VARIACIJSKI SAMOKODIRNIKI KOT VERJETNOSTNI OKVIR ZA OCENJEVANJE GOSTOTE VERJETNOSTI

JAKOB ŠEGA

Fakulteta za matematiko in fiziko
Univerza v Ljubljani

Modeliranje porazdelitve podatkov oziroma ocenjevanje gostote verjetnosti je pomemben problem statističnega sklepanja. Članek predstavi variacijske samokodirnike, eno izmed sodobnih metod globokega strojnega učenja za reševanje tega problema. Obravnava matematično ozadje njihovega delovanja, pojasni postopek učenja modela porazdelitve ter pokaže, kako je naučeni model mogoče uporabiti za ustvarjanje novih podatkov.

VARIATIONAL AUTOENCODERS AS A PROBABILISTIC FRAMEWORK FOR DENSITY ESTIMATION

Data distribution modeling or probability density estimation is an important problem of statistical inference. This article presents variational autoencoders, one of the modern deep machine learning methods for solving this problem. It addresses the mathematical background of their operation, explains the training process of the distribution model, and demonstrates how the learned model can be used to generate new data.

1. Uvod

Modeliranje porazdelitve podatkov oziroma ocenjevanje gostote verjetnosti (angl. *density estimation*) predstavlja enega izmed osrednjih problemov statističnega sklepanja in strojnega učenja. Problem lahko matematično formuliramo takole: na podlagi končne množice neodvisnih in enako porazdeljenih vzorcev $\boldsymbol{X} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)}\}$, generiranih iz neznane porazdelitve $\mathcal{P}^*$, želimo poiskati model $\mathcal{P}$, ki čim bolje aproksimira porazdelitev $\mathcal{P}^*$.

V statistiki poznamo več pristopov k reševanju tega problema, ki jih v osnovi delimo na parametrične in neparametrične. Med parametrične pristope sodi npr. cenilka največjega verjetja (angl. *Maximum Likelihood Estimation*) za vnaprej izbrane parametrične družine porazdelitev ali mešane modele (angl. *mixture models*), kot so npr. Gaussovi mešani modeli. Med neparametrične pristope pa uvrščamo histograme in ocenjevanje gostote z jedrnimi funkcijami (angl. *Kernel Density Estimation*). Čeprav so ti klasični pristopi teoretično dobro utemeljeni, se v praksi soočajo z resnimi računskimi in konceptualnimi omejitvami, ko imamo opravka z visokodimenzionalnimi podatki.

Kot odgovor na te omejitve so se razvili sodobni pristopi, ki problem ocenjevanja gostote rešujejo z uvajanjem nelinearnih transformacij in latentnih spremenljivk (angl. *latent variables*). V Bayesovem smislu lahko te pristope razdelimo v dve glavni skupini. Prva se osredotoča na eksplicitno verjetnostno modeliranje in aproksimacijo porazdelitve podatkov, kar omogoča neposredno vzorčenje iz naučene gostote. Sem med drugimi sodijo variacijski samokodirniki, normalizacijski tokovi in difuzijski modeli. Druga skupina pa se eksplicitnemu modeliranju gostote izogne in se osredotoča zgolj na mehanizem generiranja vzorcev prek minimizacije določenih razdalj med porazdelitvami (npr. generativne nasprotniške mreže).

V tem članku se osredotočamo na matematično ozadje variacijskih samokodirnikov (angl. *variational autoencoders*, VAE), ki predstavljajo verjetnostni okvir za učenje nelinearnih modelov latentnih spremenljivk. Delovanje variacijskih samokodirnikov predstavimo na podlagi 17. poglavja učbenika [1] in izvirnega članka [2]. Variacijski samokodirniki porazdelitev podatkov modelirajo posredno z uporabo latentnih spremenljivk v prostoru nižje dimenzije, pri čemer apriorna (angl. *prior*) porazdelitev določa njihovo osnovno strukturo, aposteriorna (angl. *posterior*) porazdelitev pa opisuje,

katere latentne reprezentacije so verjetne za posamezen opazovani podatek. To ponuja učinkovito in matematično elegantno rešitev za problem ocenjevanja gostote nad kompleksnimi visokodimenzionalnimi podatki.

2. Samokodirniki

Samokodirniki (angl. *autoencoders*, AE) so nevronske mreže, ki se učijo učinkovitega stiskanja podatkov s pomočjo samonadzorovanega učenja. Samokodirnik učimo tako, da vhodne podatke preslikamo v nižjedimenzionalni latentni prostor, nato pa jih poskušamo iz te reprezentacije čim bolj natančno obnoviti.

2.1 Arhitektura samokodirnikov

Samokodirnik sestavljata dva dela (Slika 1):

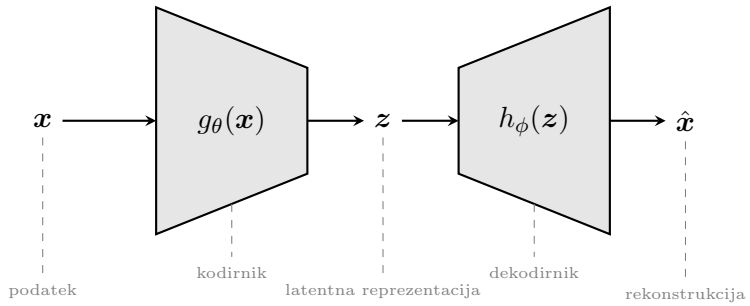
1. **Kodirnik**: funkcija g_θ, ki vhodni podatek $\boldsymbol{x}$ preslika v latentni vektor $\boldsymbol{z}$ nižje dimenzije.

$$\boldsymbol{z} = g_\theta(\boldsymbol{x})$$

2. **Dekodirnik**: funkcija h_ϕ, ki iz latentnega vektorja $\boldsymbol{z}$ ustvari rekonstrukcijo $\hat{\boldsymbol{x}}$ vhodnega podatka $\boldsymbol{x}$.

$$\hat{\boldsymbol{x}} = h_\phi(\boldsymbol{z}) = h_\phi(g_\theta(\boldsymbol{x}))$$

Tukaj θ in ϕ predstavljata parametre (uteži) nevronskih mrež kodirnika in dekodirnika, nizkodimenzionalni prostor, v katerega slikamo, pa imenujemo latentni prostor. Ključna značilnost samokodirnika je, da je dimenzija latentnega prostora bistveno nižja od dimenzije vhodnih podatkov, kar sili mrežo, da v latentnem prostoru obdrži le najpomembnejše značilnosti podatkov (npr. oblike, barve ali teksture na slikah), šum in nepomembne podrobnosti pa zavrže.



Slika 1. Shematski prikaz arhitekture samokodirnika.

Samokodirnik učimo z minimizacijo rekonstrukcijske napake, ki meri razliko med originalnim vhodom $\boldsymbol{x}$ in njegovo rekonstrukcijo $\hat{\boldsymbol{x}}$. Najpogostejša izbira za zvezne podatke je povprečna kvadratna napaka (MSE):

$$\mathcal{L}(\phi, \theta) = \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{x}^{(i)} - h_\phi(g_\theta(\boldsymbol{x}^{(i)}))\|^2,$$

kjer je N število učnih primerov. Cilj optimizacije je najti takšne parametre ϕ in θ, da bo napaka čim manjša. V idealnem primeru bi bila funkcija $h_\phi \circ g_\theta$ identiteta, vendar to preprečuje nizka dimenzija latentnega prostora, ki prisili model v iskanje stisnjene reprezentacije.

2.2 Problem deterministočnega latentnega prostora

Samokodirnik učinkovito stisne in rekonstruira podatke, vendar ni generativni model, ker ne zagotavlja latentnega prostora, iz katerega bi lahko vzorčili. Samokodirnik se namreč uči le preslikave iz učne množice v latentni prostor in nazaj v prostor, iz katerega so učni primeri – brez implicitnega modeliranja porazdelitve latentnih kod. Zaradi tega je latentni prostor pri klasičnih samokodirnikih deterministočen in običajno nelogično razporejen, kar preprečuje smiselno generiranje novih podatkov iz latentnih kod, v katere niso bili preslikani učni primeri.

Glavna težava je, da učni proces modela ne vodi do zveznega latentnega prostora, kjer bi bile podobne latentne kode povezane s podobnimi podatki. Ker model minimizira le rekonstrukcijsko napako posameznih učnih primerov, pogosto pride do preprileganja na učni množici.

To vodi do dveh ključnih težav (Slika 2):

1. **Neurejenost:** Latentne kode $\boldsymbol{z}^{(i)}$, ki pripadajo podobnim vhodom $\boldsymbol{x}^{(i)}$, se lahko v latentnem prostoru nahajajo daleč narazen.

2. **Razpršenost:** Latentni prostor lahko vsebuje območja, v katera se ni preslikal noben učni podatek, in ki ne vsebujejo smiselnih latentnih kod. Če vzorčimo točko $\boldsymbol{z}^{(\text{new})}$ iz takega območja, bo izhod dekodirnika $h_\phi(\boldsymbol{z}^{(\text{new})})$ verjetno nesmiseln (šum), ki se bo bistveno razlikoval od podatkov v učni množici.

(a) Primer neurejenosti (b) Primer razpršenosti

Slika 2. Vizualizacija pomanjkljivosti deterministočnega latentnega prostora: (a) bližina v latentnem prostoru ne zagotavlja podobnosti podatkov v originalnem prostoru, (b) območja brez latentnih kod učnih podatkov onemogočajo smiselno generiranje novih podatkov.

Samokodirnik se torej nauči identitete za učne podatke, ne nauči pa se strukture osnovne porazdelitve podatkov $f_{\boldsymbol{X}}(\boldsymbol{x})$.

3. Variacijski samokodirniki

V tem poglavju bomo vpeljali variacijske samokodirnike, ki za razliko od klasičnih samokodirnikov temeljijo na verjetnostnem okviru. Namesto učenja deterministične preslikave bomo skušali modelirati osnovno porazdelitev podatkov.

3.1 Modeli latentnih spremenljivk

Modeli latentnih spremenljivk (angl. *latent variable models*) poskušajo opisati porazdelitveno gostoto večrazsežne opazovane spremenljivke $\boldsymbol{X}$ posredno s pomočjo uvedbe neopazovane latentne spremenljivke $\boldsymbol{Z}$. Model, namesto da bi neposredno podal porazdelitveno gostoto opazovanih podatkov $f_{\boldsymbol{X}}(\boldsymbol{x})$, opiše skupno porazdelitveno gostoto $f_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z})$ podatkov $\boldsymbol{X}$ in latentne spremenljivke $\boldsymbol{Z}$. Porazdelitveno gostoto podatkov nato dobimo z integracijo po latentnem prostoru:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int f_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}.$$

Običajno skupno porazdelitveno gostoto z uporabo pravil za pogojno verjetnost razbijemo na pogojno porazdelitev $f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{x})$ in apriorno porazdelitev $f_{\boldsymbol{Z}}(\boldsymbol{z})$:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{x}) f_{\boldsymbol{Z}}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}.$$

Ta pristop omogoča modeliranje zapletenih porazdelitev podatkov z uporabo relativno enostavne pogojne porazdelitve podatkov in apriorne porazdelitve latentnih spremenljivk.

Osnovna predpostavka je, da so zapleteni vzorci v podatkih $\boldsymbol{x}$ posledica enostavnejših procesov, ki jih lahko modelirajo te latentne spremenljivke.

V primeru diskretne latentne spremenljivke Z s K možnimi vrednostmi je integracija po latentnem prostoru obtežena vsota:

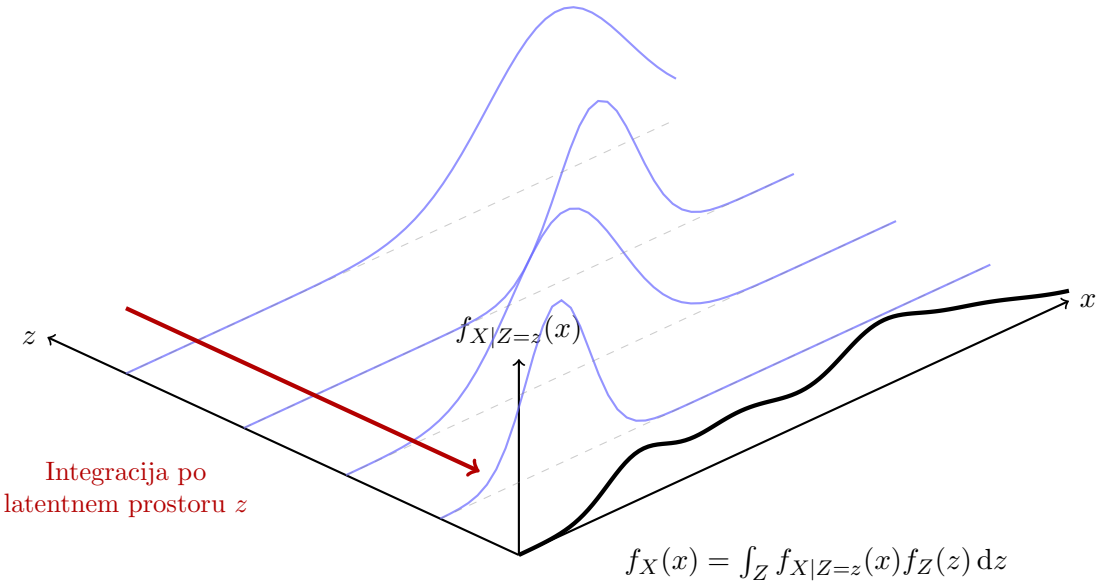
$$f_X(x) = \sum_{z=1}^{K} f_{X|Z=z}(x) P(Z = z).$$

Oglejmo si primer, kjer so pogojne porazdelitvene gostote $f_{X|Z=z}(x)$ normalne, apriorna porazdelitev pa je podana z verjetnostjo $P(Z = z)$:

$$\begin{aligned} P(Z = z) &= \lambda_z \\ X|Z = z &\sim \mathcal{N}(\mu_z, \sigma_z^2). \end{aligned}$$

V tem primeru je robna porazdelitvena gostota podatkov $f_X(x)$ obtežena vsota normalnih porazdelitev (Slika 3):

$$f_X(x) = \sum_{z=1}^{K} \lambda_z \frac{1}{\sigma_z \sqrt{2\pi}} \exp\left(-\frac{(x-\mu_z)^2}{2\sigma_z^2}\right).$$



Slika 3. Porazdelitvena gostota podatkov $f_X(x)$ je rezultat integracije skupne porazdelitvene gostote po latentnem prostoru.

3.2 Nelinearni modeli latentnih spremenljivk

V nelinearnih modelih latentnih spremenljivk so podatki $\boldsymbol{X}$ in latentne spremenljivke $\boldsymbol{Z}$ zvezne in večrazsežne. Najpogosteje predpostavimo, da je apriorna porazdelitev standardna večrazsežna normalna:

$$\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Pogojna porazdelitev podatkov je prav tako večrazsežna normalna. Njene parametre določata nelinearna funkcija latentne spremenljivke $\boldsymbol{\mu}(z, \phi)$ in diagonalna kovariančna matrika $\sigma^2\mathbf{I}$:

$$\boldsymbol{X}|\boldsymbol{Z} = \boldsymbol{z}, \phi \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{z}, \phi), \sigma^2\mathbf{I}).$$

Funkcijo $\boldsymbol{\mu}(z, \phi)$ določa nevronska mreža dekodirnika $h_\phi(\boldsymbol{z})$ s parametri ϕ in vhodom $\boldsymbol{z}$. Latentna spremenljivka $\boldsymbol{Z}$ je nižje dimenzije od podatkov $\boldsymbol{X}$. Model $\boldsymbol{\mu}(z, \phi)$ opisuje pomembne lastnosti podatkov, preostale nezajete lastnosti pa pripišemo šumu $\sigma^2\mathbf{I}$. Pogojevanje na parametre ϕ je zloraba notacije, ki si jo bomo privoščili, da poudarimo odvisnost pogojne porazdelitve od parametrov dekodirnika.

Porazdelitveno gostoto podatkov $f_{\boldsymbol{X}|\phi}(\boldsymbol{x})$ ponovno dobimo z integracijo po latentnem prostoru (Slika 4):

$$\begin{aligned} f_{\boldsymbol{X}|\phi}(\boldsymbol{x}) &= \int f_{\boldsymbol{X},\boldsymbol{Z}|\phi}(\boldsymbol{x}, \boldsymbol{z})\, \mathrm{d}\boldsymbol{z} \\ &= \int f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z},\phi}(\boldsymbol{x}) f_{\boldsymbol{Z}}(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z}. \end{aligned}$$

To si lahko predstavljamo kot neskončno obteženo vsoto večrazsežnih normalnih porazdelitev z različnimi pričakovanimi vrednostmi $\boldsymbol{\mu}(\boldsymbol{z}, \phi)$ in utežmi $f_{\boldsymbol{Z}}(\boldsymbol{z})$.

Zgled 1 (Latentni prostor za generiranje krogov)**.** Za boljšo predstavo si oglejmo preprost primer, kjer so naši podatki slike krogov različnih velikosti in na različnih položajih. Čeprav je slika sestavljena iz več tisoč pikslov, lahko vsak krog popolnoma opišemo s tremi parametri: koordinatama središča (x_0, y_0) in polmerom r (Slika 5).

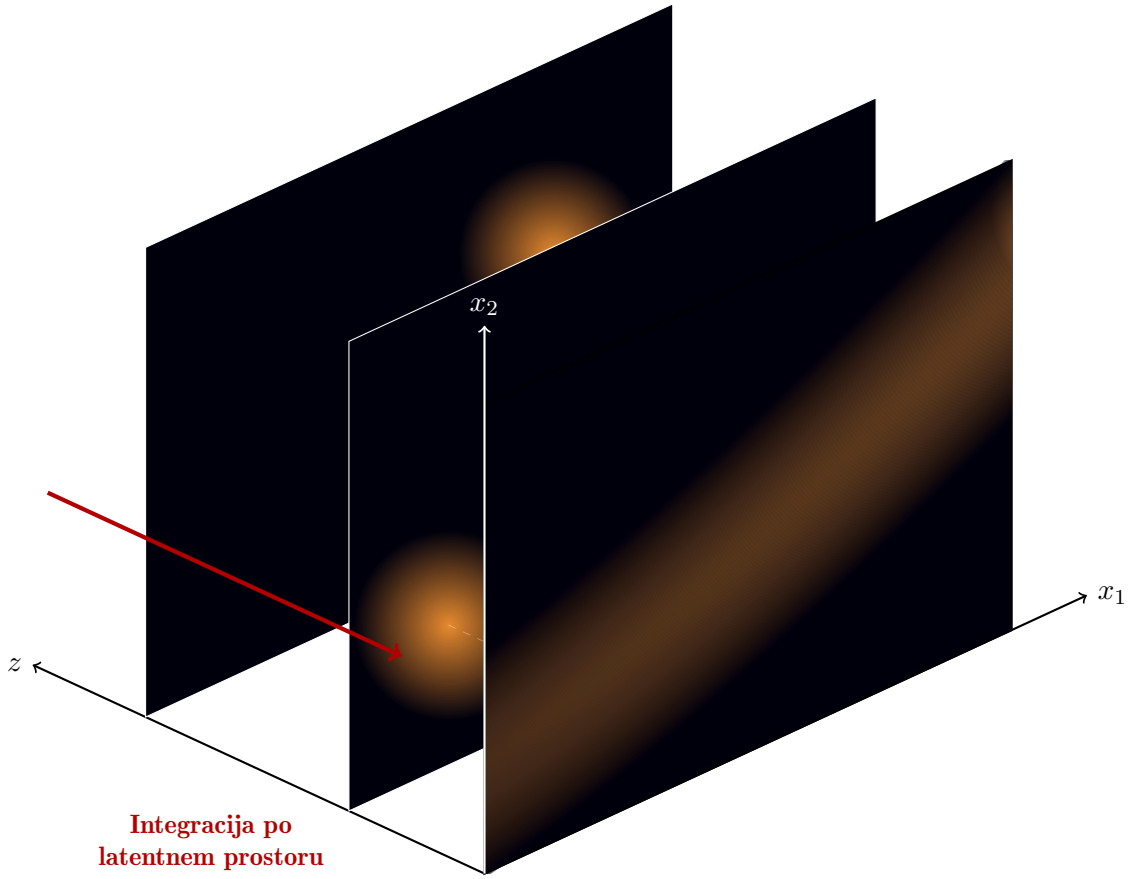
V idealnem modelu bi bila latentna spremenljivka $\boldsymbol{Z}$ tridimenzionalna, kjer bi vsaka komponenta z_i nadzorovala eno od teh lastnosti. Zvezni latentni prostor nam omogoča, da z majhno spremembo vrednosti $\boldsymbol{z}$ zvezno spreminjamo videz generiranega kroga (npr. ga premikamo po sliki ali mu večamo polmer), kar rešuje problem razpršenosti, ki smo ga opazili pri klasičnih samokodirnikih.

3.2.1 Generiranje novih podatkov

Ena najpomembnejših lastnosti modelov latentnih spremenljivk je njihova sposobnost generiranja novih, umetnih podatkov, ki so podobni podatkom iz učne množice. Postopek generiranja (angl. *ancestral sampling*) poteka v dveh korakih (Slika 6):

1. Izberemo naključno točko iz latentnega prostora glede na apriorno porazdelitev: $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

2. To točko z dekodirnikom preslikamo v podatkovni prostor $\boldsymbol{X}$, da dobimo parametre pogojne porazdelitve $\boldsymbol{X}|\boldsymbol{Z} = \boldsymbol{z}, \phi \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{z}, \phi), \sigma^2\mathbf{I})$, iz katere vzorčimo $\hat{\boldsymbol{x}}$.



Slika 4. Vizualizacija porazdelitvene gostote podatkov kot integral skupne porazdelitvene gostote $\boldsymbol{X}$ in $\boldsymbol{Z}$ po zveznem latentnem prostoru spremenljivke $\boldsymbol{Z}$.

3.3 Učenje

Pri učenju nelinearnega modela latentnih spremenljivk je naš osnovni cilj najti takšne parametre dekodirnika ϕ, da bo model čim bolje opisoval dejansko porazdelitev podatkov $\boldsymbol{X}$ s porazdelitveno gostoto $f^*_{\boldsymbol{X}}(\boldsymbol{x})$. V statističnem smislu to pomeni, da iščemo parametre, ki maksimizirajo verjetnost, da so bili učni podatki $\boldsymbol{X}$ vzorčeni iz porazdelitve s porazdelitveno gostoto $f_{\boldsymbol{X}|\phi}(\boldsymbol{x})$.

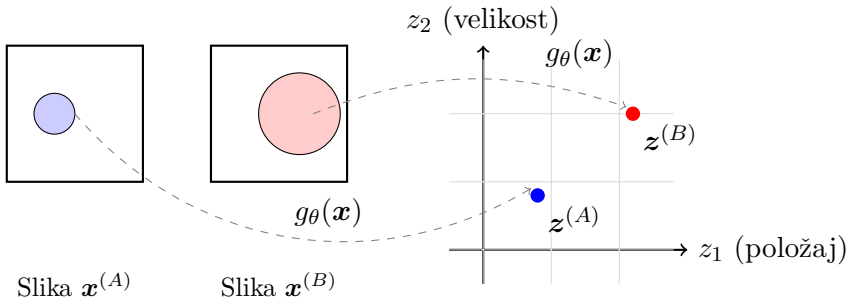
3.3.1 Cenilka največjega verjetja

Postopek učenja temelji na cenilki največjega verjetja. Želimo maksimizirati verjetje (angl. *likelihood*) našega modela glede na učno množico $\{\boldsymbol{x}^{(i)}\}_{i=1}^N$. Predpostavimo, da je varianca σ^2 znana in se osredotočimo le na učenje parametrov ϕ. Zaradi numerične stabilnosti in lažjega odvajanja namesto neposrednega verjetja maksimiziramo njegov logaritem, kar produkt verjetnosti posameznih neodvisnih vzorcev pretvori v vsoto:

$$\hat{\phi} = \arg\max_{\phi} \sum_{i=1}^{N} \ln f_{\boldsymbol{X}|\phi}(\boldsymbol{x}^{(i)}).$$

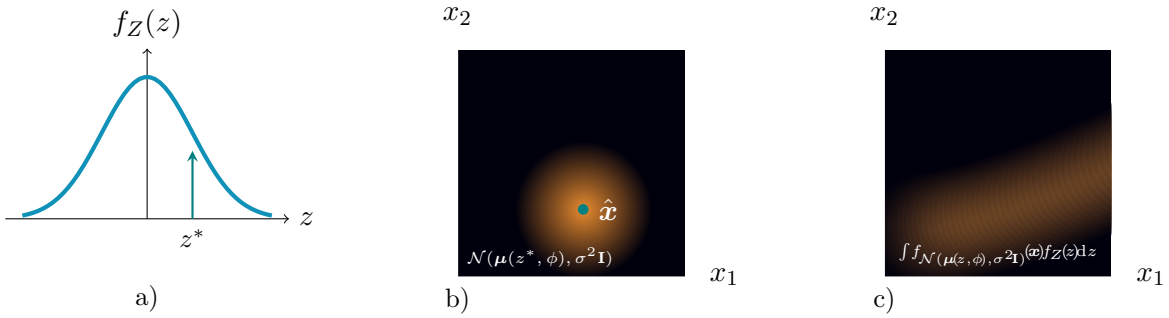
Kot smo videli v poglavju o nelinearnih modelih 3.2, je vrednost $f_{\boldsymbol{X}|\phi}(\boldsymbol{x})$ določena z integralom po celotnem latentnem prostoru:

$$f_{\boldsymbol{X}|\phi}(\boldsymbol{x}) = \int f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z},\phi}(\boldsymbol{x}) f_{\boldsymbol{Z}}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}. \tag{1}$$



(a) Opazovani podatki $\boldsymbol{X}$ (b) Latentni prostor $\boldsymbol{Z}$

Slika 5. Povezava med podatkovnim in latentnim prostorom. Vsaki sliki kroga ustreza točka v latentnem prostoru. Zveznost prostora $\boldsymbol{Z}$ omogoča, da z majhno spremembo vrednosti $\boldsymbol{z}$ zvezno spreminjamo velikost in položaj generiranega kroga.



Slika 6. Generacija z uporabo nelinearnega modela latentnih spremenljivk: a) vzorčenje iz apriorne porazdelitve $Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, b) nato $\hat{\boldsymbol{x}}$ vzorčimo iz pogojne porazdelitve v prostoru podatkov za izbrani z^* in c) če to ponovimo večkrat, dobimo robno porazdelitev X.

3.3.2 Analitična neizračunljivost integrala in spodnja meja (ELBO)

Integral v enačbi (1) je analitično neizračunljiv, nima rešitve v zaprti obliki in ni preprostega načina, da bi ga ovrednotili za posamezen $\boldsymbol{x}^{(i)}$. To težavo rešimo tako, da namesto neposredne maksimizacije logaritemskega verjetja maksimiziramo spodnjo mejo verjetja (angl. *Evidence Lower Bound*, ELBO).

Čeprav je koncept ELBO danes ključen v modernem strojnem učenju, je bil matematikom in statistikom pod tem ali sorodnimi imeni na voljo že desetletja. Izpeljava tovrstnih spodnjih mej je namreč tesno povezana z algoritmom EM (angl. *Expectation-Maximization*), ki so ga v klasičnem delu za iskanje cenilk največjega verjetja pri nepopolnih ali skritih podatkih vpeljali Dempster, Laird in Rubin [3]. Za sodobno izpeljavo ELBO v okviru variacijskih samokodirnikov pa potrebujemo Jensenovo neenakost.

3.3.3 Jensenova neenakost

Izrek 2 (Jensenova neenakost)**.** Naj bo X slučajna spremenljivka, ki zavzame vrednosti na intervalu $I \subset \mathbb{R}$ in naj bo $\varphi : I \to \mathbb{R}$ konkavna funkcija. Če je:

$$\mathbb{E}[|X|] < \infty \quad \text{in} \quad \mathbb{E}[|\varphi(X)|] < \infty,$$

potem velja:

$$\varphi(\mathbb{E}[X]) \geq \mathbb{E}[\varphi(X)].$$

Dokaz. Naj bo φ konkavna funkcija na intervalu I in naj bo $\mu = \mathbb{E}[X]$. Ker je φ konkavna, v vsaki točki x obstaja premica, tako da celotna funkcija leži pod njo ali na njej in se v točki x dotika funkcije. Enačba te premice je:

$$L(z) = \varphi(x) + a(z - x),$$

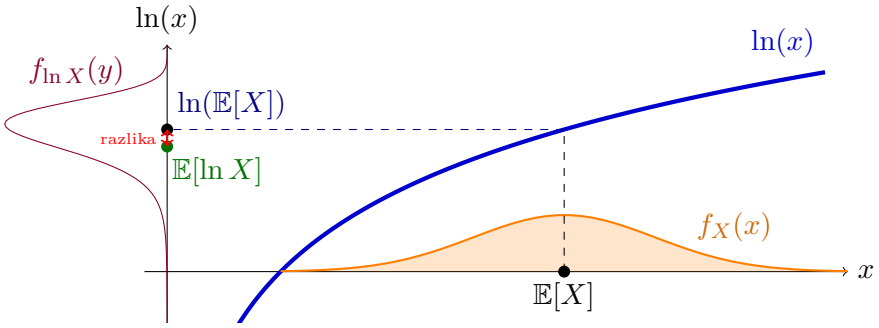
kjer je a naklon premice v točki x. Zaradi konkavnosti velja $\varphi(z) \leq L(z)$ za vse $z \in I$:

$$\varphi(z) \leq \varphi(x) + a(z - x).$$

Če namesto z v neenakost vstavimo slučajno spremenljivko X in vzamemo $x = \mu$ ter na obeh straneh neenakosti uporabimo pričakovano vrednost, zaradi linearnosti pričakovane vrednosti dobimo:

$$\begin{aligned}
\mathbb{E}[\varphi(X)] &\leq \mathbb{E}[\varphi(\mu) + a(X - \mu)] \\
\mathbb{E}[\varphi(X)] &\leq \varphi(\mu) + a(\mathbb{E}[X] - \mu) \\
\mathbb{E}[\varphi(X)] &\leq \varphi(\mu) + a(\mu - \mu) \\
\mathbb{E}[\varphi(X)] &\leq \varphi(\mu).
\end{aligned}$$

Ker je $\mu = \mathbb{E}[X]$, dobimo želeno neenakost: $\mathbb{E}[\varphi(X)] \leq \varphi(\mathbb{E}[X])$. ∎



Slika 7. Ilustracija Jensenove neenakosti za funkcijo $\ln(x)$.

Sedaj si oglejmo Jensenovo neenakost za primer $\varphi = \ln$ (Slika 7):

$$\ln(\mathbb{E}[X]) \geq \mathbb{E}[\ln(X)]$$

in zapišimo pričakovano vrednost kot integral:

$$\ln\left(\int f_X(x) x \,\mathrm{d}x\right) \geq \int f_X(x) \ln(x) \,\mathrm{d}x. \tag{2}$$

Velja še nekoliko bolj splošna trditev:

$$\ln\left(\int f_X(x) \psi(x) \,\mathrm{d}x\right) \geq \int f_X(x) \ln(\psi(x)) \,\mathrm{d}x,$$

kjer $\psi(x)$ predstavlja (nelinearno) transformacijo vhoda x. Če je $\psi(x) > 0$ za vse x, lahko definiramo $Y = \psi(X)$ in uporabimo Jensenovo neenakost (2) za ln na spremenljivki Y. S tem dobimo zgornjo neenakost. Posledica tega je, da trditev drži za poljubno gostoto $f_X(x)$, pod pogojem da so ustrezni integrali definirani.

3.3.4 Izpeljava spodnje meje

Uporabimo Jensenovo neenakost, da izpeljemo spodnjo mejo logaritemskega verjetja. Za začetek porazdelitveno gostoto pomnožimo in delimo s poljubno pogojno porazdelitveno gostoto $q_{\boldsymbol{Z}}(\boldsymbol{z})$:

$$\begin{aligned}\ln(f_{\boldsymbol{X}|\phi}(\boldsymbol{x})) &= \ln\left(\int f_{\boldsymbol{X},\boldsymbol{Z}|\phi}(\boldsymbol{x},\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}\right)\\ &= \ln\left(\int q_{\boldsymbol{Z}}(\boldsymbol{z})\frac{f_{\boldsymbol{X},\boldsymbol{Z}|\phi}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{Z}}(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}\right),\end{aligned}$$

uporabimo Jensenovo neenakost z $\varphi = \ln$ in $\psi(\boldsymbol{z}) = \frac{f_{\boldsymbol{X},\boldsymbol{Z}|\phi}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{Z}}(\boldsymbol{z})}$:

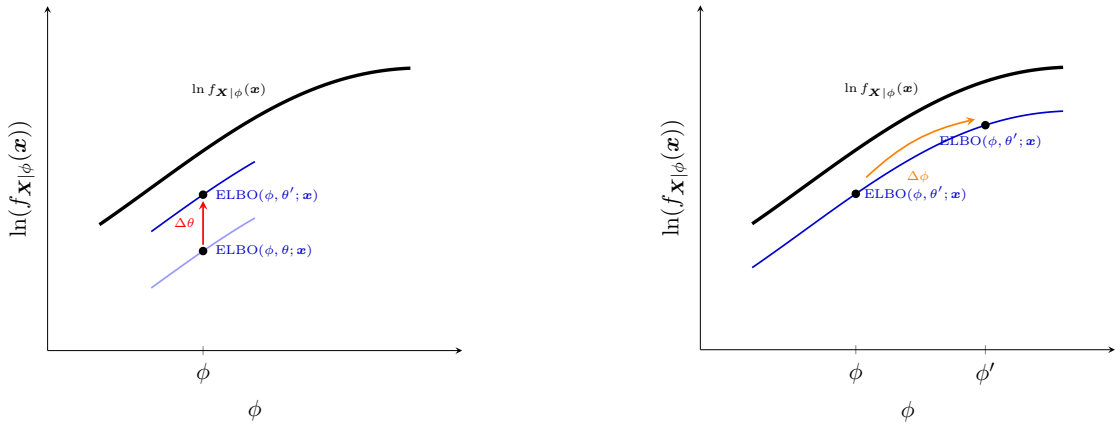
$$\ln(f_{\boldsymbol{X}|\phi}(\boldsymbol{x})) \geq \int q_{\boldsymbol{Z}}(\boldsymbol{z})\ln\left(\frac{f_{\boldsymbol{X},\boldsymbol{Z}|\phi}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{Z}}(\boldsymbol{z})}\right)\mathrm{d}\boldsymbol{z},$$

desna stran neenakosti je spodnja meja logaritemskega verjetja. Ime izhaja iz poimenovanja za $f_{\boldsymbol{X}|\phi}(\boldsymbol{x})$, ki se v Bayesovi statistiki imenuje robno verjetje oziroma evidenca (angl. *evidence*) (Izrek 5).

V praksi je porazdelitvena gostota $q_{\boldsymbol{Z}}(\boldsymbol{z})$ odvisna od parametrov θ. Tako dobimo končno obliko ELBO:

$$\mathrm{ELBO}(\phi,\theta;\boldsymbol{x}) = \int q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})\ln\left(\frac{f_{\boldsymbol{X},\boldsymbol{Z}|\phi}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})}\right)\mathrm{d}\boldsymbol{z}. \tag{3}$$

Nelinearni model latentnih spremenljivk se torej uči tako, da maksimizira spodnjo mejo verjetja ELBO kot funkcijo parametrov ϕ in θ (Slika 8). Nevronski arhitekturi, ki izračuna ELBO, pravimo variacijski samokodirnik.



a) Optimizacija θ b) Optimizacija ϕ

Slika 8. Vizualizacija dvostopenjske optimizacije spodnje meje verjetja (ELBO). a) prikazuje korak optimizacije parametrov θ: pri fiksnem ϕ posodobimo $\theta \to \theta'$, s čimer ELBO postane tesnejša meja logaritemskega verjetja $\ln f_{\boldsymbol{X}|\phi}(\boldsymbol{x})$. b) prikazuje korak optimizacije parametrov ϕ: pri fiksnem θ' posodobimo $\phi \to \phi'$, s čimer se premaknemo v točko z višjo vrednostjo ELBO.

3.4 Lastnosti ELBO

Funkcija ELBO, ki jo maksimiziramo med učenjem, ima nekaj pomembnih lastnosti, ki upravičujejo njeno uporabo namesto logaritemskega verjetja. Za vsak fiksen θ je ELBO funkcija parametrov ϕ in leži pod logaritemskim verjetjem. S spreminjanjem θ se ta funkcija približa ali oddalji od logaritemskega verjetja. S spreminjanjem ϕ pa se premikamo vzdolž funkcije spodnje meje (Slika 8).

3.4.1 Statistična razdalja

V nadaljevanju bomo potrebovali način za merjenje razlike med dvema verjetnostnima porazdelitvama. V ta namen vpeljemo koncept statistične razdalje.

Definicija 3 (Mera razhajanja)**.** Mera razhajanja (angl. *dissimilarity measure*) je funkcija $d(\mathcal{P}, \mathcal{Q})$, ki določa stopnjo različnosti med dvema verjetnostnima objektoma $\mathcal{P}$ in $\mathcal{Q}$. Splošna mera razhajanja med verjetnostnima gostotama zadošča pogojema:

1. $d(\mathcal{P}, \mathcal{Q}) \geq 0$,
2. $d(\mathcal{P}, \mathcal{Q}) = 0 \Leftrightarrow \mathcal{P} = \mathcal{Q}$.

Za razliko od metričnih razdalj mere razhajanja niso nujno simetrične in ne zadoščajo nujno trikotniški neenakosti [4, 5].

Cha [4] v svojem pregledu razvrsti več kot 40 različnih mer v osem družin na podlagi njihove sintaktične strukture. Za variacijske samokodirnike je ključna družina f-divergenc (tudi *Shannonova entropijska družina*). Slednja vključuje mere razhajanja, ki temeljijo na logaritmu razmerja gostot.

3.4.2 Kullback-Leiblerjeva divergenca

Kullback-Leiblerjeva divergenca (D_{KL}) je primer mere razhajanja iz družine f-divergenc.

Definicija 4 (Kullback-Leiblerjeva divergenca)**.** Naj bosta $\mathcal{Q}$ in $\mathcal{P}$ verjetnostni porazdelitvi z zveznima gostotama $q(\boldsymbol{z})$ in $p(\boldsymbol{z})$. KL-divergenca med porazdelitvama $\mathcal{Q}$ in $\mathcal{P}$ je definirana kot:

$$D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) = D_{\mathrm{KL}}(q\|p) = \int q(\boldsymbol{z}) \ln\left(\frac{q(\boldsymbol{z})}{p(\boldsymbol{z})}\right) \mathrm{d}\boldsymbol{z} = \mathbb{E}_{\mathcal{Q}}\left[\ln q(\boldsymbol{z}) - \ln p(\boldsymbol{z})\right].$$

Da je izraz dobro definiran, moramo integrirati po nosilcu porazdelitvene gostote $p(\boldsymbol{z})$ in nosilec porazdelitvene gostote $q(\boldsymbol{z})$ mora biti vsebovan v nosilcu porazdelitvene gostote $p(\boldsymbol{z})$. Ker je integral enak 0, kjer je $q(\boldsymbol{z}) = 0$, je to enako integralu po nosilcu $q(\boldsymbol{z})$.

Nenegativnost izhaja iz Jensenove neenakosti:

$$\begin{aligned}
-D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) &= -\int q(\boldsymbol{z}) \ln \frac{q(\boldsymbol{z})}{p(\boldsymbol{z})} \mathrm{d}\boldsymbol{z} \\
&= \int q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} \mathrm{d}\boldsymbol{z} \\
&\leq \ln \int q(\boldsymbol{z}) \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} \mathrm{d}\boldsymbol{z} \\
&= \ln \int p(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \\
&= \ln 1 = 0.
\end{aligned}$$

Po množenju z -1 dobimo $D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) \geq 0$.

Velja tudi, da je $D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) = 0$ natanko tedaj, ko sta porazdelitveni gostoti enaki, torej $q(\boldsymbol{z}) = p(\boldsymbol{z})$ za vse $\boldsymbol{z}$:

($\Leftarrow$) Če je $q(\boldsymbol{z}) = p(\boldsymbol{z})$ za vse $\boldsymbol{z}$, je logaritem kvocienta enak 0:

$$D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) = \int q(\boldsymbol{z}) \ln\left(\frac{q(\boldsymbol{z})}{q(\boldsymbol{z})}\right) \mathrm{d}\boldsymbol{z} = \int q(\boldsymbol{z}) \cdot 0 \, \mathrm{d}\boldsymbol{z} = 0.$$

($\Rightarrow$) Če je $D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) = 0$, v Jensenovi neenakosti velja enakost:

$$\mathbb{E}_{\mathcal{Q}}\left[\ln\left(\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}\right)\right] = \ln\left(\mathbb{E}_{\mathcal{Q}}\left[\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}\right]\right).$$

Ker je funkcija $t \mapsto \ln t$ strogo konkavna, velja enakost natanko tedaj, ko je kvocient $\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}$ konstanten:

$$\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} = c.$$

Konstanto c določimo z integracijo:

$$1 = \int p(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = \int c \cdot q(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = c\int q(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = c \cdot 1 = c \implies c = 1,$$

iz česar sledi $p(\boldsymbol{z}) = q(\boldsymbol{z})$ za vse $\boldsymbol{z}$.

Kasneje bomo potrebovali eksplicitno obliko KL-divergence med dvema normalnima porazdelitvama.

Funkcija porazdelitvene gostote za večrazsežno normalno porazdelitev dimenzije d s pričakovano vrednostjo $\boldsymbol{\mu}$ in kovariančno matriko $\boldsymbol{\Sigma}$ je podana z naslednjo formulo:

$$f_{\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}(\boldsymbol{z}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right).$$

Sedaj si oglejmo dve d-dimenzionalni normalni porazdelitvi $\mathcal{P} = \mathcal{N}(\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})$ in $\mathcal{Q} = \mathcal{N}(\boldsymbol{\mu}_{\mathcal{Q}}, \boldsymbol{\Sigma}_{\mathcal{Q}})$. KL-divergenca med tema dvema porazdelitvama je podana z naslednjo enačbo:

$$\begin{aligned}
&D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P})\\
&= \mathbb{E}_{\mathcal{Q}}[\ln q(\boldsymbol{z}) - \ln p(\boldsymbol{z})]\\
&= \mathbb{E}_{\mathcal{Q}}\left[\frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_{\mathcal{P}}|}{|\boldsymbol{\Sigma}_{\mathcal{Q}}|}\right) - \frac{1}{2}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}})^T\boldsymbol{\Sigma}_{\mathcal{Q}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}}) + \frac{1}{2}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})^T\boldsymbol{\Sigma}_{\mathcal{P}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})\right]\\
&= \frac{1}{2}\left(\mathbb{E}_{\mathcal{Q}}\left[\ln\left(\frac{|\boldsymbol{\Sigma}_{\mathcal{P}}|}{|\boldsymbol{\Sigma}_{\mathcal{Q}}|}\right)\right] - \mathbb{E}_{\mathcal{Q}}\left[(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}})^T\boldsymbol{\Sigma}_{\mathcal{Q}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}})\right] + \mathbb{E}_{\mathcal{Q}}\left[(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})^T\boldsymbol{\Sigma}_{\mathcal{P}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})\right]\right)\\
&= \frac{1}{2}\left(\ln\left(\frac{|\boldsymbol{\Sigma}_{\mathcal{P}}|}{|\boldsymbol{\Sigma}_{\mathcal{Q}}|}\right) - \mathbb{E}_{\mathcal{Q}}\left[(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}})^T\boldsymbol{\Sigma}_{\mathcal{Q}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}})\right] + \mathbb{E}_{\mathcal{Q}}\left[(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})^T\boldsymbol{\Sigma}_{\mathcal{P}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})\right]\right),
\end{aligned}$$

ker je $(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}})^T\boldsymbol{\Sigma}_{\mathcal{Q}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}}) \in \mathbb{R}$, ga lahko zapišemo kot $\mathrm{sl}\left((\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}})^T\boldsymbol{\Sigma}_{\mathcal{Q}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}})\right)$, kjer je $\mathrm{sl}(\cdot)$ operator sledi matrike:

$$\frac{1}{2}\left(\ln\left(\frac{|\boldsymbol{\Sigma}_{\mathcal{P}}|}{|\boldsymbol{\Sigma}_{\mathcal{Q}}|}\right) - \mathbb{E}_{\mathcal{Q}}\left[\mathrm{sl}((\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}})^T\boldsymbol{\Sigma}_{\mathcal{Q}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}}))\right] + \mathbb{E}_{\mathcal{Q}}\left[(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})^T\boldsymbol{\Sigma}_{\mathcal{P}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})\right]\right),$$

sedaj uporabimo $\mathrm{sl}(AB) = \mathrm{sl}(BA)$:

$$\frac{1}{2}\left(\ln\left(\frac{|\boldsymbol{\Sigma}_{\mathcal{P}}|}{|\boldsymbol{\Sigma}_{\mathcal{Q}}|}\right) - \mathbb{E}_{\mathcal{Q}}\left[\mathrm{sl}\left(\boldsymbol{\Sigma}_{\mathcal{Q}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}})(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}})^T\right)\right] + \mathbb{E}_{\mathcal{Q}}\left[(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})^T\boldsymbol{\Sigma}_{\mathcal{P}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})\right]\right),$$

in ker je $\mathrm{sl}(A) = \mathrm{sl}(A^T)$ in je $\boldsymbol{\Sigma}_{\mathcal{Q}}$ simetrična, dobimo:

$$\frac{1}{2}\left(\ln\left(\frac{|\boldsymbol{\Sigma}_{\mathcal{P}}|}{|\boldsymbol{\Sigma}_{\mathcal{Q}}|}\right) - \mathbb{E}_{\mathcal{Q}}\left[\mathrm{sl}((\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}})(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}})^T\boldsymbol{\Sigma}_{\mathcal{Q}}^{-1})\right] + \mathbb{E}_{\mathcal{Q}}\left[(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})^T\boldsymbol{\Sigma}_{\mathcal{P}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})\right]\right),$$

sled je linearen operator, zato lahko pričakovano vrednost premaknemo znotraj sledi in nesemo deterministični člen $\mathbf{\Sigma}_{\mathcal{Q}}^{-1}$ iz pričakovane vrednosti:

$$\begin{aligned}
&= \frac{1}{2}\left(\ln\left(\frac{|\mathbf{\Sigma}_{\mathcal{P}}|}{|\mathbf{\Sigma}_{\mathcal{Q}}|}\right) - \operatorname{sl}\left(\mathbb{E}_{\mathcal{Q}}\left[(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}})(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}})^T\right]\mathbf{\Sigma}_{\mathcal{Q}}^{-1}\right) + \mathbb{E}_{\mathcal{Q}}\left[(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})^T\mathbf{\Sigma}_{\mathcal{P}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})\right]\right)\\
&= \frac{1}{2}\left(\ln\left(\frac{|\mathbf{\Sigma}_{\mathcal{P}}|}{|\mathbf{\Sigma}_{\mathcal{Q}}|}\right) - \operatorname{sl}\left(\mathbf{\Sigma}_{\mathcal{Q}}\mathbf{\Sigma}_{\mathcal{Q}}^{-1}\right) + \mathbb{E}_{\mathcal{Q}}\left[(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})^T\mathbf{\Sigma}_{\mathcal{P}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})\right]\right)\\
&= \frac{1}{2}\left(\ln\left(\frac{|\mathbf{\Sigma}_{\mathcal{P}}|}{|\mathbf{\Sigma}_{\mathcal{Q}}|}\right) - \operatorname{sl}(\mathbf{I}_d) + \mathbb{E}_{\mathcal{Q}}\left[(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})^T\mathbf{\Sigma}_{\mathcal{P}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{P}})\right]\right),
\end{aligned}$$

nato razbijemo tretji člen in s podobnimi utemeljitvami dobimo:

$$\begin{aligned}
&= \frac{1}{2}\left(\ln\left(\frac{|\mathbf{\Sigma}_{\mathcal{P}}|}{|\mathbf{\Sigma}_{\mathcal{Q}}|}\right) - \operatorname{sl}(\mathbf{I}_d) + \mathbb{E}_{\mathcal{Q}}\left[(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}})^T\mathbf{\Sigma}_{\mathcal{P}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\mathcal{Q}}) + (\boldsymbol{\mu}_{\mathcal{Q}}-\boldsymbol{\mu}_{\mathcal{P}})^T\mathbf{\Sigma}_{\mathcal{P}}^{-1}(\boldsymbol{\mu}_{\mathcal{Q}}-\boldsymbol{\mu}_{\mathcal{P}})\right]\right)\\
&= \frac{1}{2}\left(\ln\left(\frac{|\mathbf{\Sigma}_{\mathcal{P}}|}{|\mathbf{\Sigma}_{\mathcal{Q}}|}\right) - \operatorname{sl}(\mathbf{I}_d) + (\boldsymbol{\mu}_{\mathcal{Q}}-\boldsymbol{\mu}_{\mathcal{P}})^T\mathbf{\Sigma}_{\mathcal{P}}^{-1}(\boldsymbol{\mu}_{\mathcal{Q}}-\boldsymbol{\mu}_{\mathcal{P}}) + \operatorname{sl}\left(\mathbf{\Sigma}_{\mathcal{P}}^{-1}\mathbf{\Sigma}_{\mathcal{Q}}\right)\right).
\end{aligned}$$

Tako dobimo končno obliko KL-divergence med dvema večrazsežnima normalnima porazdelitvama:

$$D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) = \frac{1}{2}\left(\ln\left(\frac{|\mathbf{\Sigma}_{\mathcal{P}}|}{|\mathbf{\Sigma}_{\mathcal{Q}}|}\right) - d + (\boldsymbol{\mu}_{\mathcal{Q}}-\boldsymbol{\mu}_{\mathcal{P}})^T\mathbf{\Sigma}_{\mathcal{P}}^{-1}(\boldsymbol{\mu}_{\mathcal{Q}}-\boldsymbol{\mu}_{\mathcal{P}}) + \operatorname{sl}\left(\mathbf{\Sigma}_{\mathcal{P}}^{-1}\mathbf{\Sigma}_{\mathcal{Q}}\right)\right). \tag{4}$$

Če je $\mathcal{P} = \mathcal{N}(\mathbf{0}, \mathbf{I})$, se izraz poenostavi v:

$$D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) = \frac{1}{2}\left(-\ln(|\mathbf{\Sigma}_{\mathcal{Q}}|) - d + \operatorname{sl}\left(\mathbf{\Sigma}_{\mathcal{Q}}\right) + \boldsymbol{\mu}_{\mathcal{Q}}^T\boldsymbol{\mu}_{\mathcal{Q}}\right).$$

3.4.3 Tesnost spodnje meje

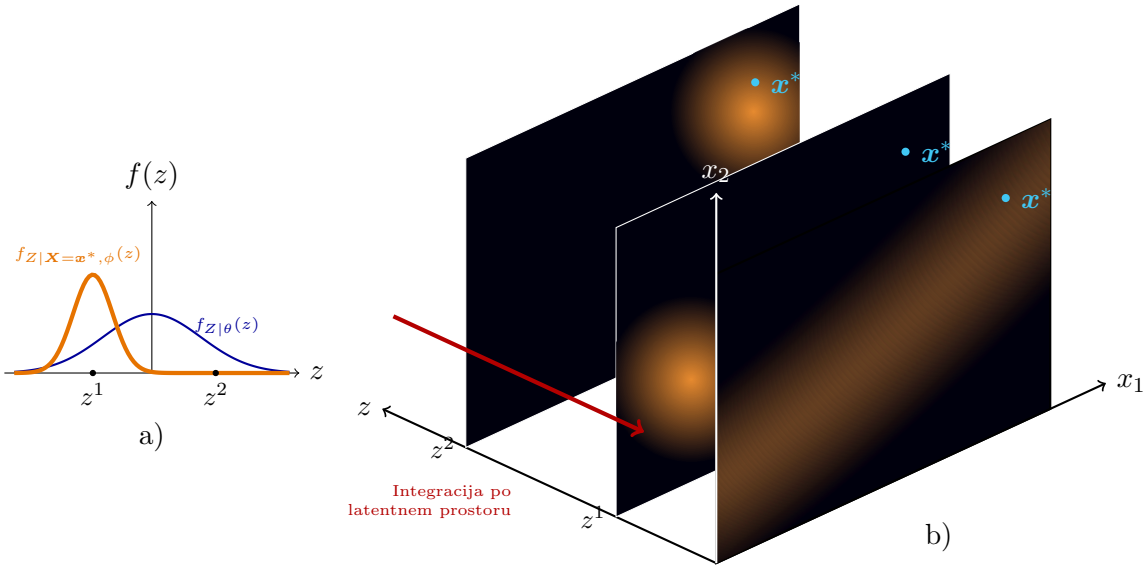
ELBO je tesna spodnja meja logaritemskega verjetja za fiksne parametre ϕ, ko ELBO in funkcija logaritemskega verjetja sovpadata. Da najdemo porazdelitveno gostoto $q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})$, pri kateri je ELBO tesna spodnja meja logaritemskega verjetja, faktoriziramo števec logaritma z uporabo definicije pogojne verjetnosti:

$$\begin{aligned}
\mathrm{ELBO}(\phi,\theta;\boldsymbol{x}) &= \int q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})\ln\left(\frac{f_{\boldsymbol{X},\boldsymbol{Z}|\phi}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})}\right)\mathrm{d}\boldsymbol{z}\\
&= \int q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})\ln\left(\frac{f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\phi}(\boldsymbol{z})f_{\boldsymbol{X}|\phi}(\boldsymbol{x})}{q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})}\right)\mathrm{d}\boldsymbol{z}\\
&= \ln\left(f_{\boldsymbol{X}|\phi}(\boldsymbol{x})\right) + \int q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})\ln\left(\frac{f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\phi}(\boldsymbol{z})}{q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})}\right)\mathrm{d}\boldsymbol{z}.
\end{aligned}$$

V zadnji vrstici opazimo definicijo KL-divergence med porazdelitvama $q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})$ in $f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\phi}(\boldsymbol{z})$, torej lahko zapišemo:

$$\mathrm{ELBO}(\phi,\theta;\boldsymbol{x}) = \ln\left(f_{\boldsymbol{X}|\phi}(\boldsymbol{x})\right) - D_{\mathrm{KL}}\left(q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})\|f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\phi}(\boldsymbol{z})\right). \tag{5}$$

Sedaj vidimo, da je ELBO enak originalnemu logaritmu verjetja, od katerega odštejemo KL-divergenco $D_{\mathrm{KL}}\left(q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})\|f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\phi}(\boldsymbol{z})\right)$. Ker je KL-divergenca nenegativna, je ELBO vedno spodnja meja logaritemskega verjetja. Spodnja meja je tesna, ko je razlika med logaritemskim verjetjem in ELBO enaka nič, torej ko se omenjeni dve porazdelitveni gostoti ujemata. $f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\phi}(\boldsymbol{z})$ je aposteriorna porazdelitvena gostota latentne spremenljivke $\boldsymbol{Z}$ ob danem opazovanem vzorcu $\boldsymbol{x}$, ki opisuje porazdelitev latentne spremenljivke, iz katere bi lahko bil generiran opazovani vzorec $\boldsymbol{x}$ (Slika 9).



Slika 9. a) Aposteriorna porazdelitvena gostota je porazdelitvena gostota $f_{Z|\boldsymbol{X}=\boldsymbol{x}^*,\phi}(z)$ latentne spremenljivke Z, ki bi lahko generirala opazovani vzorec $\boldsymbol{x}^*$. Pri izračunu uporabimo Bayesov izrek $f_{Z|\boldsymbol{X}=\boldsymbol{x}^*,\phi}(z) \propto f_{\boldsymbol{X}|Z=z,\phi}(\boldsymbol{x}^*) f_Z(z)$. b) Verjetje $f_{\boldsymbol{X}|Z=z,\phi}(\boldsymbol{x}^*)$ dobimo kot vrednost pogojne porazdelitvene gostote v $\boldsymbol{x}^*$ za fiksno vrednost latentne spremenljivke z. V prikazanem primeru je vrednost pogojne porazdelitvene gostote v $\boldsymbol{x}^*$ večja za z^1 kot za z^2, kar pomeni, da je latentna spremenljivka z^1 bolj verjeten vzrok za generacijo opazovanega vzorca $\boldsymbol{x}^*$. $f_Z(z)$ predstavlja apriorno porazdelitveno gostoto latentne spremenljivke Z. Z upoštevanjem obeh faktorjev in normalizacijo, s katero zagotovimo, da je integral porazdelitvene gostote enak ena, dobimo aposteriorno porazdelitveno gostoto $f_{Z|\boldsymbol{X}=\boldsymbol{x}^*,\phi}(z)$.

3.4.4 ELBO kot razlika rekonstrukcijske napake in KL-divergence

Enačbi (3) in (5) sta dva različna zapisa ELBO. Oglejmo si še tretji zapis, kjer bomo ELBO izrazili kot razliko med rekonstrukcijsko napako in KL-divergenco:

$$\begin{aligned}
\mathrm{ELBO}(\phi, \theta; \boldsymbol{x}) &= \int q_{\boldsymbol{Z}|\theta}(\boldsymbol{z}) \ln\left(\frac{f_{\boldsymbol{X},\boldsymbol{Z}|\phi}(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})}\right) \mathrm{d}\boldsymbol{z} \\
&= \int q_{\boldsymbol{Z}|\theta}(\boldsymbol{z}) \ln\left(f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z},\phi}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{z} + \int q_{\boldsymbol{Z}|\theta}(\boldsymbol{z}) \ln\left(\frac{f_{\boldsymbol{Z}}(\boldsymbol{z})}{q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})}\right) \mathrm{d}\boldsymbol{z} \\
&= \int q_{\boldsymbol{Z}|\theta}(\boldsymbol{z}) \ln\left(f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z},\phi}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{z} - D_{\mathrm{KL}}\left(q_{\boldsymbol{Z}|\theta}(\boldsymbol{z}) \| f_{\boldsymbol{Z}}(\boldsymbol{z})\right).
\end{aligned}$$

Prvi člen predstavlja pričakovano logaritemsko verjetje rekonstrukcije opazovanega vzorca $\boldsymbol{x}$ iz latentne spremenljivke $\boldsymbol{z}$, kar lahko interpretiramo kot negativno rekonstrukcijsko napako. Drugi člen je KL-divergenca med variacijsko porazdelitveno gostoto $q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})$ in apriorno porazdelitveno gostoto latentne spremenljivke $f_{\boldsymbol{Z}}(\boldsymbol{z})$. Tako ELBO maksimizira natančnost rekonstrukcije vzorca $\boldsymbol{x}$ iz latentne spremenljivke $\boldsymbol{z}$, hkrati pa sili aposteriorno porazdelitveno gostoto, da ostane blizu apriorni porazdelitveni gostoti latentne spremenljivke. To obliko ELBO uporabljamo pri implementaciji variacijskih samokodirnikov.

3.5 Variacijska aproksimacija

V enačbi (5) smo videli, da je ELBO tesna spodnja meja logaritemskega verjetja, ko je porazdelitvena gostota $q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})$ enaka pravi aposteriorni porazdelitveni gostoti $f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\phi}(\boldsymbol{z})$.

Za izračun aposteriorne porazdelitvene gostote uporabimo Bayesov izrek.

Izrek 5 (Bayesov izrek za zvezne spremenljivke)**.** Naj bosta $\boldsymbol{X}$ in $\boldsymbol{Z}$ zvezni slučajni spremenljivki s

skupno gostoto porazdelitve $f_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z})$. Potem za vsak $\boldsymbol{x}$, za katerega je $f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$, velja:

$$f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{z}) = \frac{f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{x}) f_{\boldsymbol{Z}}(\boldsymbol{z})}{f_{\boldsymbol{X}}(\boldsymbol{x})},$$

kjer je $f_{\boldsymbol{X}}(\boldsymbol{x}) = \int f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{x}) f_{\boldsymbol{Z}}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}$ robna porazdelitvena gostota.

Dokaz. Po definiciji pogojne porazdelitvene gostote za zvezne slučajne spremenljivke lahko skupno gostoto $f_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z})$ zapišemo na dva načina:

$$f_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z}) = f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{z}) f_{\boldsymbol{X}}(\boldsymbol{x})$$

in hkrati:

$$f_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z}) = f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{x}) f_{\boldsymbol{Z}}(\boldsymbol{z}).$$

Z enačenjem desnih strani dobimo:

$$f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{z}) f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{x}) f_{\boldsymbol{Z}}(\boldsymbol{z}).$$

Če enačbo delimo s $f_{\boldsymbol{X}}(\boldsymbol{x})$ (ob predpostavki, da je $f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$), dobimo želeno obliko Bayesovega izreka:

$$f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{z}) = \frac{f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{x}) f_{\boldsymbol{Z}}(\boldsymbol{z})}{f_{\boldsymbol{X}}(\boldsymbol{x})}.$$

Robno gostoto $f_{\boldsymbol{X}}(\boldsymbol{x})$ v imenovalcu, ki deluje kot normalizacijska konstanta, dobimo z integracijo skupne gostote po celotnem latentnem prostoru:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int f_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = \int f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{x}) f_{\boldsymbol{Z}}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}. \qquad \blacksquare$$

S pomočjo Bayesovega izreka 5 lahko zapišemo aposteriorno porazdelitveno gostoto latentne spremenljivke $\boldsymbol{Z}$ ob danem opazovanem vzorcu $\boldsymbol{x}$ kot:

$$f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\phi}(\boldsymbol{z}) = \frac{f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z},\phi}(\boldsymbol{x}) f_{\boldsymbol{Z}}(\boldsymbol{z})}{f_{\boldsymbol{X}|\phi}(\boldsymbol{x})},$$

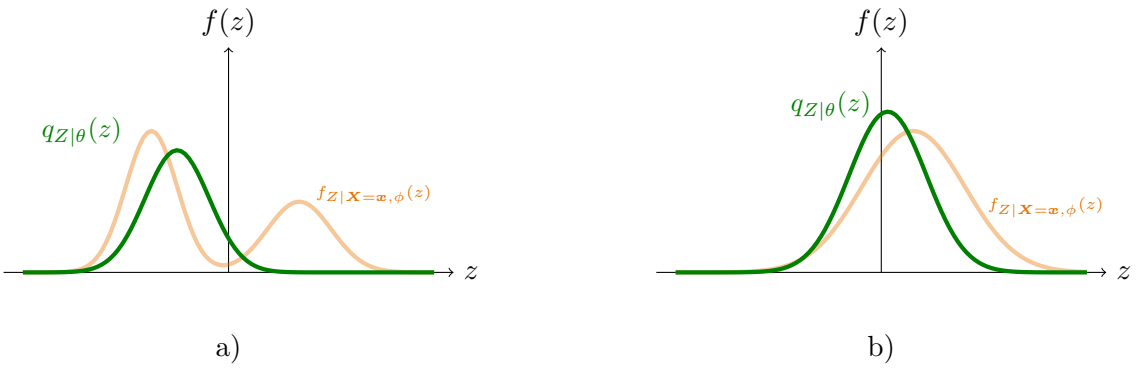
vendar pa je v praksi izračun tega izraza neizvedljiv zaradi verjetja v imenovalcu, kot smo videli v poglavju 3.3.2.

Ta problem rešimo z variacijsko aproksimacijo: izberemo preprosto parametrično obliko $q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})$ in z njo aproksimiramo aposteriorno porazdelitveno gostoto $f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\phi}(\boldsymbol{z})$. Izbrali bomo večrazsežno normalno porazdelitev s pričakovano vrednostjo $\boldsymbol{\mu}_\theta$ in diagonalno kovariančno matriko $\boldsymbol{\Sigma}_\theta$. Ta aproksimacija se ne bo vedno ujemala z aposteriorno porazdelitveno gostoto, a ji bo za nekatere vrednosti $\boldsymbol{\mu}_\theta$ in $\boldsymbol{\Sigma}_\theta$ bližje kot za druge. Med učenjem bomo našli takšne parametre θ, da bo aproksimacija $q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})$ čim bližje pravi aposteriorni porazdelitveni gostoti $f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\phi}(\boldsymbol{z})$ (Slika 10). To sovpada z minimizacijo KL-divergence med tema dvema porazdelitvenima gostotama v enačbi (5) in se odraža v izboljšanju ELBO kot funkcije spodnje meje (Slika 8).

Vemo, da je optimalna izbira za $q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})$ aposteriorna porazdelitvena gostota $f_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\phi}(\boldsymbol{z})$, ki je odvisna od podatka $\boldsymbol{x}$, torej mora biti tudi parametrična oblika $q_{\boldsymbol{Z}|\theta}(\boldsymbol{z})$ odvisna od podatka $\boldsymbol{x}$:

$$q_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\theta}(\boldsymbol{z}) = f_{\mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x},\theta),\boldsymbol{\Sigma}(\boldsymbol{x},\theta))}(\boldsymbol{z}), \tag{6}$$

kjer parametre normalne variacijske aproksimacije $\boldsymbol{\mu}$ in $\boldsymbol{\Sigma}$ določa nevronska mreža kodirnika $g_\theta(\boldsymbol{x})$ s parametri θ in vhodom $\boldsymbol{x}$.



Slika 10. Variacijska aproksimacija. Aposteriorne porazdelitvene gostote $f_{Z|\boldsymbol{X}=\boldsymbol{x},\phi}(z)$ ne moremo zapisati v zaprti obliki. Variacijska aproksimacija izbere družino porazdelitvenih gostot $q_{\boldsymbol{Z}|\theta}$ (primer z normalno) in skuša najti predstavnika, ki je najbližje aposteriorni porazdelitveni gostoti. a) Variacijska aproksimacija aposteriorne porazdelitve ni vedno enako uspešna, v primeru kompleksne multimodalne aposteriorne porazdelitve normalna družina težko ustvari primerno aproksimacijo. b) V primeru preprostejše aposteriorne porazdelitve je lahko aproksimacija zelo dobra.

3.6 Variacijski samokodirniki

Sedaj lahko končno definiramo arhitekturo variacijskega samokodirnika. Imamo vse, kar potrebujemo za izračun ELBO:

$$\mathrm{ELBO}(\phi,\theta;\boldsymbol{x}) = \int q_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\theta}(\boldsymbol{z}) \ln\left(f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z},\phi}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{z} - D_{\mathrm{KL}}(q_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\theta}(\boldsymbol{z}) \| f_{\boldsymbol{Z}}(\boldsymbol{z})),$$

kjer je $q_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\theta}(\boldsymbol{z})$ aproksimacija iz enačbe (6).

Prvi člen je še vedno neizračunljiv integral, ampak opazimo, da je to pričakovana vrednost glede na porazdelitveno gostoto $q_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\theta}(\boldsymbol{z})$, ki jo lahko ocenimo z vzorčenjem. Za vsako funkcijo $\psi(\boldsymbol{z})$ velja:

$$\mathbb{E}_{q_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\theta}(\boldsymbol{z})}[\psi(\boldsymbol{z})] = \int q_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\theta}(\boldsymbol{z})\psi(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} \approx \frac{1}{L}\sum_{l=1}^{L}\psi(\boldsymbol{z}^{(l)}),$$

kjer so $\boldsymbol{z}^{(l)}$ neodvisni vzorci iz porazdelitve s porazdelitveno gostoto $q_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\theta}(\boldsymbol{z})$. Ta metoda se imenuje integracija Monte Carlo. Zelo grobo oceno dobimo z uporabo enega vzorca $\boldsymbol{z}^*$:

$$\mathrm{ELBO}(\phi,\theta;\boldsymbol{x}) \approx \ln\left(f_{\boldsymbol{X}|\boldsymbol{Z}=\boldsymbol{z}^*,\phi}(\boldsymbol{x})\right) - D_{\mathrm{KL}}(q_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\theta}(\boldsymbol{z}) \| f_{\boldsymbol{Z}}(\boldsymbol{z})). \tag{7}$$

Drugi člen je KL-divergenca med dvema večrazsežnima normalnima porazdelitvenima gostotama dimenzije $d_{\boldsymbol{Z}}$ z diagonalnima kovariančnima matrikama, poleg tega je $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0},\mathbf{I})$, zato KL-divergenco zapišemo kot (4):

$$\begin{aligned} &D_{\mathrm{KL}}(q_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\theta}(\boldsymbol{z}) \| f_{\boldsymbol{Z}}(\boldsymbol{z})) = \\ &= \frac{1}{2}\left(\mathrm{sl}(\boldsymbol{\Sigma}(\boldsymbol{x},\theta)) + \boldsymbol{\mu}(\boldsymbol{x},\theta)^T\boldsymbol{\mu}(\boldsymbol{x},\theta) - \ln(|\boldsymbol{\Sigma}(\boldsymbol{x},\theta)|) - d_{\boldsymbol{Z}}\right). \end{aligned}$$

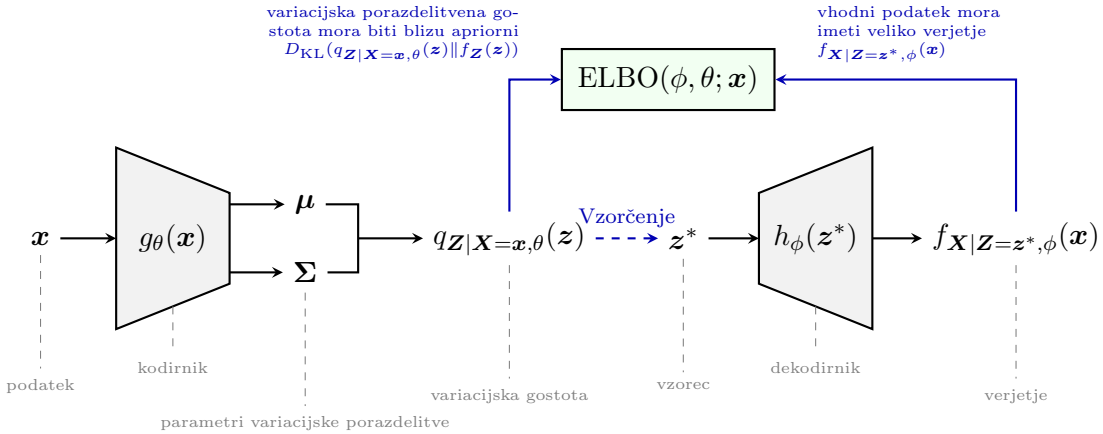
3.6.1 Algoritem VAE

Naš cilj je zgraditi model, ki za posamezen podatek $\boldsymbol{x}^{(i)}$ izračuna $\mathrm{ELBO}(\phi,\theta;\boldsymbol{x}^{(i)})$, in z optimizacijo parametrov ϕ in θ maksimizirati celotni ELBO na učni množici $\{\boldsymbol{x}^{(i)}\}_{i=1}^N$, ki je definiran kot vsota:

$$\mathrm{ELBO}(\phi,\theta) = \sum_{i=1}^{N}\mathrm{ELBO}(\phi,\theta;\boldsymbol{x}^{(i)}).$$

S tem maksimiziramo tudi spodnjo mejo logaritemskega verjetja na učni množici. Za izračun $\mathrm{ELBO}(\phi, \theta; \boldsymbol{x}^{(i)})$ pri posameznem podatku $\boldsymbol{x}^{(i)}$:

- izračunamo parametra $\boldsymbol{\mu}$ in $\boldsymbol{\Sigma}$ variacijske aposteriorne porazdelitvene gostote $q_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}^{(i)},\theta}(\boldsymbol{z})$ za podatek $\boldsymbol{x}^{(i)}$ z uporabo kodirnika $g_\theta(\boldsymbol{x}^{(i)})$,
- vzorčimo $\boldsymbol{z}^*$ iz te porazdelitve,
- izračunamo $\mathrm{ELBO}(\phi, \theta; \boldsymbol{x}^{(i)})$ z uporabo formule (7).



Slika 11. Shematski prikaz arhitekture variacijskega samokodirnika. Kodirnik $g_\theta(\boldsymbol{x})$ iz učnega podatka $\boldsymbol{x}$ napove parametre variacijske porazdelitve $\boldsymbol{\mu}$ in $\boldsymbol{\Sigma}$. Iz te porazdelitve vzorčimo $\boldsymbol{z}^*$ in z dekodirnikom $h_\phi(\boldsymbol{z}^*)$ poskušamo rekonstruirati vhodni podatek $\boldsymbol{x}$. Funkcija izgube je negativni ELBO, kjer prvi člen oceni natančnost rekonstrukcije in drugi člen oceni kako blizu je variacijska porazdelitvena gostota $q_{\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x},\theta}(\boldsymbol{z})$ apriorni porazdelitveni gostoti $f_{\boldsymbol{Z}}(\boldsymbol{z})$.

Arhitektura modela je prikazana na sliki 11. Sedaj je jasno, zakaj se model imenuje variacijski samokodirnik. Variacijski je zaradi variacijske aproksimacije aposteriorne porazdelitvene gostote latentne spremenljivke z normalno porazdelitvijo. Samokodirnik pa zato, ker ima podobno arhitekturo kot običajni samokodirnik: kodirnik $g_\theta(\boldsymbol{x})$ preslika vhodni podatek $\boldsymbol{x}$ v latentni prostor, dekodirnik $h_\phi(\boldsymbol{z}^*)$ pa iz vzorca latentne spremenljivke $\boldsymbol{z}^*$ rekonstruira vhodni podatek.

VAE izračuna ELBO kot funkcijo parametrov ϕ in θ. Maksimizacijo izvajamo s poganjanjem manjših skupkov vzorcev skozi model in posodabljanjem teh parametrov z optimizacijskim algoritmom, kot sta SGD in Adam. Gradienti ELBO glede na parametre se izračunajo kot običajno z uporabo avtomatskega odvajanja. Med tem postopkom se premikamo med obarvanimi krivuljami in vzdolž njih (Slika 8). Med tem postopkom se parametri ϕ spremenijo tako, da nelinearni model latentnih spremenljivk podatkom iz učne množice dodeli večje verjetje.

3.6.2 Reparametrizacijski trik

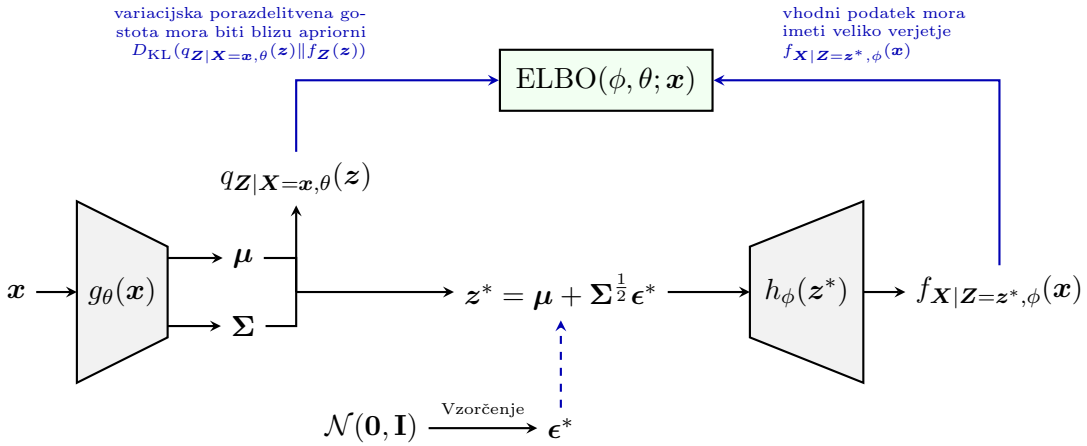
Na tej točki ostaja še ena težava: model vsebuje korak vzorčenja $\boldsymbol{Z}^* \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, ki je stohastičen in preprečuje izračun gradientov z avtomatskim odvajanjem. Odvajanje preko tega koraka je potrebno za učenje parametrov θ, ki se v modelu pojavijo pred vzorčenjem.

K sreči za to obstaja preprosta rešitev: vzorčenje prestavimo v stransko vejo modela, kjer vzorčimo $\boldsymbol{\epsilon}^*$ iz $\mathcal{N}(\mathbf{0}, \mathbf{I})$ in zapišemo:

$$\boldsymbol{z}^* = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\epsilon}^*,$$

da dobimo želeni vzorec iz variacijske aposteriorne porazdelitve. Sedaj lahko odvajamo preko celotnega modela, saj stranska veja ni odvisna od parametrov θ. Gradient se zato lahko pretaka skozi

$\boldsymbol{\mu}$ in $\boldsymbol{\Sigma}$ po deterministični transformaciji $\boldsymbol{z}^* = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}^*$, naključnost pa je izražena z neodvisno slučajno spremenljivko ϵ. Ta metoda se imenuje reparametrizacijski trik in je prikazana na sliki 12.



Slika 12. Reparametrizacijski trik: stohastični proces je izoliran v zunanji spremenljivki $\boldsymbol{\epsilon}^*$, kar omogoča nemoten pretok gradientov skozi parametre $\boldsymbol{\mu}$ in $\boldsymbol{\Sigma}$. Namesto da vzorčimo $\boldsymbol{z}^*$ neposredno iz variacijske aposteriorne porazdelitve, vzorčimo $\boldsymbol{\epsilon}^*$ iz standardne normalne porazdelitve in nato uporabimo deterministično transformacijo, da dobimo $\boldsymbol{z}^* = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\epsilon}^*$. Tako dobljeni vzorec $\boldsymbol{z}^*$ nato preslikamo z dekodirnikom $h_\phi(\boldsymbol{z}^*)$, da dobimo pogojno porazdelitev $\boldsymbol{X}|\boldsymbol{Z} = \boldsymbol{z}^*, \phi$ za rekonstrukcijsko napako – verjetje v $\boldsymbol{x}$.

Reparametrizacijski trik je pomemben del učenja variacijskih samokodirnikov, uporaba dekodirnika in vzorčenja iz latentnega prostora za potrebe generiranja novih podatkov pa ostaja taka, kot je opisana v poglavju o generiranju novih podatkov 3.2.1.

4. Zaključek

V članku smo obravnavali variacijske samokodirnike kot verjetnostni okvir za reševanje problema ocenjevanja gostote z uporabo nelinearnih modelov latentnih spremenljivk. Predstavili smo eksplicitno matematično formulacijo modela, variacijsko aproksimacijo aposteriorne porazdelitve, vlogo apriorne porazdelitve v nizkodimenzionalnem latentnem prostoru ter optimizacijo spodnje meje verjetja (ELBO), ki se neposredno naslanja na klasične statistične principe maksimizacije verjetja. Povezava med globokimi nevronskimi mrežami ter Bayesovim verjetnostnim sklepanjem odpira pot do učinkovitega generiranja novih podatkov z vzorčenjem iz zveznega latentnega prostora.

Zahvala

Zahvaljujem se mentorju prof. dr. Ljupču Todorovskem in somentorju asist. Sebastianu Mežnarju za pomoč. Zahvaljujem se tudi anonimnemu recenzentu za temeljit pregled in popravke.

Literatura

[1] S. J. D. Prince, *Understanding Deep Learning*, MIT Press, 2023, dostopno na: https://udlbook.github.io/udlbook/.

[2] D. P. Kingma in M. Welling, Auto-Encoding Variational Bayes, V *2nd International Conference on Learning Representations, ICLR 2014*, Banff, Kanada, 2014.

[3] A. P. Dempster, N. M. Laird in D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B (Methodological)*, zv. 39, št. 1, str. 1–38, 1977.

[4] S.-H. Cha, Comprehensive survey on distance/similarity measures between probability density functions, *International Journal of Mathematical Models and Methods in Applied Sciences*, zv. 1, št. 4, str. 300–307, 2007.

[5] Penn State University, *STAT 857: Dissimilarity Measures*, spletno gradivo, 2024, dostopno na: https://online.stat.psu.edu/stat857/node/3/ (dostopano 29. 1. 2026).