\cdot\tilde{g}(t,z_t)\| + \|(1-w(z_t))\cdot c_g(z_t)\| && (85)\\
&\leq \|1\cdot\tilde{g}(t,z_t)\| + \|(1-0)\cdot c_g(z_t)\| && (86)\\
&= \|\tilde{g}(t,z_t)\| + \|c_g(z_t)\| && (87)\\
&= \|\tilde{g}(t,z_t)\| + \|0\| && (88)\\
&= \|\tilde{g}(t,z_t)\| && (89)\\
&\leq \sqrt{C_T\cdot(1+\|z_t\|^2)} && (90)
\end{aligned}$$

Thus, $\|g(t,z_t)\|^2 \leq C_T\cdot(1+\|z_t\|^2)$.

Proof that WSP satisfies (ii) from Theorem 3.2. We now prove that for all $T>0$, $z_t, z_t' \in K$, and $t\in[0,T]$, $\|h(t,z_t)-h(t,z_t')\| + \|g(t,z_t)-g(t,z_t')\| \leq C_T\cdot\|z_t - z_t'\|$.

We do this as follows:

$$\begin{aligned}
\|h(t,z_t)-h(t,z_t')\| &= \|\mathrm{WSP}(\tilde{h}, c_h, t, z_t) - \mathrm{WSP}(\tilde{h}, c_h, t, z_t')\| && (91)\\
&= \left\|\Big(w(z_t)\cdot\tilde{h}(t,z_t) + (1-w(z_t))\cdot c_h(z_t)\Big) - \Big(w(z_t')\cdot\tilde{h}(t,z_t') + (1-w(z_t'))\cdot c_h(z_t')\Big)\right\| && (92)\\
&= \left\|\Big(w(z_t)\cdot\tilde{h}(t,z_t) - w(z_t')\cdot\tilde{h}(t,z_t')\Big) + \big((1-w(z_t))\cdot c_h(z_t) - (1-w(z_t'))\cdot c_h(z_t')\big)\right\| && (93)\\
&\leq \|w(z_t)\cdot\tilde{h}(t,z_t) - w(z_t')\cdot\tilde{h}(t,z_t')\| + \|(1-w(z_t))\cdot c_h(z_t) - (1-w(z_t'))\cdot c_h(z_t')\| && (94)
\end{aligned}$$

Using the trick by Esmayli (2017), we have:

$$\begin{aligned}
\|h(t,z_t)-h(t,z_t')\| &\leq \|w(z_t)-w(z_t')\|\cdot\|\tilde{h}(t,z_t)\| + \|w(z_t')\|\cdot\|\tilde{h}(t,z_t)-\tilde{h}(t,z_t')\| && (95)\\
&\quad + \|(1-w(z_t))-(1-w(z_t'))\|\cdot\|c(z_t)\| + \|1-w(z_t')\|\cdot\|c(z_t)-c(z_t')\|\\
&\leq \|w(z_t)-w(z_t')\|\cdot\|\tilde{h}(t,z_t)\| + 1\cdot\|\tilde{h}(t,z_t)-\tilde{h}(t,z_t')\| && (96)\\
&\quad + \|w(z_t)-w(z_t'))\|\cdot\|c(z_t)\| + 1\cdot\|c(z_t)-c(z_t')\|\\
&= \|w(z_t)-w(z_t')\|\cdot\underbrace{(\|\tilde{h}(t,z_t)\| + \|c(z_t)\|)}_{\text{bounded by const.}} + \underbrace{\|\tilde{h}(t,z_t)-\tilde{h}(t,z_t')\|}_{<C_T'\cdot\|z_t-z_t'\|} + \|c(z_t)-c(z_t')\| && (97)
\end{aligned}$$

Since K is compact and $\tilde{h}(z_t)$ is continuous and linearly bounded (i.e. $\|\tilde{h}(z_t)\|^2 \leq C_T(1+\|z_t\|^2)$), we know that $\|\tilde{h}(t,z_t)\|$ is bounded above by a constant. Similarly, $c(z_t)$ is continuous and bounded,

$$\|c(z_t)\|^2 = \left\|\frac{z^*-z_t'}{\|z^*-z_t'\|+\epsilon}\right\|^2 \leq \left\|\frac{z^*-z_t'}{\epsilon}\right\|^2 = \epsilon^{-2}\cdot\|z^*-z_t'\|^2, \tag{98}$$

so $\|c(z_t)\|$ is bounded above by a constant. This leaves us to show that $w(z_t)$ and $c(z_t)$ are Lipschitz. This is true since both functions are comprised of either composition of Lipschitz functions, or of multiplications of bounded Lipschitz functions, and both of these operations are closed under Lipschitz continuity.

Similarly for $\|g(t,z_t)-g(t,z_t')\|$, we have,

$$\begin{aligned}
\|g(t,z_t)-g(t,z_t')\| &= \|\mathrm{WSP}(\tilde{g}, c_h, t, z_t) - \mathrm{WSP}(\tilde{g}, c_h, t, z_t')\| && (99)\\
&= \|(w(z_t)\cdot\tilde{g}(t,z_t) + (1-w(z_t))\cdot c_g(z_t)) - (w(z_t')\cdot\tilde{g}(t,z_t') + (1-w(z_t'))\cdot c_g(z_t'))\| && (100)\\
&= \|w(z_t)\cdot\tilde{g}(t,z_t) - w(z_t')\cdot\tilde{g}(t,z_t')\| && (101)
\end{aligned}$$

Since $w(z_t)\cdot\tilde{g}(t,z_t)$ is the product of a bounded Lipschitz function and a Lipschitz function, we know that $g(t,z_t)$ is also Lipschitz.

Proof that WSP satisfies (iii) from Theorem 3.2. Here, we prove that for each $z_t \in K$, $h(t, z_t)$ and $g(t, z_t)$ are continuous.

Since all functions involved are continuous and continuity is closed under addition, subtraction, multiplication and composition, $h(t, z_t)$ and $g(t, z_t)$, defined for $t \geq 0$, are continuous for each $z_t \in K$

Proof that WSP satisfies (a) from Theorem 3.2. Here we prove that for all $s \in [1, \ldots, S]$ and $z_t \in K$ such that when $\langle z_t - u_s, v_s \rangle = 0$, we have $\langle h(t, z_t), v_s \rangle \geq 0$.

First, when $\langle z_t - u_s, v_s \rangle = 0$,

$$d(u_s, v_s, z_t) = \frac{\langle z_t - u_s, v_s \rangle}{\|v_s\|} = \frac{0}{\|v_s\|} = 0. \tag{102}$$

This means that,

$$w(z_t) = \tanh\left(\beta \cdot \prod_s \frac{e^{-d(u_s, v_s, z_t)}}{\sum_{s'} e^{-d(u_{s'}, v_{s'}, z_t)}} \cdot \tanh\left(\alpha \cdot d(u_s, v_s, z_t)\right)\right) = 0. \tag{103}$$

Plugging this into $\langle h(t, z_t), v_s \rangle$, we get:

$$\begin{aligned}
\langle h(t, z_t), v_s \rangle &= \langle \mathrm{WSP}(\tilde{h}, c_h, t, z_t), v_s \rangle && (104)\\
&= \langle w(z_t) \cdot \tilde{h}(t, z_t) + (1 - w(z_t)) \cdot c_h(z_t), v_s \rangle && (105)\\
&= \langle 0 \cdot \tilde{h}(t, z_t) + (1 - 0) \cdot c_h(z_t), v_s \rangle && (106)\\
&= \langle c_h(z_t), v_s \rangle && (107)\\
&= \left\langle \gamma \cdot \frac{z^* - z_t}{\|z^* - z_t\| + \epsilon}, v_s \right\rangle && (108)\\
&= \underbrace{\frac{\gamma}{\|z^* - z_t\| + \epsilon}}_{>0} \cdot \langle z^* - z_t, v_s \rangle && (109)
\end{aligned}$$

Because polyhedra are convex, and $z^* \in K, z_t \in K$, we know that $z^* - z_t$ points to the interior of K; thus, $\langle z^* - z_t, v_s \rangle \geq 0$, completing the proof.

Proof that WSP satisfies (b) from Theorem 3.2. Here we prove that for all $s \in [1, \ldots, S]$ and $z_t \in K$ such that when $\langle z_t - u_s, v_s \rangle = 0$, we have $\langle g(t, z_t) \odot e_d, v_s \rangle = 0$ for $t \geq 0$ and $d \in [1, \ldots, D_z]$. We do this as follows:

$$\begin{aligned}
\langle g(t, z_t) \odot e_d, v_s \rangle &= \langle \mathrm{WSP}(\tilde{g}, c_g, t, z_t) \odot e_d, v_s \rangle && (110)\\
&= \langle (w(z_t) \cdot \tilde{g}(t, z_t) + (1 - w(z_t)) \cdot c_g(z_t)) \odot e_d, v_s \rangle && (111)\\
&= \langle (0 \cdot \tilde{g}(t, z_t) + (1 - 0) \cdot c_g(z_t)) \odot e_d, v_s \rangle && (112)\\
&= \langle c_g(z_t) \odot e_d, v_s \rangle && (113)\\
&= \langle 0 \odot e_d, v_s \rangle && (114)\\
&= \langle 0, v_s \rangle && (115)\\
&= 0 && (116)
\end{aligned}$$

□

D. Discussion of Assumptions

The assumptions in Theorems 3.2 and 3.3 are easily satisfied when h, g, and $\log \tilde{p}(z_t)$ are parameterized by NNs.

Lipschitz continuity with respect to inputs. Lipschitz continuous functions are closed under composition, making a large class of neural networks Lipschitz continuous by construction. Additionally, there exist many easy and empirically

effective methods for explicitly obtaining Lipschitz continuity, for example via weight normalization (e.g. Miyato et al. (2018)), regularization (e.g. Liu et al. (2022)), and architecture design (e.g. Anil et al. (2019)). Altogether, this allows us to conveniently satisfy the Lipschitz continuity assumptions for h, g, and $\log \tilde{p}(z_t)$—(ii) in Theorem 3.2, and (ii) and (iii) in Theorem 3.3.

Next, Hurault et al. (2022) (Appendix B) proved that any composition of bounded, Lipschitz functions has Lipschitz gradients with respect to the inputs, thereby making a large class of neural network satisfy the second half of (ii) from Theorem 3.3. This is property is known as "Lipschitz smoothness," and can also be obtained explicitly, for example via mixup regularization (Gyawali et al., 2020).

Linearly bounded NNs. We parameterize all neural networks here with a composition of continuous functions, thereby making them continuous. And since continuous functions on compact spaces are bounded, we easily satisfy (i) from Theorems 3.2 and 3.3.

Differentiability and continuity of partials of NNs. All neural networks here use continuously differentiable activation functions, so they are continuously differentiable with continuous partials.

E. Experimental Setup

Dynamics. As we argue in Section 1, initialization plays a crucial role in the success of expressive SDE-based models. This is because many data sets (e.g. images) lie on compact Euclidean subspaces. In early-stage training, SDE trajectories often leave the region, requiring a large number of gradient steps just to return to it (while not necessarily fitting the data well), causing optimization to get stuck in poor local optima. In late-stage training, small perturbations to the dynamics may, again, yield trajectories that lie outside the region. As such, to empirically compare the inductive bias of WSP (Eq. 5) against baselines (Eqs. 1–3), we solve SDEs given by NNs h and g with randomly sampled weights. We define the viable region, $K = [0, 1]$, to be a compact rectangle, and specifically choose to set $z_0 = 0.99$ near the boundary to stress-test the chain-rule based SDEs in Eqs. 2 and 3 to show that once close to the boundary, they will struggle to return to the interior of K. While simple, *these preliminary experiments already show WSP boasts a stark improvement in inductive bias in comparison to baselines.*

Architecture. In all experiments presented here, we used 3-layer NNs with 64 hidden units and CELU activation (Barron, 2017). We repeated these experiments with 2-layer and 4-layer NNs and observed the *exact same behavior*, so we have omitted them for brevity. We also repeated these experiments with other continuous activation functions—GeLU (Hendrycks & Gimpel, 2016), ELU (Clevert et al., 2015), SELU (Klambauer et al., 2017), and SiLU (Elfwing et al., 2018)—and we observed the *exact same type of behavior*, so we have omitted them for brevity.

Setup. For each SDE in Eqs. 1–3 and 5, we randomly drew the weights using the normal Glorot initialization (Glorot & Bengio, 2010). In each plot, we repeated this initialization 5 times, drawing 3 samples for each initialization.

Differential Equation Solver. In Fig. 2, we used the Ito-Milstein SDE solver. In Fig. 3, we used the Dormand-Prince 8/7 ODE solver. In all experiments, we simulated the dynamics for $t \in [0, 5]$ with a step size of 0.001. We purposefully chose a small step size to ensure the faithfulness of the SDE solutions to the dynamics.

Pathwise Expansion. We used a truncation of $R = 40$ terms in the pathwise expansion (Eq. 117) for the experiment in Fig. 3. We repeated the experiments with $R = 20$, 100, and 200 and observed the *exact same behavior*, so we have omitted them for brevity.

Software. All experiments were conducted in Jax (Bradbury et al., 2018) with NumPyro (Phan et al., 2019), Diffrax (Kidger, 2021) and Chex.

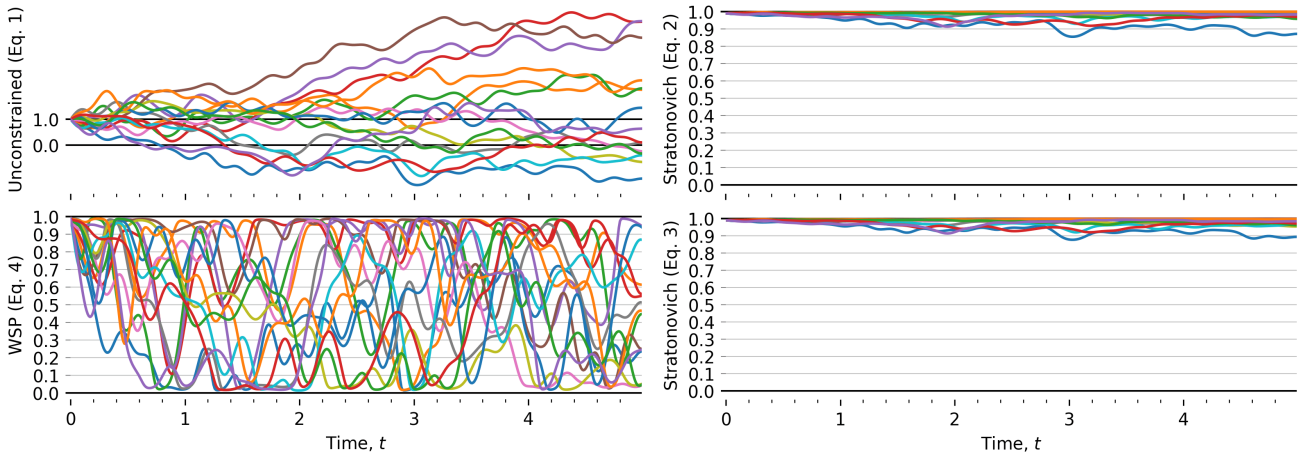

Figure 3. **WSP exhibits better inductive bias than baselines given smooth, pathwise expansion of Brownian motion.** Top left: Stratonovich-SDE with NN quickly leaves $K = [0, 1]$. Top & bottom right: Stratonovich-SDE transformed via sigmoid sticks to the boundary. Bottom left: Stratonovich-SDE with WSP successfully remains in K. Note: for Eqs. 2 and 3, we used the Stratonovich chain-rule instead of Ito's lemma.

F. Results

F.1. Inductive Bias with Smooth, Pathwise Expansion of Brownian Motion

Since SDE solvers are slow and unstable, prior work focused on finding mechanisms to use ODE solvers instead. ODE solvers are known to be more numerically stable, accurate, and well-behaved when used with adaptive step-sizes. Prior work has used several strategies to accomplish this. For example, prior work approximates the first two moments of the time-marginal using the Fokker-Planck-Kolmogorov (FPK) equation, which can then be solved using an ODE solver (this is known as the "Gaussian Assumed Approximation", Särkkä & Solin (2019)). In diffusion models, prior work derived an FPK-based fast, numerically stable process that samples from the same distribution as the SDE using an ODE solver (the "probability flow ODE") (Song et al., 2021). Finally, prior work (e.g. Ghosh et al. (2022)) replaces Brownian motion with the Karhunen–Loeve Expansion (Särkkä & Solin, 2019), which takes the following form:

$$\begin{aligned} \xi^r &\sim \mathcal{N}(0, 1), \\ d\widehat{B}_t | \xi^1, \dots, \xi^R &= \sum_{r=1}^{R} \sqrt{\frac{2}{T}} \cos\left(\frac{(2 \cdot r - 1) \cdot \pi \cdot t}{2T}\right) \cdot \xi^r \cdot dt, \end{aligned} \tag{117}$$

where T is the end-time of the process. This expansion replaces Brownian motion, dB_t, with a randomly weighted sum of ODEs, allowing us to use an ODE solver. As $R \to \infty$, the distribution of $d\widehat{B}_t$ converges to that of dB_t, and overall differential equation converges to the Stratonovich SDE (Wong & Zakai, 1965).

In Fig. 3, we empirically demonstrate that WSP boasts the same favorable inductive bias in comparison to baselines, even under this pathwise expansion.

F.2. Stationary SDEs

In Fig. 4, we show that, for any neural diffusion with WSP (Eq. 5) with randomly generated weights, we can always construct a drift, given by Theorem 3.3, with dynamics that are viable in $K = [0, 1]$ and induce the target time-marginal. Moreover, like their non-stationary counterparts, these stationary dynamics overcome the shortcomings of the baselines dynamics in Eqs. 1–3.

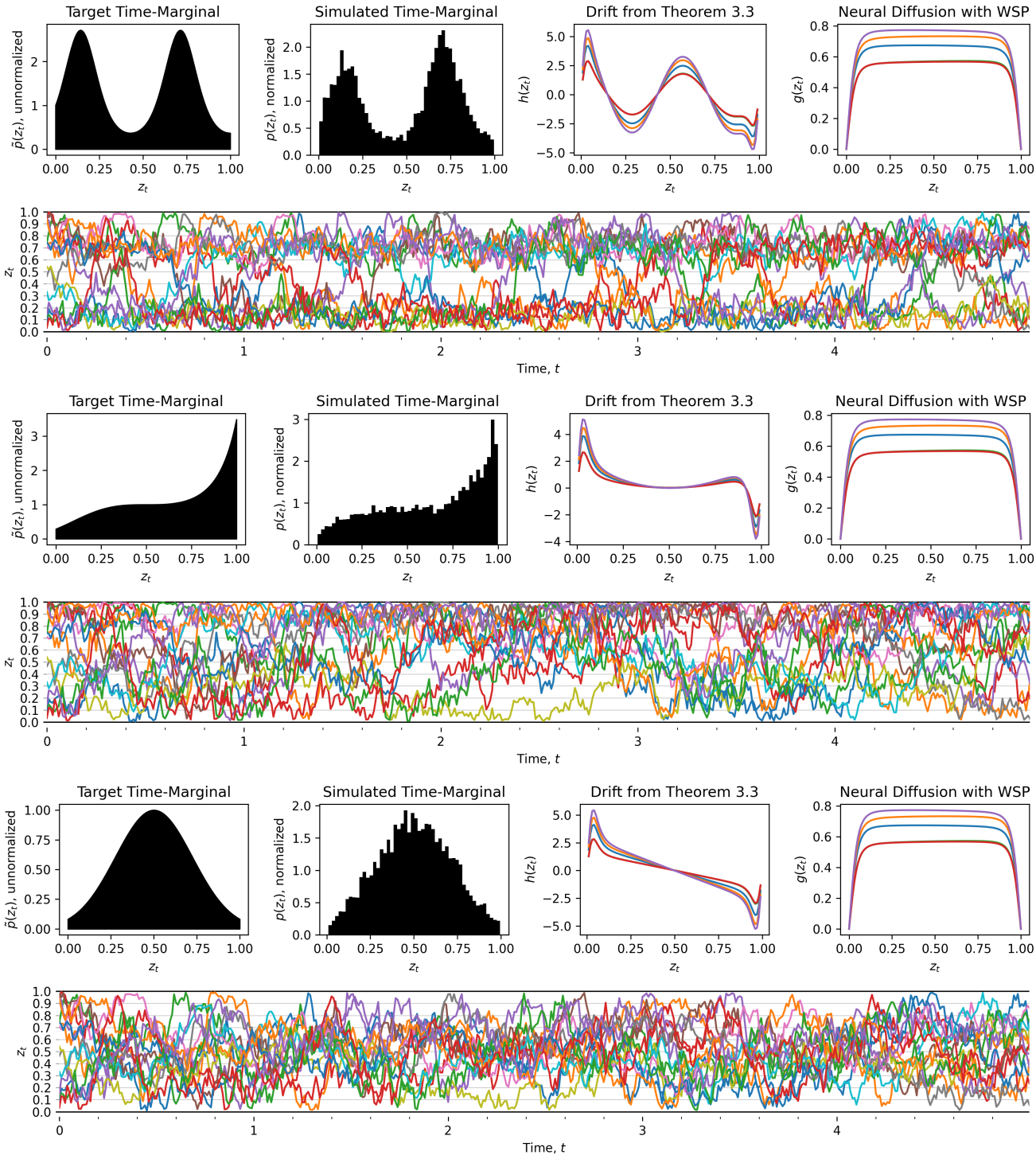

Figure 4. **Stationary SDE exhibits better inductive bias than baselines.** Given a target time-marginal, given any diffusion with WSP, we can always derive a corresponding drift via Theorem 3.3 that is viable in K and has the target stationary distribution. Like the non-stationary dynamics, these dynamics overcome the shortcomings of the baselines dynamics in Eqs. 1–3. Here, our diffusion is a NN with randomly initialized weights, with each color corresponding to a different seed. Note: the target time-marginal is *not* normalized.