MITSUBISHI ELECTRIC
Changes for the Better

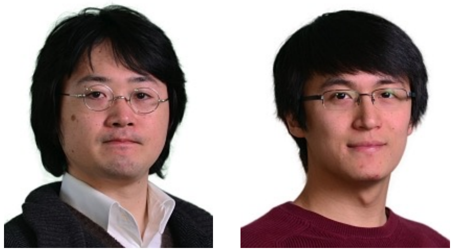
2022 IEEE INTERNATIONAL SYMPOSIUM ON INFORMATION THEORY
JUNE 26–JULY 1 AT AALTO UNIVERSITY IN ESPOO, FINLAND

H.9.4: Statistics and Machine Learning III

AutoVAE: Mismatched Variational Autoencoder with Irregular Posterior-Prior Pairing

Toshiaki Koike-Akino

Ye Wang



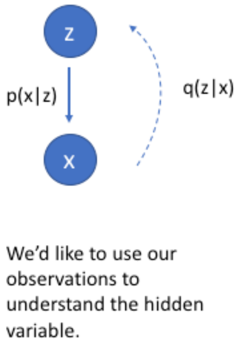
June 2022

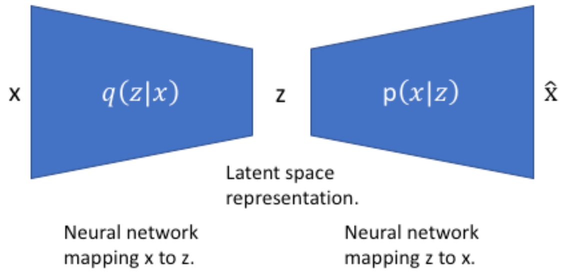
MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
Cambridge, Massachusetts, USA
http://www.merl.com

© MERL

Outline

- Trends of generative artificial intelligence (AI)
- Bayesian inference
  - Variational auto-encoder (VAE)
  - Dimensionality reduction
  - Probabilistic generative model
- Generalized variational inference (GVI)
  - Posterior-prior-likelihood beliefs
  - Discrepancy measure: divergence
  - Mismatched irregular pairing
  - **Automated VAE: AutoVAE**
- Experiments
- Summary

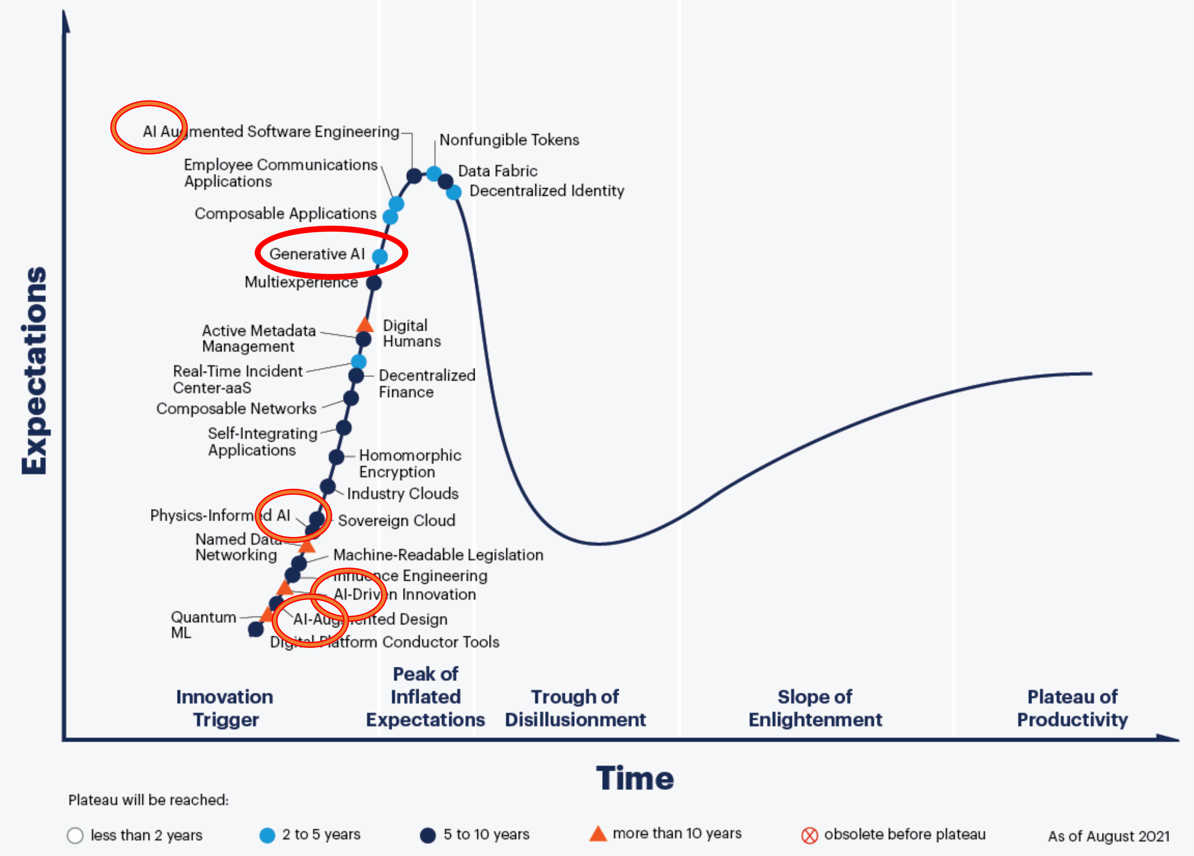




How to select stochastic model?

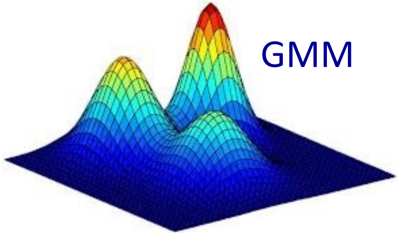
Emerging Technologies

- Gartnar's Hype Cycle for Emerging Technologies (2021 August): AI, **Generative AI**



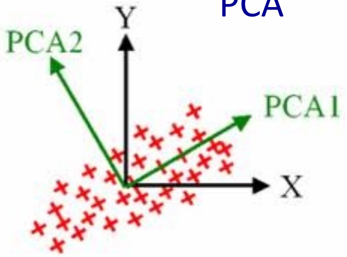
Artificial Intelligence (AI)

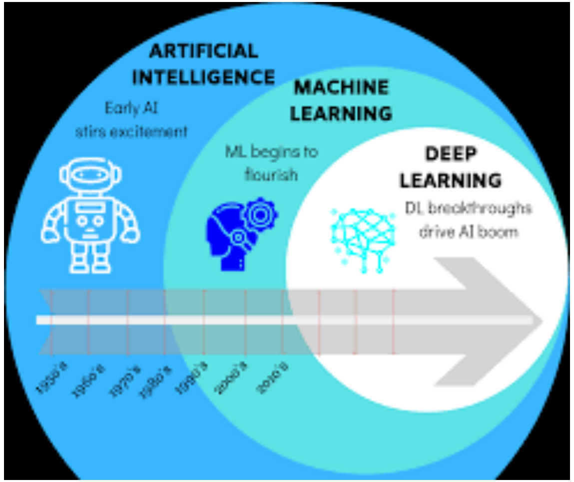
- K-means
- Gaussian mixture model (GMM)
- Principal component analysis (PCA)
- Independent component analysis (ICA)
- Logistic regression (LR)
- **Support vector machine (SVM)**
- Self-organizing map (SOM)
- Hidden Markov model (HMM)
- Artificial neural networks (ANN)
- **Deep learning (DL)**
- **QML**



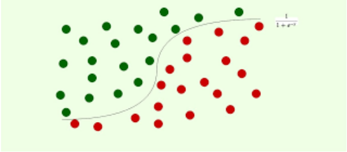
GMM

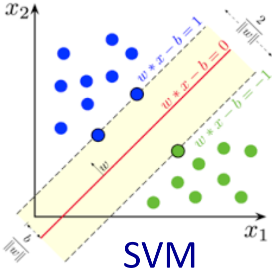
PCA



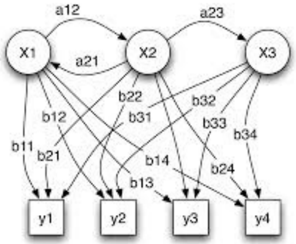


LR

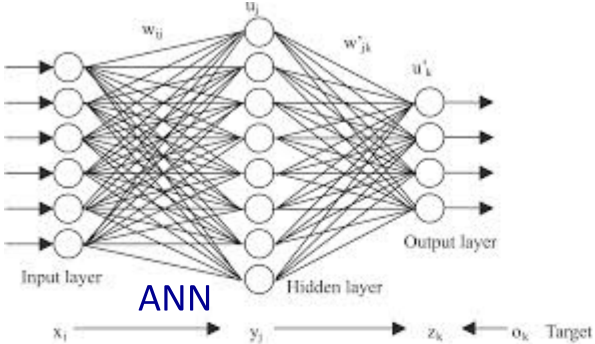




SVM



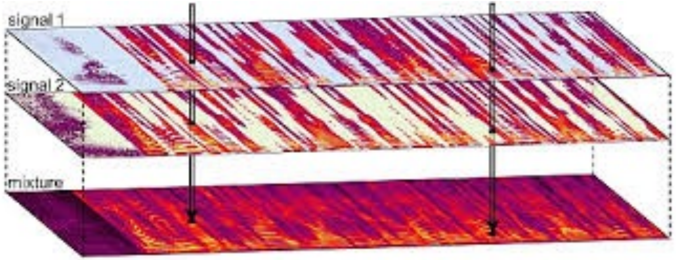
HMM

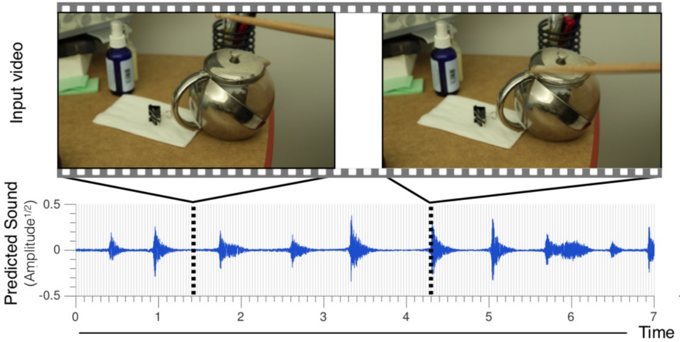


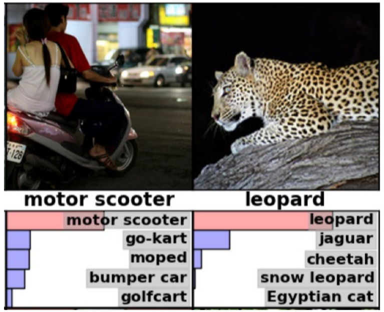
ANN

AI for Media Signal Processing

- Audio & Visual Applications

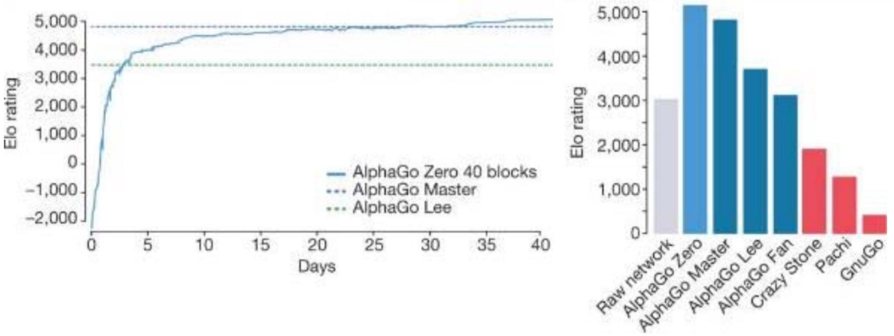






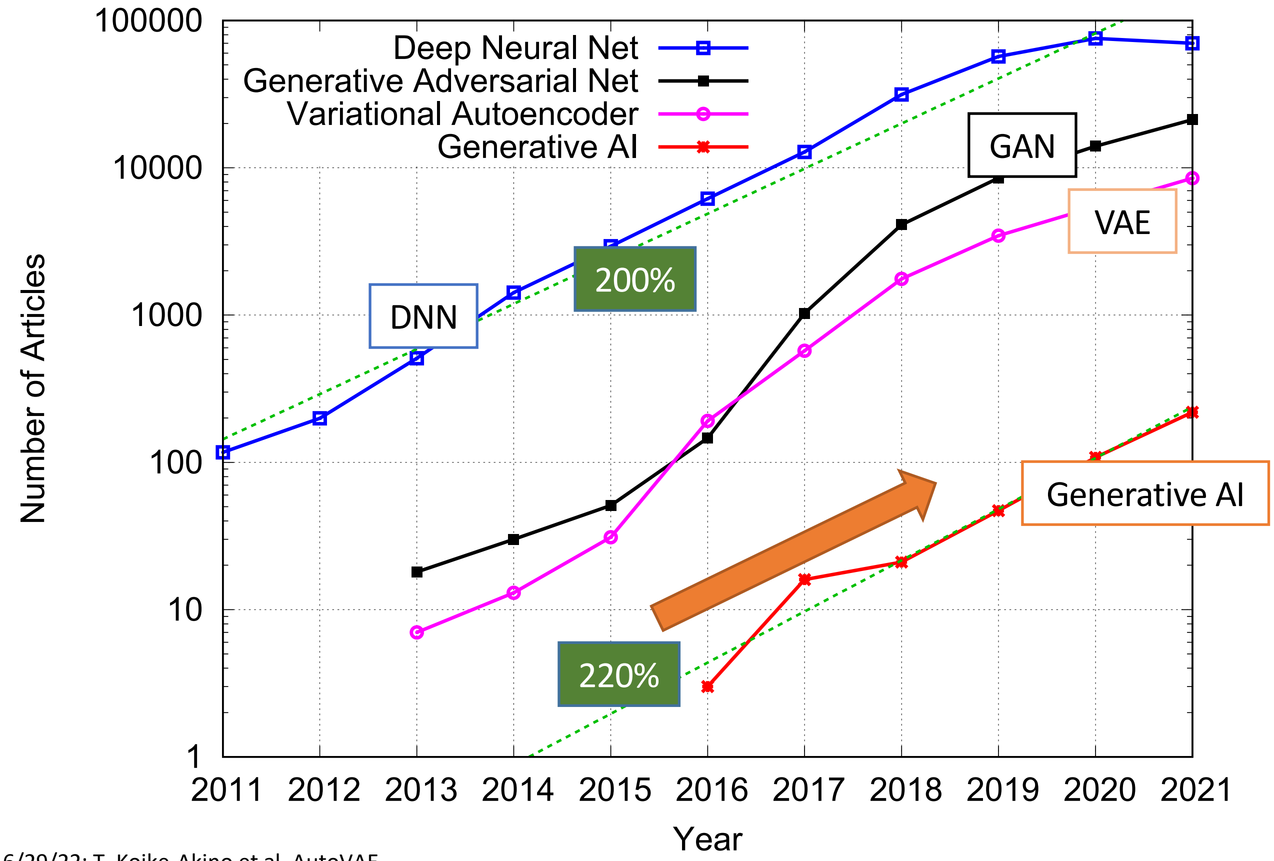
"man in black shirt is playing guitar."

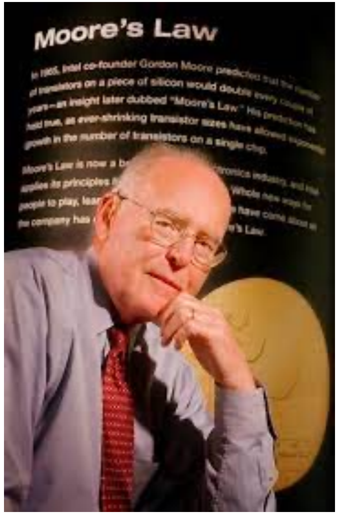
AI Surpassing Human-Level Performance



Moore's Law: Exponential Growth

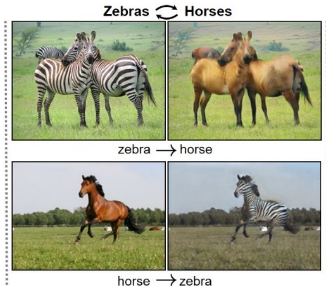
- Number of articles has been doubling every year in Google Scholar: **Generative AI**



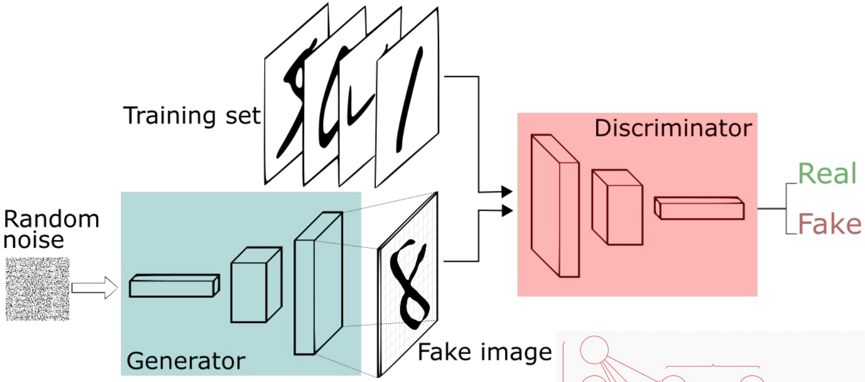


Generative AI Model

- Generative Adversarial Networks (GAN) [Goodfellow et al, 2014]
  - Train two **competing** neural networks
  - Generator learns to fake images by trying to fool discriminator
- Denoising diffusion probabilistic model (DDPM) [Ho et al., 2020]
- **Variational Auto-Encoder (VAE)** [Kingma et al, 2014]



CycleGAN [Zhu et al, 2017]



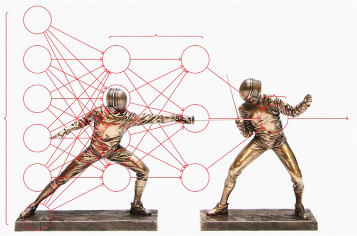


Photo-realistic face picture synthesis [Karras et al, 2018]

DDPM [Ho et al, 2020]

Variational Autoencoder (VAE)

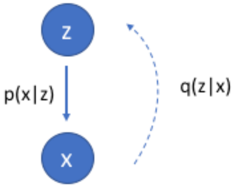
- Encoder (Inference model)

$$Z \sim q_\theta(z|x)$$

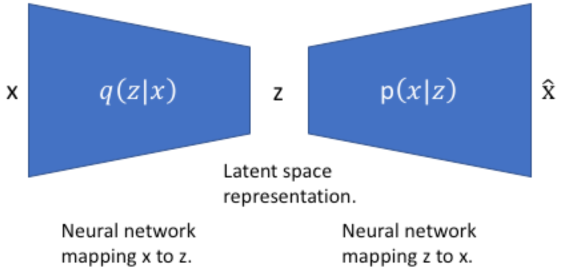
- Decoder (Generative model)

$$X' \sim p_\phi(x|z)$$

- Evidence lower-bound (ELBO)



We'd like to use our observations to understand the hidden variable.



$$\log \Pr(\boldsymbol{x}) = \log \mathop{\mathbb{E}}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\frac{p_\psi(\boldsymbol{x}|\boldsymbol{z})\pi(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] \geq \mathop{\mathbb{E}}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\log p_\psi(\boldsymbol{x}|\boldsymbol{z})\right] - D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \| \pi(\boldsymbol{z})\right)$$

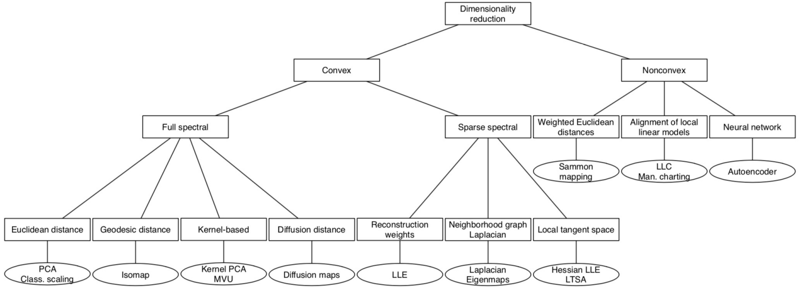
Discrepancy

Likelihood

Posterior

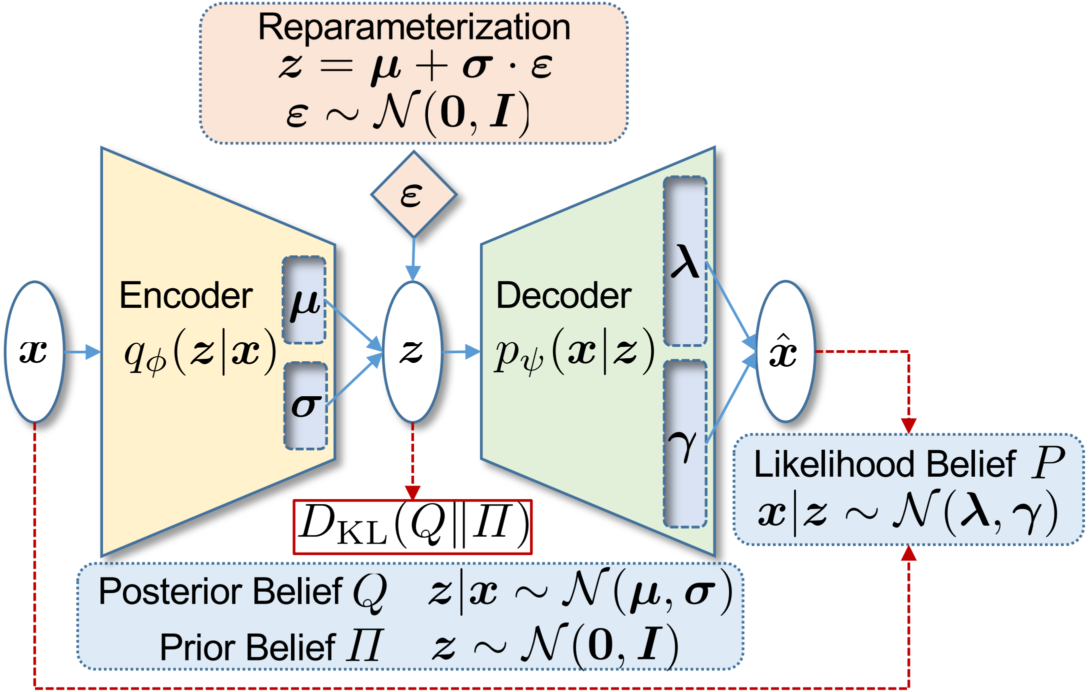
Prior

- VAE has been used in a myriad of applications:
  - Generative model
  - Bayesian inference
  - Dimensionality reduction
  - …



Standard VAE

- Typically, posterior distribution uses the same member of prior distribution family
  - Typical choice:
    - Normal prior *N*(0,1) and normal posterior *N*(mu, sigma)
    - Unspecified normal likelihood → mean-square error (MSE)
    - Bernoulli likelihood → binary cross entropy (BCE)
- What if we use mismatched posterior-prior pair?



$$\mathop{\mathbb{E}}_{\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\log p_\psi(\boldsymbol{x}|\boldsymbol{z})\right] - D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \| \pi(\boldsymbol{z})\right)$$

Variational Inference (VI) Methods

- Typical setting

$$\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\psi(\boldsymbol{x}|\boldsymbol{z})\right] - D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|\pi(\boldsymbol{z})\right)$$

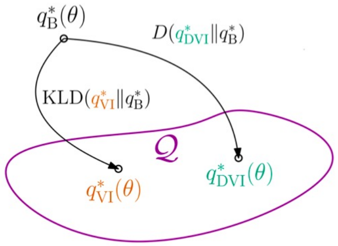
Discrepancy | Likelihood | Posterior | Prior

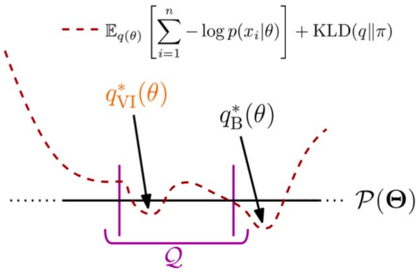
Method	Likelihood	Discrepancy	Posterior	Prior
Standard VAE [1, 2]	$\mathcal{B}$, $\mathcal{N}$	KLD	$\mathcal{N}$	$\mathcal{N}$
β-VAE [3]	$\mathcal{B}$, $\mathcal{N}$	$\beta\times$KLD	$\mathcal{N}$	$\mathcal{N}$
$\mathcal{CB}$-VAE [4]	$\mathcal{CB}$	KLD	$\mathcal{N}$	$\mathcal{N}$
Sparse-VAE [5]	$\mathcal{B}$, $\mathcal{N}$	KLD	$\mathcal{L}_{\mathrm{a}}$, $\mathcal{C}$	$\mathcal{L}_{\mathrm{a}}$, $\mathcal{C}$
IAF-VI [6]	$\mathcal{B}$, $\mathcal{N}$	KLD	IAF-$\mathcal{N}$	$\mathcal{N}$
IWAE [7]	$\mathcal{B}$, $\mathcal{N}$	KLD	IW-$\mathcal{N}$ [8]	$\mathcal{N}$
Rényi-VAE [10]	$\mathcal{B}$, $\mathcal{N}$	D_α	IW-$\mathcal{N}$	$\mathcal{N}$
Gibbs VI [9]	Any	KLD	Gibbs	$\mathcal{N}$
Generalized VI [11]	Any	Any	Any	Any
Mismatched VAE	P	D_α	$Q \neq \Pi$	Π
AutoVAE	$\mathcal{P}$	$\mathcal{D}$	$\mathcal{Q}$	$\mathbf{\Pi}$

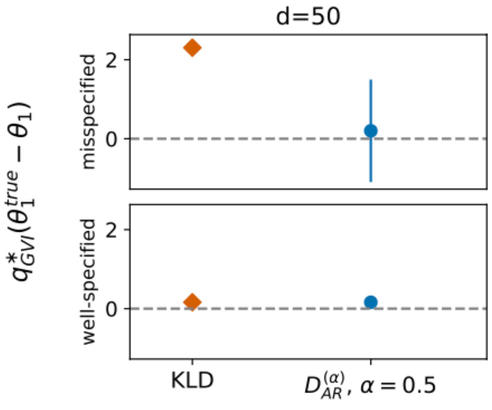
Ours

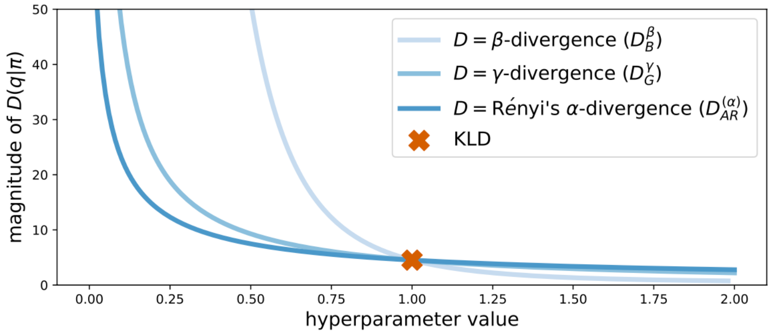
Generalized VI

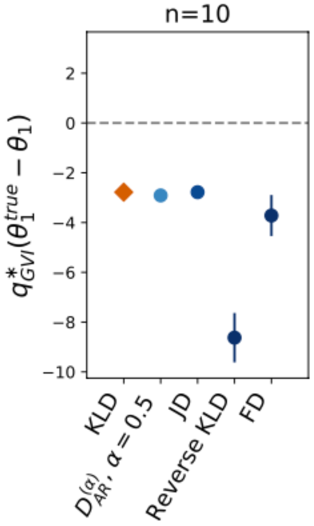
- Standard VAE is optimal if posterior/prior/likelihood beliefs are well specified
- However, real-world data do not follow specified beliefs in general
- Standard ELBO and KLD are no longer optimal for mis-specified posterior/prior/likelihood
- GVI [11] compared various discrepancy measures:
  - Renyi-alpha, beta, gamma
  - Jeffrey
  - Fisher
  - …







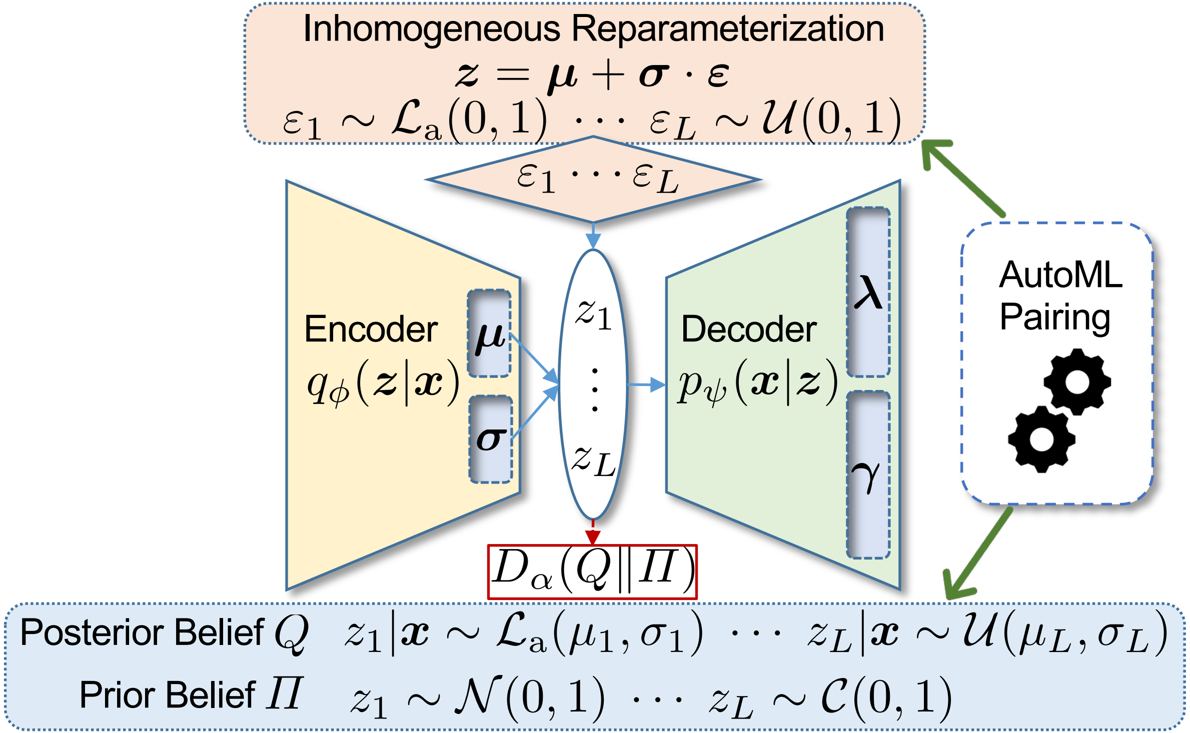




AutoVAE

- Automated machine learning (**AutoML**) for irregular mismatched posterior-prior pairing (beside architecture search)

$$\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\psi(\boldsymbol{x}|\boldsymbol{z})\right] - D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|\pi(\boldsymbol{z})\right)$$



Matched Posterior-Prior Pair

Posterior: $\boldsymbol{z} \mid \boldsymbol{x} \sim \mathcal{L}_\mathrm{a}(\boldsymbol{\mu},\boldsymbol{\sigma})$

Prior: $\boldsymbol{z} \sim \mathcal{L}_\mathrm{a}(\boldsymbol{0},\boldsymbol{I})$

Mismatched Posterior-Prior Pair

Posterior: $\boldsymbol{z} \mid \boldsymbol{x} \sim \mathcal{L}_\mathrm{a}(\boldsymbol{\mu},\boldsymbol{\sigma})$

Prior: $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$

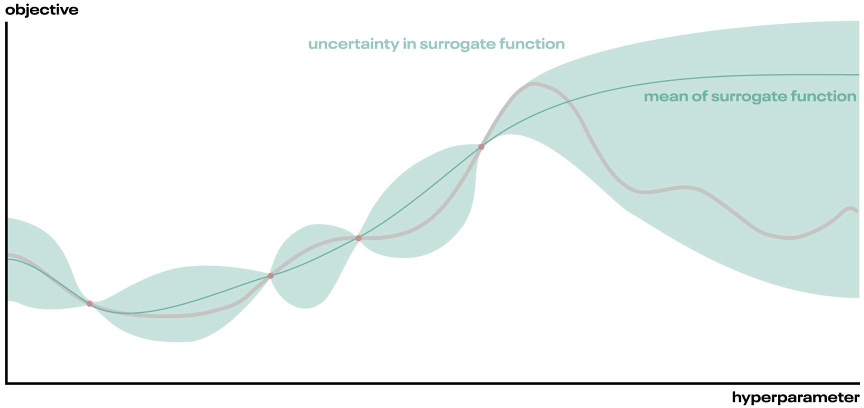
Heterogenous Posterior-Prior Pair

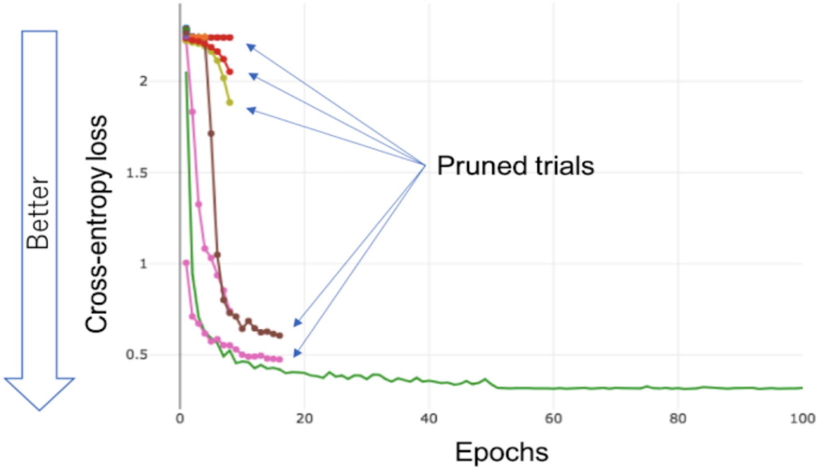
Posterior: $\boldsymbol{z}_1 \mid \boldsymbol{x} \sim \mathcal{L}_\mathrm{a}(\boldsymbol{\mu}_1,\boldsymbol{\sigma}_1)$, $\boldsymbol{z}_2 \mid \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\sigma}_2)$

Prior: $\boldsymbol{z}_1 \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$, $\boldsymbol{z}_2 \sim \mathcal{L}_\mathrm{a}(\boldsymbol{0},\boldsymbol{I})$

AutoML

- We propose to use **AutoML** framework to automate posterior-prior pairing
- We use Optuna
  - Sampler: CMA-ES, TPE (Bayesian Optimization), …
  - Pruner: Hyperband, Median, Successive Halving
  - Analysis: functional analysis of variance (fANOVA)
  - Interface: compatible to Pytorch, SK-learn, etc.
  - Parallelization: SQL-based data sharing
  - Multi-objective optimization





Reparameterization Trick: Location-Scale Family (LSF)

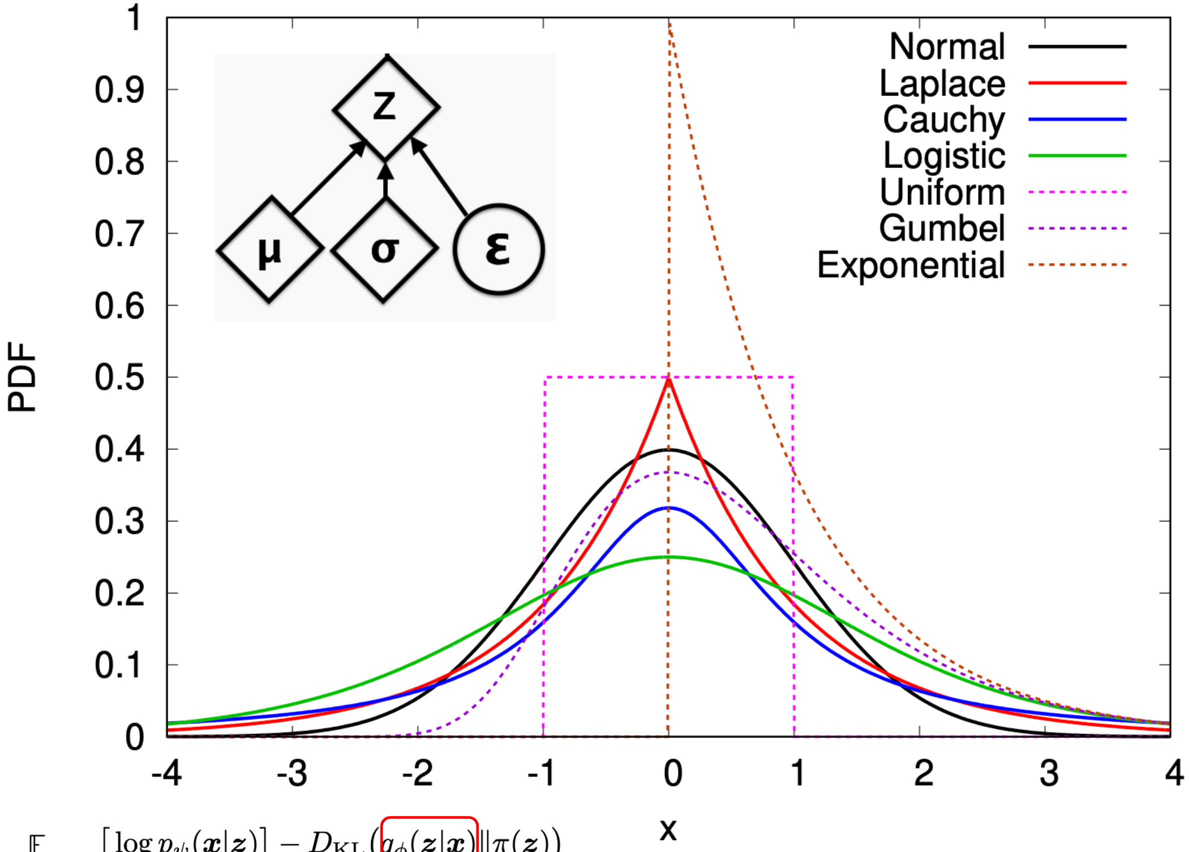
- Choice of posterior beliefs should allow differentiable **reparameterization trick**

$$Z = \mu + \sigma \cdot \varepsilon$$

- **LSF** is a natural candidate
  - Normal
  - Laplace
  - Cauchy
  - Logistic
  - Uniform
  - Gumbel
  - Exponential (scale family)
  - …

$$\varepsilon \sim \mathbb{LSF}(0,1)$$

$$Z \sim \mathbb{LSF}(\mu,\sigma)$$

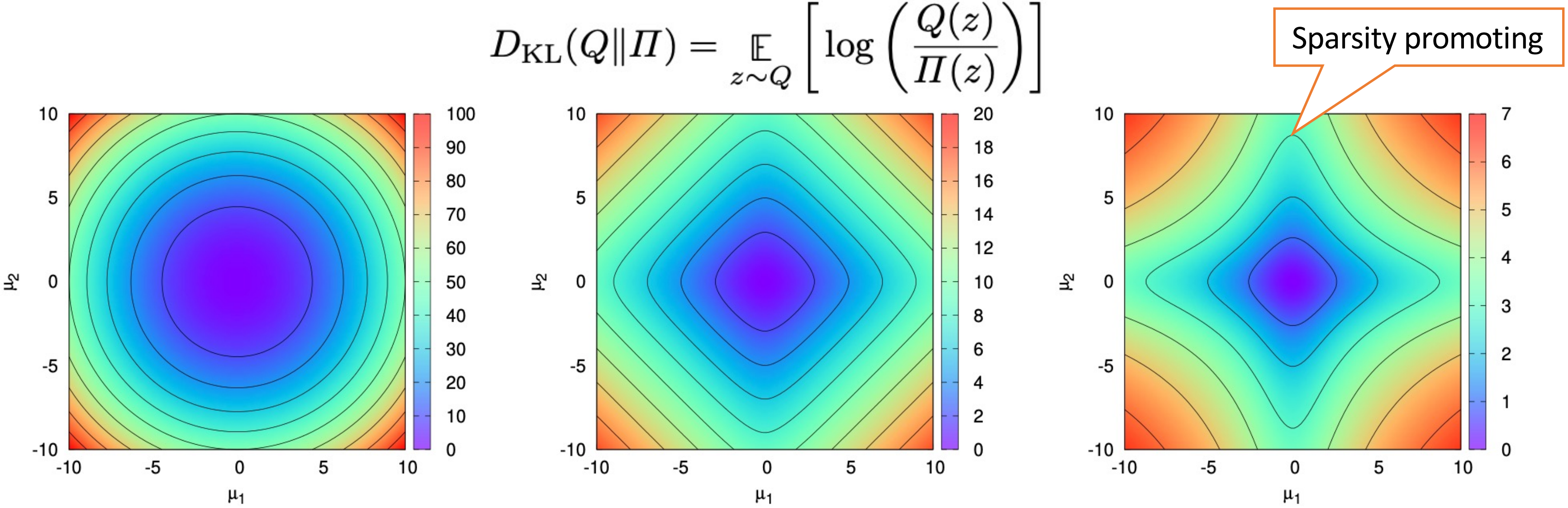


$$\mathop{\mathbb{E}}_{z\sim q_\phi(z|x)}\left[\log p_\psi(x|z)\right] - D_{\mathrm{KL}}\left(q_\phi(z|x) \| \pi(z)\right)$$

KLD Expression

- For choice of prior beliefs, KLD should be computed efficiently (closed-form expression)
- c.f) 2D landscape of KLD for matched normal, Laplace and Cauchy beliefs

$$D_{\mathrm{KL}}(Q\|\Pi) = \mathop{\mathbb{E}}_{z\sim Q}\left[\log\left(\frac{Q(z)}{\Pi(z)}\right)\right]$$



(a) Normal-Normal $\min_{\boldsymbol{\sigma}} D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\sigma})\|\mathcal{N}(\mathbf{0},\boldsymbol{I}))$

(b) Laplace-Laplace $\min_{\boldsymbol{\sigma}} D_{\mathrm{KL}}(\mathcal{L}_{\mathrm{a}}(\boldsymbol{\mu},\boldsymbol{\sigma})\|\mathcal{L}_{\mathrm{a}}(\mathbf{0},\boldsymbol{I}))$

(c) Cauchy-Cauchy $\min_{\boldsymbol{\sigma}} D_{\mathrm{KL}}(\mathcal{C}(\boldsymbol{\mu},\boldsymbol{\sigma})\|\mathcal{C}(\mathbf{0},\boldsymbol{I}))$

$$\mathop{\mathbb{E}}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\psi(\boldsymbol{x}|\boldsymbol{z})] - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|\pi(\boldsymbol{z}))$$

KLD Expression of Matched/Mismatched Pairing

- KLD for 16 posterior-prior pairs (matched and mismatched)

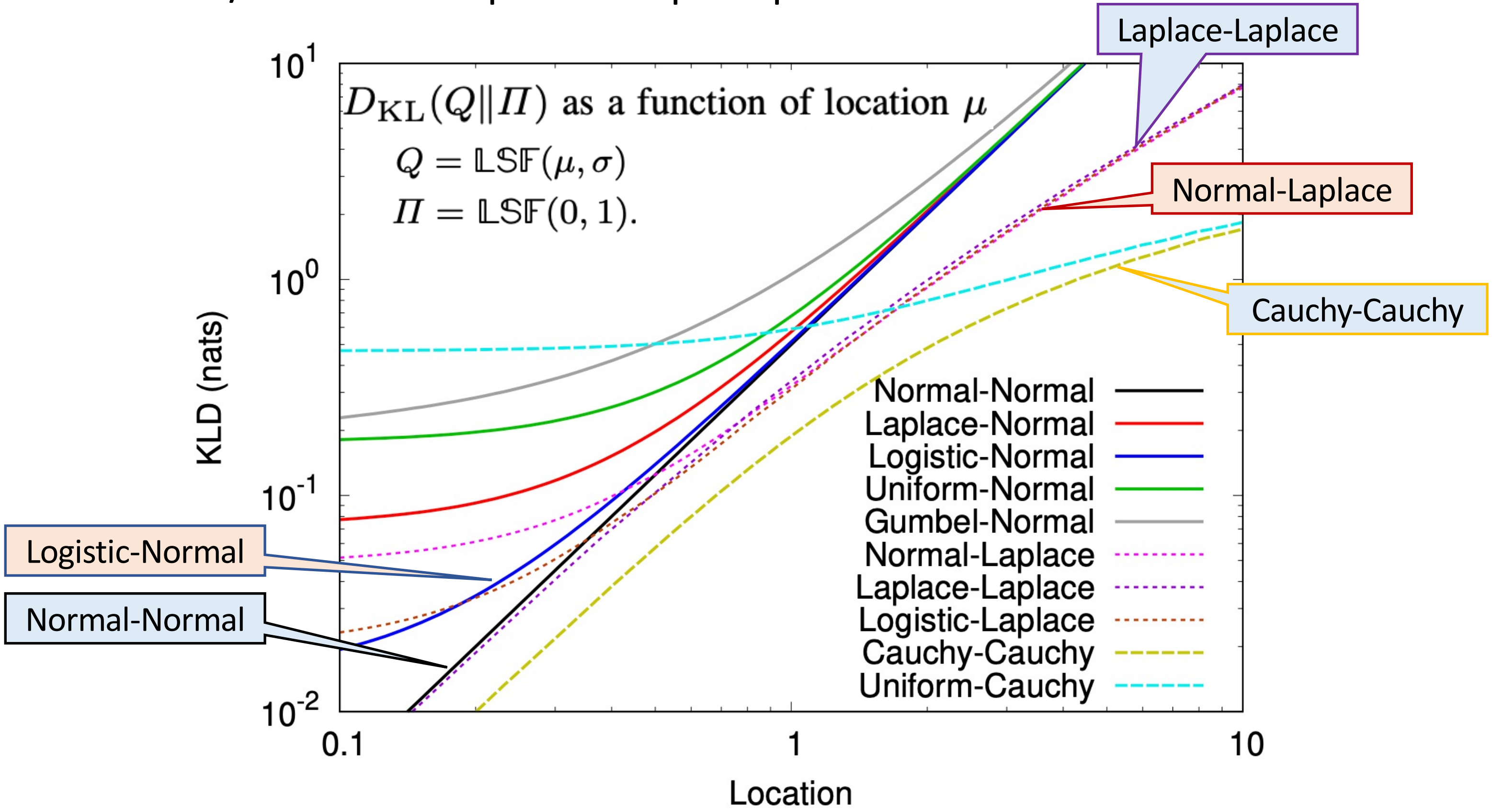
Posterior Q	Prior Π	KLD $D_{\mathrm{KL}}(Q\|\Pi)$
$\mathcal{N}(\mu,\sigma)$	$\mathcal{N}(0,1)$	$\frac{1}{2}(\mu^2+\sigma^2-1-\log(\sigma^2))$
$\mathcal{L}_{\mathrm{a}}(\mu,\sigma)$	$\mathcal{N}(0,1)$	$\frac{1}{2}\mu^2+\sigma^2-1-\frac{1}{2}\log\left(\frac{2\sigma^2}{\pi}\right)$
$\mathcal{L}_{\mathrm{o}}(\mu,\sigma)$	$\mathcal{N}(0,1)$	$\frac{1}{2}\mu^2+\frac{\pi^2}{6}\sigma^2-2-\frac{1}{2}\log\left(\frac{\sigma^2}{2\pi}\right)$
$\mathcal{U}(\mu,\sigma)$	$\mathcal{N}(0,1)$	$\frac{1}{2}\mu^2+\frac{1}{6}\sigma^2-\frac{1}{2}\log\left(\frac{2\sigma^2}{\pi}\right)$
$\mathcal{G}(\mu,\sigma)$	$\mathcal{N}(0,1)$	$\log(\frac{\sqrt{2\pi}}{\sigma})+\frac{\pi^2\sigma^2}{12}+\frac{(\mu+\sigma\gamma_0)^2}{2}-\gamma_0-1$
$\mathcal{E}(\sigma)$	$\mathcal{N}(0,1)$	$\sigma^2-1-\frac{1}{2}\log\left(\frac{\sigma^2}{2\pi}\right)$
$\mathcal{N}(\mu,\sigma)$	$\mathcal{L}_{\mathrm{a}}(0,1)$	$\mu\cdot\mathrm{erf}\frac{\mu}{\sqrt{2\sigma^2}}+\sqrt{\frac{2\sigma^2}{\pi}}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)-\frac{1}{2}-\frac{1}{2}\log\left(\frac{\pi\sigma^2}{2}\right)$
$\mathcal{L}_{\mathrm{a}}(\mu,\sigma)$	$\mathcal{L}_{\mathrm{a}}(0,1)$	$\|\mu\|+\sigma\exp\left(-\frac{\|\mu\|}{\sigma}\right)-1-\log(\sigma)$
$\mathcal{L}_{\mathrm{o}}(\mu,\sigma)$	$\mathcal{L}_{\mathrm{a}}(0,1)$	$2\sigma\log\left(2\cosh\left(\frac{\mu}{2\sigma}\right)\right)-2-\log\left(\frac{\sigma}{2}\right)$
$\mathcal{E}(\sigma)$	$\mathcal{L}_{\mathrm{a}}(0,1)$	$\sigma-\log(\sigma)-1+\log(2)$
$\mathcal{C}(\mu,\sigma)$	$\mathcal{C}(0,1)$	$\log(\mu^2+(1+\sigma)^2)-\log(4\sigma)$
$\mathcal{U}(\mu,\sigma)$	$\mathcal{C}(0,1)$	$\frac{1}{\sigma}\tan^{-1}(\sigma-\mu)+\frac{1}{\sigma}\tan^{-1}(\sigma+\mu)-2-\log\left(\frac{2\sigma}{\pi}\right)+\frac{\sigma-\mu}{2\sigma}\log\left(1+(\sigma-\mu)^2\right)+\frac{\sigma+\mu}{2\sigma}\log\left(1+(\sigma+\mu)^2\right)$
$\mathcal{N}(\mu,\sigma)$	$\mathcal{G}(0,1)$	$-\log(\sigma)+\mu+\exp(-\mu+\frac{\sigma^2}{2})-\frac{1+\log(2\pi)}{2}$
$\mathcal{U}(\mu,\sigma)$	$\mathcal{G}(0,1)$	$\mu+\frac{1}{\sigma}\exp(-\mu)\sinh(\sigma)-\log(2\sigma)$
$\mathcal{G}(\mu,\sigma)$	$\mathcal{G}(0,1)$	$\mu-\log(\sigma)+\Gamma(\sigma+1)\mathrm{e}^{-\mu}-1+\gamma_0(\sigma-1)$
$\mathcal{E}(\sigma)$	$\mathcal{G}(0,1)$	$\sigma+(1+\sigma)^{-1}-1-\log(\sigma)$

Matched Pairs

Mismatched Pairs

Mismatched Pairs

- KLD of matched/mismatched posterior-prior pairs



Renyi Divergence: Variational Renyi (VR) Bound

- Renyi divergence variational inference
  - https://arxiv.org/abs/1602.02311

$$D_\alpha(Q\|\Pi) = \frac{1}{\alpha-1}\log \mathop{\mathbb{E}}_{z\sim Q}\left[\left(\frac{Q(z)}{\Pi(z)}\right)^{\alpha-1}\right]$$

RÉNYI DIVERGENCE $D_\alpha(Q\|\Pi)$ SPECIAL CASES [10]

Order α	Definition	Correspondence
$\alpha \to 0$	$-\log \int_{Q(z)>0} \Pi(z)\mathrm{d}z$	Overlap (i.e., IWAE [7])
$\alpha = 0.5$	$-2\log(1-\mathrm{Hel}^2[Q\|\Pi])$	Square Hellinger distance
$\alpha \to 1$	$\int Q(z)\log\frac{Q(z)}{\Pi(z)}\mathrm{d}z$	KLD (i.e., standard VAE [1])
$\alpha = 2$	$-\log(1-\chi^2[Q\|\Pi])$	χ^2-divergence
$\alpha \to \infty$	$\log\max\frac{Q(z)}{\Pi(z)}$	Worst-case regret

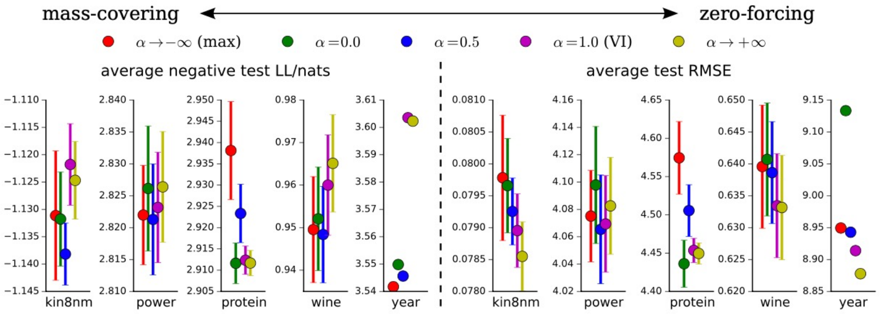
VI bound (ELBO)

$$\mathcal{L}_{\mathrm{VI}}(q;\mathcal{D},\boldsymbol{\varphi}) = \log p(\mathcal{D}|\boldsymbol{\varphi}) - \mathrm{KL}[q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D},\boldsymbol{\varphi})] = \mathbb{E}_q\left[\log\frac{p(\boldsymbol{\theta},\mathcal{D}|\boldsymbol{\varphi})}{q(\boldsymbol{\theta})}\right]$$

VR bound

$$\mathcal{L}_\alpha(q;\mathcal{D}) := \frac{1}{1-\alpha}\log\mathbb{E}_q\left[\left(\frac{p(\boldsymbol{\theta},\mathcal{D})}{q(\boldsymbol{\theta})}\right)^{1-\alpha}\right]$$

alpha=0: importance-weighted AE (IWAE)



Gradient weighting

$$\nabla_\phi\mathcal{L}_\alpha(q_\phi;\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\epsilon}}\left[w_\alpha(\boldsymbol{\epsilon};\boldsymbol{\phi},\boldsymbol{x})\nabla_\phi\log\frac{p(g_\phi(\boldsymbol{\epsilon}),\boldsymbol{x})}{q(g_\phi(\boldsymbol{\epsilon}))}\right],$$

$$w_\alpha(\boldsymbol{\epsilon};\boldsymbol{\phi},\boldsymbol{x}) = \left(\frac{p(g_\phi(\boldsymbol{\epsilon}),\boldsymbol{x})}{q(g_\phi(\boldsymbol{\epsilon}))}\right)^{1-\alpha} \Big/ \mathbb{E}_{\boldsymbol{\epsilon}}\left[\left(\frac{p(g_\phi(\boldsymbol{\epsilon}),\boldsymbol{x})}{q(g_\phi(\boldsymbol{\epsilon}))}\right)^{1-\alpha}\right]$$

$$\mathop{\mathbb{E}}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\psi(\boldsymbol{x}|\boldsymbol{z})\right] - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|\pi(\boldsymbol{z}))$$

Reconstruction Loss: Generalized NLL for Various Likelihood Beliefs

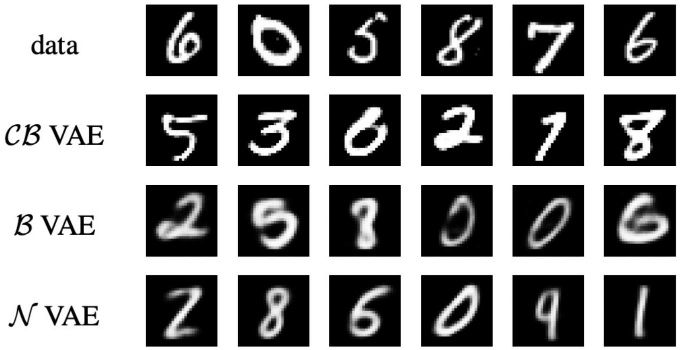
- Various choice for likelihood belief *P*
- E.g., Loaiza-Ganem et al. “The continuous Bernoulli: fixing a pervasive error in variational autoencoders”: comparing Bernoulli, cont. Bernoulli, normal, beta NLL

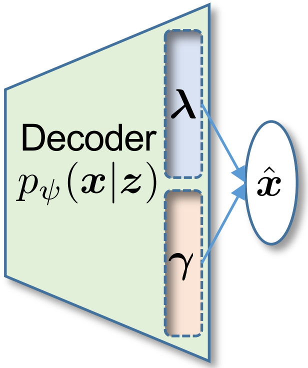
Generalized NLL for various likelihood beliefs P

Likelihood P	Generalized NLL Loss ℓ
$\mathcal{B}(\lambda)$	$\mathsf{BCE}(x;\lambda) = -x\log(\lambda) - (1-x)\log(1-\lambda)$
$\mathcal{CB}(\lambda)$	$\mathsf{NLL}(x;\lambda) = \mathsf{BCE}(x;\lambda) - \log C(\lambda)$
$\mathcal{N}(\lambda,*)$	$\mathsf{MSE}(x;\lambda) = (x-\lambda)^2$ (omitting unspecified variance)
$\mathcal{L}_\mathrm{a}(\lambda,*)$	$\mathsf{MAE}(x;\lambda) = \lvert x-\lambda\rvert$ (omitting unspecified variance)
$\mathcal{N}(\lambda,\gamma)$	$\mathsf{NLL}(x;\lambda,\gamma) = \frac{1}{2\gamma^2}\mathsf{MSE}(x;\lambda) + \frac{1}{2}\log(2\pi\gamma^2)$
$\mathcal{L}_\mathrm{a}(\lambda,\gamma)$	$\mathsf{NLL}(x;\lambda,\gamma) = \frac{1}{\gamma}\mathsf{MAE}(x;\lambda) + \log(2\gamma)$
$\mathcal{B}_\mathrm{e}(\lambda,\gamma)$	$\mathsf{NLL}(x;\lambda,\gamma) = (1-\lambda)\log(x) + (1-\gamma)\log(1-x) + \log\Gamma(\lambda) + \log\Gamma(\gamma) - \log\Gamma(\lambda+\gamma)$

Univariate

Bivariate

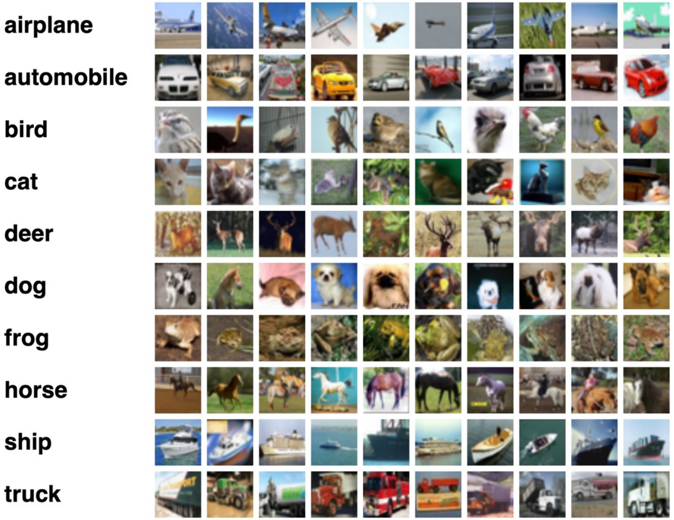


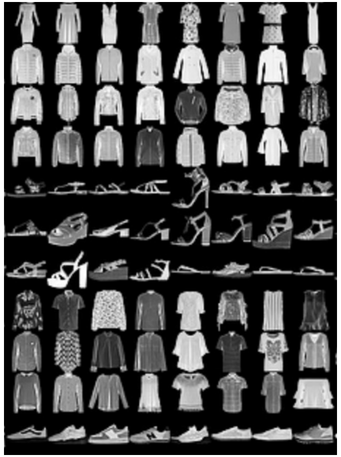


$$\mathop{\mathbb{E}}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\psi(\boldsymbol{x}|\boldsymbol{z})\right] - D_\mathrm{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})\,\|\,\pi(\boldsymbol{z})\right)$$

Experiments

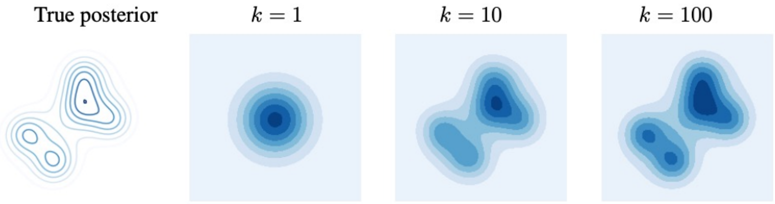
- VAE architecture
  - 3 layers 400 hidden nodes
  - 20 latent variables
  - Adam (0.0001)
  - Mini-batch 1000
  - 100 epochs
- Datasets
  - **MNIST**
  - CIFAR-10
  - FMNIST
  - KMNIST
  - SVHN
  - CIFAR-100
  - …





Multi-Sample ELBO: Importance-Weighted AE (IWAE)

- Multi-sample ELBO



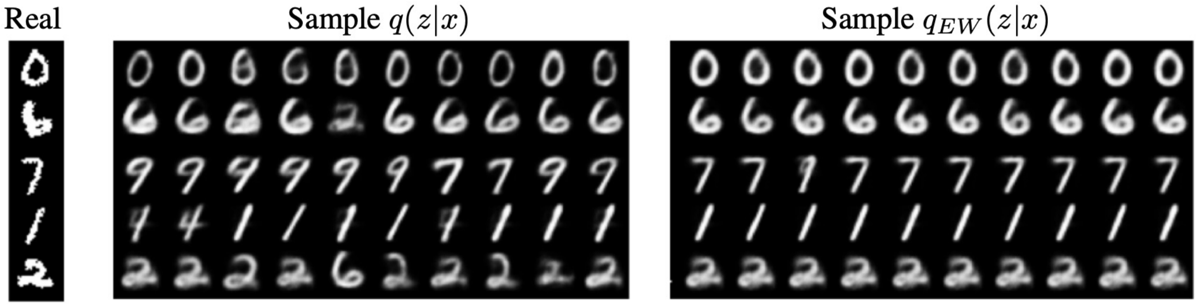
$$\log p(x) \geq E_{z\sim q(z|x)}\left[\log\left(\frac{p(x,z)}{q(z|x)}\right)\right] = L_{VAE}[q]. \qquad \text{(VAE ELBO)}$$

- IWAE: Tighter ELBO than standard VI
  - Burda et al. "Importance weighted autoencoders": https://arxiv.org/pdf/1509.00519.pdf
  - Cremer et al. "Reinterpreting importance weighted autoencoders": https://arxiv.org/pdf/1704.02916.pdf

$$\log p(x) \geq E_{z_1\ldots z_k\sim q(z|x)}\left[\log\left(\frac{1}{k}\sum_{i=1}^{k}\frac{p(x,z_i)}{q(z_i|x)}\right)\right] = L_{IWAE}[q] \qquad \text{(IWAE ELBO)}$$

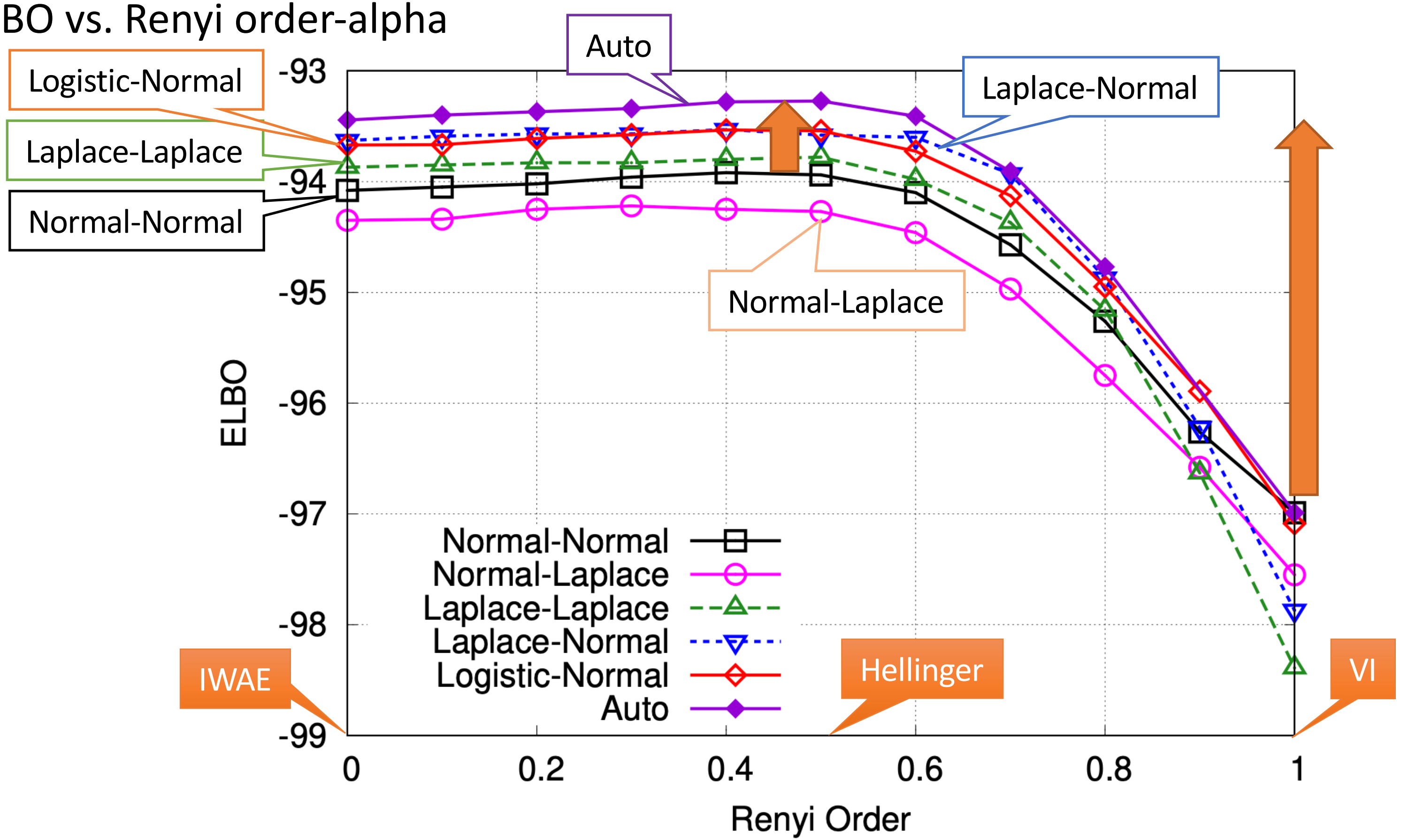
Algorithm 1 Sampling $q_{EW}(z|x)$

1: $k \leftarrow$ *number of importance samples*
2: **for** i in 1...k **do**
3: $\quad z_i \sim q(z|x)$
4: $\quad w_i = \frac{p(x,z_i)}{q(z_i|x)}$
5: Each $\tilde{w}_i = w_i / \sum_{i=1}^{k} w_i$
6: $j \sim Categorical(\tilde{\boldsymbol{w}})$
7: Return z_j



ELBO Performance: Variational Renyi (VR) Bound

- ELBO vs. Renyi order-alpha



Generated Image Snapshots

(a) $\mathcal{N}||\mathcal{N}$

(b) $\mathcal{L}_\mathrm{a}||\mathcal{N}$

(c) $\mathcal{L}_\mathrm{o}||\mathcal{N}$

(d) $\mathcal{U}||\mathcal{N}$

(e) $\mathcal{E}||\mathcal{N}$

(f) $\mathcal{N}||\mathcal{L}_\mathrm{a}$

(g) $\mathcal{L}_\mathrm{a}||\mathcal{L}_\mathrm{a}$

(h) $\mathcal{L}_\mathrm{o}||\mathcal{L}_\mathrm{a}$

(i) $\mathcal{E}||\mathcal{L}_\mathrm{a}$

(j) $\mathcal{C}||\mathcal{C}$

(k) $\mathcal{U}||\mathcal{C}$

(l) Auto

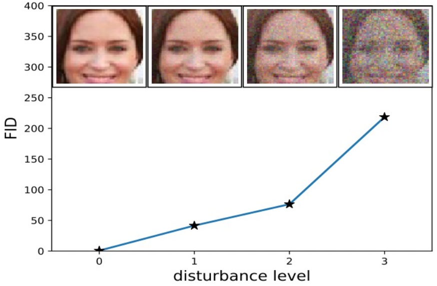
Inception Score for Synthetic Data Quality Measure

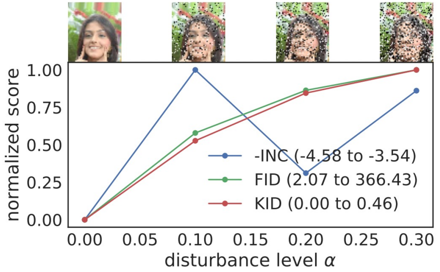
- We use torch-fidelity for inception score: https://github.com/toshas/torch-fidelity
- Inception score (**IS**): https://arxiv.org/pdf/1606.03498.pdf
  - Salimans et al. “Improved Techniques for Training GANs“
  - Perceptual score to evaluate GAN images based on inception-v3 pre-trained model
- Frechet inception distance (**FID**): https://arxiv.org/pdf/1706.08500.pdf
  - Heusel et al. “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium”

$$d^2((\boldsymbol{m}, \boldsymbol{C}), (\boldsymbol{m}_w, \boldsymbol{C}_w)) = \|\boldsymbol{m} - \boldsymbol{m}_w\|_2^2 + \mathrm{Tr}\left(\boldsymbol{C} + \boldsymbol{C}_w - 2\left(\boldsymbol{C}\boldsymbol{C}_w\right)^{1/2}\right)$$

- Kernel inception distance (**KID**): https://arxiv.org/pdf/1801.01401.pdf
  - Binkovski et al. “Demystifying MMD GANs”

$$k(x, y) = \left(\tfrac{1}{d} x^{\mathsf{T}} y + 1\right)^3$$





VAE Stochastic Model Comparisons

- ELBO, NLL, inception scores for various posterior-prior pairs with different likelihood beliefs

	$\mathcal{N}\|\mathcal{N}$	$\mathcal{L}_a\|\mathcal{N}$	$\mathcal{L}_o\|\mathcal{N}$	$\mathcal{U}\|\mathcal{N}$	$\mathcal{G}\|\mathcal{N}$	$\mathcal{E}\|\mathcal{N}$	$\mathcal{N}\|\mathcal{L}_a$	$\mathcal{L}_a\|\mathcal{L}_a$	$\mathcal{L}_o\|\mathcal{L}_a$	$\mathcal{E}\|\mathcal{L}_a$	$\mathcal{C}\|\mathcal{C}$	$\mathcal{U}\|\mathcal{C}$	$\mathcal{G}\|\mathcal{G}$	Auto
(a) Likelihood Belief $P=\mathcal{N}(\lambda,*)$: Unspecified Normal Distribution, i.e., MSE Loss														
$\hat{\mathcal{L}}_{1,1}$	-19.74	-20.35	-19.23	-22.47	-20.60	-39.76	-19.74	-19.42	-19.39	-34.06	-26.56	-26.33	-19.49	**-19.01**
MSE	12.71	12.76	**12.59**	12.84	12.91	20.0	12.61	13.00	12.72	16.76	26.44	12.84	13.14	12.74
FID	119.0	119.4	**113.2**	142.6	139.2	126.0	119.4	126.9	120.7	223.5	348.4	147.2	134.7	126.1
KID	0.125	0.127	**0.118**	0.138	0.154	0.164	0.145	0.133	0.126	0.246	0.528	0.142	0.149	0.135
(b) Likelihood Belief $P=\mathcal{B}(\lambda)$: Bernoulli Distribution, i.e., BCE Loss														
$\hat{\mathcal{L}}_{1,1}$	-102.5	-103.9	-102.9	-105.9	-103.5	-163.6	-102.7	-103.9	-103.2	-148.1	-203.4	-108.6	-103.4	**-101.6**
BCE	77.15	77.01	76.62	76.66	76.20	124.4	76.81	78.26	77.35	121.6	202.5	76.67	76.90	**76.30**
FID	42.91	43.50	44.01	42.76	41.82	113.27	40.80	41.59	42.13	152.6	389.3	42.42	40.88	**40.19**
KID	0.0369	0.0370	0.0378	0.0359	0.0349	0.1236	0.0347	0.0348	0.0360	0.1908	0.6302	0.0338	0.0344	**0.0337**
(c) Likelihood Belief $P=\mathcal{L}_a(\lambda,*)$: Unspecified Laplace Distribution, i.e., MAE Loss														
$\hat{\mathcal{L}}_{1,1}$	-65.34	-62.34	-61.83	-64.64	-66.26	-98.29	-62.18	-62.54	-62.27	-88.07	-98.73	-76.29	-65.47	**-61.08**
MAE	49.86	46.71	46.86	46.54	50.86	74.10	46.75	48.32	48.17	69.11	98.54	**44.80**	51.39	46.54
FID	46.02	46.91	48.85	46.41	52.19	159.8	50.48	48.00	48.55	174.2	219.9	102.7	54.58	**44.02**
KID	0.0343	0.0357	0.0375	0.0340	0.0428	159.8	0.0388	0.0342	0.0348	0.1767	0.2233	0.0845	0.0456	**0.0313**
(d) Likelihood Belief $P=\mathcal{N}(\lambda,\gamma^2)$: Normal Distribution														
$\hat{\mathcal{L}}_{1,1}$	1888.6	1899.2	1774.5	1819.8	-8000.1	658.6	2079.9	1521.5	2122.6	-30000	177.6	1969.7	2399.1	**2490.4**
NLL	-1954.8	-1976.0	-1839.4	-1886.3	552.1	-705.1	-2114.4	-1584.2	-2195.1	-748.3	-193.4	-2042.6	-2469.9	**-2575.0**
FID	170.2	167.9	161.5	162.3	294.3	267.7	182.0	167.9	177.2	321.1	423.4	283.4	**87.67**	98.19
KID	0.1982	0.1968	0.1840	0.1932	0.3624	0.3431	0.2195	0.2176	0.2057	0.7301	0.6555	0.3733	**0.0816**	0.0976
(e) Likelihood Belief $P=\mathcal{CB}(\lambda)$: Continuous Bernoulli Distribution														
$\hat{\mathcal{L}}_{1,1}$	1838.0	1835.4	1837.2	1833.2	1838.2	1656.3	**1840.8**	1837.4	1838.4	1656.3	1355.2	1834.8	1838.2	**1840.8**
NLL	-1882.1	-1880.9	-1881.4	-1881.2	-1883.3	-1678.3	-1885.4	-1883.5	-1883.1	-1696.0	-1356.9	**-1885.7**	-1882.4	-1883.9
FID	61.25	63.05	62.21	64.17	57.27	116.13	62.28	60.06	59.21	132.9	318.1	66.85	**52.85**	55.86
KID	0.0576	0.0589	0.0586	0.0609	0.0514	0.1223	0.0597	0.0565	0.0546	0.1467	0.7745	0.0611	**0.0471**	0.0501
(f) Likelihood Belief $P=\mathcal{B}_e(\lambda,\gamma)$: Beta Distribution														
$\hat{\mathcal{L}}_{1,1}$	6970.2	6062.9	4136.2	4916.3	5362.9	—	5450.7	4301.3	6258.4	7214.3	—	5510.9	5866.2	**8044.7**
NLL	-7053.0	-6158.5	-4223.9	-5002.0	-5430.4	—	-5501.7	-4381.6	-6353.5	-7275.0	—	-5588.8	-5946.9	**-8122.7**
FID	136.44	134.61	135.23	141.30	103.69	—	127.74	134.68	133.45	204.28	—	119.23	**98.15**	98.64
KID	0.1540	0.1525	0.1533	0.1591	0.1119	—	0.1423	0.1514	0.1516	0.2134	—	**0.1036**	0.1050	0.1070
(g) Likelihood Belief $P=\mathcal{B}(\lambda)$: Bernoulli Distribution, i.e., BCE Loss; Rényi order $\alpha=0$, i.e., IWAE														
$\hat{\mathcal{L}}_{0,50}$	-94.08	-93.63	-93.67	-98.54	-94.44	-132.6	-94.35	-93.87	-94.04	-119.6	-98.06	-100.34	-94.28	**-93.44**
BCE	77.04	76.56	76.70	77.73	**76.55**	103.9	77.50	77.01	77.45	97.02	78.12	79.16	77.10	76.81
FID	38.18	37.01	37.24	42.76	38.13	80.30	41.35	38.09	40.29	111.99	**35.71**	36.82	36.99	36.43
KID	0.0310	0.0299	0.0304	0.0359	0.0321	0.0803	0.0354	0.0316	0.0342	0.1180	**0.0256**	0.0268	0.0312	0.0293

Unspecified Normal NLL

Bernoulli NLL

Unspecified Laplace NLL

Normal NLL

Continuous Bernoulli NLL

Beta NLL

Bernoulli NLL 50-IWAE

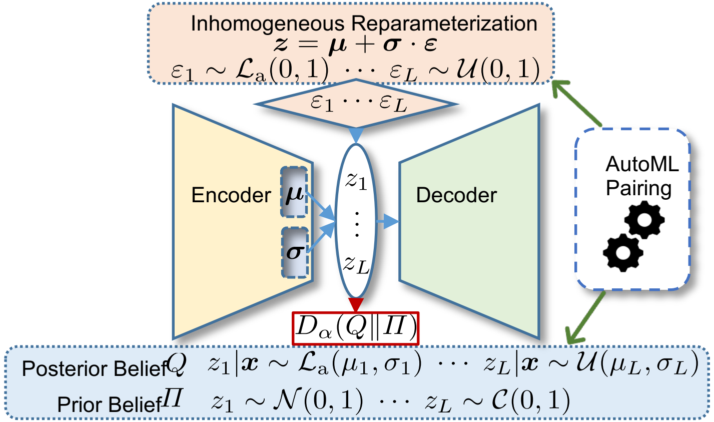
Brute-force Pairing: 16^{20}

AutoML

Latent:
50% $\mathcal{L}_o\|\mathcal{N}$
30% $\mathcal{L}_a\|\mathcal{N}$
20% $\mathcal{L}_a\|\mathcal{L}_a$

Summary

- We overviewed trends of **generative AI**
- We proposed **AutoVAE** framework:
  - Automated search of **posterior/prior/likelihood beliefs** besides architecture exploration
  - **Mismatched** posterior-prior pairing (e.g., logistic posterior for normal prior)
  - Heterogenous **irregular** posterior-prior pairing (e.g., 70% logistic-normal; 30% Cauchy-Cauchy)
  - Auto-selection of **Renyi order** for alpha divergence as an extended KLD discrepancy measure
  - Diverse negative likelihood beliefs as a reconstruction loss
- Proposed AutoVAE demonstrated the benefit for some benchmark datasets
  - **ELBO** (variational Renyi bound) analysis
  - Image synthesis snapshot
  - **Inception** score analysis
- Questions?
  - koike@merl.com



Probability Distribution Notations

- Notations and probability distribution functions (PDE)

Distribution	Notation	PDF $f(x)$
Normal	$\mathcal{N}(\mu,\sigma)$	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$
Laplace	$\mathcal{L}_{\mathrm{a}}(\mu,\sigma)$	$\frac{1}{2\sigma}\exp\left(-\frac{\lvert x-\mu\rvert}{\sigma}\right)$
Cauchy	$\mathcal{C}(\mu,\sigma)$	$\frac{1}{\pi}\frac{\sigma}{\sigma^2+(x-\mu)^2}$
Logistic	$\mathcal{L}_{\mathrm{o}}(\mu,\sigma)$	$\frac{1}{\sigma}\left(\exp\left(\frac{x-\mu}{2\sigma}\right)+\exp\left(\frac{\mu-x}{2\sigma}\right)\right)^{-2}$
Uniform	$\mathcal{U}(\mu,\sigma)$	$\frac{1}{2\sigma},\quad \mu-\sigma\leq x\leq\mu+\sigma$
Gumbel	$\mathcal{G}(\mu,\sigma)$	$\frac{1}{\sigma}\exp\left(-\frac{x-\mu}{\sigma}-\exp\left(-\frac{x-\mu}{\sigma}\right)\right)$
Exponential	$\mathcal{E}(\sigma)$	$\frac{1}{\sigma}\exp\left(-\frac{x}{\sigma}\right),\quad x\geq 0$
Bernoulli	$\mathcal{B}(\lambda)$	$\lambda^x(1-\lambda)^{1-x},\quad x\in\{0,1\}$
Cont. Bernoulli [4]	$\mathcal{CB}(\lambda)$	$C(\lambda)\lambda^x(1-\lambda)^{1-x},\quad 0\leq x\leq 1$
Beta	$\mathcal{B}_{\mathrm{e}}(\lambda,\gamma)$	$\frac{\Gamma(\lambda+\gamma)}{\Gamma(\lambda)\Gamma(\gamma)}x^{\lambda-1}(1-x)^{\gamma-1}$

References

- [1] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,”*arXiv preprint arXiv:1312.6114*, 2013.
- [2] D. J. Rezende, S. Mohamed, and D. Wierstra, “Stochastic backpropagation and approximate inference in deep generative models,” in *International conference on machine learning*. PMLR,2014,pp.1278– 1286.
- [3] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, “β-VAE: Learning basic visual concepts with a constrained variational framework,” in *International Conference on Learning Representations*, 2017.
- [4] G. Loaiza-Ganem and J. P. Cunningham, “The continuous Bernoulli: fixing a pervasive error in variational autoencoders,” *arXiv preprint arXiv:1907.06845*, 2019.
- [5] G. Barello, A. S. Charles, and J. W. Pillow, “Sparse-coding variational auto-encoders,” *bioRxiv*, p. 399246, 2018.
- [6] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, “Improved variational inference with inverse autoregressive flow,” *Advances in neural information processing systems*, vol. 29, pp. 4743–4751, 2016.
- [7] Y. Burda, R. Grosse, and R. Salakhutdinov, “Importance weighted autoencoders,” *arXiv preprint arXiv:1509.00519*, 2015.
- [8] C. Cremer, Q. Morris, and D. Duvenaud, “Reinterpreting importance- weighted autoencoders,” *arXiv preprint arXiv:1704.02916*, 2017.
- [9] P. Alquier, J. Ridgway, and N. Chopin, “On the properties of variational approximations of Gibbs posteriors,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 8374–8414, 2016.
- [10] Y. Li and R. E. Turner, “Renyi divergence variational inference,”*arXiv preprint arXiv:1602.02311*, 2016.
- [11] J. Knoblauch, J. Jewson, and T. Damoulas, “Generalized variational inference: Three arguments for deriving new posteriors,” *arXiv preprint arXiv:1904.02063*, 2019.
- [12] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next generation hyperparameter optimization framework,” in *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.
- [13] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: A review for statisticians,” *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.
- [14] T. Van Erven and P. Harremos, “Renyi divergence and Kullback–Leibler divergence,” *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.
- [15] M. Gil, F. Alajaji, and T. Linder, “Renyi divergence measures for commonly used univariate continuous distributions,” *Information Sciences*, vol. 249, pp. 124–131, 2013.
- [16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local Nash equilibrium,” *Advances in neural information processing systems*, vol. 30, 2017.
- [17] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton, “Demystifying MMD GANs,” *arXiv preprint arXiv:1801.01401*, 2018.

MITSUBISHI ELECTRIC

Changes for the Better