4/28 - Probabilistic physics w/ variational inference

- Last day of class → wed 4/28
  - → everyone should be making progress
  - → last chance for feedback
  - → final presentations 5/9 3-5 in DRLB-A2
  - → report due by 5/12
- Final Course Feedback
  - Please be honest - first time class and feedback will drastically reshape next iteration
    - → More HW
    - → Stricter pre-reqs

Recall ELBO from last time

$$\mathcal{E} = -\mathbb{E}_{x \sim z}\left[\log p(x|z)\right] - KL\left(q(z|x)\,\|\,p(z)\right)$$

We discussed Kingma + Welling "vanilla" VAE:

single-sample Monte Carlo ① $\mathbb{E}_{x \sim z}\left[\log p(x|z)\right] \approx \log p(x_d | z_d)$

② $p(z) = N(\mu = 0, \Sigma = I)$

Which allowed generative modeling

① Sample $z \sim N(0, I)$

② Decode using $p(x|z)$

There is a craft to designing architectures around different choices of embeddings z, and priors

GOAL

Computationally tractable ELBO!

Gaussian Gaussian Gaussian!

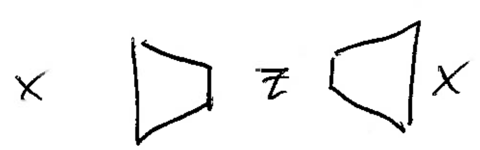
Two examples

- Denoising Diffusion
- Physical Priors

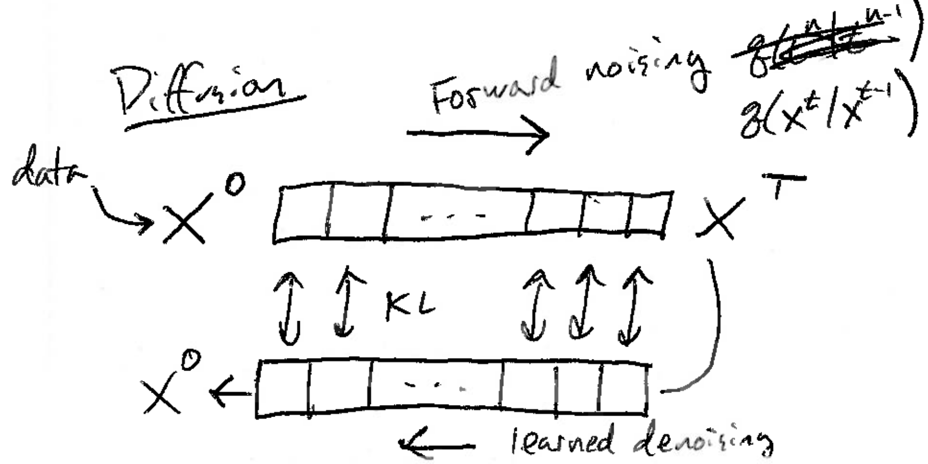
Denoising Diffusion (Sohl-Dickstein)

Idea Don't jump from x to z, instead decode in increments

Vanilla



Diffusion



Ingredients

Forward/Noising

$q(x_0)$ – data distribution

$$q(x^t | x^{t-1}) = T(x^t | x^{t-1}; \beta_t)$$

a kernel to add noise

we'll restrict to simple → $= N(x^t; x^{t-1}\sqrt{1-\beta_t}, \beta_t I)$

which is equiv to $\quad x_t = \sqrt{1-\beta_t}\, x_{t-1} + \sqrt{\beta_t}\, \varepsilon$ ← $N(0,I)$

And we get joint dist $\quad q(x^0, \dots x^T) = q(x^0) \prod_{t=1}^{T} q(x^t | x^{t-1})$ ← gaussian increments

Lemma $\quad \lim_{T\to\infty} x_t = N(0,I)$

Note Before we prove this lemma, note that we get a unit gaussian embedding by construction instead of using KL to get it by penalty

Pf To prove by induction, check 2 steps

$$x_t = \sqrt{1-\beta_t}\, x_{t-1} + \sqrt{\beta_t}\, \varepsilon_t$$

$$x_{t+1} = \underbrace{\sqrt{1-\beta_{t+1}}\sqrt{1-\beta_t}}_{\text{products}}\, x_{t-1} + \underbrace{\sqrt{\beta_{t+1}\beta_t}\, \varepsilon_t + \sqrt{\beta_{t+1}}\, \varepsilon_{t+1}}_{\text{Additive gaussian noise w/ weights}}$$

Let $\alpha_s = 1-\beta_s$

$$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$$

By induction can show

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sum_{s=1}^{t} \sqrt{\bar{\alpha}_{s-1}\beta_s}\, \varepsilon_s$$

For gaussians, variance of sum is sum of var

$$x_t \sim N\left(\sqrt{\bar{\alpha}_t}\, x_0,\ \sum_{s=1}^{t} \bar{\alpha}_{s-1}\beta_s I\right)$$

Lemma $\displaystyle\sum_{s=1}^{t} \bar{\alpha}_{s-1}\beta_s = 1 - \bar{\alpha}_t$

Pf

$$\bar{\alpha}_s = \bar{\alpha}_{s-1}\alpha_s = \bar{\alpha}_{s-1}(1-\beta_s) \rightarrow \bar{\alpha}_{s-1}\beta_s = \bar{\alpha}_{s-1} - \bar{\alpha}_s$$

So that

$$\sum_s \bar{\alpha}_{s-1}\beta_s = \sum_s \bar{\alpha}_{s-1} - \bar{\alpha}_s = 1 - \bar{\alpha}_t$$

telescoping series

Finally, take limits

Let $0 < \beta_s << 1$

$$\lim_{t\to\infty} |\bar{\alpha}_t| = \lim_{t\to\infty} \left| \prod_{s=1}^{t} (1-\beta_s) \right|$$

$$\leq \lim_{t} \left| \prod_{s=1}^{t} (1-\beta_{min}) \right|$$

$$\leq \lim_{t} (1-\beta_{min})^t$$

$$= 0$$

So $$\lim_{t\to\infty} N\left(\sqrt{\bar{\alpha}_t}\, x_0,\ \sum_{s=1}^{t} \bar{\alpha}_{s-1} \beta_s I\right)$$

$$= \lim_{t\to\infty} N\left(\sqrt{\bar{\alpha}_t}\, x_0,\ 1-\bar{\alpha}_t\right)$$

$$= N(0, I)$$

$\blacksquare$

~~Back to forward ingredient~~

$$\cancel{q(x^0, \ldots x^T) = q(x^0) \prod_{t=1}^{T} q(x^t | x^{t-1})}$$

gaussian!

Ingredient 2

$p(x^T) =$ prior distribution

$$p(x^0, \dots x^T) = p(x^T) \prod_{t=1}^{T} p(x^{t-1} | x^t)$$

going backward in time now

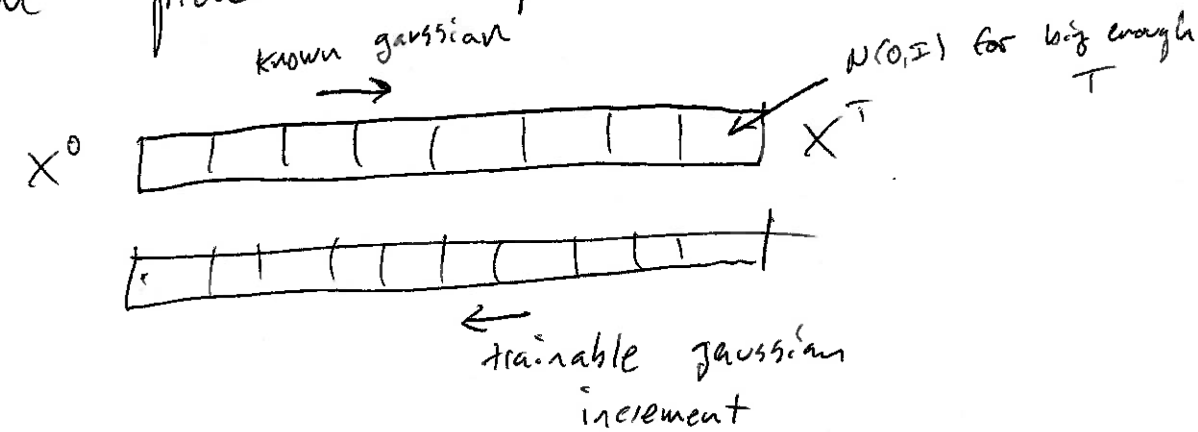
Claim For small β, $p(x^{t-1}|x^t)$ is Gaussian

we will prove this as justification for the parameterization

$$p(x^{t-1}|x^t) = N\left(\mu = f(x^t; \theta_f),\ \Sigma = g(x^t; \theta_g)\right)$$

Learnable NN's

Before we prove recap where we're at



Final step will be to make steps match w/ KL

We'll finally show

$\cancel{q(x_{t-1}|x_t, x_0) = q}$

$$q(x_{t-1}|x_t) = q(x_{t-1}|x_t, x_0) + O(\beta_t)$$

Justification of reverse transition being Gaussian

By Bayes $q(x^{t-1} \mid x^{t}, x^{0}) = \dfrac{q(x^{t} \mid x^{t-1}, x^{0})\, q(x^{t-1} \mid x^{0})}{q(x^{t} \mid x^{0})}$

From forward $q(x^{t} \mid x^{t-1}) = N(\sqrt{1-\beta_t}\, x_{t-1}, \beta_t I)$

As we derived already $q(x^{t-1} \mid x^{0}) = N(\sqrt{\bar{\alpha}_{t-1}}\, x_0, (1-\bar{\alpha}_{t-1}) I)$

product of Gaussians is an un-normalized Gaussian for some μ^*, Σ^*, C that we <u>could</u> derive

$\Rightarrow \quad q(x^{t} \mid x^{t-1})\, q(x^{t-1} \mid x^{0}) \sim C\, N(\mu^*, \Sigma^*)$

Skipping details

$$q(x_{t-1} \mid x_t, x_0) = N(\mu_t^*(x_0, x_t), \beta_t^* I)$$

$$\mu^* = \frac{\sqrt{\bar{\alpha}_{t-1}}\, \beta_t}{1-\bar{\alpha}_t} x_0 + \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t$$

$$\beta^* = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}$$

Q - ~~How do they depend on β_t when approximating $q(x_{t-1} \mid x_t)$?~~

To remove the conditioning on X_0 we'll use the laws of total expectation + variance:

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$$

$$\mathrm{var}[X] = \mathbb{E}[\mathrm{var}[X|Y]] + \mathrm{var}[\mathbb{E}[X|Y]]$$

On the mean

$$\mathbb{E}[x_{t-1}|x_t] = \mathbb{E}[\mathbb{E}[x_{t-1}|x_t, x_0]]$$

$$= \mathbb{E}[\mu^*] \sim \mathcal{O}(\beta_t^2)$$

On the variance

$$\mathrm{var}[x_{t-1}|x_t] = \mathbb{E}[\mathrm{var}(x_{t-1}|x_t, x_0)] + \mathrm{var}[\mathbb{E}[x_{t-1}|x_t, x_0]]$$

$$= \underbrace{\mathbb{E}[\beta^* I]}_{\mathcal{O}(\beta_t^2)} + \underbrace{\mathrm{var}[\mu^*]}_{\mathcal{O}(\beta_t)}$$

(we're skipping the estimates on β_t)

~~To summarize~~

Final Ingredient #3 - the ELBO

$$\mathcal{E} = \mathbb{E}_{q(x^T|x^0)}\left[\log p(x_0|x_T)\right] - \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\left[KL\left(q(x_{t-1}|x_t,x_0) \,\|\, p(x_{t-1}|x_t)\right)\right]$$

- Learn to denoise single diffusion steps
- Generate by sampling a gaussian + denoising

Some takeaways

- Designed an encoding which
  - required no learning
  - took us to a latent gaussian
  - has gaussian increments
- Designed a decoding which
  - is indirectly supervised by encoding
  - gives gaussian increments
- When combined in ELBO
  - Lots of closed form expressions for KL's

How can you use this in your research?
(How much math do you need to track?)

V1 Off-the-shelf
- just grab one of the implementations
- understand that if you mess w/ the architecture, you may break the theory
  - maybe leave ELBO loss alone
  - don't play w/ input/output of denoising,
  - do play w/ architectures, β schedules

V2 Improve
- Many other physical models can give $\lim_{T\to\infty} X^T \approx N(0, I)$
- Replace noising w/ something physical
- Think about challenges + requirements
  - Expensive to generate - how can we use shorter T
  - What other SDEs allow gaussian reverse process?
  - SPDEs?