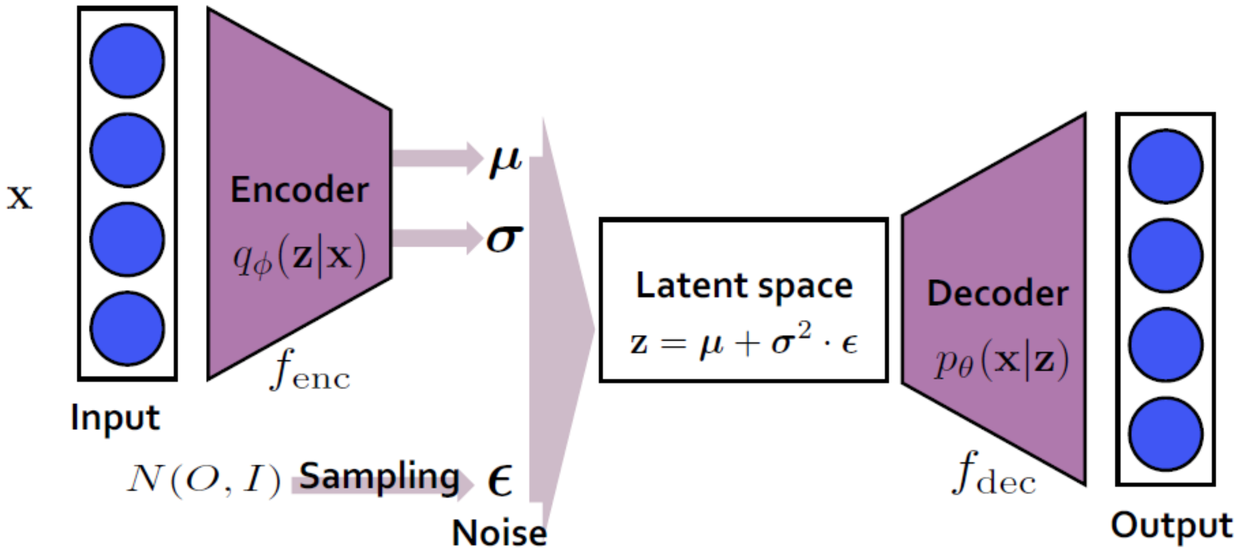
VAE and Variance

Presenter: Kim Seung Hwan (overnap@khu.ac.kr)

VAE Architecture (recall)

- The encoder inferences the mean and variance in the latent space of a sample
- From its (estimated) latent distribution, Decoder reconstruct the sample
- The latent space is set to a prior (e.g. normal or uniform)
- So the decoder can generate samples



ELBO (recall)

- It is optimized by ELBO (Evidence Lower Bound)
- ELBO consists of the reconstruction term and regularization term

$$-\log p_\theta(\mathbf{x}) \leq E_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[-\log p_\theta(\mathbf{x}|\mathbf{z})] + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}))$$

- Since ELBO is the bound of the negative log-likelihood,
- VAE is viewed as a deep maximum likelihood model

Don’t blame the ELBO

- ELBO has been blamed for many of the issues the VAE has
- Numerous studies claim the term-balance problem of ELBO
- Or/and the problem of the term itself

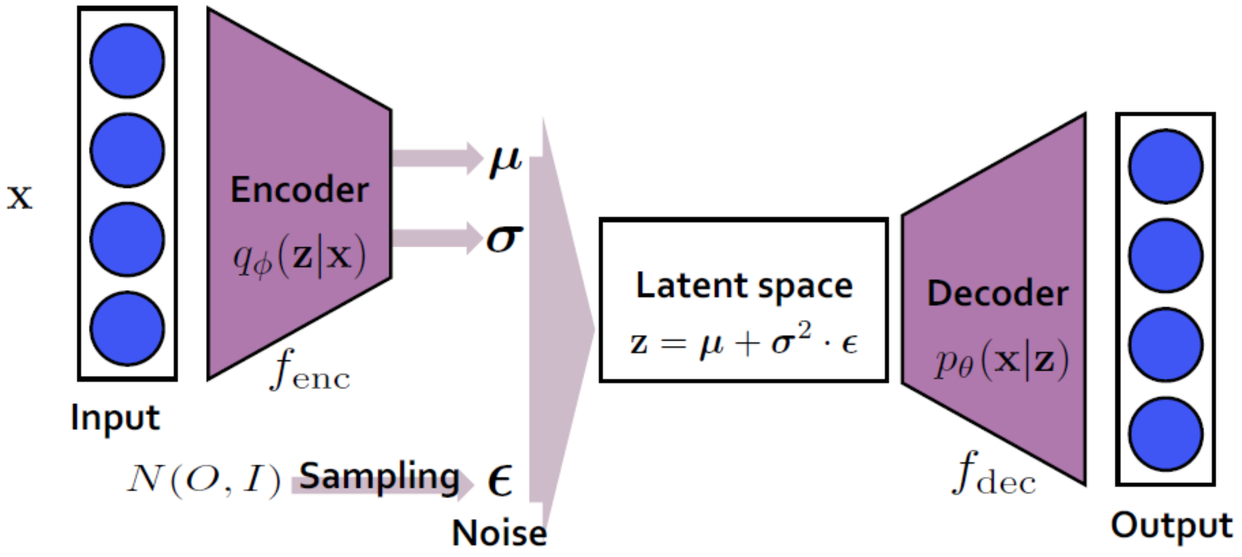
- But the local maxima issue also occurs exact log-likelihood optimization:

> “Unexpectedly, we show that spurious local maxima may arise even in the optimization of exact marginal likelihood, and such local maxima are linked with a collapsed posterior”

Lucas, James, et al. "Don't blame the elbo! a linear vae perspective on posterior collapse." Advances in Neural Information Processing Systems 32 (2019).

Blame the Encoder

- I think ELBO is both mathematically and intuitively correct
- If there is enough sampling(i.e. training data), ELBO can estimate the distribution
- I pay attention to the implementation of a VAE



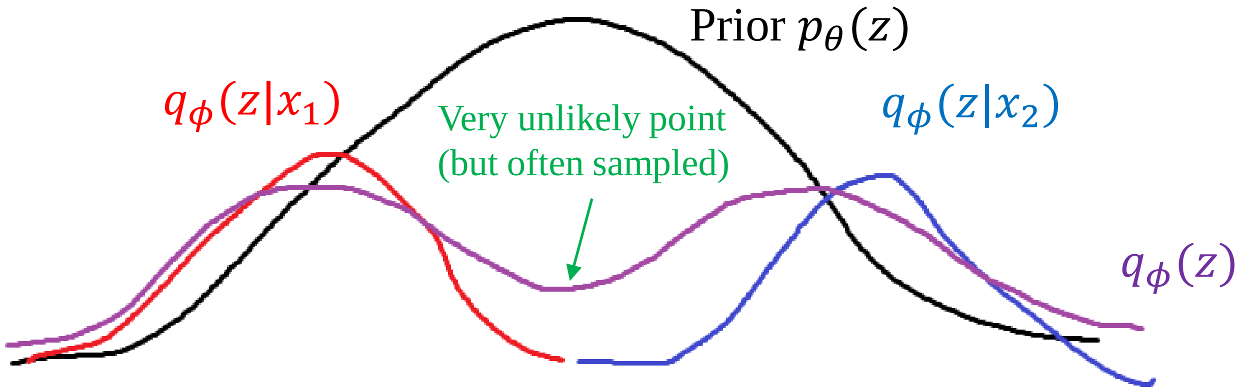
Blame the Encoder (contd.)

- We sample z from a prior distribution $p_\theta(z)$ for the generation
- Thus when $q_\phi(z)$ of the encoder must be equal to $p_\theta(z)$
- If it is different, it would not be good sampling

- The (typical) VAE estimates the mean and variance in the latent space of a sample
- Formally, The Encoder of the VAE estimates $q_\phi(z|x)$
- It is parameterized as $N(\mu_\phi(x), \sigma_\phi(x)^2 I)$

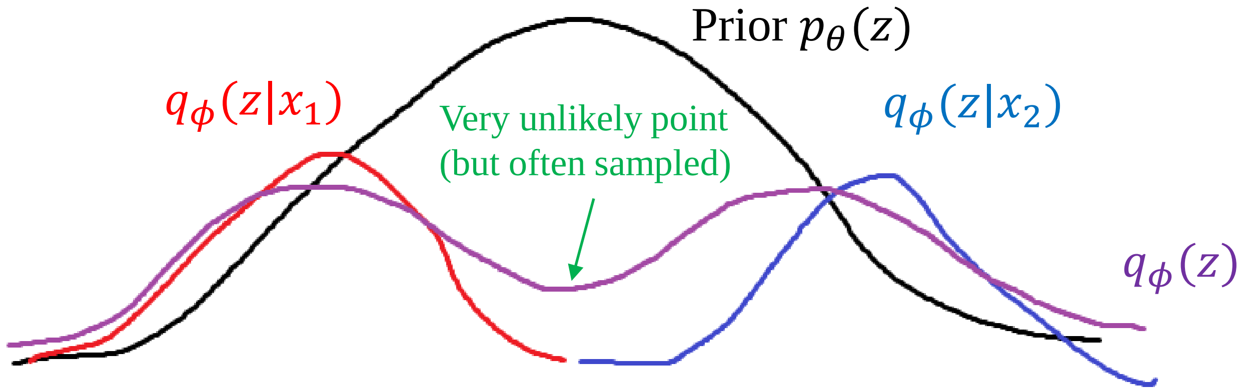
Blame the Encoder (contd.)

- Since $q_\phi(z) = \int q_\phi(z|x)q_\phi(x)dx$ and $q_\phi(z|x)$ is modeled as $N(\mu_\phi(x), \sigma_\phi(x)^2 I)$,
- $q_\phi(z)$ is closer to a Gaussian Mixture than the prior $p_\theta(z) \sim N(0, I)$
- This means that the z sampled during generation may not be sufficiently likely (from the encoder's point of view)
- i.e., it can be the latent which is difficult to exist in practice



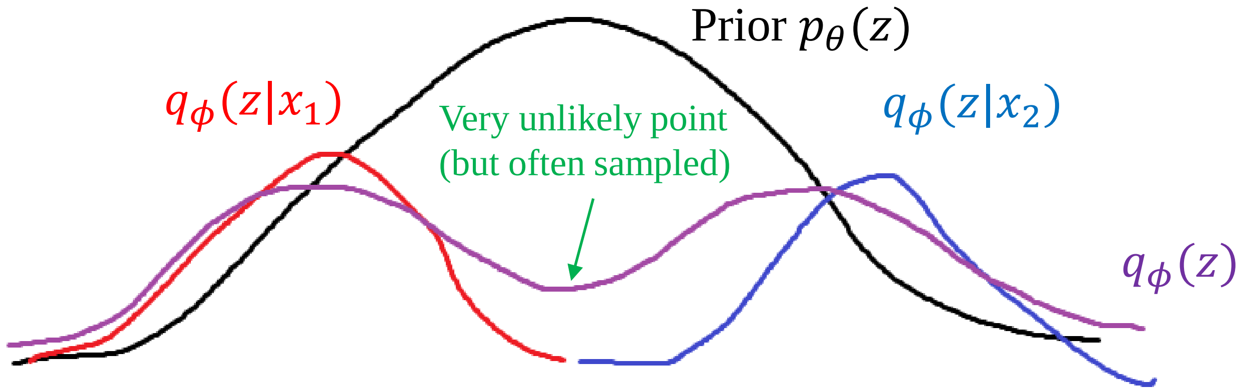
Blame the Encoder (contd.)

- Therefore, rather than modifying ELBO as before,
- The divergence of $q_\phi(z)$ and $p_\theta(z)$ should be added to the loss
- It seems very difficult…



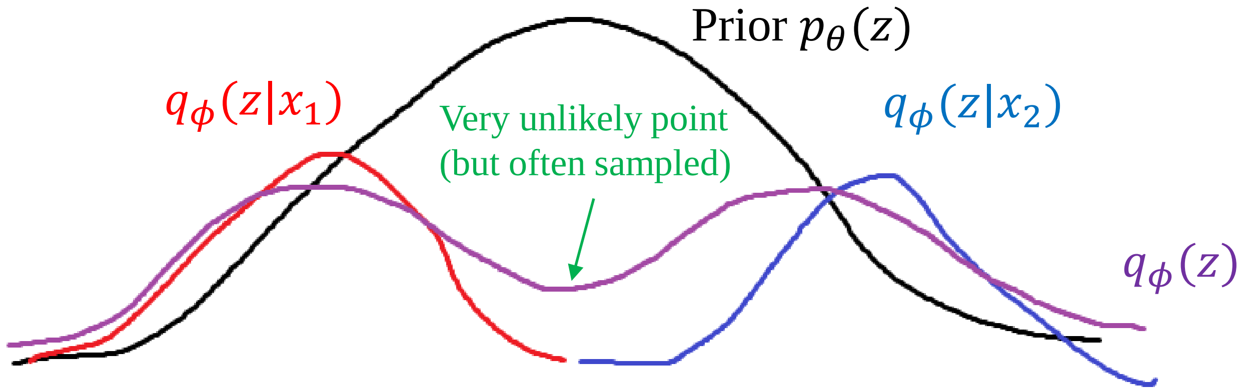
Alternative Approach

- The decoder is usually implemented as deterministic
- At another perspective, the generated one can be thought as “the mean point” that would result from the sampled latent
- This is why VAEs with deterministic decoders are blurred,
- While probabilistic decoders are noisy



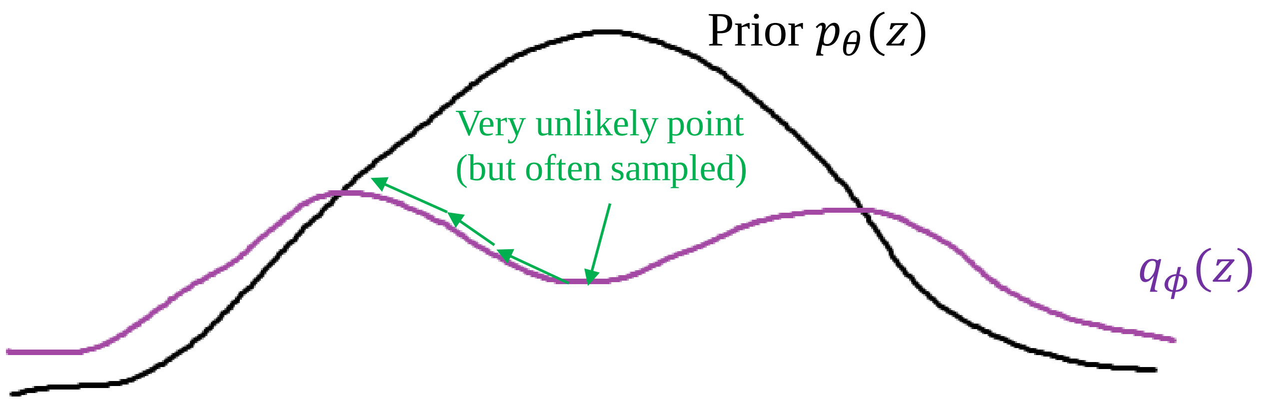
Alternative Approach (contd.)

- Let's model the decoder as probabilistic: $p_\theta(x|z) \sim N(\mu_\theta(z), \sigma_\theta(z)^2 I)$
- What kind of distribution is $\sigma_\theta(z)$?
- It will probably have a large value at an unlikely point *(need experiments)*
- Because unlikely points have more varieties/possibilities
- So Sampling is noisier than reconstruction in probabilistic decoder VAE probably *(need experiments)*



Alternative Approach (contd.)

- Then, how about gradual improvement to a likely point?
- This is taken from the idea of iterative refinement of the diffusion model
- We can implement the iterative refinement with SGD on $\sigma_\theta(z)$
- This looks like the act of going back to the training data
- However, with success in the diffusion model, it is worth experimenting *(TODO!)*



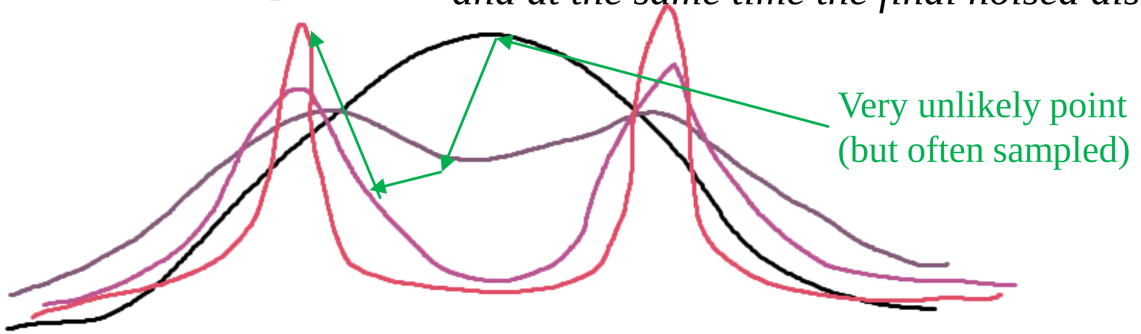
Alternative Approach (contd.)

- Note the difference with the diffusion model
  - Diffusion is performed on data (and its gradual noised) distributions
  - Whereas VAE operates on the latent (but 1-step various noised) distribution

  - Diffusion directly approximates the score, the gradient-log of the distribution
  - While VAE only knows the variance, the proxy for likelihood

Diffusion Model

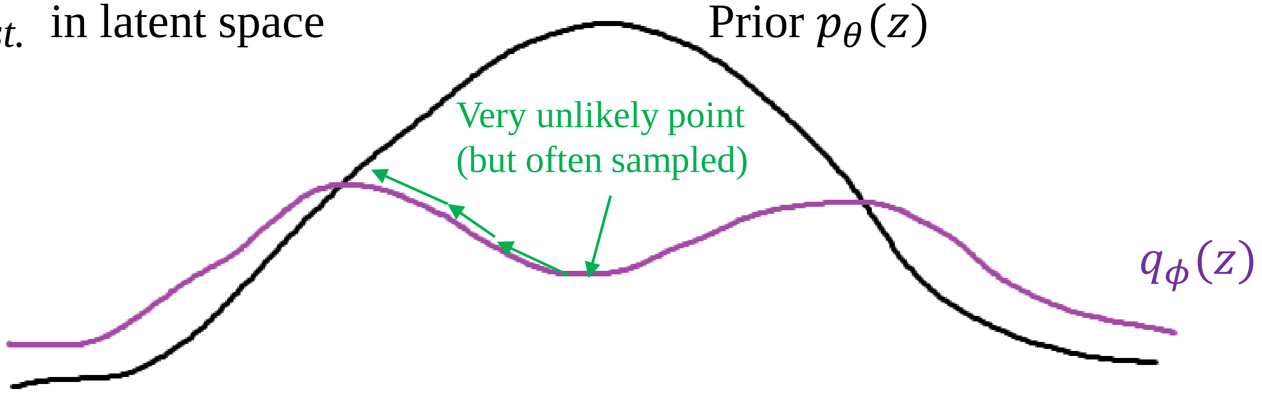
in noised data space



The redder the line, the lighter the noise

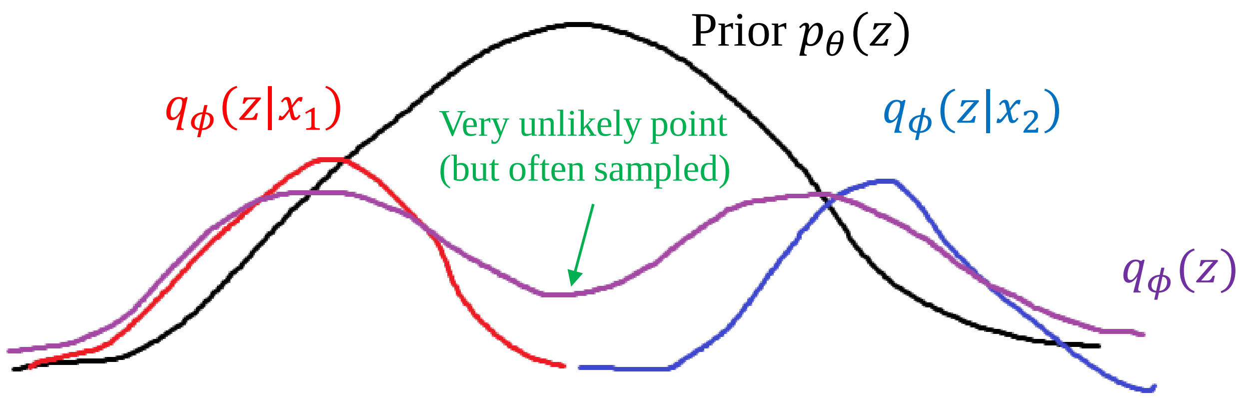
VAE

in latent space



Expected Problems

- $\sigma_\theta(z)$ is going to be smaller at the likely point (this helps improve loss)
- But is $\sigma_\theta(z)$ bigger at the unlikely point?
- Looks like some mathematical proof or additional loss is needed



Expected Problems (contd.)

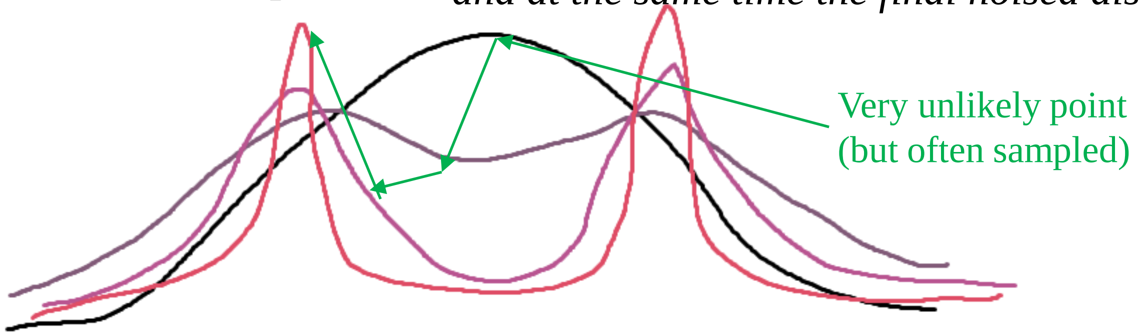
- VAE is good for logically clean normal-sampling with variational inference
- Refinement of the latent z means making a sampling method
- It is just like a different structure than ordinary VAE
- Would not this dilute the advantage of VAE?
- Why not just use diffusion?

Diffusion Model

in noised data space

Prior $p_\theta(z)$

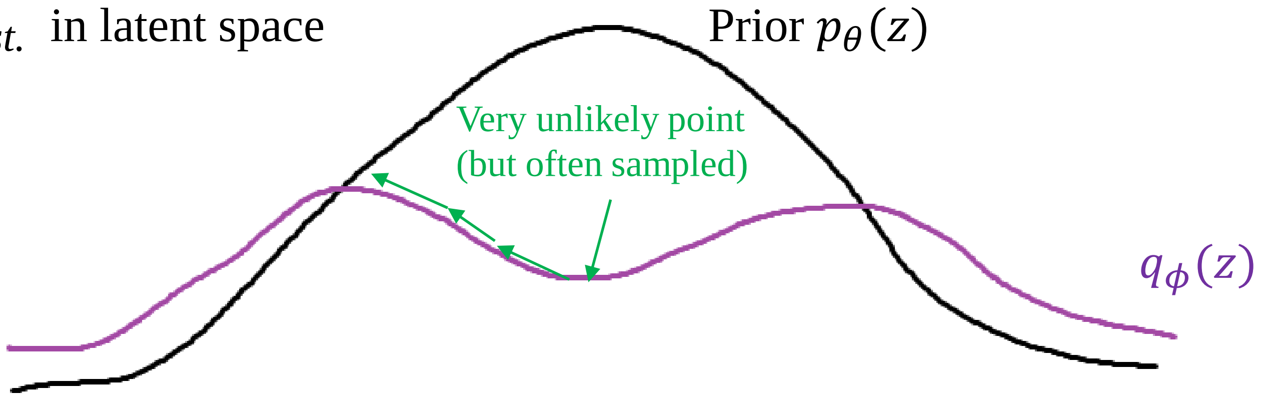
and at the same time the final noised dist.



The redder the line, the lighter the noise

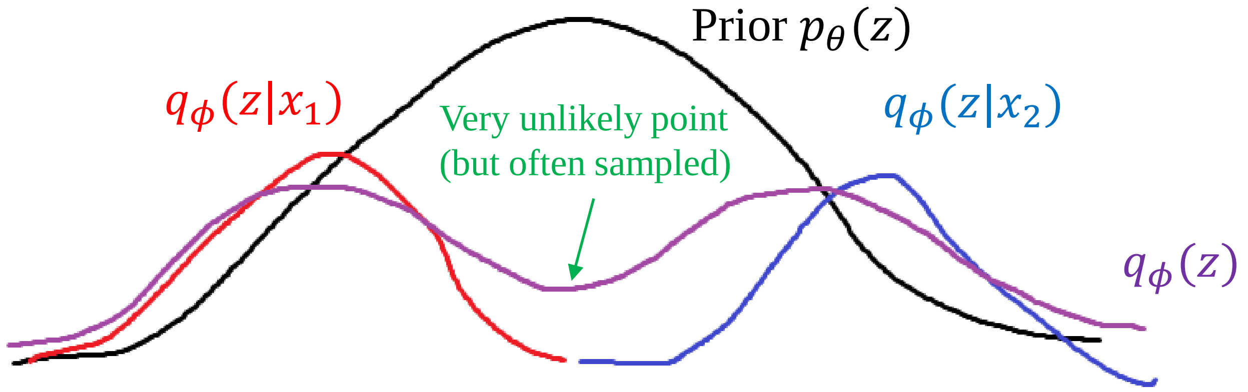
VAE

in latent space



Expected Problems (contd.)

- $\sigma_\theta(z)$ means the expected variance in the data distribution $p_\theta(x)$
- Output x has very large dimension compared to the input z
- Can this be optimized well?



Other Discussions

- I thought the variance of an encoder's output (i.e. $\sigma_\phi^2(x)$ of $q_\theta(z|x)$) was:
- "How common a sample is in that feature channel"
- But "common" is more like having a small average (i.e. $\mu_\phi(x) \approx 0$) exactly
- What does the variance mean?!

Other Discussions (contd.)

- In the first place, is the variance an important information to reconstruct/sampling?
- Already only diagonal covariance are assumed in VAE in general
- So each channel in the latent space is independent
- This is very strong assumption, but VAE works quite well...

Other Discussions (contd.)

- Theoretically, VAE with isotropic variance can represent any normalizing flow model (the computational cost for optimization is inefficient though)
- What if we just removed the degrees-of-freedom of the variance?
- Maybe we can manually set the variance per channel
- If we put it in gradual variance, can we get the benefit of the gradual noise we took in diffusion model? *(very naïve idea)*

Schedule and Plan

- Siggraph asia 2023 poster Submission Deadline: 14 August 2023, 23:59 AoE
- The deadline is very tight, and (light) VAE is very fast to experiment with
- Plan to finish all implementation and initial experiments within June (~next week)

- If the results are not promising…
- Somehow produce ideas related to VAE or diffusion
- Even if it is a bad topic, the goal is to write a paper as an experience

Schedule and Plan (contd.)

- If the results are good,
- Continue the experiment and write a poster in July
- And have to go through the previous research very meticulously

Thank you for listening

Presenter: Kim Seung Hwan (overnap@khu.ac.kr)