Supplementary to: Diffusion Models for Constrained Domains

Introduction

In this supplementary, we first recall some key concepts from Riemannian geometry in Appendix A. In Appendix B we remind the expression of the Brownian motion in local coordinates. Details about the geodesic Brownian motion are given in Appendix C. Background on the Skorokhod problem is given in Appendix D. In Appendix E we derive the implicit score matching loss. We give details about the likelihood evaluation in Appendix F. Then in Appendix G we prove the time-reversal formula of reflected Brownian motion. In Appendix I we give some background on the conformational modelling of proteins backbone for the experiment in Section 5.3. In Appendix H we detail the geometrical constraint arising from the configurational robotics arm modelling experiment from Section 5.2. Additional results, training and miscalleneous experimental details are reported in Appendix J.

A Manifold concepts

For readers unfamiliar with Riemannian geometry here we give a brief overview of some of the key concepts. This is not a technical introduction, but a conceptual one for the understanding of terms. For a technical introduction, we refer readers to Lee (2013). A Riemannian manifold is a tuple $(\mathcal{M}, \mathfrak{g})$ with $\mathcal{M}$ a smooth manifold and $\mathfrak{g}$ a metric which defines an inner product on tangent spaces.

A *smooth manifold* is a topological space which locally can be described by Euclidean space. It is characterised by a family of *charts* $\{U \subset \mathcal{M}, \phi : U \to \mathbb{R}^d\}$, homeomorphic mappings between subsets of the manifold and Euclidean space. The collection of charts must cover the whole manifold. For the manifolds to be smooth the charts must be smooth, and the transition between charts where their domains overlap must also be smooth.

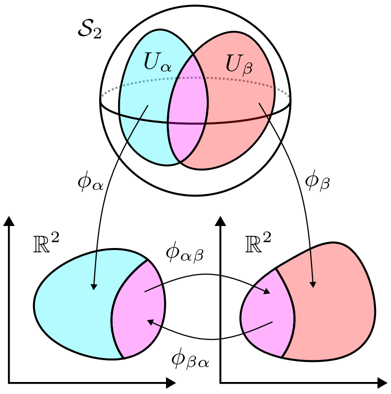


Figure 12: Example charts of the 2D manifold $\mathcal{S}^2$.

The *metric* on a Riemannian manifold gives the manifold a notion of distance and curvature. The same underlying smooth manifold with different metrics can look wildly different. The metric is defined as a smooth choice of positive definite inner product on each of the tangent spaces of the manifold. That is we have at every point a symmetric bilinear map

$$\mathfrak{g}(p) : \mathrm{T}_p\mathcal{M} \times \mathrm{T}_p\mathcal{M} \to \mathbb{R}$$

The tangent space of a point on a manifold is the generalisation of the notion of tangent planes and can be thought of as the space of derivatives of scalar functions on the manifold at that point. The collection of all tangent spaces is written as $\mathrm{T}\mathcal{M} = \bigcup_{p\in\mathcal{M}} \mathrm{T}_p\mathcal{M}$ and is called the tangent bundle. It is also manifold. Vector fields on manifolds are defined by making a choice of tangent vector at every point on the manifold in a smooth fashion. The space of vector fields is written at $\Gamma(\mathrm{T}\mathcal{M})$ and is more technically the space of *sections* of the tangent bundle.

One thing that the metric itself does not immediately define is how different tangent spaces at points on the manifold relate to one another. For this, we need additionally the concept of a *connection.* A connection is a map that takes two vector fields and produces a derivative of the first with respect to the second, that is a function $\nabla : \Gamma(\mathrm{T}\mathcal{M}) \times \Gamma(\mathrm{T}\mathcal{M}) \to \Gamma(\mathrm{T}\mathcal{M})$ and it typically written as $\nabla(X, Y) = \nabla_X Y$. Such a connection must for $X, Y, Z \in \Gamma(\mathrm{T}\mathcal{M})$ and smooth functions on the manifold $a, b : \mathcal{M} \to \mathbb{R}$ obey the following conditions:

(a) $\nabla_{aX+bY} Z = a\nabla_X Z + b\nabla_Y Z$,

(b) $\nabla_X(Y+Z) = \nabla_X Y + \nabla_X Z$,

(c) $\nabla_X(aY) = \partial_X aY + a\nabla_X Y$,

where $\partial_X aY$ is the regular *directional derivative* of aY in the direction X. These conditions ensure the connection is a well-defined derivative.

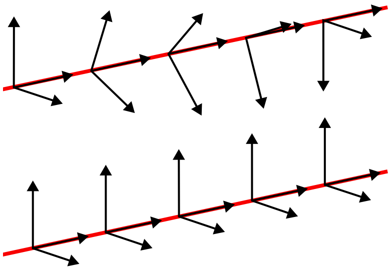


Figure 13: *Top:* Parallel transport of vectors along the red path with non-zero torsion. *Bottom:* Parallel transport of vectors along the red path with zero torsion. Both under the Euclidean metric.

On a given manifold, there are infinitely many connections. Fortunately, there is a natural choice, called the *Levi-Cevita* connection if we impose two additional conditions:

(a) $X \cdot (\mathfrak{g}(Y,Z)) = \mathfrak{g}(\nabla_X Y, Z) + \mathfrak{g}(Y, \nabla_X Z)$,

(b) $[X,Y] = \nabla_X Y - \nabla_Y X$,

where $[\cdot,\cdot]$ is the Lie bracket. The first condition ensures that the metric is preserved by the connection. That is to say, the *parallel transport* (to be defined shortly) using the connection leaves inner products unchanged on the manifold. The second ensures the connection is *torsion-free.* The change in tangent space along a geodesic (again to be defined shortly) can be described in two parts, the *curvature*, how the tangent space rotates perpendicular to the direction of travel, and the torsion, how the tangent space rotates around the axis of the direction of travel. The curvature of the connection is uniquely fixed by the other 4 conditions (the well-defined derivative and preservation of the metric). The torsion however is not fixed. By requiring it to be zero we ensure a unique connection. The requirement of zero torsion also has implications for ensuring integrability on the manifold.

With the metric and the Levi-Cevita connection in hand, we can define a number of key concepts.

We say that a vector field X is *parallel* to a curve $\gamma : (0,1) \to \mathcal{M}$ if

$$\nabla_{\gamma'} X = 0$$

where $\gamma' : (0,1) \to \mathrm{T}_{\gamma(t)}\mathcal{M}$ is the derivative of the path. For two points on the manifold $p, q \in \mathcal{M}$ and a curve between them, γ, $\gamma(0) = p$, $\gamma(1) = q$, for an initial vector $X_0 \in \mathrm{T}_p\mathcal{M}$ there is a unique vector field X that is parallel to γ such that $X(p) = X_0$. This induces a map between the tangent spaces at p and q

$$\tau_\gamma : \mathrm{T}_p\mathcal{M} \to \mathrm{T}_q\mathcal{M}$$

This map is referred as the *parallel transport* of tangent vectors between p and q, and this satisfies the condition that for $\boldsymbol{v}, \boldsymbol{u} \in \mathrm{T}_p\mathcal{M}$

$$\mathfrak{g}(p)(\boldsymbol{v}, \boldsymbol{u}) = \mathfrak{g}(q)(\tau_\gamma(\boldsymbol{v}), \tau_\gamma(\boldsymbol{u})).$$

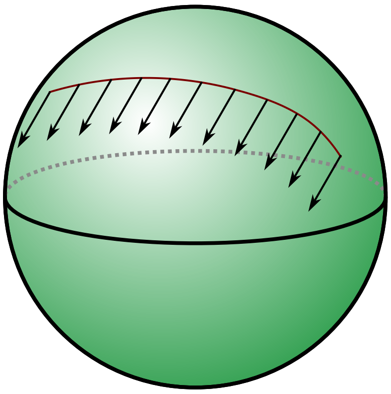


Figure 14: Parallel transport of vectors along a path on the sphere.

A *geodesic* on a manifold is the unique path on the manifold $\gamma : (0,1) \to \mathcal{M}$ such that $\nabla_{\gamma'}\gamma' = 0$. It is also the shortest path between two points on a manifold in the sense that

$$L(\gamma) = \int_0^1 \sqrt{\mathfrak{g}(\gamma(t))(\gamma'(t), \gamma(t))},$$

is minimal out of any path between the start and end of the geodesic. Geodesics give the notion of 'straight lines' on manifolds. We define the *exponential map* on a manifold as the mapping between an element of the tangent space at point p, $\boldsymbol{v} \in T_p\mathcal{M}$ and the endpoint of the unique geodesic γ with $\gamma(0) = p$ and $\gamma'(0) = \boldsymbol{v}$. In the tangent space of a manifold, we require the notion of a reflection in order to reflect geodesics off of boundary constraints. If at a point in the manifold with have $\boldsymbol{v} \in T_p\mathcal{M}$ and a unit vector normal to the constraint $\boldsymbol{n} \in T_p\mathcal{M}$, then the reflection of this vector is given by $\boldsymbol{v'} = \boldsymbol{v} - 2\mathfrak{g}(\boldsymbol{v}, \boldsymbol{n})\boldsymbol{n}$.

B Brownian motion in local coordinates

We consider a smooth function $f \in \mathrm{C}^\infty(\mathcal{M})$. The Laplace-Beltrami operator $\Delta_\mathcal{M}$ is given by $\Delta_\mathcal{M}(f) = \mathrm{div}(\mathrm{grad}(f)))$. In local coordinates we have

$$\mathrm{div}(X) = (\det(\mathfrak{g})^{-1/2}) \textstyle\sum_{i=1}^d \partial_i (\det(\mathfrak{g})^{1/2} X_i),$$

as well as

$$\mathrm{grad}(f) = \mathfrak{g}^{-1} \nabla f.$$

Therefore, the Laplace-Beltrami operator is given by

$$\Delta_\mathcal{M}(f) = \textstyle\sum_{i,j=1}^d \mathfrak{g}_{i,j}^{-1} \partial_{i,j} f + (\det(\mathfrak{g})^{-1/2}) \textstyle\sum_{i,j=1}^d \partial_i (\det(\mathfrak{g})^{1/2} \mathfrak{g}_{i,j}^{-1}) \partial_j f.$$

Therefore, in local coordinates the infinitesimal generator associated with the Laplace-Beltrami operator is given by

$$\mathcal{A}(f) = \textstyle\sum_{i,j=1}^d \mathfrak{g}_{i,j}^{-1} \partial_{i,j} f + \langle b^i, \nabla f \rangle,$$

with

$$b^i = (\det(\mathfrak{g})^{-1/2}) \textstyle\sum_{j=1}^d \partial_j (\det(\mathfrak{g})^{1/2} \mathfrak{g}_{i,j}^{-1}). \tag{11}$$

Therefore, the dual operator associated with $\mathcal{A}$ is given by

$$\mathcal{A}^\star(f) = \textstyle\sum_{i,j=1}^d \partial_{i,j} (\mathfrak{g}_{i,j}^{-1} f) - \textstyle\sum_{i=1}^d \partial_i (b^i f).$$

Note that by letting $f = \det(\mathfrak{g})^{1/2}$ we get that $\mathcal{A}^\star(f) = 0$ and therefore we recover that $p \propto \det(\mathfrak{g})$ is the invariant distribution of the Brownian motion.

Langevin dynamics on $\mathcal{M}$. We know that the Brownian motion targets $\det(\mathfrak{g})^{1/2}$. Therefore in order to correct and sample from the uniform distribution we consider the Langevin dynamics

$$\mathrm{d}\mathbf{X}_t = -\mathrm{grad} \log(\det(\mathfrak{g})^{1/2})(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t^\mathcal{M}.$$

Note that in the previous equation grad and $\mathbf{B}_t^\mathcal{M}$ are defined w.r.t. the metric of the manifold. In local coordinates we have

$$\mathrm{d}\mathbf{X}_t = \{b - \mathrm{grad} \log(\det(\mathfrak{g})^{1/2})\}(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathfrak{g}(\mathbf{X}_t)^{-1/2}\mathrm{d}\mathbf{B}_t. \tag{12}$$

where $b = \{b^i\}_{i=1}^d$ is given in equation 11. In addition, we have

$$\mathrm{grad} \log(\det(\mathfrak{g})^{1/2}) = \det(\mathfrak{g})^{-1/2} \mathfrak{g}^{-1} \nabla \det(\mathfrak{g}). \tag{13}$$

Using equation 11 we have

$$b^i = (\det(\mathfrak{g})^{-1/2}) \textstyle\sum_{j=1}^d \partial_j (\det(\mathfrak{g})^{1/2} \mathfrak{g}_{i,j}^{-1}) = \textstyle\sum_{j=1}^d \partial_j \mathfrak{g}_{i,j}^{-1} + \mathrm{grad} \log(\det(\mathfrak{g})^{1/2})_i$$

This can also be rewritten as

$$\mathrm{div}_\mathcal{M}(\mathfrak{g}^{-1}) = \mu + \mathrm{grad} \log(\det(\mathfrak{g})^{1/2}),$$

with

$$\mu_i = \textstyle\sum_{j=1}^d \partial_j \mathfrak{g}_{i,j}^{-1}.$$

Combining this result and equation 13 we get that equation 12 can be rewritten as

$$\mathrm{d}\mathbf{X}_t = \mu(\mathbf{X}_t) + \sqrt{2}\mathrm{d}\mathbf{B}_t.$$

Note that (up to a factor 2) this is the same SDE as the one considered in Lee & Vempala (2017).

C Geodesic Brownian Motion

In this section, we provide some details on the geodesics Brownian motion introduced in Section 3.1. In the rest of the section, we make the following assumption.

***A*1.** $\overline{\mathcal{M}} \subset \mathbb{R}^d$ is compact and $\mathfrak{g}^{-1} : \mathcal{M} \to \mathcal{S}_d^{++}$ can be $\mathrm{C}^\infty(\mathbb{R}^d, \mathbb{R}^{d\times d})$.

First, we start by showing that the process $(\mathbf{X}_t)_{t\geq 0}$ defined in equation 6 exists and that we have for any $t \geq 0$, $\mathbf{X}_t \in \mathcal{M}$. We recall that $\mathfrak{g} = \nabla^2\phi$ and $\lim_{x\to\partial\mathcal{M}} \Phi(x) = +\infty$.

Proposition C.1. *Assume* **A**1*. For any $x_0 \in \mathcal{M}$, there exists a unique strong solution to equation 6 denoted $(\mathbf{X}_t)_{t\geq 0}$. In addition, we have that for any $t \geq 0$, $\mathbf{X}_t \in \mathcal{M}$ almost surely. More precisely, we have $\mathbb{E}[\phi(\mathbf{X}_t)] \leq \phi(x_0) + t$.*

Proof. A unique strong solution $(\mathbf{X}_t)_{t\geq 0}$ of equation 6 with starting point $x_0 \in \mathcal{M}$ exists since the coefficients are smooth, see (Ikeda & Watanabe, 2014, Theorem 3.1, p.165). For any $A \geq 0$, we define $\tau_A = \inf\{t \geq 0 : \Phi(\mathbf{X}_t) \geq A\}$. Note that for any $t \in [0, \tau_A]$, $\Phi(\mathbf{X}_t) \in \mathcal{M}$. Using Itô formula, we have for any $t \geq 0$

$$\mathbb{E}[\Phi(\mathbf{X}_{t\wedge\tau_A})] = \Phi(x_0) + \mathbb{E}[\textstyle\int_0^{t\wedge\tau_A} \mathrm{Tr}(\mathfrak{g}^{-1}(\mathbf{X}_s)\nabla^2\Phi(\mathbf{X}_s))\mathrm{d}s] = \Phi(x_0) + \mathbb{E}[t \wedge \tau_A].$$

Using Fatou's lemma, and letting $A \to +\infty$, we conclude the proof. □

In the next result, we show that the uniform distribution is the *unique* invariant probability distribution for $(\mathbf{X}_t)_{t\geq 0}$ and that $(\mathbf{X}_t)_{t\geq 0}$ converges to this invariant distribution. We refer to (Meyn & Tweedie, 1993, Section 2, p.490) for a definition of irreducibility. We recall that the total variation of a finite (not necessarily non-negative) measure μ over $\mathbb{R}^d$ is given by $\|\mu\|_{\mathrm{TV}} = \sup\{\mu(\mathsf{A}) : \mathsf{A} \in \mathcal{B}(()\mathbb{R}^d)\}$.

Proposition C.2. *Assume* **A**1*. $(\mathbf{X}_t)_{t\geq 0}$ is π-irreducible, the uniform distribution over $\mathcal{M}$ is the only invariant probability distribution and $\lim_{t\to+\infty} \|\mathrm{P}_t - \pi\|_{\mathrm{TV}} = 0$, where P_t is the distribution of $\mathbf{X}_t$ for any $t \geq 0$ and π is the uniform distribution over $\mathcal{M}$.*

Proof. Since $x \mapsto \mathrm{div}(\mathfrak{g}^{-1})(x)$ and $x \mapsto \mathfrak{g}^{-1}$ are smooth and $\mathfrak{g}^{-1}(x)$ is positive definite for any $x \in \mathcal{M}$, we have that $(\mathbf{X}_t)_{t\geq 0}$ is π-irreducible, extending (Bhattacharya, 1978, Lemma 1.4) to $\mathcal{M}$ and using (Meyn & Tweedie, 1993, Proposition 2.1). In addition, $(\mathbf{X}_t)_{t\geq 0}$ is T-Feller using (Meyn & Tweedie, 1993, Proposition 3.3). Combining these results and the fact that $\mathcal{M}$ is bounde, we get that $(\mathbf{X}_t)_{t\geq 0}$ is positive Harris recurrent (Meyn & Tweedie, 1993, Theorem 3.2). The uniform distribution π is an invariant distribution for equation 6. Since $(\mathbf{X}_t)_{t\geq 0}$ is π-irreducible, we get that this invariant measure is unique. Hence, we conclude using (Meyn & Tweedie, 1993, Theorem 6.1). □

Note that the convergence result in total variation could be improved. In particular, quantitative geometric results could be derived. We finish this section, by applying results from the Malliavin calculus to show that for any $t > 0$, $\mathbf{X}_t$ admits a density w.r.t. the Lebesgue measure.

Proposition C.3. *Assume* **A**1*. Then, for any $t \geq 0$, $\mathbf{X}_t$ admits a smooth density p_t w.r.t. the Lebesgue measure.*

Proof. This is a direct consequence of (Nualart, 2006, Theorem 2.3.3). □

D Reflected Brownian Motion and Skorokhod problems

In this section, we provide the basic definitions and results to derive the time-reversal of the reflected Brownian motion in Appendix G. We follow closely the presentation of Lions & Sznitman (1984) and Burdzy et al. (2004). We first define the *Skorokhod problem* for deterministic problems. We consider $\mathcal{M}$ to be a smooth open bounded domain. We recall that the normal vector n is defined on $\partial\mathcal{M}$ and we set $n(x) = 0$ for any $x \notin \partial\mathcal{M}$.

Before giving the definition of the *Skorokhod problem*, we recall what is the space of functions of *bounded variations.*

Definition D.1. Let $a, b \in (-\infty, +\infty)$ and $f :\in \mathrm{C}([a,b], \mathbb{R})$. We define the *total variation* of f as

$$\mathrm{V}_{a,b}(f) = \sup\{\textstyle\sum_{i=0}^{n-1} \|f(x_{i+1}) - f(x_i)\| \ : \ (x_i)_{i=0}^{n-1}, \ a = x_0 \leq x_1 \leq \cdots \leq x_{n-1} \leq x_n = b, \ n \in \mathbb{N}\}.$$

f has bounded variations over $[a, b]$ if $\mathrm{V}_{a,b}(f) < +\infty$. Let $f \in \mathrm{C}([0, +\infty), \mathbb{R})$. f has bounded variations over $[0, +\infty)$ if for any $b > 0$, f has bounded variations over $[0, b]$.

The notion of bounded variation is a relaxation of the differentiability requirement. In particular, if $f \in \mathrm{C}^1([a,b], \mathbb{R})$, we have $\mathrm{V}_{a,b}(f) = \int_a^b \|f'(t)\|\mathrm{d}t$. In the definition of the *Skorokhod problem*, we will see that this relaxation is necessary, even in the deterministic setting.

For any function of bounded variation $f \in \mathrm{C}([a,b], \mathbb{R})$ on $[a, b]$, we define $|f| : \ [a, b] \to [0, +\infty)$ given for any $t \in [a, b]$ by $|f|_t = \mathrm{V}_{a,t}(f)$. Note that $|f|$ is non-decreasing and right-continuous. Therefore, we can define the measure $\mu_{|f|}$ on $[a, b]$, given for any $s, t \in [a, b]$ with $t \geq s$ by $\mu_{|f|}([s, t]) = |f|(t) - |f|(s)$. In particular, for any $\varphi : [a, b] \to \mathbb{R}_+$, we define

$$\textstyle\int_a^b \varphi(t)\mathrm{d}|f|_t = \int_a^b \varphi(t)\mathrm{d}\mu_{|f|}(t).$$

In addition, f can be decomposed in a difference of two non-decreasing processes right continuous processes g_1, g_2, where for any $t \in [a, b]$, $f(t) = g_1(t) - g_2(t)$, $g_1(t) = |f|_t$ and $g_2(t) = |f|_t - f(t)$. Hence, for every φ bounded on $[a, b]$, we can define

$$\textstyle\int_a^b \varphi(t)\mathrm{d}f(t) = \int_a^b \varphi(t)\mathrm{d}g_1(t) - \int_a^b \varphi(t)\mathrm{d}g_2(t).$$

Note that these definitions can be extended to the setting where $f : \ [a, b] \to \mathbb{R}^d$.

We begin with the following result, see Lions & Sznitman (1984).

Theorem D.2. *Let $(x_t)_{t\geq 0} \in \mathrm{C}([0, +\infty), \mathbb{R})$. Then, there exists a unique couple $(\bar{x}_t, k_t)_{t\geq 0}$ such that*

(a) $(k_t)_{t\geq 0}$ has bounded variation over $[0, +\infty)$.

(b) $(\bar{x}_t)_{t\geq 0} \in \mathrm{C}([0, +\infty), \overline{\mathcal{M}})$.

(c) For any $t \geq 0$, $x_t + k_t = \bar{x}_t$.

(d) For any $t \geq 0$, $|k|_t = \int_0^t \mathbf{1}_{\bar{x}_s \in \partial\mathcal{M}}(\bar{x}_s)\mathrm{d}|k|_s$ and $k_t = \int_0^t n(\bar{x}_s)\mathrm{d}|k|_s$.

Let us discuss Theorem D.2. First, Theorem D.2-(c) states the original (unconstrained) process $(x_t)_{t\geq 0}$ can be decomposed into a constrained version $(\bar{x}_t)_{t\geq 0}$ and a bounded variation process $(k_t)_{t\geq 0}$. The process $(|k|_t)_{t\geq 0}$ counts the number of times the constrained process $(\bar{x}_t)_{t\geq 0}$ hits the boundary. More formally, we have $|k|_t = \int_0^t \mathbf{1}_{x\in\partial\mathcal{M}}(\bar{x}_s)\mathrm{d}|k|_s$. When, we hit the boundary, we reflect the process. This condition is expressed in $k_t = \int_0^t n(\bar{x}_s)\mathrm{d}|k|_s$.

We now consider the extension to stochastic processes. We are given $(\mathbf{X}_t)_{t\geq 0}$ such that

$$\mathrm{d}\mathbf{X}_t = b(\mathbf{X}_t)\mathrm{d}t + \sigma(t)\mathrm{d}\mathbf{B}_t,$$

where $(\mathbf{B}_t)_{t\geq 0}$ is a d-dimensional Brownian motion. We also assume that b and σ are Lipschitz which implies the existence and strong uniqueness of $(\mathbf{X}_t)_{t\geq 0}$. We have the following result Lions & Sznitman (1984).

Theorem D.3. *There exists a unique process* $(\bar{\mathbf{X}}_t, \mathbf{k}_t)_{t \geq 0}$ *such that*

(a) $(\mathbf{k}_t)_{t\geq 0}$ *has bounded variation over* $[0, +\infty)$ *almost surely.*

(b) $(\bar{\mathbf{X}}_t)_{t\geq 0} \in \mathrm{C}([0, +\infty), \overline{\mathcal{M}})$.

(c) For any $t \geq 0$, $\bar{\mathbf{X}}_t = \bar{\mathbf{X}}_0 + \int_0^t b(\bar{\mathbf{X}}_s)\mathrm{d}s + \int_0^t \sigma(\bar{\mathbf{X}}_s)\mathrm{d}\mathbf{B}_s - \mathbf{k}_t$.

(d) For any $t \geq 0$, $|\mathbf{k}|_t = \int_0^t \mathbf{1}_{\bar{x}_s \in \partial\mathcal{M}}(\bar{x}_s)\mathrm{d}|\mathbf{k}|_s$ *and* $\mathbf{k}_t = \int_0^t n(\bar{x}_s)\mathrm{d}|\mathbf{k}|_s$.

The process $(\mathbf{X}_t)_{t\geq 0}$ is almost surely continuous, so we could apply the previous theorem almost surely for all the realizations of the process,. However, this does not tell us if the obtained solutions $(\bar{\mathbf{X}}_t, \mathbf{k}_t)_{t\geq 0}$ form themselves a process. The main difference with Theorem D.2 is in Theorem D.3-(c) which differs from Theorem D.3-(c). Note that in the case where $b = 0$ and $\sigma = \mathrm{Id}$ we recover Theorem D.3-(c). This is not true in the general case. However, it can be seen that for any realization of the process $(\bar{\mathbf{X}}_t)_{t\geq 0}$, we have that $(\bar{\mathbf{X}}_t, \mathbf{k}_t)_{t\geq 0}$ is solution of the *deterministic* Skorokhod problem by letting $x_t = \bar{\mathbf{X}}_0 + \int_0^t b(\bar{\mathbf{X}}_s)\mathrm{d}s + \int_0^t \sigma(\bar{\mathbf{X}}_s)\mathrm{d}\mathbf{B}_s$. The backward and forward Kolmogorov equations can be found in Burdzy et al. (2004). Note that the presence of the process $(\mathbf{k}_t)_{t\geq 0}$ incurs notable complications compared to unconstrained processes. In particular, there is no martingale problem associated with weak solutions of reflected SDEs but only sub-martingale problems, see Kang & Ramanan (2017) for instance.

E Implicit Score Matching Loss

E.1 Proof of ISM

Using the divergence theorem, we have

$$\begin{aligned}&(1/2)\int_{\mathcal{M}}\|\mathbf{s}_\theta(x)-\nabla\log p_t(x)\|^2 p_t(x)\mathrm{d}\mu(x)\\ &=(1/2)\int_{\mathcal{M}}\|\mathbf{s}_\theta(x)\|^2 p_t(x)\mathrm{d}\mu(x)-\int_{\mathcal{M}}\langle\mathbf{s}_\theta(x),\nabla\log p_t(x)\rangle p_t(x)\mathrm{d}\mu(x)+(1/2)\int_{\mathcal{M}}\|\nabla\log p_t(x)\|^2 p_t(x)\mathrm{d}\mu(x)\\ &=(1/2)\int_{\mathcal{M}}\|\mathbf{s}_\theta(x)\|^2 p_t(x)\mathrm{d}\mu(x)-\int_{\partial\mathcal{M}}\langle\mathbf{s}_\theta(x),\mathbf{n}\rangle p_t(x)\mathrm{d}\nu(x)\\ &\qquad+\int_{\mathcal{M}}\mathrm{div}(\mathbf{s}_\theta)(x)p_t(x)\mathrm{d}\mu(x)+(1/2)\int_{\mathcal{M}}\|\nabla\log p_t(x)\|^2 p_t(x)\mathrm{d}\mu(x).\end{aligned}$$

Using that $\mathbf{s}_\theta(x)=0$ for all $x\in\partial\mathcal{M}$, we get that

$$\begin{aligned}&(1/2)\int_{\mathcal{M}}\|\mathbf{s}_\theta(x)-\nabla\log p_t(x)\|^2 p_t(x)\mathrm{d}\mu(x)\\ &\qquad=(1/2)\int_{\mathcal{M}}\|\mathbf{s}_\theta(x)\|^2 p_t(x)\mathrm{d}\mu(x)+\int_{\mathcal{M}}\mathrm{div}(\mathbf{s}_\theta)(x)p_t(x)\mathrm{d}\mu(x)+(1/2)\int_{\mathcal{M}}\|\nabla\log p_t(x)\|^2 p_t(x)\mathrm{d}\mu(x),\end{aligned}$$

which concludes the proof.

E.2 Importance of scaling function

As discussed in Section 3.3, we include a monotone scaling function h which is zero close to the boundary to ensure the relevant conditions are met for the score matching loss and the boundary conditions. This may seem like a technical detail, but is of significant practical importance. Upon removal of the scaling function, we observe that the learned score functions behave strangely around the boundary in the reverse process, leading to samples that do not match the forward process. The problems are apparent when comparing the top three plots of Fig. 15 and Fig. 16. Interestingly, we found that these issues early on in the sampling are smoothed out by the end of the reverse process, but still lead to a failure to recover the target density.

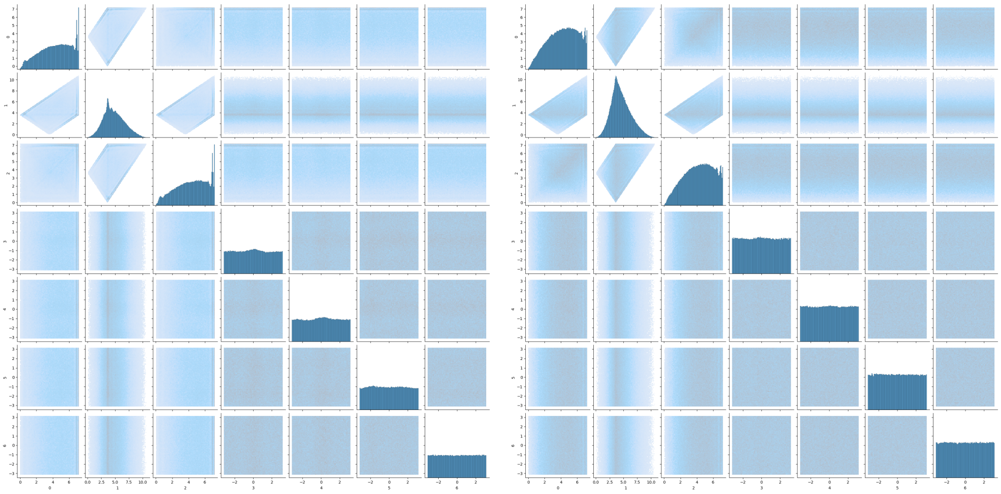


Figure 15: Reverse process samples for the cyclic peptide dataset from Section 5.3 at $t = 1.0, 0.9$ (left and right respectively) trained without the scaling function.

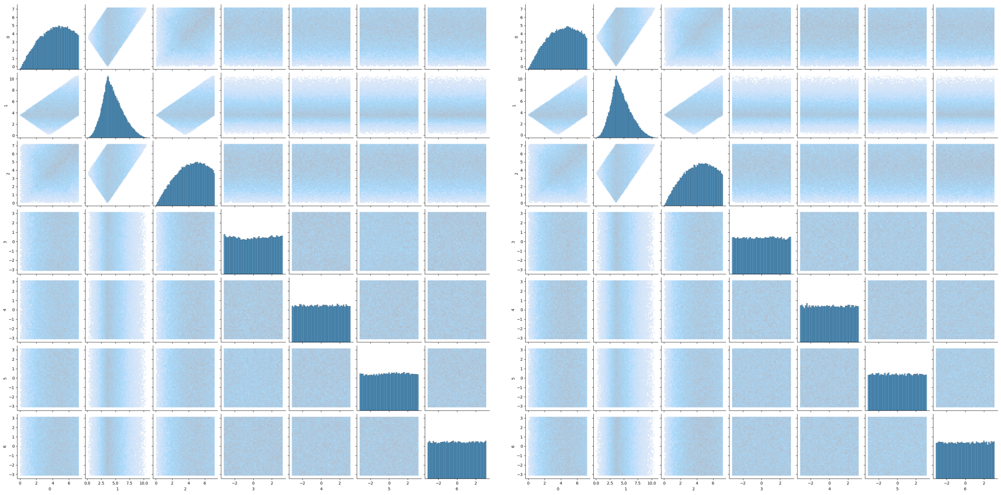


Figure 16: Reverse process samples for the cyclic peptide dataset from Section 5.3 at $t = 1.0, 0.9$ (left and right respectively) trained with the scaling function.

F Likelihood evaluation

One key advantage of constructing a continuous noising process is that, similarly to Song et al. (2021), we can evaluate the model's likelihood via the following *probability flow* Ordinary Differential Equation (ODE). In particular, for the Langevin dynamics (6) which we recall

$$\mathrm{d}\mathbf{X}_t = \tfrac{1}{2}\mathrm{div}(\mathfrak{g}^{-1})(\mathbf{X}_t)\mathrm{d}t + \mathfrak{g}(\mathbf{X}_t)^{-\frac{1}{2}}\mathrm{d}\mathbf{B}_t,$$

the following ODE has the same marginal density

$$\mathrm{d}\mathbf{Y}_t = \left[\tfrac{1}{2}\nabla \cdot \mathfrak{g}^{-1}(\mathbf{Y}_t) - \tfrac{1}{2}\mathfrak{g}^{-1}(\mathbf{Y}_t)\nabla \log p_t(\mathbf{Y}_t)\right]\mathrm{d}t.$$

We conclude this section with a derivation of the equivalent ODE. We highlight that the ODE representation for reflected diffusion models was first derived in Lou & Ermon (2023). We recall that if $\mathcal{M} \subset \mathbb{R}^d$ is a bounded open set with smooth boundary then (Burdzy et al., 2004, Theorem 2.2) ensures that the reflected Brownian motion admits a density w.r.t. the Lebesgue measure. We denote p_t this smooth density.

Proposition F.1. *Assume that $\mathcal{M} \subset \mathbb{R}^d$ is a bounded open set with smooth boundary. Assume that $(t, x) \mapsto \nabla \log p_t(x)$ is smooth on $[0, +\infty) \times \partial\mathcal{M}$. Let $(\bar{\mathbf{B}}_t)_{t\geq 0}$ be the reflected Brownian motion with $\bar{\mathbf{B}}_0 \sim p_0$ smooth and supported in $\mathcal{M}$. Let $(\mathbf{X}_t)_{t\geq 0}$ be given for any $t \geq 0$ by $\mathrm{d}\mathbf{X}_t = \frac{1}{2}\nabla \log p_t(\mathbf{X}_t)\mathrm{d}t$ and $\mathbf{X}_0 \sim p_0$, where p_t denotes the density of $\bar{\mathbf{B}}_t$ w.r.t. the Lebesgue measure for any $t > 0$. Then for any $t \in [0, T]$, $\bar{\mathbf{B}}_t$ and $\mathbf{Y}_t$ have the same distribution.*

Proof. Since the distributions of $(\bar{\mathbf{B}}_t)_{t\in[0,T]}$ and $(\mathbf{Y}_t)_{t\in[0,T]}$ satisfy the same Fokker-Planck equation whenever these processes are well-defined. Therefore, we first show that the process $(\mathbf{Y}_t)_{t\in[0,T]}$ is well-defined and stay in $\mathcal{M}$ at all times. Using (Burdzy et al., 2004, Theorem 2.2), we have that $\partial p_t(x) = \frac{1}{2}\mathrm{div}(\nabla \log p_t)(x)$, for any $t > 0$ and $x \in \mathcal{M}$. Next, we define $\mathrm{d}\mathbf{X}_t = \frac{1}{2}\nabla \log p_t(\mathbf{X}_t)\mathrm{d}t$. Note that $(\mathbf{X}_t)_{t\geq 0}$ is defined up to an explosion time T_∞ after which we fix $\mathbf{X}_t = \infty$. Denote T_0 the first time such that $\mathbf{X}_t \in \partial\mathcal{M}$. Note that since p_0 is supported on $\mathcal{M}$ we have $T_0 > 0$. We denote $(\mathbf{Y}_t)_{t\in[0,T_0]} = (\mathbf{X}_{T_0 - t})_{t\in[0,T_0]}$. We have that for any $t \in [0, T_0]$, $\mathrm{d}\mathbf{Y}_t = -\frac{1}{2}\nabla \log p_{T_0 - t}(\mathbf{Y}_t)\mathrm{d}t$. Since $(t, x) \mapsto \nabla \log p_t(x)$ is smooth on $[0, +\infty) \times \partial\mathcal{M}$, we get that for any $t \in [0, T_0]$, $\mathbf{Y}_t \in \partial\mathcal{M}$. In particular, we have that $\mathbf{Y}_{T_0/2} = \mathbf{X}_{T_0/2} \in \partial\mathcal{M}$ which is absurd. Therefore $T_0 = +\infty$ (which also implies that $T_\infty = +\infty$). Hence, $(\mathbf{X}_t)_{t\geq 0}$ is a flow on $\mathcal{M}$ and therefore for any $t \geq 0$, the density q_t of $\mathbf{X}_t$ is smooth and satisfies $\partial_t q_t(x) = -\frac{1}{2}\mathrm{div}(q_t \nabla \log p_t)(x)$. We conclude using the uniqueness of the solutions to the transport equation for smooth initialisation and coefficients on $\mathbb{R}^d$. □

G Time-reversal for reflected Brownian motion

We start with the following definition.

Definition G.1. Let $\mathcal{M} \subset \mathbb{R}^d$ be an open set. $\mathcal{M}$ has a smooth boundary if for any $x \in \partial\mathcal{M}$, there exists $\mathsf{U} \subset \mathbb{R}^d$ open and $f \in \mathrm{C}^\infty(\mathsf{U}, \mathbb{R})$ such that $x \in \mathsf{U}$ and (a) $\mathrm{cl}(\mathcal{M}) \cap \mathsf{U} = \{x \in \mathsf{U} : f(x) \leq 0\}$, (b) $\nabla f(x) \neq 0$ for any $x \in \mathsf{U}$ where $\mathrm{cl}(\mathcal{M})$ is the closure of $\mathcal{M}$.

We will make the use of the following lemma which is a straightforward extension of Burdzy et al. (2004, Theorem 2.6). The surface measure is defined in (Lee, 2006, Proposition 2.43). Under mild regularity assumptions, it corresponds to the Hausdorff measure of $\partial\mathcal{M}$, see Evans & Gariepy (2015).

Lemma G.2. *Let u such that $s \mapsto u(s, x) \in \mathrm{C}^1((0, T), \mathbb{R})$, for any $s \in (0, T)$, $x \mapsto u(s, x) \in \mathrm{C}^2(\mathcal{M}, \mathbb{R})$ and $u \in \mathrm{C}^1(\mathrm{cl}(\mathcal{M}), \mathbb{R})$. Then for any $T \geq 0$, $s, t \in [0, T]$, we have*

$$\mathbb{E}[\textstyle\int_s^t u(w, \bar{\mathbf{B}}_w)\mathrm{d}\mathbf{k}_w] = \frac{1}{2}\int_s^t \int_{\partial\mathcal{M}} u(x) p_w(x) \mathrm{d}\sigma(x) \mathrm{d}w.$$

Note that we recover Burdzy et al. (2004, Theorem 2.6) if we set $u = 1$. We also emphasize that the result of Burdzy et al. (2004, Theorem 2.6) is stronger than Theorem G.2 as it holds not only in expectation but also in L^2 and almost surely.

We are now ready to prove Theorem 3.2. We follow the approach of Petit (1997) which itself is based on an extension of Haussmann & Pardoux (1986). We refer to Cattiaux et al. (2021) for recent entropic approaches of time-reversal. Recall that $(\bar{\mathbf{B}}_t, \mathbf{k}_t)_{t\geq 0}$ is a solution to the *Skorokhod problem* (Skorokhod, 1961) if $(\mathbf{k}_t)_{t\geq 0}$ a bounded variation process and $(\bar{\mathbf{B}}_t)_{t\geq 0}$ a continuous adapted process such that for any $t \geq 0$, $\mathbf{B}_t = \bar{\mathbf{B}}_t + \mathbf{k}_t \in \mathcal{M}$, $(\bar{\mathbf{B}}_t)_{t\geq 0}$ and

$$|\mathbf{k}|_t = \textstyle\int_0^t \mathbf{1}_{\bar{\mathbf{B}}_s \in \partial\mathcal{M}} \mathrm{d}|\mathbf{k}|_s, \quad \mathbf{k}_t = \int_0^t \mathbf{n}(\bar{\mathbf{B}}_s) \mathrm{d}|\mathbf{k}|_s, \tag{14}$$

In what follows, we define $(\mathbf{Y}_t)_{t\in[0,T]}$ such that for any $t \in [0, T]$, $\mathbf{Y}_t = \bar{\mathbf{B}}_{T-t}$. Let us consider the process $(\tilde{\mathbf{B}}_t)_{t\in[0,T]}$ defined for any $t \in [0, T]$ by

$$\tilde{\mathbf{B}}_t = -\bar{\mathbf{B}}_T + \bar{\mathbf{B}}_{T-t} + \mathbf{k}_T - \mathbf{k}_{T-t} - \textstyle\int_{T-t}^T \nabla \log p_s(\bar{\mathbf{B}}_s) \mathrm{d}s.$$

First, note that $t \mapsto \tilde{\mathbf{B}}_t$ is continuous. Denote by $\mathcal{F}$, the filtration associated with $(\bar{\mathbf{B}}_{T-t})_{t\in[0,T]}$. We have that $(\tilde{\mathbf{B}}_t)_{t\in[0,T]}$ is adapted to $(\bar{\mathbf{B}}_{T-t})_{t\in[0,T]}$. Even more so, we have that $(\tilde{\mathbf{B}}_t)_{t\in[0,T]}$ satisfies the strong Markov property since $(\bar{\mathbf{B}}_t)_{t\in[0,T]}$ also satisfies the strong Markov property. Let $g \in \mathrm{C}_c^\infty(\mathrm{cl}(\mathcal{M}))$ and consider for any $0 \leq s \leq t \leq T$, $\mathbb{E}[(\tilde{\mathbf{B}}_t - \tilde{\mathbf{B}}_s)g(\bar{\mathbf{B}}_{T-t})]$. For any $0 \leq s \leq t \leq T$ we have

$$\mathbb{E}[(\tilde{\mathbf{B}}_t - \tilde{\mathbf{B}}_s)g(\bar{\mathbf{B}}_{T-t})] = \mathbb{E}[(-\bar{\mathbf{B}}_{T-s} + \bar{\mathbf{B}}_{T-t} + \mathbf{k}_{T-s} - \mathbf{k}_{T-t} - \textstyle\int_{T-t}^{T-s} \nabla \log p_u(\bar{\mathbf{B}}_u)\mathrm{d}u)g(\bar{\mathbf{B}}_{T-t})]. \tag{15}$$

In what follows, we prove that for any $0 \leq s \leq t \leq T$ we have $\mathbb{E}[(\tilde{\mathbf{B}}_t - \tilde{\mathbf{B}}_s)g(\bar{\mathbf{B}}_{T-t})] = 0$. Therefore, we only need to prove that for any $0 \leq s \leq t \leq T$ we have

$$\mathbb{E}[(-\bar{\mathbf{B}}_t + \bar{\mathbf{B}}_s + \mathbf{k}_t - \mathbf{k}_s - \textstyle\int_s^t \nabla \log p_u(\bar{\mathbf{B}}_u)\mathrm{d}u)g(\bar{\mathbf{B}}_t)] = 0. \tag{16}$$

Let $t \in (0, T]$. We introduce $u : [0, t] \times \mathcal{M}$ such that for any $s \in [0, t]$ and $x \in \mathcal{M}$, $u(s, x) = \mathbb{E}[g(\bar{\mathbf{B}}_t)|\bar{\mathbf{B}}_s = x]$. Using Burdzy et al. (2004, Theorem 2.8) we get that for any $x \in \mathcal{M}$, $s \mapsto u(s, x) \in \mathrm{C}^1((0, t), \mathbb{R})$ and for any $s \in (0, t)$, $x \mapsto u(s, x) \in \mathrm{C}^2(\mathcal{M}, \mathbb{R})$ and $x \mapsto u(s, x) \in \mathrm{C}^1(\mathrm{cl}(\mathcal{M}), \mathbb{R})$. In addition, we have that for any $s \in (0, t)$ and for any $x \in \mathcal{M}$ and $x_0 \in \partial\mathcal{M}$

$$\partial_s u(s, x) + \tfrac{1}{2}\Delta u(s, x) = 0, \qquad \langle \nabla u(s, x_0), \mathbf{n}(x_0) \rangle = 0. \tag{17}$$

This equation is called the backward Kolmogorov equation. Using equation 17, $\bar{\mathbf{B}}_t = \mathbf{B}_t - \mathbf{k}_t$ for any $t \geq 0$ and the Itô formula for semimartingale (Revuz & Yor, 2013, Chapter IV, Theorem 3.3) we have that for any

$s \in (0,t)$

$$\begin{aligned}\mathbb{E}[u(t,\bar{\mathbf{B}}_t)\bar{\mathbf{B}}_t] =& \mathbb{E}[u(s,\bar{\mathbf{B}}_s)\bar{\mathbf{B}}_s] + \mathbb{E}[\tfrac{1}{2}\int_s^t \bar{\mathbf{B}}_w \Delta u(w,\bar{\mathbf{B}}_w)\mathrm{d}w] + \mathbb{E}[\int_s^t \nabla u(w,\bar{\mathbf{B}}_w)\mathrm{d}w] \\ & - \mathbb{E}[\int_s^t \bar{\mathbf{B}}_w \langle \nabla u(w,\bar{\mathbf{B}}_w), \mathbf{n}(\bar{\mathbf{B}}_w)\rangle \mathrm{d}|\mathbf{k}|_w] \\ & - \mathbb{E}[\int_s^t u(w,\bar{\mathbf{B}}_w)\mathbf{n}(\bar{\mathbf{B}}_w)\mathrm{d}|\mathbf{k}|_w] \\ & + \mathbb{E}[\int_s^t \bar{\mathbf{B}}_w \partial_w u(w,\bar{\mathbf{B}}_w)\mathrm{d}w] \\ =& \mathbb{E}[u(s,\bar{\mathbf{B}}_s)\bar{\mathbf{B}}_s] + \mathbb{E}[\int_s^t \nabla u(w,\bar{\mathbf{B}}_w)\mathrm{d}w] - \mathbb{E}[\int_s^t u(w,\bar{\mathbf{B}}_w)\mathbf{n}(\bar{\mathbf{B}}_w)\mathrm{d}|\mathbf{k}|_w] \end{aligned} \tag{18}$$

In addition, using the Fubini theorem and the definition of $\mathbf{k}_t$ we have that for any $s \in (0,t)$

$$\mathbb{E}[\int_s^t u(w,\bar{\mathbf{B}}_w)\mathbf{n}(\bar{\mathbf{B}}_w)\mathrm{d}|\mathbf{k}|_w] = \mathbb{E}[\int_s^t \mathbb{E}[g(\bar{\mathbf{B}}_t)|\bar{\mathbf{B}}_w]\mathbf{n}(\bar{\mathbf{B}}_w)\mathrm{d}|\mathbf{k}|_w] = \mathbb{E}[g(\bar{\mathbf{B}}_t)(\mathbf{k}_t - \mathbf{k}_s)]. \tag{19}$$

Finally, using the divergence theorem and Burdzy et al. (2004, Theorem 2.6) we have that for any $s \in (0,t)$

$$\begin{aligned}\mathbb{E}[\int_s^t \nabla u(w,\bar{\mathbf{B}}_w)\mathrm{d}w] &= \int_s^t \int_{\mathcal{M}} \nabla u(w,x) p_w(x)\mathrm{d}x\mathrm{d}w \\ &= -\int_s^t \int_{\mathcal{M}} u(w,x)\nabla \log(p_w(x)) p_w(x)\mathrm{d}x\mathrm{d}w + \int_s^t \int_{\partial\mathcal{M}} u(w,x) p_w(x)\mathrm{d}x\mathrm{d}\sigma(w),\end{aligned}$$

where σ is the surface area measure on $\partial\mathcal{M}$, see Burdzy et al. (2004). Using Theorem G.2 and the Fubini theorem we get that

$$\begin{aligned}\mathbb{E}[\int_s^t \nabla u(w,\bar{\mathbf{B}}_w)\mathrm{d}w] &= -\int_s^t \int_{\mathcal{M}} u(w,x)\nabla \log(p_w(x)) p_w(x)\mathrm{d}x\mathrm{d}w + \mathbb{E}[\int_s^t u(w,\bar{\mathbf{B}}_w)\mathrm{d}\mathbf{k}_w] \\ &= -\mathbb{E}[\int_s^t g(\bar{\mathbf{B}}_t)\nabla \log(p_w(\bar{\mathbf{B}}_w))\mathrm{d}w] + 2\mathbb{E}[g(\bar{\mathbf{B}}_t)(\mathbf{k}_t - \mathbf{k}_s)]\end{aligned} \tag{20}$$

Combining equation 18, equation 19 and equation 20 we get that

$$\mathbb{E}[u(t,\bar{\mathbf{B}}_t)] = \mathbb{E}[u(s,\bar{\mathbf{B}}_s)] - \mathbb{E}[g(\bar{\mathbf{B}}_t)\int_s^t \nabla \log p_w(\bar{\mathbf{B}}_w)\mathrm{d}w] + \mathbb{E}[g(\bar{\mathbf{B}}_t)(\mathbf{k}_t - \mathbf{k}_s)].$$

Therefore, we get equation 16 and equation 15. Hence, $(\tilde{\mathbf{B}}_t)_{t\in[0,T]}$ is a continuous martingale. In addition, we have that for any $t \in [0,T]$, $\mathbb{E}[\tilde{\mathbf{B}}_t\tilde{\mathbf{B}}_t^\top] = t\,\mathrm{Id}$ and therefore, $(\tilde{\mathbf{B}}_t)_{t\in[0,T]}$ is a Brownian motion using the Lévy characterisation of Brownian motion (Revuz & Yor, 2013, Chapter IV, Theorem 3.6). Denote $(\mathbf{j}_t)_{t\in[0,T]} = (\mathbf{k}_T - \mathbf{k}_{T-t})_{t\in[0,T]}$. Using equation 15, we have that for any $t \in [0,T]$

$$\bar{\mathbf{B}}_{T-t} = \bar{\mathbf{Y}}_0 + \tilde{\mathbf{B}}_t + \int_0^t \nabla \log p_{T-s}(\mathbf{Y}_s)\mathrm{d}s - \mathbf{j}_t.$$

Using equation 14, we have for any $t \in [0,T]$

$$|\mathbf{j}|_t = \int_0^t \mathbf{1}_{\mathbf{Y}_s\in\partial\mathcal{M}}\mathrm{d}|\mathbf{j}|_s, \quad \mathbf{j}_t = \int_0^t \mathbf{n}(\bar{\mathbf{Y}}_s)\mathrm{d}|\mathbf{j}|_s,$$

which concludes the proof.

H Configurational modelling of robotic arms under manipulability constraints

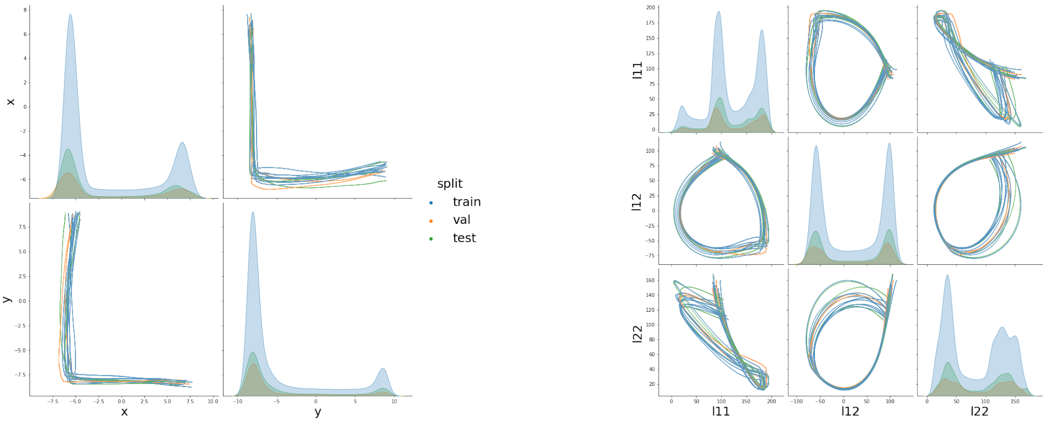
Accurately determining and specifying the movement of a robotic arm and the forces it exerts is a fundamental problem in many real-world robotics applications. A widely-used set of descriptors for modelling the flexibility of a given joint configuration are so-called manipulability ellipsoids (Yoshikawa, 1985), which are kinetostatic performance measures that quantify the ability to move or exert forces along different directions. Jaquier et al. (2021) present a geometric framework to learn trajectories of manipulability ellipsoids by making use of the fact any ellipsoid $M \in \mathbb{R}^N$ is defined by the set of points $\{\mathbf{x}|\mathbf{x}^T\mathbf{A}\mathbf{x} = 1\}$ where $\mathbf{A}$ lies on the manifold of $N \times N$ symmetric positive definite matrices S^N_{++}.

In many practical settings, it is desirable to constrain the minimal or maximal volume of a manipulability ellipsoid to retain motional flexibility or limit the magnitude of the exerted force. This necessitates lower or upper limits on the determinant of $\mathbf{A}$, translating into constraints on S^N_{++}. To model this, we make use of one of the datasets introduced by Jaquier et al. (2021), containing demonstrations of a robotic arm drawing different letters in the plane, providing the respective positional trajectories ($\mathbb{R}^2$) and velocity manipulability ellipsoids (S^2_{++}).

We use the processing routines provided by Jaquier et al. (2021) to interpolate the trajectories into 10^4 distinct points, for each of which we derive the position $\mathbf{x} \in \mathbb{R}^2$ and the PSD matrix

$$\mathbf{A} = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \in S^2_{++}.$$

parametrising the velocity manipulability ellipsoid $M \in \mathbb{R}^2$. The resulting data is split into training, validation, and test sets by trajectory and visualised in Figure 17. We add a small amount of Gaussian noise to these trajectories, which is shown as the target distribution in J.2.



(a)

(b)

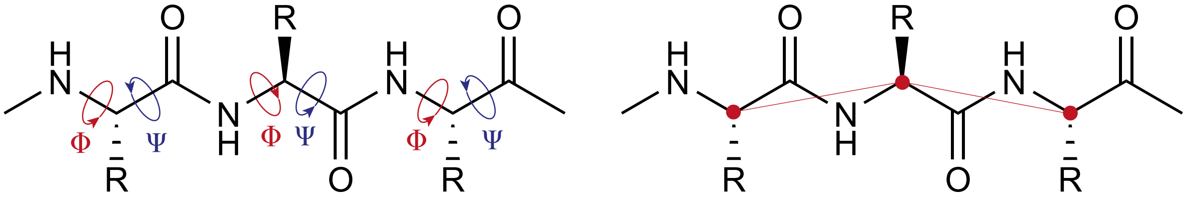
Figure 17: Positional trajectories $\mathbf{x} \in \mathbb{R}^2$ (a) and the parameters l_{11}, l_{12}, l_{22} of the the SPD matrix $\mathbf{A} \in S^2_{++}$ (b) for the letter L.

I Conformational modelling of polypeptide backbones under end point constraints

Polypeptides and proteins constitute an important class of biogenic macromolecules that underpin most aspects of organic life. Accurately modelling their conformational ensembles, i.e. the set of three-dimensional structures they assume under physiological conditions, is essential to both understanding the biological function of existing and designing the enzymatic or therapeutic properties of novel proteins (Lane, 2023). Motivated by the success of diffusion models in computer vision and natural language processing, there has been considerable interest in applying them to learn and sample from distributions over the conformational space of protein structures (Watson et al., 2022; Trippe et al., 2022; Wu et al., 2022).

I.1 Problem parameterisation

Proteins are biopolymers in which a sequence of N amino acids is joined together through $N-1$ peptide bonds, resulting in a so-called polypeptide backbone with protruding amino acid residues. As the deviation of chemical bond lengths and angles from their theoretical optimum is generally negligible, the problem of modelling the three-dimensional structure of this polypeptide chain is often reframed in the space of the internal torsion angles Φ and Ψ (see Figure 18a for an illustration), which can be modelled on a $(2N-2)$-dimensional torus $\mathbb{T}^{2N-2}$.



(a) A commonly-used approximate parameterisation of backbone geometry only considers the C_α torsion angles Φ and Ψ.

(b) As peptide bond orientations can be inferred relatively reliably, researchers often only model the C_α traces.

Figure 18: Standard approaches to modelling the conformations of polypeptide backbones.

In many data-scarce practical settings such as antibody or enzyme design, it is often unnecessary or even undesirable to model the structure of an entire protein, as researchers are primarily interested in specific functional sites with distinct biochemical properties. However, generating conformational ensembles for such substructural elements necessitates positional constraints on their endpoints to ensure that they can be accommodated by the remaining scaffold. While it is conceivable that a diffusion model could derive such constraints from limited experimental data, we argue that it is much more efficient and precise to encode them explicitly.

For this purpose, we adopt the distance constraint formulation from Han & Rudolph (2006) and interpret the backbone as a spatial chain with N spherical joints and fixed-length links (see Figure 19a for an illustration). After selecting a suitable anchor point, the geometry of the polypeptide chain is fully specified by (a) the set of link lengths $\ell = \{\ell_j\}_{j=1}^{N}$, (b) the set of vectors $\mathbf{r} = \{r(1,j)\}_{j=2}^{N}$ between the anchor point and each atom in the chain, and (c) the set of dihedral angles

$$\mathbf{T} = \left\{ \arccos\left(\frac{|\langle r(1,j) \times r(1,j+1), r(1,j+1) \times r(1,j+2)\rangle|}{|r(1,j) \times r(1,j+1)||r(1,j+1)r(1,j+2)|} \right) \right\}_{j=2}^{N-2} \in \mathbb{T}^{N-3}$$

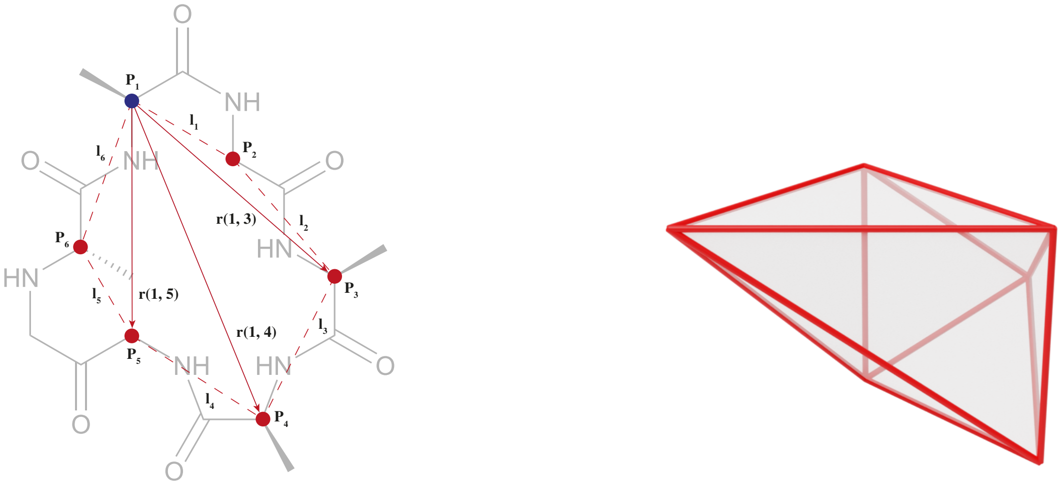
between any three consecutive vectors. After specifying the fixed bond lengths ℓ, including an arbitrary anchor point distance $d_{\text{anchor}} = \ell_N = r(1,N)$, the set of valid vectors $\mathbf{r}$ is given by the convex polytope $\mathbb{P} \subseteq \mathbb{R}^3$ defined by the following linear constraints (see Figure 19b for an illustration):

$$\begin{aligned} r(1,3) &\leq \ell_1 + \ell_2, \\ -r(1,3) &\leq -\left|\ell_1 - \ell_2\right|, \end{aligned}$$

$$\left.\begin{aligned} r(1,j) - r(1,j+1) &\leq \ell_j, \\ -r(1,j) + r(1,j+1) &\leq \ell_j, \\ -r(j) - r(j+1) &\leq -\ell_j, \end{aligned}\right\} 3 \leq j \leq N-2,$$

$$\begin{aligned} r(1,N-1) &\leq \ell_{N-1} + d_{\text{anchor}}, \\ -r(1,N-1) &\leq -\left|\ell_{N-1} - d_{\text{anchor}}\right|. \end{aligned}$$

This means that the set of all valid polypeptide backbone conformations is defined by the product manifold $\mathbb{P} \times \mathbb{T}^{N-3}$, enabling us to train diffusion models that exclusively generate conformations with a fixed anchor point distance d_{anchor}.



(a) An illustrative diagram of the proposed parameterisation for modelling the C_α trace geometry of the cyclic peptide c-AAGAGG.

(b) The convex polytope constraining the diagonals of the triangles for the given bond lengths in the illustrated molecule. The total design space is the product of this polytope with the 4D flat torus.

Figure 19: Parameterising the conformational space of polypeptide backbones under anchor point distance constraints.

I.2 Data generation and model training

As a proof-of-concept for the practicality of our methods, we chose to model the conformational distribution of the cyclic peptide c-AAGAGG. Cyclic peptides are an increasingly important drug modality with therapeutic uses ranging from antimicrobials to oncology, exhibiting circular polypeptide backbones (i.e. $d_{\text{anchor}} = 0$) that confer a range of desirable pharmacodynamic and pharmacokinetic properties (Dougherty et al., 2019). To reduce the dimensionality of the problem, we only consider the C_α traces (with fixed C_α-C_α link distances of 3.6 Å) instead of the full polypeptide backbone (see Figure 18b), although we note that our framework is fully applicable to both settings.

To derive a suitable dataset, the product manifold $\mathbb{P} \times \mathbb{T}^3$ describing the conformations of cyclic C_α traces of length $N = 6$ was constructed (see Figures 19a and 19b for an illustration) and used to generate 10^7 uniform samples satisfying the anchor point distance constraint $d_{\text{anchor}} = 0$. Subsequently, an estimate of the free energy E_i of each sample i was obtained by (1) reconstructing the full-atom peptide from each C_α trace using the PULCHRA algorithm (Rotkiewicz & Skolnick, 2008), (2) relaxing all non-C_α backbone and side-chain atoms (keeping the C_α positions fixed), and (3) quantifying the potential energy of each of the resulting conformations using the OpenMM suite of molecular dynamics tools (Eastman et al., 2017), and the AMBER force field (Hornak et al., 2006). These free energy estimates were then used to approximate the Boltzmann distribution over conformational states

$$p_B(i) \propto \exp\left(-\tfrac{E_i}{k_B T}\right),$$

where temperature was set to $T = 273.15\,\mathrm{K}$ and $k_B = 1.380\,649 \times 10^{-23}\,\mathrm{J\,K^{-1}}$ is the Boltzmann constant. We then apply a very minor amount of smoothing to the resulting distribution by running forward Brownian motion on both the polytope and the torus for 10 steps, using a small step size of 5×10^{-3} and the respective metrics. Finally, a subsample of 10^6 C_α traces was drawn from this distribution and used for training and evaluating our models.

J Experimental details

In what follows we describe the experimental settings used to generate results introduced in Section 5. The models and experiments have been implemented in Jax (Bradbury et al., 2018), using a modified version of the Riemannian geometry library Geomstats (Miolane et al., 2020).

Architecture. The architecture of the score network $\boldsymbol{s}_\theta$ is given by a multilayer perceptron with 6 hidden layers with 512 units each. We use sinusoidal activation functions.

Training. All models are trained by the stochastic optimizer Adam (Kingma & Ba, 2014) with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, batch-size of 256 data-points. The learning rate is annealed with a linear ramp from 0 to 1000 steps, reaching the maximum value of $2e-4$, and from then with a cosine schedule down to 0 after $100k$ iterations in total.

Diffusion. Following Song et al. (2021), the diffusion models diffusion coefficient is parametrized as $g(t) = \sqrt{\beta(t)}$ with $\beta: \ t \mapsto \beta_{\min} + (\beta_{\max} - \beta_{\min}) \cdot t$, where we found $\beta_{\min} = 0.001$ and $\beta_{\max} = 6$ to work best.

Metrics. We measure the performance of trained models via the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012), which is a kernel based metric between two distributions P and Q. The MMD can be empirically approximated with the following U-statistics $\mathrm{MMD}^2(P,Q) = \frac{1}{m(m-1)} \sum_i \sum_{j \neq i} k(x_i, x_j) + \frac{1}{m(m-1)} \sum_i \sum_{j \neq i} k(y_i, y_j) - 2\frac{1}{m^2} \sum_i \sum_j k(x_i, y_j)$ with $x_i \sim P$ and $y_i \sim Q$, where k is a kernel. For synthetic experiments we use a sum of weighted RBF kernels matching the generating distributions for the Gaussian mixtures. For the other experiments we use an RBF kernel. For all experiments we use 100,000 samples to compute the MMD.

J.1 Synthetic data on polytopes

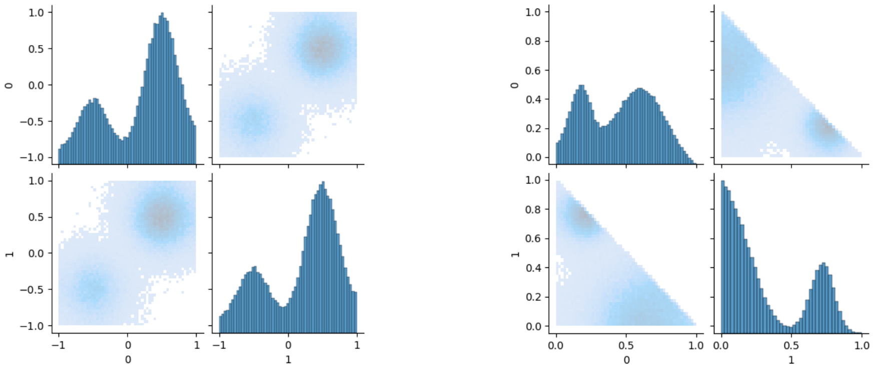
Hypercube $[-1,1]^n$. The hypercube is a specific instance of a convex polytope where the affine constraints are given by the following coefficients:

$$A = \begin{pmatrix} 1 & \dots & 0 \\ -1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \\ 0 & \dots & -1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix}.$$

Where A is a $2n \times n$ matrix and b an n dimensional vector.

We construct the training and test datasets by sampling for both 100000 points from a mixture of 'wrapped normal' distributions illustrated in Figure 20a and which density is given by

$$p_0(x) = 0.7\,\mathrm{ReflectedStep}[(0.5,0.5),\cdot,\{f_i\}_{i\in\mathcal{I}}]\#\mathcal{N}(0,0.25)+0.3\,\mathrm{ReflectedStep}[(-0.5,-0.5),\cdot,\{f_i\}_{i\in\mathcal{I}}]\#\mathcal{N}(0,0.25).$$



(a) On the hypercube.

(b) On the simplex.

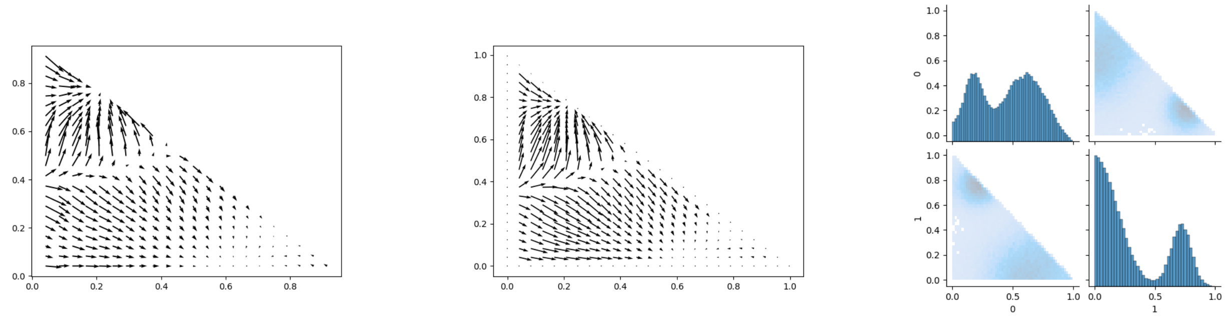
Figure 20: Pairwise and marginals samples from the synthetic data distribution.

Simplex Δ^n. Similarly, to parameterise the simplex as a convex polytope we set the matrix and constraints to be given by

$$A = \begin{pmatrix} -1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

Where A will be a $n-1 \times n$ matrix. Essentially we perform diffusion over the first $n-1$ components of the simplex, allowing the last component to be determined by the one minus the sum of the first $n-1$.

Similarly than for the hypercube, we construct the training and test datasets from generated data points which are illustrated in Figure 20b. The score network at different times is illustrated in Figure 21.



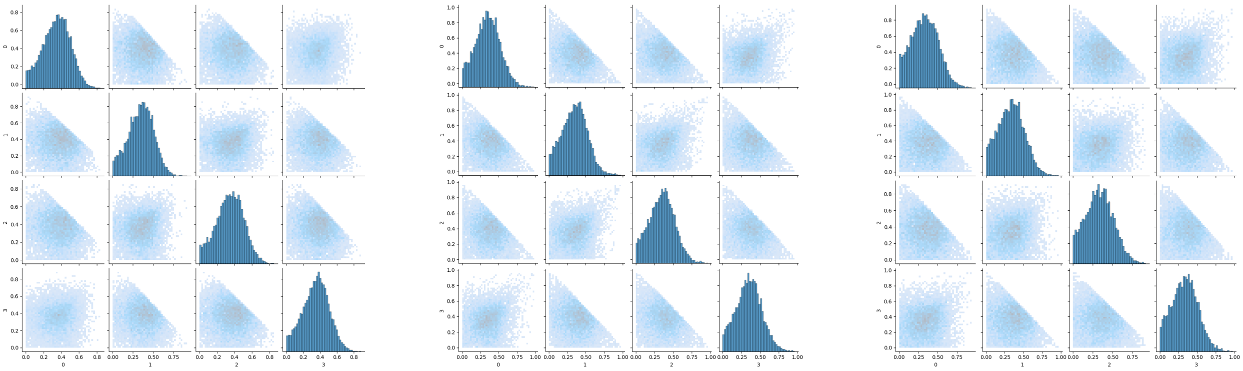
(a) $t = 0.01$.

(b) $t = 0.5$.

(c) Generated distribution.

Figure 21: Evolution of the score on the simplex and generated distribution.

The Birkhoff polytope. The Birkhoff polytope is the space of doubly stochastic matrices, i.e. $B_n = \{\mathrm{P} \in [0,1]^{n\times n} : \sum_i^n P_{i,j} = 1, \sum_j^n P_{i,j} = 1\}$. It is a convex polytope in $\mathbb{R}^{n^2}$ and has dimension $d = (n-1)^2$.



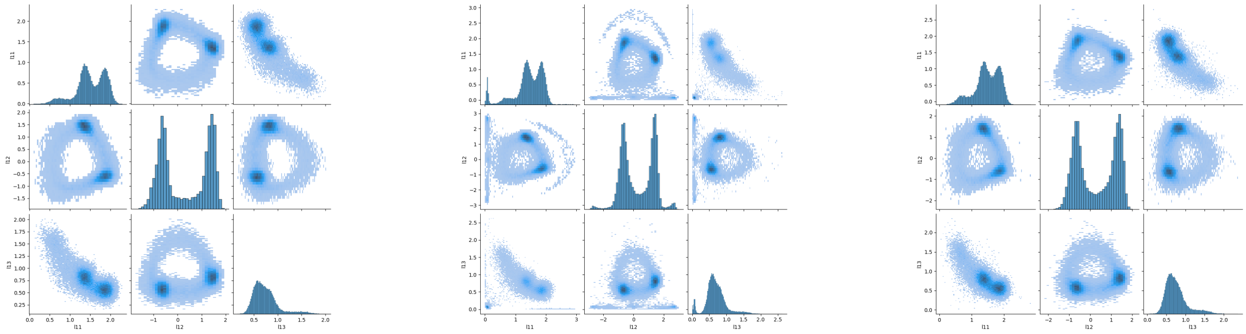
(a) Data.

(b) Log-barrier.

(c) Reflected.

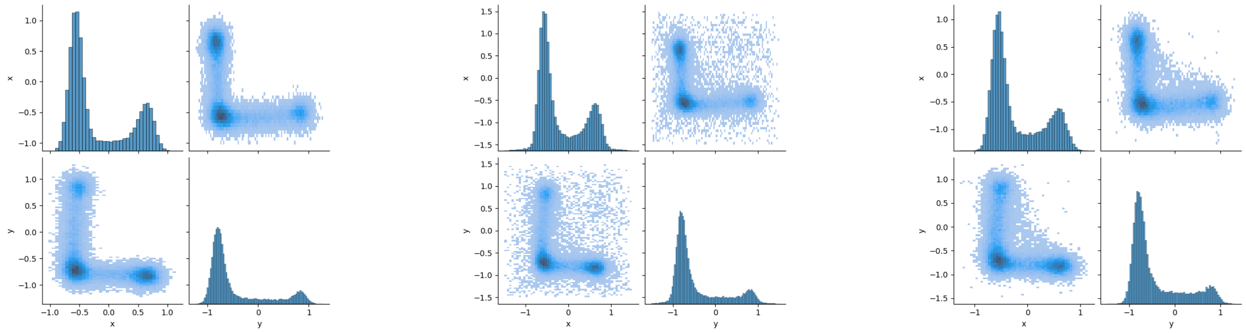
Figure 22: Pairwise and marginals samples on the Birkhoff polytope from synthetic data distribution and from trained constrained diffusion models.

J.2 Constrained SPD matrices for robotic arms modelling



(a) Data. (b) Log-barrier. (c) Reflected.

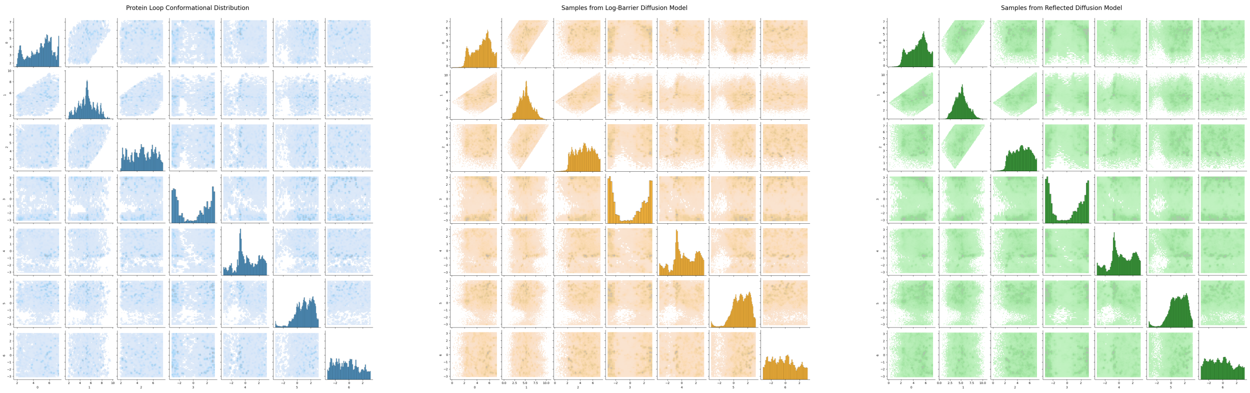
Figure 23: Pairwise and marginals distributions over the coefficients L_{11}, L_{21}, L_{22} of the lower triangle matrix parameterising SPD matrices $M = LL^\top$ (which represent the manipulability ellipsoids of the robotic arms).



(a) Data. (b) Log-barrier. (c) Reflected.

Figure 24: Pairwise and marginals distributions over the (x, y) locations of the robotic arms.

J.3 Conformational modelling of polypeptide backbones under anchor point constraints



(a) Data. (b) Log-barrier. (c) Reflected.

Figure 25: Pairwise and marginals distributions over the dimensions of the polytope and torus used to model the conformational ensembles of cyclic peptides generated by the reflected diffusion model.