Transportation of Probability measures and its application in Generative models

Hyemin Gu

hgu@umass.edu

March 10, 2023

Outline

1. Optimal transport and Gradient flows of Probability measures
   - Optimal transport problem
   - Gradient flow of measures and flow of transport maps
   - Variational representation and its optimization

2. Generative models
   - What is generative modeling and How it is related to transportation of probability measures
   - Generative models based on transportation of probability measures
   - Common difficulties of these models

Optimal transport and Gradient flows of Probability measures

Notations

$\mathcal{X}$	A domain. ex) $\mathbb{R}^d$
$\mathcal{P}(\mathcal{X})$	Set of probability measures on $\mathcal{X}$. $\int_{\mathcal{X}} dP = 1$
P, Q	Input / target probability measures : $\sigma_{\mathcal{X}} \to [0, \infty)$
$\Pi(P, Q)$	Set of couplings between P and Q $\int_x d\gamma(x, y) = dQ(y)$ and $\int_y d\gamma(x, y) = dP(x)$.
T	Transport map. $T : \mathcal{X} \to \mathcal{X}$
$\mathrm{T}_{\#}P$	Pushforward measure. $P(T^{-1}(A))$ for $A \in \sigma_{\mathcal{X}}$
F	Free energy functional. $F : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$
f^c	c-transform of f. $f^c(y) = \inf_x c(x, y) - f(x)$
f^*	Legendre transform of f. $f^*(t) = \sup_x < t, x > -f(x)$
C_b	Set of continuous bounded functions $C_b(\mathcal{X})$
$\mathcal{H}_k$	RKHS generated by kernel $k(x, y)$

Optimal transport problem

Monge

$$\inf_{T:\mathcal{X}\to\mathcal{X}} \int c(x, T(x))dP(x) : P, Q \in \mathcal{P}(\mathcal{X}), T_{\#}P = Q \tag{1}$$

Kantorovich

$$\inf_{\gamma\in\Pi(P,Q)} \int\int c(x,y)d\gamma(x,y) \tag{2}$$

ex) $c(x,y) = |x-y|^p, p \geq 1$: We obtain Wasserstein$_p$ distance $W_p^p(P,Q)$

Kantorovich Dual

$$\sup_{\phi:\mathcal{X}\to\mathbb{R}} \int \phi(x)dP(x) + \int \phi^c(y)dQ(y) : \phi(x) + \phi^c(y) \leq c(x,y) \tag{3}$$

ex) $W_1(P,Q) = \sup_{\phi:1-\text{Lipschitz}} \int \phi(x)dP(x) - \int \phi(y)dQ(y)$

Gradient flows - Continuity equation

Consider a flow of transport maps T_t which gives a flow of probability measures $P_t, t \geq 0$ transported from P to Q.
Gradient flows endowed with Wasserstein distance can model this problem as:

$$\partial_t P_t + \nabla \cdot (P_t \mathbf{V_t}) = 0, P_0 = P, P_\infty = Q \tag{4}$$

where the vector field $\mathbf{V_t}$ is the direction to get P_t closer to Q.
Question) How to find if a probability measure is close to another?

Divergences

$P = Q$ if $\int \phi dP = \int \phi dQ$ for all bounded measurable functions $\phi \in \mathcal{M}_b(\mathcal{X})$. Divergence $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to [0, \infty]$ is a function which satisfies

$$D(P, Q) = 0 \text{ } iff \text{ } P = Q. \tag{5}$$

So, consider a functional $F : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ such that $F(P_t) = D(P_t, Q)$ and let $F(P_t) \to 0$. We say F, free energy functional.

- ex) $F(P_t) = D_{KL}(P_t \| Q) = \mathbb{E}_{P_t}\left[\log \frac{dP_t}{dQ}\right] = \mathbb{E}_Q\left[\frac{dP_t}{dQ} \log \frac{dP_t}{dQ}\right]$ whenever $P_t << Q$, otherwise $+\infty$.

f-divergences

Definition) f-divergence

$f : (0, \infty) \to \mathbb{R}$ convex, $f(1) = 0$, lower semi-continuous
$F(P) = D_f(P\|Q) := \mathbb{E}_Q\left[f(\frac{dP}{dQ})\right]$
Consider super linear f i.e. $\lim_{t\to+\infty} \frac{f(t)}{t} = +\infty$ so that $D_f(P\|Q) < \infty$ only if $P << Q$.

- ex) $f_{KL}(x) = x \log x$, $f_\alpha(x) = \frac{x^\alpha - 1}{\alpha(\alpha-1)}, \alpha > 1$
- $(P, Q) \mapsto D_f(P\|Q)$ is convex.
- $P \mapsto D_f(P\|Q)$ is strictly convex.
- Asymmetric.
- Variational representation of f-divergences

$$D_f(P\|Q) = \sup_{\phi\in C_b} \mathbb{E}_P[\phi] - \mathbb{E}_Q[f^*(\phi)] \tag{6}$$

f-divergence is asymmetric

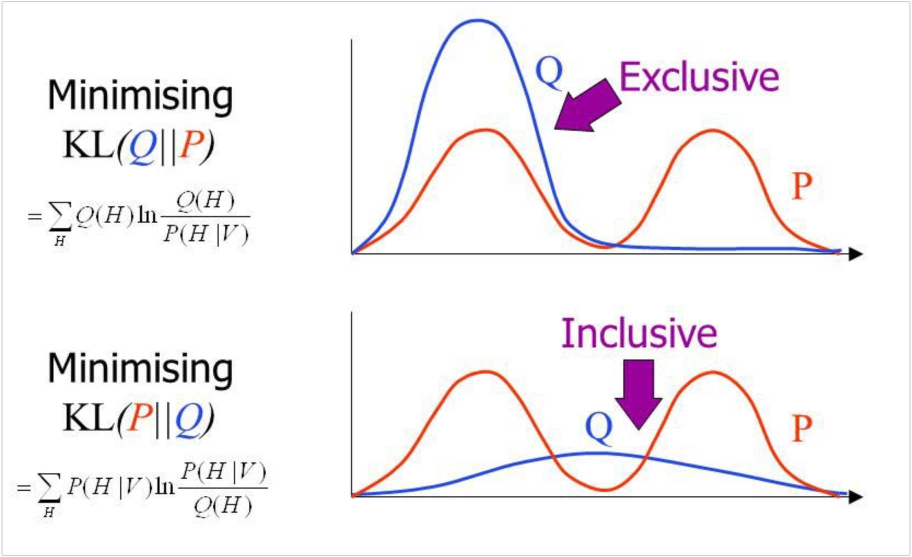

Figure by John Winn.

Integral probability metrics

Definition) Integral probability metric

For some function space $\mathcal{F}$,

$$F(P) = d_{\mathcal{F}}(P, Q) := \sup_{\phi \in \mathcal{F}} E_P[\phi] - E_Q[\phi] \tag{7}$$

- ex) Maximum Mean Discrepancy: $\mathcal{F} = \{\phi \in \mathcal{H}_k : \|\phi\|_{\mathcal{H}_k} \leq 1, \mathcal{H}_k : RKHS\}$
- ex) W_1: $\mathcal{F} = \{\phi : \ 1 - \text{Lipschitz}\}$
- A distance.
- Not require absolute continuity between probability measures.
- Not strictly convex.

Gradient flow which minimizes KL divergence

Return to the problem of finding a transport map T_t which transports $P_t, t \geq 0$ over time by the gradient flow

$$\partial_t P_t + \nabla \cdot (P_t \mathbf{V_t}) = 0, P_0 = P, P_\infty = Q. \tag{8}$$

ex) [BVE22] Choose $\mathbf{V_t}$ in order to minimize the KL divergence of P_t and Q. Then P_t can be written as

$$P_t(x) = P(T_t(x)) \exp\left(- \int_0^t \nabla \cdot \mathbf{V}_\tau \left(T_\tau(x) \right) d\tau \right), x \sim P. \tag{9}$$

Now let's see how to select $\mathbf{V_t}$ in order to minimize F.

Free energy functionals and physical meaning

Consider free energy functional $F : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ consisting of three terms: internal energy, potential energy, and interaction energy

$$F(P) = \int U(P(x))dx + \int V(x)dP(x) + \int \int W(x,y)dP(x)dP(y) \quad (10)$$

where $U \in C^1$, U, U' of polynomial growths.
We can write KL divergence adjusted to the form above.

- $F(P) = D_{KL}(P\|Q)$:
  $D_{KL}(P\|Q) = \mathbb{E}_P\left[\frac{dP}{dQ}\right] = \int P(x)\log P(x)dx - \int \log Q(x)dP(x)$
- $F(P) = D_\alpha(P\|Q)$:
  $D_\alpha(P\|Q) = \frac{1}{\alpha(\alpha-1)}\mathbb{E}_Q\left[\left(\frac{dP}{dQ}\right)^\alpha - 1\right] = \frac{1}{\alpha(\alpha-1)}\mathbb{E}_P\left[\left(\frac{dP}{dQ}\right)^{\alpha-1} - 1\right]$

First variation of free energy functional

The free energy functional F of the form (10) admits the (Gateaux) derivative of F w.r.t. P: First variation $\frac{\delta F}{\delta P}$ satisfies

$$\frac{d}{d\epsilon} F(P+\epsilon\rho)|_{\epsilon=0} = \int \frac{\delta F}{\delta P} d\rho. \tag{11}$$

For (10), closed form exists and is unique up to a constant:

$$\frac{\delta F}{\delta P}(x) = U'(P(x)) + V(x) + \int W(y,x)dP(y) + \int W(x,y')dP(y'). \tag{12}$$

- $F(P) = D_{KL}(P\|Q)$: $U(x) = x\log x$, $V(x) = -\log Q(x)$
  $\Rightarrow \frac{\delta F}{\delta P}(x) = \log P(x) - \log Q(x)$
- Other case? Calculate (11) directly.

First variation of variational representation

Once we solve the variational representations over functions ϕ
$F(P) = D_f(P\|Q) = \sup_{\phi \in C_b} \mathbb{E}_P[\phi] - \mathbb{E}_Q[f^*(\phi)]$ or
$F(P) = d_{\mathcal{F}}(P, Q) = \sup_{\phi \in \mathcal{F}} E_P[\phi] - E_Q[\phi]$, we get

$$F(P + \epsilon\rho) = F(P) + \epsilon \int \phi d\rho \tag{13}$$

where ϕ is an optimizer. And so the first variation $\frac{\delta F}{\delta P} = \phi$.

Gradient flow which minimizes F

Determine the vector field as $\mathbf{V_t} = -\nabla \frac{\delta F}{\delta P_t}$. Then the continuity equation

$$\partial_t P_t + \nabla \cdot (P_t \mathbf{V_t}) = 0, P_0 = P, P_\infty = Q \tag{14}$$

reduces to

$$\partial_t P_t = \nabla \cdot \left(P_t \nabla \frac{\delta F}{\delta P_t} \right), P_0 = P, P_\infty = Q. \tag{15}$$

Moreover, in case we use the variational representations of f-divergences or IPMs and obtain an optimizer ϕ_t, we have

$$\partial_t P_t = \nabla \cdot (P_t \nabla \phi_t), P_0 = P, P_\infty = Q. \tag{16}$$

Numerical schemes I

Minimizing movement scheme [JKO98]

$$P_{t+1} = arginf_R \, \frac{W_2^2(P_t, R)}{2\Delta t} + F(R) \tag{17}$$

- F minimizing property : $\frac{W_2^2(P_t, P_{t+1})}{2\Delta t} + F(P_{t+1}) \leq F(P_t)$
- Optimality condition: $\frac{\phi}{\Delta t} + \frac{\delta F}{\delta P_{t+1}} = constant$ where ϕ denotes the Kantorovich potential of (3) with cost $\frac{1}{2}|x-y|^2$.
- $\frac{T(x)-x}{\Delta t} = \frac{\nabla \phi}{\Delta t} = -\nabla \frac{\delta F}{\delta P_{t+1}}$ and get (Implicit) Proximal gradient

$$P_{t+1} = T_\# P_t = \left(I - \Delta t \nabla \frac{\delta F}{\delta P_{t+1}} \right)_\# (P_t). \tag{18}$$

Numerical schemes II

Forward Euler

Exchange the implicit term with (less costly) explicit $\Rightarrow$ Gradient descent

$$P_{t+1} = \left(I - \Delta t \nabla \frac{\delta F}{\delta P_t} \right)_{\#} (P_t). \tag{19}$$

Solving the variational representations for the functional ϕ is much easier than directly solving for the transport map T_t or the measure P_{t+1}. Then recover T_t and P_{t+1} by the (Forward) Euler,

- $T_t = (I - \Delta t \nabla \phi_t) \circ \cdots \circ (I - \Delta t \nabla \phi_0)$
- $P_{t+1} = (T_t)_{\#} P_0$.

<u>Idea</u>: Given finite number of samples, optimize ϕ_t over some function spaces $\mathcal{F} \subset C_b$ parametrized by Neural networks, Reproducing kernel Hilbert spaces, etc.

Neural networks

- Given i.i.d. samples $X^{(i)} \sim Q$ and $Y_t^{(i)} \sim P_t$ for $i = 1, \cdots, N$, approximate $D_f(P_t \| Q)$ by optimizing $\phi_t(x) = \mathcal{N}\mathcal{N}(x; W_t)$

$$\sup_{W_t} \frac{1}{N} \sum_i \phi(Y_t^{(i)}; W_t) - \frac{1}{N} \sum_i f^*(\phi(X^{(i)}; W_t)) + regularizer. \quad (20)$$

- Choose right activation functions and/or regularizer to approximate (subsets of) continuous real valued functions.
  - Lipschitz: $\|W^l\|_2 \leq L^{1/D}, l = 1, \cdots, D$ with $\mathrm{ReLU}(x) = \max(x, 0)$ activation functions (Spectral normalization [MKKY18])
  - Lipschitz: add gradient penalty term in the loss $regularizer = \int \max(|\nabla \phi(x)|^2 / L^2 - 1, 0) dP(x)$ [BDK+22]
  - Smoother function: smoother activation functions
  - Bounded function: bounded activation functions in the last layer
- $\nabla \phi_t(x; W_t)$ can be attained by Automatic Differentiation.

Reproducing kernel Hilbert spaces I

- RKHS $\mathcal{H}_k$ with kernel k. $k(\cdot, x)$ is continuous and $\sup_x \sqrt{k(x,x)} < \infty$ so that $\mathcal{H}_k \subset C_b(\mathcal{X})$.
- Choose a kernel k and use Representer theorem on a subset of data.

Representer theorem

In RKHS $\mathcal{H}_k$ with kernel k, given

- m training samples $x_i, i = 1, \cdots, m$
- a strictly increasing function $g : [0, \infty) \to \mathbb{R}$
- an arbitrary error function E

any minimizer of the empirical risk

$$\phi^* = argmin_{\phi \in \mathcal{H}_k} \left\{ E\left((x_i, \phi(x_i))_{i=1}^m \right) + g(\|\phi\|) \right\} \tag{21}$$

admits a representation of the form $\phi^*(x) = \sum_{i=1}^m \alpha_i k(x, x_i)$, $\alpha_i \in \mathbb{R}$.

Reproducing kernel Hilbert spaces II

- Given i.i.d. samples $X^{(i)} \sim Q$ and $Y_t^{(i)} \sim P_t$ for $i = 1, \cdots, N$, choose m samples $\left\{Z_t^{(j)}, j = 1, \cdots, m\right\}$ from $\left\{X^{(i)}, i = 1, \cdots, N\right\} \cup \left\{Y_t^{(i)}, i = 1, \cdots, N\right\}$.
- To approximate $D_f(P_t \| Q)$, optimize $\boldsymbol{\alpha_t} \in \mathbb{R}^m$ by

$$\sup_{\boldsymbol{\alpha_t}} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{m} \alpha_t^j k\left(Y_t^{(i)}, Z_t^{(j)}\right) - \frac{1}{N} \sum_{i=1}^{N} f^* \left(\sum_{j=1}^{m} \alpha_t^j k\left(X^{(i)}, Z_t^{(j)}\right) \right) + \textit{regularizer} \tag{22}$$

  and get optimizer $\hat{\phi}_t(\mathbf{x}) = \sum_{j=1}^{m} \hat{\alpha}_t^j k\left(\mathbf{x}, Z_t^{(j)}\right)$.
- $\nabla \hat{\phi}_t(\mathbf{x}) = \sum_{j=1}^{m} \hat{\alpha}_t^j \nabla_{\mathbf{x}} k\left(\mathbf{x}, Z_t^{(j)}\right)$

Another approach in RKHS

KL Approximate Lower bound Estimator [GAG21] - "Primal"

$$KALE_\lambda^P(P\|Q) = (1+\lambda) \max_{\phi \in \mathcal{H}_k} \mathbb{E}_P[\phi] - \mathbb{E}_Q[\exp(\phi) - 1] - \frac{\lambda}{2}\|\phi\|^2_{\mathcal{H}_k} \quad (23)$$

Write $J(\phi) = -KALE_\lambda^P(P\|Q)$. (convex)
The optimal value of (23) is $\hat{J} = \max_\phi < 0, \phi > -J(\phi) = J^*(0)$.
Apply infimal convolution theorem $(f_1 * f_2)^* = f_1^* + f_2^*$ on $J^*(0)$.

KL Approximate Lower bound Estimator [GAG21] - "Dual"

$$(D) : \min_{\psi > 0} \mathbb{E}_Q[\psi(\log\psi - 1) + 1] + \frac{1}{2\lambda}\left\| \int \psi(x)k(x,\cdot)dQ(x) - \mu_P \right\|^2_{\mathcal{H}_k} \quad (24)$$

Generative models

Generative models and transportation of measures

Generative models

In Machine learning discipline, generative modeling is to model a target probability distribution and produce "new" samples from the model which are thought to be from the true probability distribution.

** It can be discriminated from "sampling" in the sense of "How much information is provided about the target probability distribution" (target density is known up to a certain point or only samples are given).*

Why transportation of probability measures?

We begin generating samples from a known distribution which is easy to sample from. Transportation of probability measures arises in this context.

Fokker-Planck equation and particles ODE model

Many generative models aim to minimize KL divergence. The KL gradient flow is well-known Fokker-Planck equation.
Given densities p_t, q, the Fokker-Planck equation reads

$$\partial_t p_t = \nabla \cdot (p_t \nabla \log (p_t/q)) = \Delta p_t - \nabla \cdot (p_t \nabla \log q). \tag{25}$$

Interacting particles systems: Particles ODE

The gradient flow formulation in (25) lead to a system of ODEs for the particles

$$\dot{Y}_t = -\nabla \log (p_t/q)(Y_t),\, Y_0 \sim P_0. \tag{26}$$

Fokker-Planck equation and particles SDE model

Given densities p_t, q, the Fokker-Planck equation reads

$$\partial_t p_t = \nabla \cdot (p_t \nabla \log (p_t/q)) = \Delta p_t - \nabla \cdot (p_t \nabla \log q). \qquad (27)$$

Langevin diffusion: Particles SDE

The diffusion Δp_t in the (RHS) of (27) can be separately modeled as Brownian motion $W_t \sim N(0, tI)$, leading to a system of SDEs for the particles

$$dY_t = \nabla \log q(Y_t)dt + \sqrt{2}dW_t,\ Y_0 \sim P_0. \qquad (28)$$

Stein variational gradient descent [Liu17]

Modify the Fokker-Planck equation further by letting $h_t = p_t/q$,

$$\partial_t h_t = (\nabla + \nabla \log q) \cdot \nabla h_t. \tag{29}$$

"$\nabla + \nabla \log q$" is named Stein operator and induces a KL gradient flow endowed with Stein-Wasserstein metric.
In RKHS $\mathcal{H}_k$ with kernel k, parametrize the Stein operator by kernels

$$g(x, y) = \nabla_x k(x, y) + \nabla \log q(x) k(x, y). \tag{30}$$

Given i.i.d. particles $Y_0^{(i)} \sim P_0, i = 1, \cdots, N$, solve a system of ODEs

$$\dot{Y}_t = \mathbf{V_t}(Y_t) \tag{31}$$

where $\mathbf{V_t}(x) = \frac{1}{N} \sum_{i=1}^{N} g(Y_t^{(i)}, x)$ from the representer theorem.

Normalizing flows [RM15]

Assume there is a smooth invertible map $f : \mathcal{X} \to \mathcal{X}$ which maps the density p_0 to q as

$$q(x) = p_0(y) \left| \det \frac{\partial f^{-1}}{\partial x} \right| = p_0(y) \left| \det \frac{\partial f}{\partial y} \right|^{-1}. \tag{32}$$

NFs maximize log-likelihood(precisely, ELBO) of the third term in (32).

Continuous normalizing flows [CRBD18] and ODEs

Parametrize temporary transport map by time t and induce particles ODEs $\dot{Y}_t = f_t(Y_t)$. $T_t(Y_0) = Y_0 + \int_0^t f_s(Y_s)ds$. The log-likelihood evolves by

$$\frac{d \log p_t(y)}{dt} = -Tr\left(\frac{\partial f_t}{\partial y}\right). \tag{33}$$

More SDE approaxches [SSDK+20]

A SDE or Itô process describes an evolution of random variable $X_t \in \mathbb{R}^d$ as

$$dX_t = b(X_t, t)dt + \sigma(X_t, t)dW_t \tag{34}$$

where $b(\mathbf{x}, t) \in \mathbb{R}^d$ and $\sigma(\mathbf{x}, t) \in \mathbb{R}^{d\times d}$.
If X_t is the solution of (34), its density $p(\mathbf{x}, t)$ satisfies the forward and backward evolutions [Øks14]:

$$\partial_t p(\mathbf{x}, t) = -\nabla_{\mathbf{x}} \cdot (b(\mathbf{x}, t)p(\mathbf{x}, t)) + \frac{1}{2}\sum_{i,j} \partial_{ij}(\boldsymbol{\sigma}_i^T \boldsymbol{\sigma}_j(\mathbf{x}, t)p(\mathbf{x}, t)) \tag{35}$$

$$-\partial_t p(\mathbf{x}, t) = b(\mathbf{x}, t)\nabla_{\mathbf{x}} \cdot p(\mathbf{x}, t) + \frac{1}{2}\sum_{i,j} \boldsymbol{\sigma}_i^T \boldsymbol{\sigma}_j(\mathbf{x}, t)\partial_{ij} p(\mathbf{x}, t) \tag{36}$$

<u>Idea</u>: Handle the forward transition probability $q(X_t|X_{t-1})$ with (35) or the backward transition probability $q(X_{t-1}|X_t)$ with (36) to be normal.

Common difficulties of these models: High dimensionality

- Gradient signals are diluted due to various regularizations.
  ex) Constraining $\|\nabla\phi(x)\| \leq L$ leads to the average axis-wise velocity component to be $|[\nabla\phi(x)]_i| \leq \frac{L}{\sqrt{d}}$. Slow.
  $L = \mathcal{O}(\sqrt{d})$? It might reduce the stability of the method.
- Target distribution is supported in low dimensional manifolds - approximation where $Q(x) \approx 0$ is inaccurate.
- $\Rightarrow$ Relieve the problem by projecting to a latent space :
  - Consider a low dimensional manifold $S \subset supp(Q)$ in $\mathbb{R}^d$ and for $d' < d$ a projection map $\mathcal{E}_{d'} : \mathbb{R}^d \to \mathbb{R}^{d'}$ which is invertible in S.
  - Call $\mathcal{E}_{d'}(\mathbb{R}^d)$ be a latent space for d' dimensional features.
  - A systematic approach to obtain feature vectors?

Self-attention [VSP+17, TJ19, LEE21]

$\mathbf{X}, \mathbf{Y}$: input and output random variables in $\mathbb{R}^d$. For simplicity, assume independency among X_i's and Y_i's. Factorization of the joint distribution using conditional independence among random variables:

$$p(x_{1:d}, y_{1:d}) = \prod_{i=1}^{d} p(x_i)p(y_i|x_{1:i-1}) = \prod_{i=1}^{d} p(x_i)p(y_i|Pa(y_i)) \tag{37}$$

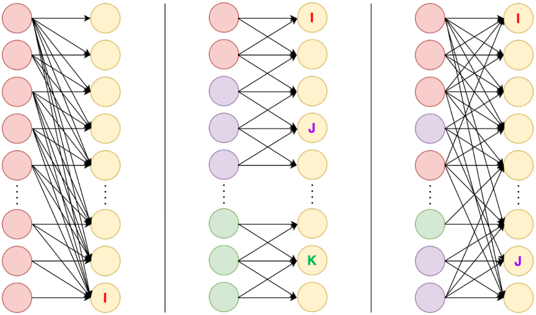

Figure: Bipartite graph of input entries and output entries.
(a) Chain rule.
(b) CNN.
(c) Self-attention in transformer.

Any questions OR Clarification

Bibliography I

Jeremiah Birrell, Paul Dupuis, Markos Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet.
(f,gamma)-divergences: Interpolating between f-divergences and integral probability metrics.
Journal of Machine Learning Research, 23:1–70, 01 2022.

Nicholas M. Boffi and Eric Vanden-Eijnden.
Probability flow solution of the fokker-planck equation, 2022.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud.
Neural ordinary differential equations.
In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Bibliography II

Pierre Glaser, Michael Arbel, and Arthur Gretton.
Kale flow: A relaxed kl gradient flow for probabilities with disjoint support.
In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8018–8031. Curran Associates, Inc., 2021.

Richard Jordan, David Kinderlehrer, and Felix Otto.
The variational formulation of the fokker–planck equation.
SIAM Journal on Mathematical Analysis, 29(1):1–17, 1998.

JUSTIN SEONYONG LEE.
Transformers: a primer, 2021.
http://www.columbia.edu/ jsl2239/transformers.html.

Bibliography III

Qiang Liu.
Stein variational gradient descent as gradient flow.
In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida.
Spectral normalization for generative adversarial networks.
02 2018.

Bernt Øksendal.
Stochastic Differential Equations: An Introduction with Applications (Universitext).
Springer, 6th edition, January 2014.

Bibliography IV

Danilo Jimenez Rezende and Shakir Mohamed.
Variational inference with normalizing flows.
In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1530–1538. JMLR.org, 2015.

Filippo Santambrogio.
Functionals on the space of probabilities, pages 249–284.
Springer International Publishing, Cham, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole.
Score-based generative modeling through stochastic differential equations, 11 2020.

Bibliography V

Mohammed Terry-Jack.
Deep learning: The transformer, 2019.
https://medium.com/@b.terryjack/deep-learning-the-transformer-9ae5e9c5a190.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin.
Attention is all you need.
In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.