GENERATIVE ASSIGNMENT FLOWS FOR REPRESENTING AND LEARNING JOINT DISTRIBUTIONS OF DISCRETE DATA

BASTIAN BOLL, DANIEL GONZALEZ-ALVARADO, STEFANIA PETRA, CHRISTOPH SCHNÖRR

ABSTRACT. We introduce a novel generative model for the representation of joint probability distributions of a possibly large number of discrete random variables. The approach uses measure transport by randomized assignment flows on the statistical submanifold of factorizing distributions, which enables to represent and sample efficiently from any target distribution and to assess the likelihood of unseen data points. The complexity of the target distribution only depends on the parametrization of the affinity function of the dynamical assignment flow system. Our model can be trained in a simulation-free manner by conditional Riemannian flow matching, using the training data encoded as geodesics on the assignment manifold in closed-form, with respect to the e-connection of information geometry. Numerical experiments devoted to distributions of structured image labelings demonstrate the applicability to large-scale problems, which may include discrete distributions in other application areas. Performance measures show that our approach scales better with the increasing number of classes than recent related work.

CONTENTS

1. Introduction 2
1.1. Overview, Motivation 2
1.2. Related Work 4
1.2.1. Statistics 4
1.2.2. Own Prior Work 5
1.2.3. Machine Learning 5
1.3. Organization 6
1.4. Basic Notation, List of Main Symbols 6
2. Background 7
2.1. Assignment Flows 7
2.2. Meta-Simplex, Flow Embedding 8
3. Approach 9
3.1. Generative Model 9
3.1.1. Goal 9
3.1.2. Representation of General Distributions 10
3.1.3. Model Learning and Model Evaluation (Sampling) 11
3.2. Riemannian Flow Matching 12
3.2.1. Representation of Labelings as Training Data 12
3.2.2. Training Criterion 12
3.2.3. Constructing Conditional Fields 13

2010 *Mathematics Subject Classification.* 49Q22, 53B12, 62H35, 68T05, 68U10, 91A22.

Key words and phrases. generative models, discrete random variables, normalizing flows, information geometry, neural ODEs, assignment flows, replicator equation.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG), grant SCHN 457/17-1, within the priority programme SPP 2298: Theoretical Foundations of Deep Learning. This work was funded by the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy EXC-2181/1-390900948 (the Heidelberg STRUCTURES Excellence Cluster).

3.2.4. Infinite Integration Time 15
3.2.5. Relation to Dirichlet Flow Matching 15
3.3. Learning Interaction between Simplices 16
3.4. Numerical Flow Integration 17
3.5. Likelihood Computation 18
3.6. Dequantization 18
4. Experiments and Discussion 19
4.1. Class Scaling 20
4.2. Generating Image Segmentations 20
4.3. Likelihood Evaluation 21
5. Conclusion 22
Appendix A. Proofs 25
A.1. Proofs of Section 3.2.3 25
A.2. Proofs of Section 3.3 27
Appendix B. Experiments: Details 28
B.1. Details of Class Scaling Experiment 28
B.2. Details of Generating Image Segmentations 28
B.2.1. Cityscapes Data Preparation 28
B.2.2. Cityscapes Training 28
B.2.3. Binarized MNIST Data Preparation 28
B.2.4. Binarized MNIST Training 29
Appendix C. Likelihood Computation: Details 29
References 31

1. Introduction

1.1. **Overview, Motivation.** *Generative models* in machine learning define an active area of research [KPB21, PNR$^+$21, RH21]. Corresponding research objectives include

(i) the representation of complex probability distributions,
(ii) efficient sampling from such distributions, and
(iii) computing the likelihoods of unseen data points.

The target probability distribution is typically not given, except for a finite sample set (empirical measure). The modeling task concerns the generation of the target distribution by transporting a simple reference measure, typically the multivariate standard normal distribution, using a corresponding pushforward mapping. This mapping is realized by a network with trainable parameters that are optimized by maximizing the likelihood of the given data or a corresponding surrogate objective which is more convenient regarding numerical optimization. This class of approaches are called *normalizing flows* in the literature.

Discrete joint probability distributions abound in applications, yet have received less attention in the literature on generative models. The recent survey paper [KPB21] concludes with a short paragraph devoted to discrete distributions and the assessment that "the generalization of normalizing flows to discrete distributions remains an open problem". Likewise, the survey paper [PNR$^+$21] briefly discusses generative models of discrete distributions in [PNR$^+$21, Section 5.3]. The authors state that "compared to flows on $\mathbb{R}^D$, discrete flows have notable theoretical limitations". The survey paper [RH21] does not mention at all generative models of discrete distributions.

This paper introduces a novel generative approach for the significant subclass of *discrete* (*categorial*) probability distributions of n random variables y_i taking values in a finite set $\{1, 2, \dots, c\}$,

$$y = (y_1, \dots, y_n)^\top \in [c]^n, \qquad y_i \in [c] := \{1, 2, \dots, c\}, \qquad i \in [n], \qquad c, n \in \mathbb{N}. \tag{1.1}$$

A corresponding distribution p is a look-up table which specifies for any realization α of the discrete random vector y the probability

$$p(\alpha) = p(\alpha_1, \dots, \alpha_n) := \Pr(y = \alpha) = \Pr(y_1 = \alpha_1 \ \wedge \dots \wedge \ y_n = \alpha_n), \qquad \alpha \in [c]^n. \tag{1.2}$$

Any such look-up table is a nonnegative tensor with the combinatorially large number

$$N := c^n \tag{1.3}$$

of entries $p(\alpha)$, $\alpha \in [c]^n$. Furthermore, since $p(\alpha) \geq 0$, $\forall \alpha$, and $\sum_{\alpha \in [c]^n} p(\alpha) = 1$, any distribution p also corresponds to a point $p \in \Delta_N$ of the probability simplex

$$\Delta_N := \{p \in \mathbb{R}^N_{\geq 0} \colon \langle \mathbb{1}_N, p \rangle = 1\}, \qquad p = (p_\alpha)_{\alpha \in [c]^n}, \quad p_\alpha := p(\alpha), \qquad \text{(meta-simplex)} \tag{1.4}$$

where $\mathbb{1}_N := (1, 1, \dots, 1)^\top \in \mathbb{R}^N$.

Thus, we denote with p discrete joint probability distributions using any of the equivalent representations

- as functions $p \colon [c]^n \to [0, 1]$, cf. Eq. (1.2);
- as nonnegative tensors with c^n components $p(\alpha_1, \dots, \alpha_n)$;
- as discrete probability vectors $p \in \Delta_N$ with $N = c^n$ components p_α, where each component specifies the probability $p_\alpha = p(\alpha) = \Pr(y = \alpha)$, cf. Eq. (1.4).

In particular, the N vertices (extreme points)

$$e_\alpha \in \{0, 1\}^N \tag{1.5}$$

of Δ_N are the unit vectors which encode the discrete Dirac measures δ_α concentrated on the realizations $\alpha \in [c]^n$.

Figure 1.1 illustrates the approach for the toy distribution of two binary variables, i.e. $c = 2$ and $N = 2^2 = 4$,

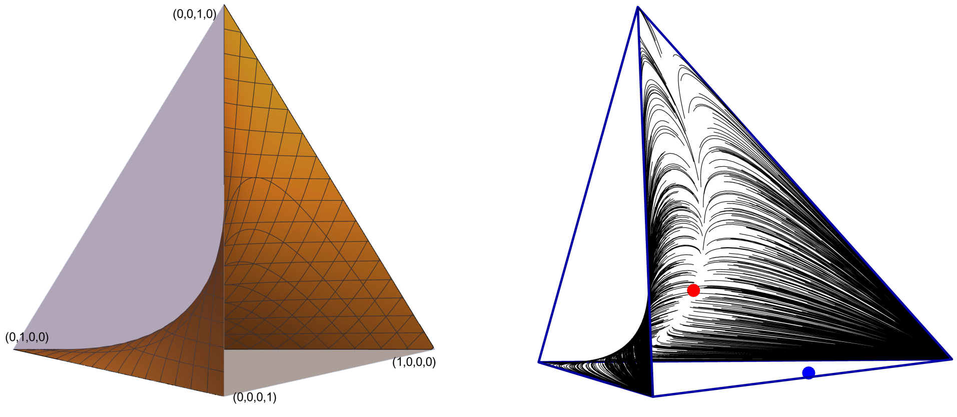
$p(\alpha_1, \alpha_2)$: (1.6)

α_1/α_2	0	1
0	0.45	0.05
1	0.05	0.45

The simplex $\Delta_4 \subset \mathbb{R}^4$ (1.4) is visualized in $\mathbb{R}^3$ in local coordinates as tetrahedron (Figure 1.1(a); see Example 2.1 (p. 8) for details). The generative model only uses the submanifold of *factorizing* discrete distributions which ensures computational efficiency of both training and sampling. Figure 1.1(a) shows that this submanifold connects all extreme points of Δ_4.

Figure 1.1(b) illustrates how sampling from p is accomplished after training, by computing on the submanifold the integral curves of a generating flow which emanates from initial random points, such that each curve converges to a vertex of the simplex which represents a realization $\alpha \sim p(\alpha)$ by (1.5). In this way, a simple reference distribution is pushed forward to p. Figure 3.1 (p. 10) gives a more detailed account of the ingredients of our approach.

Training concerns the parameters of the vector field of the dynamical system, which generates the aforementioned flow on the submanifold. This is achieved by matching the flow to closed-form geodesics on the submanifold which encode given training data. This *flow matching approach* has been recently proposed by [LCBH+23, CL23]. Our paper elaborates this approach for *discrete* joint probability distributions using the geometric approach outlined above.



(a) (b)

FIGURE 1.1. **(a)** The simplex Δ_N (1.4), for $N = 4$, depicted in local coordinates, and the submanifold of factorizing discrete distributions which connects all extreme points of Δ_4. **(b)** Visualization of 1000 samples from the target distribution $p(\alpha_1, \alpha_2)$ given by (1.6), corresponding to the blue point $p \in \Delta_4$. Each sample corresponds to an integral curve of a flow which evolves on the submanifold and can be computed efficiently by geometric integration. The parametrized vector field of the dynamical system which generates the flow has been trainined by matching the flow to geodesics on the submanifold which encode given training data. As a result, each component p_α of the target distribution corresponds to the relative frequency of integral curves converging to the vertex e_α, such that the entire distribution p is represented by the convex combination $\sum_\alpha p_\alpha e_\alpha = p$. In this way, the flow realizes the pushforward of a simple reference distribution, centered at 0 in the tangent space at the barycenter (red point), to the discrete target distribution p. Figure 3.1 (p. 10) provides a more detailled illustration of the approach.

1.2. Related Work.

The central theme of our paper are large joint distributions of discrete random variables which has been a core topic in *multivariate* and *algebraic statistics*, with numerous applications in terms of discrete graphical models in various fields. In addition, our paper contributes to research on *generative models* in *machine learning*. Related work is accordingly reported in Sections 1.2.1 and 1.2.3, respectively, in view of own prior work briefly reported in Section 1.2.2 which combines both viewpoints. The recent related work discussed in Section 1.2.3 reflects the fact that generative models for discrete probability distributions has become an active field of research recently.

1.2.1. *Statistics.* Joint distributions of discrete random variables have a long history in multivariate statistics [Agr13]. This includes the study of subsets of such distributions known as *discrete graphical models* [Lau96, CDLS99, KF09]. Here, conditional independency assumptions encoded by the structure of an underlying graph [Stu05] effectively reduce the degree of freedoms (1.3) of general discrete distributions p and imply their factorization of once realizations of conditioning variables are observed. From the algebraic viewpoint, such statistical assumptions about p give rise to monomial constraints. The study of the topology and geometry of the resulting algebraic varieties which support corresponding subfamilies of distributions, is the subject of the fields of *algebraic statistics* [GMS06, LSX09, DSS09, Zwi16, Sul18]. The special case of fully factorizing discrete distributions

$$p(\alpha) = \prod_{i \in [n]} p_i(\alpha_i) \tag{1.7}$$

is particularly relevant for this paper. For example, the subfamily of all such distributions for the toy case $n = c = 2$, depicted by Figure 1.1, is known as Wright manifold in mathematical game theory [HS98] and more generally as *Segre variety* $\Sigma_{1,1}$ in algebraic geometry [Har92, Lan12].

1.2.2. *Own Prior Work.* Our approach utilizes *assignment flows* [ÅPSS17] that evolve on the relative interior of the product of n probability simplices Δ_c, called *assignment manifold*, one factor for each random variable $y_i,\ i \in [n]$ conforming to the factorization (1.7). As summarized in Section 2.1, the restriction to strictly positive discrete distribution with full support enables to turn these domains into elementary statistical manifolds equipped with the Fisher-Rao geometry and the e-connection of information geometry [AN00]. The corresponding exponential map and the geodesics can be specified in closed form.

Assignment flows are turned into a generative model for discrete random variables as illustrated by Figure 3.1, which generalizes the toy example (1.6) and Figure 1.1: Geometric integration of the assignment flow realizes a map which pushes forward a standard reference measure on the tangent space at the barycenter to the extreme points of the (closure) of the assignment manifold. By embedding the assignment manifold into the simplex (1.4) of all discrete joint distributions, the pushforward measure concentrates on the extreme points and hence represents a more complex *non-factorizing* discrete joint distribution by convex combination of Dirac measures.

Our recent work [BCA$^+$24] characterizes assignment flows as multi-population games and studies multigame dynamics via the aforementioned embedding approach. Some results established in this work regarding the embedding map will be employed in Section 3.3.

1.2.3. *Machine Learning.* The lack of work on generative models for *discrete* distributions stated in the survey papers [KPB21, PNR$^+$21] has stimulated corresponding research recently.

The paper [SJW$^+$24] employs the parametric Dirichlet distribution on the probability simplex [Fer73, JK77, Ait82] as intermediate conditional distributions in a flow matching approach. A similarity to our method is the use of infinite transport time, which achieves favorable scaling in the regime of many classes. A detailed comparison is discussed in Section 3.2.5.

The paper [DKP$^+$24] refers to [ÅPSS17] and a preliminary version [BGAS24] of our generative model and uses geodesics with respect to the Riemannian connection rather than the e-connection, corresponding to $\alpha = 0$ and $\alpha = 1$ in the family of α-connections, respectively [AN00]. By virtue of the sphere map [ÅPSS17, Def. 1] as isometry, the former geodesics on the simplex correspond to the geodesics (great circles) on the sphere with radius 2, restricted to the intersection with the open positive orthant. The authors of [DKP+24] argue that their approach avoids numerical instability at the boundary of the manifold, which is indeed relevant when working on the sphere. However, this issue does not arise on the simplex either, provided that proper geometric numerical integration schemes are used, as demonstrated in [ZSPS20]. The focus of [DKP$^+$24] is on improving the training dynamics using optimal transport, due to the close relation on the simplex of the geometry induced by the Wasserstein distance and the Fisher-Rao geometry [LM18].

Another line of research, called *dequantization*, concerns the approximation of *discrete* probability distributions by *continuous* distributions [UML13, TvdOB16, DSDB17, SKCK17, HCS$^+$19]. A dequantization approach for general discrete data, i.e. similar in scope to our approach, was recently proposed by [CAN22]. We discuss this paper in Section 3.6 and point out differences by showing that our approach can be characterized as dequantization procedure. In particular, we indicate that a key component of the approach [CAN22], learning an embedding of class configurations, can be replicated using our approach, by defining an payoff function of our generative assignment flow approach accordingly.

Regarding the training of our generative model, our approach builds on the recent work [LCBH$^+$23, CL23]. The authors introduced a *flow-matching approach* to the training of continuous generative models which enables more stable and efficient training and hence an attractive alternative to established maximum likelihood training. We adopt this criterion and adapt it to our generative model for discrete distributions and the underlying geometry. In particular, we encode given training data as e-geodesics on the assignment manifold which makes flow matching convenient and effective.

1.3. **Organization.** Section 1.4 fixes the basic notation. Section 2 summarizes the assignment flow approach and specifies the flow embedding into the simplex (1.4), along with mappings and their properties required in the remainder of this paper.

The core Section 3 introduces and details our approach. Section 3.1 introduces the generative model. The flow-matching approach is described in Section 3.2 and how it relates to the recent work [LCBH+23, CL23] which inspired the training component of our approach. Section 3.4 details the particular geometric integration used in all experiments for computing the assignment flow, based on the methods worked out by [ZSPS20]. Section 3.5 explains how the trained generative model is evaluated for computing the likelihoods of a novel unseen data points. Section (3.6) explains dequantization and characterizes our approach from this viewpoint.

Experimental results are presented and discussed in Section 4. We conclude in Section 5.

1.4. **Basic Notation, List of Main Symbols.** We set $[n] := \{1, 2, \dots, n\}$ for $n \in \mathbb{N}$. The canonical Euclidean inner product as well as the matrix inner product which induces the Frobenius norm, are denoted by $\langle \cdot, \cdot \rangle$. The mapping $\operatorname{Diag}(\cdot)$ takes a vector to the diagonal matrix with the vector component as main diagonal entries. e_k, $k \in \mathbb{N}$, denotes a unit vector with single non-zero k-th component equal to 1 and dimension, that is clear from the context.

Data, labelings. $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with vertex set $\mathcal{V} = [n]$, denotes an arbitrary graph on which data x_i are observed at every vertex $i \in \mathcal{V}$. $c \in \mathbb{N}$ possible class labels of the data x_i are represented by discrete random variables $y_i \in [c]$. Realizations of the variables y_i are denoted by $\alpha_i \in [c]$. This results in $N = c^n$ labelings configurations $\alpha = \{\alpha_1, \dots, \alpha_n\}$ for given data $x = \{x_1, \dots, x_n\}$.

Assignment flows, dynamical labelings. The probability simplex is denoted by

$$\Delta_n = \{p \in \mathbb{R}^n_{\geq 0} \colon \langle \mathbb{1}_n, p \rangle = 1\}, \tag{1.8}$$

where $\mathbb{1}_n = (1, 1, \dots, 1)^\top \in \mathbb{R}^n$. *Assignment flows* (Section 2.1) work with the relative interior $\mathring{\Delta}_c$ of Δ_c, denoted by $\mathcal{S}_c := \mathring{\Delta}_c$, containing the strictly positive probability vectors of dimension c, and with the n-fold product conforming to $\mathcal{V}$,

$$\mathcal{W}_c := \mathcal{S}_c \times \dots \times \mathcal{S}_c, \qquad (n = |\mathcal{V}| \text{ factors}) \qquad (\text{assignment manifold}) \tag{1.9}$$

Points on $\mathcal{W}$ are denoted by

$$W = (W_1, \dots, W_n)^\top \in \mathcal{W}_c \subset \mathbb{R}^{n \times c}_{>0}, \qquad W_i \in \mathcal{S}_c, \qquad i \in [n]. \tag{1.10}$$

The evolution $W(t)$ of these assignment vectors, obtained by integrating the assignment flow equation, determines the label assignments α_i to the data point x_i at every $i \in \mathcal{V}$, by convergence to the corresponding unit vectors

$$\lim_{t \to \infty} W_i(t) = e_{\alpha_i} \in \{0, 1\}^c, \quad i \in \mathcal{V}, \tag{1.11}$$

which are the extreme points of the closure of the assignment manifold $\overline{\mathcal{W}_c}$. Further spaces and mappings defined in connection with assignment flows in Section 2.1 are: The tangent spaces $T_0, \mathcal{T}_0$ to $\mathcal{S}_c, \mathcal{W}_c$ with orthogonal projections π_0, Π_0, the barycenters $\mathbb{1}_{\mathcal{S}}, \mathbb{1}_{\mathcal{W}}$ of $\mathcal{S}_c, \mathcal{W}_c$, the Fisher-Rao metric g_p, g_W on $T_0, \mathcal{T}_0$, the replicator maps R_p, R_W and the lifting maps $\exp_p, \exp_W$ which play the role of exponential maps.

Besides the underlying geometry, the essential part of the assignment flow equation, whose integration results in (1.11), is the

$$F_\theta \colon \mathcal{W}_c \to \mathbb{R}^{n \times c}, \qquad (\text{affinity function}) \tag{1.12}$$

whose parameters θ are learned from data.

Meta-simplex, assignment manifold embedding. We overload the symbol p to denote discrete probability distributions using any of the equivalent representations specified after Eq. (1.4), as well as discrete probability vectors whose dimension should be unambigous from the context. Major examples are $p \in \mathcal{S}_c \subset \mathbb{R}^c_{\geq 0}$ and $p \in \Delta_N$ (cf. (1.4)).

Since the embedding

$$\mathcal{T} := T(\mathcal{W}_c) \subset \mathcal{S}_N := \mathring{\Delta}_N \qquad \text{(meta-simplex embedding)} \tag{1.13}$$

of the assignment manifold defined in Section (2.2) yields the submanifold of factorizing distributions in Δ_N, as depicted for a toy scenario by Figure 1.1(a), we call Δ_N as defined by (1.4) "meta-simplex", to distinguish the product of simplices $\mathcal{W}_c$ (1.9) before and after the embedding $T(\mathcal{W}_c)$ (1.13).

We denote by

$$\mathcal{P}(\mathcal{S}_c),\ \mathcal{P}(\mathcal{W}_c),\ \text{etc.} \tag{1.14}$$

the set of probability measures supported on the space $\mathcal{S}_c, \mathcal{W}_c$, etc.

2. Background

Section 2.1 defines spaces and mappings required in the remainder of the paper. Section 2.2 defines a key ingredient of our approach, the embedding (1.13) and related mappings. We refer to the basic notation introduced in Section 1.4.

2.1. **Assignment Flows.**

The basic state space of discrete distributions is the relative interior of the probability simplex

$$\mathcal{S}_c := \mathring{\Delta}_c = \{p \in \mathbb{R}^c \colon p_j > 0,\ \langle \mathbb{1}_c, p\rangle = 1,\ \forall j \in [c]\} \tag{2.1a}$$

with its

$$\mathbb{1}_{\mathcal{S}} := \frac{1}{c}\mathbb{1}_c \in \mathcal{S}_c, \qquad \text{(barycenter)} \tag{2.1b}$$

which becomes the Riemannian manifold $(\mathcal{S}_c, g)$ with trivial tangent bundle $T\mathcal{S}_c = \mathcal{S}_c \times T_0$, comprising the

$$T_0 := T_{\mathbb{1}_{\mathcal{S}}}\mathcal{S}_c := \{v \in \mathbb{R}^c \colon \langle \mathbb{1}_c, v\rangle = 0\} \qquad \text{(tangent space)} \tag{2.1c}$$

with the orthogonal projection

$$\pi_0 \colon \mathbb{R}^c \to T_0, \qquad \pi_0 := I_c - \mathbb{1}_c \mathbb{1}_{\mathcal{S}}^\top \qquad \text{(orthogonal projection)} \tag{2.1d}$$

and carrying the

$$g_p(u,v) := \langle u, \operatorname{Diag}(p)^{-1} v\rangle, \qquad u, v \in T_0, \quad p \in \mathcal{S}_c. \qquad \text{(Fisher-Rao metric)} \tag{2.1e}$$

This naturally extends to the product manifold $(\mathcal{W}_c, g)$ given by (1.9), with trivial tangent bundle $T\mathcal{W}_c = \mathcal{W}_c \times \mathcal{T}_0$, and

$$\mathbb{1}_{\mathcal{W}} = (\mathbb{1}_{\mathcal{S}}, \dots, \mathbb{1}_{\mathcal{S}})^\top, \qquad \text{(barycenter)} \tag{2.2a}$$

$$\mathcal{T}_0 := T_{\mathbb{1}_{\mathcal{W}}}\mathcal{W}_c := T_0 \times \cdots \times T_0, \qquad (n = |\mathcal{V}| \text{ factors}) \qquad \text{(tangent space)} \tag{2.2b}$$

with points denoted by

$$V = (V_1, \dots, V_n)^\top \in \mathbb{R}^{n\times c} \in \mathcal{T}_0, \qquad V_i \in T_0, \qquad i \in [n], \tag{2.2c}$$

the orthogonal projection

$$\Pi_0 \colon \mathbb{R}^{n\times c} \to \mathcal{T}_0, \qquad \Pi_0 U := (\pi_0 U_1, \dots, \pi_0 U_n)^\top \qquad \text{(orthogonal projection)} \tag{2.2d}$$

and the

$$g_W(U,V) = \sum_{i\in[n]} g_{W_i}(U_i,V_i), \qquad U,V\in\mathcal{T}_0, \quad W\in\mathcal{W}_c. \qquad \text{(Fisher-Rao metric)} \tag{2.2e}$$

Assignment flows are dynamical systems of the general form

$$\dot W(t) = R_{W(t)}\big[F_\theta\big(W(t)\big)\big], \qquad W(0)=W_0\in\mathcal{W}_c, \qquad \text{(assignment flow)} \tag{2.3}$$

parametrized by an affinity function (1.12) and comprising the linear mappings

$$R_p\colon \mathbb{R}^c\to T_0, \qquad R_p = \operatorname{Diag}(p)-pp^\top, \qquad p\in\mathcal{S}_c \qquad \text{(replicator map)} \tag{2.4a}$$

$$R_W\colon \mathbb{R}^{n\times c}\to\mathcal{T}_0, \qquad R_W[F_\theta] = (R_{W_1}F_{\theta,1},\dots,R_{W_n}F_{\theta,n})^\top, \qquad W\in\mathcal{W}_c. \quad \text{(replicator map)} \tag{2.4b}$$

The *exponential maps* with respect to the e-connection reads

$$\operatorname{Exp}_p(v) = \frac{p\cdot e^{\frac{v}{p}}}{\langle p, e^{\frac{p}{v}}\rangle}, \qquad p\in\mathcal{S}_c, \quad v\in T_0, \tag{2.5a}$$

$$\operatorname{Exp}_W(V) = \big(\operatorname{Exp}_{W_1}(V_1),\dots,\operatorname{Exp}_{W_n}(V_n)\big)^\top \qquad W\in\mathcal{W}_c, \quad V\in\mathcal{T}_0, \tag{2.5b}$$

where both the multiplication $\cdot$ and the exponential function apply componentwise. Composition with the replicator maps (2.4) yields the

$$\exp_p\colon T_0\to\mathcal{S}_c, \qquad \exp_p := \operatorname{Exp}_P\circ R_p, \qquad p\in\mathcal{S}_c, \qquad \text{(lifting map)} \tag{2.6a}$$

$$\exp_W\colon \mathcal{T}_0\to\mathcal{W}_c, \qquad \exp_W := \operatorname{Exp}_W\circ R_W, \qquad W\in\mathcal{W}_c. \qquad \text{(lifting map)} \tag{2.6b}$$

2.2. **Meta-Simplex, Flow Embedding.** The embedding (1.13) is defined by the map

$$T\colon \mathcal{W}_c\to\mathcal{T}=T(\mathcal{W}_c)\subset\mathcal{S}_N, \qquad T(W)_\alpha := \prod_{i\in[n]} W_{i,\alpha_i}, \qquad \alpha\in[c]^n. \tag{2.7}$$

Denoting the tangent space to $\mathcal{S}_N$ defined by (1.13) by

$$\mathcal{T}_0\mathcal{S}_N := \{z\in\mathbb{R}^N\colon \langle\mathbb{1}_N,z\rangle=0\}, \qquad \text{(meta-tangent space)} \tag{2.8}$$

we also require the map

$$Q\colon \mathbb{R}^{n\times c}\to\mathbb{R}^N, \qquad Q\colon\mathcal{T}_0\to\mathcal{T}_0\mathcal{S}_N, \qquad (QV)_\alpha := \sum_{i\in[n]} V_{i,\alpha_i}, \qquad \alpha\in[c]^n. \tag{2.9}$$

The mappings T,Q have been studied by [BSGA$^+$23, BCA$^+$24].

Every point $W\in\mathcal{W}$ on the assignment manifold is represented through (2.7) by the combinatorially large vector $T(W)$ with $N=c^n$ components $T(W)_\alpha$, consisting of monomials of degree n in the variables $W_{i,\alpha_i}\in(0,1)$. A labeling determined by the assignment flow by (1.11) corresponds to

$$\lim_{t\to\infty} T\big(W(t)\big) = T\big((e_{\alpha_1},\dots,e_{\alpha_n})^\top\big) = e_\alpha, \tag{2.10}$$

that is, the unit vector (vertex) of the meta-simplex $\Delta_N=\overline{\mathcal{S}_N}$ corresponding to the Dirac measure δ_α concentrated on the labelling $\alpha\in[c]^n$.

Example 2.1. We reconsider the toy scenario (1.6) of joint distributions of two binary variables. Such distributions correspond on the assignment manifold to points of the form

$$W = \Big(\big(\begin{smallmatrix} w_1\\ 1-w_1\end{smallmatrix}\big), \big(\begin{smallmatrix} w_2\\ 1-w_2\end{smallmatrix}\big)\Big)^\top, \qquad w_1,w_2\in(0,1). \tag{2.11}$$

Embedding this point by (2.7) yields the vector

$$T(W) = \big(w_1w_2, w_1(1-w_2), (1-w_1)w_2, (1-w_1)(1-w_2)\big)^\top, \tag{2.12}$$

with components $T(W)_{\alpha}$ indexed by the four possible labeling $\alpha \in \{(1,1),(1,0),(0,1),(0,0)\}$. Since any distribution on the assignment manifold factorizes, this vector is determined by merely two parameters w_1, w_2. Accordingly, the embedded assignment manifold $\mathcal{T} = T(\mathcal{W}_c) \subset \mathcal{S}_N$ is the two-dimensional submanifold depicted by Figure 1.1(a).

In mathematics, such embedded sets are known as *Segre varieties* at the intersection of algebraic geometry and statistics [LSX09, DSS09].

The following proposition highlights the specific role of the submanifold of $\mathcal{S}_N$ corresponding to the *embedded assignment manifold* $\mathcal{T} = T(\mathcal{W}) \subset \mathcal{S}_N$.

Proposition 2.2 ([BCA+24, Prop. 3.2])**.** *For every $W \in \mathcal{W}_c$, the distribution $T(W) \in \mathcal{S}_N$ has maximum entropy*

$$H\big(T(W)\big) = -\sum_{\alpha \in [c]^n} T(W)_{\alpha} \log T(W)_{\alpha} \tag{2.13}$$

among all $p \in \mathcal{S}_N$ subject to the marginal constraint

$$Mp = W, \tag{2.14a}$$

where the marginalization map is given by

$$M \colon \mathbb{R}^N \to \mathbb{R}^{n \times c}, \qquad (Mp)_{i,j} := \sum_{\alpha \in [c]^n \colon \alpha_i = j} p_{\alpha}, \qquad \forall (i,j) \in [n] \times [c]. \tag{2.14b}$$

As a consequence, any *general* distribution $p \in \mathcal{S}_N \setminus T(\mathcal{W}_c)$ which is *not* in $T(\mathcal{W}_c)$, has *non*-maximal entropy and hence is *more* informative by encoding additional statistical dependencies [CT06].

Our approach for *generating* such general distributions $p \in \mathcal{S}_N$, by combining simple factorizing distributions $W \in \mathcal{W}_c$ via the embedding (2.7) and assignment flows (2.3), is introduced in following Section 3.

3. Approach

Section 3.1 introduces our generative model for representing and learning a discrete joint distribution $p = p(\alpha) \in \mathcal{S}_N$ of label configurations $\alpha = (\alpha_1, \dots, \alpha_n)$ as realizations of discrete random variables $y = (y_1, \dots, y_n) \sim p$. The approach is illustrated by Figure 3.1. The training procedure for simulation-free training of the generative model is worked out in Section 3.2. Section 3.3 specifies precisely how the approximation of p is achieved in the meta-simplex by measure transport on the embedded nonlinear submanifold of factorizing distributions.

We conclude with short Sections 3.4–3.6 on the geometric integration method that we employed for the discretization of our time-continuous generative model in numerical experiments, on the computation of the likelihood $\widetilde{p}(\alpha)$ of arbitrary label configurations using the learned generative model, and on the characterization of our approach as a dequantization procedure.

3.1. **Generative Model.**

3.1.1. *Goal.* The goal is to learn an approximation

$$\widetilde{p} \approx p, \qquad\qquad \text{(approximation)} \tag{3.1}$$

as convex combination of *factorizing* joint distributions. The submanifold $\mathcal{T} = T(\mathcal{W}_c) \subset \mathcal{S}_N$ shown in Figure 1.1(a) spans all factorizing distributions $T(W) \in \mathcal{S}_N$, which are efficiently represented by their marginals $W \in \mathcal{W}_c$ due to (2.14a). In particular, since the dimension of $\mathcal{W}_c$ only grows linearly in the number of variables n, factorizing distributions are tractable to work with numerically. However, only *independent* random variables follow factorizing distributions, posing the question of how statistical *coupling* between such variables can be represented through convex combination.

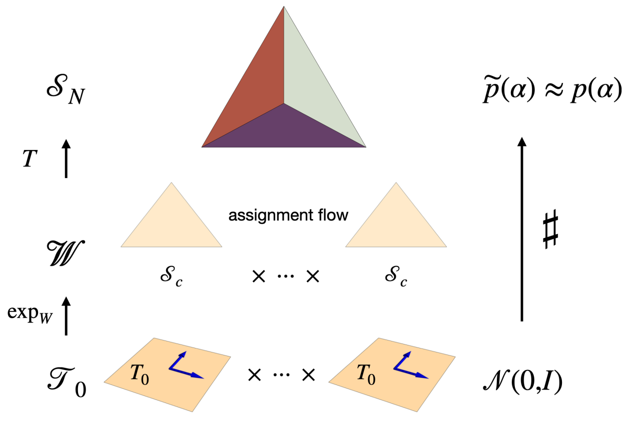


FIGURE 3.1. **Overview of the approach:** The standard Gaussian reference measure $\mathcal{N}(0,I)$ is pushed forward by the lifting map $\exp_W$ from the flat tangent product space $\mathcal{T}_0$ to the assignment manifold $\mathcal{W}_c$, and further to the meta-simplex $\mathcal{S}_N$ via the embedding map T (2.7), by geometrically integrating the assignment flow equation (2.3). Since the assignment flow converges to the extreme points of $\overline{\mathcal{W}_c}$ which after embedding agree with the extreme points of $\Delta_N = \overline{\mathcal{S}_N}$, an approximation $\widetilde{p}(\alpha)$ of a general *discrete* target measure $p(\alpha)$ can be learned in terms of a corresponding convex combination of extreme points. This is achieved by matching the flow of e-geodesics which encode given training samples to the generating assignment flow, by empirical expectation, and by learning the parameters of the affinity function F_θ (1.12). Since factorizing distributions $T(W)$, $W \in \mathcal{W}_c$, are only required, the approach is computationally feasible also in high dimensions.

3.1.2. *Representation of General Distributions.* Note that the submanifold of factorizing distributions $\mathcal{T} \subseteq \mathcal{S}_N$ is *nonconvex*. Thus, convex combinations of two factorizing distributions $T(W_1)$ and $T(W_2)$ generally lie *outside* of $\mathcal{T}$ and hence form a *non-factorizing* distribution.

In addition, we observe that every Dirac measure e_α factorizes. Intuitively, this is because each variable has a deterministic value, independent of all others. Because Dirac measures are the extreme points of the convex set $\overline{\mathcal{S}_N}$, *every* joint distribution $\widetilde{p} \in \mathcal{S}_N$ representing an *arbitrary* coupling between variables can be written as a convex combination of Dirac measures

$$\widetilde{p} = \sum_{\alpha \in [c]^n} \widetilde{p}_\alpha e_\alpha. \tag{3.2}$$

This particular representation of $\widetilde{p}$ is intractable, however, because it involves a combinatorially large number of mixture coefficients $\widetilde{p}_\alpha$. To tame this complexity, the **key idea** is to *represent mixtures $\widetilde{p} \in \mathcal{S}_N$ of factorizing distributions as measures $\nu \in \mathcal{P}(\mathcal{W}_c)$* by

$$\widetilde{p} = \mathbb{E}_{W \sim \nu}[T(W)]. \tag{3.3}$$

This shifts the problem of parameterizing useful subsets of combinatorially many mixture coefficients in (3.2) to the problem of parameterizing a preferably large subset of measures $\nu \in \mathcal{P}(\mathcal{W}_c)$, supported on the comparatively low-dimensional manifold $\mathcal{W}_c$. The latter can be achieved by *parameterized measure transport* on the *assignment manifold* $\mathcal{W}_c$.

Specifically, a simple reference measure

$$\nu_0 \in \mathcal{P}(\mathcal{W}_c) \qquad \text{(reference measure)} \tag{3.4a}$$

is chosen and transported by the assignment flow (2.3), reaching

$$\nu = \nu_\infty \quad \text{for} \quad t \to \infty. \qquad \text{(transported measure)} \tag{3.4b}$$

Parameterization of measures

$$\nu_\theta \in \mathcal{P}(\mathcal{W}) \qquad \text{(parametrized measure)} \tag{3.4c}$$

is achieved by choosing an appropriate class of affinity functions $F_\theta \colon \mathcal{W} \to \mathbb{R}^{n\times c}$ (1.12) driving the assignment flow (2.3). Note that, while the support of $\widetilde{p}$ in (3.2) was directly associated with the number of mixture coefficients, the complexity of representing $\widetilde{p}$ via the ansatz (3.3) is no longer associated with its support.

The simplest example of (3.3) is the representation of

$$\widetilde{p} = \mathbb{1}_{\mathcal{S}_N} \tag{3.5}$$

by choosing $F_\theta \equiv 0$ and a *product* reference distribution

$$\nu_0 = \prod_{i\in[n]} \nu_{0;i} \in \mathcal{P}(\mathcal{W}_c) \tag{3.6}$$

with mean $\mathbb{E}_{W_i\sim\nu_{0;i}}[W_i] = \mathbb{1}_{\mathcal{S}_c}$, which through the embedding (3.3) yields (3.5), which has *full* support on the very high-dimensional space $[c]^n$. We make this connection more explicit.

Lemma 3.1 (**convex combination of embedded nodewise measures**). *Suppose the reference measure ν_0 has the product form (3.6) with $\nu_i \in \mathcal{P}(\mathcal{S}_c)$. Then the joint distribution represented by the mixture (3.3) reads*

$$\widetilde{p} = \mathbb{E}_{W\sim\nu}[T(W)] = T(\widehat{W}), \qquad \widehat{W}_i = \mathbb{E}_{W_i\sim\nu_i}[W_i], \quad i\in[n]. \tag{3.7}$$

Proof. Let $\alpha \in [c]^n$ be an arbitrary multi-index. Since ν factorizes in the described manner, $W \sim \nu$ is independently distributed on each node which implies

$$\widetilde{p}_\alpha = \mathbb{E}_{W\sim\nu}[T(W)_\alpha] = \mathbb{E}_{W\sim\nu}\Big[\prod_{i\in[n]} W_{i,\alpha_i}\Big] = \prod_{i\in[n]} \mathbb{E}_{W\sim\nu_i}[W_{i,\alpha_i}] = T(\widehat{W})_\alpha. \tag{3.8}$$

□

Lemma 3.1 shows that, if ν is independent on every node, then $\widetilde{p} \in \mathcal{T}$. In particular, *coupling* between variables, to be represented by the joint distribution $\widetilde{p}$, has necessarily to be induced by the *interaction* of node states over the course of integrating the assignment flow.

3.1.3. *Model Learning and Model Evaluation (Sampling).* The target distribution p is unknown, in practice, and only independently drawn training samples $\beta \sim p$ are available. After choosing a class of payoff functions F_θ, the task is to learn parameters θ such that

$$\widetilde{p} = \mathbb{E}_{W\sim\nu_\theta}[T(W)], \tag{3.9}$$

i.e. a parametrization of the right-hand side of (3.3), approximates the empirical distribution of samples. To this end, we identify samples β with the corresponding extremal points $Me_\beta \in \overline{\mathcal{W}_c}$ (Section 3.2.1) and use *flow matching* on $\mathcal{W}_c$ to learn θ in a numerically stable and efficient way (Section 3.2).

After learning has converged, new samples $\beta \sim \widetilde{p}$ from the approximate distribution $\widetilde{p} \approx p$ can be drawn by a two-stage process:

(i) First, an initialization $W_0 \sim \nu_0$ is drawn and evolved over time $W(t) \in \mathcal{W}_c$ by integrating the learned assignment flow until either the desired time $t_{\max}$ is reached, or $W(t)$ approaches an extreme point of $\overline{\mathcal{W}_c}$.

(ii) The new data is subsequently drawn from the factorizing distribution $T(W(t_{\max}))$. At extreme points $Me_{\beta'}$, this distribution is a Dirac measure and sampling from it always yields β'.

3.2. **Riemannian Flow Matching.** in this section, we work out details of the procedure for training generative assignment flows.

3.2.1. *Representation of Labelings as Training Data.* Our approach to training the generative model utilizes *labelings* as training data of the form

$$\overline{W} \in \overline{\mathcal{W}_c}, \qquad \overline{W}_i = e_{\alpha_i}, \qquad \alpha_i \in [c], \qquad \forall i \in [n]. \tag{3.10}$$

Any such point $\overline{W}$ assigns a label (category) α_i to each vertex $i \in \mathcal{V}$ in terms of a corresponding unit vector $e_{\alpha_i} \in \{0,1\}^c$. The flow-matching criterion, specified in the following section, is optimized to find θ such that F_θ drives the assignment flow to labelings in the limit $\lim_{t\to\infty} W(t) = \overline{W}$. In practice, the assignment flow is integrated up to a sufficiently large point of time

$$t_{\max} > 0 \tag{3.11}$$

followed by trivial rounding of $W_i(t_{\max}) \mapsto e_{\alpha_i}$ at every vertex i.

3.2.2. *Training Criterion.* This section details the approach schematically depicted by Figure 3.1. In the following,

$$\beta \sim p \tag{3.12}$$

denotes labeling configurations for training, drawn from the *unknown* underlying discrete joint data distribution p. β corresponds to the Dirac measure $e_\beta \in \mathcal{S}_N$ (extreme point) of the meta-simplex $\mathcal{S}_N$ and to a corresponding point $\overline{W}_\beta = Me_\beta \in \overline{\mathcal{W}_c}$, to which the assignment flow (2.3) may converge.

The idea of flow matching is to directly fit the model vector field, in our case the assignment flow vector field (2.3),

$$V_\theta(W,t) := R_W[F_\theta(W,t)], \tag{3.13}$$

to a vector field whose flow realizes a desired measure transport. Let $\nu_0 \in \mathcal{P}(\mathcal{W})$ be a simple reference measure and define *conditional probability paths*

$$\nu_t(\beta) \tag{3.14}$$

satisfying

$$\nu_0(\beta) := \nu_0 \tag{3.15a}$$

$$\nu_\infty(\beta) := \lim_{t\to\infty} \nu_t(\beta) = \delta_{\overline{W}_\beta}(W) \quad \text{for all} \quad \beta \in [c]^n, \tag{3.15b}$$

where $\overline{W}_\beta = (e_{\beta_1}, \dots, e_{\beta_n})^\top \in \{0,1\}^{n\times c}$ is the extreme point of $\overline{\mathcal{W}_c}$ corresponding to β, such that $T(\overline{W}_\beta) = e_\beta \in \overline{\mathcal{S}_N}$. Then the *marginal probability path*

$$\nu_t = \mathbb{E}_{\beta\sim p}[\nu_t(\beta)] \tag{3.16}$$

represents the target data distribution p in the limit $t \to \infty$ by ν_∞ and

$$\mathbb{E}_{W\sim\nu_\infty}[T(W)] = \mathbb{E}_{\beta\sim p}[e_\beta] = p. \tag{3.17}$$

In principle, we can now define a vector field

$$u_t \colon \mathcal{W}_c \to \mathcal{T}_0 \tag{3.18}$$

which generates the path $t \mapsto \nu_t$ in the sense that the flow of u_t pushes forward ν_0 to ν_t, for all times $t \geq 0$. Let $\rho \in \mathcal{P}([0,\infty))$ be a distribution with full support on the non-negative time axis. Regression of the assignment flow vector field (3.13),

$$V_\theta(\cdot,t) \colon \mathcal{W}_c \to \mathcal{T}_0, \tag{3.19}$$

with respect to u_t, amounts to minimizing the **Riemannian flow matching criterion**

$$\mathcal{L}_{\text{RFM}}(\theta) = \mathbb{E}_{t\sim\rho, W\sim\nu_t}\Big[\big\|u_t(W) - V_\theta(W,t)\big\|_W^2\Big], \tag{3.20}$$

where $\|\cdot\|_W^2 = g_W(\cdot,\cdot)$ (cf. (2.2e)).

In this form, flow matching is intractable, however, because we do not have access to the required field u_t. On the other hand, since we are at liberty to define *conditional* paths that conform to the constraints (3.15), we can choose $\nu_t(\beta)$ that are generated by *conditional* vector fields $u_t(\cdot|\beta)$ with *known* form. The key insight in [CL23], based on [LCBH+23] and provided that each $\nu_t(\beta)$ is generated by $u_t(\cdot|\beta)$, is that the loss function (3.20) has the same gradient with respect to θ as the **Riemannian *conditional* flow matching criterion**

$$\mathcal{L}_{\mathrm{RCFM}}(\theta) = \mathbb{E}_{t\sim\rho,\beta\sim p,W\sim\nu_t(\beta)}\Big[\big\|u_t(W|\beta) - V_\theta(W,t)\big\|_W^2\Big] \tag{3.21a}$$

$$\overset{(3.13)}{=} \mathbb{E}_{t\sim\rho,\beta\sim p,W\sim\nu_t(\beta)}\Big[\big\|u_t(W|\beta) - R_W[F_\theta(W,t)]\big\|_W^2\Big]. \tag{3.21b}$$

By contrast to (3.20), conditional vector fields $u_t(W|\beta)$ generating a path

$$t \mapsto \nu_t(\beta) \tag{3.22}$$

with the required properties (3.15) can be specified in closed form (cf. Proposition 3.28 below), and the conditional loss function (3.21) can be evaluated efficiently. Ultimately, by minimizing (3.21), the measure ν_t generated from the reference measure ν_0 by the assignment flow vector field $R_W[F_\theta(W,t)]$ approximates ν_∞ in the limit $t\to\infty$, which represents the unknown data distribution p through (3.17).

3.2.3. *Constructing Conditional Fields.* This section specifies the conditional vector fields $u_t(W|\beta)$ that generate the paths (3.22) conforming to (3.15) and define the conditional flow matching objective (3.21).

Let

$$\mathcal{N}_0(V) := \mathcal{N}(V;0,\Pi_0) \tag{3.23}$$

denote the standard Gaussian centered in the tangent space at $0\in\mathcal{T}_0$, with the orthogonal projection (2.2d) respresenting the identity map on $\mathcal{T}_0\subset\mathbb{R}^{n\times c}$. Pushing forward $\mathcal{N}_0$ by the lifting map (2.6b) at the barycenter yields a simple *reference distribution*

$$\nu_0 = (\exp_{\mathbb{1}_\mathcal{W}})_\sharp\mathcal{N}_0 \in \mathcal{P}(\mathcal{W}). \tag{3.24}$$

The distribution (3.24) is *simple* in the sense that it is easy to draw samples and the conditions of Lemma 3.1 are satisfied; in particular, ν_0 *factorizes* node-wise. For each labeling $\beta\in[c]^n$ and the corresponding extreme point $\overline{W}_\beta = (e_{\beta_1},\dots,e_{\beta_n})\in\overline{\mathcal{W}_c}$, and a

$$\lambda > 0, \qquad \text{(rate parameter)} \tag{3.25}$$

define the probability path

$$t\mapsto \mathcal{N}_{t,\beta} := \mathcal{N}(\cdot;t\lambda V_\beta,\Pi_0)\in\mathcal{P}(\mathcal{T}_0), \qquad V_\beta := \Pi_0\overline{W}_\beta, \tag{3.26}$$

and lift it to $\mathcal{W}_c$, defining

$$\nu_t(\beta) := (\exp_{\mathbb{1}_\mathcal{W}})_\sharp\mathcal{N}_{t,\beta}. \tag{3.27}$$

The parameter λ controls the *rate* at which $\nu_t(\beta)$ moves probability mass closer to $\overline{W}_\beta$. Small values of λ move the mass slowly; this is useful in settings with many labels $c\gg 1$, enabling the process to make class decisions during a longer time period. Figure 3.2 illustrates quantitatively the influence of λ.

The following proposition makes explicit the conditional vector field $u_t(W|\beta)$ that generates (3.27) and hence defines the training objective function (3.21). Recall the notation of Section 2.2 and the first paragraph of Section 3.2.2 explaining the one-to-one correspondence between

- a labelling configuration β,
- the corresponding Dirac measure $e_\beta\in\mathcal{S}_N$ of the meta simplex, and
- the corresponding point $\overline{W}_\beta\in\overline{\mathcal{W}_c}$ of the closure of the assignment manifold.

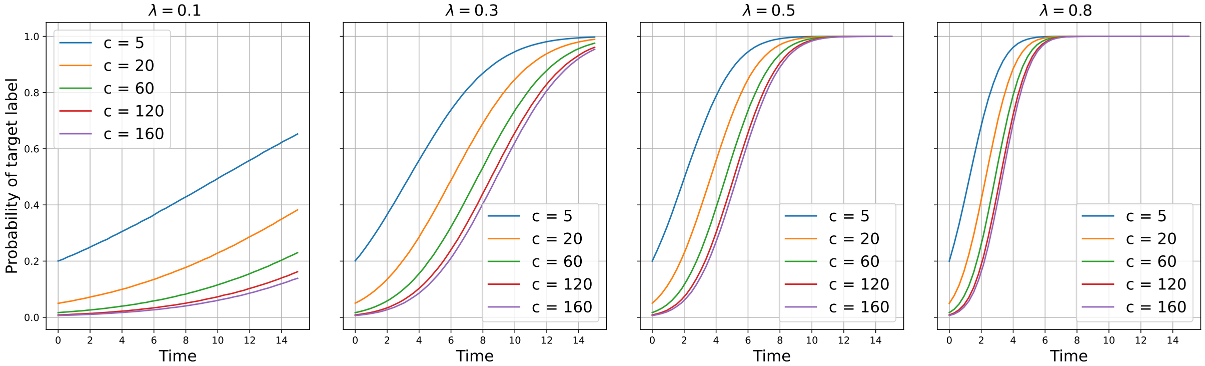


FIGURE 3.2. Influence of the parameter λ controlling in (3.26) and (3.30), respectively, the *rate* of assignment of mass of the pushforward probability measure (3.27) to a target label, depending on the number c of labels (classes, categories).

Proposition 3.2 (conditional vector fields). *The probability paths defined in* (3.27) *are generated through the smooth flow*

$$\psi_\cdot(\cdot|\beta)\colon \mathbb{R}_{\geq 0} \times \mathcal{T}_0 \to \mathcal{W}_c, \qquad \psi_t\big(V|\beta\big) = \exp_{\mathbb{1}_\mathcal{W}}(V + t\lambda V_\beta). \tag{3.28}$$

It is invertible and has the smooth inverse

$$\psi_t^{-1}(W|\beta) = \exp_{\mathbb{1}_\mathcal{W}}^{-1}(W) - t\lambda V_\beta. \tag{3.29}$$

In particular, the conditional vector field that generates (3.27) *is given by*

$$u_t(W|\beta) = R_W[\lambda V_\beta]. \tag{3.30}$$

Proof. See Appendix A.1, page 25. □

Proposition 3.3 (conditional path constraints). *The conditional probability paths $\nu_t(\beta)$ defined by* (3.27) *satisfy the constraints* (3.15).

Proof. See Appendix A.1, page 26. □

The path $\mathcal{N}_t$ is generated on the tangent space $\mathcal{T}_0$ by the constant vector field $V \mapsto \lambda V_\beta$ given by (3.26). The related vector field on $\mathcal{W}_c$, which generates the path (3.27), is given by (3.30). Comparing the shape of this field to (2.3) makes clear that assignment flows are natural candidate dynamics for matching conditional paths of the described form. The Riemannian conditional flow matching objective (3.21) consequently reads

$$\mathcal{L}_{\mathrm{RCFM}}(\theta) = \mathbb{E}_{t\sim\rho,\beta\sim p,W\sim\nu_t(\beta)}\Big[\big\|R_W[\lambda V_\beta - F_\theta(W,t)]\big\|_W^2\Big]. \tag{3.31}$$

We point out that this criterion is 'simulation free', i.e. *no integration* of the assignment flow is required for loss evaluation, which makes training computationally efficient.

Our approach (3.31) constitutes a novel instance of the flow-matching approach to generative modeling, introduced by [LCBH$^+$23] and recently extended to Riemannian manifolds by [CL23]. This instance uses the assignment manifold (1.9) and the corresponding Riemannian flow (2.3), along with the meta-simplex embedding (2.7), to devise a generative model whose underlying information geometry tailors the model to the representation and learning of *discrete joint* probability distributions.

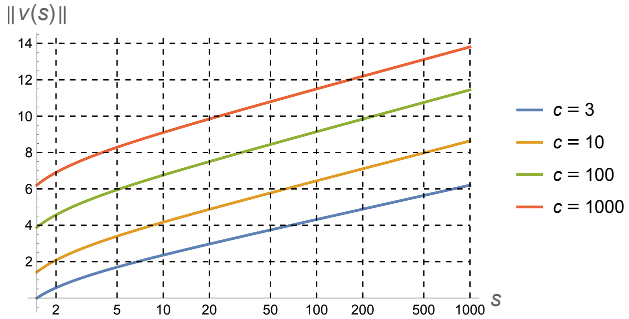


FIGURE 3.3. Norms $\|v(s)\|$ of the tangent vectors $v(s) = \exp^{-1}_{\mathbb{1}_{\mathcal{S}}}\big(p(s)\big)$ with $p(s) = \big(\frac{s-1}{s}, \frac{1}{(c-1)s}, \dots, \frac{1}{(c-1)s}\big) \to e_1 \in \mathbb{R}^c$ if $s \to \infty$, for numbers of labels $c \in \{3, 10, 100, 1000\}$. Since $\|e_1 - p(s)\| = \big(\frac{c}{c-1}\big)^{1/2}\frac{1}{s} \approx \frac{1}{s}$, the simplex Δ_c is covered, up to a very small distance to its boundary, by $\exp_{\mathbb{1}_{\mathcal{S}}}(B_0(r)) \subset \mathcal{S}_c$ and tangent vectors $v \in B_0(r) \subset T_0$ within a ball $B_0(r)$ centered at $0 \in T_0$ with radius $r = 15$.

3.2.4. *Infinite Integration Time.* A notable difference between our approach and previous Riemannian flow matching methods is that the target distribution is reached for $t \to \infty$ rather than after finite time. This corresponds to the fact that e-geodesics do not reach boundary points of $\overline{\mathcal{W}_c}$ after finite time and thus avoids two problems faced in prior work.

First, unlike the preliminary version presented in [BGAS24], data points $\beta \in [c]^n$ do not need to be smoothed in order to present targets in the interior of $\mathcal{W}$. Instead, we can directly approach extreme points $\overline{W}_\beta \in \overline{\mathcal{W}_c}$, even though they are at infinity in the tangent space $\mathcal{T}_0$ at $\mathbb{1}_{\mathcal{W}}$. Figure 3.3 shows that working within a ball in T_0 with radius 15 suffices to represent 'infinity' in practice.

Second, by not moving all mass of the reference distribution (close) to $\overline{W}_\beta$ in finite time, we avoid a pathological behavior which can arise in flow matching on the simplex. Denote by

$$r_\beta = \Big\{W \in \mathcal{W}_c \colon \beta_i \in \arg\max_{j\in[c]} W_{i,j},\ \forall i \in [n]\Big\} \tag{3.32}$$

the subset of points in $\mathcal{W}$ which assign their largest probability to the labels β. [SJW+24, Proposition 1] lays out that moving all mass of the reference distribution (close) to $\overline{W}_\beta$ in *finite* time forces the model to make class decisions very early because the probability of r_β under $\nu_t(\beta)$ increases too quickly. The effect is exacerbated by increasing the number of classes c that the model is asked to discriminate.

However, by opting for large integration time $t \to \infty$ and a corresponding construction (3.27) of conditional probability paths, our approach is able to scale to many classes $c \gg 1$, avoiding the pathology described in [SJW+24, Proposition 1]. Formally, this is because $\nu_t(\beta)$ defined in (3.27) has full support on $\mathcal{W}_c$ for every $t \geq 0$. In practice, the parameter λ in (3.26) can be used to control the speed at which the probability of r_β under $\nu_t(\beta)$ increases, allowing the model to make class decisions gradually over time.

Figure 4.5 (page 24) displays probability density paths for illustration. The corresponding impact on model accuracy is quantitatively shown in Figure 4.1 (page 20), with experimental details elaborated in Section 4.1.

3.2.5. *Relation to Dirichlet Flow Matching.* The construction of [SJW+24] specifically addresses pathological behavior of flow matching on the simplex, by choosing conditional probability paths $\nu_t(\beta)$ as paths of Dirichlet distributions. They demonstrate that this approach scales to at least $c = 160$ classes, by allowing the model to make class decisions gradually over time. However, the explicit definition of $\nu_t(\beta)$ as paths of Dirichlet distributions makes it non-trivial to find corresponding vector fields $u_t(\cdot|\beta)$ for flow matching, which leads them to make an ansatz for fields which move mass along straight lines in the ambient Euclidean space in which the probability simplex is embedded.

While we also make an explicit choice for $\nu_t(\beta)$ in (3.27), our construction is notably simpler than the approach of [SJW+24], allowing to easily compute the vector fields $u_t(\cdot|\beta)$ by pushforward (Proposition 3.2). The resulting flow moves mass along e-geodesics on $\mathcal{W}$, which is much more natural with respect to the information geometry of discrete probability distributions. To illustrate this point, consider a straight path $\widehat{p}(t) \in \mathbb{R}^n$ with direction $\frac{d}{dt}\widehat{p}(t) = v \in \mathbb{R}^n$ at all times t. The trajectory $\widehat{p}(t)$ is generated by maximizing $\langle v, \widehat{p}\rangle$ along its gradient direction. On $\mathcal{W}_c$, the quantity $\langle V_\beta, W\rangle$ can be interpreted as correlation between $W \in \mathcal{W}_c$ and the direction V_β. The Riemannian gradient of this correlation with respect to the product Fisher-Rao geometry on $\mathcal{W}_c$ is $R_W[V_\beta]$, i.e. precisely the direction of the conditional vector field (3.30).

3.3. **Learning Interaction between Simplices.**

Our prior work [BCA+24] has studied the relationship between assignment flows on the product manifold $\mathcal{W}_c$ and replicator dynamics on the meta-simplex $\mathcal{S}_N$. We now use core results of [BCA+24] to derive the flow matching approach of Section 3.2 from *first principles of flow matching in $\mathcal{S}_N$*, that is in the combinatorially large space of *all* discrete joint distributions. This demonstrates, in particular, that the proposed approach is suitable for *structured prediction* settings, in which multiple *coupled* random variables are of interest.

The result is surprising because *direct* flow matching of joint distributions in $\mathcal{S}_N$ is *intractable* due to the combinatorial dimension $N = c^n$. However, by leveraging the submanifold $\mathcal{T}$ (defined by (1.13) and illustrated by Figure 1.1) and the compatibility of assignment flows with its geometry, we show that our construction can effectively break down combinatorial complexity and define a *numerically tractable method*.

The map $T\colon \mathcal{W}_c \to \mathcal{S}_N$ defined in (2.7) associates a marginal distribution of n discrete random variables $W \in \mathcal{W}_c$ with a factorizing joint distribution $T(W) \in \mathcal{S}_N$. Define with slight abuse of notation[1] the orthogonal projection

$$\pi_0\colon \mathbb{R}^N \to \mathcal{T}_0\mathcal{S}_N \tag{3.33}$$

and formally denote the scaled standard normal distribution on $\mathcal{T}_0\mathcal{S}_N$ with variance c^{n-1} by

$$\mathcal{N}_0^{\mathcal{S}_N} = (\sqrt{c^{n-1}}\pi_0)_\sharp\mathcal{N}(0, I_N) = \mathcal{N}(0, c^{n-1}\pi_0\pi_0^\top) = \mathcal{N}(0, c^{n-1}\pi_0). \tag{3.34}$$

Analogous to the construction of conditional measures in Section 3.2.3, we define the path of conditional measures

$$\mathcal{N}_t^{\mathcal{S}_N}(\cdot|\beta) = \mathcal{N}(\cdot; tc^{n-1}\lambda\pi_0 e_\beta, c^{n-1}\pi_0) \tag{3.35}$$

given a labeling $\beta \in [c]^n$ and a rate parameter $\lambda > 0$, scaled by the constant c^{n-1}. It follows from Proposition 3.3 that

$$\nu_t^{\mathcal{S}_N}(\beta) = (\exp_{\mathbb{1}_{\mathcal{S}_N}})_\sharp\mathcal{N}_t^{\mathcal{S}_N}(\cdot|\beta) \tag{3.36}$$

satisfies the conditions (3.15) on $\mathcal{S}_N$ and is thus suitable for flow matching on $\mathcal{S}_N$ with reference distribution $\nu_0^{\mathcal{S}_N} = \mathcal{N}_0^{\mathcal{S}_N}$. Formally, the Riemannian conditional flow matching criterion analogous to (3.31) reads

$$\mathcal{L}_{\mathrm{RCFM}}^{\mathcal{S}_N}(\theta) = \mathbb{E}_{t\sim\rho,\beta\sim p,q\sim\nu_t^{\mathcal{S}_N}(\beta)}\Big[\big\|R_q[\lambda\pi_0 e_\beta - f_\theta(q,t)]\big\|_w^2\Big] \tag{3.37}$$

for an affinity function $f_\theta\colon \mathcal{S}_N \times [0,\infty) \to \mathcal{T}_0\mathcal{S}_N$.

The task of minimizing (3.37) is numerically intractable, because we are not even able to easily represent general *points* $q \in \mathcal{S}_N \setminus \mathcal{T}$ in the complement of the embedded assignment manifold $\mathcal{T} = T(\mathcal{W}_c)$ given by (2.7). To break down this complexity, we will define a projection onto $\mathcal{T}$ by using the *lifting map lemma* [BCA+24, Lemma 3.3], which states

$$\exp_{\mathbb{1}_{\mathcal{S}_N}}(QV) = T\big(\exp_{\mathbb{1}_{\mathcal{W}}}(V)\big) \tag{3.38}$$

[1] π_0 is defined by (2.1d) as orthogonal projection onto the tangent space $T_0\mathcal{S}_c$ of the *single* simplex $\mathcal{S}_c$ with trivial tangent bundle $\mathcal{S}_c \times T_0$. Here, to simplify notation, we overload π_0 to denote analogously the orthogonal projection onto the tangent space $\mathcal{T}_0\mathcal{S}_N$.

for all tangent vectors $V \in T_0\mathcal{W}$, with the mappings T and Q defined by (2.7) and (2.9). We start with an orthogonal projection $\mathcal{T}_0\mathcal{S}_N \to \operatorname{img}(Q) \cap \mathcal{T}_0\mathcal{S}_N$.

Lemma 3.4 (**orthogonal projection onto** $\operatorname{img}(Q) \cap \mathcal{T}_0\mathcal{S}_N$)**.** *The orthogonal projection* proj_0 *of tangent vectors in* $\mathcal{T}_0\mathcal{S}_N$ *to the subspace* $\operatorname{img} Q \cap T_0\mathcal{S}_N$ *reads*

$$\operatorname{proj}_0 \colon \mathcal{T}_0\mathcal{S}_N \to \operatorname{img} Q \cap \mathcal{T}_0\mathcal{S}_N, \qquad\qquad \operatorname{proj}_0(v) := Q_c \Pi_0 {Q_c}^\top, v \qquad \textit{for}\;\; v \in \mathcal{T}_0\mathcal{S}_N, \tag{3.39}$$

in terms of the linear operator

$$Q_c := \frac{1}{\sqrt{c^{n-1}}} Q. \tag{3.40}$$

Since (3.38) ensures that $\exp_{\mathbb{1}_{\mathcal{S}_N}}(\operatorname{img} Q) \subseteq \mathcal{T}$, we can now define the projection

$$\operatorname{proj}_{\mathcal{T}} := \exp_{\mathbb{1}_{\mathcal{S}_N}} \circ \operatorname{proj}_0 \circ \exp^{-1}_{\mathbb{1}_{\mathcal{S}_N}} : \mathcal{S}_N \to \mathcal{T}. \tag{3.41}$$

Under this projection, the conditional measures $\nu_t^{\mathcal{S}_N}(\beta) \in \mathcal{P}(\mathcal{S}_N)$ precisely induce the conditional probability paths $\nu_t(\beta) \in \mathcal{P}(\mathcal{W}_c)$ defined by (3.27). Note that every extreme point of $\mathcal{S}_N$ lies in (the closure of) $\mathcal{T}$. Thus, projecting to $\mathcal{T}$ preserves the Dirac measures δ_{e_β} reached by the conditional distributions (3.36) in the limit $t \to \infty$. In particular, the projection transforms the intractable conditional flow matching criterion (3.37) on $\mathcal{S}_N$ into the numerically tractable criterion (3.31).

Theorem 3.5 (**projected flow matching on** $\mathcal{S}_N$)**.** *For any* $\beta \in [c]^n$*, the pushforward of the conditional measure* $\nu_t^{\mathcal{S}_N}(\beta)$ *defined in* (3.36) *under the projection* $\operatorname{proj}_{\mathcal{T}} \colon \mathcal{S}_N \to \mathcal{T}$ *defined in* (3.41) *is*

$$(\operatorname{proj}_{\mathcal{T}})_\sharp \nu_t^{\mathcal{S}_N}(\beta) = T_\sharp \nu_t(\beta) \tag{3.42}$$

with $\nu_t(\beta) \in \mathcal{P}(\mathcal{W}_c)$ *defined in* (3.27) *and the embedding map* T *given by* (2.7)*. Furthermore, the flow matching criterion on* $\mathcal{T}$*, induced by the conditional paths* (3.42)*, reads*

$$\mathcal{L}^{\mathcal{T}}_{\mathrm{RCFM}}(\theta) = \mathbb{E}_{t\sim\rho,\beta\sim p,q\sim(\operatorname{proj}_{\mathcal{T}})_\sharp \nu_t^{\mathcal{S}_N}(\beta)} \Big[\big\| R_q[\lambda \pi_0 e_\beta - \widetilde{f}_\theta(q,t)] \big\|_w^2 \Big] \tag{3.43}$$

and, using the ansatz $\widetilde{f}_\theta = Q \circ F_\theta \circ M$ *with* Q *and* M *defined by* (2.9) *and* (2.14b) , (3.43) *is equal to the criterion* (3.31) *for matching assignment flows on* $\mathcal{W}_c$.

Theorem 3.5 shows that the constructed flow matching on $\mathcal{W}_c$, which operates separately on multiple simplices, is *induced* by flow matching in the *single* meta-simplex $\mathcal{S}_N$, with conditional distribution paths and vector fields projected to the submanifold $\mathcal{T} = T(\mathcal{W}_c)$.

This result provides a geometric justification of the fact that *interaction* between simplices is learned through flow matching, even though all conditional probability paths $\nu_t(\beta)$ used for training can be *separately* constructed on individual simplices.

3.4. **Numerical Flow Integration.** We point out again that learning our generative model by Riemannian flow matching is 'simulation free': numerical integration is not required since only vector fields have to be matched which are defined on the tangent bundle of the assignment manifold and on the corresponding tangent-subspace distribution of the meta simplex (Prop. 3.5), respectively. On the other hand, numerical integration of the flow is required for evaluating the learned generative model, in order to sample as illustrated by Figure 1.1, or for likelihood computation (Section 3.5).

Since the flow corresponds to an ODE on a Riemannian manifold, *geometric* numerical integration utilizes various representations of the ODE on the tangent bundle in order to apply established methods for numerical integration on Euclidean spaces [HLW06]. In the case of the assignment flow, this has been thoroughly studied by [ZSPS20] using the extension of the lifting map (2.6a) to the product manifold (2.6b), regarded as action of the respective tangent space (regarded as additive abelian Lie group) on the assignment manifold. From the general viewpoint of geometric numerical integration, the resulting schemes for geometric numerical integration categorize as Runge-Kutta schemes of Munthe-Kaas type [MK99].

Specifically, in this paper, numerical integration was carried out using the classical explicit embedded Dormand & Prince Runge-Kutta method [DP80] of orders 4 & 5 with stepsize control (cf. [ZSPS20, Section 5.2] and [HNW08, Section II.5]).

3.5. **Likelihood Computation.** The likelihood of test data under the model distribution $\widetilde{p}$ is commonly used as a surrogate for Kullback-Leibler divergence between $\widetilde{p}$ and the true data distribution p, due to the relationship

$$\mathrm{KL}(p,\widetilde{p}) = \mathbb{E}_p\Big[\log\frac{p}{\widetilde{p}}\Big] = -H(p) - \mathbb{E}_p[\log\widetilde{p}]. \tag{3.44}$$

The entropy $H(p)$ is a property of the data distribution, which is not typically known, but can be treated as a constant which does not depend on the model used to approximate $\widetilde{p}$. For continuous normalizing flows, likelihood under the model is directly used as a training criterion, for this reason. Using the instantaneous change-of-variables formula [CRBD18]

$$\frac{\partial}{\partial t}\log\nu_t(x) = -\operatorname{tr} J(x,t), \tag{3.45}$$

log-likelihood under continuous normalizing flows can, on continuous state spaces, be computed by integrating (3.45) backward in time. In (3.45), $J(x,t)$ denotes the vector field Jacobian, whose trace is commonly approximated by using Hutchinson's estimator [Hut89]

$$\operatorname{tr} J = \mathbb{E}_v[\langle v, Jv\rangle] \tag{3.46}$$

with v drawn from a fixed normal or Rademacher distribution. The use of this estimator in the context of likelihood under continuous normalizing flows was proposed by [GCB$^+$19]. The authors use a single sample v for each integration of (3.45), which yields an unbiased estimator for log-likelihood of independent test data. In order to use likelihood as a training criterion, numerical integration of (3.45) is required. This entails many forward and backward passes through the employed network architecture in order to compute a single parameter update.

Therefore, we do not use likelihood as a training criterion, opting instead for the simulation-free flow matching approach of Section 3.2. Since the learned model is still a normalizing flow, (3.45) remains a useful tool for computing likelihoods under our model. However, because we are modeling discrete data while working on continuous state spaces, likelihood of discrete data can not be computed as a point estimate on $\mathcal{W}_c$. Further details are provided in Appendix C.

3.6. **Dequantization.** Approximation of discrete data distributions by continuous distributions has been studied through the lens of *dequantization*. Choose a latent space $\mathcal{F}^n$ and an embedding of class label configurations $\beta \in [c]^n$ as prototypical points $f^*_\beta \in \mathcal{F}^n$. Suppose the choice of these points is fixed before training and associate disjoint sets $A_\beta \subseteq \mathcal{F}^n$ with label configurations such that they form a partition of $\mathcal{F}^n$ and $f^*_\beta \in A_\beta$. We can then define the continuous surrogate model

$$\vartheta = \sum_{\beta\in[c]^n} p_\beta \mathcal{U}_{A_\beta} \in \mathcal{P}(\mathcal{F}^n) \tag{3.47}$$

which represents $p \in \mathcal{S}_N$ as a mixture of uniform distributions $\mathcal{U}_{A_\beta}$, supported on the disjoint subsets A_β. The underlying idea is that

$$\mathbb{P}_\vartheta(A_\beta) = \int_{A_\beta}\vartheta(y)dy = p_\beta\int_{A_\beta}\mathcal{U}_{A_\beta}(y)dy = p_\beta \tag{3.48}$$

due to the disjoint support of mixture components in (3.47). Denote a continuous model distribution by $\nu \in \mathcal{P}(\mathcal{F}^n)$. Using Jensen's inequality, we find

$$
\begin{aligned}
-H(\vartheta) - \mathrm{KL}(\vartheta, \nu) &= \int \vartheta(y) \log \nu(y) dy = \sum_{\beta \in [c]^n} p_\beta \int_{A_\beta} \log \nu(y) dy && \text{(3.49a)} \\
&\leq \sum_{\beta \in [c]^n} p_\beta \log \int_{A_\beta} \nu(y) dy && \text{(3.49b)} \\
&= -H(p) - \mathrm{KL}(p, q) && \text{(3.49c)}
\end{aligned}
$$

for the discrete model distribution q defined by

$$
q_\beta = \int_{A_\beta} \nu(y) dy = \mathbb{P}_\nu(A_\beta). \tag{3.50}
$$

Thus, fitting ν to ϑ by maximizing log-likelihood of smoothed data drawn from ϑ implicitly minimizes an upper bound on the relative entropy $\mathrm{KL}(p, q)$. In practice, drawing smoothed data from ϱ amounts to adding noise to the prototypes $f^*_{\beta_k} \in \mathcal{F}^n$ of discrete data $\{\beta_k\}_{k \in [m]}$.

The above *dequantization approach* was first proposed by [TvdOB16]. Their reasoning justifies the previously known heuristic of adding noise to dequantize data [UML13]. It has thenceforth become common practice for training normalizing flows as generative models of images [DSDB17, SKCK17] and was generalized to non-uniform noise distributions by [HCS$^+$19]. These authors focus on image data which, although originally continuous, are only available discretized into 8-bit integer color values for efficient digital storage. In this case, the underlying continuous color imparts a natural structure on the set of discrete classes. Similar colors are naturally represented as prototypes which are close to each other with respect to some metric on the feature space $\mathcal{F}^n$.

For the *general* discrete data considered here, such a structure is not available. As a remedy, [CAN22] present an approach to learn the embedding jointly with likelihood maximization and defining the partition of $\mathcal{F}^n$ into subsets A_β through Voronoi tesselation. The rounding model variant (C.1) of our approach can be seen as dequantization on the space $\mathcal{F}^n = \mathcal{W}_c$ with prototypical points $f^*_\beta = \overline{W}_\beta$. The sets A_β generated by Voronoi tesselation then coincide with the sets r_β defined by (3.32). However, our approach differs from [CAN22] by using flow matching instead of likelihood-based training and by explicit consideration of information geometry on $\mathcal{W}_c$.

A natural question is whether the ability to learn an embedding of class configurations as prototypical points f^*_β, thereby representing similarity relations between classes, can be replicated in our setting. Indeed, because points in $\mathcal{S}_c$ have a clear interpretation as categorical distributions, it is easy to achieve this goal by extending the affinity function F_θ of the assignment flow (2.3).

For some $L > 0$, let $E \in \mathbb{R}^{L \times c}$ be a learnable embedding matrix. The columns of E can be seen as prototypical points in the Euclidean latent space $\mathbb{R}^L$. The action of E on an integer probability vector $e_j \in \mathcal{S}_c$ precisely selects one of these points, associating it with the class $j \in [c]$. Learning E now allows to represent relationships between classes in the latent space $\mathbb{R}^L$. Let $\mathcal{E} \colon \mathbb{R}^{n \times c} \to \mathbb{R}^{n \times c}$ denote the linear operator which applies E nodewise. We now choose a parameterized function $\widetilde{F}_\theta \colon \mathbb{R}^L \to \mathbb{R}^L$ that operates on $\mathbb{R}^L$ and define the extended payoff function

$$
F_\theta = \mathcal{E}^\top \circ \widetilde{F}_\theta \circ \mathcal{E} \colon \mathcal{W}_c \to \mathbb{R}^{n \times c}. \tag{3.51}
$$

4. Experiments and Discussion

As outlined in Section 3, we perform Riemannian flow-matching (3.20) via the conditional objective (3.31) to learn assignment flows (2.3). These in turn approximate p_∞ in the limit $t \to \infty$ and thereby the unknown data distribution p through (3.17).

4.1. **Class Scaling.** First, we replicate the experiment of [SJW$^+$24, Figure 4] to verify that our model is able to make decisions gradually over longer integration time and can scale to many classes c. Details of the training procedure are relegated to Appendix B.1. For each c, the data distribution is a randomly generated, factorizing distribution on $n = 4$ simplices.

Figure 4.1 shows the relative entropy between the learned models (histogram of 512k samples) and the known target distribution. Our proposed approach is able to outperform our earlier method [BGAS24] (green) as well as Dirichlet flow matching [SJW$^+$24] (orange) and the linear flow matching baseline (blue) in terms of scaling to many classes c. In Figure 4.1, the linear flow matching baseline scales better to many classes than in [SJW$^+$24, Figure 4], but the qualitative statement that linear flow matching is ill-suited to this end is still supported by our empirical findings. Our preliminary approach [BGAS24] (green) also scales comparatively well, even outperforming Dirichlet flow matching. Figures 4.5 and 4.6 illustrate probability paths $\nu_t(\beta)$ for our approach (cf. (3.27)) and Dirichlet flow matching [SJW$^+$24] at different time scales.

A property of assignment flow approaches, possibly linked to observed performance, is to transport probability mass relative to the underlying Fisher Rao geometry (recall Section 3.2.5). For example, this leads to little probability mass in regions close to the simplex boundaries (Figure 4.5).

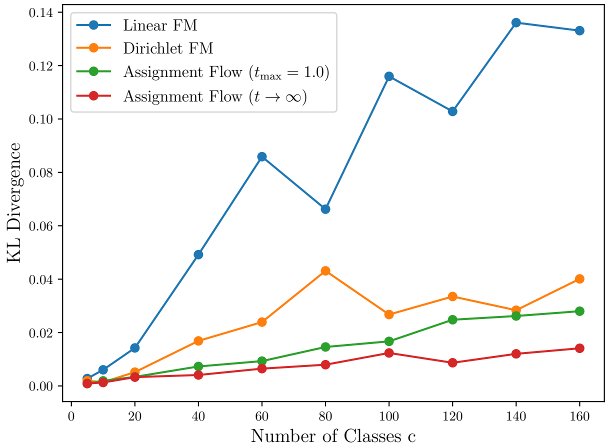


FIGURE 4.1. Relative entropy between learned models (histogram of 512k samples) and a known, factorizing target distribution on $n = 4$ simplices with varying number of classes c. By leveraging information geometry and gradual decision-making over time, our proposed approach (red) is able to outperform our earlier method [BGAS24] as well as Dirichlet flow matching [SJW$^+$24] in terms of scaling to many classes c.

4.2. **Generating Image Segmentations.** In image segmentation, a joint assignment of classes to pixels is usually sought conditioned on the pixel values themselves. Here, we instead focus on the *unconditional* discrete distribution of segmentations, without regard to the original pixel data. These discrete distributions are very high-dimensional in general, with $N = c^n$ increasing exponentially in the number of pixels.

To this end, we parametrize F_θ by the UNet architecture of [DN21] and train on downsampled segmentations of Cityscapes [COR$^+$16] images, as well as MNIST [LCB10], regarded as binary $c = 2$ segmentations after thresholding continuous pixel values at 0.5. Details of the training procedure are relegated to Appendix B.2.

Figures 4.2 and 4.3 show samples from the learned distribution of binarized MNIST and Cityscapes segmentations respectively, next to the closest training data. This illustrates that our model is able to interpolate the data distribution, without simply memorizing training data. Additional samples from our Cityscapes model are shown in Figure 4.4.

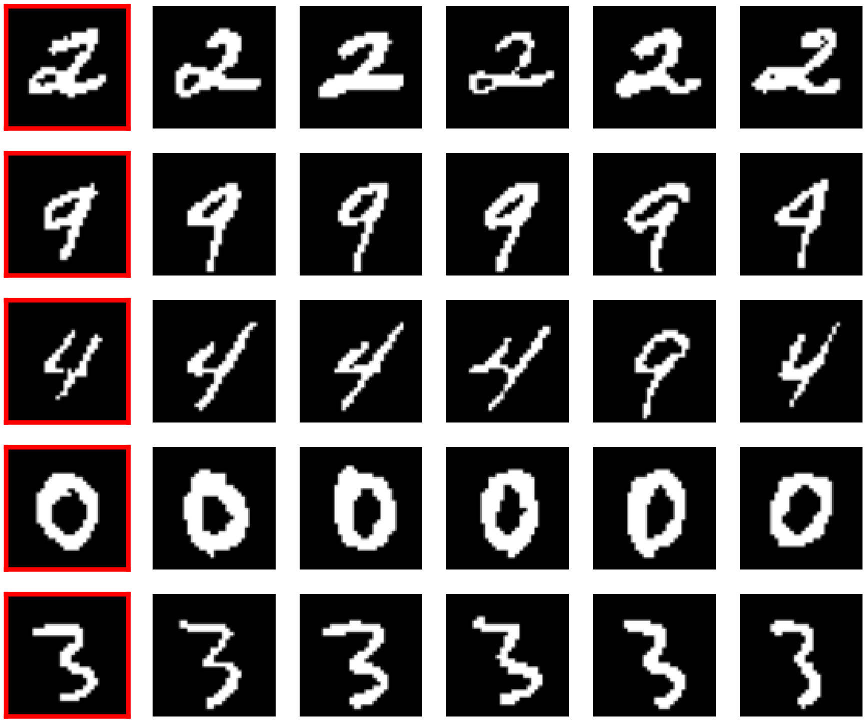


FIGURE 4.2. Comparison of model samples to the closest training data. *Left with red border*: samples drawn from our model of the binarized MNIST distribution. *Right*: training data closest to the sample in terms of pixel-wise distance.

4.3. **Likelihood Evaluation.** We compute the likelihood of test data from the MNIST dataset (binarized by thresholding) using the method described in Section 3.5. We use 100 priority samples per datum and, as is common practice for normalizing flows, only a single Hutchinson sample. The result is shown in Table 1, compared to our earlier approach [BGAS24] ($t \to 1$). For comparison, we show likelihood of MNIST test data (from the *continuous*, non-binarized distribution) under several normalizing flow methods from the literature which were trained using likelihood maximization.

Note that, although much prior work on generative modelling has been applied to continuous gray value MNIST images, binarization (in our case through thresholding) substantially changes the data distribution. Thus, likelihood of test data, which is commonly used as a surrogate for relative entropy to the data distribution in normalizing flows, is not comparable between these methods and ours. In addition, since we do not use likelihood maximization as a training criterion, it is not to be expected that our model is competitive on this measure. Still, the results of Table 1 indicate that the proposed model ($t \to \infty$) fits the binarized MNIST data distribution better in terms of relative entropy than our previous approach [BGAS24] ($t \to 1$).

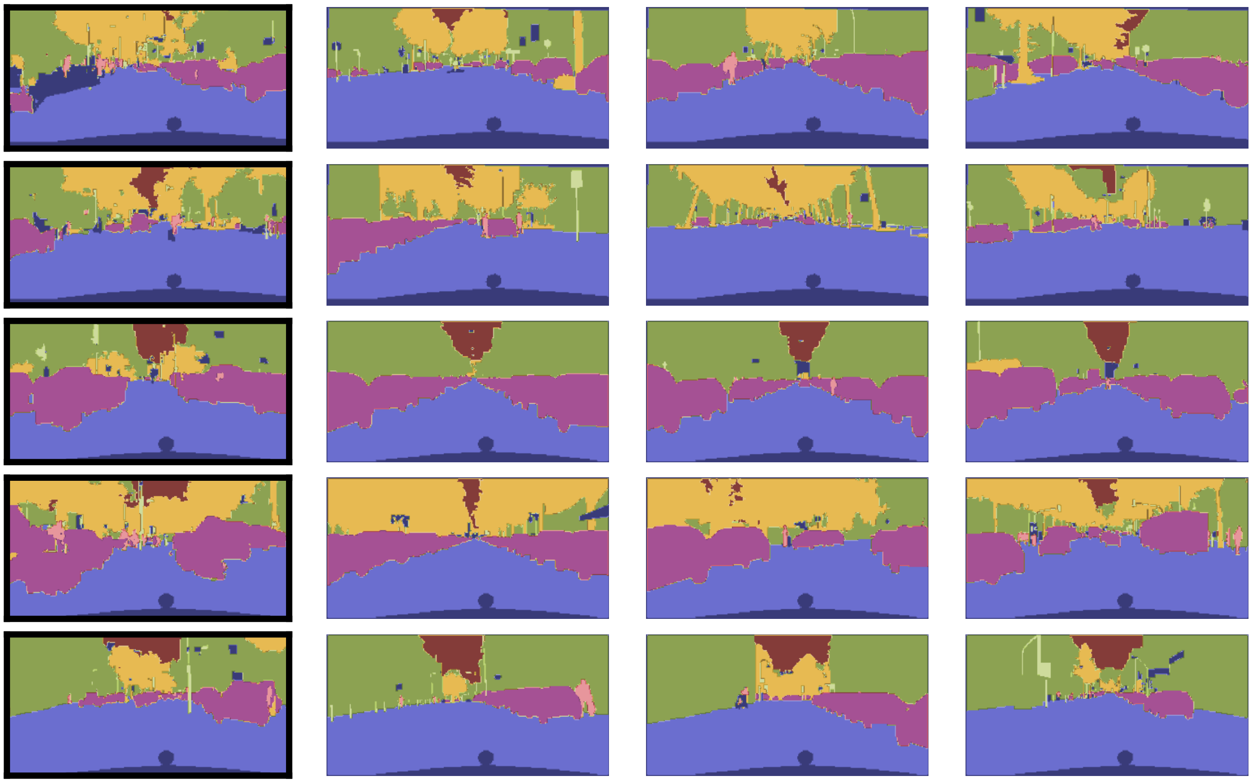


FIGURE 4.3. Comparison of model samples to the closest training data. *Left with black border*: samples drawn from our model of the Cityscapes segmentation distribution. *Right*: training data closest to the sample in terms of pixel-wise distance.

TABLE 1. Likelihood of binarized MNIST test data under our proposed model ($t \to \infty$) and the earlier version [BGAS24] ($t \to 1$). Both methods are trained by flow matching rather than likelihood maximization.

Method	AF ($t \to \infty$)	AF ($t \to 1$)
Likelihood (bits / dim)	1.01 ± 0.17	4.05 ± 0.83

5. CONCLUSION

We introduced a novel generative model for the representation and evaluation of joint probability distributions of discrete random variables. The approach employs an embedding of the assignment manifold in the meta-simplex of all joint probability distributions. Corresponding measure transport by randomized assignment flows approximates joint distributions of discrete random variables in a principled manner. The approach enables to learn the statistical dependencies of any set of discrete random variables and using the resulting model for structured prediction, independent of the area of application.

Inference using the approach is computationally efficient, since sampling can be accomplished by parallel geometric numerical integration. Training the generative model using given empirical data is computationally efficient, since matching the flow of corresponding e-geodesics is used as training criterion, which does not require sampling as a subroutine.

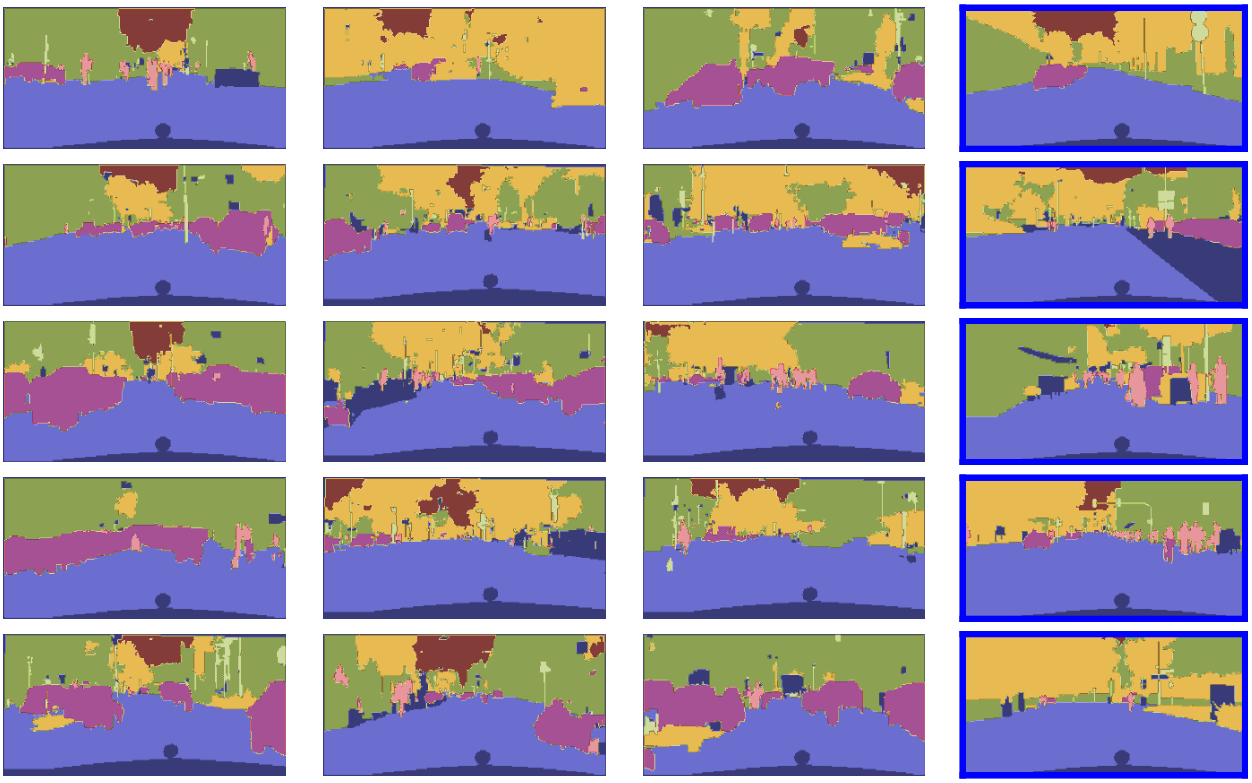


FIGURE 4.4. *Left*: Samples from our model of the Cityscapes segmentation distribution. *Right with blue border*: randomly drawn training data.

Numerical experiments showed superior performance in comparison to recent related work, which we attribute to consistently using the underlying information geometry of assignment flows and the corresponding measure transport along conditional probability paths. On the other hand, the fact that even our *preliminary* approach [BGAS24] can outperform Dirichlet flow matching [SJW+24] with respect to scaling to many classes in Figure 4.1, is surprising, because the approach [BGAS24] uses a *finite* integration time and moves all mass of the reference distribution to a Dirac measure close to $\overline{W}_\beta$ within this finite time. The core assumptions of [SJW+24, Proposition 1], therefore, apply to this approach, and the fact that it still performs well empirically suggests that further inquiry into this topic is warranted.

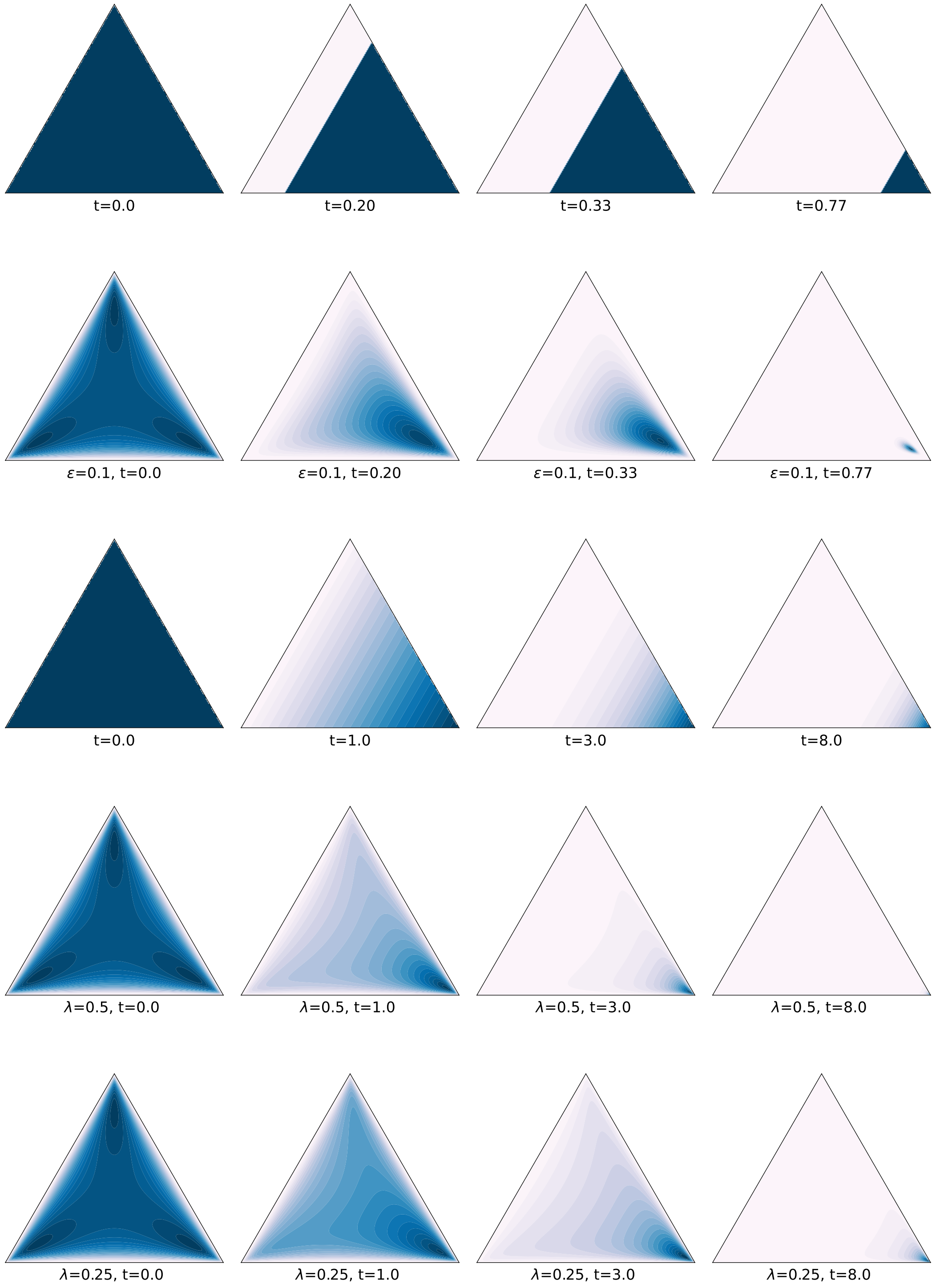


FIGURE 4.5. Plots of conditional densitities $\nu_t(\beta)$ for different points of time t. Darker colors indicate higher concentration within the densities. *From top to bottom:* Linear Flow Matching [SJW+24, Equation 11], the approach [BGAS24, Equation 18], Dirichlet Flow Matching [SJW+24, Equation 14], our approach (3.27) using two different values of the rate parameter λ. Note the different time periods $t \in [0, 0.77]$ used for the first two and $t \in [0, 8]$ for the latter approaches. See Section 4.1 for a discussion.

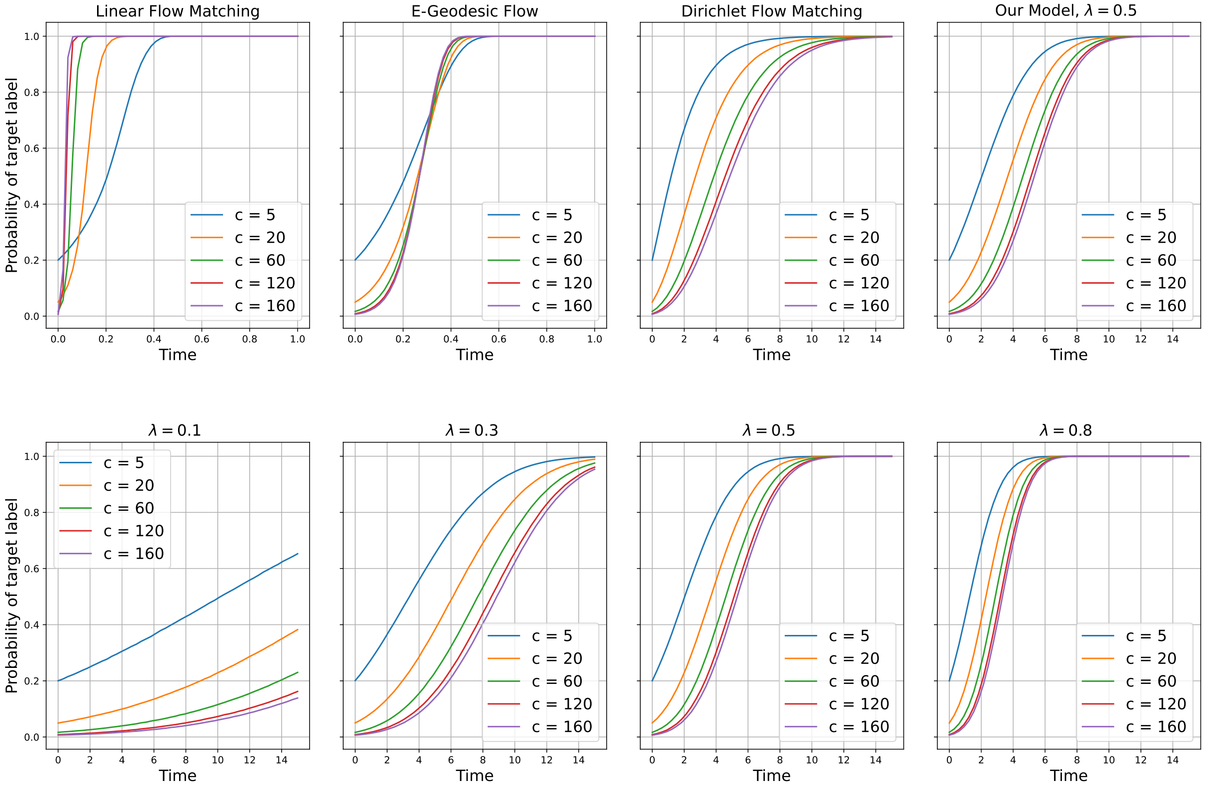


FIGURE 4.6. *Top row:* Plots of conditional densitities paths $t \mapsto \nu_t(\beta)$ for various models. *Bottom row:* Impact of the rate parameter λ of our approach (replication of Figure 3.2 to ease visual comparison).

APPENDIX A. PROOFS

A.1. Proofs of Section 3.2.3.

Proof of Proposition 3.2. Since V_β is determined by β and does not depend on V, the map $V \mapsto V + \lambda t \lambda V_\beta$ is affine. Hence, Eq. (3.28) conforms to (3.27), because affine transformations of normal distributions are again normal distributions. The mapping $\exp_{\mathbb{1}_\mathcal{W}}(\cdot) : \mathcal{W} \to \mathcal{T}_0$ is a diffeomorphism. Consequently, the inverse of (3.28) can be computed from

$$\begin{aligned} & W := \psi_t(V|\beta) = \exp_{\mathbb{1}_\mathcal{W}}\big(V + t\lambda V_\beta\big) && \text{(A.1a)} \\ \Leftrightarrow \qquad & \psi_t^{-1}(W|\beta) = V = \exp_{\mathbb{1}_\mathcal{W}}^{-1}(W) - t\lambda V_\beta, && \text{(A.1b)} \end{aligned}$$

which verifies (3.29). Regarding (3.30), recall that the conditional flow is determined by the conditional vector field through the ODE

$$\frac{d}{dt}\psi_t(V|\beta) = u_t\big(\psi_t(V|\beta)\big|\beta\big), \qquad \psi_0(V|\beta) = \psi_0(V) = \exp_{\mathbb{1}_\mathcal{W}}(V). \tag{A.2}$$

On the other hand, direct computation of the time derivative of (A.1a) using the closed-form expression

$$d\exp_W(V)[U] = R_{\exp_W(V)}[U] \tag{A.3}$$

for the differential of the lifting map (2.6b), yields

$$\frac{d}{dt}\psi_t(V|\beta) = R_{\psi_t(V|\beta)}[\lambda V_\beta]. \tag{A.4}$$

Equating (A.2) and (A.4) and using $W = \psi_t(V|\beta)$ from (A.1b) proves (3.30).

□

Proof of Proposition 3.3. Equation (3.15a) is immediate due to (3.23), (3.24) and (3.27). Writing short

$$\psi_t := \psi_t(\cdot|\beta) \tag{A.5}$$

for the flow map defined by (3.28), it remains to show that

$$\lim_{t\to\infty} \nu_t(\beta) = \lim_{t\to\infty} (\psi_t)_\sharp \nu_0 = \delta_{\overline{W}_\beta}. \tag{A.6}$$

To this end, we demonstrate that every marginal of the conditional probability path (A.6) converges to a Dirac measure supported on the assignment vector corresponding to the labeling configuration β, i.e.

$$\lim_{t\to\infty} \nu_{t;i}(\beta) = \lim_{t\to\infty} (\psi_{t;i})_\sharp \nu_{0;i} = \delta_{\overline{W}_{\beta;i}}, \quad i \in [n], \tag{A.7}$$

where $\nu_{0;i},\ i \in [n]$ denote the marginals of ν_0 given by (3.24).

First, we observe that by fixing an orthonormal basis of T_0 as column vectors of the matrix $\mathcal{B}$, every marginal $\nu_{0;i}$ of (3.24) with Gaussian $\mathcal{N}_0$ defined by (3.23) can be expressed as the lifted image measure of a standard normal distribution $\mathcal{N}\big(0_{c-1}, I_{c-1}\big)$ on $\mathbb{R}^{c-1}$ with respect to the basis $\mathcal{B}$,

$$\nu_{0;i} = (\exp_{\mathbb{1}_\mathcal{S}})_\sharp \mathcal{B}_\sharp \mathcal{N}\big(\cdot; 0_{c-1}, I_{c-1}\big) = (\exp_{\mathbb{1}_\mathcal{S}})_\sharp \mathcal{N}(\cdot; 0_c, \pi_0), \tag{A.8}$$

since $\mathcal{B}\mathcal{B}^\top = \pi_0$. Consequently, by Proposition 3.2,

$$\nu_{t;i}(\beta) = \big(\psi_{t;i}\big)_\sharp \mathcal{N}(\cdot; 0_c, \pi_0) \tag{A.9}$$

and hence using the change-of-variables formula and (A.1b), one has for any $p \in \mathcal{S}_c$,

$$\nu_{t;i}(p|\beta) = \mathcal{N}\big(\exp^{-1}_{\mathbb{1}_\mathcal{S}}(p) - t\lambda V_{\beta;i}; 0_c, \pi_0\big)|\det d\psi^{-1}_{t;i}|. \tag{A.10}$$

Equation (3.29) shows that the differential $d\psi^{-1}_{t;i}$ does not depend on t. Neither does the normalizing factor of the normal distribution, due to the covariance matrix $\pi_0 = \mathrm{id}_{T_0}$. Consequently, since $\psi^{-1}_{t;i}$ maps to T_0,

$$\nu_t(p|\beta) \propto \exp\Big(-\frac{1}{2}\big\langle \exp^{-1}_{\mathbb{1}_\mathcal{S}}(p) - t\lambda V_{\beta;i}, \pi_0\big(\exp^{-1}_{\mathbb{1}_\mathcal{S}}(p) - t\lambda V_{\beta;i}\big)\big\rangle\Big) \tag{A.11a}$$

$$= \exp\Big(-\frac{1}{2}\big\langle \exp^{-1}_{\mathbb{1}_\mathcal{S}}(p) - t\lambda V_{\beta;i}, \big(\exp^{-1}_{\mathbb{1}_\mathcal{S}}(p) - t\lambda V_{\beta;i}\big)\big\rangle\Big) \to 0 \quad \text{as} \quad t \to \infty, \tag{A.11b}$$

for any $p \neq \overline{W}_{\beta;i} \in \overline{\mathcal{S}_c}$ and $i \in [n]$, due to the choice (3.26) of the tangent vector V_β. We conclude that the image measure $\nu_{\infty;i}(\beta)$ is a Dirac measure concentrated on $\overline{W}_{\beta;i}$. □

A.2. **Proofs of Section 3.3.**

Proof of Lemma 3.4. By [BSS21, Lemma 4], one has $Q^\top Q V = c^{n-1} V$ for all $V \in \mathcal{T}_0$. Thus, Q_c defined by (3.40) has the property

$$Q_c{}^\top Q_c V = V, \qquad \text{for all } \; V \in \mathcal{T}_0. \tag{A.12}$$

To show that (3.39) indeed defines the orthogonal projection onto $\operatorname{img} Q \cap \mathcal{T}_0\mathcal{S}_N$, note that

$$Q_c \Pi_0 = \pi_0 Q_c \tag{A.13}$$

by [BCA+24, Lemma A.3] and accordingly

$$Q_c{}^\top \pi_0 = (\pi_0 Q_c)^\top = (Q_c \Pi_0)^\top = \Pi_0 Q_c{}^\top \tag{A.14}$$

by using the symmetry of Π_0 and π_0. We can use this to show $\operatorname{img} \operatorname{proj}_0 \subseteq \operatorname{img} Q \cap \mathcal{T}_0\mathcal{S}_N$, because for any $x \in \mathbb{R}^{n\times c}$, we have

$$Q_c\Pi_0 x \in \operatorname{img} Q \qquad \text{and} \qquad Q_c \Pi_0 x \overset{\text{(A.13)}}{=} \pi_0 Q_c x \in \mathcal{T}_0\mathcal{S}_N. \tag{A.15}$$

Now let $v \in \mathcal{T}_0\mathcal{S}_N$ and $y \in \operatorname{img} Q \cap \mathcal{T}_0\mathcal{S}_N$ be arbitrary. Then y can be written as $y = Q_c y'$ and we have

$$\begin{aligned} \langle v - \operatorname{proj}_0(v), y\rangle &= \langle v - Q_c\Pi_0 Q_c{}^\top v, Q_c y'\rangle = \langle Q_c{}^\top v - Q_c{}^\top Q_c \Pi_0 Q_c{}^\top v, y'\rangle && \text{(A.16a)}\\ &\overset{\text{(A.12)}}{=} \langle Q_c{}^\top v - \Pi_0 Q_c{}^\top v, y'\rangle \overset{\text{(A.14)}}{=} \langle Q_c{}^\top v - Q_c{}^\top \pi_0 v, y'\rangle && \text{(A.16b)}\\ &= 0, && \text{(A.16c)} \end{aligned}$$

which shows that proj_0 projects orthogonally. □

Proof of Theorem 3.5. We use the representation of proj_0 (Lemma 3.4) to compute the pushforward (3.42).

$$\begin{aligned} (\operatorname{proj}_{\mathcal{T}})_\sharp \nu_t^{\mathcal{S}_N}(\beta) &\overset{\text{(3.41)}}{=} (\exp_{\mathbb{1}_{\mathcal{S}_N}} \circ \operatorname{proj}_0 \circ \exp^{-1}_{\mathbb{1}_{\mathcal{S}_N}})_\sharp \nu_t^{\mathcal{S}_N}(\beta) && \text{(A.17a)}\\ &\overset{\text{(3.36)}}{=} (\exp_{\mathbb{1}_{\mathcal{S}_N}} \circ \operatorname{proj}_0)_\sharp \mathcal{N}_t^{\mathcal{S}_N}(\cdot|\beta) && \text{(A.17b)}\\ &\overset{\text{(3.39)}}{=} (\exp_{\mathbb{1}_{\mathcal{S}_N}} \circ Q_c \Pi_0 Q_c{}^\top)_\sharp \mathcal{N}_t^{\mathcal{S}_N}(\cdot|\beta) && \text{(A.17c)}\\ &\overset{\text{(3.35)}}{=} (\exp_{\mathbb{1}_{\mathcal{S}_N}} \circ Q_c \Pi_0 Q_c{}^\top)_\sharp \mathcal{N}(\cdot; t c^{n-1}\lambda \pi_0 e_\beta, c^{n-1}\pi_0) && \text{(A.17d)}\\ &= (\exp_{\mathbb{1}_{\mathcal{S}_N}})_\sharp \mathcal{N}(\cdot; t c^{n-1}\lambda Q_c \Pi_0 Q_c{}^\top \pi_0 e_\beta, c^{n-1} Q_c \Pi_0 Q_c{}^\top \pi_0 (Q_c \Pi_0 Q_c{}^\top)^\top) && \text{(A.17e)}\\ &\overset{\substack{\text{(3.40)}\\ \text{(A.14)}}}{=} (\exp_{\mathbb{1}_{\mathcal{S}_N}})_\sharp \mathcal{N}(\cdot; t\lambda Q \Pi_0 Q^\top e_\beta, c^{n-1} Q_c \Pi_0 Q_c{}^\top Q_c \Pi_0 Q_c{}^\top) && \text{(A.17f)}\\ &\overset{\substack{\text{(3.40)}\\ \text{(A.12)}}}{=} (\exp_{\mathbb{1}_{\mathcal{S}_N}})_\sharp \mathcal{N}(\cdot; t\lambda Q \Pi_0 Q^\top e_\beta, Q \Pi_0 Q^\top) && \text{(A.17g)}\\ &= (\exp_{\mathbb{1}_{\mathcal{S}_N}} \circ Q)_\sharp \mathcal{N}(\cdot; t\lambda \Pi_0 Q^\top e_\beta, \Pi_0). && \text{(A.17h)} \end{aligned}$$

By [BCA+24, Lemma 3.4], we have $Q^\top e_\beta = M e_\beta$, with Q and M defined by (2.9) and (2.14b). Using the shorthand V_β defined by (3.26) and the lifting map lemma (3.38), this shows

$$\begin{aligned} (\operatorname{proj}_{\mathcal{T}})_\sharp \nu_t^{\mathcal{S}_N}(\beta) &= (\exp_{\mathbb{1}_{\mathcal{S}_N}} \circ Q)_\sharp \mathcal{N}(\cdot; t\lambda V_\beta, \Pi_0) && \text{(A.18a)}\\ &\overset{\text{(3.38)}}{=} (T \circ \exp_{\mathbb{1}_{\mathcal{W}}})_\sharp \mathcal{N}(\cdot; t\lambda V_\beta, \Pi_0) && \text{(A.18b)}\\ &\overset{\text{(3.26)}}{=} (T \circ \exp_{\mathbb{1}_{\mathcal{W}}})_\sharp \mathcal{N}_{t,\beta} && \text{(A.18c)}\\ &\overset{\text{(3.27)}}{=} T_\sharp \nu_t(\beta) && \text{(A.18d)} \end{aligned}$$

which is the assertion (3.42).

Returning to (A.18a), we compute the conditional vector field whose flow generates the path $(\mathrm{proj}_{\mathcal{T}})_\sharp \nu_t^{\mathcal{S}_N}(\beta)$ by

$$u_t^{\mathcal{T}}(q|\beta) = d\exp_{\mathbb{1}_{\mathcal{S}_N}}(v)[\lambda Q V_\beta] = R_q[\lambda Q V_\beta] \tag{A.19}$$

with $v = \exp^{-1}_{\mathbb{1}_{\mathcal{S}_N}}(q)$, analogous to (3.30). This shows the shape of the flow matching criterion (3.43). It remains to show that it is equal to (3.31).

Substituting the ansatz $\widetilde{f}_\theta = Q \circ F_\theta \circ M$ into this criterion gives

$$\mathcal{L}^{\mathcal{T}}_{\mathrm{RCFM}} = \mathbb{E}_{t\sim\rho,\beta\sim p,W\sim\nu_t(\beta)}\Big[\big\|R_{T(W)}[\lambda Q(V_\beta) - (Q\circ F_\theta)(W,t)]\big\|^2_{T(W)}\Big]. \tag{A.20}$$

By [BCA+24, Theorem 3.1], $T\colon \mathcal{W}_c \to \mathcal{T} \subseteq \mathcal{S}_N$ defined by (2.7) is a Riemannian isometry. Thus, for any vector field $X\colon \mathcal{W}_c \to \mathcal{T}$ and any $W \in \mathcal{W}_c$, it holds that

$$\langle R_W[X], R_W[X]\rangle_W = \big\langle dT_W\big[R_W[X]\big], dT_W\big[R_W[X]\big]\big\rangle_{T(W)}. \tag{A.21}$$

Furthermore, by [BCA+24, Theorem 3.5], one has

$$dT_W\big[R_W[X]\big] = R_{T(W)}[QX]. \tag{A.22}$$

Taking (A.21) and (A.22) together, (A.20) transforms to

$$\mathcal{L}^{\mathcal{T}}_{\mathrm{RCFM}} = \mathbb{E}_{t\sim\rho,\beta\sim p,W\sim\nu_t(\beta)}\Big[\big\|R_W[\lambda V_\beta - F_\theta(W,t)]\big\|^2_W\Big] \tag{A.23}$$

which is (3.31). □

Appendix B. Experiments: Details

B.1. **Details of Class Scaling Experiment.** To parameterize F_θ, we use the same convolutional architecture used in [SJW+24]. We train for 500k steps of the Adam optimizer with constant learning rate $3\cdot 10^{-4}$ and batch size 128. We reproduce the Dirichlet flow matching results and linear flow matching baseline by using the code of [SJW+24]. The experiment shown in Figure 4.1 is slightly harder than the version in [SJW+24], because we limit training to 64k steps at batch size 512 for Dirichlet- and linear flow matching. Accordingly, both assignment flow methods are trained for 250k steps at batch size 128, such that around 32M data are seen by each model during training.

B.2. **Details of Generating Image Segmentations.**

B.2.1. *Cityscapes Data Preparation.* Rather than the original $c = 33$ classes, we only use the $c = 8$ classes specified as *categories* in *torchvision*. The same subsampling of classes was used in the related work [HNJ+21]. They additionally perform spatial subsampling to 32×64. Instead, we subsample the spatial dimensions (*NEAREST* interpolation) to 128×256.

B.2.2. *Cityscapes Training.* For the Cityscapes experiment, we employ the UNet architecture of [DN21] with *attention_resolutions* (32, 16, 8), *channel_mult* (1,1,2,3,4), 4 attention heads, 3 blocks and 64 channels. We trained for 250 epochs using Adam with cosine annealing learning rate scheduler starting at learning rate 0.0003 and batch size 4. The distribution ρ of times t used during training is an exponential distribution with rate parameter $\lambda = 0.25$. For sampling, we integrate up to $t_{\max} = 15$.

B.2.3. *Binarized MNIST Data Preparation.* We pad the original 28×28 images with zeros to size 32×32 to be compatible with spatial downsampling employed by the UNet architecture. Binarization is performed by pixelwise thresholding at grayvalue 0.5.

B.2.4. *Binarized MNIST Training.* We modify the same architecture used for Cityscapes to *attention_resolutions* (16), *channel_mult* (1,2,2,2), 4 attention heads, 2 blocks and 32 channels. The same training regimen is used as for Cityscapes except for an increase in batch size to 256. The distribution ρ of times t used during training is an exponential distribution with rate parameter $\lambda = 0.5$. For sampling, we integrate up to $t_{\max} = 10$. In table 1, we use the same UNet architecture and training regimen for the comparison method [BGAS24] $(t \to 1)$.

Appendix C. Likelihood Computation: Details

Assume we have learned a probability path ν_t and a final pushfoward distribution ν_∞. In practice, numerical integration needs to be stopped after a finite time $t = t_{\max}$, reaching a numerical pushforward distribution $\nu_{t_{\max}} \approx \nu_\infty$. Drawing samples from $\widetilde{p} = \mathbb{E}_{W\sim\nu_{t_{\max}}}[T(W)]$ is a two-stage process: $W \sim \nu_{t_{\max}}$ is drawn first, followed by sampling $\beta \sim T(W)$. Due to the numerical need to stop integration at finite time, $T(W)$ may in practice not have fully reached a discrete Dirac distribution. For long sequences of random variables, such as text or image modalities, this can lead to undesirable noise in the output samples. A way to combat this numerical problem is by rounding to a Dirac measure before sampling. This procedure can be interpreted within the framework of *dequantization*, which we elaborate in Section 3.6.

In practice, $W \sim \nu_{t_{\max}}$ is typically close to a discrete Dirac already (cf. Figure 3.3), so rounding has little impact on the represented joint distribution. Nevertheless, the rounding process is formally a different model than $\widetilde{p} = \mathbb{E}_{W\sim\nu_{t_{\max}}}[T(W)]$, which we explicitly distinguish for the purpose of computing likelihoods. Recall the definition (3.32) of subsets $r_\beta \subseteq \mathcal{W}$ with each $W \in r_\beta$ assigning the largest probability to the labels β. The points in r_β are also the ones which round to $\overline{W}_\beta$[2]. Thus, the labeling $\beta \in [c]^n$ has likelihood

$$\widetilde{p}^r_\beta = \mathbb{E}_{W\sim\nu_{t_{\max}}}[1_{r_\beta}(W)] = \mathbb{P}_{\nu_{t_{\max}}}(r_\beta) \tag{C.1}$$

under the rounding model $\widetilde{p}^r$, with 1_{r_β} denoting the indicator function of r_β. This is numerically similar to the likelihood under our original model

$$\widetilde{p}_\beta = \mathbb{E}_{W\sim\nu_\infty}[T(W)_\beta] \tag{C.2}$$

and matches it in the limit $t \to \infty$, provided that (almost) every trajectory $W(t)$ approaches an extreme point of $\overline{\mathcal{W}_c}$ under the learned assignment flow dynamics.

We will now devise an importance sampling scheme for efficient and numerically stable approximation of the integral in (C.1), that analogously applies to (C.2). Let ϱ be a proposal distribution with full support on $\mathcal{W}_c$ which has most of its mass concentrated around a point $q_\beta \in \mathcal{W}_c$ close to $\overline{W}_\beta$. Then

$$\mathbb{P}_{\nu_{t_{\max}}}(r_\beta) = \mathbb{E}_{W\sim\varrho}\Big[1_{r_\beta}(W)\frac{\nu_{t_{\max}}(W)}{\varrho(W)}\Big] \tag{C.3}$$

where we assumed that both $\nu_{t_{\max}}$ and ϱ have densities with respect to the Lebesgue measure and used again the symbols $\nu_{t_{\max}}$ and ϱ to denote these densities. The rationale behind this construction is that, since we learned $\nu_{t_{\max}}$ to concentrate close to points $\overline{W}_\beta$, drawing most samples close to q_β will reduce the estimator variance compared to sampling (C.1) directly. In high dimensions, the quantities in (C.3) are prone to numerical underflow, which motivates the transformation

$$\log \mathbb{P}_{\nu_{t_{\max}}}(r_\beta) = \log \mathbb{E}_{W\sim\varrho}\Big[1_{r_\beta}(W)\frac{\nu_{t_{\max}}(W)}{\varrho(W)}\Big] \tag{C.4a}$$

$$= \log \mathbb{E}_{W\sim\varrho}\big[\exp\big(\log 1_{r_\beta}(W) + \log \nu_{t_{\max}}(W) - \log \varrho(W)\big)\big]. \tag{C.4b}$$

After replacing the expectation with a mean over samples drawn from ϱ, we can evaluate (C.4) by leveraging stable numerical implementations of the logsumexp function.

[2] The sets r_β technically overlap on the boundary, but all intersections have measure zero.

For every evaluation of the integrand, we evaluate log-likelihood under ϱ in closed form as well as log-likelihood under $\nu_{t_{\max}}$ through numerical integration backward in time, leveraging the instantaneous change of variables (3.45) and Hutchinson's trace estimator (3.46). Note the conventions $\log 0 = -\infty$ and $\exp(-\infty) = 0$ employed in (C.4). The analogous expression for (C.1) reads

$$\log \widetilde{p}_{\beta} = \log \mathbb{E}_{W \sim \varrho}\big[\exp\big(\log T(W)_{\beta} + \log \nu_{t_{\max}}(W) - \log \varrho(W)\big)\big] \tag{C.5}$$

and we can further expand

$$\log T(W)_{\beta} = \log \prod_{i\in[n]} W_{i,\beta_i} = \sum_{i\in[n]} \log W_{i,\beta_i} \tag{C.6}$$

to avoid numerical underflow.

REFERENCES

[Agr13] A. Agresti, *Categorial Data Analysis*, 3rd ed., Wiley, 2013.

[Ait82] J. Aitchinson, *The Statistical Analysis of Compositional Data*, J. Royal Statistical Soc. B **2** (1982), 139–177.

[AN00] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry*, Amer. Math. Soc. and Oxford Univ. Press, 2000.

[ÅPSS17] F. Åström, S. Petra, B. Schmitzer, and C. Schnörr, *Image Labeling by Assignment*, Journal of Mathematical Imaging and Vision **58** (2017), no. 2, 211–238.

[BCA$^+$24] B. Boll, J. Cassel, P. Albers, S. Petra, and C. Schnörr, *A Geometric Embedding Approach to Multiple Games and Multiple Populations*, preprint arXiv:2401.05918 (2024).

[BGAS24] B. Boll, D. Gonzalez-Alvarado, and C. Schnörr, *Generative Modeling of Discrete Joint Distributions by E-Geodesic Flow Matching on Assignment Manifolds*, preprint arXiv:2402.07846 (2024).

[BSGA$^+$23] B. Boll, J. Schwarz, D. Gonzalez-Alvarado, D. Sitenko, S. Petra, and C. Schnörr, *Modeling Large-scale Joint Distributions and Inference by Randomized Assignment*, Scale Space and Variational Methods in Computer Vision (SSVM) (L. Calatroni, M. Donatelli, S. Morigi, M. Prato, and M. Santacesaria, eds.), LNCS, no. 14009, Springer, 2023, pp. 730–742.

[BSS21] B. Boll, J. Schwarz, and C. Schnörr, *On the Correspondence between Replicator Dynamics and Assignment Flows*, SSVM 2021: Scale Space and Variational Methods in Computer Vision, LNCS, vol. 12679, Springer, 2021, pp. 373–384.

[CAN22] R. T. Q. Chen, B. Amos, and M. Nickel, *Semi-Discrete Normalizing Flows through Differentiable Tesselation*, NeurIPS, 2022.

[CDLS99] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter, *Probabilistic Networks and Expert Systems*, Springer, 1999.

[CL23] R. T. Q. Chen and Y. Lipman, *Riemannian Flow Matching on General Geometries*, preprint arXiv:2302.03660 (2023).

[COR$^+$16] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, *The Cityscapes Dataset for Semantic Urban Scene Understanding*, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[CRBD18] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, *Neural Ordinary Differential Equations*, Proc. NeurIPS, 2018.

[CT06] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., John Wiley & Sons, 2006.

[DKP$^+$24] O. Davis, S. Kessler, M. Petrache, I. I Ceylan, M. Bronstein, and A. J. Bose, *Fisher Flow Matching for Generative Modeling over Discrete Data*, preprint arXiv:2405.14554 (2024).

[DN21] Prafulla Dhariwal and Alexander Quinn Nichol, *Diffusion Models Beat GANs on Image Synthesis*, NeurIPS, 2021.

[DP80] J. R. Dormand and P. J. Prince, *A Family of Embedded Runge-Kutta Formulae*, J. Comput. Appl. Math. **6** (1980), no. 1, 19–26.

[DSDB17] L. Dinh, J. Sohl-Dickstein, and S. Bengio, *Density Estimation Using Real NVP*, ICLR, 2017.

[DSS09] M. Drton, B. Sturmfels, and S. Sullivant, *Lecture on Algebraic Statistics*, Oberwolfach Seminars, vol. 39, Birkhäuser, 2009.

[Fer73] T. S. Ferguson, *A Bayesian Analysis of Some Nonparametric Problems*, Ann. Statistics **2** (1973), no. 1, 209–230.

[GCB$^+$19] W. Grathwohl, R. T. Q. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud, *FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models*, ICLR, 2019.

[GMS06] D. Geiger, C. Meek, and B. Sturmfels, *On the Toric Algebra of Graphical Models*, The Annals of Statistics **34** (2006), no. 3, 1463–1492.

[Har92] J. Harris, *Algebraic Geometry: A First Course*, Springer, 1992.

[HCS$^+$19] J. Ho, X. Chen, A. Srinivas, Y. Duan, and P. Abbeel, *Flow++: Improving flow-based generative models with variational dequantization and architecture design*, Proc. ICML, vol. PMLR 97, 2019, pp. 2722–2730.

[HLW06] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration*, Springer, 2006.

[HNJ$^+$21] E. Hoogeboom, D. Nielsen, P. Jaini, P. Forré, and M. Welling, *Argmax Flows and Multinomial Diffusion: Learning Categorial Distributions*, NeurIPS, 2021.

[HNW08] E. Hairer, S.P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I*, 3rd ed., Springer, 2008.

[HS98] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, 1998.

[Hut89] M. F. Hutchinson, *A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines*, Communications in Statistics-Simulation and Computation **18** (1989), no. 3, 1059–1076.

[JK77] N. L. Johnson and S. Kotz, *Urn Models and Their Application*, John Wiley & Sons, 1977.

[KF09] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, 2009.

[KPB21] I. Kobyzev, S.J. D. Prince, and M. A. Brubaker, *Normalizing Flows: An Introduction and Review of Current Methods*, IEEE Trans. Pattern Anal. Mach. Intell. **43** (2021), no. 11, 3964–3979.

[Lan12] J. M. Landsberg, *Tensors: Geometry and Applications*, Amer. Math. Soc., 2012.

[Lau96] S. L. Lauritzen, *Graphical Models*, Clarendon Press, Oxford, 1996.

[LCB10] Y. LeCun, C. Cortes, and C. J. Burges, *MNIST Handwritten Digit Database*, ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist **2** (2010).

[LCBH$^+$23] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, *Flow Matching for Generative Modeling*, ICLR, 2023.

[LM18] W. Li and G. Montufar, *Natural Gradient via Optimal Transport*, Information Geometry **2** (2018), no. 1, 181–214.

[LSX09] S. Lin, B. Sturmfels, and Z. Xu, *Marginal Likelihood Integrals for Mixtures of Independence Models*, J. Machine Learning Research **10** (2009), 1611–1631.

[MK99] H. Munthe-Kaas, *High Order Runge-Kutta Methods on Manifolds*, Applied Numerical Mathematics **29** (1999), no. 1, 115–127.

[PNR$^+$21] G. Papamakarios, E. Nalisnick, D.J. Rezende, S. Mohamed, and B. Lakshminarayanan, *Normalizing Flows for Probabilistic Modeling and Inference*, J. Machine Learning Research **22** (2021), no. 57, 1–64.

[RH21] L. Ruthotto and E. Haber, *An Introduction to Deep Generative Modeling*, GAMM Mitt. **44** (2021), no. 2, 24 pages.

[SJW$^+$24] H. Stark, B. Jing, C. Wang, G. Corso, B. Berger, R. Barzilay, and T. Jaakkola, *Dirichlet Flow Matching with Applications to DNA Sequence Design*, preprint arXiv:2402.05841 (2024).

[SKCK17] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, *PixelCNN++: Improving the pixelCNN with discretized logistic mixture likelihood and other modifications*, ICLR, 2017.

[Stu05] M. Studený, *On Probabilistic Conditional Independence Structures*, Springer, 2005.

[Sul18] S. Sullivant, *Algebraic Statistics*, American Mathematical Society, 2018.

[TvdOB16] L. Theis, A. van den Oord, and M. Bethge, *A Note on the Evaluation of Generative Models*, ICLR, 2016.

[UML13] B. Uria, I. Murray, and H. Larochelle, *RNADE: The Real-Valued Neural Autoregressive Density-Estimator*, NIPS, 2013.

[ZSPS20] A. Zeilmann, F. Savarino, S. Petra, and C. Schnörr, *Geometric Numerical Integration of the Assignment Flow*, Inverse Problems **36** (2020), no. 3, 034004 (33pp).

[Zwi16] P. Zwiernik, *Semialgebraic Statistics and Latent Tree Models*, CRC Press, 2016.

(B. Boll, D. Gonzalez-Alvarado) INSTITUTE FOR MATHEMATICS, IMAGE AND PATTERN ANALYSIS GROUP, HEIDELBERG UNIVERSITY, GERMANY
CORRESPONDING AUTHOR: DANIEL GONZALEZ-ALVARADO
Email address: daniel.gonzalez@iwr.uni-heidelberg.de
URL: https://ipa.math.uni-heidelberg.de

(S. Petra) MATHEMATICAL IMAGING GROUP, DEPARTMENT OF MATHEMATICS & CENTRE FOR ADVANCED ANALYTICS AND PREDICTIVE SCIENCES (CAAPS), UNIVERSITY OF AUGSBURG, UNIVERSITÄTSSTR. 14, 86159 AUGSBURG, GERMANY
Email address: Stefania.Petra@uni-a.de
URL: https://www.uni-augsburg.de/de/fakultaet/mntf/math/prof/mig/

(C. Schnörr) INSTITUTE FOR MATHEMATICS & RESEARCH STATION GEOMETRY AND DYNAMICS, IMAGE AND PATTERN ANALYSIS GROUP, HEIDELBERG UNIVERSITY, GERMANY
Email address: schnoerr@math.uni-heidelberg.de
URL: https://ipa.math.uni-heidelberg.de