

## Homework 2

### Autoregressive models

1. Рассмотрим модель MADE с одним скрытым слоем. Обозначим за  $\mathbf{W}$  матрицу весов между входным и скрытым слоем, а за  $\mathbf{V}$  матрицу весов между скрытым и выходным слоем. Пусть мы сгенерировали корректные авторегрессионные маски  $\mathbf{M}_\mathbf{W}$  и  $\mathbf{M}_\mathbf{V}$  (алгоритм генерации приведен в лекции 1) для прямого порядка переменных ( $p(\mathbf{x}) = p(x_1) \cdot p(x_2|x_1) \cdots p(x_m|x_{m-1}, \dots, x_1)$ ). Каждая маска является бинарной матрицей из 0 и 1. Введем матрицу  $\mathbf{M} = \mathbf{M}_\mathbf{W}\mathbf{M}_\mathbf{V}$ . Докажите, что:

- (a) **(2 pt)**  $\mathbf{M}$  строго верхняя треугольная (имеет нули на диагонали и ниже диагонали);
  - (b) **(2 pt)**  $M_{ij}$  равно числу путей в графе сети между выходным нейроном  $\hat{x}_i$  и входным нейроном  $x_j$ .
2. Пусть у нас есть 2 генеративные модели для изображений размера  $W \times H \times C$ , где  $W$  - ширина изображения,  $H$  - высота,  $C$  - число каналов. Первая модель  $p_1(\mathbf{x}|\boldsymbol{\theta})$  выдает дискретное распределение для каждого пикселя  $\text{Categorical}(\boldsymbol{\pi})$ , где  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_{256})$ . Вторая модель  $p_2(\mathbf{x}|\boldsymbol{\theta})$  моделирует дискретное распределение непрерывной смесью логистических функций

$$p(\nu|\boldsymbol{\mu}, \mathbf{s}, \boldsymbol{\pi}) = \sum_{i=1}^K \pi_i p(\nu|\mu_i, s_i).$$

$$P(x|\boldsymbol{\mu}, \mathbf{s}, \boldsymbol{\pi}) = P(x + 0.5|\boldsymbol{\mu}, \mathbf{s}, \boldsymbol{\pi}) - P(x - 0.5|\boldsymbol{\mu}, \mathbf{s}, \boldsymbol{\pi}).$$

Каждая из моделей выдает параметры распределений пикселей.

- (a) **(1 pt)** Посчитайте размерность выходного тензора для модели  $p_1(\mathbf{x}|\boldsymbol{\theta})$  и для модели  $p_2(\mathbf{x}|\boldsymbol{\theta})$ .
- (b) **(1 pt)** При каком числе компонент смеси  $K$  число элементов выходного тензора для  $p_2(\mathbf{x}|\boldsymbol{\theta})$  становится больше, чем для  $p_1(\mathbf{x}|\boldsymbol{\theta})$ .

### Latent Variable models

1. **(2 pt)** Пусть имеется два распределения  $p_1(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ ,  $p_2(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ . Выведите формулу для  $KL(p_1||p_2)$ .
2. На лекции 3 при выводе градиента ELBO на Е-шаге мы столкнулись с проблемой при Монте-Карло оценивании, так как функция распределения зависела от параметров дифференцирования.

$$\begin{aligned} \nabla_\phi \mathcal{L}(\phi, \boldsymbol{\theta}) &= \nabla_\phi \int q(\mathbf{z}|\mathbf{x}, \phi) [\log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) - \log q(\mathbf{z}|\mathbf{x}, \phi)] d\mathbf{z} \\ &\neq \int q(\mathbf{z}|\mathbf{x}, \phi) \nabla_\phi [\log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) - \log q(\mathbf{z}|\mathbf{x}, \phi)] d\mathbf{z} \end{aligned}$$

Reparametrization trick позволил пробросить градиент и получить Монте-Карло оценку. Но есть и другой способ, который использует log-derivative trick:

$$\nabla_\xi \log q(\eta|\xi) = \frac{\nabla_\xi q(\eta|\xi)}{q(\eta|\xi)}.$$

- (a) **(2 pt)** Используя формулу для производной логарифма получите Монте-Карло оценку градиента.
- (b) **(2 pt)** Полученная оценка работает существенно хуже, чем reparametrization trick. А именно обладает огромной дисперсией. Попробуйте описать интуицию, почему оценка обладает высоким разбросом (для этого нужно подумать какого порядка и знака будут иметь члены, участвующие в оценке).

3. **(3 pt)** В курсе нам встретятся дивергенции, отличные от  $KL$ . Поэтому давайте познакомимся с целым классом  $\alpha$ -дивергенций:

$$D_{\alpha}(p||q) = \frac{4}{1-\alpha^2} \left( 1 - \int p(x)^{\frac{1+\alpha}{2}} q(x)^{\frac{1-\alpha}{2}} dx \right).$$

Для любого значения  $\alpha \in [-\infty; +\infty]$  функция  $D_{\alpha}(p||q)$  будет задавать некоторую меру схожести двух распределений, обладающую свои свойствами.

Докажите, что при  $\alpha \rightarrow 1$  дивергенция  $D_{\alpha}(p||q) \rightarrow KL(p||q)$ , а при  $\alpha \rightarrow -1$  дивергенция  $D_{\alpha}(p||q) \rightarrow KL(q||p)$ . При доказательстве используйте факт, что  $t^{\epsilon} = \exp(\epsilon \ln t) = 1 + \epsilon \ln t + O(\epsilon^2)$ .