Fixing a Broken ELBO

Ольга Гребенькова

Байесовское мультимоделирование

24 ноября 2021 г.

О чем пойдет речь?

Обучение без учителя

Для разных моделей типа VAE основной задачей является нахождения какого-то "полезного "представления данных. То есть мы хотим найти модель следующего вида $p(x, z|\Theta) = p(z|\Theta)p(x|z, \Theta)$.

Классический подход

Обычно находим параметры модели минимизируя KL дивергенцию, что аналогично максимизации правдоподобия модели. Если выражение невычислимо, можем максимизировать нижнюю оценку этого выражения (ELBO) или рассмотреть другие виды дивергенций (RKL).

Проблема

Функции потерь зависят от $p(x|\Theta)$, а не от $p(x, z|\Theta)$. Пример: PixelVAE, где получаем хорошее маргинальное правдоподобие $p(x|\Theta)$. Получение хорошей нижней оценки (а значит и хорошеего значения маргинального правдоподобия $p(x|\Theta)$) не достаточно.

x — вектор данных; z — скрытое представление; $e(z|x)$ — выбранный нами энкодер.

Совместное распредление

$$p_e(x, z) = p(x)e(z|x).$$

Маргинальное апостериорное распределение

$$p_e(z) = \int p(x)e(z|x)dx.$$

Условное распределение

$$p_e(x|z) = \frac{p_e(x, z)}{p_e(z)}.$$

Взаимная информация

$$I_e(X;Z) = \int\int dxdz p_e(x,z) \log \frac{p_e(x,z)}{p(x)p_e(z)} \qquad (1)$$

1. $I(X,Z) = I(Z,X)$
2. $I(X,Z) \geq 0$
3. не зависит от репараметризаций
4. показывает как много информации одна случайная величина содержит о другой
5. e — указывает на зависимость от энкодера
6. если X, Y независимы, $I(X,Y) = 0$
7. если $X = Y$, $I(X,Y) = H(X)$
8. скорее не вычислима из-за $p(x)p_e(z)$ (эмпирическое распределение и оценки)

Оценки для взаимной информации

$$H - D \leq I_e(X; Z) \leq R$$

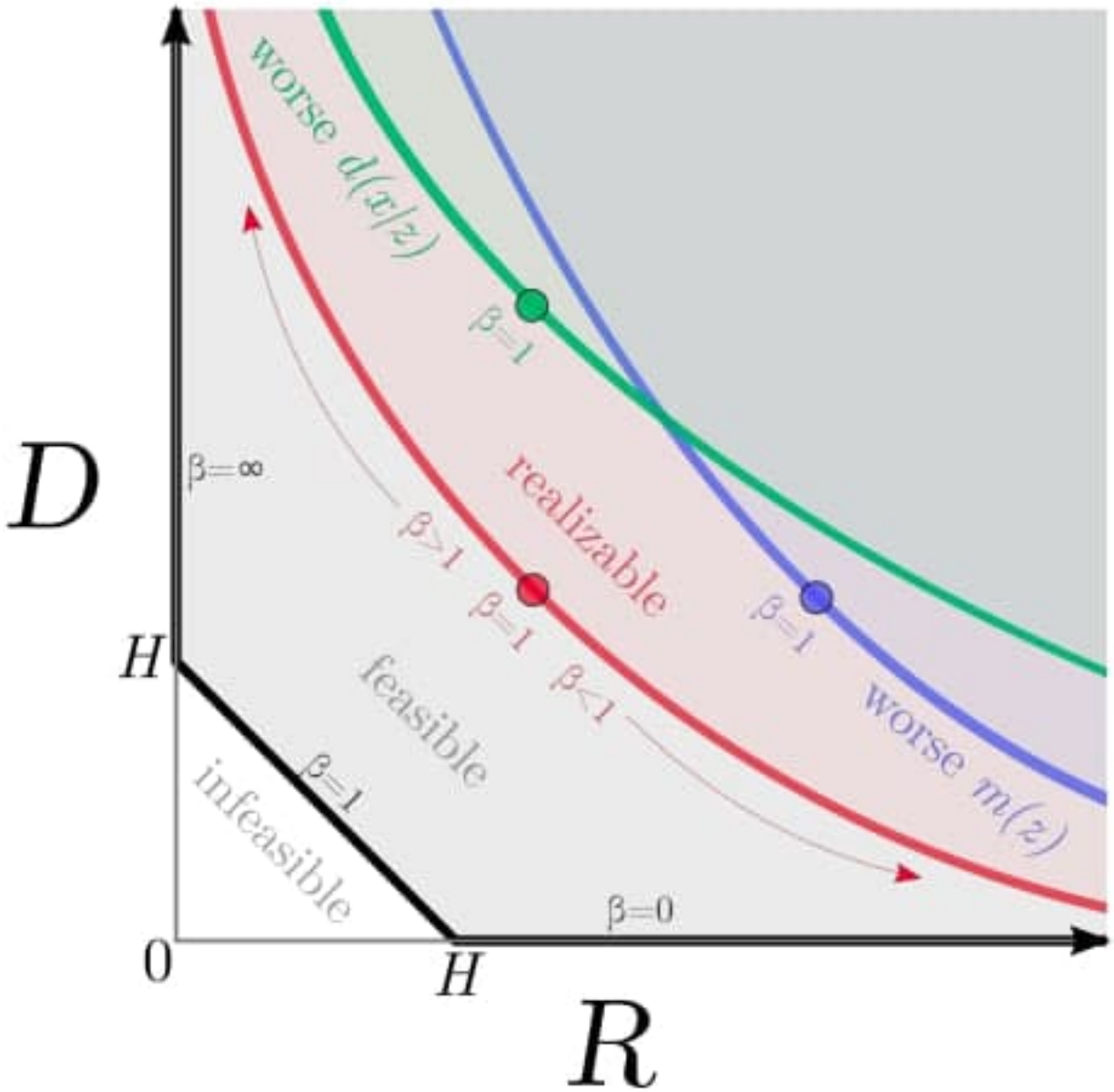
$$H = -\int dx p(x) \log p(x)$$

$$D = -\int dx p(x) \int dz e(z|x) \log d(x|z)$$

$$R = \int dx p(x) \int dz e(z|x) \log \frac{e(z|x)}{m(z)}$$

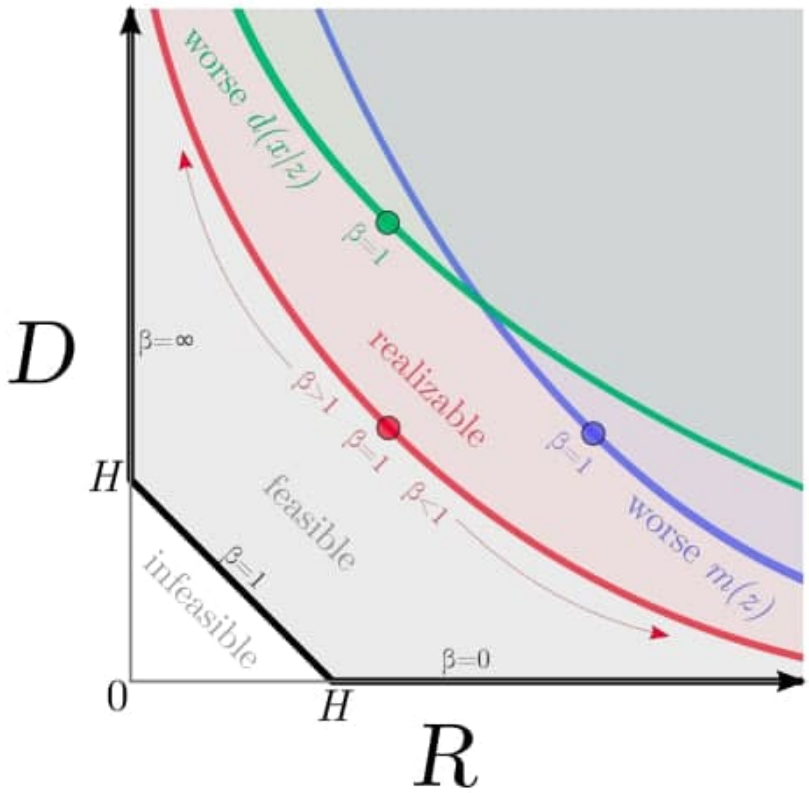
- $d(x|z)$ (Decoder) — аппроксимация $p_e(x|z)$
- $m(z)$ (Marginal) — аппроксимация $p_e(z)$
- H — энтропия данных, показывающая сложность датасета
- D — искажение, через кодировщик и декодер, равный обратному логарифму правдоподобия. отрицательная логарифмическая вероятность.
- R — значение, зависящее только от кодировщика и аппроксимации $p_e(z)$. По сути это средняя KL дивергенция между распрделенеим кодировщика и выученым маргинальным распредлением.
- Для дискретных данных $H \geq 0, D \geq 0, R \geq 0$.

Диаграмма


D
R
H
H
0
β=∞
β=1
β=1
β=1
β=1
β<1
β>1
β=0
worse d(x|z)
worse m(z)
realizable
feasible
infeasible

Вместо поиска оптимального $D(R)$, воспользуемся преобразованием Лежандра и будем искать оптимальные значения D и R для фикисрованного значения $\beta = \frac{\partial D}{\partial R}$ минимизируя $\min_{e(z|x),m(z),d(x|z)} D + \beta R$:

$$\min_{e(z|x),m(z),d(x|z)} \int dx p(x) \int dz e(z|x)[-\log d(x|z) + \beta \log \frac{e(z|x)}{m(z)}].$$



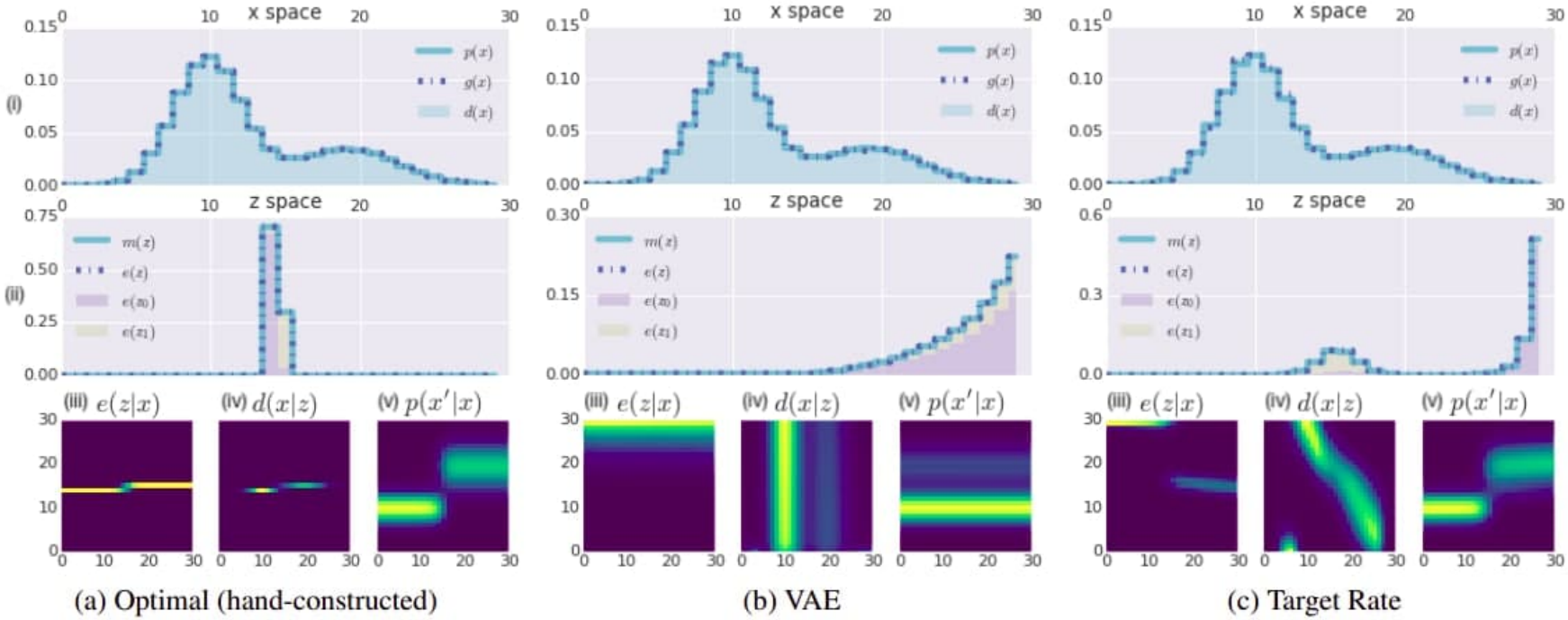


Figure 2. Toy Model illustrating the difference between fitting a model by maximizing ELBO (b) vs minimizing distortion for a fixed rate (c). **Top (i):** Three distributions in data space: the true data distribution, $p^*(x)$, the model's generative distribution, $g(x) = \sum_z m(z)d(x|z)$, and the empirical data reconstruction distribution, $d(x) = \sum_{x'} \sum_z \hat{p}(x')e(z|x')d(x|z)$. **Middle (ii):** Four distributions in latent space: the learned (or computed) marginal $m(z)$, the empirical induced marginal $e(z) = \sum_x \hat{p}(x)e(z|x)$, the empirical distribution over z values for data vectors in the set $\mathcal{X}_0 = \{x_n : z_n = 0\}$, which we denote by $e(z_0)$ in purple, and the empirical distribution over z values for data vectors in the set $\mathcal{X}_1 = \{x_n : z_n = 1\}$, which we denote by $e(z_1)$ in yellow. **Bottom:** Three $K \times K$ distributions: (iii) $e(z|x)$, (iv) $d(x|z)$ and (v) $p(x'|x) = \sum_z e(z|x)d(x'|z)$.