1 Discrete Flow Matching

This section develops *Discrete Flow Matching (DFM)*: a flow-matching-style recipe for discrete data based on *continuous-time Markov chains (CTMCs)*. The construction mirrors the continuous-state flow matching framework, but replaces deterministic ODE transport with stochastic jumps on a finite state space. This section largely follows the presentation of Lipman et al. (2024), Gat et al. (2024), and Campbell et al. (2024).

1.1 CTMCs on discrete state spaces

1.1.1 Discrete state space and PMFs

We consider a finite discrete version of $\mathbb{R}^d$ as our state space:

$$S = \mathcal{T}^d, \qquad \mathcal{T} = [K] = \{1, 2, \ldots, K\},$$

sometimes referred to as a *vocabulary*. Samples and states are denoted by sequences $x = (x^1, \ldots, x^d) \in \mathcal{S}$, where $x^i \in \mathcal{T}$ is single coordinate or a token.

Let X be a random variable valued in S, with a probability mass function (PMF) $p_X : S \to \mathbb{R}_{\geq 0}$ such that $\sum_{x \in S} p_X(x) = 1$. For any event $A \subseteq S$,

$$\mathbb{P}(X \in A) = \sum_{x \in A} p_X(x). \tag{1.1}$$

Definition 1 (Discrete delta PMF)

For $x, z \in S$, define the (discrete) delta PMF

$$\delta(x, z) \triangleq \begin{cases} 1, & x = z, \\ 0, & \text{else.} \end{cases} \tag{1.2}$$

We will also use delta PMFs on tokens, such as $\delta(x^i, y^i)$ for $x^i, y^i \in \mathcal{T}$.

1.1.2 CTMC generative model and rate conditions

A CTMC model is an S-valued time-dependent family of random variables $(X_t)_{0 \leq t \leq 1}$ forming a (time-inhomogeneous) Markov chain. It is characterized by a *probability transition kernel* $p_{t+h \mid t}$

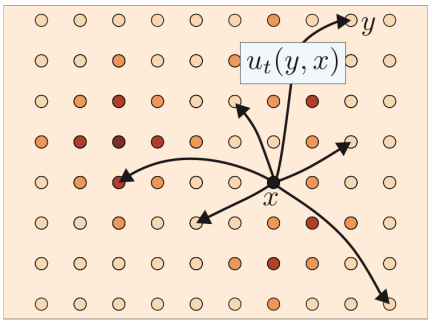

Figure 1: A CTMC is defined by prescribing rates/velocities $u_t(y, x)$ of transitioning from state x to state y.

defined (for small $h > 0$) by

$$p_{t+h\,|\,t}(y|x) \triangleq \mathbb{P}(X_{t+h} = y \mid X_t = x) = \delta(y, x) + h\, u_t(y, x) + o(h), \qquad \mathbb{P}(X_0 = x) = p(x). \quad (1.3)$$

Here $u_t(y, x)$ are called *rates* or *velocities* and $o(h)$ is any function satisfying $o(h)/h \to 0$ as $h \to 0$.

In order for (1.3) to define a valid PMF for each fixed x and small h, the velocities must satisfy the *rate conditions*:

$$u_t(y, x) \geq 0 \quad \text{for all } y \neq x, \qquad \sum_{y \in S} u_t(y, x) = 0. \quad (1.4)$$

The second condition forces $u_t(x, x) = -\sum_{y \neq x} u_t(y, x) \leq 0$. If one of these conditions were to fail, then the transition probabilities $p_{t+h\,|\,t}(|x)$ would become negative or sum to $c \neq 1$ for arbitrary small $h > 0$.

We say that a velocity field u_t *generates* a marginal probability path $(p_t)_{t \in [0,1]}$ if there exists a kernel $p_{t+h\,|\,t}$ satisfying (1.3) whose marginals are p_t.

1.1.3 Simulating a CTMC

To simulate a CTMC, a naive forward-Euler step would use

$$\mathbb{P}(X_{t+h} = y \mid X_t) = \delta(y, X_t) + h\, u_t(y, X_t). \quad (1.5)$$

However, since (1.3) only holds up to $o(h)$, the right-hand side of (1.5) may fail to be a valid PMF unless h is sufficiently small.

A standard remedy is to use the exponential-form Euler step

$$\mathbb{P}(X_{t+h} = y \mid X_t) = \begin{cases} \exp\big(h\, u_t(X_t, X_t)\big), & y = X_t, \\ \dfrac{u_t(y, X_t)}{|u_t(X_t, X_t)|}\Big(1 - \exp\big(h\, u_t(X_t, X_t)\big)\Big), & y \neq X_t, \end{cases} \quad (1.6)$$

which preserves the $o(h)$ local error while guaranteeing a valid PMF for all $h > 0$.

1.1.4 Probability paths and the Kolmogorov equation

Let p_t denote the marginal PMF of X_t. Similarly to Continuity Equation in the continuous case, the marginal probabilities $(p_t)_{t\in[0,1]}$ are characterized by the *Kolmogorov forward equation*:

$$\frac{\mathrm{d}}{\mathrm{d}t}\, p_t(y) = \sum_{x\in S} u_t(y,x)\, p_t(x). \tag{1.7}$$

The following classical theorem (see also Theorems 5.1 and 5.2 in Coddington et al. (1956)) describes the existence of unique solutions for this linear homogeneous system of ODEs.

Theorem 2 (Linear ODE existence and uniqueness)

If $u_t(y,x)$ is continuous in t for each $x, y \in S$, then there exists a unique solution $p_t(x)$ to the Kolmogorov equation (1.7) for $t \in [0,1)$ satisfying the initial condition $p_0(x) = p(x)$.

For the CTMC, the solution is guaranteed to exist for all times $t \in [0,1)$ and no extra conditions are required. Using the rate conditions (1.4), the right-hand side of (1.7) can be rewritten in a "flux divergence" form. Define the *probability flux*

$$j_t(y,x) \triangleq u_t(y,x)\, p_t(x),$$

describing the probability of moving from state x to state y per unit of time. Then

$$\sum_{x\in S} u_t(y,x)p_t(x) = \sum_{x\neq y} u_t(y,x)p_t(x) - \sum_{x\neq y} u_t(x,y)p_t(y) = -\sum_{x\neq y} \big[j_t(x,y) - j_t(y,x)\big],$$

where the first term is the *incoming* flux to y and the second term is the *outgoing* flux from y.

The following result is the main tool to build probability paths and velocities in the CTMC framework.

Theorem 3 (Discrete mass conservation)

Let $u_t(y,x)$ be continuous in t and let $p_t(x)$ be a PMF that is differentiable in t. Then the following are equivalent:

1. p_t and u_t satisfy the Kolmogorov equation (1.7) on $t \in [0,1)$ and u_t satisfies the rate conditions (1.4).
2. u_t generates p_t (i.e. there exists a kernel $p_{t+h|t}$ satisfying (1.3) with marginals p_t).

Proof of Theorem 3. We first need the following lemma.

Lemma 4 (PMF solutions to Kolmogorov with rate conditions)

Consider a solution $f_t(x)$ to Kolmogorov Equation (1.7) with initial condition $f_0(x) = p(x)$, where p is a PMF. $u_t(y, x)$ is $C([0, 1])$ in time and satisfies the rate conditions (1.4). Then $f_t(x)$ is a probability mass function (PMF) for all $t \in [0, 1]$.

Proof of Lemma 4. Let $f_t(x)$, $t \in [0, 1]$, be the solution to the Kolmogorov Equation, the existence and uniqueness of which is guaranteed by Theorem 2. Now, $f_t(x)$ is a PMF if and only if it satisfies

$$f_t(x) \geq 0, \qquad \text{and} \qquad \sum_x f_t(x) = 1. \tag{1.8}$$

The latter condition is shown to hold by summing both sides of the Kolmogorov Equation to get that the solution satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t} \sum_x f_t(x) = \sum_x \sum_z u_t(x, z)\, f_t(z) = 0,$$

where the second equality is due to $\sum_y u_t(y, x) = 0$ in the rate conditions. Since $\sum_x f_0(x) = \sum_x p(x) = 1$ we have that $\sum_x f_t(x) \equiv 1$ for all $t \in [0, 1]$.

To prove that $f_t(x) \geq 0$ for all $x \in \mathcal{S}$ we will use a result on convex invariant sets of dynamical systems. In particular, Theorem 7.3.4 in Prüss et al. (2010) asserts that as long as $f_0 = p$ satisfies this condition (which it does) and whenever $w(z)$ is on the boundary of this constraint, i.e., w is a PMF and $w(z) = 0$ for some $z \in \mathcal{S}$, then a nonnegative inner product with the outer normal to the constraint, i.e.,

$$\sum_{x,y} u_t(y, x)\, w(x)\, \delta(y, z) \geq 0,$$

implies that the solution $f_t(x) \geq 0$ for all $t \in [0, 1]$ and $x \in \mathcal{S}$. Let us check this condition:

$$\sum_{x,y} u_t(y, x)\, w(x)\, \delta(y, z) = \sum_x u_t(z, x)\, w(x) = \sum_{x \neq z} u_t(z, x)\, w(x) \geq 0,$$

where in the second equality we use the fact that $w(z) = 0$ and in the last inequality we used the rate condition (1.4) that $u_t(z, x) \geq 0$ for $z \neq x$ and $w(x) \geq 0$ for all x. □

Let us start by assuming 2. In this case, the probability transition kernel $p_{t+h|t}(y \mid x)$ satisfies (1.3),

$$p_{t+h|t}(y \mid x) = \delta(y, x) + h\, u_t(y, x) + o(h). \tag{1.9}$$

By expressing the marginal $p_t(y)$ using the Law of Total Probability, we obtain

$$p_{t+h}(y) = \sum_x p_{t+h|t}(y \mid x)\, p_t(x). \tag{1.10}$$

By plugging (1.9) into (1.10) and rearranging, we get

$$\frac{p_{t+h}(y) - p_t(y)}{h} = \sum_x u_t(y, x)\, p_t(x) + o(1),$$

where now $o(1) = o(h)/h \to 0$ as $h \to 0$, as per the definition of $o(h)$. Taking the limit $h \to 0$, we get that the pair (p_t, u_t) satisfies the Kolmogorov Equation (1.7). Next, let us prove that u_t satisfies the rate conditions (1.4). If $u_t(y, x) < 0$ for some $y \neq x$, it follows from (1.9) that $p_{t+h|t}(y \mid x) < 0$ for small $h > 0$, and this contradicts $p_{t+h|t}$ being a probability kernel. If $\sum_y u_t(y, x) = c \neq 0$, it follows from (1.9) that

$$1 = \sum_y p_{t+h|t}(y \mid x) = 1 + h\, c + o(h),$$

leading to a contradiction for small $h > 0$.

Conversely, assume now condition 1. That is, the pair (u_t, p_t) satisfies the Kolmogorov Equation (1.7) with initial condition $p_0 = p$. Fix $t \in [0, 1)$ and $x \in S$. By Theorem 2, let $p_{s|t}(y \mid x)$ be the unique solution of

$$\frac{\mathrm{d}}{\mathrm{d}s} p_{s|t}(y \mid x) = \sum_z u_s(y, z)\, p_{s|t}(z \mid x), \tag{1.11}$$

with initial condition $p_{t|t}(y \mid x) = \delta(y, x)$, where $0 \leq t \leq s < 1$ and t and x are fixed. By Lemma 4, $p_{s|t}(\cdot \mid x)$ is a PMF for each x.

Now define

$$g_s(y) \triangleq \sum_x p_{s|0}(y \mid x)\, p(x).$$

Differentiating and using (1.11) (with $t = 0$) gives

$$\frac{\mathrm{d}}{\mathrm{d}s} g_s(y) = \sum_z u_s(y, z)\, g_s(z), \qquad g_0(y) = p(y). \tag{1.12}$$

Hence g_s satisfies the same Kolmogorov equation and initial condition as p_s. By uniqueness (Theorem 2), $g_s(y) = p_s(y)$ for all s, i.e.

$$\sum_x p_{s|0}(y \mid x)\, p(x) = p_s(y).$$

Lastly, the semigroup property of the transition kernel, $\sum_z p_{s|r}(y \mid z) p_{r|t}(z \mid x) = p_{s|t}(y \mid x)$ for $0 \leq t \leq r \leq s < 1$, can be shown by repeating the argument in (1.12) with $p_{r|t}$ as initial condition at time r. In conclusion, we found a transition kernel $p_{t+h|t}$ that generates p_t, as required.

□

1.1.5 Probability-preserving velocities

If a velocity field u_t generates p_t, then we can add a *divergence-free* component without changing the marginal path.

Fact 5 (Divergence-free perturbations)

If u_t generates p_t, then $\widetilde{u}_t(y, x) = u_t(y, x) + v_t(y, x)$ also generates p_t provided:

a. v_t satisfies the rate conditions (1.4), and

b. v_t satisfies the *divergence-free* condition

$$\sum_{x \in S} v_t(y, x)\, p_t(x) = 0 \qquad \text{for all } y \in S. \tag{1.13}$$

Indeed, (1.13) implies that $\sum_x \widetilde{u}_t(y,x) p_t(x) = \sum_x u_t(y,x) p_t(x) = \dot{p}_t(y)$, so $\widetilde{u}_t$ satisfies the same Kolmogorov equation as u_t.

1.2 Discrete Flow Matching (DFM)

Discrete Flow Matching adapts the flow matching blueprint to discrete state spaces. The goal is to transport samples from a *source* PMF p to a *target* PMF q by learning a CTMC velocity field that generates a prescribed probability path p_t.

1.2.1 Data coupling

Let $X_0 \sim p$ and $X_1 \sim q$ be random variables valued in S. DFM allows either:

- *independent coupling:* $(X_0, X_1) \sim p(X_0)\, q(X_1)$, or
- a general coupling $(X_0, X_1) \sim \pi_{0,1}(x_0, x_1)$.

Couplings are useful, for instance, when x_0 and x_1 represent aligned objects (e.g. translation pairs), or when p is a simple "noise" distribution (e.g. uniform over S).

1.2.2 Discrete probability paths

A probability path is a family of PMFs $(p_t)_{t \in [0,1]}$ interpolating between $p_0 = p$ and $p_1 = q$. As in the continuous setting, it is convenient to define such paths through a conditioning random variable $Z \sim p_Z$ taking values in some space $\mathcal{Z}$. The *marginal probability path* is

$$p_t(x) = \sum_{z \in \mathcal{Z}} p_{t|Z}(x|z)\, p_Z(z), \tag{1.14}$$

where $p_{t|Z}(\cdot|z)$ is a conditional PMF.

1.2.3 The marginalization trick in the discrete setting

Assume that for each z, a conditional velocity $u_t(\cdot, \cdot|z)$ generates the conditional path $p_{t|Z}(\cdot|z)$. Then we can obtain a *marginal* velocity by averaging over the posterior $p_{Z|t}(\cdot|x)$:

$$u_t(y, x) = \sum_{z \in \mathcal{Z}} u_t(y, x|z)\, p_{Z|t}(z|x) = \mathbb{E}[u_t(y, X_t|Z) \,|\, X_t = x]\,. \tag{1.15}$$

Bayes' rule gives

$$p_{Z|t}(z|x) = \frac{p_{t|Z}(x|z)\, p_Z(z)}{p_t(x)}. \tag{1.16}$$

Assumption 6 (Regularity and positivity)

Assume

- $p_{t|Z}(x|z) \in C^1([0,1))$ for all x, z,
- $u_t(y, x|z) \in C([0,1))$ for all x, y, z, and
- $p_t(x) > 0$ for all $x \in S$ and $t \in [0,1)$.

Remark 1

The positivity condition $p_t(x) > 0$ for all x, t is mild in practice. One can enforce it, for example, by mixing the path with a small uniform component (or any full-support distribution) so that every state has nonzero probability at intermediate times.

Theorem 7 (Discrete marginalization trick)

Under Assumption 6, if $u_t(y, x|z)$ generates $p_{t|Z}(x|z)$ for each z, then the marginal velocity (1.15) generates the marginal path (1.14).

Proof. Differentiate (1.14):

$$\frac{\mathrm{d}}{\mathrm{d}t} p_t(y) = \sum_z \frac{\mathrm{d}}{\mathrm{d}t} p_{t|Z}(y|z)\, p_Z(z).$$

Since $u_t(\cdot, \cdot|z)$ generates $p_{t|Z}(\cdot|z)$, it satisfies the Kolmogorov equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} p_{t|Z}(y|z) = \sum_x u_t(y, x|z)\, p_{t|Z}(x|z).$$

Substituting and rearranging,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} p_t(y) &= \sum_z \Big[\sum_x u_t(y, x|z) p_{t|Z}(x|z) \Big] p_Z(z) \\
&= \sum_x \Big[\sum_z u_t(y, x|z) \frac{p_{t|Z}(x|z) p_Z(z)}{p_t(x)} \Big]\, p_t(x) \\
&= \sum_x u_t(y, x)\, p_t(x),
\end{aligned}$$

where we used Bayes' rule (1.16) in the last step. Thus u_t satisfies the Kolmogorov equation for p_t. Moreover, since each conditional velocity satisfies the rate conditions and u_t is a convex combination of them, u_t also satisfies the rate conditions. By discrete mass conservation (Theorem 3), u_t generates p_t. □

1.2.4 Discrete flow matching loss

In DFM we parameterize a learnable velocity field $u_t^\theta(y, x)$ (e.g. a neural network) and train it to match the *target* velocity that generates the prescribed path. Since $u_t(\cdot, x)$ must satisfy the rate constraints (1.4), define for each $x \in S$ the convex set

$$\Omega_x \triangleq \left\{ v \in \mathbb{R}^S \;\middle|\; v(y) \geq 0 \;\forall y \neq x, \quad \text{and} \quad v(x) = -\sum_{y \neq x} v(y) \right\}. \tag{1.17}$$

Let $D_x(\cdot, \cdot)$ be a Bregman divergence on Ω_x (generated by some convex $\Phi_x : \Omega_x \to \mathbb{R}$). The *Discrete Flow Matching loss* is

$$\mathcal{L}_{\text{DFM}}(\theta) = \mathbb{E}_{t \sim \text{Unif}[0,1],\, X_t \sim p_t} \Big[D_{X_t}\big(u_t(\cdot, X_t),\, u_t^\theta(\cdot, X_t)\big) \Big]. \tag{1.18}$$

The *conditional* DFM loss is

$$\mathcal{L}_{\text{CDFM}}(\theta) = \mathbb{E}_{t \sim \text{Unif}[0,1],\, Z \sim p_Z,\, X_t \sim p_{t|Z}(\cdot|Z)} \Big[D_{X_t}\big(u_t(\cdot, X_t|Z),\, u_t^\theta(\cdot, X_t)\big) \Big]. \tag{1.19}$$

As in continuous flow matching, these two losses yield the same learning gradients.

Theorem 8 (CDFM and DFM have the same gradients)

The gradients of (1.18) and (1.19) coincide:

$$\nabla_\theta \mathcal{L}_{\text{DFM}}(\theta) = \nabla_\theta \mathcal{L}_{\text{CDFM}}(\theta). \tag{1.20}$$

In particular, the minimizer of the conditional loss satisfies

$$u_t^\theta(y, x) = \mathbb{E}[u_t(y, X_t|Z) \,|\, X_t = x]. \tag{1.21}$$

1.3 Factorized paths and velocities

A direct parameterization of $u_t^\theta(y, x)$ is infeasible when $S = \mathcal{T}^d$ is large, since it would require outputting rates for all $y \in S$ (dimension K^d). A standard remedy is to use *factorized velocities*, which only allow transitions that change a single coordinate. Factorized velocities have been used extensively in discrete diffusion/flow models (e.g. Campbell et al. (2022, 2024)).

1.3.1 Factorized velocities

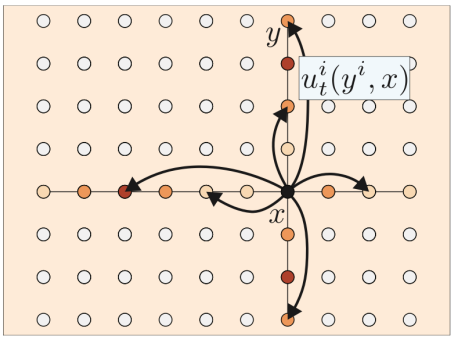

Figure 2: Factorized CTMC: nonzero rates only connect states that differ in at most one coordinate (token).

Definition 9 (Factorized velocity)

Let $\bar{i}$ denote all indices except i. For $x, y \in S$, define $\bar{x}^i = (x^1, \ldots, x^{i-1}, x^{i+1}, \ldots, x^d)$ and similarly $\bar{y}^i$. A velocity field u_t is *factorized* if it can be written as

$$u_t(y, x) = \sum_{i=1}^{d} \delta(\bar{y}^i, \bar{x}^i)\, u_t^i(y^i, x), \tag{1.22}$$

where each coordinate velocity $u_t^i(\cdot, x)$ assigns rates to changing token x^i into y^i while keeping all other coordinates fixed.

The coordinate-wise rate constraints become, for each $i \in [d]$,

$$u_t^i(y^i, x) \geq 0 \quad \forall\, y^i \neq x^i, \qquad \sum_{y^i \in \mathcal{T}} u_t^i(y^i, x) = 0 \quad \forall\, x \in S. \tag{1.23}$$

1.3.2 Coordinate-wise sampling

With factorized velocities, the infinitesimal transition kernel factorizes coordinate-wise (up to $o(h)$). Starting from (A coordinate-wise CTMC simulation viewpoint is given in Campbell et al. (2024).)

$$\mathbb{P}(X_{t+h} = y \mid X_t = x) = \delta(y, x) + h \sum_i \delta(\bar{y}^i, \bar{x}^i)\, u_t^i(y^i, x) + o(h),$$

one obtains

$$\mathbb{P}(X_{t+h} = y \mid X_t = x) = \prod_{i=1}^{d} \Big[\delta(y^i, x^i) + h\, u_t^i(y^i, x) + o(h)\Big],$$

using $\delta(y, x) = \prod_i \delta(y^i, x^i)$ and the identity

$$\prod_i (a^i + h b^i) = \prod_i a^i + h \sum_i \Big(\prod_{j \neq i} a^j\Big) b^i + o(h).$$

Therefore, the coordinate-wise transition is

$$\mathbb{P}(X^i_{t+h} = y^i \mid X_t = x) = \delta(y^i, x^i) + h\, u^i_t(y^i, x) + o(h). \tag{1.24}$$

Each coordinate can be sampled using the CTMC Euler method (1.6).

1.3.3 Factorized probability paths induce factorized velocities

Factorization can also arise on the *path* side.

Definition 10 (Factorized probability path)

A probability path $(q_t)_{t\in[0,1]}$ on $S = \mathcal{T}^d$ is *factorized* if

$$q_t(x) = \prod_{i=1}^{d} q^i_t(x^i). \tag{1.25}$$

Proposition 11 (Factorized path ⇒ factorized velocity)

Let q_t be factorized as in (1.25). Suppose for each i there exists a (token-level) velocity $u^i_t(y^i, x^i)$ that generates $q^i_t(x^i)$. Then q_t has a factorized generating velocity

$$u_t(y, x) = \sum_{i=1}^{d} \delta(\bar{y}^i, \bar{x}^i)\, u^i_t(y^i, x^i). \tag{1.26}$$

Proof. Fix $y \in S$ and differentiate $q_t(y) = \prod_{i=1}^d q^i_t(y^i)$:

$$\frac{\mathrm{d}}{\mathrm{d}t} q_t(y) = \sum_{i=1}^{d} \left(\frac{\mathrm{d}}{\mathrm{d}t} q^i_t(y^i) \right) \prod_{j\neq i} q^j_t(y^j).$$

Since u^i_t generates q^i_t, we have $\frac{\mathrm{d}}{\mathrm{d}t} q^i_t(y^i) = \sum_{x^i\in\mathcal{T}} u^i_t(y^i, x^i)\, q^i_t(x^i)$. Substituting,

$$\frac{\mathrm{d}}{\mathrm{d}t} q_t(y) = \sum_{i=1}^{d} \sum_{x^i\in\mathcal{T}} u^i_t(y^i, x^i)\, q^i_t(x^i) \prod_{j\neq i} q^j_t(y^j).$$

For each i, identify a full state $x = (y^1, \ldots, y^{i-1}, x^i, y^{i+1}, \ldots, y^d)$ so that $\delta(\bar{y}^i, \bar{x}^i) = 1$ and note that $q_t(x) = q^i_t(x^i) \prod_{j\neq i} q^j_t(y^j)$. Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t} q_t(y) = \sum_{i=1}^{d} \sum_{x\in S} \delta(\bar{y}^i, \bar{x}^i)\, u^i_t(y^i, x^i)\, q_t(x) = \sum_{x\in S} u_t(y, x)\, q_t(x).$$

Thus u_t satisfies the Kolmogorov equation (1.7) for q_t. The rate conditions follow from those of the coordinate rates u^i_t, hence u_t generates q_t by Theorem 3. □

For a general PMF $q(x)$ on S, it is useful to denote its marginals

$$q^i(x^i) \triangleq \sum_{\bar{x}^i} q(x), \qquad q^{\bar{i}}(\bar{x}^i) \triangleq \sum_{x^i} q(x). \tag{1.27}$$

1.3.4 Factorized marginalization trick

Now consider a marginal path defined by conditioning, but where the conditional paths are factorized.

Theorem 12 (Discrete factorized marginalization trick)

Consider a marginal probability path constructed as

$$p_t(x) = \sum_z p_{t|Z}(x|z)\, p_Z(z), \qquad \text{with} \qquad p_{t|Z}(x|z) = \prod_{i=1}^{d} p^i_{t|Z}(x^i|z). \tag{1.28}$$

Assume $p_t(x) > 0$ for all x, t, and for each i there exists $u^i_t(y^i, x^i|z) \in C([0,1))$ generating $p^i_{t|Z}(\cdot|z) \in C^1([0,1))$. Then the marginal velocity is factorized:

$$u_t(y, x) = \sum_{i=1}^{d} \delta(\bar{y}^i, \bar{x}^i)\, u^i_t(y^i, x), \tag{1.29}$$

where

$$u^i_t(y^i, x) = \sum_z u^i_t(y^i, x^i|z)\, p_{Z|t}(z|x) = \mathbb{E}\big[u^i_t(y^i, X^i_t|Z) \,\big|\, X_t = x\big]. \tag{1.30}$$

Moreover, u_t generates p_t.

Proof. For each fixed z, the conditional path $p_{t|Z}(\cdot|z)$ factorizes by assumption. Applying Proposition 11 (conditionally on z) shows that it admits a factorized generating velocity

$$u_t(y, x \mid z) = \sum_{i=1}^{d} \delta(\bar{y}^i, \bar{x}^i)\, u^i_t(y^i, x^i \mid z).$$

Now apply the discrete marginalization trick (Theorem 7) to the mixture representation (1.28): the marginal generating velocity is

$$u_t(y, x) = \sum_z u_t(y, x \mid z)\, p_{Z|t}(z|x).$$

Expanding the factorized form and exchanging sums yields

$$\begin{aligned} u_t(y, x) &= \sum_z \sum_{i=1}^{d} \delta(\bar{y}^i, \bar{x}^i)\, u^i_t(y^i, x^i \mid z)\, p_{Z|t}(z|x) \\ &= \sum_{i=1}^{d} \delta(\bar{y}^i, \bar{x}^i) \left[\sum_z u^i_t(y^i, x^i \mid z)\, p_{Z|t}(z|x) \right]. \end{aligned}$$

Defining the bracketed term as $u^i_t(y^i, x)$ gives (1.30) and proves the factorized marginal form (1.29). □

1.3.5 Constructing factorized paths

Theorem 12 suggests the following recipe to design factorized-velocity paths between p and q:

1. Choose factorized conditional paths $p_{t|Z}(x|z) = \prod_i p^i_{t|Z}(x^i|z)$ such that the induced marginal path satisfies $p_0 = p$ and $p_1 = q$.
2. For each coordinate i and each z, find token-level velocities $u^i_t(y^i, x^i|z)$ solving the (token-level) Kolmogorov system

$$\sum_{x^i \in \mathcal{T}} u^i_t(y^i, x^i|z)\, p^i_{t|Z}(x^i|z) = \frac{\mathrm{d}}{\mathrm{d}t}\, p^i_{t|Z}(y^i|z), \qquad \forall\, y^i \in \mathcal{T}. \tag{1.31}$$

   This is an underdetermined linear system with $|\mathcal{T}|$ unknowns (much smaller than $|S|$).

1.3.6 Conditional DFM loss for factorized velocities

When using factorized velocities, we can train coordinate-wise models $u^{\theta,i}_t$ via the loss

$$\mathcal{L}_{\mathrm{CDFM}}(\theta) = \mathbb{E}_{t,Z,X_t\sim p_{t|Z}}\left[\sum_{i=1}^{d} D^i_{X^i_t}\big(u^i_t(\cdot, X^i_t|Z),\, u^{\theta,i}_t(\cdot, X_t)\big)\right], \tag{1.32}$$

where each $D^i_{x^i}$ is a Bregman divergence on the token-level convex set

$$\Omega_\alpha \triangleq \left\{ v \in \mathbb{R}^{\mathcal{T}} \;\middle|\; v(\beta) \geq 0\ \forall \beta \in \mathcal{T}\setminus\{\alpha\},\ \text{ and }\ v(\alpha) = -\sum_{\beta\neq\alpha} v(\beta) \right\}, \qquad \alpha \in \mathcal{T}. \tag{1.33}$$

1.4 Mixture paths for arbitrary couplings

A practical and popular construction is the *mixture path*, which conditions on a coupling pair $Z = (X_0, X_1)$ (with $(X_0, X_1) \sim \pi_{0,1}$). We then define factorized conditional paths Following Gat et al. (2024), we condition on a coupling pair $Z = (X_0, X_1)$ to accommodate arbitrary couplings $\pi_{0,1}$.

$$p_{t|0,1}(x \mid x_0, x_1) = \prod_{i=1}^{d} p^i_{t|0,1}(x^i \mid x_0, x_1). \tag{1.34}$$

1.4.1 Token-level mixture path

Let $\kappa : [0,1] \to [0,1]$ be a C^1 *scheduler* (with $\kappa_0 = 0$ and $\kappa_1 = 1$). Define the token-level conditional path

$$p^i_{t|0,1}(x^i \mid x_0, x_1) = \kappa_t\, \delta(x^i, x^i_1) + (1-\kappa_t)\, \delta(x^i, x^i_0). \tag{1.35}$$

Equivalently, if $X^i_t \sim p^i_{t|0,1}(\cdot \mid x_0, x_1)$ then

$$X^i_t = \begin{cases} x^i_1, & \text{with prob. } \kappa_t, \\ x^i_0, & \text{with prob. } 1-\kappa_t. \end{cases} \tag{1.36}$$

1.4.2 Conditional velocity for the mixture path

To apply Theorem 12, we need a conditional token-level velocity $u_t^i(\cdot, \cdot \mid x_0, x_1)$ generating (1.35). Differentiating (1.35),

$$\frac{\mathrm{d}}{\mathrm{d}t} p_{t|0,1}^i(y^i \mid x_0, x_1) = \dot{\kappa}_t \big[\delta(y^i, x_1^i) - \delta(y^i, x_0^i)\big] = \frac{\dot{\kappa}_t}{1-\kappa_t} \big[\delta(y^i, x_1^i) - p_{t|0,1}^i(y^i \mid x_0, x_1)\big].$$

Using the Kolmogorov form (1.31), one convenient choice is

$$u_t^i(y^i, x^i \mid x_0, x_1) = \frac{\dot{\kappa}_t}{1-\kappa_t} \Big[\delta(y^i, x_1^i) - \delta(y^i, x^i)\Big]. \tag{1.37}$$

1.4.3 Posterior parameterization

Instead of parameterizing velocities directly, we can parameterize the *posterior* distribution of X_1 given X_t. For mixture paths, combining (1.30) with (1.37) yields the marginal token-level velocity

$$\begin{aligned} u_t^i(y^i, x) &= \sum_{x_0, x_1} \frac{\dot{\kappa}_t}{1-\kappa_t} \Big[\delta(y^i, x_1^i) - \delta(y^i, x^i)\Big] \, p_{0,1|t}(x_0, x_1 \mid x) \\ &= \sum_{x_1^i} \frac{\dot{\kappa}_t}{1-\kappa_t} \Big[\delta(y^i, x_1^i) - \delta(y^i, x^i)\Big] \, p_{1|t}^i(x_1^i \mid x), \end{aligned} \tag{1.38}$$

where $p_{0,1|t}(\cdot, \cdot \mid x)$ is the posterior of (X_0, X_1) given $X_t = x$ and

$$p_{1|t}^i(x_1^i \mid x) = \sum_{x_0, \bar{x}_1^i} p_{0,1|t}(x_0, x_1 \mid x) = \mathbb{E}\big[\delta(x_1^i, X_1^i) \,\big|\, X_t = x\big]. \tag{1.39}$$

We can learn $p_{1|t}^i(\cdot \mid x)$ with a neural network $p_{1|t}^{\theta,i}(\cdot \mid x)$ (a discrete analogue of x_1-prediction).

1.4.4 Losses for learning the posterior

One option is a *conditional matching* loss:

$$\mathcal{L}_{\mathrm{CM}}(\theta) = \mathbb{E}_{t, X_0, X_1, X_t} \Big[D_{X_t}\big(\delta(\cdot, X_1^i),\, p_{1|t}^{\theta,i}(\cdot \mid X_t)\big) \Big], \tag{1.40}$$

where D_{X_t} compares PMFs on $\mathcal{T}$. Choosing D to be the KL-divergence gives

$$\mathcal{L}_{\mathrm{CM}}(\theta) = -\,\mathbb{E}_{t, X_0, X_1, X_t} \Big[\log p_{1|t}^{\theta,i}(X_1^i \mid X_t) \Big] + \text{const.} \tag{1.41}$$

Alternatively, we can use the factorized CDFM loss (1.32) with $u_t^{\theta,i}$ parameterized by $p_{1|t}^{\theta,i}$. A convenient Bregman divergence in this setting is the *generalized KL* for nonnegative vectors $u, v \in \mathbb{R}_{\geq 0}^{\mathcal{T}}$:

$$D(u, v) = \sum_j u^j \log \frac{u^j}{v^j} \;-\; \sum_j u^j \;+\; \sum_j v^j. \tag{1.42}$$

For mixture paths, this yields

$$D\big(u_t^i(\cdot, x^i \mid x_0, x_1),\, u_t^{\theta,i}(\cdot, x)\big) = \frac{\dot\kappa_t}{1-\kappa_t}\Big[(\delta(x_1^i, x^i) - 1)\log p_{1|t}^{\theta,i}(x_1^i \mid x) + \delta(x_1^i, x^i) - p_{1|t}^{\theta,i}(x^i \mid x)\Big]. \tag{1.43}$$

This choice also provides an ELBO-style bound (useful for evaluation), as discussed for instance by Shaul et al. (2024):

$$-\log p_1^\theta(x_1) \le \mathbb{E}_{t, X_0, X_t \sim p_{t|0,1}}\left[\sum_{i=1}^d D\big(u_t^i(\cdot, X_t^i \mid X_0, x_1),\, u_t^{\theta,i}(\cdot, X_t)\big)\right], \tag{1.44}$$

where p_1^θ is the model marginal at $t = 1$.

1.4.5 Sampling mixture paths

With the posterior parameterization, coordinate-wise sampling follows from (1.24) and (1.38):

$$\mathbb{P}(X_{t+h}^i = y^i \mid X_t = x) = \delta(y^i, x^i) + h\, u_t^i(y^i, x) + o(h) \tag{1.45}$$

$$= \sum_{x_1^i}\Big[\delta(y^i, x^i) + h\frac{\dot\kappa_t}{1-\kappa_t}\big(\delta(y^i, x_1^i) - \delta(y^i, x^i)\big) + o(h)\Big] p_{1|t}^i(x_1^i \mid x). \tag{1.46}$$

Given $X_t = x$, a single step can be implemented by, for each coordinate i:

(i) draw $X_1^i \sim p_{1|t}^i(\cdot \mid x)$, then

(ii) apply the Euler step (1.6) with velocity $\frac{\dot\kappa_t}{1-\kappa_t}\big[\delta(y^i, X_1^i) - \delta(y^i, x^i)\big]$.

1.4.6 One-sided mixture paths and probability-preserving velocities

The sampling design space can be enlarged by adding divergence-free components (Key Idea 5). For factorized conditional paths, a token-level divergence-free velocity v_t^i should satisfy

$$\sum_{x^i \in \mathcal{T}} v_t^i(y^i, x^i \mid z)\, p_{t|Z}^i(x^i \mid z) = 0. \tag{1.47}$$

A useful case is when the source distribution factorizes, $p(x) = \prod_i p(x^i)$, and the coupling is independent, $\pi_{0,1}(x_0, x_1) = p(x_0)q(x_1)$. Then the marginal mixture path can be written

$$p_t(x) = \sum_{x_1} p_{t|1}(x \mid x_1)\, q(x_1), \qquad p_{t|1}(x \mid x_1) = \prod_i p_{t|1}^i(x^i \mid x_1),$$

with

$$p_{t|1}^i(x^i \mid x_1) = \kappa_t\, \delta(x^i, x_1^i) + (1-\kappa_t)\, p(x^i).$$

The conditional velocity

$$u_t^i(y^i, x^i \mid x_1) = \frac{\dot\kappa_t}{1-\kappa_t}\big[\delta(y^i, x_1^i) - \delta(y^i, x^i)\big] \tag{1.48}$$

generates $p^i_{t|1}(\cdot \mid x_1)$. A backward-time velocity can be constructed as

$$\widetilde{u}^i_t(y^i, x^i \mid x_1) = \frac{\dot{\kappa}_t}{\kappa_t}\big[\delta(y^i, x^i) - p(x^i)\big], \tag{1.49}$$

and the divergence-free component is then

$$v^i_t(y^i, x^i \mid x_1) = u^i_t(y^i, x^i \mid x_1) - \widetilde{u}^i_t(y^i, x^i \mid x_1). \tag{1.50}$$

Thus, a generalized sampling step may use the velocity

$$\begin{aligned} u^i_t(y^i, x^i \mid x_1) = {} & \frac{\dot{\kappa}_t}{1-\kappa_t}\big[\delta(y^i, X^i_1) - \delta(y^i, x^i)\big] \\ & + c_t\left[\frac{\dot{\kappa}_t}{1-\kappa_t}\big[\delta(y^i, x^i_1) - \delta(y^i, x^i)\big] - \frac{\dot{\kappa}_t}{\kappa_t}\big[\delta(y^i, x^i) - p(x^i)\big]\right], \end{aligned} \tag{1.51}$$

where $c_t > 0$ is a time-dependent constant and $X^i_1 \sim p^i_{1|t}(\cdot \mid x)$ is drawn in step (i).

References

Campbell, A., Yim, J., Barzilay, R., Rainforth, T., and Jaakkola, T. (2024). Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv preprint arXiv:2402.04997*.

Coddington, E. A., Levinson, N., and Teichmann, T. (1956). Theory of ordinary differential equations.

Gat, I., Remez, T., Shaul, N., Kreuk, F., Chen, R. T., Synnaeve, G., Adi, Y., and Lipman, Y. (2024). Discrete flow matching. *Advances in Neural Information Processing Systems*, 37:133345–133385.

Lipman, Y., Havasi, M., Holderrieth, P., Shaul, N., Le, M., Karrer, B., Chen, R. T., Lopez-Paz, D., Ben-Hamu, H., and Gat, I. (2024). Flow matching guide and code. *arXiv preprint arXiv:2412.06264*.