LATENT PROCESS GENERATOR MATCHING

Lukas Billera[*], Hedwig Nora Nordlinder[*] & Ben Murrell
Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden
lukas.billera@ki.se, hedwig.nordlinder@ki.se, benjamin.murrell@ki.se

ABSTRACT

Many recent flow-matching and diffusion-style generative models rely on auxiliary stochastic dynamics during training: a richer process is simulated to define conditional targets, but the auxiliary state is either intractable to sample at generation time or simply not part of the desired output. Existing Generator Matching theory formalises conditioning on static latent random variables, and several recent papers prove special cases of projection results for particular augmented-state constructions. We introduce latent process generator matching, a general framework that treats the observed generative state as a deterministic image $X_t = \Phi(Y_t)$ of a tractable Markov process Y_t. We show that in this setting one may learn the generator of a stochastic process on the image space which has the same one-time marginal distributions as the projected process. This generalizes and subsumes the discrete latent process results from the literature, and extends Generator Matching from static latent variables to a rich family of time-dependent latent conditional processes.

1 INTRODUCTION

In recent extensions of flow matching and diffusion approaches to generative modelling, one constructs a Markov process on an extended state space $\mathcal{X} \times \mathcal{Z}$ representing the conditional paths of a generative model, but at generation time one wishes to simulate a process on $\mathcal{X}$ alone whose one-time marginals coincide with those of the $\mathcal{X}$-component of the joint process. This is the case for Ifriqi et al. (2025), Nguyen et al. (2025), and Havasi et al. (2025). A related situation arises when an auxiliary process is introduced to aid training but modelling its dynamics at generation time is unnecessary or difficult, as in Billera et al. (2025b) and Kim et al. (2025). In each of these works, the projection result and its associated loss are derived on a case-by-case basis, and all theorems are restricted to marginalization over a discrete component of the extended state space. We introduce a general framework that removes these restrictions: given a time-inhomogeneous Feller process $(Y_t)_{0\leq t\leq 1}$ on an arbitrary state space $\mathcal{Y}$ and a map $\Phi\colon \mathcal{Y} \to \mathcal{X}$, one may learn a linear parametrisation of the generator of a Feller process on $\mathcal{X}$ whose one-time marginals coincide with those of $(\Phi(Y_t))_{0\leq t\leq 1}$. For $\mathcal{Y} = \mathcal{X} \times \mathcal{Z}$ and Φ the projection onto the first coordinate, this subsumes these prior works as special cases, allowing for a general class of latent processes $(Z_t)_{0\leq t\leq 1}$ in a nearly arbitrary state space $\mathcal{Z}$, using the formalism of generator matching to allow for continuous, discrete, or manifold-valued processes.

In particular, the learnt process at $t = 1$ samples from the distribution of $\Phi(Y_1)$, which is the desired data distribution. We give sufficient conditions for a loss function to be valid in this general setting, recovering the results of the works cited above as corollaries. This result has broad applicability, enabling the construction of a wide array of new flow matching schemes by allowing for a more general class of latent spaces. As a concrete new application, we outline a non-projection $\Phi\colon \mathcal{Y} \to \mathcal{X}$ with manifold-valued latents for protein structure generation that separates chain-level rigid-body motion from internal flexibility (§4), where the particular chain-level versus residue-level or internal state is latent, and the model only sees the world state, which we plan to implement in future work.

[*]Equal contribution.

2 EARLIER WORK

Several recent generative models train with the aid of a latent stochastic process that is marginalised out at generation time. We briefly survey these and note how the pushforward framework developed herein subsumes the special-case justifications given in each.

2.1 GENERATOR MATCHING

Generator Matching (Holderrieth et al., 2025) unifies a large class of diffusion and flow matching models by showing that one may learn the marginal generator of a time-inhomogeneous Feller process by training against conditional generators, where conditioning is on a static latent random variable Z (typically a pair of endpoints). The present work extends this to conditioning on a latent *stochastic process*.

2.2 EDIT FLOWS, ONEFLOW AND FLOWCEPTION

Edit Flows (Havasi et al., 2025) solves variable-length discrete generation by adjoining a null token ε to the alphabet and learning joint rates on an extended state space $\mathcal{X} \times \mathcal{Z}$. The projection onto the first coordinate (stripping the null tokens) induces rates on $\mathcal{X}$ alone. Theorem 3.1 therein justifies this projection for CTMCs. The same technique is employed by OneFlow (Nguyen et al., 2025) but for interleaved token image generation and by Flowception (Ifriqi et al., 2025) for video generation via learned frame-insertion rates. In the latter two, a continuous process depends non-trivially on a latent discrete process which must be marginalised out. The theorems in Edit Flows, however, are concerned only with the case where both the main and latent process are discrete.

2.3 BRANCHING FLOWS

Branching Flows (Billera et al., 2025b) introduces tree-structured conditional paths for variable-length generation on manifold, Euclidean and discrete state spaces. The branching structure creates an ambiguity over which element belongs to which branch, necessitating an auxiliary discrete process. The resulting *auxiliary generator matching* result extends Theorem 3.1 of Havasi et al. (2025) to the case of an arbitrary "main" state space, while still requiring the latent process to be discrete.

2.4 ANY-ORDER FLEXIBLE LENGTH MASKED DIFFUSION

Similarly to Havasi et al. (2025), "Any-order flexible length masked diffusion" (Kim et al., 2025) solves the diffusion-model variable-length generation task. It uses flexible masked diffusion models, extending the continuous stochastic interpolants of Albergo et al. (2025) to the discrete, variable-length setting by training a neural network to learn the unmasking posterior and an insertion expectation, which determine the insertion/unmasking rates. During training one augments with an auxiliary process s_t which keeps track of the indices of tokens as they are inserted. This auxiliary process is used during training and subsequently marginalised out for generation, similar to the branch-index tracking process of Billera et al. (2025b), except specialised for the case of a discrete state space.

3 PUSHFORWARD GENERATOR MATCHING

3.1 GENERATOR MATCHING

We briefly recall the Generator Matching framework of Holderrieth et al. (2025), which provides the theoretical foundation for the present work. Let $(\mathcal{X}, d_{\mathcal{X}})$ be a Polish space. A time-inhomogeneous Feller process $(X_t)_{0\le t\le 1}$ on $\mathcal{X}$ is characterised by its infinitesimal generator L_t, which acts on test functions $f \in \mathcal{T}(\mathcal{X}) \subset C(\mathcal{X})$ and determines the evolution of the time-marginals $(p_t)_{0\le t\le 1}$ via the Kolmogorov forward equation (KFE)

$$\partial_t \langle p_t, f\rangle = \langle p_t, L_t f\rangle, \qquad \forall f \in \mathcal{T}(\mathcal{X}), \quad t \in [0,1).$$

This framework encompasses a wide range of generative processes. For instance, when $\mathcal{X} = \mathbb{R}^d$ and $L_t f(x) = b_t(x) \cdot \nabla f(x) + \frac{1}{2}\sigma_t^2(x)\Delta f(x)$, the KFE is the Fokker–Planck equation of a stochastic

differential equation. When $\mathcal{X}$ is a finite set and $L_t f(x) = \sum_{x'} u_t(x' \mid x)[f(x') - f(x)]$, it is the forward equation of a continuous-time Markov chain with rates u_t.

However, one does not need to learn the marginal generator L_t directly. Instead, given a latent random variable Z, a conditional generator L_t^z parametrised by $F_t^z(x)$ for each realisation $Z = z$ is defined and a neural network $F_t^\theta(x)$ is trained against these conditional targets. Training against the conditional loss recovers the correct marginal generator via a gradient equality (Holderrieth et al., 2025, Theorem 1).

3.2 Linear Parametrisations of Infinitesimal Generators

The notion of a linear parametrisation, introduced in Holderrieth et al. (2025) and extended to time- and state-dependent form in Billera et al. (2025a), is central to the training objectives of Generator Matching.

Definition 3.1 (Linear parametrisation of a generator)**.** Let $L_t \colon \mathcal{T}(\mathcal{X}) \to C(\mathcal{X})$ be the generator of a Feller process on a Polish space $\mathcal{X}$. A *(time- and state-dependent) linear parametrisation* of L_t consists of:

(i) for each time t and $x \in \mathcal{X}$, an inner product space $(V_{t,x}, \langle \cdot, \cdot \rangle_{V_{t,x}})$;

(ii) a linear map $\mathcal{K}_{t,x} \colon \mathcal{T}(\mathcal{X}) \to V_{t,x}$;

(iii) a closed convex set $\Omega_{t,x} \subset V_{t,x}$ and a target function $F_t(x) \in \Omega_{t,x}$;

such that for all $f \in \mathcal{T}(\mathcal{X})$ it holds that

$$L_t f(x) = \langle \mathcal{K}_{t,x} f, F_t(x) \rangle_{V_{t,x}}. \tag{1}$$

The map $\mathcal{K}_{t,x}$ encodes the "structural" part of the generator (e.g. finite differences for a CTMC, or differential operators for a diffusion), while $F_t(x)$ captures the learnable parameters (e.g. rates or velocity fields). Training a neural network $F_t^\theta(x)$ to approximate $F_t(x)$ then recovers the generator L_t.

3.3 Marginal and Conditional Generators

Theorem 3.2 (Pushforward Kolmogorov Forward Equation)**.** *Let $(\mathcal{Y}, d_{\mathcal{Y}})$ be a Polish space and let $(Y_t)_{0 \le t \le 1}$ be a time-inhomogeneous Feller process on $\mathcal{Y}$ with infinitesimal generator W_t and time-marginals $(p_t^{\mathcal{Y}})_{0 \le t \le 1}$, satisfying the regularity conditions of (Holderrieth et al., 2025, Appendix A.2) (Assumption A.15). Let $\Phi \colon \mathcal{Y} \to \mathcal{X}$ be a measurable map into a measurable space $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$ satisfying the domain compatibility conditions of Assumption A.16, and let $p_t^{\mathcal{X}} := \Phi_{\#} p_t^{\mathcal{Y}}$. Assume the following integrability condition (Assumption A.18): for every $f \in \mathcal{T}(\mathcal{X})$ and every $t \in [0, 1)$,*

$$\mathbb{E}_{y_t \sim p_t^{\mathcal{Y}}}[|W_t(f \circ \Phi)(y_t)|] < \infty. \tag{2}$$

Define, for each $t \in [0, 1)$ and each test function $f \in \mathcal{T}(\mathcal{X})$,

$$L_t f(x_t) := \mathbb{E}[W_t(f \circ \Phi)(Y_t) \mid \Phi(Y_t) = x_t].$$

Then $p_t^{\mathcal{X}}$ satisfies the Kolmogorov forward equation with generator L_t:

$$\partial_t \langle p_t^{\mathcal{X}}, f \rangle = \langle p_t^{\mathcal{X}}, L_t f \rangle, \qquad \forall f \in \mathcal{T}(\mathcal{X}), \quad t \in [0, 1),$$

where $\langle \mu, g \rangle := \int g \, d\mu$. If additionally the KFE with generator L_t uniquely determines $(p_t^{\mathcal{X}})_{0 \le t \le 1}$ (Assumption A.19), then L_t may be used within the Generator Matching framework of Holderrieth et al. (2025). The proof is given in Appendix A.11; all assumptions are stated formally in Appendix A.10.

Remark 3.3. *Assumption A.19 is non-trivial and a regularity-violating example is given in §3.7.2.*

For many cases we will be concerned with "marginalising" a latent process to learn a base process. We therefore state this special case as a corollary.

Corollary 3.4 (Generator of a projected process)**.** *Specialising Theorem 3.2 to $\mathcal{Y} = \mathcal{X} \times \mathcal{Z}$ and $\Phi = \pi_{\mathcal{X}}$: let (X_t, Z_t) be a Feller process on $\mathcal{X} \times \mathcal{Z}$ with infinitesimal generator W_t, satisfying the hypotheses of Theorem 3.2. Then*

$$L_t f(x_t) := \mathbb{E}[W_t(f \circ \pi_{\mathcal{X}})(X_t, Z_t) \mid X_t = x_t]$$

generates a process on $\mathcal{X}$ with marginal measure $p_t^{\mathcal{X}} = [\pi_{\mathcal{X}}]_{\#} p_t^{\mathcal{X}\times\mathcal{Z}}$, that is, L_t solves the Kolmogorov Forward Equation

$$\partial_t \langle p_t^{\mathcal{X}}, f \rangle = \langle p_t^{\mathcal{X}}, L_t f \rangle, \qquad t \in [0, 1).$$

Remark 3.5 (Compact-$\mathcal{Z}$.)**.** *When $\mathcal{Z}$ is compact, integrability along the $\mathcal{Z}$-fibre is automatic by boundedness, and the domain condition reduces to the $\mathcal{X}$-factor: for flows and diffusions ($\mathcal{D}(W_t) = C_0^k$, $k \in \{1, 2\}$) it suffices that $f \in C_0^k(\mathcal{X})$, and for CTMCs it is automatic. This covers any finite-state latent Z_t (§3.6.1, §3.7.1) and flow matching and diffusion on compact Riemannian manifolds such as* $\mathrm{SO}(3)$.

Remark 3.6. *In Havasi et al. (2025) it is proven that if $u_t(x', z' \mid x_t, z_t)$ is a rate on $\mathcal{X} \times \mathcal{Z}$ that generates $p_t(x, z)$ then*

$$u_t(x' \mid x_t) := \sum_{z' \in \mathcal{Z}} \mathbb{E}_{z_t \sim p_t(\cdot \mid x_t)}[u_t(x', z' \mid x_t, z_t)]$$

generates an x-marginal path $p_t^{\mathcal{X}}(x)$. This result follows as a special case of Corollary 3.4. For a detailed discussion, refer to §3.6.1.

3.4 Latent Process Conditional Generator Matching Loss

In this section we establish which loss functions are valid for training a neural network to learn the pushforward generator L_t of Theorem 3.2. One may train against a *conditional* target which is defined pointwise for each realisation $Y_t = y$ of the latent process and still recover the correct *marginal* generator via an equality of gradients. This generalises the conditional generator matching paradigm of Holderrieth et al. (2025) from conditioning on a static latent variable to conditioning on a latent stochastic process, and subsumes the special cases proven in Havasi et al. (2025); Billera et al. (2025b).

Definition 3.7. (Latent process conditional generator). In the setting of Theorem 3.2, for each $y_t \in \mathcal{Y}$ define the *conditional generator given* $Y_t = y_t$ at the point $x_t := \Phi(y_t)$ by

$$L_t^{y_t} f(x_t) := W_t(f \circ \Phi)(y_t), \qquad f \in \mathcal{T}(\mathcal{X}).$$

This defines $L_t^{y_t} f$ only at the single point $x_t = \Phi(y_t)$, not as a function on all of $\mathcal{X}$. When two distinct $y, y' \in \mathcal{Y}$ satisfy $\Phi(y) = \Phi(y')$, the conditional generators L_t^{y} and $L_t^{y'}$ may assign different values at the same point. This is expected, as they correspond to different conditioning events (since W_t may introduce y-dependence beyond $\Phi(y)$). By the definition of the pushforward generator in Theorem 3.2, the marginal and conditional generators are related by

$$L_t f(x_t) = \mathbb{E}[L_t^{Y_t} f(x_t) \mid \Phi(Y_t) = x_t]. \tag{3}$$

Definition 3.8. (Conditional linear parametrisation). Suppose the conditional generator $L_t^{y_t}$ of Definition 3.7 admits a time- and state-dependent linear parametrisation

$$L_t^{y_t} f(\Phi(y_t)) = \langle \mathcal{K}_{t,\Phi(y_t)} f, F_t^{y_t}(\Phi(y_t)) \rangle_{V_{t,\Phi(y_t)}}$$

where $\mathcal{K}_{t,x_t}$ and V_{t,x_t} are as in the linear parametrisation of L_t and $F_t^{y_t}(\Phi(y_t)) \in \Omega_{t,\Phi(y_t)}$. Then by equation 3 and linearity of the inner product, the marginal generator inherits a linear parametrisation with

$$F_t(x_t) = \mathbb{E}[F_t^{Y_t}(x_t) \mid \Phi(Y_t) = x_t]. \tag{4}$$

Remark 3.9. *Observe that in the above, the linear parametrisation may only depend on the value $x_t := \Phi(y_t)$ for each $y_t \in \mathcal{Y}$.*

Definition 3.10. (Pushforward Conditional Generator Matching Loss). In the setting of Definition 3.8, let $D_{t,x_t} : \Omega_{t,x_t} \times \Omega_{t,x_t} \to \mathbb{R}$ be a Bregman divergence and let F_t^θ be a neural network with $F_t^\theta(x_t) \in \Omega_{t,x_t}$ for all t, x_t. Assume that for all $t \in [0, 1)$,

$$\mathbb{E}_{y_t \sim p_t^{\mathcal{Y}}}[D_{t,\Phi(y_t)}(F_t^{y_t}(\Phi(y_t)), F_t^\theta(\Phi(y_t)))] < \infty.$$

The latent process conditional generator matching loss is

$$L_{\mathrm{cgm}}(\theta) := \mathbb{E}_{t\sim U[0,1], y_t\sim p_t^{\mathcal{Y}}(dy_t)}[D_{t,\Phi(y_t)}(F_t^{y_t}(\Phi(y_t)), F_t^\theta(\Phi(y_t)))].$$

Equivalently, by disintegrating $p_t^{\mathcal{Y}}$ along Φ (Corollary A.13),

$$L_{\mathrm{cgm}}(\theta) = \mathbb{E}_{t\sim U[0,1], x_t\sim p_t^{\mathcal{X}}, y_t\sim p_t^{\mathcal{Y}}(dy_t\mid\Phi(Y_t)=x_t)}[D_{t,x_t}(F_t^{y_t}(x_t), F_t^\theta(x_t))].$$

Definition 3.11. (Marginal Generator Matching Loss). In the setting of Definition 3.8, the *marginal generator matching loss* is

$$L_{\mathrm{gm}}(\theta) := \mathbb{E}_{t\sim U[0,1], x_t\sim p_t^{\mathcal{X}}}[D_{t,x_t}(F_t(x_t), F_t^\theta(x_t))].$$

Theorem 3.12. *(Gradient equality of Conditional and Marginal Generator Matching Losses). Let $L_{\mathrm{cgm}}(\theta), L_{\mathrm{gm}}(\theta)$ be as in Definitions 3.10 and 3.11, and suppose the integrability conditions of Definition 3.10 hold. Then*

$$\nabla_\theta L_{\mathrm{cgm}}(\theta) = \nabla_\theta L_{\mathrm{gm}}(\theta).$$

In particular, training against the conditional loss L_{cgm} (which only requires samples from the process Y_t) recovers the same stationary points as training against the marginal loss L_{gm}, which depends on the generally intractable marginal target F_t.

Proof. See Appendix B.2. □

3.5 Conditioning on a latent process Z_t

In many applications, the process of interest X_t evolves jointly with an auxiliary *latent* process Z_t that is used during training but discarded at generation time. This corresponds to the product-space specialisation of the pushforward framework: we set $\mathcal{Y} = \mathcal{X}\times\mathcal{Z}$, take $\Phi = \pi_{\mathcal{X}}$ to be the canonical projection, and write $Y_t = (X_t, Z_t)$. The pushforward marginals are then $p_t^{\mathcal{X}} = (\pi_{\mathcal{X}})_{\#}p_t^{\mathcal{Y}}$, which is simply the X-marginal of the joint law, and the conditional law $p_t^{\mathcal{Y}}(dy_t \mid x_t)$ from Corollary A.13 reduces to the conditional law of Z_t given $X_t = x_t$, which we denote $p_t(dz \mid x_t)$.

In this setting, the definitions of §3.4 take the following concrete form. The conditional generator (Definition 3.7) for $y_t = (x_t, z_t) \in \mathcal{X}\times\mathcal{Z}$ is

$$L_t^{y_t} f(x_t) = W_t(f\circ\pi_{\mathcal{X}})(x_t, z_t).$$

Remark 3.13. *Since $\Phi = \pi_{\mathcal{X}}$, the evaluation point $\Phi(y_t) = x_t$ is already determined by the projection, so the dependence of the conditional generator on the x_t-component of $y_t = (x_t, z_t)$ is vacuous: the right-hand side $W_t(f\circ\pi_{\mathcal{X}})(x_t, z_t)$ depends on x_t only through the test function, not through the conditioning. Accordingly, we write $L_t^{z_t}$ in place of $L_t^{y_t}$:*

$$L_t^{z_t} f(x_t) := W_t(f\circ\pi_{\mathcal{X}})(x_t, z_t).$$

The marginal generator (Corollary 3.4) is

$$L_t f(x_t) = \mathbb{E}[L_t^{Z_t} f(x_t) \mid X_t = x_t] = \int_{\mathcal{Z}} W_t(f\circ\pi_{\mathcal{X}})(x_t, z) p_t(dz \mid x_t).$$

If the conditional generator admits a linear parametrisation with $F_t^{z_t}(x_t)$ (Definition 3.8), the marginal target is

$$F_t(x_t) = \mathbb{E}[F_t^{Z_t}(x_t) \mid X_t = x_t].$$

The conditional generator matching loss (Definition 3.10) becomes

$$L_{\mathrm{cgm}}(\theta) = \mathbb{E}_{t\sim U[0,1], (x_t,z_t)\sim p_t^{\mathcal{X}\times\mathcal{Z}}}[D_{t,x_t}(F_t^{z_t}(x_t), F_t^\theta(x_t))],$$

and Theorem 3.12 guarantees $\nabla_\theta L_{\mathrm{cgm}}(\theta) = \nabla_\theta L_{\mathrm{gm}}(\theta)$. Thus, one may sample jointly from (X_t, Z_t), use the latent state Z_t to compute the conditional training target $F_t^{Z_t}$, and train a network F_t^θ that receives only X_t as input. Equivalently,

$$L_{\mathrm{cgm}}(\theta) = \mathbb{E}_{t\sim U[0,1], x_t\sim p_t^{\mathcal{X}}(dx), z_t\sim p_t^{\mathcal{Z}}(dz\mid X_t=x_t)}[D_{t,x_t}(F_t^{z_t}(x_t), F_t^\theta(x_t))],$$

Remark 3.14 (Fixed-endpoint conditioning with a latent process)**.** *The conditioning on Z_t is compatible with additionally conditioning on a* static *latent variable Z (e.g. a pair of endpoints) in the sense of Holderrieth et al. (2025). This double conditioning perspective, developed in §3.6, provides a natural sampling procedure for the loss.*

3.6 FIXED-ENDPOINT CONDITIONING WITH A LATENT PROCESS

The conditional generator matching loss of Definition 3.10 requires sampling from the joint law $p_t^{\mathcal{Y}}(dy_t)$. In many practical settings, this joint law is itself constructed by first choosing a static latent variable z (typically an endpoint or pair of endpoints) and then running the process conditionally on z. Indeed, suppose that

$$p_t^{\mathcal{Y}}(dy_t) = \int p^{\mathcal{Y}}(dy_t \mid z) p_{\mathcal{Z}}(dz),$$

where $z \in \mathcal{Z}$ is a static random variable (e.g., points used to steer the conditional path) with law $p_{\mathcal{Z}}$, and $p^{\mathcal{Y}}(dy_t \mid z)$ is the conditional law of the process at time t given z. For each realisation z and each y_t in the support of $p^{\mathcal{Y}}(\cdot \mid z)$, the conditional generator

$$L_t^{z,y_t} f(\Phi(y_t)) := W_t^z (f \circ \Phi)(y_t)$$

depends both on the fixed latent state z and the current process state y_t. This may be viewed from two complementary angles:

(i) *As a latent-process problem (§3.5):* marginalise over z to obtain the law $p_t^{\mathcal{Y}}(dy_t)$. In this case, the conditional generator is $L_t^{y_t}$ and the framework of §3.4 applies directly.

(ii) *As a fixed-endpoint problem with process-valued conditionals:* for each z, marginalise over $y_t \mid z$ to obtain a generator $L_t^z f(x_t) := \mathbb{E}_{y_t \sim p^{\mathcal{Y}}(\cdot|z)}[L_t^{z,y_t} f(x_t)]$ conditioned only on the static latent, recovering the setting of Holderrieth et al. (2025). Marginalising further over z recovers the full marginal generator L_t.

Both viewpoints lead to the same marginal generator and the same gradient equality (Theorem 3.12), but the double-conditioning perspective makes explicit a practical sampling procedure for the conditional loss. Substituting the disintegration into the loss of Definition 3.10 gives

$$L_{\mathrm{cgm}}(\theta) = \mathbb{E}_{t\sim U[0,1], z\sim p_{\mathcal{Z}}(dz), y_t\sim p^{\mathcal{Y}}(dy_t|z)}[D_{t,\Phi(y_t)}(F_t^{z,y_t}(\Phi(y_t)), F_t^\theta(\Phi(y_t)))], \tag{5}$$

which suggests the following training procedure: **(1)** sample fixed latents $z \sim p_{\mathcal{Z}}$; **(2)** given z, sample the latent process state $y_t \sim p^{\mathcal{Y}}(dy_t \mid z)$; **(3)** compute the conditional target F_t^{z,y_t} and evaluate the Bregman divergence.

3.6.1 EDIT FLOWS

We now show how the theory of Edit Flows (Havasi et al., 2025) is recovered as a special case of the pushforward conditional generator matching framework of §3.4–§3.5. Let $\mathcal{X}$ and $\mathcal{Z}$ be finite sets and let (X_t, Z_t) be a CTMC on the joint state space $\mathcal{X} \times \mathcal{Z}$ with rates $u_t(x', z' \mid x_t, z_t) \geq 0$. Applying the joint generator W_t of (X_t, Z_t) to a test function of the form $f \circ \pi_{\mathcal{X}}$ and using Corollary 3.4, one can show that

$$u_t(x' \mid x_t) = \sum_{z'} \mathbb{E}_{z_t\sim p_t(\cdot|x_t)}[u_t(x', z' \mid x_t, z_t)] \qquad \text{generates} \qquad p_t^{\mathcal{X}}(x) = \sum_z p_t(x, z), \tag{6}$$

recovering the first part of Theorem 3.1 of (Havasi et al., 2025). Moreover, the conditional generator admits a linear parametrisation (Definition 3.8) in which the $\mathcal{Z}$-dependence is confined entirely to the target vector $F_t^{z_t}(x_t) = (\tilde{u}_t(x' \mid x_t, z_t))_{x'\in\mathcal{X}}$, where $\tilde{u}_t(x' \mid x_t, z_t) := \sum_{z'} u_t(x', z' \mid x_t, z_t)$ is the total rate of the $\mathcal{X}$-component jumping to x'. The gradient equality of the conditional and marginal losses (the second part of Theorem 3.1 in (Havasi et al., 2025)) then follows directly from Theorem 3.12. The complete derivations are given in Appendix D.

3.7 EXAMPLES

3.7.1 A ONE-DIMENSIONAL EXAMPLE

We illustrate the latent-process framework of §3.5 with a regime-switching diffusion whose conditional paths exhibit discontinuous changes of drift direction, yet whose marginal generator is a standard diffusion generator that can be learnt by a network receiving only the continuous component X_t. Consider a joint process (X_t, Z_t) on $\mathbb{R} \times \{-1, +1\}$, with fixed endpoints $x_0, x_1 \in \mathbb{R}$,

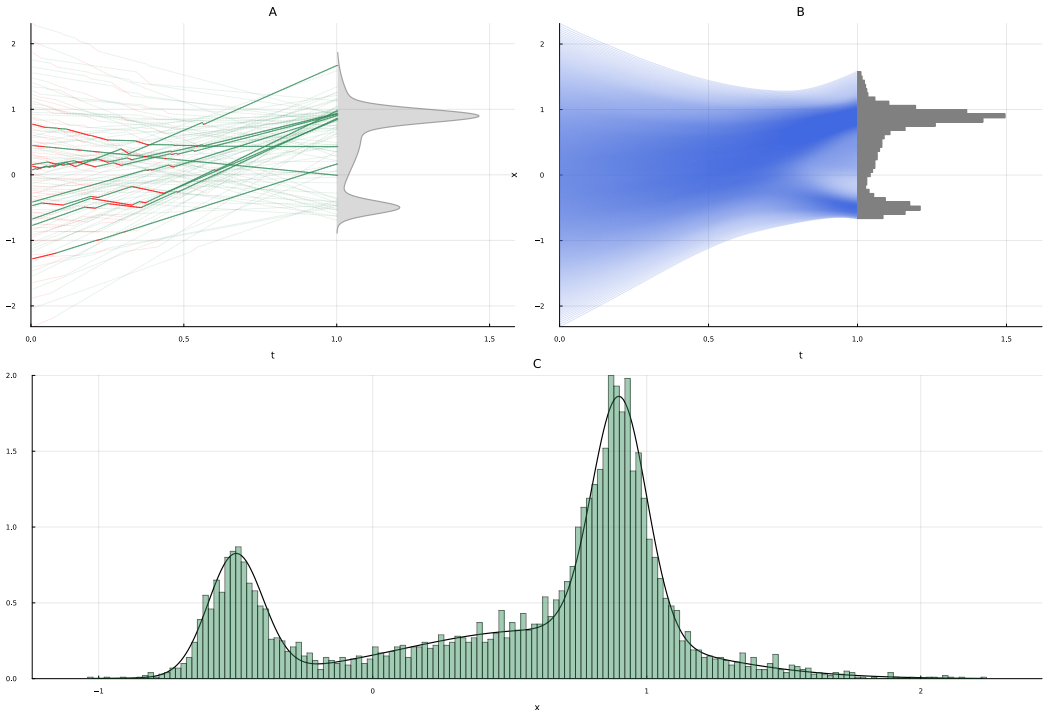

Figure 1: A) Conditional trajectories (training target) with switching, colored by the state of the latent process, where 10 conditional sample paths are foregrounded, B) Model-learnt marginal trajectories, C) Marginal distribution at $t = 1$, with histogram of generated samples matching the target distribution.

evolving according to the following dynamics:

$$Z_t \mid (X_t = x_t, X_0 = x_0, X_1 = x_1) \sim \begin{pmatrix} -\lambda_1(x_t, t) & \lambda_1(x_t, t) \\ \lambda_2(x_t, t) & -\lambda_2(x_t, t) \end{pmatrix};$$

$$X_t \mid (Z_t = z_t, X_0 = x_0, X_1 = x_1) \sim \frac{z_t x_1 - X_t}{1 - t} dt + \sigma_t dB_t.$$

where B_t is a standard Brownian motion. That is, conditional on the continuous component X_t and the endpoints, Z_t evolves as a CTMC with rates $\lambda_1(x_t, t), \lambda_2(x_t, t)$. Conditional on the CTMC value and the endpoints, X_t evolves as a Brownian bridge towards $z_t x_1$, that is, towards the target x_1 when $Z_t = +1$ and towards $-x_1$ when $Z_t = -1$. The rates λ_1, λ_2 are chosen so that $X_1 = x_1$ a.s. in the conditional paths (see Appendix C for the construction). See Figure 1 A for representative trajectories.

Since Z_t and X_t evolve in a dependent manner, this falls outside the scope of Holderrieth et al. (2025), which treats conditioning on a static latent variable Z. By Corollary 3.4, the marginal generator on $\mathbb{R}$ is nonetheless well-defined, and by Theorem 3.12 a network that receives only X_t may be trained against the conditional loss. This is an instance of the double conditioning of §3.6: the static latent is the pair of endpoints (x_0, x_1) and the process-valued latent is z_t. In Appendix C we derive the conditional generator and construct an x_1-prediction linear parametrisation (Definition 3.8). Taking the Bregman divergence to be the squared Euclidean norm, the training loss reduces to

$$\mathbb{E}_{t \sim U[0,1], (x_0, x_1) \sim q, (x_t, z_t) \sim p_t(\cdot \mid x_0, x_1)} [\| z_t x_1 - x_1^\theta(x_t) \|^2],$$

where the latent state z_t appears in the target but is not passed to the network. At generation time, the velocity is recovered via $u_t^\theta(x) = (x_1^\theta(x) - x)/(1 - t)$, with no reference to the latent dynamics. Figure 1 confirms that the learnt marginal paths are smooth, and the switching dynamics of the conditional training target are effectively integrated out, while still retaining the property that the marginal distribution at $t = 1$ matches the target.

3.7.2 A REGULARITY-VIOLATING EXAMPLE

One may ask whether regularity of the base process Y_t and smoothness of the map Φ suffice to guarantee that the pushforward process $\Phi(Y_t)$ satisfies the KFE sufficiency condition (Assumption A.19). The following example shows that this need not be the case, even for $\Phi \in C^\infty$.

Example 3.15. Define $\Phi\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ by

$$\Phi(x) := \begin{cases} e^{-1/x} & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

This function belongs to $C^\infty(\mathbb{R})$. Let $(B_t)_{t\geq 0}$ be a standard Brownian motion, so that $(B_t, \frac{1}{2}\partial_{xx})$ satisfies the Generator Matching regularity conditions of Holderrieth et al. (2025). The pushforward $\Phi(B_t)$ has an atom at 0 for every $t > 0$, since $\mathbb{P}[\Phi(B_t) = 0] = \mathbb{P}[B_t \leq 0] = \frac{1}{2}$.

The corresponding marginal generator L_t satisfies $L_t f(0) = 0$ for every test function f. In Appendix E, we describe a one-parameter family of probability paths satisfying the KFE, violating the uniqueness assumption on the KFE.

4 NON-PROJECTION PUSHFORWARDS

All examples in §3.7 and in the prior work discussed in §2 take $\Phi = \pi_{\mathcal{X}}$ to be a canonical projection. [1]

Protein complexes are formed by multiple 'chains', where residues within a chain are covalently linked. We might wish to train a model via a conditional path that respects this structure. Each residue in a protein backbone can be modelled by a point in $\mathbb{R}^3$ and an orientation in $\mathrm{SO}(3)$ (Jumper et al., 2021). Further, let each chain have a shared orientation $R_{\mathrm{ch},t} \in \mathrm{SO}(3)$, a centroid $\mu_t \in \mathbb{R}^3$, per-residue internal rotations $R^{(i)}_{\mathrm{res},t} \in \mathrm{SO}(3)$, and per-residue centroid-relative displacements $p^{(i)}_t \in \mathbb{R}^3$. For each chain, the latent space is

$$\mathcal{Y} = \mathrm{SO}(3) \times \mathbb{R}^3 \times \mathrm{SO}(3)^n \times (\mathbb{R}^3)^n,$$

and the map to observed (world-frame) coordinates is $\Phi\colon \mathcal{Y} \to \mathrm{SO}(3)^n \times (\mathbb{R}^3)^n$,

$$\Phi(R_{\mathrm{ch},t}, \mu_t, (R^{(i)}_{\mathrm{res},t})_i, (p^{(i)}_t)_i) = (R_{\mathrm{ch},t} R^{(i)}_{\mathrm{res},t}, R_{\mathrm{ch},t} p^{(i)}_t + \mu_t)_i.$$

The chain rotation $R_{\mathrm{ch},t}$ acts on the internal geometry (orientations and displacements) while leaving the centroid μ_t invariant. When Y_t follows an SDE on $\mathcal{Y}$, the pushforward through Φ induces correlated rotational and translational noise at the observed level: all residue orientations are simultaneously rotated, and all positions are coherently tumbled about the centroid. This captures chain-level rigid-body motion, while the per-residue components $R^{(i)}_{\mathrm{res},t}$ and $p^{(i)}_t$ handle internal flexibility. Preliminary calculations show that the resulting conditional generator admits a velocity-only parametrisation in the sense of Definition 3.8, with all diffusion coefficients determined by the observed state.

5 DISCUSSION

Here we have expanded Generator Matching to allow for conditional processes with latent time-dependent variables with rich state spaces and dynamics. Recent examples have demonstrated the utility of discrete-valued time-dependent latent variables, and we anticipate the value of an expanded latent state space. In future work we plan to attempt to relax the regularity conditions for this, deepen our understanding of when they might be violated in real practical examples, and to develop and implement chain-level rotation example from section 4.

[1] Any pushforward $\Phi\colon \mathcal{Y} \to \mathcal{X}$ can formally be subsumed by the projection framework: one embeds $Y_t \mapsto (\Phi(Y_t), Y_t) \in \mathcal{X} \times \mathcal{Y}$ and projects onto the first coordinate. However, this reduction is artificial. The "latent" component becomes the entire original process Y_t, the joint law on $\mathcal{X} \times \mathcal{Y}$ is degenerate, and the geometric structure of Φ is obscured.

6 ACKNOWLEDGEMENTS

This project received support from the Swedish Research Council (2024-00390 and 2023-02516) and the Knut and Alice Wallenberg Foundation (2024.0039) to B.M.

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2025. URL https://arxiv.org/abs/2303.08797.

Lukas Billera, Hedwig Nora Nordlinder, and Ben Murrell. Time dependent loss reweighting for flow matching and diffusion models is theoretically justified, 2025a. URL https://arxiv.org/abs/2511.16599.

Lukas Billera, Hedwig Nora Nordlinder, Jack Collier Ryder, Anton Oresten, Aron Stålmarck, Theodor Mosetti Björk, and Ben Murrell. Branching flows: Discrete, continuous, and manifold flow matching with splits and deletions, 2025b.

Vladimir I. Bogachev. *Measure Theory*. Measure Theory ; 1. Springer Berlin Heidelberg, Berlin, Heidelberg, 1st ed. 2007. edition, 2007. ISBN 1-280-74570-3.

Marton Havasi, Brian Karrer, Itai Gat, and Ricky T. Q. Chen. Edit flows: Flow matching with edit operations, 2025. URL https://arxiv.org/abs/2506.09018.

Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky T. Q. Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary markov processes, 2025. URL https://arxiv.org/abs/2410.20587.

Tariq Berrada Ifriqi, John Nguyen, Karteek Alahari, Jakob Verbeek, and Ricky T. Q. Chen. Flowception: Temporally expansive flow matching for video generation, 2025.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Jaeyeon Kim, Lee Cheuk-Kit, Carles Domingo-Enrich, Yilun Du, Sham Kakade, Timothy Ngotiaoco, Sitan Chen, and Michael Albergo. Any-order flexible length masked diffusion, 2025. URL https://arxiv.org/abs/2509.01025.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code, 2024. URL https://arxiv.org/abs/2412.06264.

John Nguyen, Marton Havasi, Tariq Berrada, Luke Zettlemoyer, and Ricky T. Q. Chen. Oneflow: Concurrent mixed-modal and interleaved generation with edit flows, 2025.

A PROOF OF THE PUSHFORWARD KOLMOGOROV FORWARD EQUATION

In this appendix we collect the measure-theoretic background needed to rigorously define the marginal generator $L_t f(x_t) := \mathbb{E}[W_t(f \circ \Phi)(Y_t) \mid \Phi(Y_t) = x_t]$ and to prove that it governs the Kolmogorov forward equation for the pushforward measure $p_t^{\mathcal{X}} = \Phi_{\#} p_t^{\mathcal{Y}}$. All results in §A.1–§A.8 are standard and are cited from Bogachev (2007). We reproduce precise statements here for completeness.

Remark A.1 (Notation)**.** *Our notation differs from Bogachev (2007) in two respects.*

Conditional expectation. *We use the standard notation $\mathbb{E}[f \mid \mathscr{B}]$ and $\mathbb{E}[f \mid \eta]$ for the conditional expectation of f with respect to a sub-σ-algebra $\mathscr{B}$ or the σ-algebra generated by a random variable η, respectively. In Bogachev (2007) the same object is written $\mathbb{E}^{\mathscr{B}} f$, or $\mathbb{E}_{\mu}^{\mathscr{B}} f$.*

Markov kernels. *In Bogachev (2007), a transition probability is written $P(\cdot \mid \cdot)\colon X_1 \times \mathscr{B}_2 \to \mathbb{R}$ with the convention that $P(x \mid B)$ places the conditioning point x in the first slot and the measurable set B in the second. We instead write $\kappa(B \mid x)$, which emphasises the conditional-probability interpretation: $\kappa(\cdot \mid x)$ is a probability measure on $\mathscr{B}_2$ for each fixed x, while $\kappa(B \mid \cdot)$ is a $\mathscr{B}_1$-measurable function for each fixed B. Integrals against the kernel are written $\int f(y)\kappa(dy \mid x)$.*

A.1 CONDITIONAL EXPECTATION: DEFINITION AND EXISTENCE

Definition A.2. (Bogachev, 2007, Def. 2.12.2) Let $\mathcal{F}$ be some collection of functions in X that have the same codomain. The smallest σ-algebra on X in which all functions belonging to $\mathcal{F}$ are measurable is called the σ-algebra generated by $\mathcal{F}$ and is denoted by $\sigma(\mathcal{F})$ (formally, this can be defined as the intersection of all σ-algebras for which each function in $\mathcal{F}$ is measurable). In particular, the σ-algebra generated by a random variable η is the smallest σ-algebra in which η is measurable and is denoted by $\sigma(\eta)$.

Definition A.3 (Conditional expectation (Bogachev, 2007, Def. 10.1.1))**.** Let $(\Omega, \mathscr{A}, \mu)$ be a probability space and let $\mathscr{B} \subset \mathscr{A}$ be a sub-σ-algebra. For $f \in L^1(\mu)$, a *conditional expectation of f given $\mathscr{B}$* is a $\mathscr{B}$-measurable, μ-integrable function $\mathbb{E}[f \mid \mathscr{B}]$ satisfying

$$\int_{\Omega} g f d\mu = \int_{\Omega} g \mathbb{E}[f \mid \mathscr{B}] d\mu \tag{7}$$

for every bounded $\mathscr{B}$-measurable function g. Equivalently,

$$\int_{B} f d\mu = \int_{B} \mathbb{E}[f \mid \mathscr{B}] d\mu, \qquad \forall B \in \mathscr{B}. \tag{8}$$

When $\mathscr{B} = \sigma(\eta)$ is generated by a measurable mapping (i.e. random variable) η, we write $\mathbb{E}[f \mid \eta]$ in place of $\mathbb{E}[f \mid \sigma(\eta)]$.

Theorem A.4 (Existence and basic properties (Bogachev, 2007, Thm. 10.1.5 (1)-(2)))**.** *Let μ be a probability measure on $(\Omega, \mathscr{A})$ and let $\mathscr{B} \subset \mathscr{A}$ be a sub-σ-algebra. For every $f \in L^1(\mu)$ there exists a $\mathscr{B}$-measurable function $\mathbb{E}[f \mid \mathscr{B}]$, unique μ-a.e., such that:*

(1) *$\mathbb{E}[f \mid \mathscr{B}]$ is a conditional expectation of f given $\mathscr{B}$ in the sense of Definition A.3.*

(2) *If f is already $\mathscr{B}$-measurable and integrable, then $\mathbb{E}[f \mid \mathscr{B}] = f$ μ-a.e.*

A.2 THE TOWER PROPERTY

Proposition A.5 (Tower property (Bogachev, 2007, Eq. (10.1.4)))**.** *Let $\mathscr{B}_1 \subset \mathscr{B} \subset \mathscr{A}$ be sub-σ-algebras. Then, for every $f \in L^1(\mu)$,*

$$\mathbb{E}[\mathbb{E}[f \mid \mathscr{B}] \mid \mathscr{B}_1] = \mathbb{E}[f \mid \mathscr{B}_1] = \mathbb{E}[\mathbb{E}[f \mid \mathscr{B}_1] \mid \mathscr{B}], \qquad \mu\text{-a.e.} \tag{9}$$

A.3 THE DOOB–DYNKIN LEMMA

The conditional expectation $\mathbb{E}[f \mid \eta]$ is defined abstractly as any $\sigma(\eta)$-measurable random variable satisfying the integral identity $\int_B f d\mu = \int_B \mathbb{E}[f \mid \mathscr{B}] d\mu$, for all $B \in \mathscr{B}$. This means $\mathbb{E}[f \mid \eta]$

is determined only up to a set of measure zero, and *a priori* it is not clear that it can be written as a deterministic function of the observed value of η. The Doob–Dynkin lemma resolves this by characterising *exactly* which random variables are $\sigma(\eta)$-measurable (they are precisely the Borel functions of η).

Theorem A.6 (Doob–Dynkin (Bogachev, 2007, Thm. 2.12.3))**.** *Let I be a countable index set and $\{f_i\}_{i\in I}$ be a family of measurable functions on a nonempty space X. A function g on X is measurable with respect to $\sigma(\{f_i\}_{i\in I})$ if and only if there exists a Borel-measurable function $\psi\colon \mathbb{R}^I \to \mathbb{R}$ such that*

$$g(x) = \psi(f_{i_1}(x), f_{i_2}(x), \ldots).$$

The consequence for conditional expectation is immediate. Since $\mathbb{E}[f \mid \eta]$ is $\sigma(\eta)$-measurable by definition, the theorem (applied to the single-function family $\{f_i\} = \{\eta\}$) guarantees the existence of a Borel-measurable function $h_f\colon Y \to \mathbb{R}$ such that

$$\mathbb{E}[f \mid \eta](\omega) = h_f(\eta(\omega)), \qquad \mathbb{P}\text{-a.e. } \omega.$$

In words: the conditional expectation, which is *a priori* just some random variable on Ω, is in fact a deterministic function of the observed value $\eta(\omega)$. This is what justifies the pointwise notation

$$\mathbb{E}[f \mid \eta = x] := h_f(x),$$

which evaluates the conditional expectation at a specific value $x \in Y$ of the conditioning variable. The function h_f is unique up to redefinition on a set of $(\mathbb{P} \circ \eta^{-1})$-measure zero.

A.4 Transition probabilities (Markov kernels)

Definition A.7 (Transition probability (Bogachev, 2007, Def. 10.7.1))**.** Let $(X_1, \mathscr{B}_1)$ and $(X_2, \mathscr{B}_2)$ be measurable spaces. A *transition probability* (or *Markov kernel*) for this pair is a function $\kappa\colon \mathscr{B}_2 \times X_1 \to [0, 1]$ such that:

(i) for every fixed $x \in X_1$, the map $B \mapsto \kappa(B \mid x)$ is a probability measure on $\mathscr{B}_2$;

(ii) for every fixed $B \in \mathscr{B}_2$, the map $x \mapsto \kappa(B \mid x)$ is $\mathscr{B}_1$-measurable.

A.5 The kernel Fubini theorem

Theorem A.8 (Kernel Fubini (Bogachev, 2007, Thm. 10.7.2))**.** *Let κ be a transition probability for $(X_1, \mathscr{B}_1)$ and $(X_2, \mathscr{B}_2)$, and let ν be a probability measure on $\mathscr{B}_1$. Then there exists a unique probability measure μ on $(X_1 \times X_2, \mathscr{B}_1 \otimes \mathscr{B}_2)$ satisfying*

$$\mu(B_1 \times B_2) = \int_{B_1} \kappa(B_2 \mid x)\nu(dx), \qquad \forall B_1 \in \mathscr{B}_1, B_2 \in \mathscr{B}_2. \tag{10}$$

Moreover, for every $f \in L^1(\mu)$, the iterated integral is well defined and

$$\int_{X_1\times X_2} f(x_1, x_2)\mu(d(x_1, x_2)) = \int_{X_1} \left(\int_{X_2} f(x_1, x_2)\kappa(dx_2 \mid x_1) \right) \nu(dx_1). \tag{11}$$

Remark A.9. *In the above, recall that $\mathscr{B}_1 \otimes \mathscr{B}_2$ is the product σ-algebra generated by the collection of rectangles $\{B_1 \times B_2 : B_1 \in \mathscr{B}_1, B_2 \in \mathscr{B}_2\}$.*

A.6 Existence of the conditional law as a transition probability

Theorem A.10 (Regular conditional distribution (Bogachev, 2007, Ex. 10.7.5))**.** *Let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probability space and let S be a Souslin space (that is, S is the image of a Polish (complete, separable, metrisable) space under a continuous mapping) and let $(Y, \mathscr{A}_Y)$ be a measurable space, and let*

$$\xi\colon (\Omega, \mathscr{A}) \to (S, \mathscr{B}(S)), \qquad \eta\colon (\Omega, \mathscr{A}) \to (Y, \mathscr{A}_Y)$$

be measurable mappings (i.e., ξ and η are random variables taking values in S and Y, respectively). Then there exists a transition probability (Markov kernel)

$$\kappa\colon \mathscr{B}(S) \times Y \to [0, 1], \qquad (B, y) \longmapsto \kappa(B \mid y),$$

such that for every $B \in \mathscr{B}(S)$,

$$\mathbb{P}(\xi \in B \mid \eta) = \kappa(B \mid \eta), \qquad \mathbb{P}\text{-a.e.} \tag{12}$$

Moreover, the family $\{\kappa(\cdot \mid y)\}_{y\in Y}$ is unique up to modification on a set of $(\mathbb{P} \circ \eta^{-1})$-measure zero.

A.7 AGREEMENT OF THE KERNEL AND σ-ALGEBRA DEFINITIONS

The σ-algebra formalism produces the conditional expectation $\mathbb{E}[f(\xi) \mid \eta]$ as an abstract $\sigma(\eta)$-measurable random variable, determined only $\mathbb{P}$-a.e. The Doob–Dynkin lemma (Theorem A.6) guarantees a Borel function h_f with $\mathbb{E}[f(\xi) \mid \eta](\omega) = h_f(\eta(\omega))$, and one *defines* $\mathbb{E}[f(\xi) \mid \eta = y] := h_f(y)$. On the other hand, the kernel from Theorem A.10 gives a pointwise formula $y \mapsto \int_S f d\kappa(\cdot \mid y)$. The following proposition, which is a special case of a general result connecting regular conditional measures to conditional expectations (Bogachev, 2007, Prop. 10.4.18), shows that the two coincide.

Proposition A.11 (Kernel representation of conditional expectation (Bogachev, 2007, Prop. 10.4.18, Eq. (10.4.13))). *In the setting of Theorem A.10, for every $f \in L^1(\mathbb{P})$ of the form $f = \varphi \circ \xi$ with φ a Borel function on S, one has*

$$\mathbb{E}[\varphi(\xi) \mid \eta = y] = \int_S \varphi(s)\kappa(ds \mid y), \qquad (\mathbb{P} \circ \eta^{-1})\text{-a.e. } y. \tag{13}$$

This is a consequence of the integral identity for regular conditional measures (Bogachev, 2007, Thm. 10.4.5).

A.8 PUSHFORWARD AND CHANGE OF VARIABLES

Theorem A.12 (Change of variables / pushforward (Bogachev, 2007, Thm. 3.6.1)). *Let μ be a non-negative measure on $(X, \mathscr{A})$ and $\Phi\colon X \to Y$ an $(\mathscr{A}, \mathscr{B})$-measurable mapping. A $\mathscr{B}$-measurable function g on Y is integrable with respect to the image (pushforward) measure $\Phi_{\#}\mu := \mu \circ \Phi^{-1}$ if and only if $g \circ \Phi$ is μ-integrable. In that case,*

$$\int_Y g(y)(\Phi_{\#}\mu)(dy) = \int_X g(\Phi(x))\mu(dx). \tag{14}$$

A.9 DISINTEGRATION OF AN INTEGRAL VIA A MARKOV KERNEL

Combining the preceding results yields the following identity, which is the key computational device in the proof of Theorem 3.2. Suppose $\xi\colon \Omega \to S$ is a random variable in a Souslin space, $\eta := \Phi \circ \xi\colon \Omega \to \mathcal{X}$ for a Borel map Φ, and κ is the regular conditional distribution of ξ given η (Theorem A.10). Then for any $\mathbb{P}$-integrable function of the form $\varphi(\xi)$, the pushforward formula (Theorem A.12) and the kernel Fubini theorem (Theorem A.8) give

$$\int_\Omega \varphi(\xi) d\mathbb{P} = \int_{\mathcal{X}} \left(\int_S \varphi(y)\kappa(dy \mid x) \right) (\Phi_{\#}p_\xi)(dx), \tag{15}$$

where $p_\xi := \mathbb{P} \circ \xi^{-1}$ denotes the law of ξ.

We now specialise the above to the setting of Theorem 3.2.

Corollary A.13 (Conditional law of Y_t given $\Phi(Y_t)$). *Let $(\mathcal{Y}, d_{\mathcal{Y}})$ be a Polish space, let $\Phi\colon \mathcal{Y} \to \mathcal{X}$ be a $(\mathscr{B}(\mathcal{Y}), \mathscr{B}(\mathcal{X}))$-measurable map, and let $(Y_t)_{0\le t\le 1}$ be a stochastic process on $\mathcal{Y}$ with time-marginals $(p_t^{\mathcal{Y}})_{0\le t\le 1}$. For each $t \in [0,1]$, define $p_t^{\mathcal{X}} := \Phi_{\#}p_t^{\mathcal{Y}}$. Then:*

(i) *Since $\mathcal{Y}$ is Polish (hence Souslin), Theorem A.10 provides a Markov kernel $\kappa_t\colon \mathscr{B}(\mathcal{Y}) \times \mathcal{X} \to [0,1]$ such that*

$$\mathbb{P}(Y_t \in B \mid \Phi(Y_t)) = \kappa_t(B \mid \Phi(Y_t)), \qquad \mathbb{P}\text{-a.e.}, \quad \forall B \in \mathscr{B}(\mathcal{Y}).$$

We denote this kernel by

$$p_t^{\mathcal{Y}}(dy_t \mid x_t) := \kappa_t(dy_t \mid x_t), \tag{16}$$

so that $p_t^{\mathcal{Y}}(\cdot \mid \Phi(Y_t) = x_t)$ is the conditional law of Y_t given $\Phi(Y_t) = x_t$. This can also be written

$$p_t^{\mathcal{Y}}(dy_t \mid \Phi(Y_t) = x_t) := p_t^{\mathcal{Y}}(dy \mid x_t).$$

(ii) *For any $p_t^{\mathcal{Y}}$-integrable function $\varphi\colon \mathcal{Y} \to \mathbb{R}$, the disintegration identity in equation 15 gives*

$$\int_{\mathcal{Y}} \varphi(y_t) p_t^{\mathcal{Y}}(dy_t) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \varphi(y_t) p_t^{\mathcal{Y}}(dy_t \mid x_t) \right) p_t^{\mathcal{X}}(dx_t). \tag{17}$$

(iii) *By Proposition A.11, the kernel integral and the conditional expectation agree:*

$$\int_{\mathcal{Y}} \varphi(y_t) p_t^{\mathcal{Y}}(dy_t \mid x_t) = \mathbb{E}[\varphi(Y_t) \mid \Phi(Y_t) = x_t], \qquad p_t^{\mathcal{X}}\text{-a.e. } x_t. \tag{18}$$

A.10 HYPOTHESES

We collect here the standing assumptions for Theorem 3.2. We first fix notation for the relevant function spaces.

Definition A.14 (Function spaces)**.** Let (S, d) be a metric space.

(i) $C_0(S)$ denotes the space of continuous functions $f\colon S \to \mathbb{R}$ that *vanish at infinity*: for every $\varepsilon > 0$ there exists a compact set $K \subset S$ such that $|f(s)| < \varepsilon$ for all $s \notin K$. When S is compact, $C_0(S) = C(S)$.

(ii) For $k \in \{1, 2, \dots, \infty\}$ and S a smooth manifold (possibly with boundary), $C_0^k(S)$ denotes the space of C^k functions that, together with all derivatives up to order k, vanish at infinity. When S is compact, $C_0^k(S) = C^k(S)$.

Assumption A.15 (Generator Matching on $\mathcal{Y}$)**.** $(\mathcal{Y}, d_{\mathcal{Y}})$ is a Polish space. $(Y_t)_{0\le t\le 1}$ is a time-inhomogeneous Feller process on $\mathcal{Y}$ with infinitesimal generator W_t, time-marginals $(p_t^{\mathcal{Y}})_{0\le t\le 1}$, and test-function space $\mathcal{D}(W_t) \subset C_0(\mathcal{Y})$, satisfying the Generator Matching regularity conditions of (Holderrieth et al., 2025, Appendix A.2, Assumptions 1–5).

Assumption A.16 (Pushforward map)**.** $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$ is a measurable space equipped with a measure-determining linear space of bounded test functions $\mathcal{T}(\mathcal{X})$. The map $\Phi\colon \mathcal{Y} \to \mathcal{X}$ is $(\mathscr{B}(\mathcal{Y}), \mathscr{B}(\mathcal{X}))$-measurable, and for every $f \in \mathcal{T}(\mathcal{X})$ and every $t \in [0, 1)$:

(i) $f \circ \Phi \in \mathcal{D}(W_t)$;

(ii) the map $t \mapsto W_t(f \circ \Phi)$ is continuous in $\|\cdot\|_\infty$ on $[0, 1)$.

Remark A.17. *Assumption A.16 places conditions on the projection Φ from $\mathcal{Y}$ to $\mathcal{X}$, and it does not restrict the choice of neural network architecture or loss.*

Assumption A.18 (Integrability)**.** For every $f \in \mathcal{T}(\mathcal{X})$ and every $t \in [0, 1)$, the generator's action on the lifted test function is integrable:

$$\mathbb{E}_{y_t \sim p_t^{\mathcal{Y}}}[|W_t(f \circ \Phi)(y_t)|] < \infty. \tag{19}$$

If the conditional generator admits a linear parametrisation (Definition 3.8) with target $F_t^{y_t}(\Phi(y_t))$, we additionally require

$$\mathbb{E}_{y_t \sim p_t^{\mathcal{Y}}}[\|F_t^{y_t}(\Phi(y_t))\|_{V_{t,\Phi(y_t)}}] < \infty, \qquad t \in [0, 1). \tag{20}$$

Assumption A.19 (KFE sufficiency on $\mathcal{X}$)**.** The KFE with the marginal generator $L_t f$ of Theorem 3.2 uniquely determines the probability path on $\mathcal{X}$: if $(q_t)_{0\le t\le 1}$ is any probability path with $q_0 = p_0^{\mathcal{X}} := \Phi_{\#} p_0^{\mathcal{Y}}$ and $\partial_t \langle q_t, f\rangle = \langle q_t, L_t f\rangle$ for all $f \in \mathcal{T}(\mathcal{X})$ and $t \in [0, 1)$, then $q_t = \Phi_{\#} p_t^{\mathcal{Y}}$ for all $t \in [0, 1]$.

We remark that Assumption A.19 is non-trivial and we give an example of it being violated in §3.7.2, despite $\Phi \in C^\infty$ and Y_t simply being a Brownian motion. Informally, KFE insufficiency may occur when the first-order dynamics of the process (such as drift, diffusion, and instantaneous rate of jumps in the Euclidean case) do not uniquely determine the marginal dynamics — that is, when there exists a stochastic process whose instantaneous dynamics are indistinguishable from the target process in $\mathcal{X}$ but which does not have the desired marginal distribution. This is exactly the failure mode of Example 3.7.2. We note, however, that this hypothesis is agnostic to the dimensionality of $\mathcal{X}$ and $\mathcal{Y}$. For further discussion of uniqueness of the KFE, we direct the reader to Generator Matching (Holderrieth et al., 2025), Appendix A.2.

Discussion of the hypotheses. Assumption A.15 is simply the statement that (Y_t, W_t) is a Generator Matching process on $\mathcal{Y}$. The proof of Theorem 3.2 uses Assumptions A.15–A.18. Assumption A.19 is not needed for the KFE itself but is required in order to *use* the marginal generator L_t within the Generator Matching framework.

Domain compatibility. Since $\mathcal{D}(W_t) \subset C_0(\mathcal{Y})$, condition (i) requires both that $f \circ \Phi$ vanish at infinity and that $f \circ \Phi$ belong to the domain of W_t, which may impose further regularity constraints on Φ, e.g., differentiability.

Compactness. When $\mathcal{Y}$ is compact, $C_0(\mathcal{Y}) = C(\mathcal{Y})$ and $C_0^k(\mathcal{Y}) = C^k(\mathcal{Y})$, so the vanishing-at-infinity requirement on $f \circ \Phi$ is vacuous The remaining question is whether $f \circ \Phi$ has sufficient regularity to belong to $\mathcal{D}(W_t)$. For many standard generators, $\mathcal{D}(W_t) = C_0^k(\mathcal{Y}) = C^k(\mathcal{Y})$ for some k (e.g. $k = 2$ for second-order differential operators, $k = 0$ for rate matrices), so it suffices that Φ and f are C^k. Moreover, integrability (Assumption A.18) is automatic on a compact space, since continuous functions are bounded and integration against a probability measure is finite. The most important special case in which compactness simplifies the verification is the product-space setting $\mathcal{Y} = \mathcal{X} \times \mathcal{Z}$, $\Phi = \pi_{\mathcal{X}}$ (see §3.5). When $\mathcal{Z}$ is compact, the domain compatibility and integrability conditions along the $\mathcal{Z}$-fibre are automatic for continuous integrands. This covers, for example, the case of a continuous main process X_t jointly evolving with a finite-state latent CTMC Z_t (as in the example of §3.7.1), since any finite set is compact. It also covers flow matching and diffusion on compact Riemannian manifolds, such as $\mathrm{SO}(3)$.

Non-compact state spaces. When $\mathcal{Y}$ is non-compact, condition (i) requires $f \circ \Phi \in \mathcal{D}(W_t) \subset C_0(\mathcal{Y})$, so in particular $f \circ \Phi$ must vanish at infinity. This is a genuine restriction on Φ. It holds whenever Φ is proper (preimages of compact sets are compact). However, it may fail for general maps such as projections $\pi_{\mathcal{X}} : \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ when $\mathcal{Z}$ is non-compact.

A.11 Proof of Theorem 3.2

Proof. By the change-of-variables formula for the pushforward (Theorem A.12),

$$\langle p_t^{\mathcal{X}}, f \rangle = \int_{\mathcal{X}} f(x_t)(\Phi_{\#} p_t^{\mathcal{Y}})(dx_t) = \int_{\mathcal{Y}} f(\Phi(y_t)) p_t^{\mathcal{Y}}(dy_t) = \langle p_t^{\mathcal{Y}}, f \circ \Phi \rangle. \tag{21}$$

By Assumption A.16(i), $f \circ \Phi \in \mathcal{D}(W_t)$. Since Y_t is generated by W_t and has time-marginals $p_t^{\mathcal{Y}}$, the KFE holds

$$\partial_t \langle p_t^{\mathcal{Y}}, f \circ \Phi \rangle = \langle p_t^{\mathcal{Y}}, W_t(f \circ \Phi) \rangle = \int_{\mathcal{Y}} W_t(f \circ \Phi)(y_t) p_t^{\mathcal{Y}}(dy_t). \tag{22}$$

Applying Corollary A.13(ii) with $\varphi := W_t(f \circ \Phi)$, noting that φ is $p_t^{\mathcal{Y}}$-integrable by Assumption A.18, we have

$$\int_{\mathcal{Y}} W_t(f \circ \Phi)(y_t) p_t^{\mathcal{Y}}(dy_t) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} W_t(f \circ \Phi)(y_t) p_t^{\mathcal{Y}}(dy_t \mid x_t) \right) p_t^{\mathcal{X}}(dx_t). \tag{23}$$

By Corollary A.13(iii), the inner integral in equation 23 is precisely the conditional expectation evaluated at the point x_t:

$$\int_{\mathcal{Y}} W_t(f \circ \Phi)(y_t) p_t^{\mathcal{Y}}(dy_t \mid x_t) = \mathbb{E}[W_t(f \circ \Phi)(Y_t) \mid \Phi(Y_t) = x_t] = L_t f(x_t). \tag{24}$$

Combining equations 21–24:

$$\partial_t \langle p_t^{\mathcal{X}}, f \rangle = \int_{\mathcal{X}} L_t f(x_t) p_t^{\mathcal{X}}(dx_t) = \langle p_t^{\mathcal{X}}, L_t f \rangle. \qquad \square$$

A.11.1 Alternative proof via the tower property.

We also remark that the tower property (Proposition A.5) gives a shorter argument that avoids explicit use of kernels. Start from the weak KFE for Y_t evaluated against $f \circ \Phi \in \mathcal{T}^W$:

$$\partial_t \mathbb{E}[(f \circ \Phi)(Y_t)] = \mathbb{E}[W_t(f \circ \Phi)(Y_t)]$$

The left-hand side equals $\partial_t \int f \, dp_t^{\mathcal{X}}$, since $p_t^{\mathcal{X}} = \Phi_{\#} p_t^{\mathcal{Y}}$. For the right-hand side, apply the tower property

$$\begin{aligned} \mathbb{E}[W_t(f \circ \Phi)(Y_t)] &= \mathbb{E}[\mathbb{E}[W_t(f \circ \Phi)(Y_t) \mid \Phi(Y_t)]] \\ &= \mathbb{E}[\mathcal{L}_t f(\Phi(Y_t))] \\ &= \int \mathcal{L}_t f \, dp_t^{\mathcal{X}}. \end{aligned}$$

Combining both sides yields $\partial_t \langle p_t^{\mathcal{X}}, f \rangle = \langle p_t^{\mathcal{X}}, \mathcal{L}_t f \rangle$.

B GRADIENTS ARE PRESERVED

B.1 A PROPERTY OF BREGMAN DIVERGENCES

Here we restate a well-known property of Bregman divergences, seen in e.g. (Lipman et al., 2024; Billera et al., 2025a).

Lemma B.1. *(Expectations commute with Bregman divergences under gradients) Let $D_\phi(a,b) = \phi(a) - \phi(b) - \langle a-b, \nabla\phi(b)\rangle$ be a Bregman divergence, where $\phi\colon \Omega_\phi \subset V \to \mathbb{R}$ is strictly convex and differentiable on the closed and convex domain $\Omega_\phi \subset V$. Let X be an integrable Ω_ϕ-valued random variable such that $\mathbb{E}[|\phi(X)|] < \infty$, and let $f_\theta(x) \in \Omega_\phi$. Then*

$$\nabla_\theta \mathbb{E}[D_\phi(X, f_\theta(x))] = \nabla_\theta D_\phi(\mathbb{E}[X], f_\theta(x)).$$

Proof. Write $b := f_\theta(x)$ for brevity. Since Ω_ϕ is closed and convex and X is an integrable Ω_ϕ-valued random variable, one has $\mathbb{E}[X] \in \Omega_\phi$, so $D_\phi(\mathbb{E}[X], b)$ is well-defined. We also remark that the inner product term $\langle X - b, \nabla\phi(b)\rangle$ has finite expectation since $\nabla\phi(b)$ is deterministic and the Cauchy–Schwarz inequality gives $\mathbb{E}[|\langle X, \nabla\phi(b)\rangle|] \le \|\nabla\phi(b)\|\mathbb{E}[\|X\|] < \infty$.

Now, note that, expanding the Bregman divergence and taking expectations, we obtain

$$\begin{aligned}\mathbb{E}[D_\phi(X,b)] &= \mathbb{E}[\phi(X)] - \phi(b) - \mathbb{E}[\langle X-b, \nabla\phi(b)\rangle] \\ &= \mathbb{E}[\phi(X)] - \phi(b) - \langle \mathbb{E}[X] - b, \nabla\phi(b)\rangle.\end{aligned}$$

On the other hand, expanding $D_\phi(\mathbb{E}[X], b)$ gives

$$D_\phi(\mathbb{E}[X], b) = \phi(\mathbb{E}[X]) - \phi(b) - \langle \mathbb{E}[X] - b, \nabla\phi(b)\rangle.$$

Comparing the two expressions yields

$$\mathbb{E}[D_\phi(X,b)] = D_\phi(\mathbb{E}[X], b) + \mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X]). \tag{25}$$

The remainder $\mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X])$ is independent of θ. Applying ∇_θ to both sides eliminates this constant and one has

$$\nabla_\theta \mathbb{E}[D_\phi(X, f_\theta(x))] = \nabla_\theta D_\phi(\mathbb{E}[X], f_\theta(x)). \qquad \square$$

B.2 PROOF OF THEOREM 3.12

Proof. We show that $\nabla_\theta L_{\mathrm{cgm}}(\theta) = \nabla_\theta L_{\mathrm{gm}}(\theta)$. Writing the conditional loss in its disintegrated form (Definition 3.10), we compute

$$\begin{aligned}
&\nabla_\theta L_{\mathrm{cgm}}(\theta) \\
&\quad= \nabla_\theta \mathbb{E}_{t\sim U[0,1]} \mathbb{E}_{x_t \sim p_t^{\mathcal{X}}} \mathbb{E}_{y_t \sim p_t^{\mathcal{Y}}(\cdot \mid \Phi(Y_t) = x_t)} [D_{t,x_t}(F_t^{y_t}(x_t), F_t^\theta(x_t))] \\
&\quad= \mathbb{E}_{t\sim U[0,1]} \mathbb{E}_{x_t \sim p_t^{\mathcal{X}}} \nabla_\theta \mathbb{E}_{y_t \sim p_t^{\mathcal{Y}}(\cdot \mid \Phi(Y_t) = x_t)} [D_{t,x_t}(F_t^{y_t}(x_t), F_t^\theta(x_t))] \\
&\quad= \mathbb{E}_{t\sim U[0,1]} \mathbb{E}_{x_t \sim p_t^{\mathcal{X}}} \nabla_\theta D_{t,x_t}(\mathbb{E}_{y_t \sim p_t^{\mathcal{Y}}(\cdot \mid \Phi(Y_t) = x_t)} [F_t^{y_t}(x_t)], F_t^\theta(x_t)) && \text{(Lemma B.1)} \\
&\quad= \mathbb{E}_{t\sim U[0,1]} \mathbb{E}_{x_t \sim p_t^{\mathcal{X}}} \nabla_\theta D_{t,x_t}(F_t(x_t), F_t^\theta(x_t)) && \text{(Eq. 4)} \\
&\quad= \nabla_\theta \mathbb{E}_{t\sim U[0,1]} \mathbb{E}_{x_t \sim p_t^{\mathcal{X}}} [D_{t,x_t}(F_t(x_t), F_t^\theta(x_t))] \\
&\quad= \nabla_\theta L_{\mathrm{gm}}(\theta).
\end{aligned}$$

In the third line, we apply Lemma B.1 to the inner expectation over y_t, noting that the the random variable corresponds to $F_t^{y_t}(x_t) \in \Omega_{t,x_t}$ and the second argument $F_t^\theta(x_t)$ depends on θ but not on y_t, so the Bregman gradient-expectation interchange applies. In the fourth line, we use the identity $F_t(x_t) = \mathbb{E}[F_t^{Y_t}(x_t) \mid \Phi(Y_t) = x_t]$ from equation 4. $\square$

C ONE-DIMENSIONAL EXAMPLE

This appendix provides a derivation for the one-dimensional example of §3.7.1. The example is an instance of the double conditioning of §3.6, in that the static latent variable is the pair of endpoints (x_0, x_1), and the latent stochastic process is the CTMC state z_t. We also note that in our empirical example, the rate $\lambda_z^{x_1}(x,t)$ does not depend on the endpoint x_1, but we present the general case.

C.1 THE CONDITIONAL GENERATOR

Fix endpoints $x_0, x_1 \in \mathbb{R}$. We define the joint process (X_t, Z_t) on $\mathbb{R} \times \{-1, +1\}$, conditioned on x_0, x_1, by specifying its infinitesimal generator and initial distribution $(X_0, Z_0) \sim (\delta_{x_0}, \mathrm{Unif}(\{-1, 1\}))$.

For test functions $g \in C^2(\mathbb{R}) \otimes C(\{-1, +1\})$, let

$$W_t^{x_1} g(x,z) = b_t^{z,x_1}(x)\partial_x g(x,z) + \frac{1}{2}\sigma_t^2 \partial_{xx} g(x,z) + \lambda_z^{x_1}(x,t)[g(x,-z) - g(x,z)], \tag{26}$$

where $b_t^{z,x_1}(x) := \frac{zx_1 - x}{1-t}$ and $\lambda_z^{x_1}(x,t) \geq 0$ is the rate of jumping from z to $-z$. Conditional on $Z_t = z$ and x_1, the continuous component X_t evolves as a Brownian bridge

$$dX_t = \frac{Z_t x_1 - X_t}{1-t} dt + \sigma_t dB_t,$$

directed towards x_1 when $Z_t = +1$ and towards $-x_1$ when $Z_t = -1$, while the discrete component jumps between ± 1 at state-dependent rate $\lambda_z^{x_1}(X_t, t)$. The rates $\lambda_z^{x_1}$ are chosen so that $X_1 = x_1$ a.s. in the conditional paths. The resulting conditional trajectories are illustrated in subfigure B of Figure 1.

Applying equation 26 to $f \circ \pi_{\mathcal{X}}$, the CTMC term vanishes, giving the conditional generator

$$L_t^{x_1,z} f(x_t) = \frac{z_t x_1 - x_t}{1-t} f'(x_t) + \frac{1}{2}\sigma_t^2 f''(x_t). \tag{27}$$

Rather than parametrising the velocity $b_t^{z,x_1}(x_t) = \frac{z_t x_1 - x_t}{1-t}$ directly (which would require training against a quantity that diverges as $t \to 1$) we adopt an x_1-*prediction* parametrisation. This admits the linear parametrisation

$$L_t^{x_1,z} f(x_t) = \langle \mathcal{K}_{t,x_t} f, F_t^{x_1,z}(x_t) \rangle_{V_{t,x_t}}$$

with $\Omega_{t,x_t} = V_{t,x_t} = \mathbb{R}^2$, equipped with the Euclidean inner product, and

$$\mathcal{K}_{t,x_t} f = \begin{pmatrix} \frac{f'(x_t)}{1-t} \\ -\frac{x_t f'(x_t)}{1-t} + \frac{1}{2}\sigma_t^2 f''(x_t) \end{pmatrix}, \qquad F_t^{x_1,z}(x_t) = \begin{pmatrix} z_t x_1 \\ 1 \end{pmatrix}. \tag{28}$$

Note that $\mathcal{K}_{t,x_t}$ depends on the joint state (x_t, z_t) only through x_t, so this is a valid conditional linear parametrisation in the sense of Definition 3.8.

C.2 TRAINING

By equation 4, the marginal target is

$$F_t(x_t) = \mathbb{E}[F_t^{x_1,Z}(x_t) \mid X_t = x_t] = \begin{pmatrix} \mathbb{E}[Z_t \mid X_t = x_t] \cdot x_1 \\ 1 \end{pmatrix}.$$

The conditional dependence on z_t has been integrated out, and the first component depends on x_t through the conditional mean $\mathbb{E}[Z_t \mid X_t = x_t]$, while the second component is constant. By Theorem 3.12, a neural network $F_t^\theta(x_t) = (x_1^\theta(x_t), f_\theta^{(2)}(x_t))^T$ that receives only x_t as input may be trained against the conditional loss

$$L_{\mathrm{cgm}}(\theta) = \mathbb{E}_{t\sim U[0,1],(x_0,x_1)\sim q,(x_t,z_t)\sim p_t(dx,dz|x_0,x_1)} \left[\left\| F_t^{x_1,z}(x_t) - F_t^\theta(x_t) \right\|^2 \right]$$

and still recover the correct marginal target $F_t(x_t)$. Since $F_t^{x_1,z}(x_t) = (z_t x_1, 1)^T$, the second component is minimised trivially by $f_\theta^{(2)} \equiv 1$ and may be fixed a priori, so it suffices to train only x_1^θ against

$$\mathbb{E}_{t\sim U[0,1],(x_0,x_1)\sim q,(x_t,z_t)\sim p_t(\cdot|x_0,x_1)} \left[\left\| z_t x_1 - x_1^\theta(x_t) \right\|^2 \right].$$

D EDIT FLOWS AS A SPECIAL CASE

In this appendix we give the detailed derivation showing that Theorem 3.1 of Havasi et al. (2025) is recovered as a special case of the pushforward conditional generator matching framework of §3.4–§3.5. We assume that the state space is finite, which has no practical consequence but simplifies the test function space to $C(\mathcal{X} \times \mathcal{Z}) \cong \mathbb{R}^{|\mathcal{X}\times\mathcal{Z}|}$.

D.1 CONDITIONAL GENERATOR AND LINEAR PARAMETRISATION

Let $\mathcal{X}$ and $\mathcal{Z}$ be finite sets and let (X_t, Z_t) be a CTMC on the joint state space $\mathcal{X} \times \mathcal{Z}$ with rates $u_t(x', z' \mid x_t, z_t) \geq 0$. Then the infinitesimal generator on the joint space acts on test functions $g \in C(\mathcal{X} \times \mathcal{Z})$ as

$$W_t g(x_t, z_t) = \sum_{(x',z')} u_t(x', z' \mid x_t, z_t)[g(x', z') - g(x_t, z_t)]. \tag{29}$$

We verify that the conditional generator $L_t^{z_t}$ of §3.5 admits a linear parametrisation in the sense of Definition 3.8 in which $\mathcal{K}_{t,x_t}$ and V_{t,x_t} depend only on x_t. Applying equation 29 to a test function of the form $f \circ \pi_{\mathcal{X}}$, we obtain

$$L_t^{z_t} f(x_t) = W_t(f \circ \pi_{\mathcal{X}})(x_t, z_t) = \sum_{x'} [f(x') - f(x_t)] \underbrace{\sum_{z'} u_t(x', z' \mid x_t, z_t)}_{=:\tilde{u}_t(x' \mid x_t, z_t)}. \tag{30}$$

The quantity $\tilde{u}_t(x' \mid x_t, z_t) := \sum_{z'} u_t(x', z' \mid x_t, z_t)$ is the total rate of the $\mathcal{X}$-component jumping to x', given the current joint state (x_t, z_t), irrespective of where z transitions. Taking $V_{t,x_t} := \mathbb{R}^{|\mathcal{X}|}$ with the standard inner product and defining

$$\mathcal{K}_{t,x_t} f := (f(x') - f(x_t))_{x' \in \mathcal{X}} \in \mathbb{R}^{|\mathcal{X}|}, \qquad F_t^{z_t}(x_t) := (\tilde{u}_t(x' \mid x_t, z_t))_{x' \in \mathcal{X}} \in \mathbb{R}^{|\mathcal{X}|}_{\geq 0},$$

the conditional generator has the linear parametrisation

$$L_t^{z_t} f(x_t) = \langle \mathcal{K}_{t,x_t} f, F_t^{z_t}(x_t) \rangle_{V_{t,x_t}}.$$

Both $\mathcal{K}_{t,x_t}$ and V_{t,x_t} depend on the current state (x_t, z_t) only through x_t, so this is a valid conditional linear parametrisation in the sense of Definition 3.8. The z_t-dependence is confined entirely to the training target $F_t^{z_t}(x_t)$.

D.2 RECOVERY OF THEOREM 3.1 OF HAVASI ET AL. (2025)

By equation 4 and the linearity of the inner product in the second argument, the linear parametrisation of the marginal generator has

$$F_t(x_t) = \mathbb{E}[F_t^{Z_t}(x_t) \mid X_t = x_t],$$

whose x'-component is

$$F_t(x_t)_{x'} = \mathbb{E}_{z_t \sim p_t(\cdot \mid x_t)} \left[\sum_{z'} u_t(x', z' \mid x_t, z_t) \right] = \sum_{z'} \mathbb{E}_{z_t \sim p_t(\cdot \mid x_t)} [u_t(x', z' \mid x_t, z_t)].$$

Writing $u_t(x' \mid x_t) := F_t(x_t)_{x'}$ for the projected marginal rate, we recover the first part of Theorem 3.1 of (Havasi et al., 2025). Indeed, as a direct consequence of Corollary 3.4, if $u_t(x', z' \mid x_t, z_t)$ generates $p_t(x, z)$, then

$$u_t(x' \mid x_t) = \sum_{z'} \mathbb{E}_{z_t \sim p_t(\cdot \mid x_t)} [u_t(x', z' \mid x_t, z_t)] \qquad \text{generates} \qquad p_t^{\mathcal{X}}(x) = \sum_{z} p_t(x, z). \tag{31}$$

The second part of Theorem 3.1 in (Havasi et al., 2025) states that if D_ϕ is a Bregman divergence, then

$$\nabla_\theta \mathbb{E}_{(x_t, z_t) \sim p_t} [D_\phi(\tilde{u}_t(\cdot \mid x_t, z_t), u_t^\theta(\cdot \mid x_t))] = \nabla_\theta \mathbb{E}_{x_t \sim p_t(x)} [D_\phi(u_t(\cdot \mid x_t), u_t^\theta(\cdot \mid x_t))].$$

In our setting, this is Theorem 3.12 applied to the conditional linear parametrisation above.

E NON-UNIQUENESS OF THE KFE IN EXAMPLE 3.15

In this appendix we prove the claim made in Example 3.15, namely that the marginal generator L_t arising from applying the map $\Phi(x) = e^{-1/x} \cdot I_{\{x>0\}}$ to a standard Brownian motion does not uniquely determine the probability path, so that Assumption A.19 fails. For this, we construct a one-parameter family of probability paths all satisfying the same KFE with the same initial condition.

E.1 THE PUSHFORWARD GENERATOR

Recall that $\Phi\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ is defined by $\Phi(x) = e^{-1/x}$ for $x > 0$ and $\Phi(x) = 0$ otherwise. Let $(B_t)_{t\geq 0}$ be a standard Brownian motion with generator $W_t f(x) = \frac{1}{2}\partial_{xx} f(x)$. For $t > 0$, the pushforward marginals are

$$p_t^{\mathcal{X}}(dx) = \frac{1}{2}\delta_0(dx) + \rho_t(x)dx, \tag{32}$$

where ρ_t is the density on $(0, 1)$ arising from the pushforward of the positive half of the Gaussian.

For $x_t \in (0, 1)$, the preimage $\Phi^{-1}(\{x_t\})$ is a single point $w = -1/\log x_t > 0$, so the conditional expectation in Theorem 3.2 simply becomes

$$\begin{aligned} L_t f(x_t) = \mathbb{E}[W_t(f \circ \Phi)(B_t) \mid \Phi(B_t) = x_t] &= \frac{1}{2}(f \circ \Phi)''(w) \\ &= \frac{1}{2}\Phi'(w)^2 f''(x_t) + \frac{1}{2}\Phi''(w) f'(x_t). \end{aligned}$$

Moreover, switching to $x_t = 0$, since Φ and all its derivatives vanish at 0, the chain rule gives $(f \circ \Phi)^{(k)}(0) = 0$ for all $k \geq 1$. Therefore

$$L_t f(0) = \mathbb{E}\big[\frac{1}{2}(f \circ \Phi)''(B_t) \mid B_t \leq 0\big] = 0 \tag{33}$$

for every $f \in \mathcal{T}(\mathbb{R}_{\geq 0})$ and every $t > 0$.

E.2 A ONE-PARAMETER FAMILY OF SOLUTIONS

By the KFE, it holds that $\partial_t \langle p_t^{\mathcal{X}}, f\rangle = \langle p_t^{\mathcal{X}}, L_t f\rangle$, which in our case becomes

$$\int_0^1 f(x)\partial_t \rho_t(x)dx = \frac{1}{2}L_t f(0) + \int_0^1 L_t f(x)\rho_t(x)dx = \int_0^1 L_t f(x)\rho_t(x)dx. \tag{34}$$

For any $c \in [0, 1]$, we define probability paths

$$q_t^{(c)}(dx) := \begin{cases} \delta_0(dx) & t = 0, \\ c\delta_0(dx) + 2(1-c)\rho_t(x)dx & t > 0. \end{cases} \tag{35}$$

Since $B_0 = 0$ a.s., we have $p_0^{\mathcal{X}} = \delta_0$, so every member of the one-parameter family shares the initial condition $q_0^{(c)} = p_0^{\mathcal{X}}$. For $t > 0$ and $f \in \mathcal{T}(\mathbb{R}_{\geq 0})$, one has

$$\partial_t \langle q_t^{(c)}, f\rangle = 2(1-c)\int_0^1 f(x)\partial_t \rho_t(x)dx = 2(1-c)\int_0^1 L_t f(x)\rho_t(x)dx, \tag{36}$$

$$\langle q_t^{(c)}, L_t f\rangle = cL_t f(0) + 2(1-c)\int_0^1 L_t f(x)\rho_t(x)dx = 2(1-c)\int_0^1 L_t f(x)\rho_t(x)dx. \tag{37}$$

so $q_t^{(c)}$ satisfies $\partial_t \langle q_t^{(c)}, f\rangle = \langle q_t^{(c)}, L_t f\rangle$ for every $c \in [0, 1]$. Therefore, in this case, Assumption A.19 fails.