Categorical Flow Matching on Statistical Manifolds

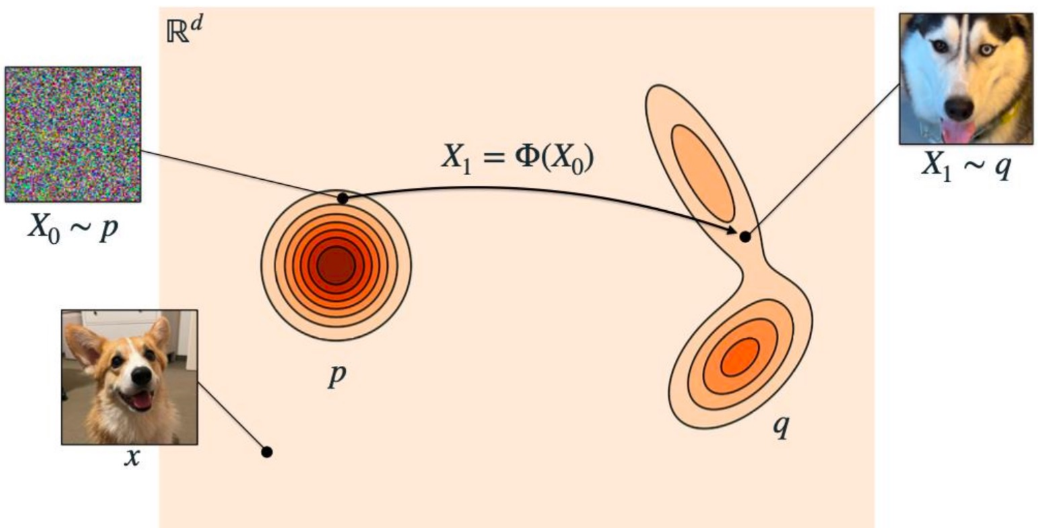
Chaoran Cheng, Jiahan Li, Jian Peng, Ge Liu

NeurIPS 2024

Presenter: Huyen Nguyen (huyentn2)

Introduction

Ways to map x_0 -> x_1



Flow matching as a generative model:

The goal is to find a flow mapping samples X_0 from a known source or noise distribution q into samples X_1 from an unknown target or data distribution q.

Design a time-continuous probability path (p_t) 0≤t≤1 interpolating between p := p_0 and q := p_1

Introduction

- **Information Geometry: F**rom information theory: all probability measures over the sample space form the structure known as ***statistical manifold*.**

- Suppose the statistical manifold $\mathcal{P} = \mathcal{P}(\mathcal{X}) = \{p: \int d\mu = \int p(x;\ \theta) d\upsilon = 1\}$ is parameterized by $\theta = (\theta_1, \theta_2, \dots, \theta_n) \in \theta$, this parameterization naturally provides a coordinate system for $\mathcal{P}$ on which each point is a probability measure μ with the corresponding probability density function $p(x;\ \theta)$

- The *Fisher information metric*

$$g_{jk}(\theta) = \mathbb{E}_X \left[\frac{\partial \log p(X;\theta)}{\partial \theta_j} \frac{\partial \log p(X;\theta)}{\partial \theta_k} \right] = \int_{\mathcal{X}} \frac{\partial \log p(x;\theta)}{\partial \theta_j} \frac{\partial \log p(x;\theta)}{\partial \theta_k} p(x;\theta)\, \mathrm{d}\nu. \qquad (1)$$

Introduction

Riemannian Manifold: A Riemannian manifold M is a real, smooth manifold equipped with a positive definite inner product g on the tangent space $T_x(\mathcal{M})$ at each point x ∈ $\mathcal{M}$. We can also define:

- ***geodesic*** γ(t) : [0, 1] → p, p ∈ $\mathcal{M}$ defines a "shortest" path (under the Riemannian metric) connecting two probability measures on the statistical manifold.

- ***geodesic distance*** between two probability measures, measures the similarity between them.

- The *tangent space* $T_x(\mathcal{M})$ at a point x ∈ $\mathcal{M}$ can be naturally identified with the affine subspace $T_x(\mathcal{M})$ = = $\{\mathrm{v} | \int dv = 0\}$ where each element v is a signed measure over sample space χ

- *exponential map* exp_x*:* $T_x(\mathcal{M}) \rightarrow \mathcal{M}$

- *logarithm map* log_x : $\mathcal{M} \rightarrow T_x(\mathcal{M})$

- Let $T\mathcal{M} = \bigcup_{x \in \mathcal{M}} T_x(\mathcal{M})$ be the *tangent bundle* of the manifold M, a time-dependent *vector field* on $\mathcal{M}$ is a mapping u_t : [0, 1] × $\mathcal{M} \rightarrow T\mathcal{M}$ where $u_t(x) \in T_x(\mathcal{M})$

Motivation

- Propose *Statistical Flow Matching* (SFM), a novel and mathematically rigorous generative **framework** on the manifold of parameterized probability measures by connecting Riemannian flow matching, information geometry, and natural gradient descent:
  - Tackle the discrete generation problem
  - Not pose any prior assumptions on the statistical manifold but instead deduces its intrinsic geometry via mathematical tools.
  - Deduce closed-form exponential and logarithm maps and develop an efficient flow- matching training algorithm that avoids numerical issues

- Further apply optimal transport during training and derive tractable exact likelihood for any given sample of probability measure, both of which are unachievable for most existing methods.

- Experiment the SFM with a toy example on simplex and on diverse real-world discrete generation tasks involving computer vision, natural language processing, and bioinformatics.

Conditional Flow Matching on Riemannian Manifold

- Consider a smooth Riemannian manifold $\mathcal{M}$ with the Riemannian metric g, a ***probability path*** $\boldsymbol{p_t}$:[0,1]→ $\mathcal{P}(\mathcal{M})$ is a curve of probability densities over $\mathcal{M}$. A ***flow*** Ψ_t :[0,1]× $\mathcal{M}$ → $\mathcal{M}$ is a time-dependent diffeomorphism defined by a **time-dependent vector field** u_t : [0, 1] × $\mathcal{M}$ → $T\mathcal{M}$ via the ordinary differential equation (ODE):

  $d\Psi_t$(x)/ dt = u_t (Ψ_t(x))

- The flow matching objective dt directly regresses the vector field ut with a time-dependent neural net $v(x_t, t)$ where x_t := Ψ_t(x)

- The Riemannian flow matching objective:

$$\mathcal{L} = \mathbb{E}_{t\sim U[0,1], x_0\sim p_0(x), x_1\sim q(x)}[\|v(x_t, t) - u_t(x_t|x_0, x_1)\|_g^2] \qquad (2)$$

Learned with a neural network

$u_t(x_t)$ = $d\Psi_t$(x)/ dt ← Derived Ψ_t(x) in the paper

Statistical Manifold of Categorical Distributions

- Consider the discrete sample space $\chi = \{1, 2, \ldots, n\}$, an n-class categorical distribution over X can be parameterized by n parameters $\mu_1, \mu_2, \ldots, \mu_n$ such that $\sum_{i=1}^{n} \mu_i = 1$, $\mu_i \geq 0$. In this way, the reference measure ν is the counting measure and the probability measure μ can be written as the convex combination of the canonical basis of Dirac measures $\{\delta^i\}_{i=1}^{n}$ over χ : μ $\sum_{i=1}^{n} \mu_i \delta^i$.

$$d_{\text{cat}}(\mu, \nu) = 2 \arccos \left(\sum_{i=1}^{n} \sqrt{\mu_i \nu_i} \right) \quad (3)$$

(5)

$$\langle u, v \rangle_\mu = \sum_{i=1}^{n} \frac{u_i v_i}{\mu_i}, \quad \mu \in \mathcal{P}_+, u, v \in T_\mu(\mathcal{P}) \quad (4)$$

- Introduce the following diffeomorphism:

$$\pi : \mathcal{P} \to S_+^{n-1}, \quad \mu_i \mapsto x_i = \sqrt{\mu_i}, \quad (5)$$

Statistical Manifold of Categorical Distributions

- Proposition 1.

$$d_S(\pi(\mu), \pi(\nu)) = \frac{1}{2} d_{\text{cat}}(\mu, \nu), \quad \mu, \nu \in \mathcal{P}. \quad (6)$$

$$\boxed{d_S(x, y) = \arccos(\langle x, y \rangle), \quad x, y \in S_+^{n-1}.} \quad (7)$$

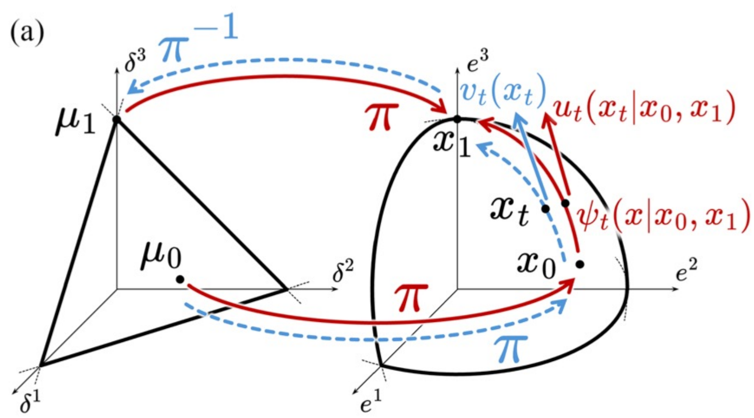


Figure 2: Statistical flow matching (SFM) framework.

Motivation: Note that the inner product is ill-defined on the boundary, causing numerical issues near the boundary.

The geodesic distance d_S and the inner product $\langle \cdot, \cdot \rangle$ are well-defined for the boundary, and we found this transform led to the practical stabilized training of the flow model.

Statistical Manifold of Categorical Distributions

- A *geodesic* is a locally distance-minimizing curve on the manifold. The existence and the uniqueness of the geodesic state that for any point $x \in \mathcal{M}$ and for any tangent vector $u \in T_x(\mathcal{M})$, there exists a unique geodesic γ : [0, 1] → $\mathcal{M}$ such that γ(0) = x and γ′(0) = u. The ***exponential map*** exp : $\mathcal{M} \times T\mathcal{M} \rightarrow \mathcal{M}$ is uniquely defined to be $exp_x(u)$:= γ(1). The ***logarithm map*** log : $\mathcal{M} \times \mathcal{M} \rightarrow T\mathcal{M}$ is defined as the inverse mapping of the exponential map such that $exp_x(log_x(y))$ ≡ y, ∀x, y ∈ $\mathcal{M}$
- With the exponential map and logarithm map, the time-dependent flow can be compactly written as time interpolation along the geodesic:

$$x_t := \psi_t(x_t | x_0, x_1) = \exp_{x_0}(t \log_{x_0} x_1), \quad t \in [0, 1]. \tag{15}$$

- It can be demonstrated that the above flow indeed traces the geodesic between x0, x1 with linearly decreasing geodesic distance $d_g(x_t, x_1)$ = (1 − t) $d_g(x_0, x_1)$

Statistical Manifold of Categorical Distributions

- Spherical Manifold

- The tangent space $T_x(S_+^{n-1}) = \{u \mid \langle u, x\rangle = 0\}$ is a (n − 1)-dimensional hyperplane perpendicular to the vector x.
- The geodesic on the sphere follows the great circle between two points, and the geodesic distance can be calculated in Eq:

$$d_S(x, y) = \arccos(\langle x, y\rangle), \quad x, y \in S_+^{n-1}. \tag{7}$$

- Exponential map, where $sinc(\theta) = \sin(\theta)/\theta$ is the unnormalized sinc function:

$$\exp_x(u) = x \cos \|u\|_2 + u \operatorname{sinc} \|u\|_2, \tag{19}$$

- The logarithm map can be calculated as:

$$\log_x(y) = \arccos(\langle x, y\rangle) \frac{P_x(y - x)}{\|P_x(y - x)\|_2}, \tag{20}$$

- Where $P(w) = w - < x, w > x$ is the projection of vector w onto the tangent space $T_x(S_+^{n-1})$.

Statistical Manifold of Categorical Distributions

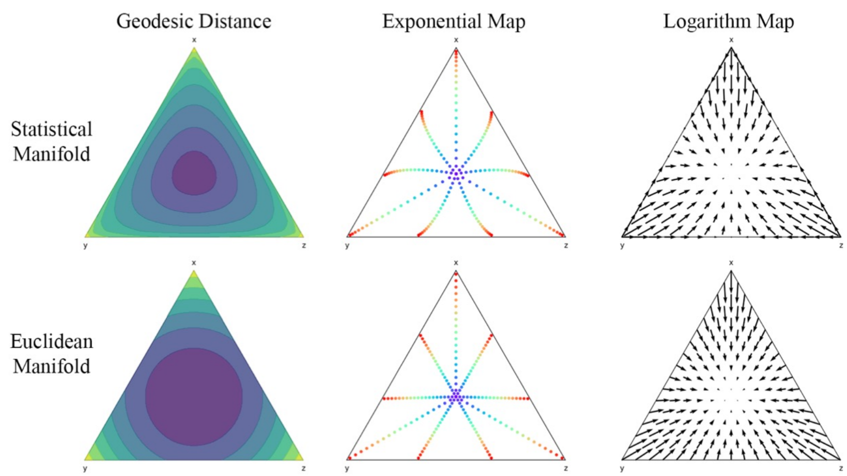


Figure 1: The Riemannian geometry of the statistical manifold for categorical distributions in comparison to Euclidean geometry on the simplex. **Left**: Contours for the geodesic distances to $\mu_0 = (1/3, 1/3, 1/3)$. **Middle**: Exponential maps (geodesics) from μ_0 to different points near the boundary. **Right**: Logarithm maps (vector fields) to μ_0.

Experimental Setup

manually project the predicted vector field onto the corresponding tangent space. For the spherical manifold, the projection can be described as

$$v_t(x_t) = \tilde{v}_t(x_t) - \langle x_t, \tilde{v}_t(x_t) \rangle x_t.$$

Algorithm 2 Training SFM

while not converged **do**
  Sample noise distribution $\mu_0 \sim p_0(\mu)$ and target distribution $\mu_1 \sim q(\mu)$.
  if optimal transport **then**
    Do batch OT assignments of μ_0 and μ_1 according to the average statistical distances.
  end if
  Apply the diffeomorphism in Eq.(5) to obtain $x_0 = \pi(\mu_0), x_1 = \pi(\mu_1)$.
  Sample $t \sim U[0,1]$ and interpolate $x_t = \exp_{x_0}(t \log_{x_0} x_1)$ using Eq.(19) and (20).
  Calculate the conditional vector field $u_t^S(x_t|x_0, x_1) = \boxed{\frac{\mathrm{d}}{\mathrm{d}t} x_t} = \log_{x_t}(x_1)/(1-t)$.
  Predict the vector field using $v(x_t, t)$ and optimize the SFM loss in Eq.(8).
end while

$$\boxed{x_t := \psi_t(x_t|x_0, x_1) = \exp_{x_0}(t \log_{x_0} x_1), \quad t \in [0,1].}$$

Experimental Setup

- Model Sampling

- The sampling process from the trained model can be described as solving the differential equation $\frac{\partial}{\partial t} x_t = v_t(x_t)$ from t = 0 to 1 with the initial conditional x_0 sampled from the prior noise distribution.

$$x_1 = x_0 + \int_0^1 v_t(x_t)\, \mathrm{d}t. \tag{49}$$

Algorithm 3 Sampling from SFM

1: Sample noise distribution $\mu_0 \sim p_0(\mu)$.
2: Apply the diffeomorphism in Eq.(5) to obtain $x_0 = \pi(\mu_0)$.
3: **if** ODE sampling **then**
4:     Solve $\frac{\partial}{\partial t} x_t = v_t(x_t)$ using Dopri5 ODE solver with initial condition x_0.
5: **else** ▷ Euler method
6:     **for** $t \leftarrow 0, 1/N, 2/N, \ldots, (N-1)/N$ **do**
7:         $x_{t+1/N} = \exp_{x_t}(v(x_t, t)/N)$
8:     **end for**
9: **end if**
10: **return** $\mu_1 = \pi^{-1}(x_1)$

Experimental Setup: NLL Calculation

- Exact Likelihood Calculation
  - For an arbitrary test sample $\mathrm{x} \in \mathcal{M}$, using the change of measure formula, the likelihood can be modeled by the continuity equation, where div_g is the Riemannian divergence and $v_t(x_t) := v(x_t, t)$ is the time-dependent vector field

$$\frac{\partial}{\partial t} \log p_t(x_t) + \mathrm{div}_g(v_t)(x_t) = 0, \qquad (11)$$

$$\log p(x_1) = \log p^{\mathrm{ODE}} + \log p_0(x_0)$$

$$\log p_1(\mu_1) = \log |\det \mathrm{d}\pi^{-1}(x_1)| + \log p^{\mathrm{ODE}}(x_1) + \log |\det \mathrm{d}\pi(\mu_0)| + \log p_0(\mu_0). \qquad (13)$$

Experimental Setup: NLL Calculation

Algorithm 1 NLL Calculation for Discrete Data

1: Sample $\tilde{\mu}_1 \sim q_t(\mu|\delta)$ in Eq.(30) and calculate $-\log q_t(\tilde{\mu}_1|\delta)$ and $\log p(\delta|\tilde{\mu}_1)$.
2: Apply the diffeomorphism in Eq.(5) to obtain $\tilde{x}_1 = \pi(\tilde{\mu}_1)$ and calculate $\log|\det \mathrm{d}\pi^{-1}(\tilde{x}_1)|$.
3: Solve the ODE system in Eq.(41) to obtain x_0 and $\log p^{\mathrm{ODE}}$.
4: Apply π^{-1} to obtain $\mu_0 = \pi^{-1}(x_0)$ and calculate $\log|\det \mathrm{d}\pi(\mu_0)|$.
5: Calculate the base log probability $\log p_0(\mu_0)$.
6: **return** NLL as in Eq.(14).

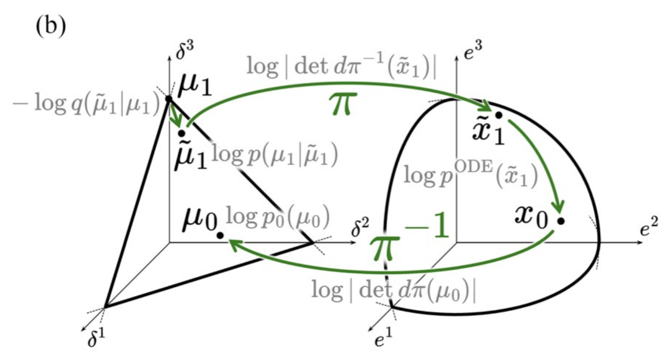


Figure 2: Statistical flow matching (SFM) framework.

In the NLL calculation for one-hot examples (Sec.3.5), the probability density is marginalized over a small neighborhood of some Dirac measure to avoid undefined behaviors at the boundary (in green).

Experiments

- Toy Example: Swiss Roll on Simplex
- Binarized MNIST
- Text8
- Promoter DNA Design

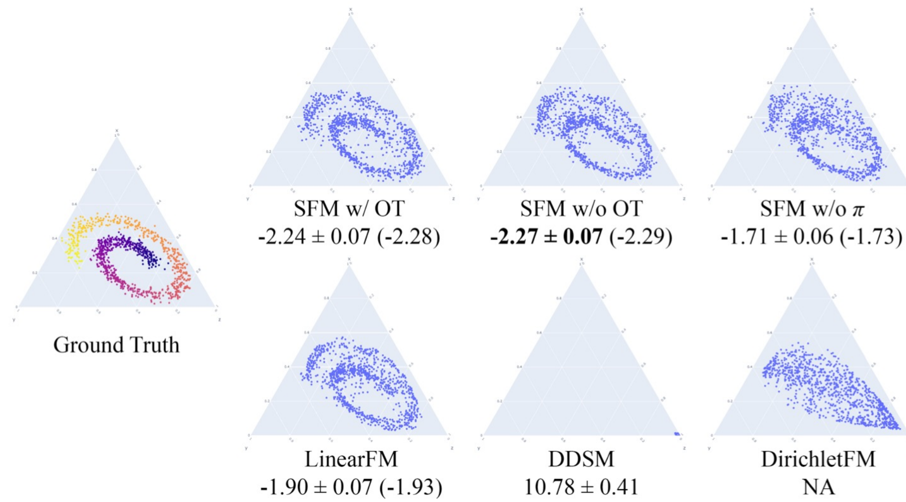


Figure 3. Generated samples of the Swiss roll on simplex dataset and NLL (lower is better). The NLLs are estimated using Hutchinson's trace estimator, whereas those in the parenthesis are exact.

Experiments and Results

Table 1: NLL and FID of different discrete
are discrete NLLs; therefore, they are not

Model	SFM w/ OT	SFM
NLL↓	**-1.687 ± 0.020**	-1.63
FID↓	**4.62**	
Model	DirichletFM	D
NLL↓	NA	0.100
FID↓	77.35	

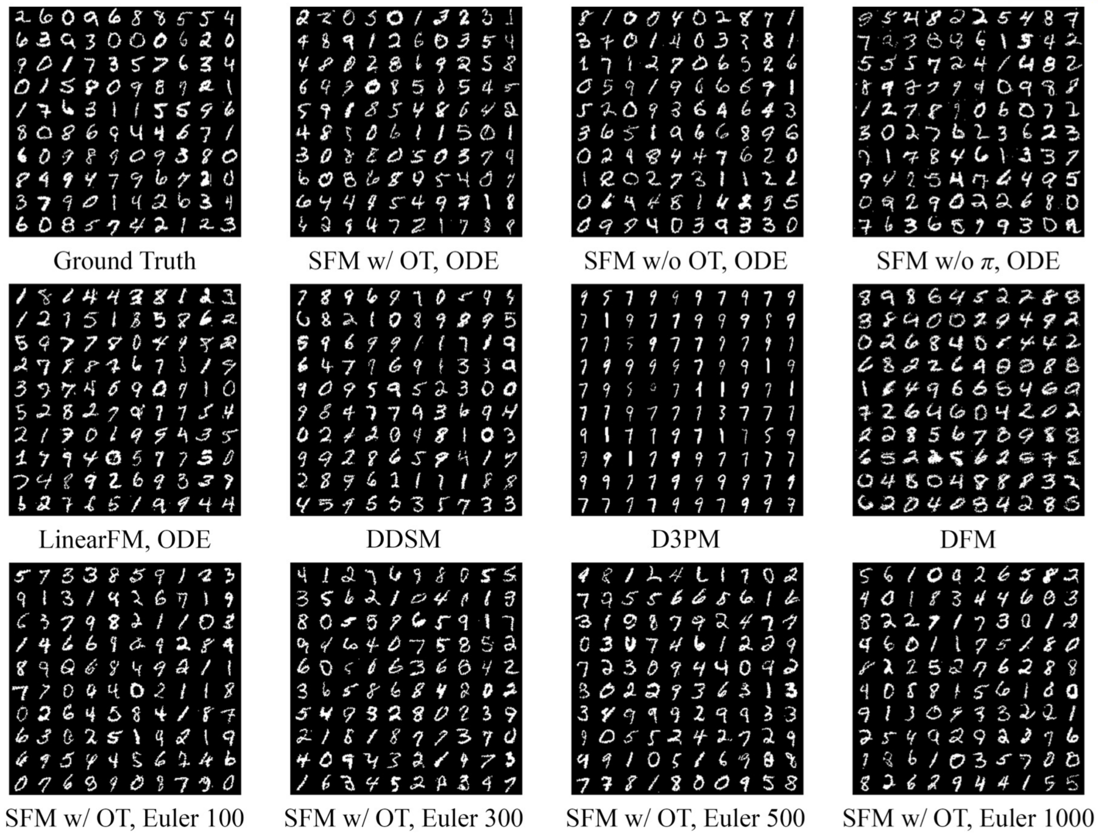


Figure 5: Generated samples of the binarized MNIST dataset from various models and different sampling settings.

Experiments and Results

Table 2: BPC on Text8. Results marked * are taken from the corresponding papers.

Model	BPC↓
SFM w/ OT	1.399 ± 0.020
SFM w/o OT	1.386 ± 0.033
LinearFM	1.651 ± 0.027
D3PM-absorb[6]	1.47*
BFN[24]	1.41*
SEDD-absorb[39]	**1.32***
MultiFlow[12]	1.41*
Argmax Flow[28]	1.80*
Discrete Flow[64]	1.23*
Transformer[6]	1.23*
Transformer XL[16]	**1.08***

Table 3: SP-MSE (as evaluated by Sei [13]) on the generated promoter DNA sequences. Results marked * are from [7] and results marked † are from [60].

Model	SP-MSE↓
SFM w/ OT	0.0279
SFM w/o OT	**0.0258**
LinearFM	0.0282
DDSM	0.0334*
D3PM-uniform	0.0375*
Bit-Diffusion (one-hot) [15]	0.0395*
Bit-Diffusion (bit) [15]	0.0414*
Language Model	0.0333†
DirichletFM	0.0269†

Experiments and Results

SFM w/ OT, ODE, NLL: 6.762, Entropy: 7.340

zero_zero_zero_more_as_well_as_the_needed_of_all_of_it_church_the_country_s_higner_upcoming_bank_the_country_comment_on_quebec_e dits_includes_the_account_of_diego_hyle_ciaspare_coes_tain_three_zero_seven_zero_millimeter_if_south_of_the_south_leo_jordan_the	NLL: 6.336
such_as_in_outcarge_of_coincination_with_mows_such_as_adler_martie_the_hilly_patt_evedhon_of_morcele_s_night_of_blood_the_tremen t_of_eliensberg_while_an_ulav_at_esrheim_that_he_had_to_proved_left_mainied_this_label_is_in_hellenistic_separatism_the_falix_ro	NLL: 6.805
t_orator_lemmoi_s_mother_toury_ghost_for_his_history_on_a_blaster_the_three_stallman_family_sources_including_the_film_that_a_ro mance_nine_author_higtly_lacaded_the_second_harmour_open_source_for_which_orrie_changed_the_bluebogs_books_moy_s_athlite_s_medit	NLL: 7.522

SFM w/o OT, ODE, NLL: 6.811, Entropy: 7.387

_became_known_as_the_shacon_valley_to_the_heaven_green_and_in_the_middle_of_the_lechneit_tracked_the_line_kej_nis_a_valley_one_p inochules_this_was_verified_by_many_charterly_brollary_applications_including_those_which_synonymous_with_orbits_some_of_the_mas	NLL: 6.407
cable_now_masi_had_little_to_port_from_six_eight_nine_made_hofavor_a_new_printer_of_disruption_this_platforv_would_be_faving_to_ the_current_country_but_this_need_for_saw_della_even_this_four_one_three_bit_moil_callers_did_soo_after_a_as_n_if_platform_for_t	NLL: 6.819
nomic_ancestor_wh_meil_berg_hiarst_red_rthonstrak_utter_upon_technology_baddendin_models_on_bendrays_hypothesies_anti_aer_dynami cs_work_have_been_intelligent_to_develop_an_european_astronomic_conifice_in_the_production_of_ten_conifices_of_develop_and_princ	NLL: 7.479

LinearFM, ODE, NLL: 6.935, Entropy: 7.356

is_resulted_in_gawzik_college_in_the_five_season_of_feason_at_twice_the_atmosphere_is_named_after_the_list_called_him_before_inn _s_college_at_stulpford_university_of_london_also_cambridge_the_burroughs_henrians_college_which_is_yelled_apollo_one_college_na	NLL: 6.466
ne_two_eight_zero_perhaps_that_one_s_stream_roman_frxwuapered_the_practices_of_telleeist_speakership_settled_and_an_army_of_the_ two_set_of_love_relationships_the_foundation_of_the_colfederation_homewater_to_during_the_civil_war_or_dan_brown_xian_john_zinso	NLL: 6.935
level_mortans_already_sick_but_evade_dissolve_the_moses_of_auctional_with_deng_about_four_sekes_there_was_a_moikade_problem_to_p eople_who_receive_signed_grief_of_culture_of_the_middle_bone_island_for_a_more_designation_of_a_kick_trade_bands_and_rangers_bom	NLL: 7.454

MultiFlow, $T = 1$, NLL: 6.728, Entropy: 7.387

er_of_the_soap_opera_by_andrew_wills_goosecat_productions_one_nine_nine_one_the_sea_monsters_of_the_late_one_nine_nine_zero_s_th e_famous_woman_stanley_goodman_jerdre_mcnabb_out_of_zoom_movie_barry_leroy_barbara_lewis_and_brenda_punceco_aka_sney_steary_aka_	NLL: 5.906
she_hill_obhalarnnach_eochrans_eileann_munthar_cearlomha_mhaonna__tardare_mho_mord_tore_lian_linn_mu_phaile_gael_cangallauig_lao thuis_guilleith_leois_glion_guildh_lara_gall_innonte_tilbonne_guilecht_shuachtansert_guillaste_guatnaoic_asthache_cuichant_conai	NLL: 6.648
lde_its_replacement_and_or_not_mist_mere_decabeod_man_and_drast_m_ek_or_ubangostrades_dialogue_or_connon_cainne_as_follows_make_ wolsey_conane_i_get_clean_to_contemplate_the_static_problem_to_reduce_it_into_perception_for_frbellist_man_jewish_views_the_othe	NLL: 7.426

Conclusion

- The authors proposed statistical flow matching (SFM) as a general generative framework for generative modeling on the statistical manifold of probability measures.
- By leveraging results from information geometry, the proposed SFM effectively captures the underlying intrinsic geometric properties of the statistical manifold.
- Applied SFM to diverse downstream discrete generation tasks across different domains to demonstrate our framework's effectiveness over the baselines.
- Future work: SFM can be further extended to non-discrete generative tasks whose targets are probability distributions.
- Limitations of SFM framework:
  - As a special case of the flow matching model, the generation is an iterative process of refinement that cannot modify the size of the initial input. This may pose limitations to generation compared with autoregressive models.
  - Imposed the assumption of independence between classes so that the canonical Riemannian structure can be induced by the Fisher metric. However, discretized data like CIFAR-10 (256 ordinal pixel values) do not follow this assumption