TIME DEPENDENT LOSS REWEIGHTING FOR FLOW MATCHING AND DIFFUSION MODELS IS THEORETICALLY JUSTIFIED

Lukas Billera[*], **Hedwig Nora Nordlinder & Ben Murrell**[*]
Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden

ABSTRACT

Here we clarify that, in Generator Matching (which subsumes a large family of flow matching and diffusion models over continuous, manifold, and discrete spaces), both the Bregman divergence loss and the linear parameterization of the generator can depend on both the current state X_t and the time t, and we show that the expectation over time in the loss can be taken with respect to a broad class of time distributions. We also show this for Edit Flows, which falls outside of Generator Matching. That the loss can depend on t clarifies that time-dependent loss weighting schemes, often used in practice to stabilize training, are theoretically justified when the specific flow or diffusion scheme is a special case of Generator Matching (or Edit Flows). It also often simplifies the construction of X_1-predictor schemes, which are sometimes preferred for model-related reasons. We show examples that rely upon the dependence of linear parameterizations, and of the Bregman divergence loss, on t and X_t.

1 INTRODUCTION

Diffusion (Ho et al., 2020; Song et al., 2021b) and Flow Matching (Lipman et al., 2023) approaches to generative modeling both train a model to transport samples from a simple distribution to the data distribution. This family of approaches was recently subsumed under 'Generator Matching', which characterizes a general space of states, processes, and losses under which conditional samples can be used to train a marginal generator.

The loss used for these models generally involves i) a term that compares the model's prediction to some function of the conditional process generator, and ii) a time-dependent scaling constant prescribed by the specific theoretical framework used to justify the method. Since the outset of the field (e.g., Ho et al. (2020)) right up until recently (Nguyen et al., 2025) the prescribed time-dependent loss scaling is discarded or adjusted in practice, usually motivated empirically and often without theoretical justification (with exceptions, e.g. Zhang et al. (2025); Oresten et al. (2025)).

With unlimited compute, if you trained a separate model for each time t and state X_t, scaling each model's loss by a different positive constant should not change what that model learns, suggesting that time and state dependent loss scaling is justified and merely a practical issue. Here we formalize this notion, and extend it in ways that are trickier to heuristically reason about.

2 PRELIMINARIES

Throughout this note, we make use of notation and definitions similar to those found in Generator Matching (GM) (Holderrieth et al., 2025) and Flow Matching Guide (FMG) (Lipman et al., 2024). Following FMG, generator matching prescribes a conditional probability path $p_t(dx|z)$ interpolating between a simple initial distribution p_0 and a target distribution p_1 on a state space S, conditioned on a latent state $z \in \mathcal{Z}$. With a distribution p_Z over $\mathcal{Z}$, we consider the marginal probability path $p_t(dx)$ defined via the two-stage sampling procedure

$$Z \sim p_Z,\ (X_t|Z=z) \sim p_{t|Z}(dx|z) \implies X_t \sim p_t(dx).$$

[*]Correspondence to: lukas.billera@ki.se & benjamin.murrell@ki.se

We assume the GM-specified regularity conditions hold (cf. Appendix D), denoting our state space by S and our class of test functions by $\mathcal{T}$. Under those conditions, the trajectories of a process X_t are determined by i) its initial value, and 2) the infinitesimal generator — meaning that if a neural network can learn to parametrize the infinitesimal generator, we then can transport from the initial distribution to the target distribution. We aim to parametrize the infinitesimal generator $\mathcal{L}_t$ of the process X_t by a neural network $\mathcal{L}_t^\theta$.

FMG suggests the following form of a linear parametrization of the generator of the process X_t:

$$\mathcal{L}_t f(x) = \langle \mathcal{K} f(x), F_t(x) \rangle_x$$

for a function F_t such that $F_t(x) \in \Omega_x$, where $\Omega_x \subset V_x$ is a closed, convex subset of the inner product space $(V_x, \langle \cdot, \cdot \rangle_x)$. Analogously, FMG defines a linear parameterization of the conditional infinitesimal generator by

$$\mathcal{L}_t^z f(x) = \langle \mathcal{K} f(x), F_t^z(x) \rangle_x$$

for a function F_t^z such that $F_t^z(x) \in \Omega_x$. Then for Bregman divergences $D_x : \Omega_x \times \Omega_x \to \mathbb{R}$, FMG suggests the following form of a conditional generator matching loss

$$L_{\text{cgm}}(\theta) = \mathbb{E}_{t, Z \sim p_Z, X_t \sim p_{t|Z}(x|Z)}[D_{X_t}(F_t^Z(X_t), F_t^\theta(X_t))],$$

and up to a constant in θ, this loss function coincides with the generator matching loss

$$L_{\text{gm}}(\theta) = \mathbb{E}_{t, X_t \sim p_t}[D_{X_t}(F_t(X_t), F_t^\theta(X_t))],$$

so, their minimums also coincide.

We suggest that the class of generator matching loss functions be broadened to include those that have i) explicitly time- and state dependent linear parameterizations, ii) time- and state dependent Bregman divergences, and iii) a nearly arbitrarily weighted time distribution $t \sim \mathcal{D}$. We also extend this characterization to Edit Flows (Havasi et al., 2025), which falls outside of the scope of GM.

3 Previous work

The theoretical aspects of time-dependent loss re-weightings for flow and diffusion models are discussed in (Kingma et al., 2023), (Song et al., 2021a), (Lipman et al., 2024), (Esser et al., 2024), and others. See App. table 1 for a brief summary.

4 Results

4.1 Time and state varying linear parametrizations

We extend the notion of a linear parametrization (Holderrieth et al., 2025) to explicitly vary with time and state, whereas in Lipman et al. (2024) it was already clarified to be able to vary with state, but to the best of our knowledge not with both.

Following their notation, let S be our state space, $\mathcal{T}$ a class of test functions on S and let $B(S)$ be the bounded functions on S. Fix a time $t \in [0, 1)$ and $x \in S$. A time and state varying linear parametrization of the subspace $W_t \subset \{\mathcal{L} : \mathcal{T} \to B(S) \mid \mathcal{L} \text{ is linear}\}$ is given by: i) a closed, convex set $\Omega_{t,x} \subset V_{t,x}$, where $(V_{t,x}, \langle \cdot, \cdot \rangle_{V_{t,x}})$ forms an inner product space; and ii) a linear operator $\mathcal{K}_{t,x} : \mathcal{T} \to V_{t,x}$ such that every $\mathcal{L}_t \in W_t$ can be written as

$$\mathcal{L}_t f(x) = \langle \mathcal{K}_{t,x} f,\ F_t(x) \rangle_{V_{t,x}},$$

for a function $F_t : S \to \bigsqcup_{x \in S} \Omega_{t,x}$ such that $F_t(x) \in \Omega_{t,x}$ for every time t and $x \in S$. If such a construction exists, then F_t is said to linearly parametrize W_t, or with the appropriate context, to linearly parametrize the generator $\mathcal{L}_t$.

Assume that for each $z \in \mathcal{Z}$, where $\mathcal{Z}$ is a latent conditioning space, the conditional Markov process X_t^z has a conditional infinitesimal generator $\mathcal{L}_t^z$. Then a time and state varying linear parametrization of $\mathcal{L}_t^z$ is given by

$$\mathcal{L}_t^z f(x) = \langle \mathcal{K}_{t,x} f,\ F_t^z(x) \rangle_{V_{t,x}},$$

where $F_t^z : S \to \bigsqcup_{x \in S} \Omega_{t,x}$ is such that $F_t^z(x) \in \Omega_{t,x}$ for every time t, $x \in S$ and $z \in \mathcal{Z}$, and we require the following regularity conditions:

1. For each time $t \in [0,1)$ and $x \in S$, it holds that $\dim(V_{t,x}) < \infty$ and
$$\mathbb{E}_{Z\sim p_{Z|t}(dz|x)}\|F_t^Z(x)\|_{V_{t,x}} < \infty,$$
where $p_{Z|t}(dz|x)$ is the posterior distribution $(Z|X_t = x)$.
2. The GM regularity assumptions listed in Appendix D hold for the marginal and model parametrized generators and their associated Markov processes. For the model parametrized generators, we write
$$\mathcal{L}_t^\theta f(x) = \langle \mathcal{K}_{t,x} f, F_t^\theta(x)\rangle_{V_{t,x}},$$
and we denote X_t^θ for the associated Markov process. For the marginal generator, we write
$$\mathcal{L}_t f(x) = \langle \mathcal{K}_{t,x} f, F_t(x)\rangle_{V_{t,x}}$$
and simply denote X_t for the associated Markov process. See Theorem 4.10 for a connection between the conditional and marginal parametrizations.

4.2 TIME AND STATE VARYING LOSS

In this section, we extend the concept of a generator matching loss (GM loss) and a conditional generator matching loss (CGM loss) as introduced in GM and FMG, to allow for a general class of time distributions $\mathcal{D}$, with an extended class of Bregman divergences $D_{t,x}$ that depend on both time and state, applied to linear parameterizations that are time and state dependent.

As in Banerjee et al. (2005) and Della Pietra et al. (2002), we consider Bregman divergences corresponding to $\phi : \Omega \to \mathbb{R}$ that might only be differentiable on a subset $\widehat{\Omega} \subseteq \Omega$ containing the relative interior $\mathrm{ri}(\Omega)$ (see Remark B.1). In particular, this is useful for rigorously defining the Poisson-like Bregman divergence and the Binary Cross Entropy Bregman divergence in Examples C.6 and C.7, where ϕ is not differentiable on the boundary of Ω. To our knowledge, Bregman divergences were only defined in association with convex functions differentiable on all of Ω in GM and FMG, making it difficult to rigorously reason about such cases.

For each time $t \in [0,1]$ and $x \in S$, let $\Omega_{t,x}$ be a closed convex subset of the inner product space $(V_{t,x}, \langle\cdot,\cdot\rangle_{V_{t,x}})$ and let $\phi_{t,x} : \Omega_{t,x} \to \mathbb{R}$ be a strictly convex continuous function. Let $\widehat{\Omega}_{t,x}$ be a convex set satisfying $\mathrm{ri}(\Omega_{t,x}) \subseteq \widehat{\Omega}_{t,x} \subseteq \Omega_{t,x}$ on which $\phi_{t,x}$ is differentiable (see Remark B.1 for the definition of $\mathrm{ri}(\Omega_{t,x})$).

We define the Bregman divergence $D_{t,x} : \Omega_{t,x} \times \widehat{\Omega}_{t,x} \to \mathbb{R}$ associated to $\phi_{t,x}$ by
$$D_{t,x}(a,b) = \phi_{t,x}(a) - \phi_{t,x}(b) - \langle a-b, \nabla\phi_{t,x}(b)\rangle_{V_{t,x}}.$$

When $\widehat{\Omega}_{t,x} \neq \Omega_{t,x}$, a closure property is required that, in practice, is nonrestrictive (cf. Assumption 3 and its succeeding remark). Note that whenever $\Omega_{t,x} = V_{t,x}$, we have $\mathrm{ri}(\Omega_{t,x}) = \Omega_{t,x}$, so $\widehat{\Omega}_{t,x} = \Omega_{t,x}$ (see Remark B.3).

Definition 4.1 (Valid time distributions)**.** We consider probability distributions $\mathcal{D}$ on $[0,1]$ satisfying $\mathcal{D} \gg \lambda$ and $\mathcal{D}(\{1\}) = 0$ where λ is the Lebesgue measure.

Definition 4.2 (Reweighting function)**.** We consider *weighting functions* $w : [0,1] \to \mathbb{R}_{\geq 0}$ satisfying $\mathbb{E}_{t\sim\mathcal{D}}[w(t)] < \infty$ and $w(t) > 0$ for λ a.e t.

Remark 4.3. *Note that reweighting the loss by such a $w(t)$ is tantamount to taking the expectation over an alternative time distribution which also dominates the Lebesgue measure and selecting an alternative Bregman divergence rescaled by a constant factor, as discussed in Lemma 4.5.*

We extend the notion of a GM loss to allow for
$$L_{\mathrm{gm}}(\theta) = \mathbb{E}_{t\sim\mathcal{D}, X_t\sim p_t(dx)}[w(t) D_{t,X_t}(F_t(X_t), F_t^\theta(X_t))]$$
and we similarly extend the CGM loss to allow
$$L_{\mathrm{cgm}}(\theta) = \mathbb{E}_{t\sim\mathcal{D}, X_t\sim p_t(dx), Z\sim p_{Z|t}(dz|x)}[w(t) D_{t,X_t}(F_t^Z(X_t), F_t^\theta(X_t))].$$

In the above, the integrands are defined for $t \in [0,1)$ (since F_t, F_t^Z, F_t^θ parametrize the generator, which exists only on $[0,1)$), and we require the following integrability conditions:

1. $\mathbb{E}_{t\sim\mathcal{D},X_t\sim p_t(dx)}[w(t)D_{t,X_t}(F_t(X_t),F_t^\theta(X_t))]<\infty$,
2. $\mathbb{E}_{t\sim\mathcal{D},X_t\sim p_t,Z\sim p_{Z|t}(dz|x)}w(t)[D_{t,X_t}(F_t^Z(X_t),F_t(X_t))]<\infty$,
3. For all $t\in[0,1)$ and $x\in S$: $\mathbb{E}_{Z\sim p_{Z|t}(dz)}[F_t^Z(x)]\in\widehat{\Omega}_{t,x}$ and $F_t^\theta(x)\in\widehat{\Omega}_{t,x}$.

Remark 4.4. *The third condition is automatically satisfied when $\widehat{\Omega}_{t,x}=\Omega_{t,x}$ (see Remark B.3). Otherwise, the condition is not particularly restrictive: if it fails, then the essential convex hull of $z\mapsto F_t^z(x)$ lies in the boundary of $\Omega_{t,x}$, as noted in the case when $\mathcal{Z}$ is finite in Banerjee et al. (2005). If this condition doesn't hold exactly, then you can only ever be able to approach the true expectation, and this assumption is implicit in e.g. parametrizing probabilities via a sigmoid function in discrete models — in which case, the model cannot ever emit a probability of exactly 0 or 1 with finite logits. In the separable case (Section 4.5), where $\widehat{\Omega}_{t,x}=\prod_{i=1}^{N_{t,x}}\widehat{\Omega}_{t,x}^i$ as in Corollary B.4, this condition reduces to $\mathbb{E}[F_t^{Z,i}(x)]\in\widehat{\Omega}_{t,x}^i$ and $F_t^{\theta,i}(x)\in\widehat{\Omega}_{t,x}^i$ for each component i, with the same remarks applying per component.*

4.3 Time dependent loss reweighting

Note: In what follows, we denote

$$\langle\mu,f\rangle=\int f(x)\mu(dx)$$

for the duality pairing between probability measures μ on S and test functions $f\in\mathcal{T}$.

Lemma 4.5. *Let $\mathcal{D}$ be as in 4.1, and let $w(t)\geq 0$ be as in 4.2. Then $\widetilde{\mathcal{D}}(dt):=\frac{w(t)}{\int w(t)\mathcal{D}(dt)}\mathcal{D}(dt)$ is a probability measure on $[0,1]$ such that $\widetilde{\mathcal{D}}\gg\lambda$ and one has*

$$\mathbb{E}_{t\sim\mathcal{D}}[w(t)f(t)]=K\cdot\mathbb{E}_{t\sim\widetilde{\mathcal{D}}}[f(t)],\qquad K=\int w(t)\mathcal{D}(dt)>0.$$

Proof. See Appendix F.1.1. □

Remark 4.6. *In fact, in the above, we only require $w(t)\geq 0$ for $\mathcal{D}$-almost every $t\in[0,1]$.*

Theorem 4.7. *Let $L_{\mathrm{gm}}(\theta)$ be the generator matching loss defined in Section 4.2,*

$$L_{\mathrm{gm}}(\theta)=\mathbb{E}_{t\sim\mathcal{D},X_t\sim p_t(dx)}[w(t)D_{t,X_t}(F_t(X_t),F_t^\theta(X_t))]$$

again with $\mathcal{D}$ as in 4.1 and $w(t)$ as in 4.2. Suppose that

$$L_{\mathrm{gm}}(\theta)=0$$

and that the GM regularity assumptions listed in Appendix D hold for $\mathcal{L}_t$ and $\mathcal{L}_t^\theta$, for the same class of test functions $\mathcal{T}$. Then the model parametrized generator $\mathcal{L}_t^\theta$ solves the Kolmogorov Forward Equation (KFE) for p_t on $[0,1)$, and the Markov process X_t^θ associated to $\mathcal{L}_t^\theta$ with the initial distribution p_0 will satisfy $X_t^\theta\sim p_t(dx)$ for all $t\in[0,1]$. In particular, $X_1^\theta\sim p_1(dx)$.

Proof. See Appendix F.2. □

Example 4.8. Suppose $\mathcal{D}$ has a density against the Lebesgue measure that is positive Lebesgue-almost everywhere in $[0,1]$. That is, it would permissible if its density was zero on a set of Lebesgue-measure zero. Then $\mathcal{D}\gg\lambda$ since $A\subset[0,1]$ and $\lambda(A)>0$ implies $\mathcal{D}(A)>0$, so we may take our loss to be weighted by $\mathcal{D}$ — i.e., it is permissible for $\mathcal{D}$ to have a probability density function on $[0,1]$ that is positive except for a set of Lebesgue-measure zero.

Example 4.9. It would, for example, be permissible if $\mathcal{D}$ had a density that vanished at finitely many points of $[0,1]$, so e.g. either or both boundary points of $[0,1]$, e.g., a $\mathrm{Beta}(\alpha,\beta)$ distribution. It would also be permissible if $\mathcal{D}$ had a density that vanished at countably many points of $[0,1]$.

4.4 GENERATOR MATCHING

Assume for every $z \in \mathcal{Z}$ there corresponds a conditional Markov process X_t^z with conditional infinitesimal generator $\mathcal{L}_t^z$ and linear parametrization

$$\mathcal{L}_t^z f = \langle \mathcal{K}_{t,x} f; F_t^z(x) \rangle_{V_{t,x}},$$

as is described in Section 4.1. By Proposition 1 in GM, we have that

$$\mathcal{L}_t f(x) = \mathbb{E}_{Z \sim p_{Z|t}(dz|x)}[\mathcal{L}_t^Z f(x)]$$

generates $p_t(dx) = \int_{\mathcal{Z}} p_t(dx|z) p_Z(dz)$ — i.e., the KFE holds for $t \in [0, 1)$:

$$\partial_t \langle p_t, f \rangle = \langle p_t, \mathcal{L}_t f \rangle.$$

The proof of Proposition 1 in GM Appendix C.2 for time and state varying linear parametrizations runs through the same and is adapted below:

Theorem 4.10. *Let*

$$\mathcal{L}_t^z f(x) = \langle \mathcal{K}_{t,x} f; F_t^z(x) \rangle_{V_{t,x}}$$

be a linear parametrization of $\mathcal{L}_t^z$, as in Section 4.1, for $F_t^z(x) \in \widehat{\Omega}_{t,x} \subset V_{t,x}$. Then it follows that

$$F_t(x) := \mathbb{E}_{Z \sim p_{Z|t}(dz|x)}[F_t^Z(x)]$$

linearly parametrizes the marginal generator.

Proof. See Appendix F.3.1. □

Next, we adapt Proposition 2 from GM to support a time and state varying Bregman divergence $D_{t,x}$ with a corresponding time and state varying linear parametrization.

Theorem 4.11. *Let $F_t^z : S \to \bigsqcup_{x \in S} \widehat{\Omega}_{t,x}$ be a time and state varying linear parametrization of the conditional generator as in Section 4.1, from which it holds that*

$$F_t(x) = \mathbb{E}_{Z \sim p_{Z|t}(dz|x)}[F_t^Z(x)]$$

linearly parametrizes the marginal generator, by Theorem 4.10. Let $D_{t,x} : \Omega_{t,x} \times \widehat{\Omega}_{t,x} \to \mathbb{R}$ be a Bregman divergence for each $t \in [0, 1]$ and each $x \in S$. Then the CGM loss and the GM loss coincide up to a constant in θ

$$\begin{aligned} &\nabla_\theta \mathbb{E}_{t \sim \mathcal{D}, X_t \sim p_t(dx), Z \sim p_{Z|t}(dz|x)}[w(t) D_{t,X_t}(F_t^Z(X_t), F_t^\theta(X_t))] \\ &\qquad = \nabla_\theta \mathbb{E}_{t \sim \mathcal{D}, X_t \sim p_t(dx)}[w(t) D_{t,X_t}(F_t(X_t), F_t^\theta(X_t))] \end{aligned}$$

where $\mathcal{D}, w(t)$ are as in 4.1, 4.2.

Proof. See Appendix F.3.2. □

4.5 SUMS OF LINEAR PARAMETRIZATIONS

Fix a time $t \in [0, 1)$ and $x \in S$. Suppose that for all $f \in \mathcal{T}$ the infinitesimal generator $\mathcal{L}_t$ can be written as a sum of linear parametrizations

$$\mathcal{L}_t f(x) = \sum_{i=1}^{N_{t,x}} \langle \mathcal{K}_{t,x}^i f, F_t^i(x) \rangle_{V_{t,x}^i}$$

for $N_{t,x} < \infty$ and functions $F_t^i(x) \in \widehat{\Omega}_{t,x}^i \subset V_{t,x}^i$, and linear operators $\mathcal{K}_{t,x}^i : \mathcal{T} \to V_{t,x}^i$, for $i = 1, \ldots, N_{t,x}$.

We show that a valid per-term conditional generator matching loss is given by a loss whose per-term contribution is a sum of Bregman divergences, one along each component. This was mentioned in GM Proposition 5 in the context of multimodal generative models with factorized probability paths, but we would like to make this note more general.

Theorem 4.12. *Under the above setup, the sum of linear parametrizations can be rewritten as a single linear parametrization* $\mathcal{L}_t f(x) = \langle \mathcal{K}_{t,x} f, F_t(x) \rangle_{V_{t,x}}$, *where* $V_{t,x} := \bigoplus_{i=1}^{N_{t,x}} V_{t,x}^i$ *and* $F_t(x) := (F_t^i(x))_{i=1}^{N_{t,x}}$. *Assuming the conditions in Sections 4.1 and 4.2 hold for the overall linear parametrization, a valid conditional generator matching loss under a separable Bregman divergence* $D_{t,x}(y,x) = \sum_{i=1}^{N_{t,x}} D_{t,x}^i(y^i, x^i)$ *is given by*

$$L_{\mathrm{cgm}}(\theta) = \mathbb{E}_{t\sim\mathcal{D}, Z\sim p_Z, X_t\sim p_t(dx|z)} \left[\sum_{i=1}^{N_{t,X_t}} D_{t,X_t}^i(F_t^{Z,i}(X_t), F_t^{\theta,i}(X_t)) \right].$$

Proof. See Appendix F.4. □

4.6 Flow matching X_1 prediction

The *endpoint prediction* formulation of flow matching losses provides a natural way to obtain a time-dependent linear parameterization of the drift, and is discussed in e.g. Zhang et al. (2025); Lipman et al. (2024); Oresten et al. (2025), and others.

In flow matching, conditioned on terminating at the endpoint $x_1 \in \mathbb{R}^n$, a time-dependent vector field $u_t(x|x_1)$ governs the drift along the conditional paths. Suppose that $u_t(x|x_1)$ is affine in x_1, i.e.

$$u_t(x|x_1) = A_{t,x} x_1 + b_{t,x},$$

with $A_{t,x} \in \mathbb{R}^{n\times n}$ and $b_{t,x} \in \mathbb{R}^n$.

Theorem 4.13. *Let* $u_t(x|x_1) = A_{t,x}x_1 + b_{t,x}$ *be an affine conditional vector field as above. Then the infinitesimal generator* $\mathcal{L}_t^{x_1} f(x) = \nabla f(x)^T (A_{t,x}x_1 + b_{t,x})$ *admits a sum of linear parametrizations as in Section 4.5, and the model endpoint prediction* $\hat{x}_1^\theta(t,x)$ *yields a valid CGM loss:*

$$L_{\mathrm{cgm}}(\theta) = \mathbb{E}_{t\sim\mathcal{D}, X_1\sim p_Z, X_t\sim p_t(\cdot|x_1)}[w(t) D_{t,X_t}(X_1, \hat{x}_1^\theta(t, X_t))]$$

where $w(t) > 0$ *for* λ*-almost every* $t \in [0,1]$.

Proof. See Appendix F.5. □

Example 4.14. One could take

$$u_t(x|x_1) = \frac{x_1 - x}{1-t}$$

in the above.

Example 4.15. Let $u_t(x|x_1) = A_{t,x}x_1 + b_{t,x}$ be as above and parametrize x_1-predictions by $\hat{x}_1^\theta(t, X_t)$. Due to the linear parametrization

$$\mathcal{L}_t^{x_1} f(x) = \langle A_{t,x}^T \nabla f(x), x_1 \rangle_{\mathbb{R}^n} + \langle b_{t,x}^T \nabla f(x), 1 \rangle_{\mathbb{R}},$$

a valid CGM loss is given by

$$L_{\mathrm{cgm}}(\theta) = \mathbb{E}_{t\sim\mathcal{D}, X_1\sim p_Z, X_t\sim p_t(dx|x_1)} \left[w(t) \| X_1 - \hat{X}_1^\theta(t, X_t) \|^2 \right],$$

for some reweighting function $w(t)$ that is positive λ-almost everywhere on $[0,1]$, using the Bregman divergence $D(y,x) := \|y - x\|^2$ (cf. Example C.2).

Example 4.16. Fix $\varepsilon > 0$, and let $c(t) = \dfrac{1}{(1-t+\varepsilon)^2}$. Then a valid CGM loss is given by

$$L_{\mathrm{cgm}}(\theta) = \mathbb{E}_{t\sim\mathcal{D}, X_1\sim p_Z, X_t\sim p_t(dx|x_1)} \left[\left\| \frac{X_1 - \hat{x}_1^\theta(t, X_t)}{1-t+\varepsilon} \right\|^2 \right].$$

This is in contrast with the usual conditional flow matching loss under x_1 prediction

$$L_{\mathrm{cfm}}(\theta) = \mathbb{E}_{t\sim\mathcal{D}, X_1\sim p_Z, X_t\sim p_t(dx|x_1)} \left[\left\| \frac{X_1 - \hat{x}_1^\theta(t, X_t)}{1-t} \right\|^2 \right]$$

whose per-term contribution has a singularity at $t = 1$.

4.7 DIFFUSION MODELS AND x_0-PREDICTION

As in GM Section H.2, to view denoising diffusion models from the perspective of generator matching, we consider a Markov noising process $\bar{X}_t$ given by $d\bar{X}_t = \sigma_t d\bar{W}_t$ from $t = 1$ to $t = 0$ backwards in time. Then the KFE holds in reverse time with

$$\partial_t \langle p_t, f \rangle = \langle p_t, \nabla f^T \frac{\sigma_t^2}{2} \nabla \log p_t \rangle$$

as it is shown in GM, corresponding to the probability flow ODE in Song et al. (2021b).

Theorem 4.17. *Suppose that $u_t(x|x_0) := \nabla_x \log p_t(x|x_0) = A_{t,x}x_0 + b_{t,x}$ is affine in x_0. Then the conditional generator $\mathcal{L}_t^z f(x) = \langle \frac{\sigma_t^2}{2} \nabla f(x), \nabla_x \log p_t(x|z) \rangle_{\mathbb{R}^n}$ admits a sum of linear parametrizations as in Section 4.5, and a valid CGM loss is given by*

$$L(\theta) = \mathbb{E}_{t\sim\mathcal{D}, x_0\sim p_{\text{data}}, x_t\sim p_t(\cdot|x_0)}[w(t)\|x_0 - \hat{x}_0^\theta(t, x_t)\|^2]$$

for a reweighting function $w(t)$ that is positive λ-almost everywhere on $[0, 1]$.

Proof. See Appendix F.6. □

4.8 RESCALING TIME-DEPENDENT CONSTANTS OUT OF JUMP MODELS

As in GM, we let $Q_t(dy; x)$ specify a time-dependent jump kernel on the state space S. It can be useful to model jump intensities in terms of continuous hazards $h_{t,j}(x) > 0$, together with individual rate multipliers $R_{t,j}(x) \geq 0$ that are continuous, and to train a model to match the rate multiplier (cf. Equation 28 and Proposition 5 in Gat et al. (2024), where in their notation this is done by predicting the posterior $\hat{w}_t^j(x^i|X_t)$ and the hazards are $a_t^{i,j}$).

Fix a time $t \in [0, 1)$ and $x \in S$. As in Branching Flows (Billera et al., 2025), consider a time-dependent jump kernel of the form:

$$Q_t(dy; x) = \sum_{j=1}^{N_{t,x}} h_{t,j}(x) R_{t,j}(x) \delta_{\Gamma_{t,j}(x)}(dy),$$

for continuous functions $\Gamma_{t,j} : S \to S$ that we call jump targets, and continuous rate multipliers $R_{t,j}(x) \geq 0$, for $j = 1, \ldots, N_{t,x}$.

Theorem 4.18. *The generator $\mathcal{L}_t f(x) = \sum_{j=1}^{N_{t,x}} (f(\Gamma_{t,j}(x)) - f(x)) h_{t,j}(x) R_{t,j}(x)$ admits a time and state dependent linear parametrization in which the rate multiplier vector $R_t(x) = (R_{t,j}(x))_{j=1}^{N_{t,x}}$ is the linearly parametrizing function, with inner product space $V_{t,x} = \mathbb{R}^{N_{t,x}}$ weighted by $\langle v, w \rangle_{V_{t,x}} = \sum_{j=1}^{N_{t,x}} h_{t,j}(x) v^j w^j$. For latent $z \in \mathcal{Z}$ with conditional rate multipliers $R_{t,j}^z(x)$ satisfying $\mathbb{E}_{Z\sim p_{Z|t}(dz|x)}[\lambda_{\text{total}}^z(t, x)] < \infty$, a valid CGM loss under the conditions in Section 4.2 is given by*

$$L_{\text{cgm}}(\theta) = \mathbb{E}_{t\sim\mathcal{D}, Z\sim p_Z(dz), X_t\sim p_t(dx|z)}[D_{t,X_t}(R_t^Z(X_t), R_t^\theta(X_t))].$$

In particular, the hazard rates $h_{t,j}(x)$ are not present in the loss, so the loss can be taken directly against X_1-predictions.

Proof. See Appendix F.7. □

Remark 4.19. *In accordance with Assumption 2 in Appendix D, the expected number of jumps should be finite, i.e. $\mathbb{E}\left[\int_0^1 \lambda_{\text{total}}(t, X_t)dt\right] < \infty$ in the above.*

Example 4.20. As used in Branching Flows (Billera et al., 2025), suppose the model predicted rate multipliers are $\rho_{t,i}^\theta \in (0, 1)$ and the ground truth conditional rate multipliers are $\rho_{t,i}^z \in [0, 1]$. A valid CGM loss is

$$\begin{aligned} L_{\text{cgm}}(\theta) &= \mathbb{E}_{t\sim\mathcal{D}, Z\sim p_Z, X_t\sim p_t(dx|z)}[\textstyle\sum_{i=1}^{N_{t,x}} D_{\text{BCE}}(\rho_{t,i}^Z(X_t), \rho_{t,i}^\theta(X_t))] \\ &= \mathbb{E}_{t\sim\mathcal{D}, Z\sim p_Z, X_t\sim p_t(dx|z)}\Big[-\textstyle\sum_{i=1}^{N_{t,x}} \big(\rho_{t,i}^Z(X_t) \log(\rho_{t,i}^\theta(X_t)) \\ &\qquad + (1 - \rho_{t,i}^Z(X_t)) \log(1 - \rho_{t,i}^\theta(X_t))\big) + \text{const.}\Big] \end{aligned}$$

via the separable Bregman divergence which takes the binary cross entropy loss on each component, considering $0 \cdot \log 0 := 0$ in the above (cf. Example C.7).

Example 4.21. Alternatively, also as used in Branching Flows, we may consider the Poisson-style Bregman divergence (cf. Example C.6) between predicted rate multipliers $R^\theta_{t,i}(x) \in (0, \infty)$ and conditional, ground-truth rate multipliers $R^z_{t,i}(x) \in [0, \infty)$,

$$\begin{aligned} L_{\text{cgm}}(\theta) &= \mathbb{E}_{t\sim\mathcal{D}, Z\sim p_Z, X_t\sim p_t(dx|z)} \left[\sum_{i=1}^{N_{t,x}} D_{\text{Poiss}}(R^Z_{t,i}(X_t), R^\theta_{t,i}(X_t))\right] \\ &= \mathbb{E}_{t\sim\mathcal{D}, Z\sim p_Z, X_t\sim p_t(dx|z)} \left[\sum_{i=1}^{N_{t,x}} \left(R^\theta_{t,i}(X_t) - R^Z_{t,i}(X_t)\log(R^\theta_{t,i}(X_t))\right) + \text{const.}\right]. \end{aligned}$$

4.9 Edit Flows Propositions

In the notation of Edit Flows (EF) (Havasi et al., 2025), we let $u_t(x, z|x_t, z_t)$ generate $p_t(x, z)$ on $\mathcal{X} \times \mathcal{Z}$. By the first part of EF Theorem 3.1, $u_t(x|x_t) := \sum_z \mathbb{E}_{p_t(z_t|x_t)} u_t(x, z|x_t, z_t)$ generates $p_t(x) := \sum_z p_t(x, z)$.

We show an extended second part of EF Theorem 3.1:

Theorem 4.22. *For each $t \in [0, 1)$ and $x \in \mathcal{X}$, it holds that*

$$\begin{aligned} &\nabla_\theta \mathbb{E}_{t\sim\mathcal{D}, x_t, z_t\sim p_t(x,z)}[w(t) D_{t,x_t}(\textstyle\sum_z a(t) u_t(\cdot, z|x_t, z_t), b(t) u^\theta_t(\cdot|x_t))] \\ &\qquad = \nabla_\theta \mathbb{E}_{t\sim\mathcal{D}, x_t\sim p_t}[w(t) D_{t,x_t}(a(t) u_t(\cdot|x_t), b(t) u^\theta_t(\cdot|x_t))], \end{aligned}$$

where $a(t), b(t) > 0$ are internally reweighting the loss and $w(t) > 0$ is externally reweighting the loss, under the integrability assumptions:

(a) $\mathbb{E}_{t\sim\mathcal{D}, x_t\sim p_t}[w(t) D_{t,x_t}(a(t) u_t(\cdot|x_t), b(t) u^\theta_t(\cdot|x_t))] < \infty$,

(b) $\mathbb{E}_{t\sim\mathcal{D}, x_t\sim p_t, z_t\sim p_t(z_t|x_t)}[w(t) D_{t,x_t}(\sum_z a(t) u_t(\cdot, z|x_t, z_t), b(t) u_t(\cdot|x_t))] < \infty$.

Proof. See Appendix F.3.3. □

Note that Theorem 4.22 recovers the original statement of the second part of EF Theorem 3.1 when for each t and x_t, we define $w(t) \equiv a(t) \equiv b(t) \equiv 1$, $D_{t,x_t} \equiv D$ and $\mathcal{D} = \delta_t$ for some choice of Bregman divergence D. We also remark that scaling both $a(t)$ and $b(t)$ by the same factor has the same effect as scaling the loss by $w(t)$.

Example 4.23. Let $h(t)$ be some overall hazard rate, e.g. $h(t) := \frac{\dot\kappa_t}{1-\kappa_t}$ for a scheduler κ_t. Recasting the per-time contribution to the loss in terms of a rescaled prediction $v^\theta_t(x|x_t)$ of $\frac{u_t(x|x_t)}{h(t)}$, we have:

$$\nabla_\theta \mathbb{E}_{x_t, z_t\sim p_t(x,z)}[D(\tfrac{1}{h(t)} \textstyle\sum_z u_t(\cdot, z|x_t, z_t), v^\theta_t(\cdot|x_t))] = \nabla_\theta \mathbb{E}_{x_t\sim p_t} D(\tfrac{1}{h(t)} u_t(\cdot|x_t), v^\theta_t(\cdot|x_t))$$

for any Bregman divergence D. The loss minimum is found at $v^{\theta^*}_t(\cdot|x_t) := \frac{1}{h(t)} u_t(\cdot|x_t)$, implying that $v^\theta_t(\cdot|x_t)$ can be used to parametrize edit rates via $u^\theta_t(\cdot|x_t) := h(t) v^\theta_t(\cdot|x_t)$, whereby we recover the marginal rates at the loss minimum:

$$u^{\theta^*}_t(\cdot|x_t) = h(t) v^{\theta^*}_t(\cdot|x_t) = u_t(\cdot|x_t).$$

5 Discussion

While time-dependent loss reweighting is often used in practice, it is sometimes only heuristically or empirically justified. Here we have clarified that, for a large class of generative models, it is also theoretically justified. We further clarify that the linear parameterization of the generator, and the Bregman divergence loss, can both depend on both the time and the state of the process.

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

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6(58):1705–1749, 2005. URL http://jmlr.org/papers/v6/banerjee05b.html.

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, 2025. URL https://arxiv.org/abs/2511.09465.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004. ISBN 0521833787. URL https://stanford.edu/~boyd/cvxbook/.

Stephen Della Pietra, Vincent Dellapietra, and John Lafferty. Duality and auxiliary functions for bregman distances. 01 2002.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching, 2024. URL https://arxiv.org/abs/2407.15595.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

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.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models, 2023. URL https://arxiv.org/abs/2107.00630.

Carl W. Lee. An introduction to convex polytopes. by arne brøndsted. *The American Mathematical Monthly*, 93(9):750–752, 1986. doi: 10.1080/00029890.1986.11971939. URL https://doi.org/10.1080/00029890.1986.11971939.

Y. 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. *ArXiv*, abs/2412.06264, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

John Nguyen, Marton Havasi, Tariq Berrada, Luke Zettlemoyer, and Ricky T. Q. Chen. Oneflow: Concurrent mixed-modal and interleaved generation with edit flows, 2025. URL https://arxiv.org/abs/2510.03506.

Anton Oresten, Kenta Sato, Aron Stålmarck, Lukas Billera, Hedwig Nora Nordlinder, Jack Collier Ryder, Mateusz Kaduk, and Ben Murrell. Spontaneous emergence of symmetry in a generative model of protein structure. *bioRxiv*, 2025. doi: 10.1101/2025.11.03.686219. URL https://www.biorxiv.org/content/early/2025/11/04/2025.11.03.686219.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models, 2021a. URL https://arxiv.org/abs/2101.09258.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021b. URL https://arxiv.org/abs/2011.13456.

Xi Zhang, Yuan Pu, Yuki Kawamura, Andrew Loza, Yoshua Bengio, Dennis L. Shung, and Alexander Tong. Trajectory flow matching with applications to clinical time series modeling, 2025. URL https://arxiv.org/abs/2410.21154.

A Previous work

Work	Result
Kingma et al. (2023)	Justifies re-weighting a variational diffusion loss by a function of the *signal to noise* ratio, which is assumed differentiable
Song et al. (2021b)	Justifies reweighting by a positive, time-dependent function $\lambda(t)$ for a score matching diffusion model. Proposes re-weighting by $\mathbb{E}[\|\nabla_{x_t} \log p_t(x_t\|x_0)\|_2^2]^{-1}$
Lipman et al. (2024)	Equation (4.27) allows taking the time expectation of a flow matching loss with respect to an arbitrary absolutely continuous distribution on $[0, 1]$. As will be shown in lemma 4.5, this corresponds to time-dependent loss reweighting.
Esser et al. (2024)	Justifies time-dependent re-weighting for rectified flow models, and gives an intuition to weight losses at the middle of the generative trajectory higher, since endpoints are simpler prediction targets.

Table 1: Instances of time-dependent loss reweighting from the literature.

B Properties of Bregman Divergences

We state and prove a general analogue of the statement of Proposition 1 in Banerjee et al. (2005).

Remark B.1. *Following Lee (1986), the relative interior of Ω, denoted $\mathrm{ri}(\Omega)$, is the interior of Ω taken within its affine hull. The affine hull of Ω, denoted $\mathrm{aff}(\Omega)$, is the set of affine combinations of elements of Ω, and an affine combination of $x_1, \dots, x_n \in \Omega$ is given by $\sum_{i=1}^n \alpha_i x_i$, where $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ are such that $\sum_{i=1}^n \alpha_i = 1$. Note that in contrast to convex combinations, the terms α_i can be negative. For example, if Ω contains two distinct points, then $\mathrm{aff}(\Omega)$ must contain the line through them.*

Lemma B.2. *Let $(V, \langle\cdot,\cdot\rangle)$ be a finite dimensional inner product space, let $\Omega \subset V$ be convex and closed, and let $\phi : \Omega \to \mathbb{R}$ be a continuous strictly convex function. Let $\widehat{\Omega}$ be a convex set satisfying $\mathrm{ri}(\Omega) \subseteq \widehat{\Omega} \subseteq \Omega$ on which ϕ is differentiable (see Remark B.1 for the definition of $\mathrm{ri}(\Omega)$). Define the Bregman divergence $D_\phi : \Omega \times \widehat{\Omega} \to \mathbb{R}$ by*

$$D_\phi(a, b) = \phi(a) - \phi(b) - \langle a - b, \nabla\phi(b)\rangle.$$

Let $(\Xi, \mathcal{F}, \mu)$ be a probability space and let $X : \Xi \to \Omega \subset V$ be a random vector such that

1. $\mathbb{E}_\mu\|X\| < \infty$ *and* $\mathbb{E}_\mu[X] \in \widehat{\Omega}$,
2. $\mathbb{E}_\mu[D_\phi(X, \mathbb{E}_\mu[X])] < \infty$.

Then it holds that

(a) $\mathbb{E}_\mu[X]$ *is the unique minimizer of* $y \mapsto \mathbb{E}_\mu[D_\phi(X, y)]$ *over* $\widehat{\Omega}$.

(b) $\mathbb{E}_\mu[D_\phi(X, y)] = D_\phi(\mathbb{E}_\mu[X], y) + \mathbb{E}_\mu[D_\phi(X, \mathbb{E}_\mu[X])]$, *for all* $y \in \widehat{\Omega}$.

(c) $D_\phi(x, y) = 0 \iff x = y$ *and* $D_\phi(x, y) > 0$ *if* $x \neq y$, *over* $x \in \Omega$ *and* $y \in \widehat{\Omega}$.

Proof. Property (c) is the first order characterization of strict convexity of ϕ at any point y where ϕ is differentiable, as seen in Equation (3.3) in Boyd & Vandenberghe (2004); see also Property 1 in Appendix A of Banerjee et al. (2005).

For (b), first note that both y and $\mathbb{E}_\mu[X]$ lie in $\widehat{\Omega}$, so $\nabla\phi$ exists at both points. A computation gives

$$\begin{aligned} D_\phi(x, y) - D_\phi(z, y) &= \phi(x) - \phi(z) - \langle x - z, \nabla\phi(y)\rangle \\ &= D_\phi(x, z) - \langle x - z, \nabla\phi(y) - \nabla\phi(z)\rangle. \end{aligned}$$

Setting $z = \mathbb{E}_\mu[X]$ and taking expectations,

$$\begin{aligned}\mathbb{E}_\mu\big[D_\phi(X,y) - D_\phi(\mathbb{E}_\mu[X],y)\big] &= \mathbb{E}_\mu\big[D_\phi(X,\mathbb{E}_\mu[X]) - \langle X - \mathbb{E}_\mu[X],\ \nabla\phi(\mathbb{E}_\mu[X]) - \nabla\phi(y)\rangle\big] \\ &= \mathbb{E}_\mu\big[D_\phi(X,\mathbb{E}_\mu[X])\big] - \langle \underbrace{\mathbb{E}_\mu[X - \mathbb{E}_\mu[X]]}_{=0},\ \nabla\phi(\mathbb{E}_\mu[X]) - \nabla\phi(y)\rangle \\ &= \mathbb{E}_\mu[D_\phi(X,\mathbb{E}_\mu[X])].\end{aligned}$$

Noting that $D_\phi(\mathbb{E}_\mu[X], y) < \infty$ and $\mathbb{E}_\mu[D_\phi(X,\mathbb{E}_\mu[X])] < \infty$, we can conclude (b). Finally, (a) follows from (c) and (b). □

Remark B.3. *When $\widehat{\Omega} = \Omega$, condition 1 reduces to $\mathbb{E}_\mu\|X\| < \infty$, since $\mathbb{E}_\mu[X] \in \Omega$ is automatic: Ω is convex and closed, and $\mathbb{E}_\mu[X]$ is the limit of convex combinations in Ω.*

Corollary B.4. *For $i = 1, \dots, N$, let $\Omega_i \subset V_i$ be closed and convex and let $\phi_i : \Omega_i \to \mathbb{R}$ be strictly convex and continuous, differentiable on $\widehat{\Omega}_i$ where $\mathrm{ri}(\Omega_i) \subseteq \widehat{\Omega}_i \subseteq \Omega_i$. Define $V = \bigoplus_{i=1}^N V_i$, $\Omega = \prod_{i=1}^N \Omega_i$, $\widehat{\Omega} = \prod_{i=1}^N \widehat{\Omega}_i$, and $\Phi(x) = \sum_{i=1}^N \phi_i(x^i)$. Then Φ is strictly convex and continuous on Ω, differentiable on $\widehat{\Omega}$, and $\mathrm{ri}(\Omega) = \prod_{i=1}^N \mathrm{ri}(\Omega_i) \subseteq \widehat{\Omega} \subseteq \Omega$. In particular, Lemma B.2 applies to $D_\Phi : \Omega \times \widehat{\Omega} \to \mathbb{R}$, and the separable decomposition $D_\Phi(y,x) = \sum_{i=1}^N D_{\phi_i}(y^i, x^i)$ holds.*

Proof. Strict convexity, continuity, and differentiability of Φ on the respective domains follow from the corresponding properties of each ϕ_i. The inclusion $\mathrm{ri}(\Omega) \subseteq \widehat{\Omega}$ follows from $\mathrm{ri}(\prod_i \Omega_i) = \prod_i \mathrm{ri}(\Omega_i) \subseteq \prod_i \widehat{\Omega}_i = \widehat{\Omega}$. The separable decomposition is a direct computation using $\nabla\Phi(x) = (\nabla\phi_1(x^1), \dots, \nabla\phi_N(x^N))$. □

C Examples of Bregman Divergences

Here, we present several examples of Bregman divergences, e.g., as in Holderrieth et al. (2025) and Lipman et al. (2024).

Example C.1 (Time-Scaled Bregman Divergence)**.** Let $\phi_t : \Omega_t \to \mathbb{R}$ be a strictly convex differentiable function from the closed convex set $\Omega_t \subset V_t$ where $(V_t, \langle\cdot,\cdot\rangle_t)$ is an inner product space. Let $w(t) > 0$ and consider the mapping $\psi_t(x) = w(t)\phi_t(x)$. Under these conditions,

$$\begin{aligned}D_{\psi_t}(y,x) &= \psi_t(y) - \psi_t(x) - \langle y - x, \nabla\psi_t(x)\rangle \\ &= w(t)\Big(\phi_t(y) - \phi_t(x) - \langle y - x, \nabla\phi_t(x)\rangle\Big) \\ &= w(t) D_{\phi_t}(y,x).\end{aligned}$$

The above shows that $w(t)D_{\phi_t}(y,x)$ defines a Bregman divergence for each $t \in [0,1]$. A critical example is when $w(t) \equiv C > 0$, some positive constant, which in particular shows that a Bregman divergence rescaled by a positive constant is still a Bregman divergence.

Example C.2 (MSE)**.** Let $\phi_{\mathrm{MSE}} : \mathbb{R}^N \to \mathbb{R}$ be the strictly convex and differentiable function $\phi_{\mathrm{MSE}}(x) = \|x\|^2 = \sum_{i=1}^N x^i x^i$, using superscripts for components. Then the Bregman divergence associated to ϕ_{MSE} is given by

$$\begin{aligned}D_{\mathrm{MSE}}(y,x) &= \phi_{\mathrm{MSE}}(y) - \phi_{\mathrm{MSE}}(x) - \langle y - x, \nabla\phi_{\mathrm{MSE}}(x)\rangle \\ &= \sum_{i=1}^N (y^i y^i - x^i x^i - 2(y^i - x^i)(x^i)) = \|y - x\|^2.\end{aligned}$$

Example C.3 (Separable Bregman Divergences)**.** Let $\phi_i : \Omega_i \to \mathbb{R}$ be strictly convex and differentiable functions for $i = 1, \dots, N$. Suppose further that $\Omega_i \subset V_i$ is convex and closed, where $(V_i, \langle\cdot,\cdot\rangle_i)$ is an inner product space. Consider $V = \bigoplus_{i=1}^N V_i$ and $\Omega = \prod_{i=1}^N \Omega_i$, and let the inner product on V be given by $\langle y, x\rangle = \sum_{i=1}^N \langle y^i, x^i\rangle_i$, using superscripts to denote the i'th-coordinate.

Notice that Ω is convex and closed, and define $\Phi : \Omega \to \mathbb{R}$ by $\Phi(x) = \sum_{i=1}^N \phi_i(x^i)$. Then Φ is strictly convex and differentiable. This gives rise to a *separable* Bregman divergence, where

$D_{\phi_1}, \ldots, D_{\phi_N}$ act separately on each of the N-components:

$$\begin{aligned} D_\Phi(y, x) &= \Phi(y) - \Phi(x) - \langle y - x, \nabla\Phi(x)\rangle \\ &= \sum_{i=1}^{N} \Big(\phi_i(y^i) - \phi_i(x^i) - \langle y^i - x^i, \nabla\phi_i(x^i)\rangle_i\Big) \\ &= \sum_{i=1}^{N} D_{\phi_i}(y^i, x^i). \end{aligned}$$

Remark C.4. *In Example C.3, each component can independently have its own $\widehat{\Omega}_i$, and by Corollary B.4 the product $\widehat{\Omega} = \prod_{i=1}^{N} \widehat{\Omega}_i$ again satisfies the hypotheses of Lemma B.2. This is useful in practice: for instance, a loss combining MSE on one component (where $\widehat{\Omega}_i = \Omega_i = \mathbb{R}$) with Binary Cross Entropy on another (where $\widehat{\Omega}_i = \mathrm{ri}([0,1]) = (0,1)$).*

Remark C.5. *In the next two examples, we show two Bregman divergences up to a constant in the y, i.e. the left slot, since the model is passed to the right slot. We consider Bregman divergences as mappings $D : \Omega \times \widehat{\Omega} \to \mathbb{R}$ with $\widehat{\Omega} = \mathrm{ri}(\Omega)$ in these examples, as discussed in Section 4.2 and Lemma B.2.*

Example C.6 (Poisson-Style Bregman)**.** Let $\phi_{\mathrm{Poiss}} : [0, \infty) \to \mathbb{R}$ be given by

$$\phi_{\mathrm{Poiss}}(x) = \begin{cases} x \log x & \text{if} \quad x \in (0, \infty), \\ 0 & \text{if} \quad x = 0. \end{cases}$$

A computation shows that ϕ is strictly convex and continuous on $[0, \infty)$ and differentiable on $\mathrm{ri}([0, \infty)) = (0, \infty)$. Its associated Bregman divergence is $D_{\mathrm{Poiss}} : [0, \infty) \times (0, \infty) \to \mathbb{R}$,

$$D_{\mathrm{Poiss}}(y, x) = y \log y - x \log x - (y - x)(1 + \log x) = x - y \log x + f(y),$$

where $f(y) = y \log y - y$ is a function only of y.

Example C.7 (Binary Cross Entropy)**.** Let $\phi_{\mathrm{BCE}} : [0, 1] \to \mathbb{R}$ be given by

$$\phi_{\mathrm{BCE}}(x) = \begin{cases} x \log x + (1 - x) \log(1 - x) & \text{if} \quad x \in (0, 1) \\ 0 & \text{if} \quad x \in \{0, 1\}. \end{cases}$$

Then ϕ is strictly convex and continuous on $[0, 1]$ and differentiable on $\mathrm{ri}([0, 1]) = (0, 1)$. Its associated Bregman divergence is $D_{\mathrm{BCE}} : [0, 1] \times (0, 1) \to \mathbb{R}$,

$$\begin{aligned} D_{\mathrm{BCE}}(y, x) &= D_{\mathrm{Poiss}}(y, x) + D_{\mathrm{Poiss}}(1 - y, 1 - x) \\ &= -[y \log x + (1 - y) \log(1 - x)] + g(y), \end{aligned}$$

where $g(y) = y \log(y) + (1 - y) \log(1 - y)$.

D Regularity conditions inherited from Generator Matching

We repeat a few basic definitions from Generator Matching (GM) (Holderrieth et al., 2025) and their regularity conditions. The infinitesimal generator $\mathcal{L}_t$ of a process X_t acts on a test function $f \in \mathcal{T}$, for $t \in [0, 1)$, via

$$\mathcal{L}_t f(x) = \frac{d}{dh}\bigg|_{h=0} \Big(\mathbb{E}[f(X_{t+h}) | X_t = x]\Big),$$

and analogously, the conditional infinitesimal generator $\mathcal{L}_t^z$ of a conditional Markov process X_t^z is defined for each $z \in \mathcal{Z}$ and $t \in [0, 1)$ by

$$\mathcal{L}_t^z f(x) = \frac{d}{dh}\bigg|_{h=0} \Big(\mathbb{E}[f(X_{t+h}^z) | X_t^z = x]\Big).$$

Also following GM, we denote

$$\langle \mu, f \rangle := \mathbb{E}_{x \sim \mu}[f(x)] = \int f(x) \mu(dx)$$

for the duality pairing between a Borel probability measure μ on our state space S (assumed to be a Polish metric space) and a test function $f \in \mathcal{T} \subset C_0(S)$, denoting $C_0(S)$ for the functions that vanish at infinity. That is, $f \in C_0(S)$ if for all $\varepsilon > 0$ there is a compact K such that $|f(x)| < \varepsilon$ for all $x \in S \setminus K$. In accordance with GM Section A.2., we make the following regularity assumptions:

1. The Markov process $(X_t)_{0\le t\le 1}$ associated to $\mathcal{L}_t$ is Feller, in the sense defined in GM Appendix A.1.2. In particular, we require that $\mathcal{L}_t f \in C_0(S)$ for all $f \in \mathcal{T}$ and $t \in [0,1)$.
2. In each time interval $[s,t]$, the expected number of discontinuities of $u \mapsto X_u$ is finite.
3. There is a dense subspace $\mathcal{T} \subset C_0(S)$ satisfying i) $\mathcal{T} \subset \mathrm{dom}(\mathcal{L}_t)$ for all $t \in [0,1)$, and ii) the function $t \mapsto \mathcal{L}_t f$ is continuous on $[0,1)$ for any $f \in \mathcal{T}$. Moreover, two probability distributions μ_1 and μ_2 are equal if and only if $\mathbb{E}_{x\sim\mu_1}[f(x)] = \mathbb{E}_{x\sim\mu_2}[f(x)]$ holds for all $f \in \mathcal{T}$.
4. Any probability path $(p_t)_{0\le t\le 1}$ satisfies that $t \mapsto \langle p_t, f\rangle$ is continuous in t for all $f \in \mathcal{T}$.
5. KFE Uniqueness: If $(p_t)_{0\le t\le 1}$ is a probability path on S and X_t is the Markov process associated to $\mathcal{L}_t$, then

$$X_0 \sim p_0 \quad \text{and} \quad \partial_t \langle p_t, f\rangle = \langle p_t, \mathcal{L}_t f\rangle \ \text{ for all } t \in [0,1) \implies X_t \sim p_t \ \text{ for all } t \in [0,1].$$

Remark D.1. *We require the GM regularity conditions to hold for the model-parametrized generator $\mathcal{L}_t^\theta$ as well. In practice, on compact state spaces — which covers all cases of practical interest, since one can always restrict to a sufficiently large compact subset of $\mathbb{R}^d$ — the condition $\mathcal{L}_t^\theta f \in C_0(S)$ reduces to $\mathcal{L}_t^\theta f$ being continuous, which holds whenever $F_t^\theta(x)$ is continuous, as is the case for standard neural network architectures. For uniqueness of the KFE, we direct the reader to the discussion in Generator Matching (Holderrieth et al., 2025), Appendix A.2.*

Remark D.2. *Note that Assumption 5 requires the KFE only on $[0,1)$, since the infinitesimal generator is not defined at the terminal time $t = 1$. The conclusion nonetheless extends to $t = 1$ by weak continuity, as shown in the following lemma.*

Lemma D.3. *Let $(p_t)_{0\le t\le 1}$ and $(q_t)_{0\le t\le 1}$ be probability paths on S such that $t \mapsto \langle p_t, f\rangle$ and $t \mapsto \langle q_t, f\rangle$ are both continuous on $[0,1]$ for all $f \in \mathcal{T}$. If $p_t = q_t$ for all $t \in [0,1)$, then $p_1 = q_1$.*

Proof. For any $f \in \mathcal{T}$, by continuity we have

$$\langle p_1, f\rangle = \lim_{t\uparrow 1}\langle p_t, f\rangle = \lim_{t\uparrow 1}\langle q_t, f\rangle = \langle q_1, f\rangle.$$

Since $\mathcal{T}$ separates probability distributions by Assumption 3, it follows that $p_1 = q_1$. □

Remark D.4. *In particular, Lemma D.3 together with Assumption 4 justifies the extension to $t = 1$ in Assumption 5: if the KFE holds on $[0,1)$ and uniquely determines the law of X_t on $[0,1)$ (so that $X_t \sim p_t$ for $t \in [0,1)$), then since both $t \mapsto \langle p_t, f\rangle$ and $t \mapsto \langle \mathrm{law}(X_t), f\rangle$ are continuous on $[0,1]$ (the former by Assumption 4, and the latter by the Feller property), Lemma D.3 gives $X_1 \sim p_1$.*

Theorem D.5. *Under the GM regularity assumptions found above, for every $f \in \mathcal{T}$ the map $t \mapsto \langle p_t, \mathcal{L}_t f\rangle$ is continuous on $[0,1)$.*

Proof. First we show that $t \mapsto \langle p_t, g\rangle$ is continuous for all $g \in C_0(S)$. By Assumption 3, there is a sequence $(g_n)_{n=1}^\infty \subset \mathcal{T}$ such that $\|g_n - g\|_\infty \to 0$ as $n \to \infty$, and we see that

$$\sup_{t\in[0,1]} |\langle p_t, g_n\rangle - \langle p_t, g\rangle| = \sup_{t\in[0,1]} \left| \int (g_n(x) - g(x)) p_t(dx) \right| \le \|g_n - g\|_\infty \to 0, \text{ as } n \to \infty.$$

This shows that $t \mapsto \langle p_t, g_n\rangle$ converges uniformly to $t \mapsto \langle p_t, g\rangle$. Since $t \mapsto \langle p_t, g_n\rangle$ is continuous for all $n \in \mathbb{N}$ and continuity is preserved by uniform limits, we have that $t \mapsto \langle p_t, g\rangle$ is continuous for any fixed $g \in C_0(S)$. Now fix $f \in \mathcal{T}$ and let $t_n \to t$ in $[0,1)$. Using the triangle inequality, we can bound

$$\begin{aligned} |\langle p_{t_n}, \mathcal{L}_{t_n} f\rangle - \langle p_t, \mathcal{L}_t f\rangle| &= |\langle p_{t_n}, \mathcal{L}_{t_n} f - \mathcal{L}_t f + \mathcal{L}_t f\rangle - \langle p_t, \mathcal{L}_t f\rangle\rangle| \\ &= |\langle p_{t_n}, \mathcal{L}_{t_n} f - \mathcal{L}_t f\rangle + \langle p_{t_n} - p_t, \mathcal{L}_t f\rangle| \\ &\le |\langle p_{t_n}, \mathcal{L}_{t_n} f - \mathcal{L}_t f\rangle| + |\langle p_{t_n} - p_t, \mathcal{L}_t f\rangle|. \end{aligned}$$

Now, we have that $\mathcal{L}_{t_n} f \to \mathcal{L}_t f$ as $n \to \infty$ by Assumption 3, so that

$$|\langle p_{t_n}, \mathcal{L}_{t_n} f - \mathcal{L}_t f\rangle| \leq \|\mathcal{L}_{t_n} f - \mathcal{L}_t f\|_\infty \to 0, \text{ as } n \to \infty.$$

Moreover, for all $f \in \mathcal{T}$ it holds $\mathcal{L}_t f \in C_0(S)$ since X_t is Feller by Assumption 1, so we have that

$$|\langle p_{t_n} - p_t, \mathcal{L}_t f\rangle| = |\langle p_{t_n}, \mathcal{L}_t f\rangle - \langle p_t, \mathcal{L}_t f\rangle| \to 0, \text{ as } n \to \infty$$

since we showed the map $t \mapsto \langle p_t, g\rangle$ was continuous for all $g \in C_0(S)$ (which implies that $\langle p_{t_n}, g\rangle \to \langle p_t, g\rangle$ if $t_n \to t$ and $g \in \mathcal{T}$). Combining the above, we get

$$\limsup_{n\to\infty} |\langle p_{t_n}, \mathcal{L}_{t_n} f\rangle - \langle p_t, \mathcal{L}_t f\rangle| \leq \limsup_{n\to\infty} \Big(|\langle p_{t_n}, \mathcal{L}_{t_n} f - \mathcal{L}_t f\rangle| + |\langle p_{t_n} - p_t, \mathcal{L}_t f\rangle|\Big) = 0,$$

so that

$$\lim_{n\to\infty} \langle p_{t_n}, \mathcal{L}_{t_n} f\rangle = \langle p_t, \mathcal{L}_t f\rangle$$

and we can conclude that $t \mapsto \langle p_t, \mathcal{L}_t f\rangle$ is continuous on $[0, 1)$. □

E Regularity conditions for linear parametrizations of conditional generators

Lemma E.1. *For all times $t \in [0, 1)$ and all $x \in S$, let $\dim V_{t,x} < \infty$ and let*

$$\mathcal{L}_t^z f(x) = \langle \mathcal{K}_{t,x} f; F_t^z(x)\rangle_{V_{t,x}}$$

be a linear parametrization of $\mathcal{L}_t^z$. Then, whenever $\mathbb{E}_{Z\sim p_{Z|t}(dz|x)} \|F_t^Z(x)\|_{V_{t,x}} < \infty$, we have

$$\mathbb{E}_{Z\sim p_{Z|t}(dz|x)}[\langle \mathcal{K}_{t,x} f, F_t^Z(x)\rangle_{V_{t,x}}] = \langle \mathcal{K}_{t,x} f, \mathbb{E}_{Z\sim p_{Z|t}(dz|x)}[F_t^Z(x)]\rangle_{V_{t,x}}.$$

Proof. Let $(e_i^{t,x})_{i=1}^{N_{t,x}}$ be an orthonormal basis for $V_{t,x}$ — such a basis can be obtained by the Gram-Schmidt algorithm. Then writing $v^i := \langle v, e_i^{t,x}\rangle_{V_{t,x}}$ and $w^j := \langle w, e_j^{t,x}\rangle_{V_{t,x}}$, we have:

$$\begin{aligned}
\langle v, w\rangle_{V_{t,x}} &= \langle \textstyle\sum_{i=1}^{N_{t,x}} v^i e_i^{t,x}, \sum_{j=1}^{N_{t,x}} w^j e_j^{t,x}\rangle \\
&= \textstyle\sum_{i=1}^{N_{t,x}} v^i \langle e_i^{t,x}, \sum_{j=1}^{N_{t,x}} w^j e_j^{t,x}\rangle \\
&= \textstyle\sum_{i=1}^{N_{t,x}} v^i \sum_{j=1}^{N_{t,x}} w^j \langle e_i^{t,x}, e_j^{t,x}\rangle \\
&= \textstyle\sum_{i=1}^{N_{t,x}} v^i \sum_{j=1}^{N_{t,x}} w^j \delta_{ij} \\
&= \textstyle\sum_{i=1}^{N_{t,x}} v^i w^i.
\end{aligned}$$

Denote $(\mathcal{K}_{t,x} f)_i := \langle \mathcal{K}_{t,x} f, e_i^{t,x}\rangle_{V_{t,x}}$, and similarly denote $(F_t^Z(x))_i := \langle F_t^Z(x), e_i^{t,x}\rangle_{V_{t,x}}$. Then, it holds

$$\mathcal{K}_{t,x} f = \sum_{i=1}^{N_{t,x}} (\mathcal{K}_{t,x} f)_i e_i^{t,x} \qquad \text{and} \qquad F_t^Z(x) = \sum_{i=1}^{N_{t,x}} (F_t^Z(x))_i e_i^{t,x}.$$

By the Cauchy-Schwarz inequality, we can bound

$$|(F_t^Z(x))_i| \leq \|F_t^Z(x)\| \cdot \|e_i^{t,x}\| = \|F_t^Z(x)\| < \infty.$$

The $V_{t,x}$-valued expectation is defined by taking the expectation along the components

$$\mathbb{E}_{Z\sim p_{Z|t}(dz|x)}[F_t^Z(x)] = \sum_{i=1}^{N_{t,x}} \mathbb{E}_{Z\sim p_{Z|t}(dz|x)}[(F_t^Z(x))_i] e_i^{t,x}.$$

In particular, we have

$$\mathbb{E}_{Z\sim p_{Z|t}(dz|x)}[F_t^Z(x)]_i := \langle \mathbb{E}_{Z\sim p_{Z|t}(dz|x)}[F_t^Z(x)], e_i^{t,x}\rangle = \mathbb{E}_{Z\sim p_{Z|t}(dz|x)}[(F_t^Z(x))_i]$$

and we therefore obtain

$$\begin{aligned}
\mathbb{E}[\langle \mathcal{K}_{t,x} f, F_t^Z(x)\rangle_{V_{t,x}}] &= \mathbb{E}_{Z\sim p_{Z|t}(dz|x)}[\textstyle\sum_{i=1}^{N_{t,x}} (\mathcal{K}_{t,x} f)_i \cdot (F_t^Z(x))_i] \\
&= \textstyle\sum_{i=1}^{N_{t,x}} (\mathcal{K}_{t,x} f)_i \mathbb{E}_{Z\sim p_{Z|t}(dz|x)}[F_t^Z(x)_i] \\
&= \textstyle\sum_{i=1}^{N_{t,x}} (\mathcal{K}_{t,x} f)_i \mathbb{E}_{Z\sim p_{Z|t}(dz|x)}[F_t^Z(x)]_i \\
&= \langle \mathcal{K}_{t,x} f, \mathbb{E}_{Z\sim p_{Z|t}(dz|x)}[F_t^Z(x)]\rangle_{V_{t,x}}.
\end{aligned}$$

□

F PROOFS

F.1 TIME DEPENDENT LOSS REWEIGHTING

F.1.1 PROOF OF LEMMA 4.5

Proof. Let $K := \int_{[0,1]} w(t)\,\mathcal{D}(dt) = \mathbb{E}_{t\sim\mathcal{D}}[w(t)]$. We first show $K > 0$ and $\widetilde{\mathcal{D}} \gg \lambda$. Fix $A \subset [0,1]$ with $\lambda(A) > 0$. Since $w > 0$ for λ-almost every t, we have $\lambda(A \cap \{w > 0\}) = \lambda(A) > 0$, and hence $\mathcal{D}(A \cap \{w > 0\}) > 0$ since $\mathcal{D} \gg \lambda$. Writing $A \cap \{w > 0\} = \bigcup_{n=1}^{\infty} A \cap \{w > n^{-1}\}$ and using continuity of measure from below, there exists $n_0 \in \mathbb{N}$ such that $\mathcal{D}(A \cap \{w > n_0^{-1}\}) > 0$, and therefore

$$\int_A w(t)\,\mathcal{D}(dt) \geq n_0^{-1}\,\mathcal{D}(A \cap \{w > n_0^{-1}\}) > 0.$$

Taking $A = [0,1]$ gives $K \in (0,\infty)$, and $\widetilde{\mathcal{D}}(A) = K^{-1}\int_A w\,d\mathcal{D} > 0$ whenever $\lambda(A) > 0$, so $\widetilde{\mathcal{D}} \gg \lambda$. The expectation identity follows directly from the definition:

$$\mathbb{E}_{t\sim\widetilde{\mathcal{D}}}[f(t)] = \int f(t)\,\widetilde{\mathcal{D}}(dt) = K^{-1}\int f(t)\,w(t)\,\mathcal{D}(dt) = K^{-1}\,\mathbb{E}_{t\sim\mathcal{D}}[w(t)f(t)].$$

□

F.2 PROOF OF THEOREM 4.7

Proof. Since $\mathcal{D} \gg \lambda$ and $w(t) > 0$ for λ-almost every $t \in [0,1]$, we use Lemma 4.5 to obtain $\widetilde{\mathcal{D}} \gg \lambda$ and a constant $K > 0$ such that

$$\begin{aligned}
L_{\mathrm{gm}}(\theta) &= \mathbb{E}_{t\sim\mathcal{D},X_t\sim p_t(dx)}[w(t)D_{t,X_t}(F_t(X_t),F_t^\theta(X_t))] \\
&= \mathbb{E}_{t\sim\mathcal{D}}[w(t)\mathbb{E}_{X_t\sim p_t(dx)}[D_{t,X_t}(F_t(X_t),F_t^\theta(X_t))] \\
&= K^{-1}\cdot\mathbb{E}_{t\sim\widetilde{\mathcal{D}}}[\mathbb{E}_{X_t\sim p_t(dx)}[D_{t,X_t}(F_t(X_t),F_t^\theta(X_t))] \\
&= K^{-1}\cdot\mathbb{E}_{t\sim\widetilde{\mathcal{D}},X_t\sim p_t(dx)}[D_{t,X_t}(F_t(X_t),F_t^\theta(X_t))].
\end{aligned}$$

Since rescaling the loss by a positive constant factor simply corresponds to a different choice of Bregman divergence (cf. Example C.1), it therefore suffices to only consider the case $w(t) \equiv 1$ and $\mathcal{D} \gg \lambda$ in the proof going forward.

Suppose that $L_{\mathrm{gm}}(\theta) = 0$. For each $t \in [0,1)$ and $f \in \mathcal{T}$, consider the set $E_t := \{x \in S : \mathcal{L}_t^\theta f(x) = \mathcal{L}_t f(x)\}$, and let $A \subset [0,1)$ be the set of all $t \in [0,1)$ such that $p_t(S \setminus E_t) = 0$.

Since Bregman divergences are non-negative, $L_{\mathrm{gm}}(\theta) = 0$ forces

$$D_{t,X_t}(F_t(X_t),F_t^\theta(X_t)) = 0$$

to hold p_t-almost surely for $\mathcal{D}$-almost every $t \in [0,1]$, and thus λ-almost every $t \in [0,1]$ since $\mathcal{D} \gg \lambda$. By Lemma B.2(c),

$$D_{t,x}(F_t(x),F_t^\theta(x)) = 0 \iff F_t(x) = F_t^\theta(x),$$

and therefore $\mathcal{L}_t^\theta f(X_t) = \mathcal{L}_t f(X_t)$ holds p_t-almost surely for λ-almost every $t \in [0,1)$, from which it follows that $\lambda(A) = 1$. Now consider $h_f^\theta : [0,1) \to \mathbb{R}$, defined by

$$h_f^\theta(t) := \langle p_t, \mathcal{L}_t^\theta f\rangle - \langle p_t, \mathcal{L}_t f\rangle.$$

Then, for every $t \in A$ we have

$$\begin{aligned}
h_f^\theta(t) = \langle p_t, \mathcal{L}_t^\theta f - \mathcal{L}_t f\rangle &= \int_S [\mathcal{L}_t^\theta f(x) - \mathcal{L}_t f(x)]p_t(dx) \\
&= \int_{E_t} [\mathcal{L}_t^\theta f(x) - \mathcal{L}_t f(x)]p_t(dx) + \int_{S\setminus E_t} [\mathcal{L}_t^\theta f(x) - \mathcal{L}_t f(x)]p_t(dx) \\
&= 0,
\end{aligned}$$

where in the last equality, we have used that $\mathcal{L}_t^\theta f(x) - \mathcal{L}_t f(x) = 0$ for all $x \in E_t$ and that $p_t(S \setminus E_t) = 0$.

By the argument of Theorem D.5, applied to $\mathcal{L}_t$ and $\mathcal{L}_t^\theta$, both $t \mapsto \langle p_t, \mathcal{L}_t f\rangle$ and $t \mapsto \langle p_t, \mathcal{L}_t^\theta f\rangle$ are continuous on $[0,1)$, and hence $h_f^\theta(t)$ is continuous on $[0,1)$. Since $h_f^\theta(t) = 0$ for λ-almost every $t \in [0,1)$, it follows that $h_f^\theta(t) = 0$ for all $t \in [0,1)$ by continuity.

This implies that for all $t \in [0,1)$, the KFE holds

$$\begin{aligned}\langle p_t, \mathcal{L}_t^\theta f\rangle &= \langle p_t, \mathcal{L}_t f\rangle \\ &= \partial_t \langle p_t, f\rangle\end{aligned}$$

which by Assumption 5 implies that $X_t^\theta \sim p_t(dx)$ for all $t \in [0,1]$. □

F.3 Generator Matching

F.3.1 Proof of Theorem 4.10

Proof. We have

$$\begin{aligned}\partial_t \langle p_t(dx), f(x)\rangle &= \mathbb{E}_{X_t\sim p_t(dx), Z\sim p_{Z|t}(dz|x)}[\mathcal{L}_t^Z f(X_t)] \\ &= \mathbb{E}_{X_t\sim p_t(dx)}\mathbb{E}_{Z\sim p_{Z|t}(dz|x)}[\langle \mathcal{K}_{t,X_t} f,\ F_t^Z(X_t)\rangle_{V_{t,X_t}}] \\ &= \mathbb{E}_{X_t\sim p_t(dx)}[\langle \mathcal{K}_{t,X_t} f,\ \mathbb{E}_{Z\sim p_{Z|t}(dz|x)} F_t^Z(X_t)\rangle_{V_{t,X_t}}] \\ &= \left\langle p_t(dx), \left\langle \mathcal{K}_{t,x} f,\ \underbrace{\mathbb{E}_{Z\sim p_{Z|t}(dz|x)} F_t^Z(x)}_{=:F_t(x)} \right\rangle_{V_{t,x}} \right\rangle,\end{aligned}$$

where we have used Lemma E.1 in the third equality, possible by Assumption 1 in Section 4.1. This shows that

$$\mathcal{L}_t f(x) := \langle \mathcal{K}_{t,x} f(x),\ F_t(x)\rangle_{V_{t,x}}$$

solves the KFE for p_t on $[0,1)$. □

F.3.2 Proof of Theorem 4.11

Proof. By the remarks in the beginning of the proof of Theorem 4.7, it suffices to consider when $w(t) \equiv 1$ and $\mathcal{D} \gg \lambda$, since the reweighting of the loss by $w(t)$ merely corresponds to an alternative time distribution $\widetilde{\mathcal{D}} \gg \lambda$ with a rescaled choice of Bregman divergence. Now, note that by the second integrability condition in Section 4.2, it follows that

$$\mathbb{E}_{Z\sim p_{Z|t}(dz|X_t)}[D_{t,X_t}(F_t^Z(X_t), F_t(X_t))] < \infty$$

holds p_t-almost surely for $\mathcal{D}$-almost every $t \in [0,1]$. Therefore, we may use Lemma B.2(b) to obtain that

$$\mathbb{E}_{Z\sim p_{Z|t}(dz|X_t)}[D_{t,X_t}(F_t^Z(X_t), F_t^\theta(X_t))] = D_{t,X_t}(F_t(X_t), F_t^\theta(X_t)) + \mathbb{E}_{Z\sim p_{Z|t}(dz|X_t)}[D_{t,X_t}(F_t^Z(X_t), F_t(X_t))].$$

holds p_t-almost surely for $\mathcal{D}$-almost every $t \in [0,1]$. Now, we can write

$$\begin{aligned}&\mathbb{E}_{t\sim\mathcal{D}, X_t\sim p_t(dx), Z\sim p_{Z|t}(dz|X_t)}[D_{t,X_t}(F_t^Z(X_t), F_t^\theta(X_t))] \\ &\quad= \mathbb{E}_{t\sim\mathcal{D}, X_t\sim p_t(dx)}\Big[\mathbb{E}_{Z\sim p_{Z|t}(dz|X_t)}[D_{t,X_t}(F_t^Z(X_t), F_t^\theta(X_t))]\Big] \\ &\quad= \mathbb{E}_{t\sim\mathcal{D}, X_t\sim p_t(dx)}\Big[D_{t,X_t}(F_t(X_t), F_t^\theta(X_t)) + \mathbb{E}_{Z\sim p_{Z|t}(dz|x)} D_{t,X_t}(F_t^Z(X_t), F_t(X_t))\Big] \\ &\quad= \mathbb{E}_{t\sim\mathcal{D}, X_t\sim p_t(dx)}[D_{t,X_t}(F_t(X_t), F_t^\theta(X_t))] + \underbrace{\mathbb{E}_{t\sim\mathcal{D}, X_t\sim p_t(dx), Z\sim p_{Z|t}(dz|X_t)}[D_{t,X_t}(F_t^Z(X_t), F_t(X_t))]}_{\text{const.}}\end{aligned}$$

from which the result follows. □

F.3.3 PROOF OF THEOREM 4.22

Proof. Note that

$$\mathbb{E}_{z_t\sim p_t(z_t|x_t)}\left[\sum_z a(t)u_t(\cdot,z|x_t,z_t)\right] = a(t)u_t(\cdot|x_t).$$

The proof then runs analogously to that of Theorem 4.11. Using Lemma B.2(b) in the second equality, and the integrability assumptions (a) and (b) to ensure well-posedness, we have

$$\begin{aligned}
&\mathbb{E}_{t\sim\mathcal{D},x_t,z_t\sim p_t(x,z)}[w(t)D_{t,x_t}(\textstyle\sum_z a(t)u_t(\cdot,z|x_t,z_t), b(t)u_t^\theta(\cdot|x_t))] \\
&\qquad = \mathbb{E}_{t\sim\mathcal{D},x_t\sim p_t}\left[w(t)\mathbb{E}_{z_t\sim p_t(z_t|x_t)}[D_{t,x_t}(\textstyle\sum_z a(t)u_t(\cdot,z|x_t,z_t), b(t)u_t^\theta(\cdot|x_t))]\right] \\
&\qquad = \mathbb{E}_{t\sim\mathcal{D},x_t\sim p_t}\Big[w(t)D_{t,x_t}(a(t)u_t(\cdot|x_t), b(t)u_t^\theta(\cdot|x_t)) \\
&\qquad\qquad + \mathbb{E}_{z_t\sim p_t(z_t|x_t)}\left[w(t)D_{t,x_t}(\textstyle\sum_z a(t)u_t(\cdot,z|x_t,z_t), b(t)u_t(\cdot|x_t))\right]\Big] \\
&\qquad = \mathbb{E}_{t\sim\mathcal{D},x_t\sim p_t}[w(t)D_{t,x_t}(a(t)u_t(\cdot|x_t), b(t)u_t^\theta(\cdot|x_t))] \\
&\qquad\quad + \underbrace{\mathbb{E}_{t\sim\mathcal{D},x_t\sim p_t,z_t\sim p_t(z_t|x_t)}[w(t)D_{t,x_t}(\textstyle\sum_z a(t)u_t(\cdot,z|x_t,z_t), a(t)u_t(\cdot|x_t))]}_{\text{const.}}
\end{aligned}$$

from which the result follows. □

F.4 SUMS OF LINEAR PARAMETRIZATIONS

F.4.1 PROOF OF THEOREM 4.12

Proof. We rewrite the sum of linear parametrizations as a single linear parametrization, and then we take our Bregman divergence to be separable.

Define $V_{t,x} := \bigoplus_{i=1}^{N_{t,x}} V_{t,x}^i$ and $\widehat{\Omega}_{t,x} := \prod_{i=1}^{N_{t,x}} \widehat{\Omega}_{t,x}^i$. Then we set $F_t(x) := (F_t^i(x))_{i=1}^{N_{t,x}} \in \widehat{\Omega}_{t,x}$. Let $v = (v^1,\dots,v^{N_{t,x}})$ and $w = (w^1,\dots,w^{N_{t,x}})$ be elements of $V_{t,x}$. An inner product on $V_{t,x}$ is given by

$$\langle v, w\rangle_{V_{t,x}} = \sum_{i=1}^{N_{t,x}} \langle v^i, w^i\rangle_{V_{t,x}^i},$$

and the map $\mathcal{K}_{t,x}f = (\mathcal{K}_{t,x}^1 f,\dots,\mathcal{K}_{t,x}^{N_{t,x}} f)$ is a linear operator from $\mathcal{T}$ to $V_{t,x}$. A linear parametrization of the generator $\mathcal{L}_t$ is therefore given by

$$\mathcal{L}_t f(x) = \langle \mathcal{K}_{t,x}f, F_t(x)\rangle_{V_{t,x}}.$$

Consider the separable Bregman divergence $D_{t,x} : \Omega_{t,x}\times\widehat{\Omega}_{t,x}\to\mathbb{R}$ acting on each component by way of an individual Bregman divergence $D_{t,x}^i : \Omega_{t,x}^i\times\widehat{\Omega}_{t,x}^i\to\mathbb{R}$, so that

$$D_{t,x}(y,x) = \sum_{i=1}^{N_{t,x}} D_{t,x}^i(y^i,x^i).$$

By Corollary B.4, Lemma B.2 applies to $D_{t,x}$ over $\Omega_{t,x}\times\widehat{\Omega}_{t,x}$, from which the result follows. □

F.5 FLOW MATCHING X_1 PREDICTION

F.5.1 PROOF OF THEOREM 4.13

Proof. We write the infinitesimal generator corresponding to the affine conditional vector field as

$$\begin{aligned}
\mathcal{L}_t^{x_1} f(x) &= \nabla f(x)^T(A_{t,x}x_1 + b_{t,x}) \\
&= \langle A_{t,x}^T\nabla f(x), x_1\rangle_{\mathbb{R}^n} + \langle b_{t,x}^T\nabla f(x), 1\rangle_{\mathbb{R}}.
\end{aligned}$$

The above is a sum of linear parametrizations as in Section 4.5. The model parametrization can be written

$$\begin{aligned}
\mathcal{L}_t^\theta f(x) &= \nabla f(x)^T(A_{t,x}\hat{x}_1^\theta(t,x) + b_{t,x}) \\
&= \langle A_{t,x}^T\nabla f(x), \hat{x}_1^\theta(t,x)\rangle_{\mathbb{R}^n} + \langle b_{t,x}^T\nabla f(x), 1\rangle_{\mathbb{R}},
\end{aligned}$$

from which the CGM loss follows by Theorem 4.12. □

F.6 Diffusion Models and x_0-Prediction

F.6.1 Proof of Theorem 4.17

Proof. We write and linearly parametrize the conditional generator as follows:

$$\begin{aligned}\mathcal{L}_t^z f(x) &= \nabla f(x)^\top \frac{\sigma_t^2}{2} \nabla_x \log p_t(x|z) \\ &= \langle \frac{\sigma_t^2}{2} \nabla f(x), \nabla_x \log p_t(x|z) \rangle_{\mathbb{R}^n}.\end{aligned}$$

Since $\nabla_x \log p_t(x|x_0) = A_{t,x} x_0 + b_{t,x}$ is affine in x_0, the argument proceeds analogously to the flow matching case (Theorem 4.13), yielding the stated CGM loss. □

F.7 Rescaling Time-Dependent Constants out of Jump Models

F.7.1 Proof of Theorem 4.18

Proof. We define

$$\lambda_{\text{total}}(t,x) = \int_S Q_t(dy;x) = \sum_{j=1}^{N_{t,x}} h_{t,j}(x) R_{t,j}(x)$$

for each $x \in S$ and $t \in [0,1)$, and we require that $\lambda_{\text{total}}(t,x) < \infty$. The associated process X_t can be described by:

$$X_{t+\Delta t} = \begin{cases} X_t & \text{with probability} \quad 1 - \Delta t \lambda_{\text{total}}(t,x) + o(\Delta t) \\ \sim J_t(dy;x) & \text{with probability} \quad \Delta t \lambda_{\text{total}}(t,x) + o(\Delta t), \end{cases}$$

where in the above, the distribution $J_t(dy;x)$ is the normalized jump distribution:

$$J_t(dy;x) = \frac{Q_t(dy;x)}{\lambda_{\text{total}}(t,x)} = \frac{1}{\lambda_{\text{total}}(t,x)} \sum_{j=1}^{N_{t,x}} h_{t,j}(x) R_{t,j}(x) \delta_{\Gamma_{t,j}(x)}(dy)$$

having infinitesimal generator

$$\mathcal{L}_t f(x) = \int (f(y) - f(x)) Q_t(dy;x) = \sum_{j=1}^{N_{t,x}} (f(\Gamma_{t,j}(x)) - f(x)) h_{t,j}(x) R_{t,j}(x).$$

Write $R_t(x) = (R_{t,j}(x))_{j=1}^{N_{t,x}}$. A time and state dependent linear parametrization of the generator is given by

$$\mathcal{L}_t f(x) = \langle \mathcal{K}_t f(x), R_t(x) \rangle_{V_{t,x}},$$

where $V_{t,x} = \mathbb{R}^{N_{t,x}}$ is equipped with the inner product

$$\langle v, w \rangle_{V_{t,x}} = \sum_{j=1}^{N_{t,x}} h_{t,j}(x) v^j w^j,$$

and the linear map $\mathcal{K}_{t,x} : \mathcal{T} \to \mathbb{R}^{N_{t,x}}$ sends $\mathcal{K}_{t,x} f = (f(\Gamma_{t,i}(x)) - f(x))_{i=1}^{N_{t,x}}$.

For latent $z \in \mathcal{Z}$ we specify a conditional atomic time-dependent jump kernel through conditional rate multipliers $R_{t,j}^z(x)$:

$$Q_t^z(dy;x) = \sum_{j=1}^{N_{t,x}} h_{t,j}(x) R_{t,j}^z(x) \delta_{\Gamma_{t,j}(x)}(dy),$$

making the same assumptions on $h_{t,j}(x)$, $R_{t,j}^z$, $\lambda_{\text{total}}^z(t,x) := \sum_{j=1}^{N_{t,x}} h_{t,j}(x) R_{t,j}^z(x)$ and on the marginal rates $R_{t,j}(x)$ and $\lambda_{\text{total}}(t,x)$. In particular, we have

$$\lambda_{\text{total}}(x) = \mathbb{E}_{Z \sim p_{Z|t}(dz|x)}[\lambda_{\text{total}}^z(t,x)] < \infty.$$

Using the elementary inequality $\sum_i a_i^2 \leq (\sum_i a_i)^2$ for $a_i \geq 0$, we can show that $\mathbb{E}_{Z\sim p_{Z|t}(dz|x)}\|R_t^Z(x)\|_{V_{t,x}} < \infty$, since

$$\|R_t^z(x)\|_{V_{t,x}} = \left(\sum_{j=1}^{N_{t,x}} h_{t,j}(x) R_{t,j}^z(x)^2\right)^{1/2} \leq \sum_{j=1}^{N_{t,x}} \sqrt{h_{t,j}(x)} R_{t,j}^z(x)$$
$$\leq C_{t,x}\lambda_{\text{total}}^z(t,x)$$

where $C_{t,x} = \max_j\{I_{h_{t,j}(x)>0} \cdot h_{t,j}(x)^{-1/2}\}$ is independent of z, so it follows that

$$\mathbb{E}_{Z\sim p_{Z|t}(dz|x)}\|R_t^Z(x)\|_{V_{t,x}} \leq C_{t,x}\mathbb{E}_{Z\sim p_{Z|t}(dz|x)}[\lambda_{\text{total}}^z(x)] < \infty.$$

This satisfies the condition for Lemma E.1, so the stated CGM loss is valid under the conditions in Section 4.2. In particular, the hazard rates $h_{t,j}(x)$ are not present in the loss, so the loss can be taken directly against X_1-predictions (e.g., in DFM, the posterior $\hat{w}_t^j(x^i|X_t)$). □