Aprendizaje profundo basado en la física

Semana 4.5: Introducción a modelos probabilísticos

Docente: José I. Robledo - 30/04/2026

Modelos probabilísticos

Distribuciones e incertidumbres

- No queremos modelar un valor específico, queremos saber su distribución!
- Modelar la incertidumbre inherente a los sistemas naturales
- Múltiples fuentes de variabilidad: ruido, condiciones desconocidas, interacciones no observables, o incluso la naturaleza estocástica de un sistema.
- Queremos aprender una **distribución de probabilidad**
- **Un modelo probabilístico define una distribución sobre los datos, ya sea explícita o implícitamente**

A partir de datos, el aprendizaje de una distribución de probabilidad permite capturar patrones, correlaciones y estructuras subyacentes que no son evidentes con modelos deterministas

Aprender una distribución de datos

Función de costo

- Tenemos datos $D = \{x_i\}_{i=1,\dots,n}$ y buscamos un modelo con parámetros θ tal que aprenda la distribución verdadera de los datos p,

$$p(D) \text{ o } p(x)$$

- Buscamos un modelo que *genera* datos con cierta probabilidad! Llamemos p_θ a nuestro modelo paramétrico (θ parámetros del modelo generativo):

Maxima verosimilitud

$$\theta^* = \arg\max_\theta p_\theta(D) \overset{i.i.d}{=} \arg\max_\theta \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i)$$

$$\implies \theta^* \approx \arg\max_\theta \mathbb{E}_{X \sim p(x)} \left[\log p_\theta(X)\right]$$

Maxima log-verosimilitud esperada

Aprender una distribución de datos

Modelos simples

Con $x \in \mathbb{R}^d$, podríamos modelar $p_\theta(x)$ con alguna distribución conocida

$$x \sim p_\theta(x) = \mathcal{N}(x \,|\, \mu, \Sigma), \text{ con } \theta = \{\mu, \Sigma\}$$

Ajustamos una distribución gaussiana multivariada, aprendiendo sus parámetros globales.

- Modelo **unimodal**
- Difícil capturar distribuciones complejas
- Necesitamos modelos más expresivos

Aprender una distribución de datos

Divergencia de Kullback-Leibler

- La **Divergencia de Kullback-Leibler** (KL) entre P y Q es la entropía relativa entre las dos distribuciones. Es una *medida de información* que cuantifica que tan similar a la distribución de probabilidad $P(X)$ es la distribución candidata $Q(X)$.
- Sabemos que la entropía es una medida de la información media o incerteza de una variable aleatoria X, y la podemos definir como (Shannon, 1948)

$$\mathbb{H}(P) = -\sum_{x \in X} P(x) \log P(x),$$

donde X se muestrea de la distribución P.

Aprender una distribución de datos

Divergencia de Kullback-Leibler

- Podemos definir a la divergencia-KL como

$$D_{KL}(P\,||\,Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} = -\,\mathbb{H}(P) + \mathbb{H}(P, Q)$$

Se puede interpretar como la cantidad de información extra promedio requerida para codificar los datos usando la distribución de probabilidad candidata en vez de la verdadera.

Continua: $D_{KL}(P\,||\,Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx = \mathbb{E}_{x \sim P}\left[\log \frac{p(x)}{q(x)}\right]$

Aprender una distribución de datos

De MLE a divergencia de KL

- Resulta que **maximizar la log-verosimilitud equivale a minimizar la divergencia de Kullback-Leibler** entre la distribución real y la del modelo paramétrico:

como $D_{KL}(p(x)\,|\,p_\theta(x)) = \mathbb{E}_{x\sim p(x)}\left[\log \frac{p(x)}{p_\theta(x)}\right]$, **entonces**

$$\theta^* = \arg\min_\theta D_{KL}\big(p(x)\,|\,p_\theta(x)\big) = \arg\max_\theta \mathbb{E}_{x\sim p(x)}\left[\log p_\theta(x)\right]$$

Veremos que en modelos de variables latentes no siempre podemos evaluar esta divergencia de KL y debemos recurrir a aproximaciones…

Mezcla de densidades Gaussianas

Caso sencillo: distribución categórica

- Por qué tomar una sola Gaussiana como en el modelo sencillo?
- Introducimos una variable latente $z \sim \text{Categorical}(\pi)$, con K categorías, i.e $\pi = (\pi_1, \ldots, \pi_K)$. Podemos entonces escribir

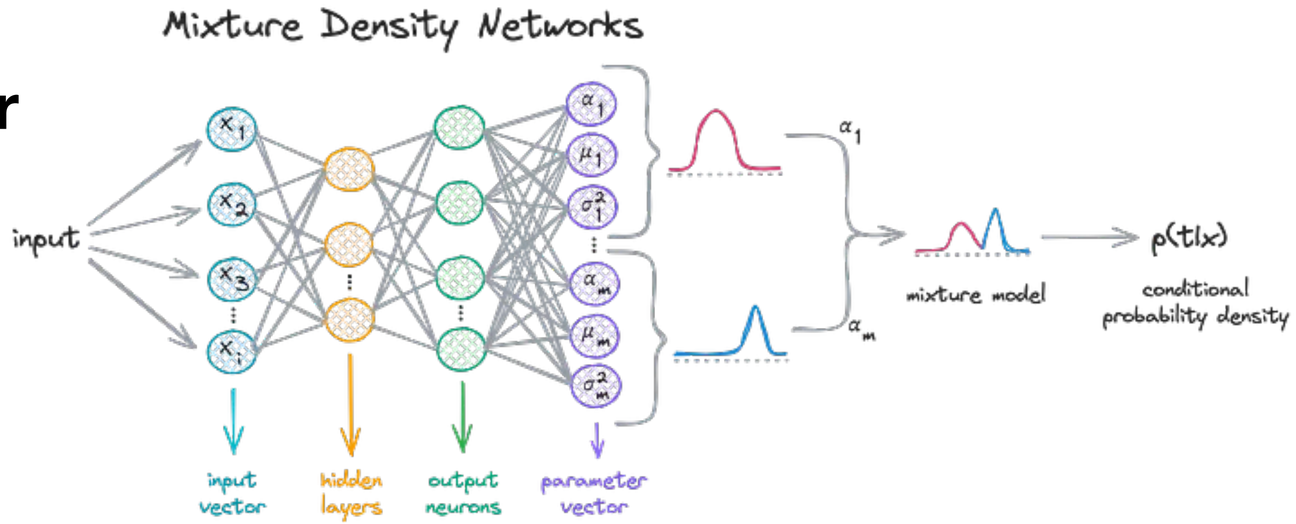
$$\log[p_\theta(x)] = \log\left[\sum_z p(z)p_\theta(x\,|\,z)\right] = \log\left[\sum_{k=1}^{K} \pi_k \mathcal{N}(x\,|\,\mu_k, \Sigma_k)\right]$$

- Parámetros a optimizar: $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1,\ldots,K}$

Mezcla de densidades Gaussianas

Red de Mezcla de Densidades (MDN)

- Supongamos que vemos una variable t que depende de x. **Queremos modelar** $p(t \mid x)$.
- Pero por qué elegimos esos valores de π_k, μ_k, Σ_k? Podemos hacer que dependan de la entrada x.



Planteamos una red neuronal que aprende

$$\pi_k(x), \mu_k(x), \Sigma_k(x)$$

$$\log[p_\theta(t \mid x)] = \log \left[\sum_{k=1}^{K} \pi_k(x) \mathcal{N}(t \mid \mu_k(x), \Sigma_k(x)) \right]$$

Podemos retropropagar porque la función es diferenciable. Funciona porque podemos sumar sobre z

Teórico-práctico MDN

Aprender una distribución de datos

Camino hacia el ELBO

RECAP

- Queremos que $p_\theta(x)$ se aproxime a la distribución real $p(x)$.
- Podemos reescribir

$$D_{KL}(p\,||\,p_\theta) = \mathbb{E}_{x\sim p(x)}\left[\log \frac{p(x)}{p_\theta(x)}\right] = \mathbb{E}_{x\sim p(x)}\left[\log p(x)\right] - \mathbb{E}_{x\sim p(x)}\left[\log p_\theta(x)\right]$$

- Si minimizamos respecto a θ, el primer termino es una constante. Por lo tanto

$$\arg\min_\theta D_{KL}(p\,||\,p_\theta) = \arg\max_\theta \mathbb{E}_{x\sim p(x)}[\log p_\theta(x)]$$

Aprender una distribución de datos

Aproximación Variacional

- $\log p_\theta(x)$ suele ser difícil de evaluar si el modelo tiene variables latentes

$$\text{Evidencia} := \log p_\theta(x) = \log \int p_\theta(x, z) dz$$

- lo que suele ser **imposible de evaluar** ya que no suele tener forma cerrada.
- Aproximación Variacional (ϕ parámetros del modelo codificador):

Introducimos Red neuronal $\quad q_\phi(z \mid x) \approx p_\theta(z \mid x)$

$$\log p_\theta(x) = \log \int q_\phi(z \mid x) \frac{p_\theta(x, z)}{q_\phi(z \mid x)} dz = \log \mathbb{E}_{z \sim q_\phi(z|x)} \left[\frac{p_\theta(x, z)}{q_\phi(z \mid x)} \right] \geq \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \frac{p_\theta(x, z)}{q_\phi(z \mid x)} \right]$$

Multiplicamos y dividimos por $q_\phi(z \mid x)$

Desigualdad de Jensen

Aprender una distribución de datos

Aproximación Variacional

- Evidence Lower Bound:

$$ELBO(\phi, \theta; x) := \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \frac{p_\theta(x, z)}{q_\phi(z \mid x)} \right]$$

- De hecho, vamos a ver que la diferencia entre la evidencia y su cota inferior, es la divergencia de Kullback-Leibler!!

$$\log p_\theta(x) = D_{KL}(q_\phi(z \mid x) \,||\, p_\theta(z \mid x)) + ELBO(\phi, \theta; x)$$

Como la evidencia es constante respecto a ϕ para un x dado, minimizar D_{KL} es equivalente a maximizar $ELBO$

Aprender una distribución de datos

Inferencia Variacional

$$
\begin{aligned}
D_{KL}(q_\phi(z\,|\,x)\,||\,p_\theta(z\,|\,x)) &= \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{q_\phi(z\,|\,x)}{p_\theta(z\,|\,x)}\right] \\
&= \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log q_\phi(z\,|\,x)\right] - \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(z\,|\,x)\right] \\
&= \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log q_\phi(z\,|\,x)\right] - \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{p_\theta(x)}\right] \\
&= \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log q_\phi(z\,|\,x)\right] - \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x,z)\right] + \log p_\theta(x) \\
&= -\,ELBO(\phi,\theta;x) + \log p_\theta(x) \\
\implies \log p_\theta(x) &= ELBO(\phi,\theta;x) + D_{KL}(q_\phi(z\,|\,x)\,||\,p_\theta(z\,|\,x))
\end{aligned}
$$

Podemos usar el ELBO como función de costo y maximizarlo

Aprender una distribución de datos

Inferencia Variacional

$$ELBO(\phi, \theta; x) := \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \frac{p_\theta(x, z)}{q_\phi(z \mid x)} \right] = \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log p_\theta(x, z) \right] - \mathbb{E}_{z \sim q_\theta(z|x)} \left[\log q_\phi(z \mid x) \right]$$

$$= \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log p_\theta(x \mid z) \right] + \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log p(z) \right] - \mathbb{E}_{z \sim q_\theta(z|x)} \left[\log q_\phi(z \mid x) \right]$$

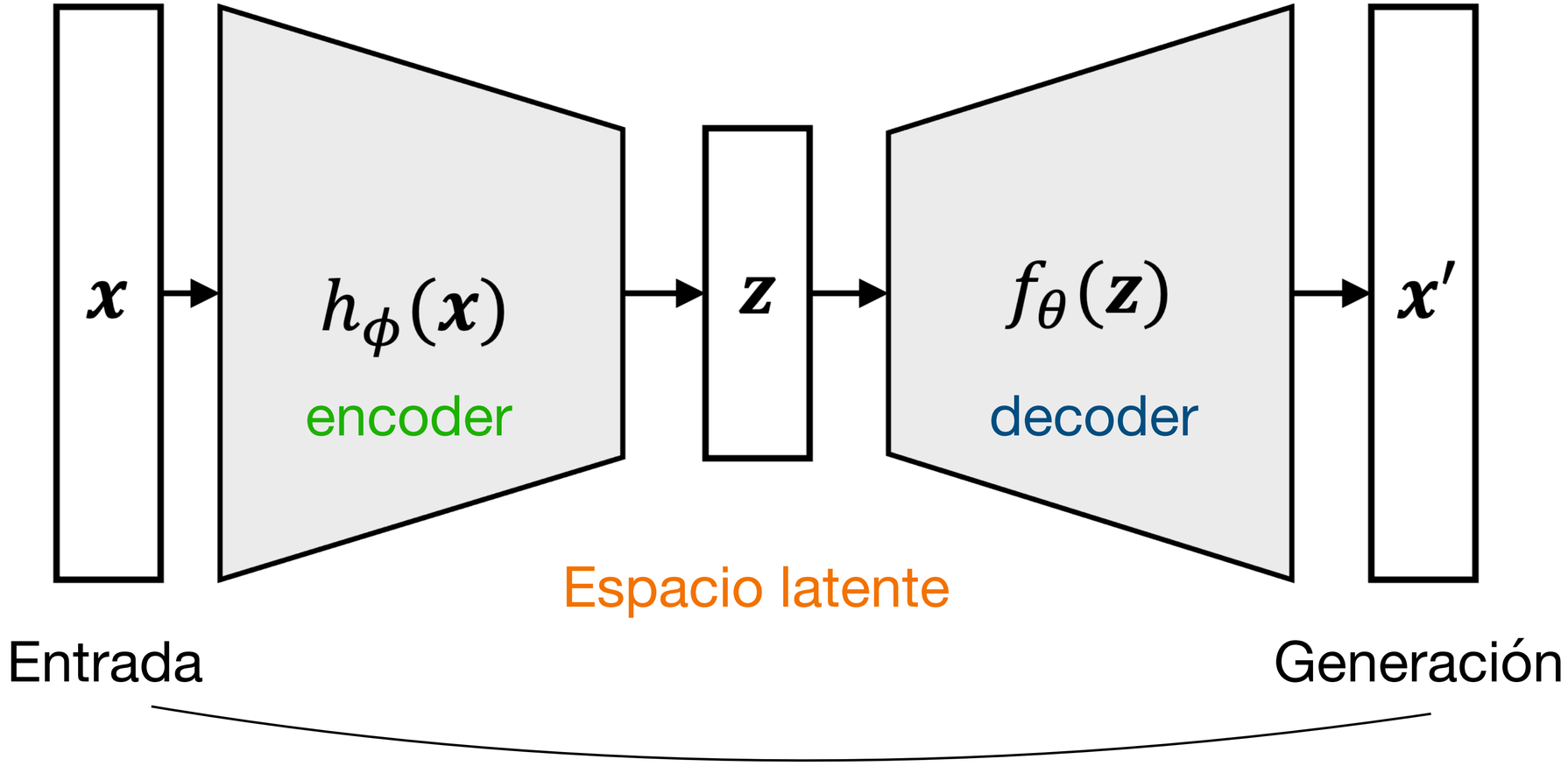
$$= \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log p_\theta(x \mid z) \right] - D_{KL}(q_\phi(z \mid x) \,||\, p(z))$$

Reconstrucción Regularización

Podemos usar el ELBO como función de costo y maximizarlo

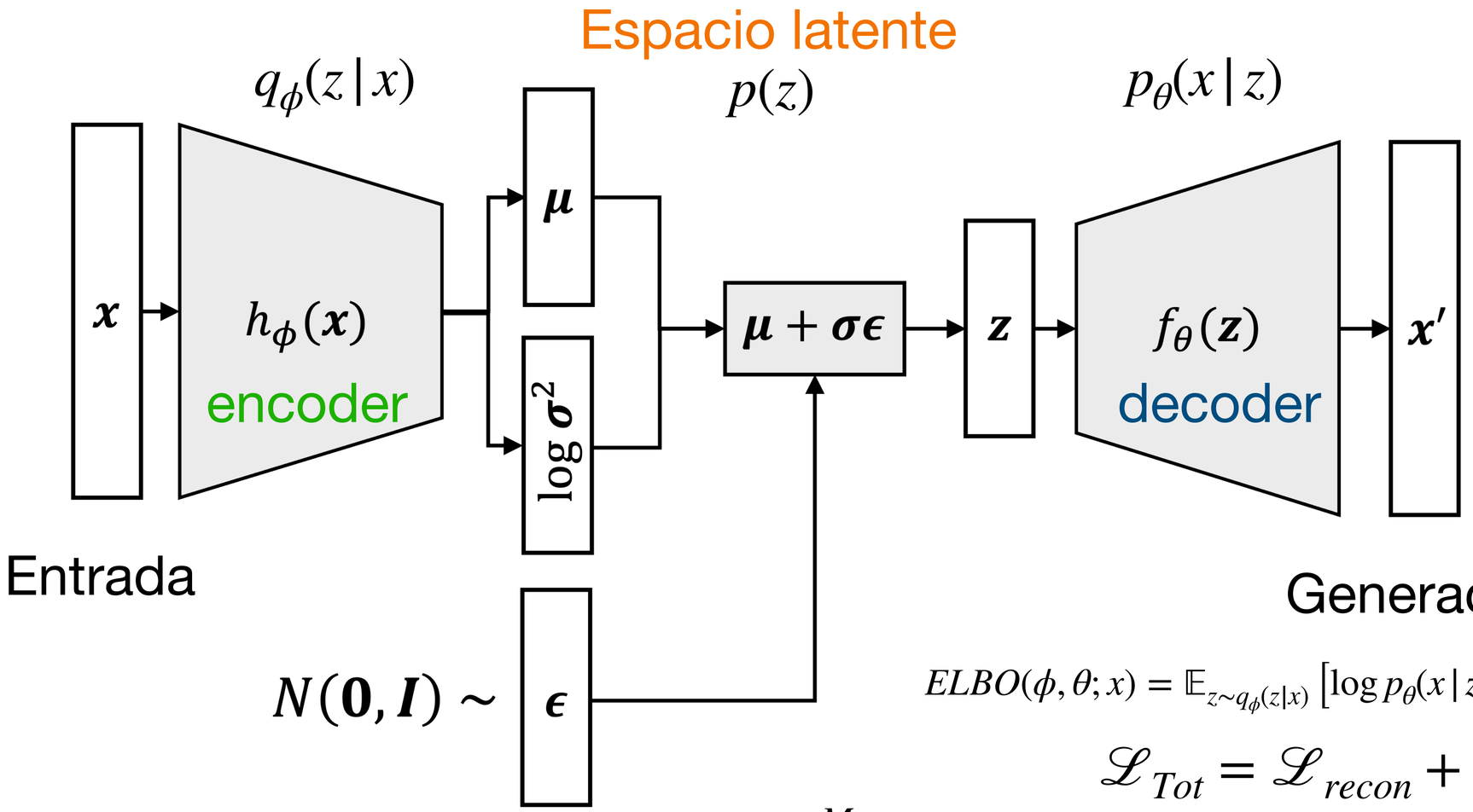
Hacia el Autoencoder Variacional

Autoencoder (AE)



$$MSE(\theta, \phi, x) = \frac{1}{N} \sum_i (x - f_\theta(h_\phi(x)))^2 = \mathscr{L}_{recon}$$

Autoencoder Variacional (VAE)



$$ELBO(\phi, \theta; x) = \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log p_\theta(x|z) \right] - D_{KL}(q_\phi(z|x) || p(z))$$

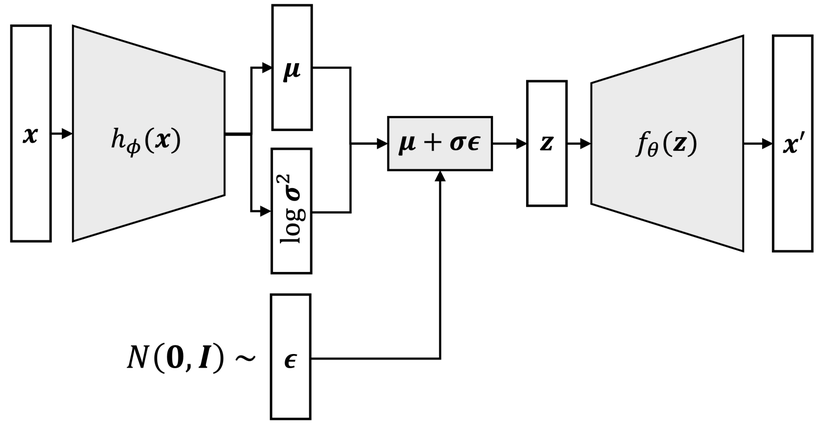
$$\mathscr{L}_{Tot} = \mathscr{L}_{recon} + \mathscr{L}_{KL}$$

$$\mathscr{L}_{KL} = KL(h_\phi(\vec{z}|\vec{x}) || p(\vec{z})) = -\frac{1}{2} \sum_{m=1}^{M} (1 + \log \sigma_\phi(\vec{x})_m^2 - \mu_\phi(\vec{x})_m^2 - \exp(\log \sigma_\phi(\vec{x})_m^2))$$

https://jorobledo.github.io/apbf/modulo_4.html

Autoencoder Variacional (VAE)

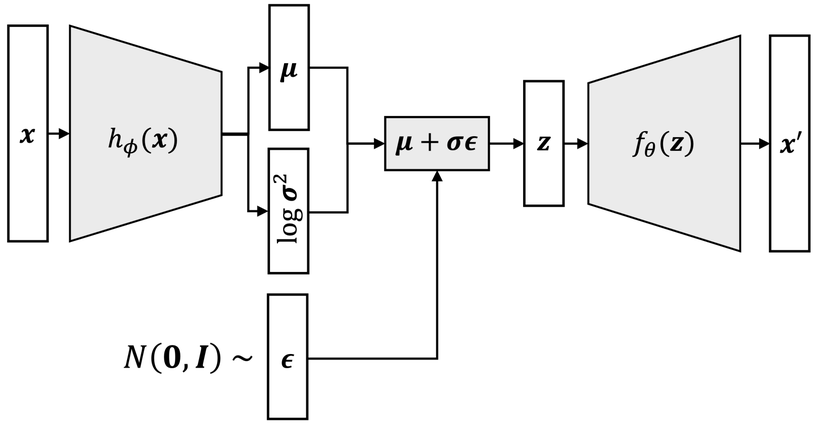
Término de la divergencia de KL e hipótesis



- Típicamente $z \in \mathbb{R}^M$ y se utiliza distribuciones Gaussianas
- $q_\phi(\vec{z}\,|\,\vec{x}) = \mathcal{N}(\mu_\phi(\vec{x}), \text{diag}\left[\sigma^2_\phi(\vec{x})\right])$ (aprox. Variacional a posterior que lleva a espacio latente)
- $\mu_\phi(\vec{x})$ y $\sigma^2_\phi(\vec{x})$ son la salida de $h_\phi(\vec{x})$ ($2M$ valores)
- $p(\vec{z}) = \mathcal{N}(0,\mathbf{I})$ (a priori para $\vec{z}$)
- La distribución del decoder dependerá de la naturaleza de los datos $\vec{x}$ (podría ser Normal, Bernoulli, Categórica, etc.) Si es Normal, utilizaríamos a MSE como $\mathscr{L}_{recon}$, para Bernoulli BCE, para Categórica entropía cruzada, etc.
- f_θ recibe $\vec{z}$ de entrada, pero $h_\phi(\vec{x})$ produjo medias y varianzas de la distribución multivariante de $\vec{z}$. Además, cómo retropropagamos después el gradiente?

Autoencoder Variacional (VAE)

Reparametrización



- Usamos truco llamado **reparametrización:**
- Muestreamos un ruido aleatorio de distribución conocida $\epsilon \sim \mathcal{N}(0,\mathbb{I})$. Como $\vec{z}\,|\,\vec{x} \sim \mathcal{N}(\vec{\mu}_\phi(\vec{x}), \sigma^2_\phi(\vec{x})\mathbb{I})$, entonces podemos utilizar

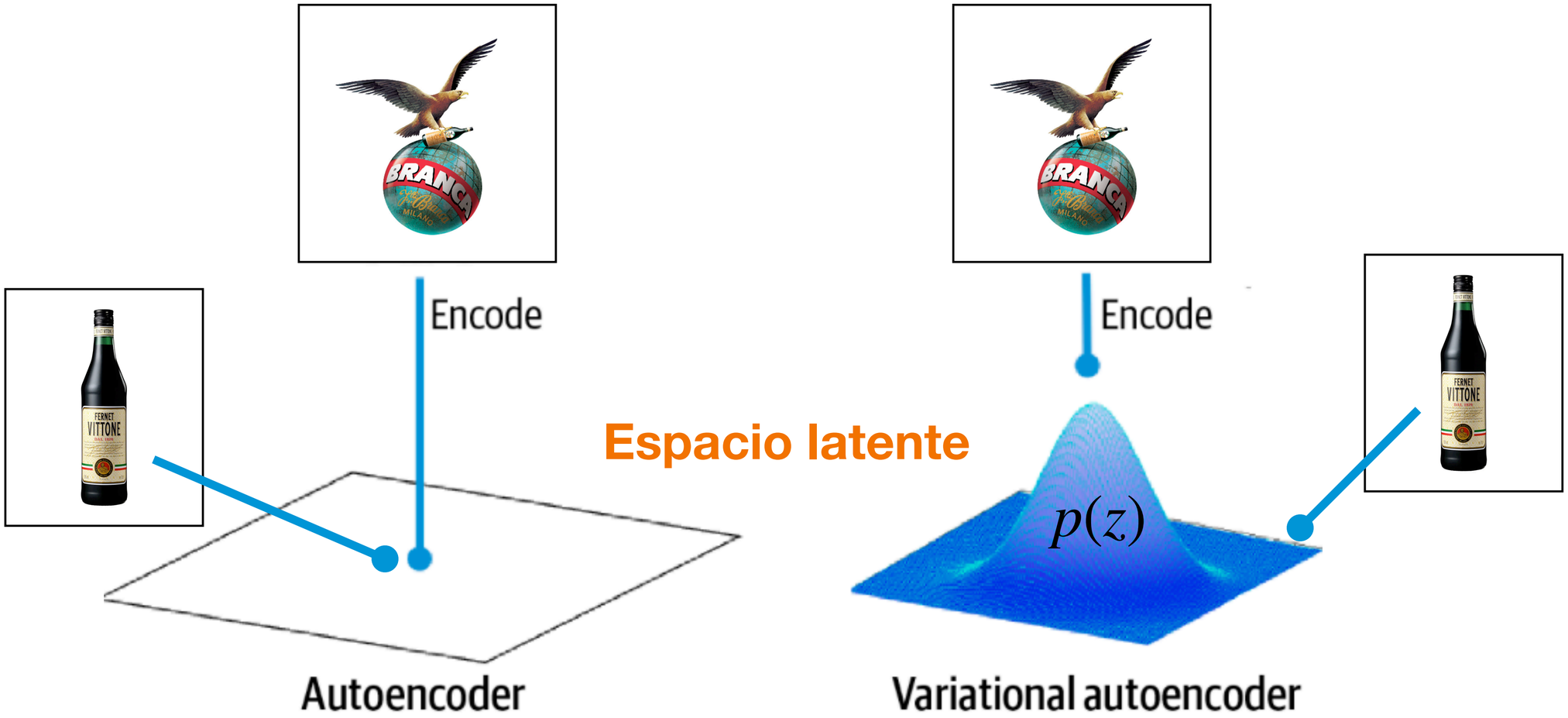
$$\vec{z} = \vec{\mu}_\phi + \vec{\sigma}_\phi \cdot \epsilon.$$ (importante, ϵ no depende de ϕ)

- En ELBO, todos los términos $\mathbb{E}_{z\sim q_\phi(z|x)}[\,\cdot\,]$ pasan a ser $\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbb{I})}[\,\cdot\,]$
- Permite usar backprop!

Ya estamos!!!!! (Creo)

Autoencoder Variacional (VAE)

Diferencia conceptual entre AE y VAE



A practicar: Notebook VAE