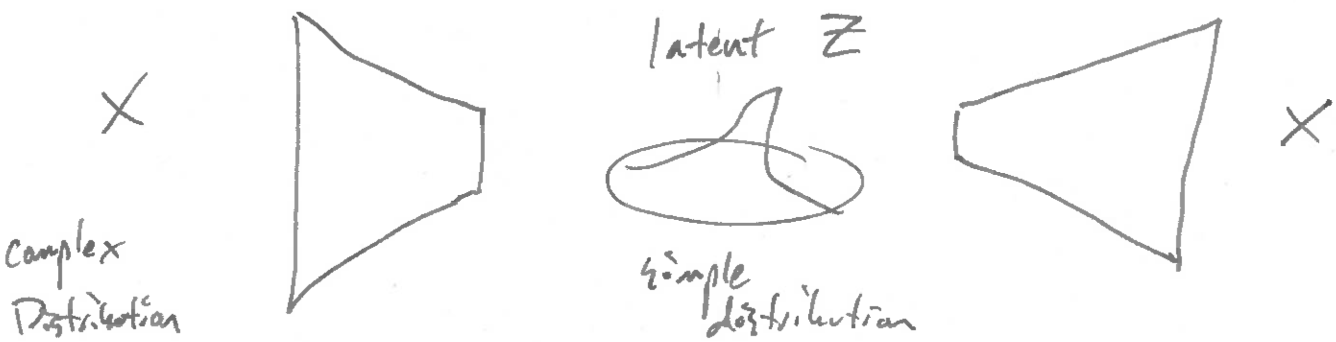
Today: Generative Modeling + Variational Inference

- How to solve for "latent" physics?
- How to use probability to perform generative modeling



<u>Idea</u> Use an identity map to sample from the data distribution

Today VAE → Next Diffusion models

Variational Inference

A new kind of probabilistic variational method to optimize over prob. dist. instead of mechanics

Some probability basics

For cont RV in $\mathbb{R}$

- probability density function

Given cont. RV X, the prob X takes a value ~~$x \in b$~~ $x \in [a, b]$

$$\mathbb{P}(a \leq X \leq b) = \int_a^b p(x)\, dx$$

for $p(x) \geq 0$

$$\int_{-\infty}^{\infty} p(x) = 1$$

- expectation

$$\mathbb{E}_p[f(x)] = \int_{-\infty}^{\infty} f(x)\, p(x)\, dx$$

sometimes drop if we are only talking about a specific dist.

<u>Ex</u> Multivariate Gaussian (our bread + butter)

Given $\vec{x} \in \mathbb{R}^d$, $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$ invertible SPD

$$p(x) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left(-\tfrac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

$$\underline{X} \sim N(x; \mu, \Sigma)$$

Gaussians are like the piecewise polynomials of probability
- easy to work with

<u>def</u> Joint, marginal, conditional distributions

<u>joint</u> $p(x,z)$ - (prob of $x \wedge z$)

<u>marginal</u> $p(x) = \int p(x,z)\, dz$

(account for all possible values of z)

<u>conditional</u> $p(z|x) = \dfrac{p(x,z)}{p(x)}$ if $p(x) > 0$

(prob of z if x happened)

Bayes theorem

$$p(x,z) = p(x|z)\,p(z) = p(z|x)\,p(x)$$

$$\Rightarrow \quad p(x|z) = \frac{p(z|x)\,p(x)}{p(z)}$$

use this to flip "input"/"output" relationships

Finally, a "pseudo" metric on distributions

KL-divergence

$$KL(q\|p) = \int q(x) \log\left(\frac{q(x)}{p(x)}\right) dx$$

- Note $KL(q\|p) \neq KL(p\|q)$
- Positive

For Gaussians $q \sim N(\mu_q, \Sigma_q)$, $p \sim N(\mu_p, \Sigma_p)$

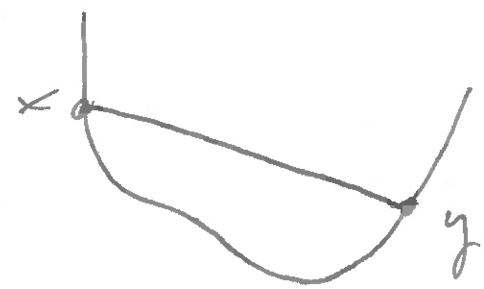
$$KL(q\|p) = \frac{1}{2}\left[\log \frac{|\Sigma_p|}{|\Sigma_q|} - d + tr(\Sigma_p^{-1}\Sigma_q) + (\mu_p - \mu_q)^T \Sigma_p^{-1} (\mu_p - \mu_q)\right]$$

Jensen's Inequality

Let ϕ be a convex function

$$\forall t \in [0,1], \quad x, y$$

$$\phi(tx + (1-t)y) \le t\,\phi(x) + (1-t)\,\phi(y)$$



Jensen $\quad \phi\,\mathbb{E}[x] \le \mathbb{E}[\phi(x)]$

To sample from data distribution

$$P(z|X) = \frac{p(x|z)\,p(z)}{p(x)} = \frac{p(x|z)\,p(z)}{\underbrace{\int p(x|z)\,p(z)\,dz}_{\text{Computationally Intractable}}}$$

Typically, we would do MLE
i.e. pose a joint dist and solve

$$\min_{\theta} -\log p_{\theta}(x, z)$$

but we don't know z

Instead, build an objective that
accounts for any possible dist. on z

Marginal log likelihood

$$\mathcal{L}_{MLL} = -\sum_d \log p(X_d ; \theta)$$

Marginalize

$$= -\sum_d \log \sum_{z_d} p(x_d, z_d ; \theta)$$

Introduce an arbitrary dist. on z, $q(z)$

$$= -\sum_d \log \sum_{z_d} q(z_d) \frac{p(x_d, z_d ; \theta)}{q(z_d)}$$

Interpret as expectation

$$= -\sum_d \log \mathbb{E}_{z\sim q}\left[\frac{p(x_d, z_d ; \theta)}{q(z_d)}\right]$$

Jensen's inequality

$$\leq -\sum_d \mathbb{E}_{z\sim q}\left[\log \frac{p(x_d, z_d ; \theta)}{q(z_d)}\right]$$

$$\mathcal{E}(\theta, q) = -\sum_d \mathbb{E}_{z\sim q}\left[\log p(x_d, z_d ; \theta)\right] - H(q_d)$$

Evidence Lower Bound (ELBO)

entropy of q

Note Don't need to plug z's in.

Since $\mathcal{L}_{MLL} \leq \mathcal{E}(\theta, q)$

We can choose q to make $\mathcal{E}$ as close as possible to $\mathcal{L}_{MLL}$

Rewriting:

$$\mathcal{E}(\theta q) = \sum_d \mathbb{E}_{z \sim q}\left[\log\left(\frac{p(z_d \mid x_d;\theta)\ p(x_d;\theta)}{q_d}\right)\right]$$

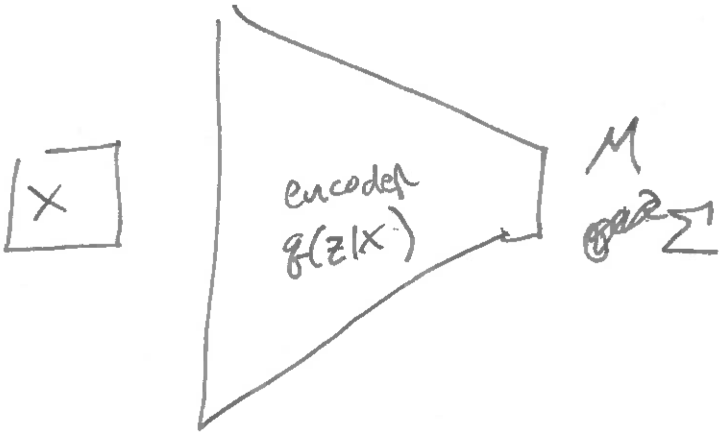
$$= \sum_d \mathbb{E}_{z \sim q}\left[\log \frac{p(z_d \mid x_d;\theta)}{q}\right] + \mathbb{E}_{z \sim q}\left[\log p(x_d;\theta)\right]$$

$$= \sum_d -KL\left(p(z_d \mid x_d;\theta) \,\|\, q_d\right) + \quad "$$

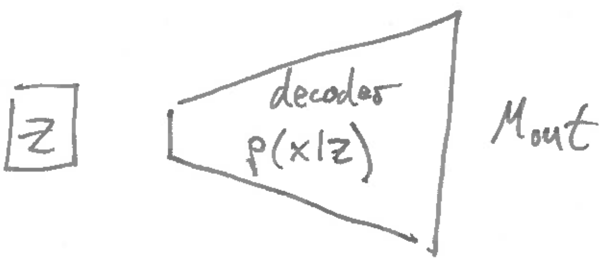
Biggest when $KL = 0$

Choose $q_d = p(z_d \mid x_d;\theta)$

Variational Autoencoder (VAE) Kingma + Welling



$$z = \mu + \sqrt{\Sigma}\,\varepsilon$$
$$\varepsilon \sim N(0, I)$$
$$\Rightarrow z \sim N(\mu, \Sigma)$$



$$p(x|z) = N(\mu_{out}, I)$$

$$\mathcal{E} = \underbrace{-\mathbb{E}\left[\log p(x|z)\right]}_{C + \|x - \mu_{out}\|^2} - KL\left(q(z|x) \,\|\, p(z)\right)$$

Reconstruction Loss

Prior penalty

Choices: → Architecture
→ Prior

Some building blocks

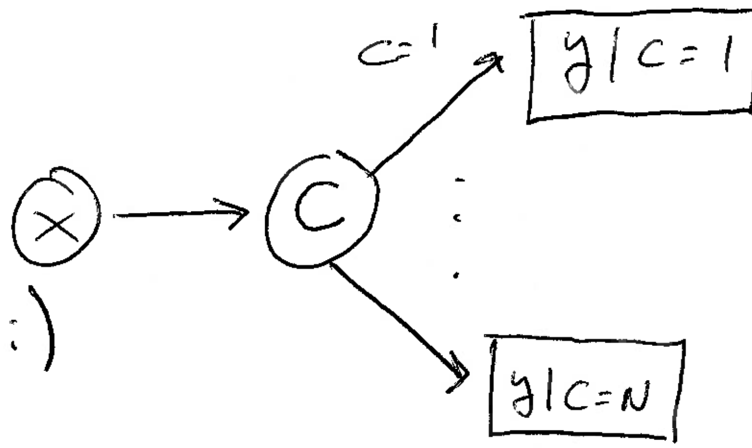
def Categorical RV $\quad c \sim \text{cat}(\pi)$

$$\pi_i > 0$$

$$\sum_i \pi_i = 1$$

Mixture of Experts (Jacobs, Jordan, Nowlan, Hinton 1991)

$$\pi_i(x) = \text{softmax}(NN(x))$$



$$P(y) = \sum_i P(c=i)\, P(y|c=i)$$

- A means to sparsely increase model param w/o increasing compute time
  (Switch Transformers: Scaling to trillion param models … Fedus 2022)

Lemma Product of Gaussian PDFs is a Gaussian

$$P_1 = N(\mu_1, \sigma_1^2)$$

$$P_2 = N(\mu_2, \sigma_2^2)$$

$$P_1 \cdot P_2 = \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \exp\left(-\frac{(x-\tilde{\mu})^2}{\tilde{\sigma}^2}\right)$$

$$\frac{\tilde{\mu}}{\tilde{\sigma}^2} \cancel{\frac{\tilde{\mu}}{\tilde{\sigma}^2}} = \frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}$$

$$\tilde{\sigma}^2 = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$$

Product-of-experts