Tutorial of Information Geometry

&

t^3-Variational Autoencoder: Learning Heavy-Tailed Data With Student's t and Power Divergence

Xin Gao

2024.4.27

Content

- Preliminary
  - VAE from Joint Minimization insights
  - Information Geometry (IG)
  - Student's t Distribution
- Introduction
- The t^3-Variational Autoencoder
- Heavy-tailed distribution experiment

Published as a conference paper at ICLR 2024

t^3-VARIATIONAL AUTOENCODER: LEARNING HEAVY-TAILED DATA WITH STUDENT'S T AND POWER DIVERGENCE

Juno Kim[1,2]* **Jaehyuk Kwon**[3]* **Mincheol Cho**[3]* **Hyunjong Lee**[3] **Joong-Ho Won**[3]
[1]Department of Mathematical Informatics, The University of Tokyo
[2]Center for Advanced Intelligence Project, RIKEN
[3]Department of Statistics, Seoul National University
junokim@g.ecc.u-tokyo.ac.jp {jh19984,code1478,hyunjong526}@snu.ac.kr wonj@stats.snu.ac.kr

Preliminary VAE Notation Recap

- A VAE models the distribution $p_{data}(x)$ of the observed variable $x \in \mathbb{R}^n$ by jointly learning a stochastic latent variable $z \in \mathbb{R}^m$.
- Generation is performed by sampling z from the prior $p_Z(z)$, then sampling x according to a probabilistic **decoder** $p_\theta(x|z)$ parametrized by $\theta \in \Theta$.
- The observed likelihood $p_\theta(x) = \int p_\theta(x|z)p_Z(z)dz$ is intractable, so we instead aim to approximate the posterior $p_\theta(z|x)$ with a parametrized **encoder** $q_\phi(z|x)$ by minimizing their KL divergence. This leads to maximizing **the evidence lower bound (ELBO)** of the log-likelihood, defined as

$$\begin{aligned}\mathcal{L}(x;\theta,\phi) :&= \log p_\theta(x) - \mathcal{D}_{\mathrm{KL}}(q_\phi(z|x) \parallel p_\theta(z|x)) \\ &= \mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[\log p_\theta(x|z)\right] - \mathcal{D}_{\mathrm{KL}}(q_\phi(z|x) \parallel p_Z(z)).\end{aligned}$$

Preliminary From EM to VAE

- The proposal posterior $q(z)$

$$\log p(x) = \int q(z) \log p(x) dz$$

$$= \int q(z) \log \frac{p(x \mid z)p(z)}{p(z \mid x)} \frac{q(z)}{q(z)} dz$$

$$= \int q(z) \log p(x \mid z) dz - KL(q(z)|p(z)) + KL(q(z) \mid p(z \mid x))$$

ELBO KL divergence

- In EM: calculate $q(z) = p(z|x)$, $KL = 0$, max $ELBO$ ($= \log P(x)$)
- In VAE: intractable $p(z|x)$, max $ELBO$ ($\leq \log P(x)$) $\Longrightarrow$ min KL

$\Longrightarrow$ Using encoder $q_\phi(z|x)$ to approximate $p(z|x)$

Preliminary VAE from Joint Minimization insights

- **Model distribution manifold**: $\mathcal{P} = \{p_\theta(x, z) = p_\theta(x|z)p_Z(z): \theta \in \Theta\}$
- **Data distribution manifold**: $\mathcal{Q} = \{q_\phi(x, z) = p_{data}(x)q_\phi(z|x): \phi \in \Phi\}$

(Both finite-dimensional submanifolds of the space of joint distributions)

- The VAE can be reinterpreted as a joint minimization process between two statistical manifolds [1].

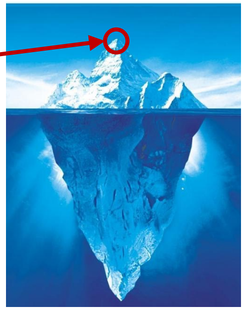
$$\begin{aligned} D_{\mathrm{KL}}(q_\phi(x, z) \parallel p_\theta(x, z)) &= \mathbb{E}_{x \sim p_{\mathrm{data}}}\left[-\log p_\theta(x) + \mathcal{D}_{\mathrm{KL}}(q_\phi(z|x) \parallel p_\theta(z|x))\right] - H(p_{\mathrm{data}}) \\ &= -\mathbb{E}_{x \sim p_{\mathrm{data}}}\left[\mathcal{L}(x; \theta, \phi)\right] - H(p_{\mathrm{data}}). \end{aligned}$$

- **Minimizing the divergence between points on the model and data distribution manifolds is equivalent to maximizing the expected ELBO.**

$$(p_{\theta^*}, q_{\phi^*}) = \underset{p \in \mathcal{P}, q \in \mathcal{Q}}{\operatorname{argmin}} \mathcal{D}_{\mathrm{KL}}(q \| p).$$

- Can be solved by *em*-projection algorithm on manifolds
- Can be substituted by other divergences / statistical families

Information Geometry (1/10)

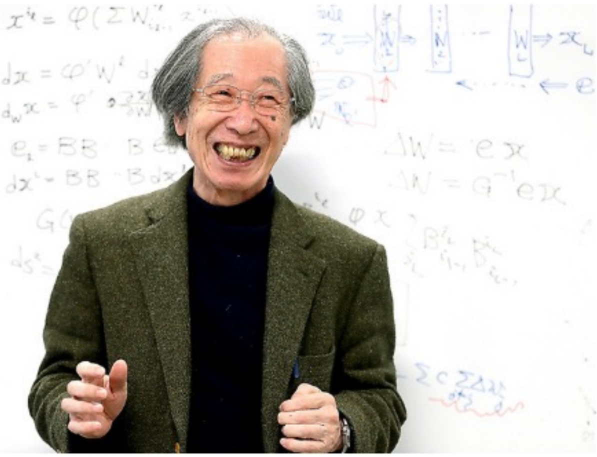


Brief Introduction

- Information geometry aims to elucidate the geometry of the space of probability distributions.
- Generative models can be understood within the framework of information geometry, where each probability distribution is treated as a point, and different families of probability distributions form different manifolds.

Application

- Applied to diverse research fields including machine learning, signal processing, neuroscience and physics, where probability distribution matters.

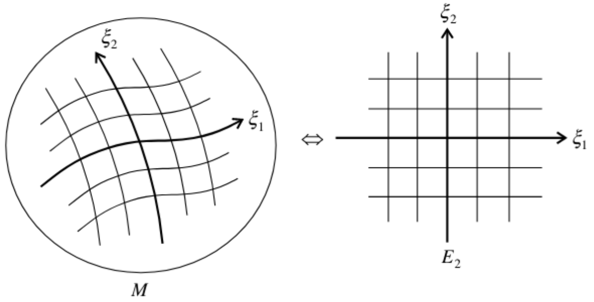


Shun-ichi Amari

Information Geometry (2/10)

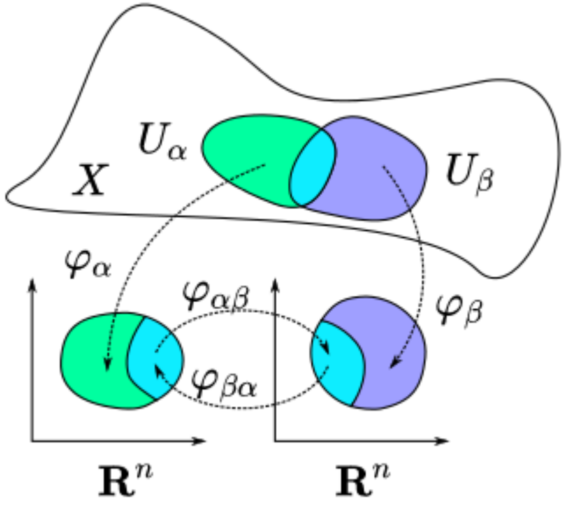
Manifold and Coordinate Systems

- **Manifold**: An n-dimensional topological manifold M is a topological Hausdorff space with a countable base which is locally homeomorphic to $\mathbb{R}^n$. This means that for every point p in M there is an open neighborhood U of p and a homeomorphism $\varphi: U \to V$ which maps the set U onto an open set $V \subset \mathbb{R}^n$.
  - The mapping $\varphi: U \to V$ is called a chart or coordinate system.
  - The image of the point $p \in U$, denoted by $\varphi(p) \in \mathbb{R}^n$, is called the coordinates or local coordinates of p in the chart.

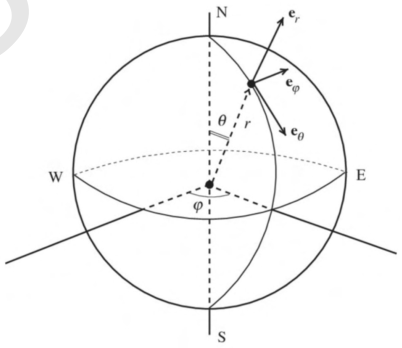


- **Statistical manifold** [2] : Each point is a probability distribution.

A family of probability distributions $M = \{p(x, \xi)\}$ specified by a vector parameter ξ. ξ is the coordinate.



- **Example: 2D surface of 3D sphere**



Information Geometry (3/10)

Why manifold for probability distributions?

- **Space of normal distributions, Coordinate $\boldsymbol{\mu}, \boldsymbol{\sigma}^2$**

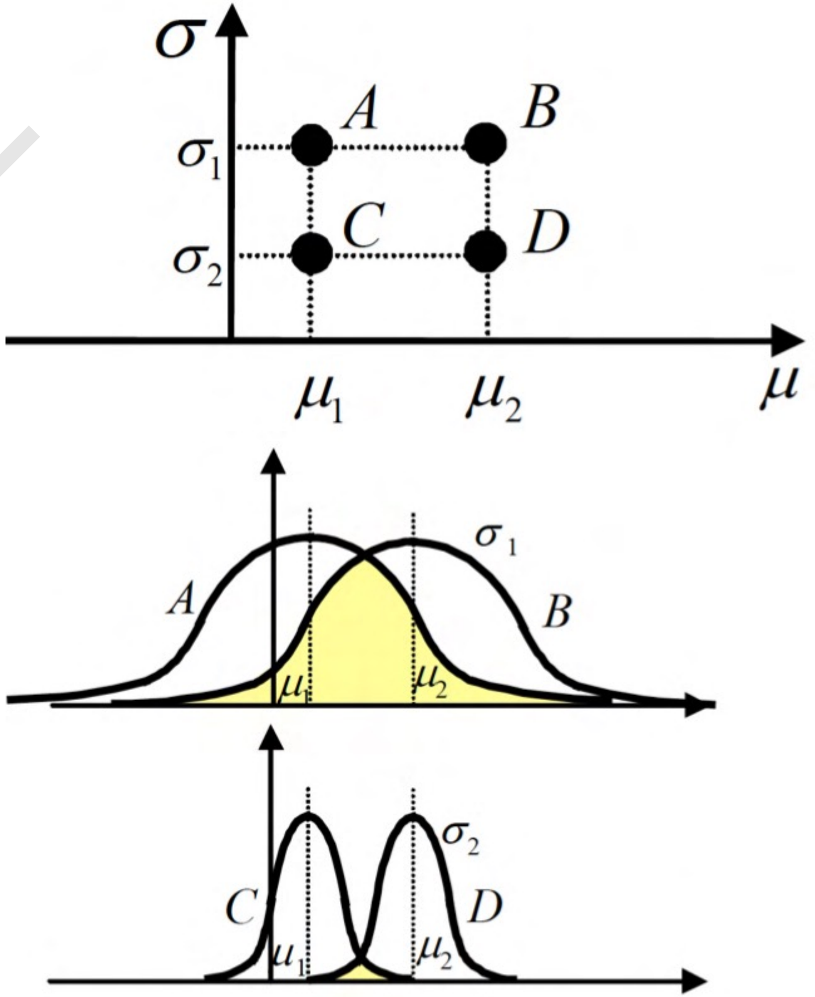
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

If a simple Euclidean distance is used, distance $|AB| = |CD|$

But...

distance $|AB|$ and $|CD|$ should not be the same.

A different metric of distance is necessary.



- What is the metric and connection in the **manifold** of probability distributions?

Information Geometry (4/10)

Divergence and Riemannian Metric

- **Divergence**

Def. A divergence is a kind of statistical distance: a binary function which establishes **the separation** from one probability distribution to another on a statistical manifold.

① Non-negativity ② Positivity ③ $\mathcal{D}(p_\theta \parallel p_{\theta+d\theta}) = \frac{1}{2}\sum_{i,j=1}^{d} g_{ij}(\theta)d\theta_i d\theta_j + O(\|d\theta\|^3)$

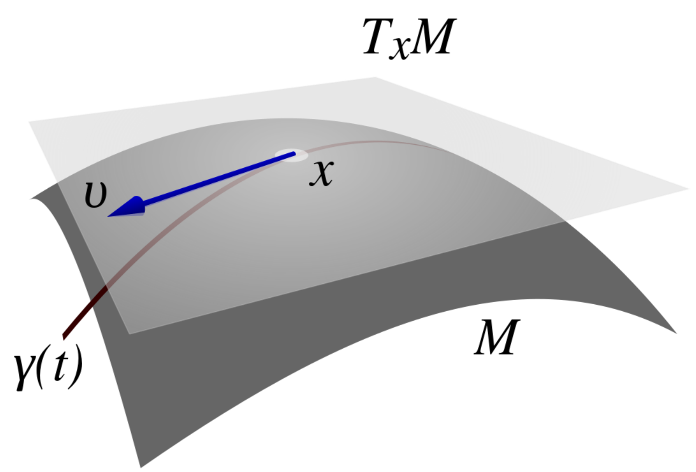
g_{ij} → **Symmetric positive-definite matrix**

- **The dual divergence** $\mathcal{D}^*$ is defined as $\mathcal{D}^*(p,q) = \mathcal{D}(q,p)$.

- **Riemannian metric**

A divergence D provides M with a Riemannian structure.

Def. Tangent space T_pM at point p, a positive-definite inner product $g_p: T_pM \times T_pM \to \mathbb{R}$. The smooth manifold endowed with this metric g is a Riemannian manifold, denoted (M, g).



Information Geometry (5/10)

Affine Connection and Parallel transport

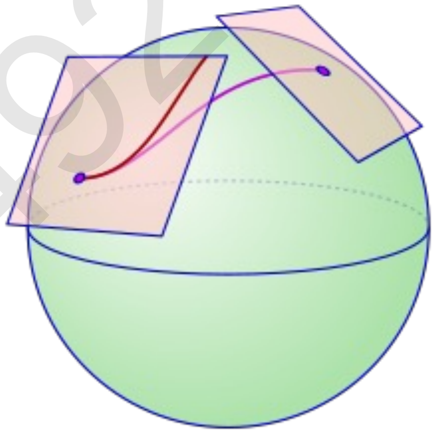
- An **affine connection** is a geometric object on a smooth manifold which connects nearby tangent spaces.

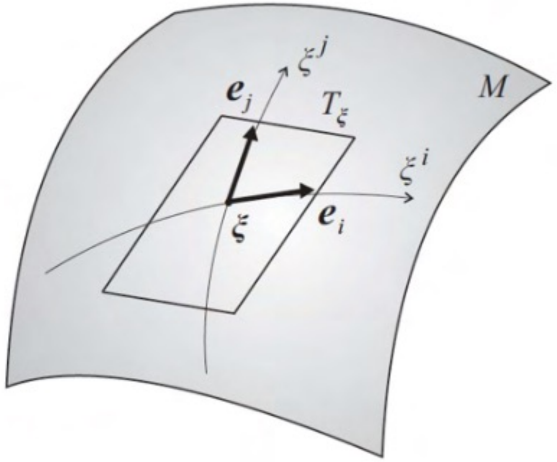
Two nearby tangent spaces T_ξ and $T_{\xi+d\xi}$:

$$T_\xi \qquad e_1(\boldsymbol{\xi}), e_2(\boldsymbol{\xi}), \dots, e_n(\boldsymbol{\xi})$$

$$T_{\xi+d\xi} \qquad e_1(\boldsymbol{\xi}+d\boldsymbol{\xi}), e_2(\boldsymbol{\xi}+d\boldsymbol{\xi}), \dots, e_n(\boldsymbol{\xi}+d\boldsymbol{\xi})$$

$$\boldsymbol{e}_i(\boldsymbol{\xi}+d\boldsymbol{\xi}) = \boldsymbol{e}_i(\boldsymbol{\xi}) + \sum \boxed{\Gamma_{ki}^{j}} \boldsymbol{e}_j(\boldsymbol{\xi}) d\xi^k$$



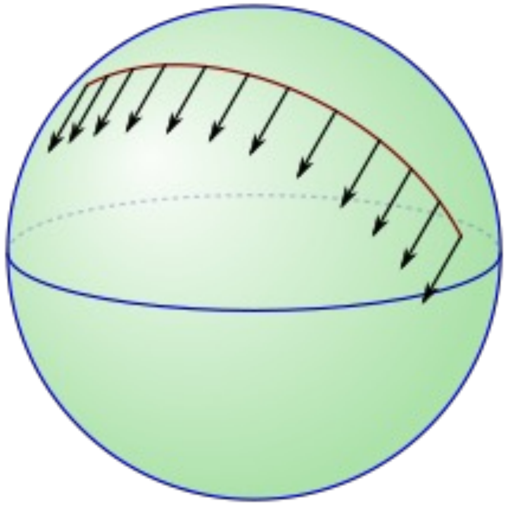


Divergence $\mathcal{D}(\cdot,\cdot)$ can define an affine connection $\Gamma_{ij,k}^{(\mathcal{D})} = -\dfrac{\partial^3}{\partial\theta_p^i \partial\theta_p^j \partial\theta_q^k} \mathcal{D}(p,q)\Big|_{p=q}$

- **Parallel transport and Geodesic**

If the manifold has an affine connection, it enables vector transport along curves to maintain parallelism relative to the connection.

- Tangent vectors along the geodesic maintain the same direction.



Information Geometry (6/10)

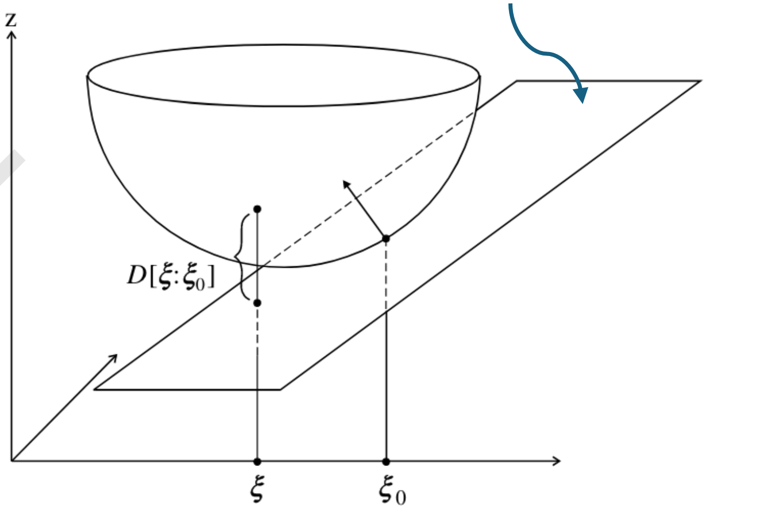
Bregman Divergence & Legendre Transformation

- **Bregman Divergence from convex function $\psi(\xi)$**

Since ψ is convex, it is always above the hyperplane, touching it at ξ_0. Hence, it is a supporting hyperplane of ψ at ξ_0

$$D_\psi\left[\boldsymbol{\xi}:\boldsymbol{\xi}_0\right]=\psi(\boldsymbol{\xi})-\psi\left(\boldsymbol{\xi}_0\right)-\nabla\psi\left(\boldsymbol{\xi}_0\right)\cdot\left(\boldsymbol{\xi}-\boldsymbol{\xi}_0\right).$$

$$z=\psi\left(\boldsymbol{\xi}_0\right)+\nabla\psi\left(\boldsymbol{\xi}_0\right)\cdot\left(\boldsymbol{\xi}-\boldsymbol{\xi}_0\right)$$



- **Legendre Transformation => A dualistic structure**

The gradient of $\psi(\xi)$: $\xi^*=\nabla\psi(\xi)$ is the normal vector n, we define a new function of ξ^* by

$$\psi^*\left(\boldsymbol{\xi}^*\right)=\boldsymbol{\xi}\cdot\boldsymbol{\xi}^*-\psi(\boldsymbol{\xi}),$$

and ξ is not free but is a function of ξ^*, $\xi=\mathrm{f}(\xi^*)$, which is the inverse function of $\xi^*=\nabla\psi(\xi)$

By differentiating $\psi^*(\xi^*)$, we have a dualistic structure

$$\boldsymbol{\xi}^*=\nabla\psi(\boldsymbol{\xi}),\quad \boldsymbol{\xi}=\nabla\psi^*\left(\boldsymbol{\xi}^*\right).$$

ψ^* is called the Legendre dual of ψ

$\psi^*(\xi^*)$ is a convex function, thus a new Bregman divergence is derived from the dual convex function $\psi^*(\xi^*)$, and

$$D_{\psi^*}\left[\boldsymbol{\xi}^*:\boldsymbol{\xi}^{*\prime}\right]=D_\psi\left[\boldsymbol{\xi}^{\prime}:\boldsymbol{\xi}\right].$$

Information Geometry (7/10)

Dually flat manifold and structure

- **Flat manifold** ⟺ Affine coordinate
- **Dual affine connections** $\Gamma_{ijk}(\boldsymbol{\xi}) = -\frac{\partial^2}{\partial \xi^i \partial \xi^j}\frac{\partial}{\partial \xi'^k} D(\boldsymbol{\xi}||\boldsymbol{\xi}')$ $\Gamma^*_{ijk}(\boldsymbol{\xi}) = -\frac{\partial}{\partial \xi^k}\frac{\partial^2}{\partial \xi'^i \partial \xi'^j} D(\boldsymbol{\xi}||\boldsymbol{\xi}')$

Dual metric condition: $\langle A, B\rangle_{\xi_0} = \langle \Pi A, \Pi^* B\rangle_{\xi_1}$ Π, Π^*: parallel transport using Γ, Γ^*

Theorem. A dually flat manifold S has two special coordinate systems denoted by $\theta = (\theta_1,\cdots,\theta_n)$ and $\eta = (\eta_1,\cdots,\eta_n)$ such that θ is an affine coordinate system of ∇-connection and η is an affine coordinate system of ∇^*-connection. There exist two potential functions $\psi(\boldsymbol{\theta})$ and $\varphi(\boldsymbol{\eta})$ which are strictly convex, and are connected by the Legendre transformation such that $\psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}) - \sum \theta^i \eta_i = 0,$

where $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are the respective coordinates of the same point. S has a canonical divergence between two points P and Q defined by

$$D[P:Q] = \psi(\boldsymbol{\theta}_P) + \varphi(\boldsymbol{\eta}_Q) - \sum \theta_P^i \eta_{Qi}$$

where $\boldsymbol{\theta}_P$ and $\boldsymbol{\eta}_Q$ are respective coordinates of points P and Q.

Information Geometry (8/10)

Example: Exponential Family

$$p(\boldsymbol{x}, \boldsymbol{\theta})dx = \exp[\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta})]\, d\mu(\boldsymbol{x})$$ **Dually flat manifold**

- **Dual convex functions**

$$\psi(\boldsymbol{\theta}) = \log \int \exp(\boldsymbol{\theta} \cdot \boldsymbol{x}) d\mu(\boldsymbol{x})$$ **Free Energy** Convex function with respect to $\boldsymbol{\theta}$

$$\varphi(\boldsymbol{\eta}) = \int p(\boldsymbol{x}, \boldsymbol{\eta}) \log p(\boldsymbol{x}, \boldsymbol{\eta}) d\boldsymbol{x}$$ **Entropy** Convex function with respect to $\boldsymbol{\eta}$

- **Dual affine coordinates**

$$\theta^i = \frac{\partial \varphi(\boldsymbol{\eta})}{\partial \eta_i} \qquad \eta_i = \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta^i} = \int x_i p(\boldsymbol{x}, \boldsymbol{\theta}) d\mu(\boldsymbol{x})$$ The expectation of x

- **Canonical divergence**

$$D[\boldsymbol{\theta}_P : \boldsymbol{\theta}_Q] = \psi(\boldsymbol{\theta}_P) + \varphi(\boldsymbol{\eta}_Q) - \theta_P^i \eta_{Qi} = \int p(\boldsymbol{x}, \boldsymbol{\theta}_P) \log \frac{p(\boldsymbol{x}, \boldsymbol{\theta}_P)}{p(\boldsymbol{x}, \boldsymbol{\theta}_Q)} d\boldsymbol{x}$$ **Kullback-Leibler divergence**

- **Fisher information matrix**

$$D\left(p(x, \boldsymbol{\xi}) || p(x, \boldsymbol{\xi} + d\boldsymbol{\xi})\right) = \frac{\partial}{\partial \xi'^i} D[\boldsymbol{\xi} || \boldsymbol{\xi}']_{\xi'=\xi} d\xi^i + \frac{1}{2} \frac{\partial^2}{\partial \xi'^i \partial \xi'^j} D(\boldsymbol{\xi} || \boldsymbol{\xi}')_{\xi'=\xi} d\xi^i d\xi^j$$
$$= \frac{1}{2} E_\xi [\partial_i \log p(x, \boldsymbol{\xi}) \partial_j \log p(x, \boldsymbol{\xi})] d\xi^i d\xi^j = \frac{1}{2} g_{ij} d\xi^i d\xi^j$$

Fisher information metric can be derived from the second derivative of KL divergence.

Information Geometry (9/10)

Generalized Pythagorean Theorem

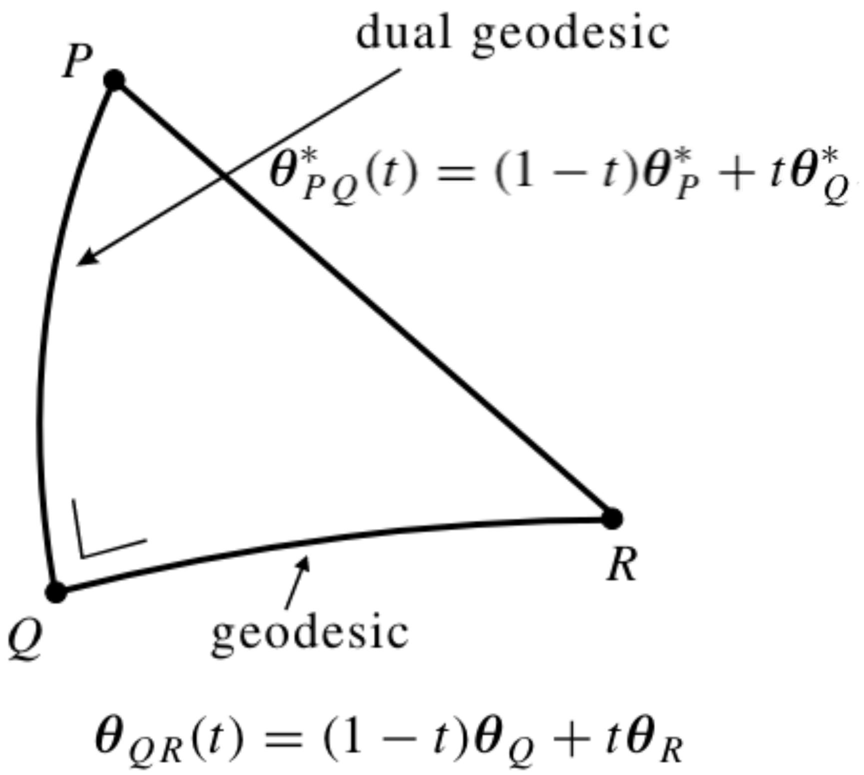
- **Theorem** (Generalized Pythagorean Theorem): When triangle PQR is orthogonal such that the dual geodesic connecting P and Q is orthogonal to the geodesic connecting Q and R, the following generalized Pythagorean relation holds:

$$D_\psi(R : P) = D_\psi(Q : P) + D_\psi(R : Q).$$

- Dual version: The geodesic connecting P and Q is orthogonal to the dual geodesic connecting Q and R, then

$$D_{\psi^*}(R : P) = D_{\psi^*}(Q : P) + D_{\psi^*}(R : Q).$$

- **Theorem:** The canonical divergence function of a dually flat manifold satisfies the Pythagorean relation, when ∇^*-geodesic connection P and Q is orthogonal at Q to ∇-geodesic connecting Q and R.



Information Geometry (10/10)

em-Projection Theorem

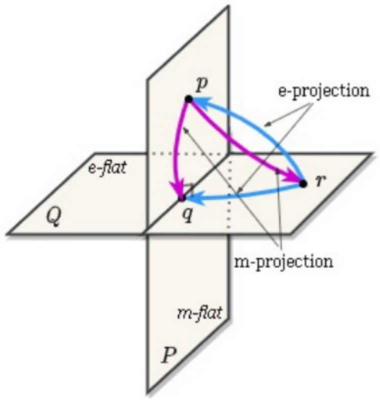
- **Theorem:** Given $P \in M$ and a smooth submanifold $S \subset M$, the point that minimizes the divergence $D_\psi[P : R]$, $R \in S$, is the dual geodesic projection of P to S. The point that minimizes the dual divergence $D_{\psi^*}[P : R]$, $R \in S$, is the geodesic projection of P to S.

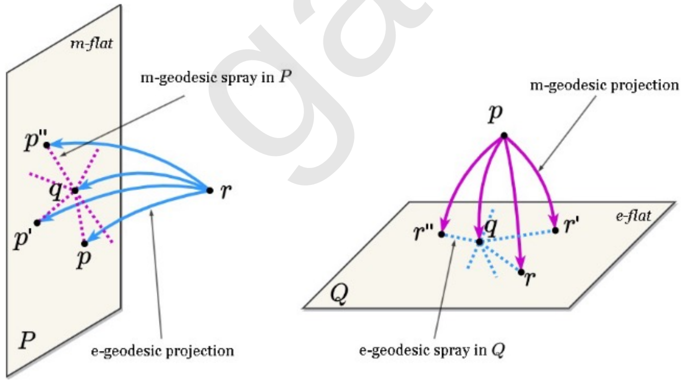
(Dually flat manifold: If exist, then unique)

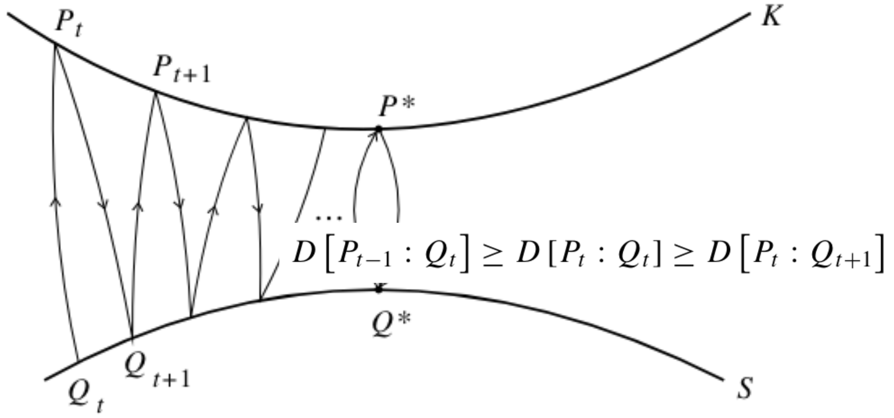
- **Divergence Between Submanifolds: Alternating Minimization Algorithm**

Two submanifolds K and S in a dually flat M, we define a divergence between K and S by

$$D[K : S] = \min_{P \in K, Q \in S} D[P : Q] = D\left[\bar{P} : \bar{Q}\right].$$





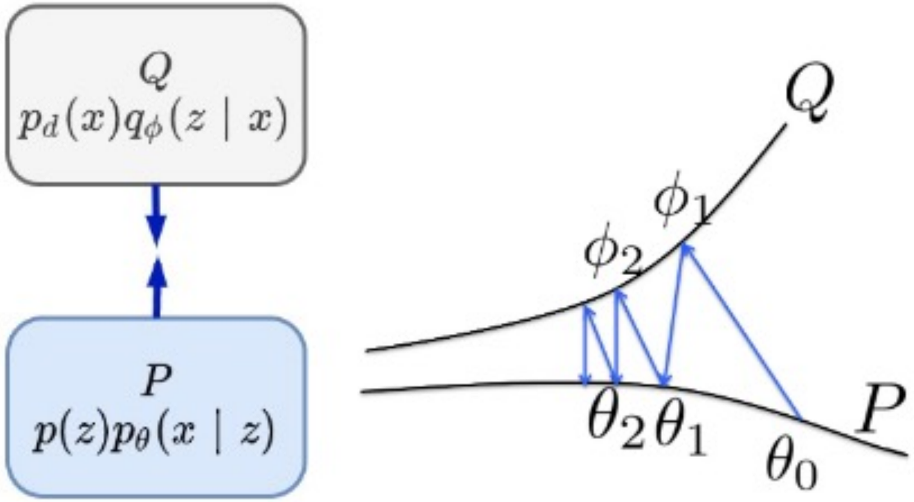


VAE from IG insights

- **Model distribution manifold**: $\mathcal{P} = \{p_\theta(x, z) = p_\theta(x|z)p_Z(z): \theta \in \Theta\}$
- **Data distribution manifold**: $\mathcal{Q} = \{q_\phi(x, z) = p_{data}(x)q_\phi(z|x): \phi \in \Phi\}$

$$(p_{\theta^*}, q_{\phi^*}) = \underset{p\in\mathcal{P}, q\in\mathcal{Q}}{\operatorname{argmin}} \mathcal{D}_{\mathrm{KL}}(q\|p).$$

Exponential family with KL divergence is a dually flat manifold, $\mathcal{P}$ and $\mathcal{Q}$ are flat submanifolds. The *em*-projection theorem guarantees its convergence.



Additionally, this framework readily accommodates alternative divergences and extends to encompass broader statistical manifolds.

Preliminary Student's t Distribution

- The VAE framework a priori does not require the prior, encoder or decoder to be a particular probability distribution; the usual choice of Gaussian is mainly due to feasibility of **the reparameterization trick** and **closed-form computation of divergence**.
- **PDF**: The family of d-variate Student's t-distributions with variable mean μ, scale matrix Σ and fixed degrees of freedom ν

$$t_d(x|\mu,\Sigma,\nu) = C_{\nu,d}|\Sigma|^{-\frac{1}{2}}\left(1+\frac{1}{\nu}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)^{-\frac{\nu+d}{2}}, \quad C_{\nu,d} = \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{\frac{d}{2}}}$$

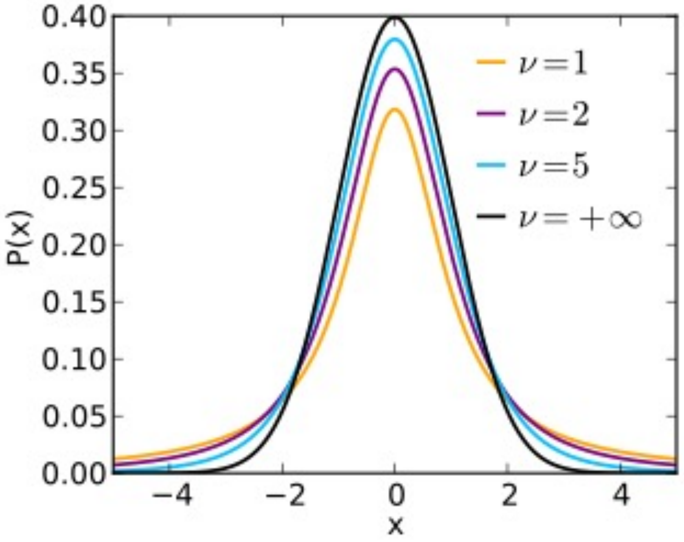
- **Relation with Gaussian distribution** $\nu \to \infty, \quad \to \mathcal{N}(\mu,\Sigma)$
  - **Visualization:** Right figure
  - **Intuition:**

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \sim \frac{1}{\sqrt{2\pi}}(1-\frac{x^2}{2}+O(x^2))$$

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1+\frac{x^2}{\nu})^{-\frac{\nu+1}{2}} \sim \sqrt{\frac{1+\nu}{2\pi\nu}}(1-\frac{\nu+1}{2\nu}x^2+O(x^2))$$

(Taylor expansion and Stirling approximation for gamma function)

  - **Theoretical proof** [3, 4]



Preliminary Student's t Distribution

- **Moments:** For $\nu > 1$, the moments of the t distribution are

$$\mathbb{E}\{\ T^k\ \} = \begin{cases} 0 & k \text{ odd}, \quad 0 < k < \nu, \\ \nu^{\frac{k}{2}} \prod_{j=1}^{k/2} \frac{2j-1}{\nu-2j} & k \text{ even}, \quad 0 < k < \nu. \end{cases}$$

Moments of order ν or higher do not exit.

$\mathbb{E}(X) = \nu, for\ \nu > 1$; $Var(X) = \frac{\nu}{\nu-1}\Sigma,\ \ for\ \nu > 2,\ \infty\ for\ 1 < \nu \leq 2$

- **How to generate an RV subjected to student's t distribution?**

A multivariate t-distribution $T \sim t_d(\mu, \Sigma, \nu)$ may be constructed from a multivariate centered Gaussian $Z \sim \mathcal{N}_d(0, \Sigma)$ and an independent chi-squared variable $V \sim \chi^2(\nu)$ via

$$T \overset{d}{=} \mu + \frac{Z}{\sqrt{V/\nu}}.$$

- **Heavy-tailed distribution:** In probability theory, heavy-tailed distributions are probability distributions whose tails are not exponentially bounded: that is, they have heavier tails than the exponential distribution.

Introduction

- Motivation
  - Real-world data frequently displays **heavy-tailed** and **imbalanced** patterns.
  - The Gaussian prior is too tight to effectively fit complex latent representations; ‘**over-regularization**’.
  - Distributing more mass to the tails allows encoded points to spread out easily.

- Contribution
  - **t^3VAE**: a complete VAE framework that incorporates **Student’s t-distributions** for the prior, encoder, and decoder.
  - **Experiments**: t^3VAE effectively models the **low-density regions** of **heavy-tailed datasets** and generates **high-dimensional images with richer detail**.
  - **Extension**: Introducing **a hierarchical architecture** enables the reconstruction of high-resolution images with enhanced sophistication.

The t³-VAE

- **γ divergence:** $D_\gamma(q\|p) := \gamma^{-1}C_\gamma(q,p) - \gamma^{-1}H_\gamma(q)$

$$\mathcal{H}_\gamma(p) := -\|p\|_{1+\gamma} = -\left(\int p(x)^{1+\gamma}dx\right)^{\frac{1}{1+\gamma}}, \quad \mathcal{C}_\gamma(q,p) := -\int q(x)\left(\frac{p(x)}{\|p\|_{1+\gamma}}\right)^{\gamma}dx$$

- Computing the dual connections yields that the totally Γ^*-geodesic submanifolds consist of power families of the form

$$\mathcal{S}_\gamma = \{p_\theta(x) \propto (1+\gamma\theta^\top s(x))^{\frac{1}{\gamma}} : \theta \in \Theta\}.$$

- The family of d-variate Student's t-distributions is Γ^*-geodesic when $\gamma = -\frac{2}{\nu+d}$.

Statistical Manifold	Bregman Divergence	Riemannian metric	Dually flat structure	Flat Submanifold
Exponential family	KL divergence	Fisher information metric	The natural parameters and expectation parameters	Gaussian distribution $p(x,z)$
Power family	γ divergence	$g_{ij}(\theta) = -\frac{\partial^2}{\partial\theta_i\partial\theta'_j}\Big\|_{\theta'=\theta}\mathcal{D}_\gamma(p_\theta \parallel p_{\theta'})$	$\Gamma^k_{ij}(\theta) = -\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta'_k}\Big\|_{\theta'=\theta}\mathcal{D}(p_\theta \parallel p_{\theta'})$ $\Gamma^{*k}_{ij}(\theta) = -\frac{\partial^3}{\partial\theta'_i\partial\theta'_j\partial\theta_k}\Big\|_{\theta'=\theta}\mathcal{D}(p_\theta \parallel p_{\theta'})$	Student t distribution $p(x,z),\ \gamma = -\frac{2}{\nu+d}$

The t^3-VAE

- Define joint distribution $p_{\theta,\nu}(x,z)$ of a power form, parametrized by the degrees of freedom $\nu > 2$ and $\gamma = -\frac{2}{\nu+d}$.

$$p_{\theta,\nu}(x,z) \propto \sigma^{-n}\left[1+\frac{1}{\nu}\left(\|z\|^2+\frac{1}{\sigma^2}\|x-\mu_\theta(z)\|^2\right)\right]^{-\frac{\nu+m+n}{2}}$$

- **Prior-decoder pair**

$$p_{Z,\nu}(z) = \int p_{\theta,\nu}(x,z)dx = t_m(z|0,I,\nu)$$

$$p_{\theta,\nu}(x|z) = \frac{p_{\theta,\nu}(x,z)}{p_{Z,\nu}(z)} = t_n\left(x \mid \mu_\theta(z), \frac{1+\nu^{-1}\|z\|^2}{1+\nu^{-1}m}\sigma^2 I, \nu+m\right)$$

- **When** $\nu \to \infty$, $p_{\theta,\nu}(x,z) \to \mathcal{N}(\mu_\theta(z), \sigma^2 I)$.

The t³-VAE

- Since the true posterior $z|x$ is t-distributed with degrees of freedom $\nu + n$ when the decoder is shallow (Linear layer): $\mu_\theta(z) = Wz + b$

$$\begin{pmatrix} x \\ z \end{pmatrix} \propto \left[1 + \frac{1}{\nu\sigma^2} \begin{pmatrix} x - b \\ z \end{pmatrix}^\top \begin{pmatrix} I & -W \\ -W^\top & W^\top W + \sigma^2 I \end{pmatrix} \begin{pmatrix} x - b \\ z \end{pmatrix} \right]^{-\frac{\nu+m+n}{2}}$$

$$\propto t_{m+n}\left(\begin{pmatrix} b \\ 0 \end{pmatrix}, \begin{pmatrix} WW^\top + \sigma^2 I & W \\ W^\top & I \end{pmatrix}, \nu \right).$$

True posterior

$$z|x \sim t_m(\tilde{\mu}(x), \tilde{\Sigma}(x), \nu + n)$$

$$\tilde{\mu}(x) = W^\top (WW^\top + \sigma^2 I)^{-1}(x - b)$$

$$\tilde{\Sigma}(x) = \frac{1 + \nu^{-1}(x-b)^\top (WW^\top + \sigma^2 I)^{-1}(x - b)}{1 + \nu^{-1} n} (I - W^\top (WW^\top + \sigma^2 I)^{-1} W).$$

- We are motivated to incorporate a t-distributed **encoder**

$$q_{\phi,\nu}(z|x) = t_m\left(z \,\middle|\, \mu_\phi(x), (1 + \nu^{-1} n)^{-1} \Sigma_\phi(x), \nu + n\right).$$

- **When** $\nu \to \infty$, $q_{\phi,\nu}(z|x) \to \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x))$.

The t³-VAE

- From the geometric relationship of γ-power divergence and power families, we are motivated to replace the KL objective in the joint minimization problem with γ-power divergence.

$$(p_{\theta^*,\nu}, q_{\phi^*,\nu}) = \underset{p\in\mathcal{P}_\nu, q\in\mathcal{Q}_\nu}{\operatorname{argmin}} \mathcal{D}_\gamma(q \parallel p)$$

where γ is coupled to ν as $\gamma = -\frac{2}{\nu+n+m}$.

- **γ- loss**

The γ-power divergence from $q_{\varphi,\nu} \in Q_\nu$ to $p_{\theta,\nu} \in P_\nu$ can be computed in closed-form after an approximation of order γ^2.

$$\mathcal{L}_\gamma(\theta,\phi) = \frac{1}{2}\mathbb{E}_{x\sim p_{\text{data}}}\left[\frac{1}{\sigma^2}\mathbb{E}_{z\sim q_{\phi,\nu}(\cdot|x)}\left\|x-\mu_\theta(z)\right\|^2 + \left\|\mu_\phi(x)\right\|^2 + \frac{\nu}{\nu+n-2}\operatorname{tr}\Sigma_\phi(x) - \frac{\nu C_1}{C_2}\left|\Sigma_\phi(x)\right|^{-\frac{\gamma}{2(1+\gamma)}}\right]$$

for constants $C_1 = \left(\frac{\nu+m+n-2}{\nu+n-2}\left(1+\frac{n}{\nu}\right)^{\frac{\gamma m}{2}} C_{\nu+n,m}^{\gamma}\right)^{\frac{1}{1+\gamma}}$ and $C_2 = \left(\frac{\nu+m+n-2}{\nu-2}\sigma^n C_{\nu,m+n}^{-1}\right)^{-\frac{\gamma}{1+\gamma}}$.

The t³-VAE

Alternative prior and balance weight

- Analogously to the ELBO, the γ-loss consists of an MSE reconstruction error and additional terms which act as a regularizer.
- In fact, the remaining terms are equivalent (up to constants) to the γ-power divergence from the posterior $q_{\varphi,\nu}(z|x)$ to **the alternative prior**:

$$p_\nu^\star(z) = t_m\left(z|0, \tau^2 I, \nu+n\right); \quad \tau^2 = \frac{1}{1+\nu^{-1}n}\left(\sigma^{-n} C_{\nu,n}(1+\frac{n}{\nu-2})^{-1}\right)^{\frac{2}{\nu+n-2}}$$

γ-loss can then be rewritten as

$$\mathcal{L}_\gamma(\theta,\phi) = \mathbb{E}_{x\sim p_{\text{data}}}\left[\frac{1}{2\sigma^2}\mathbb{E}_{z\sim q_{\phi,\nu}(\cdot|x)}\left\|x-\mu_\theta(z)\right\|^2 + \alpha\mathcal{D}_\gamma(q_{\phi,\nu}\,\|p_\nu^\star)\right] + \text{const.}$$

Hence, γ-loss can be interpreted similarly as a balance between reconstruction and regularization, and ν controls both the target scale τ^2 and the regularizer coefficient $\alpha = -\frac{\gamma\nu}{2C_2}$.

- As $\nu \to \infty$, t^3VAE converges to the Gaussian VAE. As $\nu \to 2$, in theory both $\tau, \alpha \to 0$ so that regularization vanishes and t^3VAE regresses to a raw autoencoder.

Equivalence to the Bayesian Hierarchical Model

- **Prior-decoder pair**

$$p_{Z,\nu}(z) = \int p_{\theta,\nu}(x,z)dx = t_m(z|0, I, \nu)$$

$$z \sim \int_0^\infty \mathcal{N}_m\left(z \middle| 0, \frac{1}{\nu^{-1}\lambda} I\right) \chi^2(\lambda|\nu)d\lambda \propto \left(1 + \frac{1}{\nu}\|z\|^2\right)^{-\frac{\nu+m}{2}}$$

$$p_{\theta,\nu}(x|z) = \frac{p_{\theta,\nu}(x,z)}{p_{Z,\nu}(z)} = t_n\left(x \mid \mu_\theta(z), \frac{1+\nu^{-1}\|z\|^2}{1+\nu^{-1}m}\sigma^2 I, \nu + m\right)$$

$$x|z \sim \int_0^\infty \mathcal{N}_n\left(x \middle| \mu_\theta(z), \frac{1}{\nu^{-1}\lambda}\sigma^2 I\right) \mathcal{N}_m\left(z \middle| 0, \frac{1}{\nu^{-1}\lambda} I\right) \chi^2(\lambda|\nu)d\lambda$$

$$\propto t_n\left(x \middle| \mu_\theta(z), \frac{1+\nu^{-1}\|z\|^2}{1+\nu^{-1}m}\sigma^2 I, \nu + m\right)$$

- It is then straightforward to add any number of latent layers $z_i|z_{<i}, \lambda \sim \mathcal{N}_{m_i}(z_i|\mu_\theta(z_{<i}), \upsilon\lambda^{-1}\sigma_i^2 I)$ to obtain a heavy-tailed hierarchical prior $(z_1, \cdots, z_L)$.

Heavy-tailed distribution experiment

- The Fréchet inception distance (FID) score (Heusel et al.,2017) is employed to evaluate image quality.

- The images in Figure 3 display rare feature combinations.
- The top left image belongs to the intersection of the Male and Heavy Make-up classes, which constitute around 1% of all images.

Table 2: Reconstruction FID scores of CelebA and CIFAR100-LT. In CelebA, both overall scores and selected classes are shown. Bald, Mustache (Mst), and Gray hair (Gray) are rare classes (less than 5% of the total), while No beard (No Bd) is common (over 50%). In CIFAR100-LT, FID is measured varying imbalance factor ρ. Complete results of tuning each model are included in Appendix C.3.

(a) CelebA

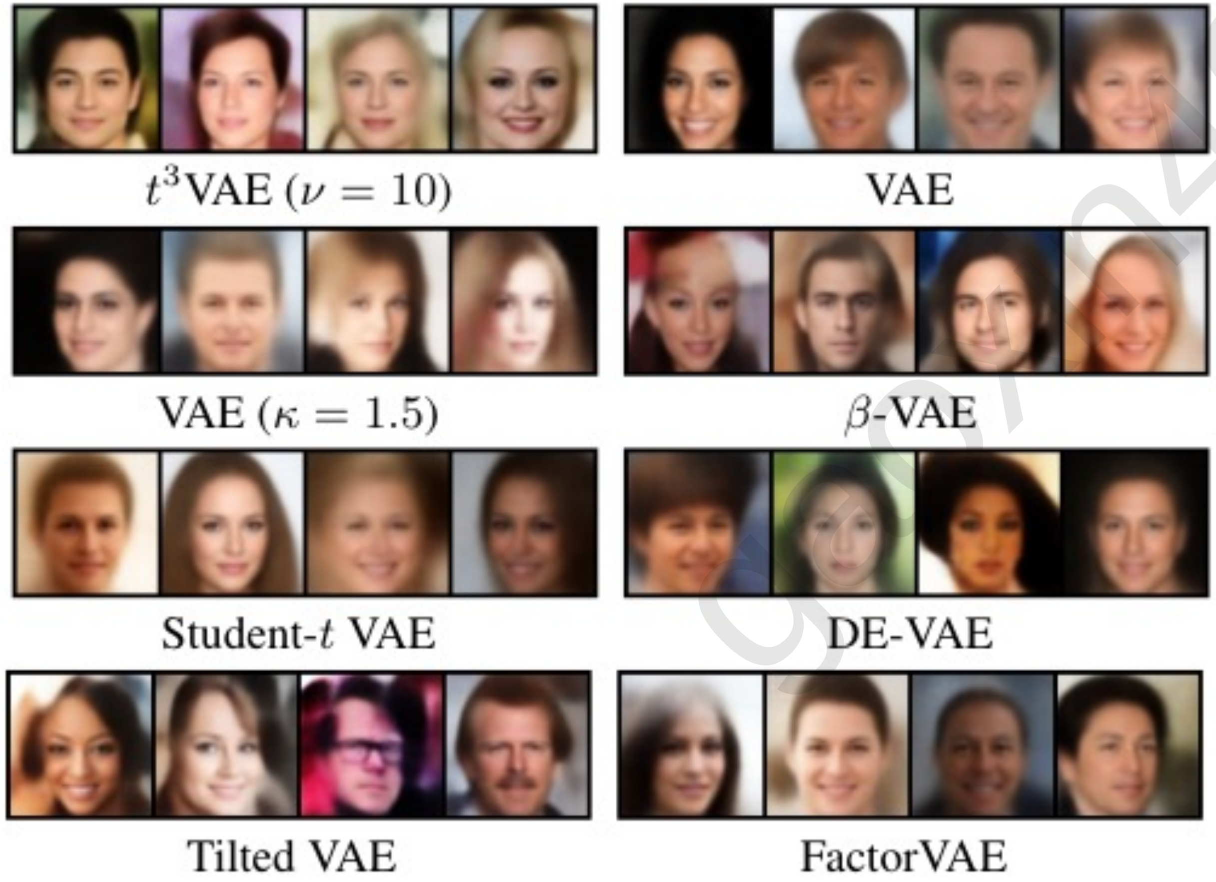
Framework	All	Bald	Mst	Gray	No Bd
t^3VAE ($\nu = 10$)	**39.4**	**66.5**	**61.5**	**67.2**	**40.1**
VAE	57.9	85.8	79.7	91.0	58.4
VAE ($\kappa = 1.5$)	73.2	105.3	96.4	114.5	73.8
β-VAE ($\beta = 0.05$)	40.4	69.3	62.7	71.1	40.9
Student-t VAE	78.4	112.0	104.2	118.7	78.6
DE-VAE ($\nu = 5$)	58.9	89.6	84.3	94.9	59.1
Tilted VAE ($\tau = 50$)	42.6	73.0	65.4	73.7	42.9
FactorVAE ($\gamma_{tc} = 5$)	59.8	91.7	85.7	95.2	60.8

(b) CIFAR100-LT

Framework	$\rho = 1$	10	50	100
t^3VAE ($\nu = 10$)	**97.5**	**102.8**	**108.3**	**128.7**
VAE	256.1	267.2	277.4	287.3
VAE ($\kappa = 1.5$)	274.2	290.5	296.7	297.7
β-VAE ($\beta = 0.1$)	114.1	130.4	138.5	160.6
Student-t VAE	259.5	314.1	323.7	333.4
DE-VAE ($\nu = 2.5$)	219.4	250.2	256.7	258.5
Tilted VAE ($\tau = 50$)	101.0	126.1	147.0	193.2
FactorVAE ($\gamma_{tc} = 5$)	232.3	272.5	275.6	270.1

Heavy-tailed distribution experiment

- Sample from the alternative t-prior $p_\nu^\star(z)$; more vivid images.



Framework	FID
t^3VAE ($\nu = 10$)	**50.6**
VAE	64.7
VAE ($\kappa = 1.5$)	79.6
β-VAE ($\beta = 0.05$)	51.8
Student-t VAE	82.3
DE-VAE ($\nu = 2.5$)	58.9
Tilted VAE ($\tau = 30$)	59.2
FactorVAE ($\gamma_{tc} = 2.5$)	67.0

Table 3: Generation FID scores for CelebA.

◀ Figure 4: Generated CelebA example images.

Heavy-tailed distribution experiment

Higher clarity and sharper detail

Original $\qquad$ t^3HVAE $\qquad$ HVAE

Figure 5: Original and reconstructed images by t^3HVAE ($\nu = 10$) and HVAE.

References

[1] Han T, Zhang J, Wu Y N. From *em*-projections to variational auto-encoder[C]//NeurIPS 2020 Workshop: Deep Learning through Information Geometry. 2020.

[2] Amari S. Information geometry and its applications[M]. Springer, 2016.

[3] Kotz S, Nadarajah S. Multivariate t-distributions and their applications[M]. Cambridge university press, 2004.

[4] Dickey J M. Expansions of t Densities and Related Complete Integrals[J]. The Annals of Mathematical Statistics, 1967, 38(2): 503-510.

[5] Amari S. Information geometry and its applications: Convex function and dually flat manifold[C]//LIX Fall Colloquium on Emerging Trends in Visual Computing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008: 75-102.

- Other resources you can refer to:

Frank Nielsen | Information Geometry, divergences, and diversities | Geometric Science of Information

Manifolds: A Gentle Introduction