

# Flow модели

## Содержание

<b>1</b>	<b>Пререквизиты</b>	<b>2</b>
1.1	ODE . . . . .	2
1.2	SDE . . . . .	3
1.3	Как генерировать? . . . . .	5
1.4	Эволюция плотности распределения . . . . .	5
1.4.1	Уравнение непрерывности . . . . .	6
1.4.2	Уравнение Фоккера-Планка-Колмогорова . . . . .	8
1.5	Единственность . . . . .	10
1.6	Резюме . . . . .	11
<b>2</b>	<b>Генеративные модели на основе ODE/SDE</b>	<b>11</b>
2.1	Непрерывные нормализующие потоки . . . . .	12
2.2	Flow Matching . . . . .	12
2.2.1	Идея . . . . .	12
2.2.2	Реализация . . . . .	13
2.2.3	Применение к генеративному моделированию . . . . .	16
2.2.4	Обуславливание на два конца . . . . .	18
2.2.5	Оптимальный транспорт . . . . .	20
2.2.6	Анализ транспортной цены Flow Matching . . . . .	21
2.2.7	Применение к парным задачам . . . . .	24

## 1 Пререквизиты

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

$$dX_t = f(X_t, t)dt$$

или стохастического (SDE) дифференциального уравнения

$$dX_t = f(X_t, t)dt + G(X_t, t)dW_t.$$

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

### 1.1 ODE

Обыкновенное дифференциальное уравнение

$$dX_t = f(X_t, t)dt$$

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:

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

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

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

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

$$dX_t = f(X_t, t)dt + \text{случайность}.$$

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  $t_1 < t_2 < \dots < t_n$  величины  $W_{t_1}, W_{t_2} - W_{t_1}, \dots, W_{t_n} - W_{t_{n-1}}$  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

$$dX_t = f(X_t, t)dt + G(X_t, t)dW_t.$$

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

$$\begin{aligned} X_{t+h} &\approx X_t + h \cdot f(X_t, t) + G(X_t, t) \cdot (W_{t+h} - W_t) = \\ &= X_t + h \cdot f(X_t, t) + G(X_t, t) \cdot \mathcal{N}(0, hI) \\ &= X_t + h \cdot f(X_t, t) + \sqrt{h} \cdot G(X_t, t) \cdot \mathcal{N}(0, I). \end{aligned}$$

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

$$\begin{cases} dX_t = f(X_t, t)dt + G(X_t, t)dW_t; \\ X_0 = Z, \end{cases}$$

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

$$dX_t = f(X_t, t)dt + g(t)dW_t,$$

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

$$\begin{cases} dX_t = f(X_t, t)dt + g(t)dW_t; \\ X_0 = Z \end{cases}$$

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 [3], говорит, что нужно максимизировать правдоподобие семплов из датасета с точки зрения плотности случайной величины  $X_1$  (которая, конечно, зависит от  $f$ ). Подход [6], 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

$$\begin{cases} dX_t = f(X_t, t)dt + g(t)dW_t; \\ X_0 = Z \end{cases}.$$

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  $\bar{o}(h)$  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:

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

В обратную сторону:

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

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  $\varphi$  и пораскрывать:

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

Сначала разберемся с определителем. Вспомним, что для подсчета определителя матрицы  $A$  нужно рассмотреть все перестановки  $\sigma$ , 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  $\bar{o}(h)$  и в пределе не дает вклада. Значит, важной здесь остается только диагональ, равная

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

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

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

$$\begin{aligned} p_{t+h}(x + h \cdot f(x,t)) &= p_{t+h}(x) + h \cdot \left\langle \frac{\partial p_{t+h}(x)}{\partial x}, f(x,t) \right\rangle + \bar{o}(h \cdot f(x,t)) = \\ &= p_{t+h}(x) + h \cdot \left\langle \frac{\partial p_{t+h}(x)}{\partial x}, f(x,t) \right\rangle + \bar{o}(h). \end{aligned}$$

Собираем вместе:

$$\begin{aligned} p_t(x) &= \left( p_{t+h}(x) + h \cdot \left\langle \frac{\partial p_{t+h}(x)}{\partial x}, f(x,t) \right\rangle + \bar{o}(h) \right) (1 + h \cdot \operatorname{div} f(x,t) + \bar{o}(h)) = \\ &= p_{t+h}(x) + h \cdot \left\langle \frac{\partial p_{t+h}(x)}{\partial x}, f(x,t) \right\rangle + h \cdot p_{t+h}(x) \cdot \operatorname{div} f(x,t) + \bar{o}(h) = \\ &= p_{t+h}(x) + h \cdot \sum_{i=1}^d \frac{\partial p_{t+h}(x)}{\partial x_i} \cdot f_i(x,t) + h \cdot \sum_{i=1}^d p_{t+h}(x) \cdot \frac{\partial f_i(x,t)}{\partial x_i} + \bar{o}(h) = \\ &= p_{t+h}(x) + h \cdot \sum_{i=1}^d \frac{\partial}{\partial x_i} (p_{t+h}(x) \cdot f_i(x,t)) + \bar{o}(h) = \\ &= p_{t+h}(x) + h \cdot \operatorname{div} (p_{t+h}(x) \cdot f(x,t)) + \bar{o}(h). \end{aligned}$$

Приводим к виду из определения производной:

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

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

$$\frac{\partial}{\partial t} p_t(x) = -\operatorname{div} (p_t(x) f(x, t)) , \quad (1)$$

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:

$$p_t(x) = p_{\varphi(X_t)}(x) + h \cdot \operatorname{div} (p_{\varphi(X_t)}(x) \cdot f(x, t)) + \bar{o}(h),$$

или

$$p_{\varphi(X_t)}(x) = p_t(x) - h \cdot \operatorname{div} (p_{\varphi(X_t)}(x) \cdot f(x, t)) + \bar{o}(h).$$

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

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

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

$$p_{t+h}(x) = \int \left( p_{\varphi(X_t)}(x) + (y - x) \frac{\partial}{\partial x} p_{\varphi(X_t)}(x) + \frac{1}{2} (y - x)^\top \frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x) (y - x) + o(\|y - x\|^2) \right) \cdot \mathcal{N}(y \mid x, h \cdot g^2(t) \cdot I) dy.$$

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  $\xi \sim \mathcal{N}(x, h \cdot g^2(t) \cdot I)$ . Оно равно нулю, так как  $\mathbb{E}\xi = x$ .



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

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

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  $\mathbb{E}\|\xi - \mathbb{E}\xi\|^2 = \text{Tr Cov}\xi = h \cdot g^2(t) \cdot d$ , где  $\text{Cov}\xi$  — ковариационная матрица  $\xi$ . Переставляя матожидание и о-малое (очень грубая и не всегда верная операция, но мы поверим), получаем  $\bar{o}(h \cdot g^2(t) \cdot d) = \bar{o}(h)$ , то есть это слагаемое не даст вклада.

Таким образом, у нас остается

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

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

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

Под матожиданием же стоит известная оценка следа Хатчинсона [5]. Воспользуясь циклическим свойством и линейностью следа, перепишем матожидание квадратичной формы:

$$\begin{aligned} \mathbb{E} \left[ \eta^\top \cdot \frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x) \cdot \eta \right] &= \mathbb{E} \text{Tr} \left[ \eta^\top \cdot \frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x) \cdot \eta \right] = \mathbb{E} \text{Tr} \left[ \frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x) \cdot \eta \eta^\top \right] = \\ &= \text{Tr} \mathbb{E} \left[ \frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x) \cdot \eta \eta^\top \right] = \text{Tr} \left[ \frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x) \mathbb{E} [\eta \eta^\top] \right] = \text{Tr} \left[ \frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x) \cdot I \right] = \\ &= \text{Tr} \left[ \frac{\partial^2}{\partial x^2} p_{\varphi(X_t)}(x) \right] = \Delta p_{\varphi(X_t)}(x), \end{aligned}$$

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

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

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

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

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

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

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

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

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

описывающее эволюцию плотности частицы, двигающейся под действием SDE.

## 1.5 Единственность

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

$$dX_t = f(X_t, t)dt,$$

удовлетворяет уравнению непрерывности

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

а плотность частицы, двигающейся под действием SDE

$$dX_t = f(X_t, t)dt + g(t)dW_t$$

удовлетворяет уравнению Фоккера-Планка-Колмогорова

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

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

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

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  $dX_t = f(X_t, t)dt + g(t)dW_t$  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

$$\begin{cases} dX_t = f(X_t, t)dt + g(t)dW_t; \\ X_0 = Z \sim p_0 \end{cases},$$

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/SDE

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 [6]. На самом деле, мы будем пользоваться результатом не только этой работы, но и множества фоллоу-апов и параллельных статей: [10, 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 [4, 9]. 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

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

где от  $\alpha_t < 1$  и  $\beta_t$  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 [6] предлагают рассмотреть для процесса генерации шум  $\rightarrow$  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 [10], 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

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

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

$$dX_t = f(X_t, t)dt.$$

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

$$p_t(x) = \int p_t(x|z)q(z)dz.$$

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

$$dX_t = f(X_t, t|z)dt,$$

то есть, условная динамика удовлетворяет уравнению непрерывности

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

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

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

где

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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:

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

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

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

и подставим это в функционал. Получаем

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

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

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

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

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

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

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

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

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

свелась к эквивалентной решаемой задаче

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

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:

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

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 [8]: в нем для нахождения безусловной скор-функции решается регрессия на условную.

### 2.2.3 Применение к генеративному моделированию

Авторы оригинальной статьи [6] 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:

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



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  $\varepsilon$ , 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

$$p_0(x) = \int p_0(x|x_1)q_1(x_1)dx_1 = \int \mathcal{N}(x|0, I)q_1(x_1)dx_1 = \mathcal{N}(x|0, I);$$

$$p_1(x) = \int p_1(x|x_1)q_1(x_1)dx_1 = \int \delta(x - x_1)q_1(x_1)dx_1 = q_1(x),$$

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:

$$\frac{d}{dt}x_t = \mu'_t(x_1) + \sigma'_t(x_1) \cdot x_0.$$

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

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

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

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

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

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

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 [6] пробуют брать динамику вероятностей из диффузионных моделей [9], 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

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

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

После обучения можно будет генерировать из модели, решая ODE

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

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 [10] 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  $\int q_{01}(x_0, x_1)dx_1 = q_0(x_0)$  и  $\int q_{01}(x_0, x_1)dx_0 = q_1(x_1)$ , 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:

$$\begin{aligned} p_0(x) &= \int \delta(x - x_0)q_{01}(x_0, x_1)dx_0dx_1 = \int q_{01}(x, x_1)dx_1 = q_0(x); \\ p_1(x) &= \int \delta(x - x_1)q_{01}(x_0, x_1)dx_0dx_1 = \int q_{01}(x_0, x)dx_0 = q_1(x). \end{aligned}$$

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

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

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

$$\int_0^1 \mathbb{E}_{q_{01}(x_0, x_1)p_t(x_t|x_0, x_1)} \|f_\theta(x_t, t) - (x_1 - x_0)\|^2 dt \rightarrow \min_\theta, \quad (4)$$

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

построению Flow Matching, решая ODE

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

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  $f_\theta$  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

Неамного 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 придумать более сложные функции, считающие расстояния, например, между эмбедингами 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, минимизируя среднее расстояние между семплом и его отображением:

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

Метрика

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

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 [7], 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  $p_t(x) = \int p_t(x|x_0, x_1) q_{01}(x_0, x_1) dx_0 dx_1$ .

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

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

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 [7] 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  $X_1 = X_0 + \int_0^1 f(X_t, t)dt$  — 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:

$$f(x, t) = \int f(x, t|x_0, x_1) p_t(x_0, x_1|x) dz.$$

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

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

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

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

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

$$X_1 = X_0 + \int_0^1 f(X_t, t) dt .$$

Тогда транспортную цену можно переписать как

$$\mathbb{E} c(X_1 - X_0) = \mathbb{E} c \left( \int_0^1 f(X_t, t) dt \right) .$$

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  $\varphi$  верно  $\varphi(\mathbb{E}\xi) \leq \mathbb{E}\varphi(\xi)$ . Тогда

$$\mathbb{E} c \left( \int_0^1 f(X_t, t) dt \right) \leq \mathbb{E} \int_0^1 c(f(X_t, t)) dt = \int_0^1 \mathbb{E} c(f(X_t, t)) dt .$$

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

$$\int_0^1 \mathbb{E} c(f(X_t, t)) dt = \int_0^1 \mathbb{E} c(f(x_t, t)) dt .$$

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

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

или при подставлении самого семпла  $x_t$  вместо  $x$ :

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

Подставляем и получаем

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

Теперь пользуемся неравенством Йенсена для условного матожидания и получаем

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

а матожидание условного матожидания — это просто матожидание:

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

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 обучение.

### 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  $p_t(x) = \int p_t(x|x_0, x_1)q_{01}(x_0, x_1)dt$ , 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 будем 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  $f_\theta$  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  $dX_t = f_\theta(X_t, t)dt$ , начиная с точки  $X_0$ . На практике это можно делать, например, используя схему Эйлера

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

## 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] 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.
- [4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [5] 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.
- [6] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- [7] 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.
- [8] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.
- [9] 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.
- [10] 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.