Flow модели

Денис Ракитин

Содержание

1 Пререквизиты

Модели, о которых идет речь в проекте, основаны на представлении порождающего процесса в виде обыкновенного (ODE)

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t$$

или стохастического (SDE) дифференциального уравнения

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + G(X_t, t)\mathrm{d}W_t.$$

Для начала вспомним, что представляют собой эти уравнения по смыслу.

1.1 ODE

Обыкновенное дифференциальное уравнение

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t$$

очень удобно трактовать как описание траектории частицы, движущейся в пространстве. В такой интерпретации t означает время, X_t — позицию частицы в момент времени t, а $f(X_t, t)$ — ее вектор скорости в момент времени t (физический смысл производной). Особенно хорошо это видно, если записать схему Эйлера для приближенного решения:

$$X_{t+h} = X_t + h \cdot f(X_t, t) + \overline{o}(h) \approx X_t + h \cdot f(X_t, t).$$

Схема Эйлера говорит, что чтобы получить позицию частицы через небольшой момент времени h нужно сдвинуть ее по направлению вектора скорости на длину, пропорциональную прошедшему времени. Подобное представление позволит нам разобраться с тем, что такое SDE, не прибегая к тяжелой теории.

Для того, чтобы описать траекторию движения частицы, недостаточно просто определить ее скорость в каждый момент времени, нужно еще знать, откуда частица начинает двигаться. Задачей Коши (initial value problem) называют систему

$$\begin{cases} \mathrm{d}X_t = f(X_t, t)\mathrm{d}t\,; \\ X_0 = z. \end{cases}$$

Добавив начальное условие $X_0 = z$, говорящее о том, что частица начинает движение из точки z, мы полностью определили физическую систему, и, значит, определили (хотелось бы) единственную траекторию, которая задается с помощью данной задачи Коши. Теория говорит, что это действительно так, и гарантирует существование и единственность решения задачи Коши на некотором отрезке времени $[0, T]$ при некоторых условиях на скорость $f(X_t, t)$. Мы же про это не думаем и считаем, что решение всегда существует.

1.2 SDE

Со стохастическими дифференциальными уравнениями все устроено сильно сложнее несмотря на то, что идейно этот объект описать несложно. Говоря про ODE, мы трактовали ее решение, как траекторию частицы, движущейся с некоторым вектором скорости. Но что, если движение частицы не может быть описано с помощью детерминированной динамики? Такое, например, случается в квантовой физике, в которой частицы зачастую действительно ведут себя случайно. В таком случае к детерминированному вектору скорости хотелось бы добавить некоторую стохастическую часть, и получить уравнение вида

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + \text{случайность}.$$

Так, в качестве «чистой» случайности (а-ля броуновское движение) хотелось бы добавить что-то, соответствующее случайному блужданию. Раз мы описываем все в терминах дифференциальных уравнений, это соответствовало бы тому, чтобы к вектору скорости добавить так называемый белый шум: случайный процесс Z_t, в котором каждая величина Z_t имеет нормальное распределение $\mathcal{N}(0, I)$, и все величины между собой независимы. С таким определением, правда, возникают проблемы из-за того, что белый шум — очень плохой и не регулярный объект, который нам хотелось бы, например, проинтегрировать, чтобы получить траекторию частицы, но с интегрируемостью у него проблемы. Поэтому заходят с другой стороны и определяют для начала случайное блуждание с непрерывным временем, которое называют Винеровским процессом или процессом броуновского движения.

d-мерный винеровский процесс W_t — случайный процесс, обладающий тремя важными свойствами:

1. $W_0 = 0$;
2. Приращения независимы: для любых моментов времени $t_1 < t_2 < \ldots < t_n$ величины $W_{t_1}, W_{t_2} - W_{t_1}, \ldots, W_{t_n} - W_{t_{n-1}}$ независимы;
3. Приращения $W_t - W_s$ имеют распределение $\mathcal{N}(0, (t-s)I)$ для $t > s$.

Второе свойство объясняет использование винеровского процесса в качестве модели случайного блуждания: следующий шаг не зависит от предыдущих шагов, а третье свойство задает некоторый классический способ определить тип шума — нормальное распределение и его магнитуду, имеющую порядок $\sqrt{t}$ за время t (так как $\mathcal{N}(0, (t-s)I)$ совпадает по распределению с $\sqrt{t-s} \cdot \mathcal{N}(0, I)$).

Вооружившись винеровским процессом, можно добавить ее производную к правой части уравнения, чтобы получить ту самую «скорость», двигающую частицу в случайном направлении. Проблема тут в том, что винеровский процесс не дифференцируем

практически ни в каком разумном смысле (нужно заходить на обобщенные функции, чтобы это сделать), поэтому дальше идет построение интеграла Ито и формальное определение SDE с его помощью. Мы же все это опустим и приведем интуицию.

Стохастическое дифференциальное уравнение имеет вид

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + G(X_t, t)\mathrm{d}W_t.$$

Как это понимать? Как было заявлено выше, это можно понять с помощью дискретизации процесса по схеме Эйлера:

$$X_{t+h} \approx X_t + h \cdot f(X_t, t) + G(X_t, t) \cdot (W_{t+h} - W_t) =$$

$$= X_t + h \cdot f(X_t, t) + G(X_t, t) \cdot \mathcal{N}(0, hI)$$

$$= X_t + h \cdot f(X_t, t) + \sqrt{h} \cdot G(X_t, t) \cdot \mathcal{N}(0, I).$$

Запись, в которой какие-то величины умножаются на $\mathcal{N}(0, I)$, конечно, неформальная, потому что непонятно, как эта нормальная величина связана с X_t. Но если вспомнить, что приращения винеровского процесса независимы и раскрутить X_t рекурсивно, то получится, что X_t зависит только от X_0 и приращений винеровского процесса вида $W_{h \cdot k} - W_{h \cdot (k-1)}$, с которыми новое приращение $W_{t+h} - W_t$ независимо. Таким образом, мы сдвигаемся по направлению вектора скорости, как и раньше, но корректируем движение, добавляя независимый нормальный шум, магнитуда которого зависит как корень от времени, умноженный на так называемый коэффициент диффузии $G(X_t, t)$, являющийся в многомерном случае матрицей.

По аналогии с ODE, для SDE определяют задачу Коши вида

$$\begin{cases} \mathrm{d}X_t = f(X_t, t)\mathrm{d}t + G(X_t, t)\mathrm{d}W_t; \\ X_0 = Z, \end{cases}$$

где Z теперь может быть случайной величиной. Для задачи Коши с SDE, опять же, есть теоремы существования и единственности, которые для нас всегда работают.

Для простоты, дальше мы будем работать с SDE вида

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + g(t)\mathrm{d}W_t,$$

в котором коэффициент диффузии $g(t)$ является скаляром и не зависит от координаты.

1.3 Как генерировать?

Мы довольно абстрактно поговорили об ODE и SDE, но как их использовать в контексте генеративных моделей? Очень просто: генерируем случайную величину Z из какого-то фиксированного распределения (например, стандартного нормального) и решаем задачу Коши

$$\begin{cases} \mathrm{d}X_t = f(X_t, t)\mathrm{d}t + g(t)\mathrm{d}W_t; \\ X_0 = Z \end{cases}$$

вплоть до некоторого фиксированного момента времени (например, до $T = 1$). Можно (и нам это дальше понадобится в парных задачах) начинать динамику не только со случайного шума, но и, например, с картинки, которую мы хотим каким-то образом изменить, сохранив часть исходных данных.

Обучать мы будем функцию f, отвечающую за вектор скорости частицы. Обучать ее можно несколькими способами. Классический способ, предложенный, например, в статье [4], говорит, что нужно максимизировать правдоподобие семплов из датасета с точки зрения плотности случайной величины X_1 (которая, конечно, зависит от f). Подход [7], с которым будем работать мы, предлагает рассмотреть произвольную условную динамику, переводящую шум в фиксированную точку из датасета и восстановить из нее безусловную динамику, переводящую шум в плотность распределения данных. Оба подхода очень сильно завязаны на том, как плотность распределения частицы меняется, когда случайно сгенерированная частица в момент времени $t = 0$ начинает двигаться под действием динамики, заданной ODE или SDE.

1.4 Эволюция плотности распределения

В этой секции мы попробуем на интуитивном физическом уровне вывести уравнение непрерывности и уравнение Фоккера-Планка(-Колмогорова), отвечающие на вопрос о том, как эволюционирует плотность распределения решения ODE или SDE.

Пусть $Z \sim p_0$, а X_t — решение задачи Коши

$$\begin{cases} \mathrm{d}X_t = f(X_t, t)\mathrm{d}t + g(t)\mathrm{d}W_t; \\ X_0 = Z \end{cases}.$$

Обозначим плотность величины X_t за p_t и попробуем разобраться как меняется плотность с течением времени. Вопрос о том, как какая-то величина меняется, эквивалентен вопросу о том, как устроена производная этой величины. Поэтому нашей целью будет понять, как при небольшом приращении по времени меняется плотность p_t.

Вспомним схему Эйлера для SDE:

$$X_{t+h} \approx X_t + h \cdot f(X_t, t) + \sqrt{h} \cdot g(t) \cdot \xi,$$

где $\xi \sim \mathcal{N}(0, I)$ — независимая от X_t величина. Чтобы понять, как изменилась плотность при переходе от t к $t+h$, нам достаточно посчитать плотность X_{t+h}. Мы сделаем это, конечно, приближенно (уже используем знак $\approx$), но все, что мы не учли, будет $\overline{o}(h)$ и не повлияет на результат.

Обозначим $Y_t = X_t + h \cdot f(X_t, t)$ и $Z_t = \sqrt{h} \cdot g(t) \cdot \xi$. В таком случае $X_{t+h} = Y_t + Z_t$, и Y_t с Z_t независимы. Плотность суммы независимых величин легко посчитать по формуле свертки. А при переходе от X_t к Y_t производится биективное дифференцируемое преобразование, для которого есть формула замены переменных. Таким образом, все сводится к этим двум шагам.

1.4.1 Уравнение непрерывности

Начинаем с плотности $Y_t = X_t + h \cdot f(X_t, t)$. Представим, что в стохастической части $g(t) = 0$. В таком случае, мы фактически работаем с ODE и из двух шагов остается только детерминированный.

Обозначим $\varphi(x) = x + h \cdot f(x, t)$. Тогда $Y_t = \varphi(X_t)$. Вспомним, как выглядит формула замены переменной:

$$p_{\varphi(X)}(y) = p_X\left(\varphi^{-1}(y)\right) \left|\det \frac{\partial \varphi^{-1}}{\partial y}\right|.$$

В обратную сторону:

$$p_X(x) = p_{\varphi(X)}\left(\varphi(x)\right) \left|\det \frac{\partial \varphi}{\partial x}\right|.$$

Здесь нам будет удобно пользоваться вторым вариантом:

$$p_t(x) = p_{X_t}(x) = p_{\varphi(X_t)}(\varphi(x)) \left|\det \frac{\partial \varphi}{\partial x}\right| = p_{X_{t+h}}(\varphi(x)) \left|\det \frac{\partial \varphi}{\partial x}\right| = p_{t+h}(\varphi(x)) \left|\det \frac{\partial \varphi}{\partial x}\right|.$$

Отлично, осталось только подставить φ и пораскрывать:

$$p_t(x) = p_{t+h}\left(x + h \cdot f(x, t)\right) \left|\det\left(I + h \cdot \frac{\partial f(x, t)}{\partial x}\right)\right|.$$

Сначала разберемся с определителем. Вспомним, что для подсчета определителя матрицы A нужно рассмотреть все перестановки σ, взять произведение элементов

$\prod_{i=1}^{d} a_{i,\sigma(i)}$ и просуммировать, домножив на знак перестановки. Заметим, что если хотя бы один элемент перестановки оказался не на диагонали, то окажется еще один такой, и в произведении будут два множителя вида $h \cdot \frac{\partial f_i(x,t)}{\partial x_j}$, что будет $\overline{o}(h)$ и в пределе не дает вклада. Значит, важной здесь остается только диагональ, равная

$$\prod_{i=1}^{d} \left(1 + h \cdot \frac{\partial f_i(x,t)}{\partial x_i}\right) = 1 + h \cdot \sum_{i=1}^{d} \frac{\partial f_i(x,t)}{\partial x_i} + \overline{o}(h) = 1 + h \cdot \mathrm{div} f(x,t) + \overline{o}(h),$$

где $\mathrm{div} g = \sum_{i=1}^{d} \frac{\partial g_i}{\partial x_i}$ — оператор дивергенции, часто встречающийся при анализе векторных полей. Заметим также, что при малых h получается заведомо положительное выражение, поэтому модуль можно опустить.

Теперь к оставшему множителю, равному $p_{t+h}(x + h \cdot f(x,t))$. С ним все просто: раскладываем по Тейлору до линейной части, остальное будет $\overline{o}(h)$:

$$p_{t+h}(x + h \cdot f(x,t)) = p_{t+h}(x) + h \cdot \left\langle \frac{\partial p_{t+h}(x)}{\partial x},\ f(x,t) \right\rangle + \overline{o}(h \cdot f(x,t)) =$$

$$p_{t+h}(x) + h \cdot \left\langle \frac{\partial p_{t+h}(x)}{\partial x},\ f(x,t) \right\rangle + \overline{o}(h).$$

Собираем вместе:

$$p_t(x) = \left(p_{t+h}(x) + h \cdot \left\langle \frac{\partial p_{t+h}(x)}{\partial x},\ f(x,t) \right\rangle + \overline{o}(h)\right)\left(1 + h \cdot \mathrm{div} f(x,t) + \overline{o}(h)\right) =$$

$$= p_{t+h}(x) + h \cdot \left\langle \frac{\partial p_{t+h}(x)}{\partial x},\ f(x,t) \right\rangle + h \cdot p_{t+h}(x) \cdot \mathrm{div} f(x,t) + \overline{o}(h) =$$

$$= p_{t+h}(x) + h \cdot \sum_{i=1}^{d} \frac{\partial p_{t+h}(x)}{\partial x_i} \cdot f_i(x,t) + h \cdot \sum_{i=1}^{d} p_{t+h}(x) \cdot \frac{\partial f_i(x,t)}{\partial x_i} + \overline{o}(h) =$$

$$= p_{t+h}(x) + h \cdot \sum_{i=1}^{d} \frac{\partial}{\partial x_i}\left(p_{t+h}(x) \cdot f_i(x,t)\right) + \overline{o}(h) =$$

$$= p_{t+h}(x) + h \cdot \mathrm{div}\left(p_{t+h}(x) \cdot f(x,t)\right) + \overline{o}(h).$$

Приводим к виду из определения производной:

$$\frac{p_{t+h}(x) - p_t(x)}{h} = -\mathrm{div}(p_{t+h}(x) f(x,t)) + \overline{o}(1)$$

и переходим к пределу. Получаем так называемое *уравнение непрерывности*

$$\frac{\partial}{\partial t}p_t(x) = -\mathrm{div}\left(p_t(x)f(x,t)\right), \tag{1}$$

описывающее эволюцию плотности частицы, двигающейся под действием ODE.

1.4.2 Уравнение Фоккера-Планка-Колмогорова

Уравнение непрерывности интересно и является предметом исследования само по себе, но нам важно и его обобщение на SDE, называемое уравнением Фоккера-Планка-Колмогорова. Для этого нам нужно проделать второй шаг, анонсированный выше. Теперь

$$X_{t+h} = \varphi(X_t) + Z_t.$$

Для первой части $\varphi(X_t)$ мы уже вычислили плотность с точностью до о-малого:

$$p_t(x) = p_{\varphi(X_t)}(x) + h \cdot \mathrm{div}\left(p_{\varphi(X_t)}(x) \cdot f(x,t)\right) + \overline{o}(h),$$

или

$$p_{\varphi(X_t)}(x) = p_t(x) - h \cdot \mathrm{div}\left(p_{\varphi(X_t)}(x) \cdot f(x,t\right) + \overline{o}(h).$$

Для вычисления же плотности X_{t+h} нужно провести свертку $p_{\varphi(X_t)}(x)$ с плотностью $p_{Z_t}(x) = \mathcal{N}\left(x \mid 0, h \cdot g^2(t) \cdot I\right)$. Записываем:

$$p_{t+h}(x) = p_{X_{t+h}}(x) = \int p_{\varphi(X_t)}(y) \cdot p_{Z_t}(x-y)\mathrm{d}y.$$

Разложим плотность $\varphi(X_t)$ в точке x по Тейлору до второго порядка:

$$p_{t+h}(x) = \int \left(p_{\varphi(X_t)}(x) + (y-x)\frac{\partial}{\partial x}p_{\varphi(X_t)}(x) + \frac{1}{2}(y-x)^\top \frac{\partial^2}{\partial x^2}p_{\varphi(X_t)}(x)(y-x) + o(\|y-x\|^2)\right) \cdot$$

$$\cdot \mathcal{N}\left(y \mid x, h \cdot g^2(t) \cdot I\right)\, \mathrm{d}y.$$

Теперь же присматриваемся к этому интегралу и замечаем много приятных моментов:

- Первое слагаемое не зависит от y, а интеграл плотности равен единице, поэтому остается просто $p_{\varphi(X_t)}(x)$;
- Во втором слагаемом производная выносится за интеграл и остается матожидание $\mathbb{E}(\xi - x)$, для $\xi \sim \mathcal{N}\left(x, h \cdot g^2(t) \cdot I\right)$. Оно равно нулю, так как $\mathbb{E}\xi = x$.

- Последнее слагаемое является о-малым от $\|y - x\|^2$, а интеграл

$$\int \|y - x\|^2 \cdot \mathcal{N}(y|x, h \cdot g^2(t) \cdot I)\mathrm{d}y$$

совпадает с $\mathbb{E}\|\xi - x\|^2$ для $\xi \sim \mathcal{N}(x, h \cdot g^2(t) \cdot I)$. Тогда его можно выразить как $\mathbb{E}\|\xi - \mathbb{E}\xi\|^2 = \operatorname{Tr}\operatorname{Cov}\xi = h \cdot g^2(t) \cdot d$, где $\operatorname{Cov}\xi$ — ковариационная матрица ξ. Переставляя матожидание и о-малое (очень грубая и не всегда верная операция, но мы поверим), получаем $\overline{o}(h \cdot g^2(t) \cdot d) = \overline{o}(h)$, то есть это слагаемое не даст вклада.

Таким образом, у нас остается

$$p_{t+h}(x) = p_{\varphi(X_t)}(x) + \mathbb{E}\left[\frac{1}{2}(\xi - x)^\top \frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x)(\xi - x)\right] + \overline{o}(h),$$

где $\xi \sim \mathcal{N}(x, h \cdot g^2(t) \cdot I)$. Для простоты введем величину

$$\eta = \frac{1}{g(t)\sqrt{h}}(\xi - x) \sim \mathcal{N}(0, I)$$

и перепишем выражение как

$$p_{t+h}(x) = p_{\varphi(X_t)}(x) + \frac{g^2(t)}{2} \cdot h \cdot \mathbb{E}\left[\eta^\top \cdot \frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x) \cdot \eta\right] + \overline{o}(h).$$

Под матожиданием же стоит известная оценка следа Хатчинсона [6]. Воспользуясь циклическим свойством и линейностью следа, перепишем матожидание квадратичной формы:

$$\mathbb{E}\left[\eta^\top \cdot \frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x) \cdot \eta\right] = \mathbb{E}\operatorname{Tr}\left[\eta^\top \cdot \frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x) \cdot \eta\right] = \mathbb{E}\operatorname{Tr}\left[\frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x) \cdot \eta\eta^\top\right] =$$

$$= \operatorname{Tr}\mathbb{E}\left[\frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x) \cdot \eta\eta^\top\right] = \operatorname{Tr}\left[\frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x)\mathbb{E}\left[\eta\eta^\top\right]\right] = \operatorname{Tr}\left[\frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x) \cdot I\right] =$$

$$= \operatorname{Tr}\left[\frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x)\right] = \Delta p_{\varphi(X_t)}(x),$$

где $\Delta f(x) = \sum_{i=1}^{d} \frac{\partial^2 f}{\partial x_i^2}$ — оператор Лапласа. В итоге, мы имеем

$$p_{t+h}(x) = p_{\varphi(X_t)}(x) + \frac{g^2(t)}{2} \cdot h \cdot \Delta p_{\varphi(X_t)}(x) + \overline{o}(h).$$

Вспоминаем, что $p_{\varphi(X_t)}(x) = p_t(x) - h \cdot \operatorname{div}(p_{\varphi(X_t)}(x) \cdot f(x,t)) + \overline{o}(h)$ и получаем

$$p_{t+h}(x) = p_t(x) - h \cdot \operatorname{div}\left(p_{\varphi(X_t)}(x) \cdot f(x,t)\right) + h \cdot \frac{g^2(t)}{2} \cdot \Delta p_{\varphi(X_t)}(x) + \overline{o}(h).$$

Переносим $p_t(x)$ в левую часть и делим на h:

$$\frac{p_{t+h}(x) - p_t(x)}{h} = -\operatorname{div}\left(p_{\varphi(X_t)}(x) \cdot f(x,t)\right) + \frac{g^2(t)}{2} \cdot \Delta p_{\varphi(X_t)}(x) + \overline{o}(1).$$

Вспоминаем, что $\varphi(X_t) = X_t + h \cdot f(X_t, t) \xrightarrow[h \to 0]{} X_t$ и, тем самым, $p_{\varphi(X_t)}(x) \xrightarrow[h \to 0]{} p_{X_t}(x) = p_t(x)$.

Переходим к пределу и получаем *уравнение Фоккера-Планка-Колмогорова*

$$\frac{\partial}{\partial t} p_t(x) = -\operatorname{div}\left(p_t(x) \cdot f(x,t)\right) + \frac{g^2(t)}{2} \cdot \Delta p_t(x), \tag{2}$$

описывающее эволюцию плотности частицы, двигающейся под действием SDE.

1.5 Единственность

Итак, мы выяснили, что плотность частицы, двигающейся под действием ODE

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t,$$

удовлетворяет уравнению непрерывности

$$\frac{\partial}{\partial t} p_t(x) = -\operatorname{div}\left(p_t(x) \cdot f(x,t)\right),$$

а плотность частицы, двигающейся под действием SDE

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + g(t)\mathrm{d}W_t$$

удовлетворяет уравнению Фоккера-Планка-Колмогорова

$$\frac{\partial}{\partial t} p_t(x) = -\operatorname{div}\left(p_t(x) \cdot f(x,t)\right) + \frac{g^2(t)}{2} \Delta p_t(x),$$

которое обобщает уравнение непрерывности.

В дополнение к этому очень важное утверждение состоит в том, что (при некоторых условиях на f и g) существует только одно решение уравнения Фоккера-Планка-Колмогорова в паре с начальным условием

$$\begin{cases} \frac{\partial}{\partial t} p_t(x) = -\text{div}\left(p_t(x) \cdot f(x,t)\right) + \frac{g^2(t)}{2} \Delta p_t(x); \\ p_0(x) = p(x), \end{cases}$$

то есть не может случиться такого, что есть несколько последовательностей плотностей $p_t(x)$, удовлетворяющих одному и тому же уравнению Фоккера-Планка-Колмогорова (и уравнению непрерывности, в частности).

Благодаря этому свойству, мы можем говорить о том, что последовательность плотностей, удовлетворяющая уравнению Фоккера-Планка-Колмогорова с вектором сдвига f, коэффициентом диффузии g и начальным распределением p_0, *порождается* SDE $\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + g(t)\mathrm{d}W_t$ вместе с начальным условием $X_0 \sim p_0$, так как эти утверждения эквивалентны. Для ODE, соответственно, все то же самое.

1.6 Резюме

Итак, в нашем распоряжении есть два способа построить генеративную модель: на основе ODE или SDE, при этом первый является частным случаем второго. В общем случае, вооружившись SDE

$$\begin{cases} \mathrm{d}X_t = f(X_t, t)\mathrm{d}t + g(t)\mathrm{d}W_t; \\ X_0 = Z \sim p_0 \end{cases},$$

мы решаем его до момента времени $T = 1$ и говорим, что X_1 — искомый семпл из распределения. Остается только вопрос о том, как выучить сдвиг $f(x,t)$, чтобы X_1 имел распределение данных (раз уж мы решаем задачу генеративного моделирования). Об этом мы дальше и поговорим.

2 Генеративные модели на основе ODE

В данном разделе мы будем говорить о том, как обучать модели на основе ODE/SDE. Разговор будет преимущественно касаться ODE, потому что с ними сильно меньше нюансов и этого вполне хватит для проекта. Если будет время, будет добавлен материал про SDE.

2.1 Непрерывные нормализующие потоки

Это не особо нужная часть, TODO.

2.2 Flow Matching

2.2.1 Идея

Наиболее интересной для нас моделью будет Flow Matching [7]. На самом деле, мы будем пользоваться результатом не только этой работы, но и множества фоллоу-апов и параллельных статей: [15, 2, 1]. Здесь будет сделана попытка объединить наиболее важные для проекта наблюдения из этих статей.

Идею этого семейства методов легче всего описать на примере задачи генеративного моделирования и сравнивая с диффузионными моделями [5, 14]. Диффузионные модели берут картинку, рассматривают процесс ее постепенного зашумления, пытаются обучить модель так, чтобы этот процесс восстановить, и в частности восстановить обратный к нему процесс, генерирующий картинки из шума. В наиболее простой формулировке процесс (дискретный по времени) постепенного зашумления можно описать как

$$x_{t+1} = \alpha_t \cdot x_t + \beta_t \cdot \varepsilon_t; \;\; \varepsilon_t \sim \mathcal{N}(0, 1),$$

где от $\alpha_t < 1$ и β_t мы требуем, чтобы итоговый семпл x_T был по распределению близок к $\mathcal{N}(0, I)$. Таким образом, каждый шаг просто уничтожает часть информации с предыдущего и добавляет вместо этого шум.

Если присмотреться к этому процессу, то в нем нет практически ничего сложного: зная исходную картинку x_0, мы понимаем, что каждый следующий шаг — семплирование из некоторого нормального распределения, и в каждый момент времени x_t имеет нормальное распределение. Таким образом, при зафиксированном x_0 мы полностью знаем эволюцию плотности процесса в каждый момент времени. Вся сложность как раз-таки сидит в x_0: мы хотим этот процесс запускать в обратном направлении и картинку x_0 мы, конечно, на этапе генерации не знаем.

На основе нехитрого наблюдения о том, что вся сложность кроется только в распределении данных, авторы [7] предлагают рассмотреть для процесса генерации шум→данные условный процесс, обусловленный на картинку, описать его в терминах условного ODE, через него выразить безусловное ODE, порождающее безусловный процесс, и выписать функцию потерь для нахождения его векторного поля.

2.2.2 Реализация

Наиболее общий фреймворк Flow Matching представлен в статье [15], поэтому будем оперировать в ее терминах. Представим, что нам задана некоторая последовательность плотностей $p_t(x)$, для которой мы хотим восстановить векторное поле $f(x,t)$ такое, что

$$\frac{\partial}{\partial t}p_t(x) = -\text{div}\left(p_t(x)f(x,t)\right),$$

то есть, наша цель — породить $p_t(x)$ с помощью ODE

$$\mathrm{d}X_t = f(X_t,t)\mathrm{d}t.$$

В данном случае (вспоминаем мотивацию с обсулавливанием на картинку) мы предполагаем, что эту динамику можно представить в виде

$$p_t(x) = \int p_t(x|z)q(z)\mathrm{d}z.$$

В этом случае у нас появляется условная динамика $p_t(x|z)$ и маргинальное распределение условия $q(z)$, не зависящее от времени. Предположим, что условная динамика $p_t(x|z)$ может быть задана с помощью ODE

$$\mathrm{d}X_t = f(X_t,t|z)\mathrm{d}t,$$

то есть, условная динамика удовлетворяет уравнению непрерывности

$$\frac{\partial}{\partial t}p_t(x|z) = -\text{div}\left(p_t(x|z)f(x,t|z)\right).$$

Тогда утверждается, что можно аналитически выразить безусловное векторное поле $f(x,t)$ через условное $f(x,t|z)$:

$$f(x,t) = \int f(x,t|z)\frac{p_t(x|z)q(z)}{p_t(x)}\mathrm{d}z = \int f(x,t|z)p_t(z|x)\mathrm{d}z,$$

где

$$p_t(z|x) = \frac{p_t(x|z)q(z)}{p_t(x)} = \frac{p_t(x|z)q(z)}{\int p_t(x|z)q(z)\mathrm{d}z}$$

является условным распределением z при условии x в совместной модели $z \sim q(z)$, $x \sim p_t(x|z)$. Покажем, что это действительно так: мы хотим, чтобы векторное поле

$$\int f(x,t|z)\frac{p_t(x|z)q(z)}{p_t(x)}\mathrm{d}z$$

в паре с $p_t(x)$ удовлетворяло уравнению непрерывности. Распишем производную по времени:

$$\frac{\partial}{\partial t}p_t(x) = \frac{\partial}{\partial t}\int p_t(x|z)q(z)\mathrm{d}z = \int \left(\frac{\partial}{\partial t}p_t(x|z)\right) q(z)\mathrm{d}z.$$

Для условной динамики $p_t(x|z)$ есть уравнение непрерывности, из которого можно выразить производную по времени в скобках:

$$\int \left(\frac{\partial}{\partial t}p_t(x|z)\right) q(z)\mathrm{d}z = \int -\mathrm{div}\left(p_t(x|z)f(x,t|z)\right) q(z)\mathrm{d}z.$$

Дивергенция содержит в себе только производные, связанные с x, поэтому можно поменять ее с интегралом и получить

$$-\mathrm{div}\int p_t(x|z)f(x,t|z)q(z)\mathrm{d}z.$$

Наконец, выражение под дивергенцией можно домножить и разделить на $p_t(x)$ и получить

$$\frac{\partial}{\partial t}p_t(x) = -\mathrm{div}\left(p_t(x)\cdot\int f(x,t|z)\frac{p_t(x|z)q(z)}{p_t(x)}\mathrm{d}z\right).$$

Таким образом, динамика $p_t(x)$ действительно удовлетворяет уравнению непрерывности с векторным полем

$$f(x,t) = \int f(x,t|z)\frac{p_t(x|z)q(z)}{p_t(x)}\mathrm{d}z.$$

Как мы выяснили в Главе 1.5, это означает, что безусловная плотность $p_t(x)$ порождается с помощью решения ODE

$$\begin{cases}\mathrm{d}X_t = f(X_t,t)\mathrm{d}t;\\ X_0 \sim p_0.\end{cases}$$

Таким образом, для генерации динамики с помощью ODE остается только восстановить векторное поле f. Чтобы его восстановить, можно было бы решить задачу оптимизации вида

$$\int_0^1 \mathbb{E}_{p_t(x)}\|f_\theta(x,t) - f(x,t)\|^2\mathrm{d}t \to \min_\theta.$$

Но вот незадача: f мы не знаем и не можем оптимизировать такой функционал на практике. Оказывается, что по аналогии с denoising score matching функционалом, можно

свести все к регрессии на условное векторное поле вместо безусловного. Для этого начнем с того, что раскроем квадрат в функционале:

$$\int_0^1 \mathbb{E}_{p_t(x)} \|f_\theta(x,t) - f(x,t)\|^2 \mathrm{d}t = \int_0^1 \mathbb{E}_{p_t(x)} \left[\|f_\theta(x,t)\|^2 - 2 \left\langle f_\theta(x,t),\ f(x,t) \right\rangle \right] \mathrm{d}t + \text{const}.$$

Последнее слагаемое на оптимизацию не влияет, и его можно отбросить. Теперь воспользуемся представлением $f(x,t)$ через матожидание условного векторного поля

$$f(x,t) = \int f(x,t|z) p_t(z|x) \mathrm{d}z = \int f(x,t|z) \frac{p_t(x|z) q(z)}{p_t(x)} \mathrm{d}z$$

и подставим это в функционал. Получаем

$$\int_0^1 \mathbb{E}_{p_t(x)} \left[\|f_\theta(x,t)\|^2 - 2 \left\langle f_\theta(x,t),\ \int f(x,t|z) p_t(z|x) \mathrm{d}z \right\rangle \right] \mathrm{d}t.$$

В силу линейности, внутренний интеграл по плотности $p_t(z|x)$ можно переставить со скалярным произведением, а потом и вовсе внесит в него не зависящее от z первое слагаемое. Получим

$$\int_0^1 \mathbb{E}_{p_t(x)} \left(\int \left[\|f_\theta(x,t)\|^2 - 2 \left\langle f_\theta(x,t),\ f(x,t|z) \right\rangle \right] p_t(z|x) \mathrm{d}z \right) \mathrm{d}t.$$

Внутренний интеграл же можно «присоединить» к матожиданию, коим он и является. Тогда по-другому можно записать функционал как

$$\int_0^1 \mathbb{E}_{p_t(x) p_t(z|x)} \left[\|f_\theta(x,t)\|^2 - 2 \left\langle f_\theta(x,t),\ f(x,t|z) \right\rangle \right] \mathrm{d}t =$$

$$= \int_0^1 \mathbb{E}_{q(z) p_t(x|z)} \left[\|f_\theta(x,t)\|^2 - 2 \left\langle f_\theta(x,t),\ f(x,t|z) \right\rangle \right] \mathrm{d}t,$$

так как $p_t(z|x)$ — условное распределение в модели $q(z)p_t(x|z)$. Теперь же видно, что это практически среднеквадратичная ошибка, в которой не хватает только нормы условного векторного поля. Оно не зависит от параметров и не влияет на оптимизацию,

поэтому можно смело вернуть его в функционал. Получим

$$\int_0^1 \mathbb{E}_{q(z)p_t(x|z)} \left[\|f_\theta(x,t)\|^2 - 2\left\langle f_\theta(x,t),\ f(x,t|z)\right\rangle + \|f(x,t|z)\|^2 \right] \mathrm{d}t + \text{const}.$$

Таким образом, изначальная не решаемая на практике из-за незнания f задача оптимизации

$$\int_0^1 \mathbb{E}_{p_t(x)} \|f_\theta(x,t) - f(x,t)\|^2 \mathrm{d}t \to \min_\theta$$

свелась к эквивалентной решаемой задаче

$$\int_0^1 \mathbb{E}_{q(z)p_t(x|z)} \|f_\theta(x,t) - f(x,t|z)\|^2 \mathrm{d}t \to \min_\theta,$$

минимум в которой (при условии, что через $f_\theta(x,t)$ можно задать произвольную функцию) достигается на истинном векторном поле $f(x,t)$. Решается такая задача на практике, разумеется, с помощью Монте-Карло и градиентного спуска:

$$\nabla_\theta \int_0^1 \mathbb{E}_{q(z)p_t(x|z)} \|f_\theta(x,t) - f(x,t|z)\|^2 \mathrm{d}t \approx \nabla_\theta \|f_\theta(x,t) - f(x,t|z)\|^2,$$

где $t \sim \mathcal{U}[0,1], z \sim q(z), x \sim p_t(x|z)$.

Получается интересная вещь: мы хотим найти безусловное векторное поле и решаем регрессию на условное, каждый раз обусловленное на разный семпл из маргинального распределения. Сравните этот результат с denoising score matching [13]: в нем для нахождения безусловной скор-функции решается регрессия на условную.

2.2.3 Применение к генеративному моделированию

Авторы оригинальной статьи [7] предлагают рассмотреть, как и в диффузионных моделях, эволюцию плотности в соответствии с процессом зашумления. А именно, авторы представляют этот процесс в обратном относительно диффузионных моделей порядке, считая x_1 картинкой из датасета, а x_0 шумом. Авторы предлагают взять в качестве переменной z картинку x_1, чье маргинальное распределение $q_1(x_1)$ берется как распределение датасета. Условная динамика определяется следующим образом:

$$p_t(x|x_1) = \mathcal{N}\left(x|\mu_t(x_1), \sigma_t^2(x_1)\cdot I\right),$$

где параметры в начальный и конечный моменты времени определяются так, чтобы при маргинализации получить распределение данных $q_1(x_1)$ в момент времени $t = 1$ и стандартное нормальное распределение в момент времени $t = 0$. Получаются следующие условия: $\mu_0(x_1) = 0, \mu_1(x_1) = x_1$ и $\sigma_0^2(x_1) = 1, \sigma_1^2(x_1) = 0$ (иногда $\sigma_1^2(x_1)$ выставляют в маленькое значение ε, чтобы даже условная динамика была невырожденной, но это не критично). Таким образом, мы определяем некоторую интерполяцию по времени между стандартным нормальным распределением и распределением, сосредоточенным в одной точке. Несложно проверить, что при соблюдении таких условий мы получим

$$p_0(x) = \int p_0(x|x_1)q_1(x_1)\mathrm{d}x_1 = \int \mathcal{N}(x|0, I)q_1(x_1)\mathrm{d}x_1 = \mathcal{N}(x|0, I);$$

$$p_1(x) = \int p_1(x|x_1)q_1(x_1)\mathrm{d}x_1 = \int \delta(x - x_1)q_1(x_1)\mathrm{d}x_1 = q_1(x),$$

то есть выученная динамика действительно будет в нулевой момент времени иметь стандартное нормальное распределение, а в единичный момент времени иметь распределение данных.

Осталось только понять, каким условным векторным полем порождается динамика $p_t(x|x_1)$. Для этого заметим, что такая динамика может быть порождена следующим процессом: генерируем x_0 из $\mathcal{N}(0, I)$, как и договаривались, а в произвольный момент времени t положим

$$x_t = \mu_t(x_1) + \sigma_t(x_1) \cdot x_0.$$

Легко убедиться в том, что x_t действительно имеет распределение $\mathcal{N}(\mu_t(x_1), \sigma_t^2(x_1)I)$. Производную же по времени такой динамики посчитать очень легко:

$$\frac{\mathrm{d}}{\mathrm{d}t}x_t = \mu_t'(x_1) + \sigma_t'(x_1) \cdot x_0.$$

Для того, чтобы формально привести производную к виду $f(x_t, t|x_1)$, можно выразить и подставить x_0:

$$f(x_t, t|x_1) = \frac{\mathrm{d}}{\mathrm{d}t}x_t = \mu_t'(x_1) + \frac{\sigma_t'(x_1)}{\sigma_t(x_1)}(x_t - \mu_t(x_1)).$$

Таким образом, у нас есть все необходимые ингредиенты для обучения безусловного векторного поля через функционал

$$\int_0^1 \mathbb{E}_{q_1(x_1)\mathcal{N}(x|\mu_t(x_1),\sigma_t(x_1)I)}\|f_\theta(x, t) - f(x, t|x_1)\|^2\mathrm{d}t \to \min_\theta.$$

На самом деле же не обязательно было выражать функцию f и достаточно было делать регрессию на изначально полученное выражение $\mu'_t(x_1) + \sigma_t(x_1) \cdot x_0$, так как в нашей модели x_0 и x_t биективно выражаются друг через друга. Получим

$$\int_0^1 \mathbb{E}_{\mathcal{N}(x_0|0,I)q_1(x_1)} \| f_\theta(x_t, t) - \left(\mu'_t(x_1) + \sigma'_t(x_1) \cdot x_0\right) \|^2 \mathrm{d}t \to \min_\theta,$$

где $x_t = \mu_t(x_1) + \sigma_t(x_1) \cdot x_0$. Единственная оставшаяся в модели степень свободы — выбор конкретных значений $\mu_t(x_1)$ и $\sigma_t(x_1)$. Авторы [7] пробуют брать динамику вероятностей из диффузионных моделей [14], но получают лучшие результаты с наиболее простым способом проинтерполировать шумовой семпл и картинку: взять их выпуклую комбинацию. То есть, условная динамика порождается процессом

$$x_t = t \cdot x_1 + (1 - t) \cdot x_0.$$

Здесь $\mu_t(x_1) = t \cdot x_1$ и $\sigma_t(x_1) = 1 - t$ и условное векторное будет равно просто-напросто

$$\mu'_t(x_1) + \sigma'_t(x_1) \cdot x_0 = x_1 - x_0.$$

В таком случае функционал превращается в

$$\int_0^1 \mathbb{E}_{\mathcal{N}(x_0|0,I)q_1(x_1)} \| f_\theta(x_t, t) - (x_1 - x_0) \|^2 \mathrm{d}t, \tag{3}$$

где $x_t = t \cdot x_1 + (1 - t) \cdot x_0$.

После обучения можно будет генерировать из модели, решая ODE

$$\begin{cases} \mathrm{d}X_t = f_\theta(X_t, t)\mathrm{d}t; \\ X_0 \sim \mathcal{N}(0, I). \end{cases}$$

Для такой генерации будет гарантировано, что в каждый момент времени t семпл X_t будет иметь такое же распределение, как линейная интерполяция случайного шума и случайной картинки из датасета (по построению модели flow matching).

2.2.4 Обуславливание на два концу

Присмотревшись к функционалу 3, можно заметить, что все, что мы делаем — генерируем независимую пару из случайного шума и датасета и регрессируем обучаемое

векторное поле $f_\theta(x_t, t)$ на разницу $x_1 - x_0$. Вообще говоря, в такой схеме в качестве q_0 совсем не обязательно должно быть стандартное нормальное распределение. Почему бы не взять, например, распределение еще одного датасета?

Авторы статьи Conditional Flow Matching [15] говорят, что можно обуславливаться не только на конечный семпл x_1, как в оригинальном Flow Matching, но и на начальный семпл x_0. Как обычно, нужно определить маргинальное распределение условия $q_{01}(x_0, x_1)$ и условную динамику $p_t(x|x_0, x_1)$. Совместное распределение на пару условий мы будем задавать так, чтобы $\int q_{01}(x_0, x_1)\mathrm{d}x_1 = q_0(x_0)$ и $\int q_{01}(x_0, x_1)\mathrm{d}x_0 = q_1(x_1)$, где $q_0(x_0)$ и $q_1(x_1)$ — распределения, которые мы хотим видеть как начальное и конечное распределение динамики. В частности, никто не мешает, как и раньше, взять независимую пару $q_{01}(x_0, x_1) = \mathcal{N}(x_0|0, I)q_1(x_1)$ и получить динамику из случайного шума в датасет q_1.

От условной динамики мы теперь хотим, чтобы $p_0(x|x_0, x_1)$ было вырожденным распределением в точке x_0, то есть $p_0(x|x_0, x_1) = \delta(x - x_0)$. Аналогично, $p_1(x|x_0, x_1) = \delta(x - x_1)$. Как и во Flow Matching, мы получим необходимые безусловные распределения:

$$p_0(x) = \int \delta(x - x_0)q_{01}(x_0, x_1)\mathrm{d}x_0\mathrm{d}x_1 = \int q_{01}(x, x_1)\mathrm{d}x_1 = q_0(x);$$

$$p_1(x) = \int \delta(x - x_1)q_{01}(x_0, x_1)\mathrm{d}x_0\mathrm{d}x_1 = \int q_{01}(x_0, x)\mathrm{d}x_0 = q_1(x).$$

Таким образом, $p_t(x|x_0, x_1)$ представляет собой некоторую интерполяцию между парой точек. Можно не заморачиваться и, как и во Flow Matching, рассмотреть динамику, порожденную простой линейной интерполяцией:

$$x_t = t \cdot x_1 + (1 - t) \cdot x_0.$$

Плотность в момент времени t тогда тоже будет дельта-функцией

$$p_t(x|x_0, x_1) = \delta\left(x - \big(t \cdot x_1 + (1 - t) \cdot x_0\big)\right),$$

а условное векторное поле тривиальным образом равно $x_1 - x_0$. Таким образом, решив задачу оптимизации

$$\int_0^1 \mathbb{E}_{q_{01}(x_0,x_1)p_t(x_t|x_0,x_1)}\|f_\theta(x_t, t) - (x_1 - x_0)\|^2\mathrm{d}t \to \min_\theta, \tag{4}$$

мы получим векторное поле f_θ, с помощью которого будем генерировать. Заметим, что функционал 3 является частным случаем полученного сейчас функционала. По

построению Flow Matching, решая ODE

$$\begin{cases} \mathrm{d}X_t = f_\theta(X_t, t)\mathrm{d}t; \\ X_0 \sim q_0, \end{cases}$$

в момент времени t мы получим семпл X_t, который будет иметь такое же распределение, как линейная интерполяция между двумя семплами x_0, x_1, сгенерированными из $q_{01}(x_0, x_1)$. В частности, в момент времени $t = 1$ мы получим семпл из $q_1(x_1)$.

Чем это хорошо? Чисто с теоретической точки зрения мы теперь можем взять два датасета и попробовать решить между ними задачу а-ля style transfer: ODE с нашим векторным полем позволяет превратить семпл из одного распределения (в том числе, задаваемого датасетом) в семпл из другого распределения. В чем же здесь проблема? У нас есть гарантии, что в начальный и конечный момент времени X_0 и X_1 из ODE действительно будут иметь распределения q_0 и q_1, соответственно.

К сожалению, никто не гарантирует, что между этими семплами вообще будет какая-то связь. То есть, вполне может получиться, что мы обучим векторное поле f_θ на паре датасетов с зимними и летними картинками, и из зимнего пейзажа в результате преобразования будет получаться летний пейзаж, не имеющий к нему никакого отношения. Именно поэтому для непарных задач вида style transfer методы напрямую неприменимы. Однако, можно немного лучше проанализировать полученный метод и понять, что применять его для парных задач все-таки можно.

2.2.5 Оптимальный транспорт

Ненамного отвлечемся от Flow Matching и поговорим об отвлеченной теме. Рассмотрим задачу style transfer. В чем она состоит? Неформально, в том, чтобы сохранить контент картинки и изменить ее стиль. Вопрос: есть ли способ сформулировать на математическом языке такую задачу?

Для начала, подумаем о том, как можно сформулировать, что такое стиль. Учитывая, что мы отождествляем наборы данных с вероятностными распределениями, так можно поступить и со стилем. Например, если мы работаем с пейзажами, стиль «летний» можно отождествить с распределением q_0, семплы из которого являются летними пейзажами, а стиль «зимний» отождествить с распределением q_1, семплы которого являются зимними пейзажами. Тогда требование о переносе стиля говорит о том, что семпл $x_0 \sim q_0$ должен под действием преобразования g превратиться в семпл $g(x_0) \sim q_1$. Это ровно то, что нам гарантирует Conditional Flow Matching, если под g подразумевать решение ODE вплоть до момента времени $t = 1$

Остается вопрос про то, какие ограничения нужно наложить на g, чтобы сохранить контент? Мы понимаем, что сохранение контента означает, что исходная и конечная

картинка должны быть в каком-то смысле похожи. Как мы можем мерить похожесть? На это в современном глубоком обучении огромное число ответов, но в наиболее общем понимании ответ в том, что нужно ввести некоторую функцию $c(x, y)$, которая будет в каком-то смысле считать расстояние между парой объектов. Наиболее простая такая функция — посчитать попиксельное L_2 расстояние $c(x, y) = \|x - y\|^2$. Можно придумывать более сложные функции, считающие расстояния, например, между эмбеддингами картинок, но простое L_2 является хорошим базовым примером.

Таким образом, помимо требования о том, что $g(x_0)$ должно иметь распределение q_1, мы еще понимаем, что расстояние $c(x_0, g(x_0))$ должно быть не слишком высоким. Ровно из этих соображений вводится задача оптимального транспорта, которая говорит о том, что отображение g между q_0 и q_1 нужно выбирать, минизируя среднее расстояние между семплом и его отображением:

$$\mathbb{E}_{q_0(x_0)} c\left(x_0, g(x_0)\right) \to \min_{g: g(x_0) \sim q_1}.$$

Метрика

$$\mathbb{E}_{q_0(x_0)} c\left(x_0, g(x_0)\right)$$

называется транспортной ценой отображения g и является хорошей метрикой сохранения контента.

2.2.6 Анализ транспортной цены Flow Matching

Теперь попробуем разобраться, можно ли каким-то образом оценить, насколько низкой получается транспортная цена между входом X_0 и выходом X_1 обученной модели Flow Matching. Для этого обратимся к статье Rectified Flows [9], которая смотрит на Conditional Flow Matching с линейной интерполяцией немного под другим углом. Нам понадобится несколько обозначений. Также, как это было ранее в Conditional Flow Matching, мы вводим произвольное распределение $q_{01}(x_0, x_1)$ и условную динамику $p_t(x|x_0, x_1)$. Условную динамику $p_t(x|x_0, x_1)$ мы задаем с помощью линейной интерполяции

$$x_t = t \cdot x_1 + (1 - t) \cdot x_0.$$

Таким образом, мы получаем случайный процесс x_t, условное распределение которого в момент t при фиксированных x_0, x_1 совпадает с $p_t(x|x_0, x_1) = \delta(x - (t \cdot x_1 + (1 - t) \cdot x_0))$, а маргинальное распределение в момент времени t совпадает с $p_t(x) = \int p_t(x|x_0, x_1) q_{01}(x_0, x_1) \mathrm{d}x_0 \mathrm{d}x_1$.

После обучения Conditional Flow Matching на функционал 4, мы (в теории) получаем такое векторное поле $f(x,t)$, что решение X_t ODE

$$\begin{cases} \mathrm{d}X_t = f(X_t, t)\mathrm{d}t; \\ X_0 \sim q_0 \end{cases}$$

в момент времени t будет иметь маргинальное распределение $p_t(x)$, то есть, распределение X_t в каждый момент времени t совпадает с распределением ранее заданного процесса x_t.

В данных обозначениях нашим отображением g из предыдущей секции является решение ODE в момент времени 1, которое мы обозначили за X_1. В контексте предыдущего разговора о сохранении контента, нам бы хотелось, чтобы транспортная цена между X_0 и X_1 была небольшой. Авторы статьи [9] показывают, что эта транспортная цена будет на самом деле не больше, чем у пары (x_0, x_1), сгенерированной из совместного распределения $q_{01}(x_0, x_1)$. Попробуем это показать.

Для начала, оговоримся что в этой секции мы не будем писать, по каким распределением берется матожидания. Сразу зафиксируем, что пара из исходного процесса (x_0, x_1) генерируется из совместного распределения q_{01}, величина x_t в момент времени t этого процесса получается с помощью линейной интерполяции $x_t = t \cdot x_1 + (1-t) \cdot x_0$. Помимо этого, $X_0 \sim q_0$ — начальная точка обученного ODE процесса, а $X_1 = X_0 + \int_0^1 f(X_t, t)\mathrm{d}t$ — решение соответствующего ODE. Тогда наша цель — показать, что транспортная цена не увеличивается при переходе от x_t к X_t:

$$\mathbb{E}\, c(X_0, X_1) \leq \mathbb{E}\, c(x_0, x_1).$$

Для этого вспомним, как выражается безусловное выучиваемое CFM векторное поле через условное:

$$f(x,t) = \int f(x, t|x_0, x_1) p_t(x_0, x_1|x)\mathrm{d}z.$$

При этом напомним, что

$$p_t(x_0, x_1|x) = \frac{q_{01}(x_0, x_1) p_t(x|x_0, x_1)}{p_t(x)}$$

является условным распределением в модели, соответствующей нашему процессу x_t. В более вероятностных терминах тогда это просто-напросто условное матожидание

$$f(x,t) = \mathbb{E}\left[f(x, t|x_0, x_1)|x_t = x\right] = \mathbb{E}\left[f(x_t, t|x_0, x_1)|x_t = x\right].$$

А вспомним, что процесс x_t у нас это простая линейная интерполяция, которую порождает условное векторное поле $x_1 - x_0$, мы получаем

$$f(x,t) = \mathbb{E}\left[x_1 - x_0 | x_t = x\right].$$

Вооружившись таким представлением, мы готовы доказывать неравенство. Начнем с того, что мы будем рассматривать функции расстояния $c(x,y)$, являющиеся выпуклыми функциями от разности аргументов: $c(x,y) = c(y-x)$. Под такое описание, в частности, подходит $c(x,y) = \|y-x\|^2$. Для них начнем раскручивать. Вспомним, что X_1 — решение ODE с начальным условием X_0, а значит, удовлетворяет равенству

$$X_1 = X_0 + \int\limits_0^1 f(X_t,t)\mathrm{d}t.$$

Тогда транспортную цену можно переписать как

$$\mathbb{E}\, c(X_1 - X_0) = \mathbb{E}\, c\left(\int\limits_0^1 f(X_t,t)\mathrm{d}t\right).$$

Интеграл по отрезку $[0,1]$ — то же самое, что матожидание по равномерному распределению, а значит, можно применить неравенство Йенсена, которое говорит, что для выпуклой функции φ верно $\varphi(\mathbb{E}\xi) \le \mathbb{E}\varphi(\xi)$. Тогда

$$\mathbb{E}\, c\left(\int\limits_0^1 f(X_t,t)\mathrm{d}t\right) \le \mathbb{E}\int\limits_0^1 c\left(f(X_t,t)\right)\mathrm{d}t = \int\limits_0^1 \mathbb{E}\, c\left(f(X_t,t)\right)\mathrm{d}t.$$

По построению мы знаем, что распределения X_t и x_t совпадают, а матожидание зависит только от распределения, поэтому

$$\int\limits_0^1 \mathbb{E}\, c\left(f(X_t,t)\right)\mathrm{d}t = \int\limits_0^1 \mathbb{E}\, c\left(f(x_t,t)\right)\mathrm{d}t.$$

Теперь же, когда мы перешли к процессу x_t, можно вспомнить, как выражалось безусловное векторное поле f:

$$f(x,t) = \mathbb{E}\left[x_1 - x_0 | x_t = x\right],$$

или при подставлении самого семпла x_t вместо x:

$$f(x_t, t) = \mathbb{E}\left[x_1 - x_0 | x_t\right].$$

Подставляем и получаем

$$\int_0^1 \mathbb{E}\left[c\left(\mathbb{E}\left[x_1 - x_0|x_t\right]\right)\right] \mathrm{d}t.$$

Теперь пользуемся неравенством Йенсена для условного матожидания и получаем

$$\int_0^1 \mathbb{E}\left[c\left(\mathbb{E}\left[x_1 - x_0|x_t\right]\right)\right] \mathrm{d}t \leq \int_0^1 \mathbb{E}\left[\mathbb{E}\left[c(x_1 - x_0)|x_t\right]\right] \mathrm{d}t,$$

а матожидание условного матожидания — это просто матожидание:

$$\int_0^1 \mathbb{E}\left[\mathbb{E}\left[c(x_1 - x_0)|x_t\right]\right] \mathrm{d}t = \int_0^1 \mathbb{E}\, c(x_1 - x_0)\mathrm{d}t = \mathbb{E}\, c(x_1 - x_0).$$

Таким образом, мы действительно доказали, что

$$\mathbb{E}\, c(X_1 - X_0) \leq \mathbb{E}\, c(x_1 - x_0),$$

что означает, что транспортная цена между входом и выходом Conditional Flow Matching не больше, чем между парами, которые мы подаем модели на обучении.

2.2.7 Применение к парным задачам

НАКОНЕЦ-ТО мы приходим к тому, чтобы применять все это дело к парным задачам. Для начала подведем итоги того, что мы имеем.

Conditional Flow Matching — это способ выучить ODE, переводящее распределение q_0 в распределение q_1 со следующими свойствами:

- Для заранее заданных совместного распределения на парах картинок $q_{01}(x_0, x_1)$ и условной динамики (интерполяции) $p_t(x|x_0, x_1)$ ODE восстанавливает безусловную динамику $p_t(x) = \int p_t(x|x_0, x_1)q_{01}(x_0, x_1)\mathrm{d}t$, то есть семпл X_t из выученного ODE имеет распределение $p_t(x)$;
- В частности, стартуя из $X_0 \sim q_0$, мы обязательно придем в $X_1 \sim q_1$;

- Для линейной интерполяции $p_t(x|x_0, x_1) = \delta(x - (t \cdot x_1 + (1-t) \cdot x_0))$ транспортная цена между X_0 и X_1 не больше, чем транспортная цена между $x_0, x_1 \sim q_{01}(x_0, x_1)$, причем это верно для всех транспортных цен вида $c(y - x)$, где c — выпуклая функция.

Что же это нам дает в применении к любым парным задачам? В практически любой разумной парной задаче (парный датасет для переноса стиля, датасет из пар вида картинка + испорченная картинка) транспортная цена между парами небольшая (так как испорченная картинка все равно будет близка к изначальной картинке и практически всегда ближе, чем к любой другой картинке), а значит, если мы будет подавать при обучении Conditional Flow Matching пары из такого датасета (и таким образом определим распределение $q_{01}(x_0, x_1)$), то после обучения мы будем переводить X_0 в X_1 так, что транспортная цена между ними будет не больше, чем было в парах из датасета! Таким образом, во всех парных задачах мы будем гарантированно переводить объект в пару, соответствующую именно ему, а не произвольному семплу из распределения q_1.

После всего этого остается только порадоваться прочтению 25 страниц и записать итоговый алгоритм обучения. На каждом шаге обучения мы (для простоты представим, что размер батча 1):

1. Берем пару (x_0, x_1) из нашего парного датасета, генерируем момент времени t и определяем $x_t = t \cdot x_1 + (1-t) \cdot x_0$;

2. Подаем в обучаемое векторное поле f_θ пару (x_t, t) в качестве аргумента;

3. Считаем лосс $\|f_\theta(x_t, t) - (x_1 - x_0)\|^2$ и делаем градиентный шаг по $\nabla_\theta \|f_\theta(x_t, t) - (x_1 - x_0)\|^2$.

На этапе тестирования мы генерируем к поданному изображению X_0 соответствующую пару X_1 посредством решения ODE $\mathrm{d}X_t = f_\theta(X_t, t)\mathrm{d}t$, начиная с точки X_0. На практике это можно делать, например, используя схему Эйлера

$$X_{t+h} = X_t + h \cdot f_\theta(X_t, t).$$

3 Генеративные модели на основе SDE

Следующий небольшой набор моделей будет представлять из себя, по сути, обобщение Flow Matching на стохастический случай. В разных статьях предлагают разные способы это делать: задать интерполянт между x_0 и x_1 с помощью некоторого стохастического моста или задавать интерполянт как $p_t(x_t|x_0, x_1) = \mathcal{N}(x_t|I_t(x_0, x_1), \gamma^2(t)I)$ и учить SDE, которое может задать соответствующую безусловную динамику.

Оба варианта, на самом деле, дают некоторые плюсы по сравнению с детерминированным случаем. Задавая динамику через SDE или добавляя нормальный шум, мы будем подавать некоторую стохастическую интерполяцию двух картинок, что будет добавлять дополнительную регуляризацию (для фиксированной пары картинок мы будем подавать каждый раз разную интерполяцию). Помимо этого, сама динамика становится проще: вместо распределения на некотором многообразии, мы получаем распределение с полноценной плотностью, с которым проще работать.

3.1 Манипуляции с SDE

Для того, чтобы пользоваться методами на основе SDE, нам понадобятся несколько манипуляций с SDE на основе уравнения Фоккера-Планка.

3.1.1 Изменение коэффициента диффузии

Первая вещь, впервые использованная в контексте диффузионных моделей в статье [14], состоит в том, что процесс, задаваемый с помощью SDE с одним коэффициентом диффузии, можно превратить в процесс, задаваемый SDE с другим коэффициентом диффузии так, чтобы они порождали одну и ту же динамику маргинальных распределений.

Допустим, прямой процесс задается с помощью SDE

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + g(t)\mathrm{d}W_t,$$

а начальная точка X_0 генерируется из распределения с плотностью p_0. Тогда плотность $p_t(x)$ процесса в момент времени t удовлетворяет уравнению Фоккера-Планка

$$\frac{\partial}{\partial t}p_t(x) = -\mathrm{div}\left(p_t(x)f(x,t)\right) + \frac{g^2(t)}{2}\Delta p_t(x).$$

Теперь представим, что нам с какого-то перепугу захотелось породить динамику $p_t(x)$ с помощью SDE с другим коэффициентом диффузии, скажем, $h(t)$. Мы знаем, как это сделать: нужно просто написать уравнение Фоккера-Планка на $p_t(x)$ так, чтобы коэффициент при втором слагаемом превратился в $h^2(t)/2$. Прибавим-вычтем:

$$\frac{\partial}{\partial t}p_t(x) = -\mathrm{div}\left(p_t(x)f(x,t)\right) + \frac{g^2(t)}{2}\Delta p_t(x) + \frac{h^2(t)}{2}\Delta p_t(x) - \frac{h^2(t)}{2}\Delta p_t(x) =$$

$$-\mathrm{div}\left(p_t(x)f(x,t)\right) + \frac{g^2(t) - h^2(t)}{2}\Delta p_t(x) + \frac{h^2(t)}{2}\Delta p_t(x).$$

Вспоминаем, что $\Delta = \text{div}\nabla$ и переписываем второе слагаемое:

$$-\text{div}\left(p_t(x)f(x,t)\right) + \frac{g^2(t) - h^2(t)}{2}\text{div}\nabla p_t(x) + \frac{h^2(t)}{2}\Delta p_t(x) =$$

$$= -\text{div}\left(p_t(x)f(x,t) + \frac{h^2(t) - g^2(t)}{2}\nabla p_t(x)\right) + \frac{h^2(t)}{2}p_t(x) =$$

$$= -\text{div}\left(p_t(x)f(x,t) + \frac{h^2(t) - g^2(t)}{2}p_t(x)\nabla \log p_t(x)\right) + \frac{h^2(t)}{2}p_t(x) =$$

$$= -\text{div}\left(p_t(x)\left[f(x,t) + \frac{h^2(t) - g^2(t)}{2}\nabla \log p_t(x)\right]\right) + \frac{h^2(t)}{2}p_t(x).$$

Таким образом, динамика, порождаемая изначальным SDE

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + g(t)\mathrm{d}W_t$$

с коэффициентом диффузии $g(t)$, задается также SDE

$$\mathrm{d}X_t = \left(f(X_t, t) + \frac{h^2(t) - g^2(t)}{2}\nabla \log p_t(x)\right)\mathrm{d}t + h(t)\mathrm{d}W_t.$$

В частности, очень важным частным случаем является $h(t) = 0$, то есть, случай, в котором SDE превращается в ODE

$$\mathrm{d}X_t = \left(f(X_t, t) - \frac{g^2(t)}{2}\nabla \log p_t(X_t)\right)\mathrm{d}t.$$

Здесь так же начинает во всей красе проявляться важность такой вещи, как скор-функция $\nabla \log p_t(x)$, без которой все эти манипуляции сделать не получается.

3.1.2 Обращение диффузионного процесса по времени

Помимо изменения коэффициента диффузии может быть очень полезно обращать SDE по времени. На этом, по сути, и работают диффузионные модели: определив прямой процесс постепенного зашумления картинки и обратив его, можно получить процесс постепенного расшумления картинки из $\mathcal{N}(0, I)$. В конце предыдущей части мы перешли от SDE к ODE с сохранением динамики $p_t(x)$. Замечательно в этом то, что,

в отличие от SDE, обращать ODE совсем просто: нужно просто взять минус вектор скорости вместо вектора скорости.

Действительно, если $p_t(x)$ — прямая динамика, то $q_t(x) = p_{1-t}(x)$ — обратная динамика. При этом,

$$\frac{\partial}{\partial t} q_t(x) = -\frac{\partial}{\partial t} p_{1-t}(x) = \operatorname{div}\left(p_{1-t}(x)\left(f(x, 1-t) - \frac{g^2(1-t)}{2}\nabla \log p_{1-t}(x)\right)\right) =$$

$$= -\operatorname{div}\left(q_t(x)\left(-f(x, 1-t) + \frac{g^2(1-t)}{2}\nabla \log p_{1-t}(x)\right)\right).$$

Таким образом, обратная динамика порождается с помощью ODE

$$\mathrm{d}Y_t = \left(-f(Y_t, 1-t) + \frac{g^2(1-t)}{2}\nabla \log p_{1-t}(Y_t)\right)\mathrm{d}t.$$

А здесь, в свою очередь, можно перейти обратно от ODE к SDE, проделав аналогичные предыдущей главе манипуляции (хорошее упражнение):

$$-\operatorname{div}\left(q_t(x)\left(-f(x, 1-t) + \frac{g^2(1-t)}{2}\nabla \log p_{1-t}(x)\right)\right) =$$

$$= -\operatorname{div}\Big(q_t(x)\big(-f(x, 1-t) + g^2(1-t)\nabla \log p_{1-t}(x)\big)\Big) + \frac{g^2(1-t)}{2}\Delta q_t(x).$$

Таким образом, обратная динамика порождается с помощью SDE

$$\mathrm{d}Y_t = \big(-f(Y_t, 1-t) + g^2(1-t)\nabla \log p_{1-t}(Y_t)\big) + g(1-t)\mathrm{d}W_t.$$

Отметим, что мы построили обратный процесс в следующем смысле: каждое маргинальное распределения обратного процесса в момент времени t совпадает с маргинальным распределением прямого процесса в момент времени $1-t$. На самом же деле, справедливо и гораздо более сильное утверждение: меры этих процессов целиком совпадают, то есть

$$\mathbb{P}((X_t, t \in [0,1]) \in B) = \mathbb{P}((Y_{1-t}, t \in [0,1]) \in B)$$

для произвольного множества $B \subset \mathcal{C}(0,1)$. Данный результат можно найти в оригинальной работе [3].

3.1.3 Обуславливание процесса на конечную точку

Если для непрерывной версии диффузионных моделей наиболее важны предыдущие рассуждения про обращение процесса по времени, то в нашем случае очень важно уметь обуславливать процесс на конечную точку, чтобы получать некоторую стхастическую интерполяцию между начальной и конечной точкой и перенести техники Flow Matching на SDE. Обращение же процесса по времени здесь будет очень полезным. Интуиция здесь такая: при рассмотрении прямого процесса конечная точка это результат целого прохода процесса вперед, а вот для обратного процесса это ни что иное, как начальное условие, ничего сложного собой не представляющее.

Итак, нам задан прямой процесс

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + g(t)\mathrm{d}W_t$$

с начальным условием $X_0 \sim p_0$ и маргинальным распределением p_t в каждый момент времени. Наша цель — определить условный процесс

$$\mathrm{d}X_t^{0,1} = f(X_t^{0,1}, t)\mathrm{d}t + g(t)\mathrm{d}W_t$$

такой, что его плотность в каждый момент времени t будет совпадать с условной плотностью исходного процесса $p_{t|1}(x|x_1)$ при зафиксированной конечной точке x_1. Для этого вспомним, что в нашем распоряжении есть обратный процесс с маргинальной плотностью $q_{1-t}(x) = p_t(x)$, для которого условная плотность с фиксированным начальным моментом времени $q_{1-t|0}(x|x_1)$ совпадает с $p_{t|1}(x|x_1)$. Величина x_1, подставляемая как условие в момент времени 0, конечно, может путать, поэтому главное ориентироваться на нижние индексы плотности.

Для сопоставления динамики некоторому SDE нужно подобрать соответствующее уравнение Фоккера-Планка для $p_{t|1}(x|x_1)$. Для этого перейдем к условной плотности обратного процесса $q_{t|0}(x|x_1)$. Обусловить диффузионный процесс на стартовую точку x_1 — все равно, что подменить начальное распределение с некоторого $q_0(x)$ на $\delta(x - x_1)$. Поэтому, динамика $q_{t|0}(x|x_1)$ все еще удовлетворяет тому же уравнению Фоккера-Планка, что и $q_t(x)$:

$$\frac{\partial}{\partial t} q_{t|0}(x|x_1) = -\mathrm{div}\left(q_{t|0}(x|x_1)\left(-f(x, 1-t) + g^2(1-t)\nabla \log q_t(x)\right)\right) + \frac{g^2(1-t)}{2}\Delta q_t(x).$$

А теперь применим все те же преобразования, позволяющие перейти от Фоккера-Планка для прямого процесса к Фоккеру-Планку для обратного процесса, только теперь считая $q_{t|0}(x|x_1)$ за распределение прямого процесса. Обратная к $q_{t|0}(x|x_1)$ динамика — ни что иное, как $p_{t|1}(x|x_1)$, так как $p_{t|1}(x|x_1) = q_{1-t|1}(x|x_1)$. Тогда мы получим

$$\frac{\partial}{\partial t}p_{t|1}(x|x_1) = -\text{div}\left(p_{t|1}(x|x_1)\left(f(x,t) - g^2(t)\nabla \log q_{1-t}(x) + g^2(t)\nabla \log p_{t|1}(x|x_1)\right)\right) + \\ + \frac{g^2(t)}{2}\Delta p_{t|1}(x|x_1).$$

Осталось только немного преобразовать первую скобку, чтобы увидеть там более простое выражение:

$$-\nabla \log q_{1-t}(x) + \nabla \log p_{t|1}(x|x_1) = -\nabla \log p_t(x) + \nabla \log p_{t|1}(x|x_1) =$$

$$= -\nabla \log p_t(x) + \nabla \log \frac{p_t(x)p_{1|t}(x_1|x)}{p_1(x_1)} = -\nabla \log p_t(x) + \nabla \log p_t(x) + \nabla \log p_{1|t}(x_1|x) =$$

$$= \nabla \log p_{1|t}(x_1|x).$$

Итого, мы получаем

$$\frac{\partial}{\partial t}p_{t|1}(x|x_1) = -\text{div}\left(p_{t|1}(x|x_1)\left(f(x,t) + g^2(t)\nabla \log p_{1|t}(x_1|x)\right)\right) + \frac{g^2(t)}{2}\Delta p_{t|1}(x|x_1),$$

а значит, условная динамика $p_{t|1}$ порождается с помощью SDE

$$\mathrm{d}X_t^{0,1} = \left(f(X_t^{0,1},t) + g^2(t)\nabla \log p_{1|t}(x_1|X_t^{0,1})\right)\mathrm{d}t + g(t)\mathrm{d}W_t.$$

Как и в предыдущем случае, здесь мы доказали только то, что распределение условного процесса в каждый момент времени совпадает с распределением $p_{t|1}$, но ничего не доказали про распределение всего процесса целиком. Но и тут оказывается, что распределение процесса $X_t^{0,1}$ совпадает с условным распределением процесса X_t при фиксированной конечной точке:

$$\mathbb{P}((X_t^{0,1}, t \in [0,1]) \in B) = \mathbb{P}((X_t, t \in [0,1]) \in B | X_1 = x_1)$$

для $B \subset \mathcal{C}(0,1)$. Условная плотность $p_{1|t}$ называется h-функцией Дуба, а сам метод носит имя h-преобразования Дуба. Более подробно о нем можно почитать, например, в [10].

3.2 Bridge Matching

Построив всю необходимую теорию, мы наконец-то возвращаемся к построению стохастических аналогов Flow Matching. Первый метод будет полностью основан на выведенном ранее преобразовании Дуба, которое позволяет привязать произвольный диффузионный процесс к конечной точке x_1, получив, таким образом, некоторую стохастическую интерполяцию между x_0 и x_1. Этот метод используется напрямую для парных задач в статьях [12, 8], а также является частью алгоритма в [11].

3.2.1 Броуновский мост.

Классика такого вида интерполяции — Броуновский мост. В достаточно общем случае выглядит он так: берем в качестве безусловного процесса

$$\mathrm{d}X_t = \sqrt{\beta_t}\mathrm{d}W_t.$$

В частности, сам процесс будет равен

$$X_t = X_0 + \int\limits_0^t \sqrt{\beta_s}\mathrm{d}W_s,$$

а если магнитуда шума β_t постоянна и равна ε, мы получим

$$X_t = X_0 + \int\limits_0^t \sqrt{\varepsilon}\mathrm{d}W_s = X_0 + \sqrt{\varepsilon}W_t,$$

то есть просто масштабированный винеровский процесс, стартующий из X_0. Из такого безусловного процесса мы получаем условный процесс, удовлетворяющий

$$\mathrm{d}X_t^{0,1} = \beta_t \nabla_{X_t^{0,1}} \log p_{1|t}(x_1|X_t^{0,1})\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}W_t.$$

В частности, при $\beta_t = \varepsilon$, получится

$$\mathrm{d}X_t^{0,1} = \varepsilon \nabla_{X_t^{0,1}} \log p_{1|t}(x_1|X_t^{0,1})\mathrm{d}t + \sqrt{\varepsilon}\mathrm{d}W_t.$$

Остается только рассчитать условную плотность изначального процесса.

Постоянная магнитуда шума. Сделаем это отдельно для $\beta_t = \varepsilon$, чтобы показать на более наглядном примере. В этом случае справедливо

$$X_1 - X_t = X_0 + \sqrt{\varepsilon}W_1 - (X_0 + \sqrt{\varepsilon}W_t) = \sqrt{\varepsilon}(W_1 - W_t).$$

Тогда X_1 выражается через X_t как

$$X_1 = X_t + \sqrt{\varepsilon}(W_1 - W_t).$$

Учитывая, что сам X_t зависит только от $(W_s, s \leq t)$, получается, что $W_1 - W_t$ независимо с X_t, при этом $\sqrt{\varepsilon}(W_1 - W_t) \sim \mathcal{N}(0, \varepsilon(1-t))$. Тогда условное распределение равно

$$p_{1|t}(x_1|x_t) = \mathcal{N}(x_1|x_t, \varepsilon(1-t)I) = \mathrm{const} \cdot \exp\left(-\frac{1}{2\varepsilon(1-t)}\|x_1 - x_t\|^2\right).$$

Берем логарифм, дифференцируем по x_t и получаем

$$\nabla_{x_t} \log p_{1|t}(x_1|x_t) = \frac{x_1 - x_t}{\varepsilon(1-t)}.$$

Подставляем в формулу из преобразования Дуба и получаем условный процесс

$$\mathrm{d}X_t^{0,1} = \varepsilon \frac{x_1 - X_t^{0,1}}{\varepsilon(1-t)}\mathrm{d}t + \sqrt{\varepsilon}\mathrm{d}W_t = \frac{x_1 - X_t^{0,1}}{1-t}\mathrm{d}t + \sqrt{\varepsilon}\mathrm{d}W_t,$$

который называется броуновским мостом (с магнитудой ε).

Пара свойств интеграла Ито. Теперь переходим к произвольной магнитуде шума

$$\mathrm{d}X_t = \sqrt{\beta_t}\mathrm{d}W_t.$$

Как и раньше, выразим X_1 через X_t:

$$X_1 - X_t = X_0 + \int_0^1 \sqrt{\beta_s}\mathrm{d}W_s - \left(X_0 + \int_0^t \sqrt{\beta_s}\mathrm{d}W_t \right) = \int_t^1 \sqrt{\beta_s}\mathrm{d}W_s.$$

Здесь, в общем, попытки избежать стоханализ проваливаются, и нужно уже и определять интеграл Ито (который по $\mathrm{d}W_t$) и узнавать про какие-то его свойства. В нашем случае все довольно просто: интеграл от неслучайной функции определяется как предел частичных сумм

$$\int_a^b f(t)\mathrm{d}W_t = \lim_{\max\limits_{i=1\dots n}\left(t_{i+1}-t_i\right)\to 0} \sum_{i=1}^n f(t_i)(W_{t_{i+1}} - W_{t_i}).$$

Единственная тонкость состоит в том, что предел здесь понимается в смысле L^2, но это не сильно важно. А важно то, что любая частичная сумма имеет нормальное распределение: приращения $W_{t_{i+1}} - W_{t_i}$ независимы и имеют распределение $\mathcal{N}(0, t_{i+1} - t_i)$. Таким образом, перед нами взвешенная сумма нормальных случайных величин — тоже нормальная случайная величина. При этом,

$$\mathbb{E}\sum_{i=1}^n f(t_i)(W_{t_{i+1}} - W_{t_i}) = \sum_{i=1}^n f(t_i)\mathbb{E}(W_{t_{i+1}} - W_{t_i}) = 0,$$

$$\mathrm{Var}\sum_{i=1}^n f(t_i)(W_{t_{i+1}} - W_{t_i}) = \sum_{i=1}^n \mathrm{Var}\left(f(t_i)(W_{t_{i+1}} - W_{t_i})\right) = \sum_{i=1}^n f^2(t_i)\mathrm{Var}(W_{t_{i+1}} - W_{t_i}) =$$

$$= \sum_{i=1}^{n} f^2(t_i)(t_{i+1} - t_i).$$

Таким образом, мы получили

$$\sum_{i=1}^{n} f(t_i)(W_{t_{i+1}} - W_{t_i}) \sim \mathcal{N}\left(0, \sum_{i=1}^{n} f^2(t_i)(t_{i+1} - t_i)\right).$$

Куда это сойдется в пределе? Любая сходимость случайных величин сохраняет в пределе нормальное распределение, а параметры являются пределами параметров. Таким образом,

$$\int_a^b f(t)\mathrm{d}W_t \sim \mathcal{N}\left(0, \lim_{\max\limits_{i=1\dots n}(t_{i+1}-t_i)\to 0} \sum_{i=1}^{n} f^2(t_i)(t_{i+1} - t_i)\right).$$

Несложно заметить, что в дисперсии записана Римановская сумма, а значит, в пределе мы получим просто интеграл:

$$\int_a^b f(t)\mathrm{d}W_t \sim \mathcal{N}\left(0, \int_a^b f^2(t)\mathrm{d}t\right).$$

Таким образом, мы показали три интересных свойства. Во-первых, интеграл Ито от неслучайной функции — гауссовская случайная величина. Во-вторых, матожидание интеграла Ито равно нулю (и это верно и для случайных функций). В-третьих, мы посчитали дисперсию и получили частный случай так называемой *изометрии Ито*, которая говорит, что

$$\mathbb{E}\left[\int_a^b X_t\mathrm{d}W_t \cdot \int_a^b Y_t\mathrm{d}W_t\right] = \int_a^b \mathbb{E}\left[X_t Y_t\right]\mathrm{d}t.$$

Переменная магнитуда шума. Snap back to reality, возвращаемся к броуновскому мосту. Мы вывели, что для процесса

$$\mathrm{d}X_t = \sqrt{\beta_t}\mathrm{d}W_t$$

имеет место представление

$$X_1 = X_t + \int_t^1 \sqrt{\beta_s}\mathrm{d}W_s.$$

Как и в случае постоянной магнитуды шума, величина X_t зависит только от $(W_s, s \leq t)$, а интеграл $\int_t^1 f(s)\mathrm{d}W_s$ выражается как предел функции от приращений винеровского процесса после момента времени t, которые независимы с $(W_s, s \leq t)$. При этом, мы выяснили, что $\int_t^1 \sqrt{\beta_s}\mathrm{d}W_s$ имеет нормальное распределение $\mathcal{N}(0, \int_t^1 \beta_s \mathrm{d}s)$. Из этого мы получаем условное распределение:

$$p_{1|t}(x_1|x_t) = \mathcal{N}\left(x_1 \middle| x_t, \int_t^1 \beta_s \mathrm{d}s\right) = \text{const} \cdot \exp\left(-\frac{1}{2\int_t^1 \beta_s \mathrm{d}s}\|x_t - x_1\|^2\right).$$

Логарифмируем, дифференцируем и получаем

$$\nabla_{x_t} \log p_{1|t}(x_1|x_t) = \frac{x_1 - x_t}{\int_t^1 \beta_s \mathrm{d}s}.$$

Тогда условный процесс, задаваемый с помощью преобразования Дуба, задается

$$\mathrm{d}X_t^{0,1} = \beta_t \frac{x_1 - X_t^{0,1}}{\int_t^1 \beta_s \mathrm{d}s}\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}W_t.$$

3.2.2 Распределение броуновского моста

Напомним, что мосты мы строим для того, чтобы с их помощью делать интерполяцию между начальной и конечной точкой стохастическим образом. В будущем мы будем применять эту технику к алгоритму, очень похожему на Flow Matching, в которой мы на каждом шаге очень просто интерполировали две картинки линейной интерполяцией. Здесь же оказывается, что вместо простенькой суммы нужно решать SDE, чего делать не хотелось бы. На самом же деле, несмотря на кажущуюся сложность самого условного процесса, его распределение в каждый момент времени t все еще будет простым. Выведем его для броуновского моста с постоянным коэффициентом. Для более общего случая делается это похожим образом, но нужна формула Ито (аналог производной сложной функции для стохастических дифференциалов).

Итак, наш условный процесс имеет вид

$$\mathrm{d}X_t^{0,1} = \frac{x_1 - X_t^{0,1}}{1-t}\mathrm{d}t + \sqrt{\varepsilon}\mathrm{d}W_t.$$

Попробуем решить такое стохастическое дифференциальное уравнение. Для этого введем процесс $Y_t^{0,1} = a_t X_t^{0,1}$, где a_t — некоторая функция от времени. Нужно понять,

какой у $Y_t^{0,1}$ дифференциал. Несложно поверить в следующую формулу:

$$\mathrm{d}\left(a_t \cdot X_t^{0,1}\right) = X_t^{0,1} \cdot a_t' \mathrm{d}t + a_t \cdot \mathrm{d}X_t^{0,1},$$

которая просто соответствует производной произведения. Тем не менее, если функция будет зависеть еще и от $X_t^{0,1}$, так делать уже будет нельзя (см. формулу Ито). Перепишем дифференциал $Y_t^{0,1}$:

$$\mathrm{d}Y_t^{0,1} = a_t' \cdot X_t^{0,1} \cdot \mathrm{d}t + a_t \left(\frac{x_1 - X_t^{0,1}}{1-t}\mathrm{d}t + \sqrt{\varepsilon}\mathrm{d}W_t\right) = \left(a_t' \cdot X_t^{0,1} + \frac{a_t}{1-t}(x_1 - X_t^{0,1})\right)\mathrm{d}t + a_t\sqrt{\varepsilon}\mathrm{d}W_t =$$

$$= \left(a_t' - \frac{a_t}{1-t}\right) X_t^{0,1}\mathrm{d}t + \frac{a_t x_1}{1-t}\mathrm{d}t + a_t\sqrt{\varepsilon}\mathrm{d}W_t$$

Такое SDE легко решить, если в правой части ничего не зависит от $X_t^{0,1}$. Поэтому возьмем, да и подберем a_t' так, чтобы первая скобка обнулилась:

$$a_t' - \frac{a_t}{1-t} = 0 \iff a_t = \frac{a_0}{1-t},$$

в частности, $a_t = 1/(1-t)$ подойдет. Тогда мы получим

$$\mathrm{d}Y_t^{0,1} = \frac{x_1}{(1-t)^2}\mathrm{d}t + \frac{\sqrt{\varepsilon}}{1-t}\mathrm{d}W_t$$

и $Y_t^{0,1}$ явно выражается:

$$Y_t^{0,1} = Y_0^{0,1} + \int\limits_0^t \frac{x_1 \mathrm{d}s}{(1-s)^2} + \int\limits_0^t \frac{\sqrt{\varepsilon}}{1-s}\mathrm{d}W_s = Y_0^{0,1} + x_1 \cdot \frac{t}{1-t} + \int\limits_0^t \frac{\sqrt{\varepsilon}}{1-s}\mathrm{d}W_s.$$

Вспоминаем, что $Y_t^{0,1} = X_t^{0,1}/(1-t)$, а значит, $X_t^{0,1} = (1-t)Y_t$ и $Y_0^{0,1} = x_0$. Тогда

$$X_t^{0,1} = (1-t) \cdot x_0 + t \cdot x_1 + (1-t)\sqrt{\varepsilon} \int\limits_0^t \frac{1}{1-s}\mathrm{d}W_s.$$

Вспоминаем, что интеграл имеет нормальное распределение:

$$\int\limits_0^t \frac{1}{1-s}\mathrm{d}W_s \sim \mathcal{N}\left(0, \int\limits_0^t \frac{1}{(1-s)^2}\mathrm{d}s\right) = \mathcal{N}\left(0, \frac{t}{1-t}\right),$$

а значит,

$$(1-t)\sqrt{\varepsilon}\int\limits_0^t \frac{1}{1-s}\mathrm{d}W_s \sim \mathcal{N}\left(0, \frac{(1-t)^2\,\varepsilon\,t}{1-t}\right) = \mathcal{N}\left(0, \varepsilon\,t\,(1-t)\right).$$

Итого, мы получаем

$$X_t^{0,1} \sim (1-t)\cdot x_0 + t\cdot x_1 + \mathcal{N}(0, \varepsilon\cdot t\cdot(1-t)) = \mathcal{N}\left(t\cdot x_1 + (1-t)\cdot x_0,\ \ \varepsilon\cdot t\cdot(1-t)\right).$$

Таким образом, с точки зрения маргинальных распределений броуновский мост — ни что иное, как простейший линейный интерполянт $t\cdot x_1 + (1-t)\cdot x_0$, который разбавляется нормальным шумом с дисперсией вида параболы, зануляющейся на краях. Получается, что такой интерполянт соответствует интерполяции с постепенным зашумлением к середине и расшумлением ближе к концам. Это звучит логично с точки зрения перевода одного изображения в другое: разбавив изображение шумом, мы убьем часть его деталей, а менее детализированное изображение легче видоизменить так, чтобы превратить его в желаемый объект.

В случае с изменяющейся по времени магнитудой шума, то есть безусловного процесса

$$\mathrm{d}X_t = \sqrt{\beta_t}\mathrm{d}W_t$$

и условного процесса

$$\mathrm{d}X_t^{0,1} = \beta_t\frac{x_1 - X_t^{0,1}}{\int_t^1 \beta_s \mathrm{d}s}\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}W_t$$

мы получим нормальное распределение

$$X_t \sim \mathcal{N}\left(\mu_t, \gamma_t^2\cdot I\right),$$

с параметрами

$$\mu_t = \frac{\sigma_t^2}{\sigma_t^2+\overline{\sigma}_t^2}\cdot x_1 + \frac{\overline{\sigma}_t^2}{\sigma_t^2+\overline{\sigma}_t^2}\cdot x_0;$$

$$\gamma_t^2 = \frac{\sigma_t^2\overline{\sigma}_t^2}{\sigma_t^2+\overline{\sigma}_t^2},$$

где

$$\sigma_t^2 = \int\limits_0^t \beta_s\mathrm{d}s;\quad \overline{\sigma}_t^2 = \int\limits_t^1 \beta_s\mathrm{d}s.$$

Несложно заметить, что здесь тоже появляется некоторая монотонная интерполяция между x_0 и x_1 в матожидании и выпуклая дисперсия, зануляющаяся на концах. Расписание β_t становится важным гиперпараметром, настраивая который можно пытаться улучшить качество.

3.2.3 Обучение SDE на мостах

Мы наконец-то переходим к непосредственно обучению модели. Сначала зададим сеттинг: у нас есть, как и во Flow Matching, некоторое распределение $q_{01}(x_0, x_1)$ на парах и условная динамика $p_t(x_t|x_0, x_1)$, которая, на этот раз, задается с помощью SDE

$$\mathrm{d}X_t^{0,1} = f(X_t^{0,1}, t|x_0, x_1)\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}W_t.$$

Наша цель — выучить безусловное векторное поле f такое, что SDE

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}W_t$$

будет задавать безусловную динамику

$$p_t(x_t) = \int p_t(x_t|x_0, x_1)q_{01}(x_0, x_1)\mathrm{d}x_0\mathrm{d}x_1,$$

то есть динамику, соответствующую в момент времени t соединению x_0 с x_1 при помощи некоторого стохастического моста. Напомним, что в случае ODE было справедливо соотношение

$$f(x, t) = \mathbb{E}\left[f(x_t, t|x_0, x_1)|x_t = x\right] = \int f(x, t|x_0, x_1)\frac{p_t(x|x_0, x_1)q_{01}(x_0, x_1)}{p_t(x)}\mathrm{d}x_0\mathrm{d}x_1,$$

и доказывалось это тем, полученная по такой формуле f удовлетворяет уравнению непрерывности с $p_t(x)$. Оказывается, что точно такая же формула остается и в случае SDE и доказывается аналогичным образом. Начнем с того, что если динамика $p_t(x|x_0, x_1)$ задается с помощью выписанного выше SDE, то она задается уравнением Фоккера-Планка-Колмогорова

$$\frac{\partial}{\partial t}p_t(x|x_0, x_1) = -\mathrm{div}\left(p_t(x|x_0, x_1)f(x, t|x_0, x_1)\right) + \frac{\beta_t}{2}\Delta p_t(x).$$

Вспомним, как условная динамика выражается через безусловную и совместим это с Ф-П-К для условной динамики:

$$\frac{\partial}{\partial t}p_t(x) = \frac{\partial}{\partial t}\int p_t(x|x_0, x_1)q_{01}(x_0, x_1)\mathrm{d}t = \int \frac{\partial}{\partial t}p_t(x|x_0, x_1)q_{01}(x_0, x_1)\mathrm{d}x_0\mathrm{d}x_1 =$$

$$= \int \left(-\mathrm{div}\left(p_t(x|x_0, x_1)f(x, t|x_0, x_1)\right) + \frac{\beta_t}{2}\Delta p_t(x|x_0, x_1)\right) q_{01}(x_0, x_1)\mathrm{d}x_0\mathrm{d}x_1.$$

Раскрываем по линейности на два интеграла, в каждом вспоминаем, что div и Δ содержат только производные по x, а значит их можно выносить за интеграл:

$$= -\mathrm{div} \int p_t(x|x_0, x_1) f(x, t|x_0, x_1) q_{01}(x_0, x_1) \mathrm{d}x_0 \mathrm{d}x_1 + \frac{\beta_t}{2} \Delta \int p_t(x|x_0, x_1) q_{01}(x_0, x_1) \mathrm{d}x_0 \mathrm{d}x_1.$$

Замечаем, что под оператором Лапласа стоит не что иное, как безусловная динамика, а под дивергенцией мы пользуемся старым трюком: домножаем и делим на $p_t(x)$. Получаем

$$-\mathrm{div} \left(p_t(x) \cdot \int f(x, t|x_0, x_1) \frac{p_t(x|x_0, x_1) q_{01}(x_0, x_1)}{p_t(x)} \mathrm{d}x_0 \mathrm{d}x_1 \right) + \frac{\beta_t}{2} \Delta p_t(x).$$

Замечаем, что интеграл внутри дивергенции в точности является заявленным кандидатом на безусловное векторное поле

$$f(x, t) = \int f(x, t|x_0, x_1) \frac{p_t(x|x_0, x_1) q_{01}(x_0, x_1)}{p_t(x)} \mathrm{d}x_0 \mathrm{d}x_1.$$

Таким образом,

$$\frac{\partial}{\partial t} p_t(x) = -\mathrm{div} \left(p_t(x) f(x, t) \right) + \frac{\beta_t}{2} \Delta p_t(x),$$

то есть, безусловная динамика удовлетворяет уравнению Фоккера-Планка-Колмогорова с векторным полем f и коэффициентом диффузии $\sqrt{\beta_t}$, а значит, порождается с помощью SDE

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}W_t.$$

Таким образом, остается только выучить безусловное векторное поле f, а это можно, как и раньше, сделать с помощью регрессии на условное векторное поле (доказывается как в случае ODE, поэтому пропустим):

$$\int_0^1 \mathbb{E}_{q_{01}(x_0, x_1) p_t(x_t|x_0, x_1)} \| f_\theta(x_t, t) - f(x_t, t|x_0, x_1) \|^2 \mathrm{d}t \to \min_\theta .$$

Остается один вопрос: откуда нам брать условное векторное поле, соединяющее две точки x_0 и x_1? А мы этим занимались всю предыдущую секцию: подойдет любой условный процесс, заданный с помощью преобразования Дуба

$$\mathrm{d}X_t^{0,1} = \left(\mu(X_t^{0,1}) + \beta_t \nabla_{X_t^{0,1}} \log p_{1|t}(x_1|X_t^{0,1}) \right) \mathrm{d}t + \sqrt{\beta_t}\mathrm{d}W_t.$$

В частности, мы вывели это преобразование для винеровского процесса и получили броуновский мост

$$\mathrm{d}X_t^{0,1} = \beta_t \frac{x_1 - X_t^{0,1}}{\overline{\sigma}_t^2}\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}W_t,$$

где $\overline{\sigma}_t^2 = \int_t^1 \beta_s \mathrm{d}s$. Помимо этого, мы еще и вывели распределение такого процесса в момент времени t (а это, как мы помним, условная динамика $p_t(x_t|x_0, x_1)$). Здесь оно равно

$$p_t(x_t|x_0, x_1) = \mathcal{N}\left(\frac{\sigma_t^2}{\sigma_t^2 + \overline{\sigma}_t^2} \cdot x_1 + \frac{\overline{\sigma}_t^2}{\sigma_t^2 + \overline{\sigma}_t^2} \cdot x_0,\ \ \frac{\sigma_t^2 \overline{\sigma}_t^2}{\sigma_t^2 + \overline{\sigma}_t^2} \cdot I\right),$$

где

$$\sigma_t^2 = \int_0^t \beta_s \mathrm{d}s,\ \ \overline{\sigma}_t^2 = \int_t^1 \beta_s \mathrm{d}s.$$

Тогда алгоритм обучения становится очень простым:

1. Генерируем пару $(x_0, x_1) \sim q_{01}(x_0, x_1)$;
2. Генерируем $t \sim \mathcal{U}[0, 1]$;
3. Семплируем интерполяцию x_t из распределения $p_t(x_t|x_0, x_1)$ (выписано выше);
4. Подаем пару (x_t, t) на вход нейросети и вычисляем $f_\theta(x_t, t)$;
5. Считаем условное векторное поле $\beta_t \cdot (x_1 - x_t)/\overline{\sigma}_t^2$;
6. Делаем шаг по антиградиенту

$$\nabla_\theta \left\| f_\theta(x_t, t) - \beta_t \frac{x_1 - x_t}{\overline{\sigma}_t^2} \right\|^2.$$

На этапе генерации нужно будет семплировать начальную точку $X_0 \sim q_0$ и пропускать ее через SDE

$$\mathrm{d}X_t = f_\theta(X_t, t)\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}W_t,$$

что можно делать, например, по методу Эйлера

$$X_{t+h} = X_t + h \cdot f_\theta(X_t, t) + \sqrt{h\beta_t} \cdot \varepsilon_t;\ \ \varepsilon_t \sim \mathcal{N}(0, I),$$

а лучше, используя солверы. У полученной модели есть теоретическая гарантия того, что в итоге точка придет в $X_1 \sim q_1$, то есть, мы действительно будем генерировать из целевого распределения. Остаются ли здесь гарантии на уменьшение транспортной цены? Это тяжелый вопрос, возможно позже он будет здесь разобран.

3.2.4 Транспортная цена у Bridge Matching

[TODO]

Список литературы

[1] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

[2] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

[3] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

[4] Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[6] Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

[7] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[8] Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I^{2}sb: Image-to-image schr\"odinger bridge. *arXiv preprint arXiv:2302.05872*, 2023.

[9] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[10] L Chris G Rogers and David Williams. *Diffusions, Markov processes and martingales: Volume 2, Itô calculus*, volume 2. Cambridge university press, 2000.

[11] Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schr\"odinger bridge matching. *arXiv preprint arXiv:2303.16852*, 2023.

[12] Vignesh Ram Somnath, Matteo Pariset, Ya-Ping Hsieh, Maria Rodriguez Martinez, Andreas Krause, and Charlotte Bunne. Aligned diffusion schr\"odinger bridges. *arXiv preprint arXiv:2302.11419*, 2023.

[13] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[14] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[15] Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Conditional flow matching: Simulation-free dynamic optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.