Problem 1. Пусть $\mathbf{X}_0 \sim p^{\text{data}}$ — объект из распределения данных, а $(\mathbf{X}_t, t = 0, \ldots, T)$ — дискретный процесс с сохраняющейся дисперсией

$$\mathbf{X}_t = \sqrt{1-\beta_t}\mathbf{X}_{t-1} + \sqrt{\beta_t}\varepsilon_t.$$

Предположим, что $\mathrm{D}_t^\theta(\boldsymbol{x})$ является теоретическим оптимумом в задаче оптимизации

$$\sum_{t=1}^{T} \mathbb{E}\|\mathrm{D}_t^\theta(\mathbf{X}_t) - \mathbf{X}_0\|^2 \to \min_\theta,$$

а генерация при помощи $\mathrm{D}_t^\theta(\boldsymbol{x})$ выглядит следующим образом:

Algorithm 1 Генерация из модели-денойзера

$\mathbf{X}_T \sim \mathcal{N}(0, I)$
for $t = T, \ldots, 1$ **do**
  $\widehat{\mathbf{X}}_0 = \mathrm{D}_t^\theta(\mathbf{X}_t)$ ▷ Предсказываем $\mathbf{X}_0$ из текущего состояния
  $\mathbf{X}_{t-1} \sim p_{t-1|t,0}(\mathbf{X}_{t-1}|\mathbf{X}_t, \widehat{\mathbf{X}}_0)$. ▷ Шаг назад, используя предсказанный $\mathbf{X}_0$
end for
return $\mathbf{X}_0$

Выпишите шаг генерации по Алгоритму 1 в терминах score-функции: покажите, что он будет иметь вид

$$\mathbf{X}_{t-1} = A_t\mathbf{X}_t + B_t\nabla \log p_t(\mathbf{X}_t) + \sqrt{C_t}\,\varepsilon_t,$$

где $\varepsilon_t \sim \mathcal{N}(0, I)$. Найдите A_t, B_t, C_t.

Покажите, что если взять $\beta_t = h\gamma_t$, то $A_t = 1 + h\frac{\gamma_t}{2} + \overline{o}(h)$ и $B_t = h\gamma_t + \overline{o}(h)$.

Замечание: если рассмотреть схему генерации из диффузионных моделей с непрерывным временем на основе VP-SDE

$$\mathrm{d}\mathbf{Y}_t = -\frac{\gamma_t}{2}\mathbf{Y}_t\mathrm{d}t + \sqrt{\gamma_t}\mathrm{d}\mathbf{W}_t$$

и соответствующую схему Эйлера для обратного VP-SDE:

$$\mathbf{Y}_{t-h} = \mathbf{Y}_t\left(1 + h\frac{\gamma_t}{2}\right) + h\gamma_t \cdot \nabla \log p_t(\mathbf{Y}_t) + \sqrt{h\gamma_t}\varepsilon_t,$$

то получится очень близкая к Алгоритму 1 схема.

1. Из прошлой дз знаем, что

$$p_{t-1|t,0}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_{t-1}\middle|\frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}\boldsymbol{x}_0 + \frac{\sqrt{1-\beta_t}(1-\alpha_{t-1})}{1-\alpha_t}\boldsymbol{x}_t, \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t\right)$$

Теперь можем выписать формулу для $\mathbf{X}_{t-1}$:

$$\mathbf{X}_{t-1} = \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}\widehat{\mathbf{X}}_0 + \frac{\sqrt{1-\beta_t}(1-\alpha_{t-1})}{1-\alpha_t}\mathbf{X}_t + \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t}\varepsilon_t =$$
$$= \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}\mathrm{D}_t^\theta(\mathbf{X}_t) + \frac{\sqrt{1-\beta_t}(1-\alpha_{t-1})}{1-\alpha_t}\mathbf{X}_t + \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t}\varepsilon_t \quad (1)$$

Так как денойзер D_t^θ обучен на VP процесс, то скор функция может быть записана следующим образом:

$$\nabla \log p_t(\boldsymbol{x}) = \frac{\sqrt{\alpha_t}\,\mathrm{D}_t^\theta(\boldsymbol{x}) - \boldsymbol{x}}{1-\alpha_t}$$

Почему именно так (обсуждалось на семинаре, но всё же): процесс VP, знаем как выглядит распределение в момент времени t при фиксированном $\mathbf{X}_0$; тогда как будто записали плотность этого распределения, взяли логарифм и производную, получили мат.ожидание минус точку и поделить на диспресию.

Попробуем подогнать выражение 1, чтобы увидеть там скор функцию:

$$\begin{aligned}\mathbf{X}_{t-1} = \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}\mathrm{D}_t^\theta(\mathbf{X}_t) + \frac{\sqrt{1-\beta_t}(1-\alpha_{t-1})}{1-\alpha_t}\mathbf{X}_t + \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t}\varepsilon_t = \\ = \frac{\sqrt{\alpha_t}\,\mathrm{D}_t^\theta(\mathbf{X}_t) - \mathbf{X}_t}{1-\alpha_t}\frac{\sqrt{\alpha_{t-1}}\beta_t}{\sqrt{\alpha_t}} - \frac{\mathbf{X}_t}{1-\alpha_t}\frac{\sqrt{\alpha_{t-1}}\beta_t}{\sqrt{\alpha_t}} + \frac{\mathbf{X}_t}{1-\alpha_t}\frac{\sqrt{\alpha_{t-1}}\beta_t}{\sqrt{\alpha_t}} + \frac{\sqrt{1-\beta_t}(1-\alpha_{t-1})}{1-\alpha_t}\mathbf{X}_t + \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t}\varepsilon_t = \\ = \nabla\log p_t(\mathbf{X}_t)\frac{\sqrt{\alpha_{t-1}}\beta_t}{\sqrt{\alpha_t}} + \mathbf{X}_t\frac{\sqrt{\alpha_{t-1}}\beta_t + \sqrt{\alpha_t}\sqrt{1-\beta_t}(1-\alpha_{t-1})}{(1-\alpha_t)\sqrt{\alpha_t}} + \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t}\varepsilon_t = \\ = \nabla\log p_t(\mathbf{X}_t)\frac{\beta_t}{\sqrt{1-\beta_t}} + \mathbf{X}_t\frac{\sqrt{\alpha_{t-1}}\beta_t + \sqrt{\alpha_t}\sqrt{1-\beta_t}(1-\alpha_{t-1})}{(1-\alpha_t)\sqrt{\alpha_t}} + \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t}\varepsilon_t \end{aligned} \tag{2}$$

Отдельно ещё расмотрим коэфф перед $\mathbf{X}_t$:

$$\begin{aligned}\frac{\sqrt{\alpha_{t-1}}\beta_t + \sqrt{\alpha_t}\sqrt{1-\beta_t}(1-\alpha_{t-1})}{(1-\alpha_t)\sqrt{\alpha_t}} = \sqrt{\alpha_{t-1}}\,\frac{\beta_t + \sqrt{1-\beta_t}\sqrt{1-\beta_t}(1-\alpha_{t-1})}{(1-\alpha_t)\sqrt{\alpha_t}} = \\ = \frac{1}{\sqrt{1-\beta_t}}\frac{\beta_t + (1-\beta_t)(1-\alpha_{t-1})}{(1-\alpha_t)} = \frac{1}{\sqrt{1-\beta_t}}\frac{\beta_t + 1 - \beta_t - \alpha_t}{(1-\alpha_t)} = \frac{1}{\sqrt{1-\beta_t}}\end{aligned} \tag{3}$$

Итого получили, что

$$\begin{aligned} A_t &= \frac{1}{\sqrt{1-\beta_t}} \\ B_t &= \frac{\beta_t}{\sqrt{1-\beta_t}} \\ C_t &= \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t \end{aligned} \tag{4}$$

2. Теперь подставим $\beta_t = h\gamma_t$:

   (a) в A_t:

   $$\frac{1}{\sqrt{1-\beta_t}} = \frac{1}{\sqrt{1-h\gamma_t}} = (1-h\gamma_t)^{-1/2} = 1 + \frac{h\gamma_t}{2} + \overline{o}(h)$$

   (b) в B_t:

   $$\frac{\beta_t}{\sqrt{1-\beta_t}} = \frac{h\gamma_t}{\sqrt{1-h\gamma_t}} = h\gamma_t\left(1 + \frac{h\gamma_t}{2} + \overline{o}(h)\right) = h\gamma_t + \overline{o}(h)$$

Problem 2. Пусть случайная величина $\mathbf{X}$ имеет распределение с плотностью $p(\boldsymbol{x})$. Покажите, что функционал обучения score функции

$$\frac{1}{2}\mathbb{E}\|s_\theta(\mathbf{X}) - \nabla \log p(\mathbf{X})\|^2 \to \min_\theta$$

1. в одномерном случае с точностью до константы, не зависящей от θ, совпадает с функционалом
$$\mathbb{E}\left(\frac{1}{2}(s_\theta(\mathbf{X}))^2 + \frac{\partial}{\partial \boldsymbol{x}} s_\theta(\mathbf{X})\right) \to \min_\theta;$$
в предположении, что
$$\lim_{\boldsymbol{x}\to+\infty} s_\theta(\boldsymbol{x})p(\boldsymbol{x}) = \lim_{\boldsymbol{x}\to-\infty} s_\theta(\boldsymbol{x})p(\boldsymbol{x}) = 0;$$
2. в общем случае с точностью до константы, не зависящей от θ, совпадает с функционалом
$$\mathbb{E}\left(\frac{1}{2}\|s_\theta(\mathbf{X})\|^2 + \operatorname{div} s_\theta(\mathbf{X})\right) \to \min_\theta$$
в предположении, что
$$\lim_{\|\boldsymbol{x}\|\to+\infty} s_\theta(\boldsymbol{x})p(\boldsymbol{x}) = 0;$$
3. выведите из этого, что score-функцию можно обучать с помощью минимизации
$$\mathbb{E}\left(\frac{1}{2}\|s_\theta(\mathbf{X})\|^2 + \left\langle \frac{\partial}{\partial \boldsymbol{x}} s_\theta(\mathbf{X})\varepsilon, \varepsilon\right\rangle\right) \to \min_\theta,$$
где ε — независимая с $\mathbf{X}$ величина с $\mathbb{E}\,\varepsilon = 0$, $\operatorname{Var}\varepsilon = I$.

Замечание: предположение про предел имеет смысл, так как идеальное решение $s_\theta(\boldsymbol{x}) = \nabla \log p(\boldsymbol{x})$ удовлетворяет $\nabla \log p(\boldsymbol{x}) \cdot p(\boldsymbol{x}) = \nabla p(\boldsymbol{x})$, что логично предположить нулем на бесконечности, так как там плотность нулевая и не меняется.

1. Перепишем, как выглядит выражение в одномерном случае:
$$\frac{1}{2}\mathbb{E}\left[\left(s_\theta(\mathbf{X}) - \frac{\partial}{\partial \boldsymbol{x}}\log p(\mathbf{X})\right)^2\right] =$$
Теперь раскроем квадрат:
$$= \frac{1}{2}\mathbb{E}\left[(s_\theta(\mathbf{X}))^2 + \left(\frac{\partial}{\partial \boldsymbol{x}}\log p(\mathbf{X})\right)^2 - 2s_\theta(\mathbf{X})\frac{\partial}{\partial \boldsymbol{x}}\log p(\mathbf{X})\right]$$
Слагаемое посередине является константой по θ:
$$= \frac{1}{2}\mathbb{E}\left[(s_\theta(\mathbf{X}))^2 - 2s_\theta(\mathbf{X})\frac{\partial}{\partial \boldsymbol{x}}\log p(\mathbf{X})\right] = \mathbb{E}\left[\frac{1}{2}(s_\theta(\mathbf{X}))^2\right] - \mathbb{E}\left[s_\theta(\mathbf{X})\frac{\partial}{\partial \boldsymbol{x}}\log p(\mathbf{X})\right]$$
Посмотрим отдельно на второе мат.ожидание:
$$\mathbb{E}\left[s_\theta(\mathbf{X})\frac{\partial}{\partial \boldsymbol{x}}\log p(\mathbf{X})\right] = \int_{-\infty}^{+\infty} s_\theta(\boldsymbol{x})\frac{\partial}{\partial \boldsymbol{x}}\log p(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \left\{\frac{\partial}{\partial \boldsymbol{x}}\log p(\boldsymbol{x}) = \frac{\frac{\partial}{\partial \boldsymbol{x}}p(\boldsymbol{x})}{p(\boldsymbol{x})}\right\} =$$
$$= \int_{-\infty}^{+\infty} s_\theta(\boldsymbol{x})\frac{\partial}{\partial \boldsymbol{x}}p(\boldsymbol{x})d\boldsymbol{x} = \int_{-\infty}^{+\infty} \frac{\partial}{\partial \boldsymbol{x}}s_\theta(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} - \int_{-\infty}^{+\infty} p(\boldsymbol{x})\frac{\partial}{\partial \boldsymbol{x}}s_\theta(\boldsymbol{x})d\boldsymbol{x} =$$
$$= \left\{\text{по предположению} \int_{-\infty}^{+\infty} \frac{\partial}{\partial \boldsymbol{x}}s_\theta(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = s_\theta(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}\Big|_{-\infty}^{+\infty} = 0\right\} =$$
$$= -\int_{-\infty}^{+\infty} p(\boldsymbol{x})\frac{\partial}{\partial \boldsymbol{x}}s_\theta(\boldsymbol{x})d\boldsymbol{x} = -\mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{x}}s_\theta(\mathbf{X})\right]$$
Возвращаясь назад, получаем:
$$\mathbb{E}\left[\frac{1}{2}(s_\theta(\mathbf{X}))^2\right] + \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{x}}s_\theta(\mathbf{X})\right] = \mathbb{E}\left[\frac{1}{2}(s_\theta(\mathbf{X}))^2 + \frac{\partial}{\partial \boldsymbol{x}}s_\theta(\mathbf{X})\right]$$

2. Опять же, раскрываем скобки с сразу выкинем член, не влияющий на оптимизацию:

$$\mathbb{E}\left[\frac{1}{2}\|s_\theta(\mathbf{X})\|^2 - \langle s_\theta(\mathbf{X}), \nabla \log p(\mathbf{X})\rangle\right] =$$

Посмотрим на второй член отдельно:

$$\mathbb{E}\left[\langle s_\theta(\mathbf{X}), \nabla \log p(\mathbf{X})\rangle\right] = \int_{-\infty}^{+\infty} p(\boldsymbol{x})\langle s_\theta(\boldsymbol{x}), \nabla \log p(\boldsymbol{x})\rangle d\boldsymbol{x} = \int_{-\infty}^{+\infty} \langle s_\theta(\boldsymbol{x}), \nabla p(\boldsymbol{x})\rangle d\boldsymbol{x} =$$

$$= \int_{-\infty}^{+\infty} \sum_{i=1}^{n} s_\theta^{(i)}(\boldsymbol{x}) \frac{\partial}{\partial x_i} p(\boldsymbol{x}) d\boldsymbol{x} = \left\{ s_\theta^{(i)}(\boldsymbol{x}) \frac{\partial}{\partial x_i} p(\boldsymbol{x}) = \frac{\partial}{\partial x_i}(s_\theta^{(i)}(\boldsymbol{x})p(x)) - p(x)\frac{\partial}{\partial x_i} s_\theta^{(i)}(\boldsymbol{x}) \right\} =$$

$$= \int_{-\infty}^{+\infty} \sum_{i=1}^{n} \frac{\partial}{\partial x_i}(s_\theta^{(i)}(\boldsymbol{x})p(x)) d\boldsymbol{x} - \int_{-\infty}^{+\infty} \sum_{i=1}^{n} p(x)\frac{\partial}{\partial x_i} s_\theta^{(i)}(\boldsymbol{x}) d\boldsymbol{x} = 0 - \int_{-\infty}^{+\infty} p(x) \text{div } s_\theta(\boldsymbol{x}) d\boldsymbol{x} = -\mathbb{E}\left[\text{div } s_\theta(\mathbf{X})\right]$$

Почем интеграл занулился: меняем сумму с интегралом, далее интеграл по $d\boldsymbol{x}$ можем записать как $dx_1 dx_2 \ldots dx_n$ и, по теореме Фубини, можем менять интегралы местами, один занулится по предположению, а за ним и остальные.

Тогда получаем:

$$= \mathbb{E}\left[\frac{1}{2}\|s_\theta(\mathbf{X})\|^2 + \text{div } s_\theta(\mathbf{X})\right]$$

3. Введём обозначение div $s_\theta(\mathbf{X}) = Tr(J)$, где J – якобиан $s_\theta(\mathbf{X})$. Тогда поработаем отдельно со вторым слагаемым:

$$\mathbb{E}\left[\text{div } s_\theta(\mathbf{X})\right] = \mathbb{E}\left[Tr(J)\right] = Tr(\mathbb{E}[J]) = Tr(\mathbb{E}[J]\mathrm{I}) = Tr(\mathbb{E}[J]\mathbb{E}[\varepsilon\varepsilon^\top]) = Tr(\mathbb{E}[J\varepsilon\varepsilon^\top]) = Tr(\mathbb{E}[\varepsilon^\top J\varepsilon]) =$$

$$= \mathbb{E}[Tr(\varepsilon^\top J\varepsilon)] = \mathbb{E}[\varepsilon^\top J\varepsilon] = \mathbb{E}[\langle J\varepsilon, \varepsilon\rangle]$$

Problem 3.

1. Пусть $\mathbf{X}, \mathbf{Z}$ — независимые величины. Покажите, что

$$\nabla \log p_{\mathbf{X}+\mathbf{Z}}(\boldsymbol{y}) = \mathbb{E}\left[\nabla \log p_{\mathbf{X}}(\mathbf{X}) | \mathbf{X}+\mathbf{Z} = \boldsymbol{y}\right].$$

2. Предложите на основе первого пункта функционал обучения score функции $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$ шумного распределения, если известна score функция чистого распределения $\nabla_{\boldsymbol{x}} \log p_0(\boldsymbol{x})$, а прямой процесс задан VE-SDE

$$\begin{cases} \mathrm{d}\mathbf{X}_t = g(t)\mathrm{d}\mathbf{W}_t; \\ \mathbf{X}_0 \sim p_0. \end{cases}$$

1.

$$\nabla \log p_{\mathbf{X}+\mathbf{Z}}(\boldsymbol{y}) = \frac{\nabla p_{\mathbf{X}+\mathbf{Z}}(\boldsymbol{y})}{p_{\mathbf{X}+\mathbf{Z}}(\boldsymbol{y})} = \frac{1}{p_{\mathbf{X}+\mathbf{Z}}(\boldsymbol{y})} \nabla_{\boldsymbol{y}} \int_{-\infty}^{+\infty} p_{\mathbf{X}}(\boldsymbol{y}-t) p_{\mathbf{Z}}(t) dt = \frac{1}{p_{\mathbf{X}+\mathbf{Z}}(\boldsymbol{y})} \int_{-\infty}^{+\infty} \nabla_{\boldsymbol{y}} p_{\mathbf{X}}(\boldsymbol{y}-t) p_{\mathbf{Z}}(t) dt =$$

$$= \{s = \boldsymbol{y} - t, \nabla_{s+t} p_{\mathbf{X}}(s) = \nabla_s p_{\mathbf{X}}(s), dt = |-1| dv\} = \frac{1}{p_{\mathbf{X}+\mathbf{Z}}(\boldsymbol{y})} \int_{-\infty}^{+\infty} \nabla_s p_{\mathbf{X}}(s) p_{\mathbf{Z}}(\boldsymbol{y} - s) ds =$$

$$= \{p_{\mathbf{Z}}(\boldsymbol{y} - s) = p_{\mathbf{X}+\mathbf{Z}|\mathbf{X}}(\boldsymbol{y}|s)\} = \frac{1}{p_{\mathbf{X}+\mathbf{Z}}(\boldsymbol{y})} \int_{-\infty}^{+\infty} \nabla_s p_{\mathbf{X}}(s) p_{\mathbf{X}+\mathbf{Z}|\mathbf{X}}(\boldsymbol{y}|s) ds =$$

$$= \frac{1}{p_{\mathbf{X}+\mathbf{Z}}(\boldsymbol{y})} \int_{-\infty}^{+\infty} \nabla_s \log p_{\mathbf{X}}(s) p_{\mathbf{X}}(s) p_{\mathbf{X}+\mathbf{Z}|\mathbf{X}}(\boldsymbol{y}|s) ds = \int_{-\infty}^{+\infty} \nabla_s \log p_{\mathbf{X}}(s) p_{\mathbf{X}|\mathbf{X}+\mathbf{Z}}(s|\boldsymbol{y}) ds =$$

$$= \mathbb{E}\left[\nabla \log p_{\mathbf{X}}(\mathbf{X}) | \mathbf{X}+\mathbf{Z} = \boldsymbol{y}\right]$$

2. У нас был следующий фукнционал (для фиксированного t):

$$\mathbb{E}\left\|s_t^\theta(\mathbf{X}_t) - \nabla \log p_t(\mathbf{X}_t)\right\|^2 \to \min_\theta$$

Подробнее рассмотрим p_t:

$$p_t(\boldsymbol{x}) = p_{\mathbf{X}_t}(\boldsymbol{x})$$

Так как процесс VE, то $\mathbf{X}_t = \mathbf{X}_0 + \varepsilon_t \sqrt{\int_0^t g(s)^2 ds}$, тогда

$$p_{\mathbf{X}_t}(\boldsymbol{x}) = p_{\mathbf{X}_0 + \varepsilon_t \sqrt{\int_0^t g(s)^2 ds}}(\boldsymbol{x})$$

Так как оба слагаемых в плотности независимые, то можем воспользоваться результатом первого пункта:

$$\nabla \log p_t(\boldsymbol{x}) = \nabla \log p_{\mathbf{X}_0 + \varepsilon_t \sqrt{\int_0^t g(s)^2 ds}}(\boldsymbol{x}) = \mathbb{E}\left[\nabla \log p_{\mathbf{X}_0}(\mathbf{X}_0) \middle| \mathbf{X}_0 + \varepsilon_t \sqrt{\int_0^t g(s)^2 ds} = \boldsymbol{x}\right]$$

По условию, $\nabla \log p_{\mathbf{X}_0}(\mathbf{X}_0) = \nabla \log p_0(\mathbf{X}_0)$ известен. Соответственно, пользуясь свойством, что минимизация квадрата разности функции и условного мат.ожидания с точностью до константы – это минимизация квадрата разности этой функции и то, почему мы матожидали, получаем:

$$\mathbb{E}\left\|s_t^\theta\left(\mathbf{X}_0 + \varepsilon_t \sqrt{\int_0^t g(s)^2 ds}\right) - \nabla \log p_0(\mathbf{X}_0)\right\|^2 \to \min_\theta$$