Complex variational autoencoders admit Kähler structure

Andrew Gracyk*

Abstract

It has been discovered that latent-Euclidean variational autoencoders (VAEs) admit, in various capacities, Riemannian structure. We adapt these arguments but for complex VAEs with a complex latent stage. We show that complex VAEs reveal to some level Kähler geometric structure. We derive the Fisher information metric in the complex case under a latent complex Gaussian regularization with trivial relation matrix. It is well known from statistical information theory that the Fisher information coincides with the Hessian of the Kullback-Leibler (KL) divergence. Thus, the metric Kähler potential relation is exactly achieved under relative entropy. We propose a Kähler potential derivative of complex Gaussian mixtures that has rough asymptotic equivalence to the Fisher information metric while still being faithful to the underlying Kähler geometry, whereas simply overriding nontrivial terms in the Fisher metric can depart from Kähler geometric structure or require a nearest neighbor algorithm. Through our potential, valid as a plurisubharmonic (PSH) function, we develop a sampling procedure for the generative stage that, not only improves quality due to incorporating the geometric structure into the sampling, but is more efficient in metric construction by placing the computational burden of automatic differentiation instead on an expectation and covariance.

Key words. Variational autoencoder, VAE, geometric VAE, complex geometry, Kähler geometry, Kähler potential, statistics on manifolds, Fisher information, Fisher metric, complex Gaussian

Contents

1 Introduction

We study the auxiliary qualities of the variational autoencoder (VAE) via the complex geometric perspective. The Euclidean VAE is a foundational machine learning archetype, and from it has stemmed the complex analog Xie et al. (2023), although this technology has less establishment in literature. We

*Purdue University; agracyk@purdue.edu

apply the overarching arguments presented in for imaginary latent data. In particular, Chadebec and Allassonnière (2022) presents justification that VAEs admit Riemannian structure in the Euclidean-immersed case. We reframe this argument in a new light, and so we argue imaginary latent data holds the same features but through Kähler geometry.

The overarching effects of geometric features for latent variational tasks Kingma and Welling (2022) has been studied in tasks for external imposition, such as in Lopez and Atzberger (2025) and Palma et al. (2025). In this work, we study the effects of geometric feature for those that form independent of outside force, and rather those that arise natural in training, consistent with the work of Chadebec and Allassonnière (2022) but in the context of complex geometry. We refer to Lobashev et al. (2025) Zeng et al. (2025) Lee et al. (2025) Chadebec et al. (2020) Kouzelis et al. (2025) for other developments regarding latent geometry, primarily for variational inference, thus geometric quality in inference and generative tasks is a widely studied problem. In particular, latent geometry certainly develops, but the natural geometry is only part of the picture, i.e. there exist geometries outside of those that can form in the model which can be enforced independently Duque et al. (2023).

2 Notation

We will use notation $\dagger$ to denote a Hermitian transpose $X^\dagger_{ij} = \overline{X}_{ji}$, $\overline{z}$ to denote the complex conjugate of z. h will denote some notion, since we will subsequently aside indices and variants, of a Hermitian metric. z will denote a point in the latent space. $\partial\overline{\partial}$ denotes the complex Hessian with Dolbeault operators. We will use Ψ to denote the (positive) quadratic form in the exponent of a Gaussian density. In general, our notation is mostly standard. For the remainder of this work, we will always assume $\mu(z)$ is holomorphic and that $\Sigma(z)$ is not.

3 The complex variational autoencoder (VAE)

The complex VAE archetype has the same fundamental underpinning as the typical VAE. Let

$$x \in \mathbb{C}^n = \{a + bi : a, b \in \mathbb{R}, i = \sqrt{-1}\} \cong \mathbb{R}^2, \tag{1}$$

$$\mathbb{E}_{q_\phi}[z] \in \tilde{\mathbb{C}}^d := \mathbb{C}^d \cap \Big\{x \in P_c : \mu^*(\{z \in \overline{P_c} : \mathrm{Im}(z) = 0 \text{ almost surely}\} = 0)\Big\}. \tag{2}$$

Here $P_c = \{z_1, \dots, z_q\}$ is a point cloud where $z_i \in \mathbb{C}^d$, and $\overline{P}_c$ is a connected extension of this point cloud, and we will assume it exists. μ is Lebesgue measure for a complex manifold space. We highlight the above because we will often take $b = 0$ regarding x, i.e. real-valued data, but the manifold is only trivial on a set of outer measure zero with probability almost surely 1 up to the randomness in training, i.e.

$$\mathbb{P}\Big(\mu\Big(\{z \in \overline{P_c} : \mathrm{Im}(z) = 0\}\Big) = 0\Big) = 1. \tag{3}$$

We remark $\{z \in \overline{P_c} : \mathrm{Im}(z) = 0\}$ need not be a μ-measurable set. We also highlight the point cloud P_c is among the means of the VAE, as the VAE latent stage will have Gaussian noise as well.

As with the typical VAE, we follow suit of Xie et al. (2023), the objective is

$$\log p(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x) \parallel p(z)) \tag{4}$$

$$\propto ||x - g_\theta(z)||_2^2 - \mathrm{KL}(q_\phi(z|x) \parallel p(z)) := \text{VAE loss}, \tag{5}$$

but now our variables are complex-valued. Our neural networks will be such that

$$f_\phi : \mathbb{C}^n \to (\mu, \sigma, \delta) \in \mathbb{C}^d \times \mathbb{R}^d \times \mathbb{R}^d, \qquad g_\theta : \mathbb{C}^d \to \mathbb{C}^n. \tag{6}$$

Continuing to follow Xie et al. (2023), we take

$$\tilde{z}_i = z_i + \psi_{\mathrm{Re}} \odot \epsilon_{\mathrm{Re}} + \psi_{\mathrm{Im}} \odot \epsilon_{\mathrm{Im}} \tag{7}$$

$$\psi_{\mathrm{Im}} = \frac{\sigma + \delta}{2\sigma + 2\mathrm{Re}(\delta)}, \psi_{\mathrm{Re}} = i\frac{\sqrt{\sigma^2 - |\delta|^2}}{2\sigma + 2\mathrm{Re}(\delta)} \tag{8}$$

Note that we have the closed form

$$\mathrm{KL}(q_\phi(z|x) \parallel p(z)) = \mu^\dagger \mu + ||\sigma - 1 - \frac{1}{2}\log(\sigma^2 - |\delta|^2)||_1. \tag{9}$$

4 Euclidean VAEs unveil Riemannian structure

The standard VAE is taken such that

$$q_\phi(z|x) \sim \mathcal{N}(\mu(x_i), \Sigma(x_i)). \tag{10}$$

The Riemannian metric

$$g_z : T_zM \times T_zM \to \mathbb{R}, g(Y, Z) = g_{ij}Y^iZ^j, Y = Y^i\frac{\partial}{\partial y^i}, Z = Z^j\frac{\partial}{\partial z^j}, \tag{11}$$

or equivalently

$$g = g_{ij}dz^i \otimes dz^j, \tag{12}$$

can be viewed such that

$$g_{\mathbb{E}[f_\phi(x_i)]} = \Sigma^{-1}(x_i). \tag{13}$$

We will operate so that Σ is positive definite, thus Σ^{-1} and g are too. The Riemannian metric of the encoder can be extended under Taylor expansions

$$G(z) \approx \Sigma^{-1}(x_i) + \sum_{j\neq i} \Sigma^{-1}(x_j) \cdot \Omega_j(\mu(x_i)) + \Sigma^{-1}(x_i) \cdot J_{\Omega_i}(\mu(x_i)) \cdot (z - \mu(x_i)) \tag{14}$$

$$\Omega_i(z) = \exp\left\{ -\frac{(z - \mu(x_i))^T\Sigma^{-1}(x_i)(z - \mu(x_i))}{\text{constant}^2} \right\}, . \tag{15}$$

The Riemannian metric of the decoder differs to the sampling of the posterior. It is known $G(z) = J_g^T(z)J_g(z)$ on the latent spaces of generative models via the decoder. This gives rise to

$$\tilde{\Sigma}^{-1}(x) = J_g(\mu(x))^T J_g(\mu(x)) + I \tag{16}$$

for the decoder geometry. This work proposes sampling according to

$$U = \sqrt{\det(G(z))} \bigg/ \int_{\mathbb{R}^d} \sqrt{\det(G(z))}. \tag{17}$$

Note this sampling is done with the encoder geometry using $G(z)$.

An important caveat with this work is here the Riemannian metric has the same intrinsic dimension $d \times d$ as the latent space d, thus G does not characterize low-dimensional geometry, but rather geometry of the same space in which the manifold exists. In particular, a manifold exists in the latent space, but it is not low-dimensional, but rather describes the latent space curvature.

It is also known the Fisher information metric is a Riemannian metric on a smooth statistical manifold, which describes the decoder metric. Previously, we discussed the encoder metric, which is sampleable. In particular, for decoder geometry

$$g = \mathbb{E}[(\nabla_z p_z(x))(\nabla_z p_z(x))^T]. \tag{18}$$

For Gaussian likelihood, the log-likelihood, assuming constant variance, is

$$\log p_\theta(x|z) = -\frac{1}{2}(x - \mu(z))^T\Sigma^{-1}(x - \mu(z)) + \text{constants}. \tag{19}$$

Notice

$$\partial_{z_i} \log p_\theta(x|z) = (\Sigma^{-1}(x - \mu(z))^T \partial_i\mu(z). \tag{20}$$

Thus the Fisher information metric is

$$g_{ij} = \mathbb{E}[(\partial_{z_i} \log p(x|z)(\partial_{z_j} \log p(x|z))] = (\partial_{z_i}\mu(z))^T\Sigma^{-1}(\partial_{z_j}\mu(z)). \tag{21}$$

The partial derivatives run over the elements of z. Equivalently,

$$g(z) = J_\mu^T(z)\Sigma^{-1}J_\mu(z), \tag{22}$$

5 Our work: complex decoder geometry

Let $K : M \to \mathbb{R}$ be a Kähler potential, where M is a complex manifold. Let h be a Hermitian metric h such that

$$h_{\alpha\overline{\beta}}(z) = \partial_\alpha \partial_{\overline{\beta}} K(z, \overline{z}), \qquad \text{with associated Kähler form} \qquad \omega = i\partial\overline{\partial}K, \tag{23}$$

where $(w, \overline{w})$ is within a local coordinate chart and

$$K(p) = (z^1(p), \dots, z^m(p), \overline{z}^1(p), \dots, \overline{z}^m(p)). \tag{24}$$

The Kähler form is such that

$$\omega = ih_{\alpha\overline{\beta}} dz^\alpha \wedge d\overline{z}^\beta \tag{25}$$

and the Hermitian metric decomposes into real and imaginary parts Rowland (2025)

$$h = g + i\omega \tag{26}$$

is the Hermitian metric for Riemannian metric g.

5.1 The Fisher information metric in the complex case

In this section, we will derive the Fisher information metric that appears in the VAE but with a complex Gaussian distribution assumption. We will use the following definition in our proof: the probability density of a complex normal random variable is, for $P = \overline{\Sigma} - C^\dagger \Sigma^{-1} C$,

$$f(z) = \frac{1}{\pi^n \sqrt{\det(\Sigma)\det(P)}} \exp\left\{ -\frac{1}{2} \begin{bmatrix} z-\mu \\ \overline{z}-\overline{\mu} \end{bmatrix}^\dagger \begin{bmatrix} \Sigma & C \\ \overline{C} & \overline{\Sigma} \end{bmatrix}^{-1} \begin{bmatrix} z-\mu \\ \overline{z}-\overline{\mu} \end{bmatrix} \right\}. \tag{27}$$

We will take $C = 0$, which is standard in a VAE, thus the factor of $1/2$ in the exponent is removed. We will always take Σ to be Hermitian and diagonal, i.e.

$$\Sigma^\dagger = \Sigma = \overline{\Sigma}. \tag{28}$$

The Hermitian quality is a standard fact for complex Gaussians Hankin (2015), and the complex conjugate quality follows since Σ is (typically) diagonal in a VAE.

Theorem 1. (a known result; mentioned in Collier (2005); the proof is ours) Suppose $x|z \sim \mathcal{CN}(\mu(z), \Sigma(z))$ (in particular, the relation matrix is $C = 0$) and that μ is holomorphic. Then the Fisher information Hermitian metric on the decoded latent space is

$$h_{\alpha\overline{\beta}}(z) = (\partial_\alpha \mu)^\dagger \Sigma^{-1} (\partial_\beta \mu) + \operatorname{Tr}(\Sigma^{-1}(\partial_\alpha \Sigma)\Sigma^{-1}(\partial_\beta \Sigma)). \tag{29}$$

Remark. The Fisher information metric for a Gaussian $X \sim \mathcal{N}(x;, \mu(x), \Sigma(x))$ is well known, but not studied as often in the complex case. The primary difference in the complex case is the fact of $1/2$ appearing in the real-valued case, whereas here it does not due to the complex multivariate Gaussian density definition. It is possible to simplify the first term by taking

$$(\partial_\alpha \mu)^\dagger \Sigma^{-1} (\partial_\beta \mu) = \partial_\alpha \partial_{\overline{\beta}} (\mu^\dagger \Sigma^{-1} \mu). \tag{30}$$

We have used μ is holomorpic. There is no immediate way to simplify $\partial_\gamma \Sigma$.

Proof. The log-likelihood with trivial relation matrix is

$$\log p(x|z) = -(x-\mu)^\dagger \Sigma^{-1} (x-\mu) - \log \det \Sigma. \tag{31}$$

Let ∂_γ denote the element-wise derivative, i.e.

$$\partial_\gamma \mu = \frac{\partial \mu}{\partial z^\gamma} = (\frac{\partial \mu_1}{\partial z^\gamma}, \dots, \frac{\partial \mu_m}{\partial z^\gamma}). \tag{32}$$

Note the following derivative identities:

$$\partial_\alpha \Sigma^{-1} = -\Sigma^{-1}(\partial_\alpha \Sigma)\Sigma^{-1} \tag{33}$$

$$\partial_\alpha \log \det \Sigma = \mathrm{Tr}(\Sigma^{-1}\partial_\alpha \Sigma). \tag{34}$$

Now,

$$\partial_\alpha \log p = (\partial_\alpha \mu)^\dagger \Sigma^{-1}(x-\mu) + (x-\mu)^\dagger \underbrace{\Sigma^{-1}(\partial_\alpha \Sigma)\Sigma^{-1}}_{=A}(x-\mu) - \underbrace{\mathrm{Tr}(\Sigma^{-1}\partial_\alpha \Sigma)}_{=t_\alpha} \tag{35}$$

$$\partial_{\overline{\beta}} \log p = (x-\mu)^\dagger \Sigma^{-1}(\partial_\beta \mu) + (x-\mu)^\dagger \underbrace{\Sigma^{-1}(\partial_{\overline{\beta}} \Sigma)\Sigma^{-1}}_{=B}(x-\mu) - \underbrace{\mathrm{Tr}(\Sigma^{-1}\partial_{\overline{\beta}} \Sigma)}_{=t_{\overline{\beta}}}. \tag{36}$$

Now, note that $\mathbb{E}[x-\mu] = 0, \mathbb{E}[(x-\mu)(x-\mu)^\dagger] = \Sigma(z)$. Now, it is known from the Fisher information

$$h_{\alpha\overline{\beta}} = \mathbb{E}[\partial_\alpha \log p \partial_{\overline{\beta}} \log p] \tag{37}$$

$$= \mathbb{E}\Bigg[(\partial_\alpha \mu)^\dagger \Sigma^{-1}(x-\mu)(x-\mu)^\dagger \Sigma^{-1}(\partial_\beta \mu) + (\partial_\alpha \mu)^\dagger \Sigma^{-1}(x-\mu)(x-\mu)^\dagger B(x-\mu) - (\partial_\alpha \mu)^\dagger \Sigma^{-1}(x-\mu)t_{\overline{\beta}} \tag{38}$$

$$+ (x-\mu)^\dagger A(x-\mu)(x-\mu)^\dagger \Sigma^{-1}(\partial_\beta \mu) + (x-\mu)^\dagger A(x-\mu)(x-\mu)^\dagger B(x-\mu) - (x-\mu)^\dagger A(x-\mu)t_{\overline{\beta}} \tag{39}$$

$$- t_\alpha (x-\mu)^\dagger \Sigma^{-1}(\partial_\beta \mu) - t_\alpha (x-\mu)^\dagger B(x-\mu) + t_\alpha t_{\overline{\beta}}\Bigg] \tag{40}$$

$$= \mathbb{E}\Bigg[(\partial_\alpha \mu)^\dagger \Sigma^{-1}(\partial_\beta \mu) + 0 + 0 + 0 + (x-\mu)^\dagger A(x-\mu)(x-\mu)^\dagger B(x-\mu) - (x-\mu)^\dagger A(x-\mu)t_{\overline{\beta}} \tag{41}$$

$$- t_\alpha (x-\mu)^\dagger B(x-\mu) + t_\alpha t_{\overline{\beta}}\Bigg] \tag{42}$$

$$= (\partial_\alpha \mu)^\dagger \Sigma^{-1}(\partial_\beta \mu) + \mathbb{E}\Bigg[(x-\mu)^\dagger A(x-\mu)(x-\mu)^\dagger B(x-\mu)\Bigg] - t_\alpha t_{\overline{\beta}} - t_\alpha t_{\overline{\beta}} + t_\alpha t_{\overline{\beta}} \tag{43}$$

$$\stackrel{*}{=} (\partial_\alpha \mu)^\dagger \Sigma^{-1}(\partial_\beta \mu) + t_\alpha t_{\overline{\beta}} + \mathrm{Tr}\Bigg(\Sigma^{-1}(\partial_\alpha \Sigma)\Sigma^{-1}(\partial_{\overline{\beta}} \Sigma)\Bigg) - t_\alpha t_{\overline{\beta}} - t_\alpha t_{\overline{\beta}} + t_\alpha t_{\overline{\beta}} \tag{44}$$

$$= (\partial_\alpha \mu)^\dagger \Sigma^{-1}(\partial_\beta \mu) + \mathrm{Tr}\Bigg(\Sigma^{-1}(\partial_\alpha \Sigma)\Sigma^{-1}(\partial_{\overline{\beta}} \Sigma)\Bigg). \tag{45}$$

For $(*)$ we have used the fact that

$$\mathbb{E}[(y^\dagger A y)(y^\dagger B y)] = \mathrm{Tr}(A\Sigma)\mathrm{Tr}(B\Sigma) + \mathrm{Tr}(A\Sigma B\Sigma). \tag{46}$$

By using trace properties under shifts, this completes the proof.
□

In practice, the metric of equation 29 is nontrivial to sample from since it requires differentiation of Σ. We will attempt to construct a metric that has greater ability to be sampled in the following section.

5.2 The Fisher information metric and the Hessian of the KL divergence as a Kähler potential

It is well known it statistical theory that the Fisher information is the Hessian of the Kullback-Leibler (KL) divergence. Let the Fisher information with respect to the paramater be defined as

$$I = \mathbb{E}\Big[(\nabla_\theta p_\theta(x))(\nabla_\theta p_\theta(x))^T\Big]. \tag{47}$$

As Riemann integral and Lebesgue integrals respectively, the KL divergence is

$$\mathrm{KL}(p_\theta(x) \parallel p_{\theta'}(x)) = \int p_\theta(x) \log \frac{p_\theta(x)}{p_{\theta'}(x)} dx = \int \log \frac{dP_\theta}{dP_{\theta'}} dP_\theta. \tag{48}$$

In particular, it is known Zuo (2017)

$$I = \mathrm{Hess}_{\theta'} \mathrm{KL}(p_\theta(x) \parallel p_{\theta'}(x)) \bigg|_{\theta'=\theta}. \tag{49}$$

However, a Kähler potential necessitates

$$h_{\alpha\overline{\beta}} = \partial_\alpha \partial_{\overline{\beta}} K(z, \overline{z}). \tag{50}$$

i.e. the partial derivative is with respect to z coordinate, whereas the Fisher information derivatives are taken with respect to the neural network parameters in the above. Thus, if we take z as the parameter in the above, we have

$$h_{\alpha\overline{\beta}} = \mathrm{Hess}_{z',\alpha,\overline{\beta}} \mathrm{KL}(p(x|z) \parallel p(x|z')\bigg|_{z'=z} = \partial_\alpha \partial_{\overline{\beta}} \mathrm{KL}(p(x|z) \parallel p(x|z')\bigg|_{z'=z} = \partial_\alpha \partial_{\overline{\beta}} K(z, \overline{z}), \tag{51}$$

and so we take our Kähler potential as

$$\boxed{K(z, \overline{z}) = \mathrm{KL}(p(x|z) \parallel p(x|z')} \tag{52}$$

subsequently evaluated at $z' = z$ once the Hessian is taken.

Moreover, $p(x|z) \sim \mathcal{N}_{\mathbb{C}}(x; \mu(z), \Sigma(z))$ is complex Gaussian, and the KL divergence between complex Gaussians has a closed form. It is known

$$\mathrm{KL}\big(p(x|z) \parallel p(x|z')\big) = \frac{1}{2}\Big[\mathrm{tr}\big(\Sigma(z')^{-1}\Sigma(z)\big) + (\mu(z') - \mu(z))^H \Sigma(z')^{-1}(\mu(z') - \mu(z)) - n + \ln \frac{|\Sigma(z')|}{|\Sigma(z)|}\Big]. \tag{53}$$

When the covariance is diagonal, we get

$$\mathrm{KL}\big(p(x|z) \parallel p(x|z')\big) = \frac{1}{2} \sum_{k=1}^{n} \left[\frac{\sigma_k^2(z)}{\sigma_k^2(z')} + \frac{|\mu_k(z) - \mu_k(z')|^2}{\sigma_k^2(z')} - 1 + \ln \frac{\sigma_k^2(z')}{\sigma_k^2(z)} \right]. \tag{54}$$

It is known

$$\partial\overline{\partial}\ \mathrm{KL}\big(p(x|z) \parallel p(x|z')\big)\bigg|_{z'=z} = (\partial_\alpha \mu)^\dagger \Sigma^{-1} (\partial_\beta \mu) + \mathrm{Tr}(\Sigma^{-1}(\partial_\alpha \Sigma)\Sigma^{-1}(\partial_\beta \Sigma)). \tag{55}$$

This is exactly what our previous derivation achieved,

If we have N latent samples, K is ambient dimension, and d latent dimension, the above sampling procedure requires $4NKd$ derivative computations when Σ is diagonal.

5.3 A log-likelihood sum potential induces a metric asymptotically close to the Fisher information metric

The primary goal of this section is that we will show there exists a Kähler potential that induces a Hermitian metric similarly to the true Fisher information metric. In particular, our method is more justified theoretically rather than simply "replacing" terms in 29, since if we swap terms for things more computable, it will not necessarily hold that Kähler geometry exists. Thus, this section will attempt to justify we can approximate the Hermitian metric of 29 with something that is valid as a metric that is simultaneously more computable under assumptions.

We provide some motivation for this section. We begin by sampling a collection in the latent space. We refer to this as the previous samples. In this section, we anticipate a sampled point z is sufficiently "close" to a previously sampled mean, i.e. the previous sample is sufficiently dense. Under this assumption, we operate so that covariance is estimated sufficiently well under this density. Moreover, the metric we provide will be an expected value analog in the previous section, more or less. We will argue, under this density of points and due to exponential convergence properties, the weights of our expected value will be arranged in such a way that the weight of the "close" points to the newly sampled point is significantly larger than those of previously sampled points far away. Thus our expectation will coincide with the terms in the previous section. Moreover, with all of this in mind, we still maintain validity of the metric. Thus, we have sound justification for approximately the metric of the previous section while maintaining efficiency.

Let z be a latent point, x the decoder, and μ be the map of a mean of a Gaussian of the map in the latent space. Let us take Gaussian weights

$$\Omega_i(z) = \exp\left\{\frac{1}{\rho^2}(x(z)-\mu_i)^\dagger \Sigma_i^{-1}(x(z)-\mu_i)\right\}, \qquad \Psi_i = (x(z)-\mu_i)^\dagger \Sigma_i^{-1}(x(z)-\mu_i), \tag{56}$$

with Kähler potential, which is a log-likelihood of Gaussian components,

$$K(z,\overline{z}) = \rho^2 \log\Big(\sum_i a_i \Omega_i(z)\Big), \qquad \sum_i a_i = 1, a_i > 0, \tag{57}$$

where we will often take $a_i = \frac{1}{N}$. This Kähler potential is distinct from statistical literature. Taking a logarithm of a mixture sum is not frequent in literature. From a statistical perspective, it poses computational issues Rudin (2015). For instance, the sum is in the interior of the logarithm and makes statistical computational difficult, if even relevant. In literature, a similar Kähler potential is studied in Xie et al. (2023) from a geometry perspective, but without the Gaussian component.

Our Kähler metric is given by

$$h_{\alpha\overline{\beta}}(z) = \partial_\alpha \partial_{\overline{\beta}} K(z,\overline{z}). \tag{58}$$

Let us derive this. Define $v_i := x - \mu_i$. First, notice, in Einstein notation and since we are working with holomorphic functions,

$$\Psi_i = (\Sigma_i^{-1})_{\alpha\overline{\beta}} v_i^\alpha \overline{v}_i^\beta \tag{59}$$

$$\partial_\alpha \Psi_i = (\partial_\alpha v_i^\delta)(\Sigma_i^{-1})_{\delta\overline{\beta}}\overline{v}_i^\beta, \qquad \partial_{\overline{\beta}}\Psi_i = (\partial_{\overline{\beta}} v_i^\delta)(\Sigma_i^{-1})_{\alpha\delta} v_i^\alpha \tag{60}$$

$$\partial_\alpha\partial_{\overline{\beta}}\Psi_i = (\partial_\alpha v_i^\delta)(\Sigma_i^{-1})_{\delta\overline{\gamma}}(\partial_{\overline{\beta}}\overline{v}_i^\gamma) = (\partial_\alpha v_i)^\dagger(\Sigma_i^{-1})(\partial_\beta \overline{v}_i). \tag{61}$$

Now, using the exponential derivative and the product rule,

$$\partial_{\overline{\beta}}\Omega_i = \partial_{\overline{\beta}}\Big(e^{\Psi_i/\rho^2}\Big) = \frac{1}{\rho^2}(\partial_{\overline{\beta}}\Psi_i)\Omega_i \tag{62}$$

$$\partial_\alpha\partial_{\overline{\beta}}\Omega_i = \frac{1}{\rho^2}\Big[\partial_\alpha\Omega_i\partial_{\overline{\beta}}\Psi_i + \Omega_i\partial_\alpha\partial_{\overline{\beta}}\Psi_i\Big] = \frac{\Omega_i}{\rho^2}\Big[\frac{1}{\rho^2}\partial_\alpha\Psi_i\partial_{\overline{\beta}}\Psi_i + \partial_\alpha\partial_{\overline{\beta}}\Psi_i\Big] \tag{63}$$

$$= \frac{\Omega_i}{\rho^2}\Big[\frac{1}{\rho^2}(\partial_\alpha v_i^\delta)(\Sigma_i^{-1})_{\delta\overline{\beta}}\overline{v}_i^\beta(\partial_{\overline{\beta}} v_i^\delta)(\Sigma_i^{-1})_{\alpha\delta}v_i^\alpha + (\partial_\alpha v_i^\delta)(\Sigma_i^{-1})_{\delta\overline{\gamma}}(\partial_\beta\overline{v}_i^\gamma)\Big] := \Xi_i. \tag{64}$$

Thus, we will endow z to be on a compact subset of $\mathbb{C}^d$, then by uniform convergence, we can exchange differentiation and the infinite sum

$$h_{\alpha\overline{\beta}} = \partial_\alpha\partial_{\overline{\beta}}\Big(\rho^2\log\Big(\sum_i a_i\Omega_i(z)\Big)\Big) \tag{65}$$

$$= \rho^2\partial_\alpha\Big(\frac{1}{\sum_i a_i\Omega_i(z)}\sum_i a_i\partial_{\overline{\beta}}\Omega_i(z)\Big) \tag{66}$$

$$= \rho^2\left(-\frac{\sum_i a_i\partial_\alpha\Omega_i(z)}{(\sum_i a_i\Omega_i(z))^2}\sum_i a_i\partial_{\overline{\beta}}\Omega_i(z) + \frac{1}{\sum_i a_i\Omega_i(z)}\sum_i a_i\partial_\alpha\partial_{\overline{\beta}}\Omega_i(z)\right) \tag{67}$$

$$= \rho^2\left(\frac{-1}{(\sum_i a_i\Omega_i(z))^2}(\sum_i a_i\frac{1}{\rho^2}\Omega_i\partial_{\overline{\beta}}\Psi_i)(\sum_i a_i\frac{1}{\rho^2}\Omega_i\partial_\alpha\Psi_i) + \frac{1}{\sum_i a_i\Omega_i(z)}\sum_i a_i\Xi_i\right). \tag{68}$$

This is the Kähler metric of the complex VAE. Let us justify this more. Note that the density of a discrete random variable is a simple function, thus we have a sequence of simple functions converging to an integrable function. Justified by the Law of Large Numbers (or equivalently, by a Monte Carlo argument; we will justify this more thoroughly after our analysis here)

$$\frac{\sum_i a_i \Omega_i(z) \partial_\alpha \Psi_i}{\sum_i a_i \Omega_i(z)} \xrightarrow{N\to\infty} \mathbb{E}_w[\partial_\alpha \Psi] \tag{69}$$

$$\frac{\sum_i a_i \Omega_i(z) \partial_{\overline{\beta}} \Psi_i}{\sum_i a_i \Omega_i(z)} \xrightarrow{N\to\infty} \mathbb{E}_w[\partial_{\overline{\beta}} \Psi] \tag{70}$$

$$\frac{1}{\sum_i a_i \Omega_i(z)} \sum_i a_i \Xi_i \xrightarrow{N\to\infty} \frac{1}{\rho^2} \mathbb{E}_w[\frac{1}{\rho^2} \partial_\alpha \Psi \partial_{\overline{\beta}} \Psi + \partial_\alpha \partial_{\overline{\beta}} \Psi]. \tag{71}$$

Thus,

$$h_{\alpha\overline{\beta}} = \rho^2 \left(\frac{-1}{\rho^4} \mathbb{E}_w[\partial_\alpha \Psi] \mathbb{E}_w[\partial_{\overline{\beta}} \Psi] + \frac{1}{\rho^2} \mathbb{E}_w[\frac{1}{\rho^2} \partial_\alpha \Psi \partial_{\overline{\beta}} \Psi - \partial_\alpha \partial_{\overline{\beta}} \Psi] \right) \tag{72}$$

$$= \mathbb{E}_w[\partial_\alpha \partial_{\overline{\beta}} \Psi] + \frac{1}{\rho^2} \Big(\mathbb{E}_w[\partial_\alpha \Psi \partial_{\overline{\beta}} \Psi] - \mathbb{E}_w[\partial_\alpha \Psi] \mathbb{E}_w[\partial_{\overline{\beta}} \Psi] \Big) \tag{73}$$

$$\boxed{h_{\alpha\overline{\beta}} = \mathbb{E}_w[\partial_\alpha \partial_{\overline{\beta}} \Psi] + \frac{1}{\rho^2} \mathrm{Cov}(\partial_\alpha \Psi, \partial_{\overline{\beta}} \Psi)} \tag{74}$$

$$= \sum_i w_i(z) \partial_\alpha \partial_{\overline{\beta}} \Psi_i(z) + \frac{1}{\rho^2} \left(\sum_i w_i (\partial_\alpha \Psi_i(z) \partial_{\overline{\beta}} \Psi_i(z)) - \Big[\sum_i w_i \partial_\alpha \Psi_i(z) \Big] \Big[\sum_i w_i \partial_{\overline{\beta}} \Psi_i(z) \Big] \right). \tag{75}$$

We have

$$w_i(z) = \frac{a_i e^{-\Psi_i(z)/\rho^2}}{\sum_j a_j e^{-\Psi_i(z)/\rho^2}}, \qquad \mathbb{E}_w[X] = \sum_i w_i(z) X(z). \tag{76}$$

It is known that Hess $= J^\dagger J$ is reasonable. Thus, let us take

$$h_{\alpha\overline{\beta}} = (1 + \frac{1}{\rho^2}) \mathbb{E}_w[\partial_\alpha \Psi \partial_{\overline{\beta}} \Psi] - \frac{1}{\rho^2} \mathbb{E}_w[\partial_\alpha \Psi] \mathbb{E}_w[\partial_{\overline{\beta}} \Psi] \tag{77}$$

Using this, as before, if we have N latent samples, I samples for the expectation, K is ambient dimension, and d latent dimension, the above sampling procedure requires $2NId$ derivative computations, which is more efficient depending on I and K. Our method is more efficient if $I < 2K$, which is reasonable since K is often quite large. For example, for MNIST data $K = 28 \times 28 = 784$, while we can easily take I to be between $100 - 200$.

Let us justify our use of the Law of Large Numbers more sufficiently. Consider a sample $\{z_i\}_{i=1}^\infty$ with weights $\mathbb{E}[w] < \infty, \mathbb{E}[w\Gamma] < \infty$. Taking the limit, we have

$$\lim_{N\to\infty} \frac{\sum_{i=1}^N w_i \Gamma(z_i)}{\sum_{i=1}^N w_i} \xrightarrow{\text{a.s.}} \mathbb{E}_\mu[\Gamma] = \int_\Omega \Gamma d\mu. \tag{78}$$

This is equivalent to a Monte Carlo argument because Monte Carlo integration uses the Law of Large Numbers.

We make some other remarks. Suppose we wanted to attempt to approximate the Fisher metric of equation 29 by using discrete substitutions μ_i, Σ_i^{-1}. This is nontrivial because attempting to find the closest point among a collection requires a nearest neighbor algorithm. This is especially nontrivial in large-scale settings, which is the focus of this work. This is already baked into the expectation calculation via the softmax function, so our proposed metric of 74, always a valid metric regardless of sampling, has no need for such a nearest-point search algorithm.

Our metric indeed discards the $\partial_\gamma \Sigma$ terms. It is of interest to attempt to replace these structures in the calculations with a fast approximation. Our proposed Kähler potential discards $\partial_\gamma \Sigma$ in the computation but not geometrically. As we show in Appendix A, these derivatives are absorbed into the expectation and covariance terms in the purely analytical case. Our proposed method is a discrete approximation to this. We threw away our $\partial_\gamma \Sigma$ terms with different sources of curvature, as now the covariance term acts as a substitute for $\partial_\gamma \Sigma$.

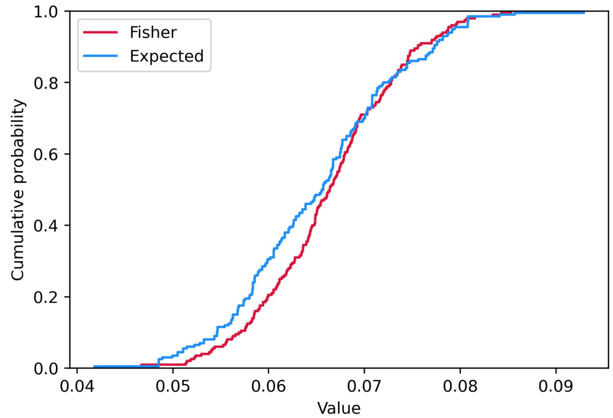

Figure 1: We show distributional equivalence by plotting the cumulative distribution of the $(1,1)$ real element of the first term Fisher information metric $(\partial_\alpha \mu)^\dagger \Sigma^{-1}(\partial_\beta \mu)$ versus its expectation analog $\mathbb{E}[\partial_\alpha \partial_{\overline{\beta}} \Psi]$ of our proposed metric derivative of the Kähler potential. We sampled 200 points z in the latent space, and we computed the expected value with 300 points. Here, we used the MNIST dataset to train the VAE.

Lemma 1. Suppose $x(z) \approx \mu_i$. Then

$$\mathbb{E}_w[\partial_\alpha \partial_{\overline{\beta}} \Psi] = \sum_i w_i \partial_\alpha \partial_{\overline{\beta}} \Psi_i \approx (\partial_\alpha \mu)^\dagger \Sigma^{-1}(\partial_\beta \mu). \tag{79}$$

Proof. The first equality follows from definition of the expectation. For a single Gaussian, notice $w_i \approx 1$, $w_{j\neq i} \approx 0$ for $x(z) \approx \mu_i$, and by definition

$$\partial_\alpha \partial_{\overline{\beta}} \Psi_i = (\partial_\alpha v_i)^\dagger (\Sigma_i^{-1})(\partial_{\overline{\beta}} v_i), \tag{80}$$

so the result follows. If Gaussians are sufficiently close to one another, take for suitable collection of indices I, $\sum_{i\in I} w_i \approx 1$, $w_{j\notin I} \approx 0$.
□

Lemma 2 (a well-known fact). $h_{\alpha\overline{\beta}}$ is valid as a Hermitian metric and g a Riemannian metric if the potential Φ is plurisubharmonic.

Proof. Since potential Φ is plurisubharmonic then

$$\partial\overline{\partial}\Phi \succeq 0 \implies \omega = \frac{i}{2}\partial\overline{\partial}\Phi \geq 0. \tag{81}$$

Thus, the bilinear form $g(u,v) = \omega(u, Jv)$ is positive semi-definite and a valid Riemannian metric. Moreover, since $h = \partial\overline{\partial}\Phi \succeq 0$, we have h is Hermitian, and so (g, J, h) is valid as a Kähler structure.
□

Theorem 2. The potential K we defined in 57 is plurisubharmonic.

Proof. It is known that that if $\Omega \subseteq \mathbb{C}^d$, and each u_i is plurisubharmonic on Ω, and Φ is real-valued, increasing, and convex, then $\Phi \circ (u_1, \ldots, u_K)$ is plurisubharmonic. This is found and proven as Proposition 2.1.1 in Çigdem Çelik (2015). In particular, u_i plurisubharmonic for all i implies is $(u_1, \ldots, u_K)$ is in PSH(Ω) in each component. It is known that $\log(\sum_i e^{x_i})$ is a convex, and increasing, function over its domain de Azevedo (2017). Thus, it suffices to show $u_i := \Psi_i/\rho^2 + \log a_i$ is plurisubharmonic. Now,

$$\partial\overline{\partial}\Psi_i = J^\dagger(z)\Sigma_i^{-1}J(z) = J^\dagger(z)\Gamma^\dagger\Gamma J(z) \succeq 0, \tag{82}$$

which gives the result, since therefore $\Psi_i/\rho^2 + \log a_i$ must be plurisubharmonic too. Here, we have decomposed $\Sigma^{-1} = \Gamma^\dagger\Gamma$, which is a square root decomposition and exists since Σ^{-1} is positive semi-definite.
□

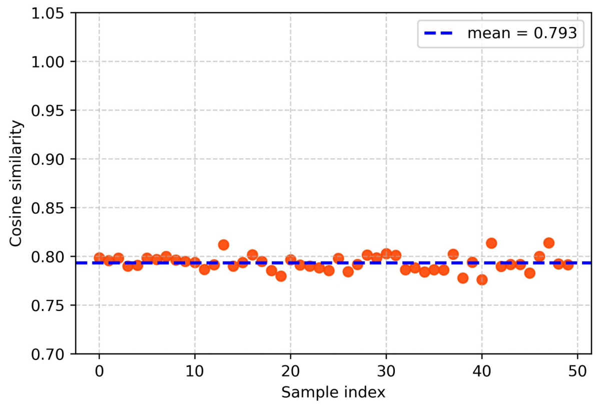

Figure 2: We illustrate the condition as in Theorem 3, $\partial v \approx M\Sigma^{-1}v$, is reasonable through a cosine similarity. In particular, a similarity of $+1 \in [-1, +1]$ is most desirable. Ours is 0.793, which means there is reasonable similarity.

Theorem 3. Let $\tilde{h}$ be the Fisher information metric as in Theorem 1, and let h be the Kähler-potential induced mertric as of equation 74. Moreover, suppose

$$\partial v \approx M\Sigma^{-1}v, \tag{83}$$

for some matrix M, where $v_i := x(z) - \mu_i$. In particular, we have checked this empirically, and this is reasonable. Then if $x(z) \approx \mu_i$ for some i, there exists a constant ρ

$$\tilde{h} \propto h. \tag{84}$$

Proof. By Lemma 1, it follows

$$\mathbb{E}_w[\partial_\alpha \partial_{\overline{\beta}} \Psi] \approx (\partial_\alpha \mu)^\dagger \Sigma^{-1} (\partial_\beta \mu). \tag{85}$$

Observe the other terms scale quadratically in Σ^{-1}, i.e.

$$\mathrm{Tr}\left(\Sigma^{-1}(\partial_\alpha \Sigma)\Sigma^{-1}(\partial_{\overline{\beta}}\Sigma)\right) \sim \mathcal{O}\Big(||\Sigma^{-1}||^2\Big), \qquad \mathrm{Cov}(\partial_\alpha \Psi, \partial_{\overline{\beta}} \Psi) \sim \mathcal{O}\Big(||\Sigma^{-1}||^2\Big). \tag{86}$$

In particular, the second line follows since

$$\begin{aligned}
\partial_\alpha \Psi_i &= (\partial_\alpha v)^\dagger \Sigma_i^{-1} v + v^\dagger \Sigma_i^{-1} (\partial_\alpha v) && (87)\\
&\approx (M\Sigma_i^{-1} v)^\dagger \Sigma_i^{-1} v + v^\dagger \Sigma_i^{-1} (M \Sigma_i^{-1} v) && (88)\\
&= v^\dagger (\Sigma_i^{-1} M^\dagger \Sigma_i^{-1} + \Sigma_i^{-1} M \Sigma_i^{-1}) v = v^\dagger P v. && (89)
\end{aligned}$$

Again, there exists a weight, or weights, sufficiently large so that one, or multiple, $i \in I$ dominate. It is the case that $\Sigma_{i\in I}^{-1} \approx \Sigma(z)$. Thus,

$$\mathrm{Cov}(\partial_\alpha \Psi, \partial_{\overline{\beta}} \Psi) \approx \mathrm{Cov}(v^\dagger P_\alpha v, v P_\beta v) = \mathrm{Tr}(P_\alpha \Sigma P_\beta \Sigma^{-1}) = \mathcal{O}(||\Sigma^{-1}||^2). \tag{90}$$

The real version of covariance formula above is found in Petersen and Pedersen (2012) on page 43. Here, we have noted $v \sim \mathcal{N}(0, \Sigma)$.
□

5.4 Sampling

We gave the general sampling procedure with respect to a metric in 17. Let us use notation

$$\text{matrix concatenation with respect to } ij := \bigoplus_{ij} \tag{91}$$

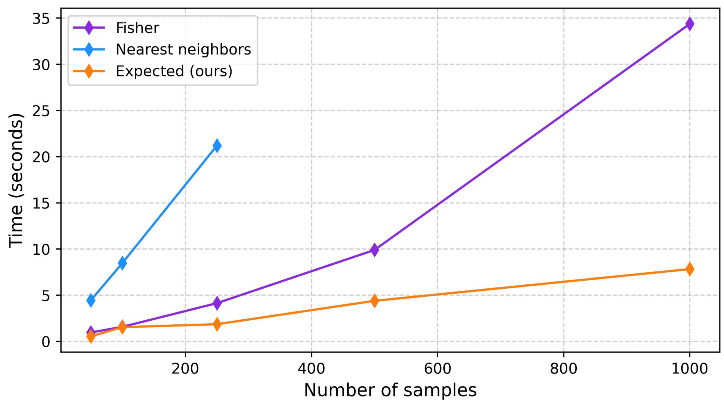

Figure 3: We illustrate runtime in constructing the Hermitian metrics via sampling. We choose $\alpha = \overline{\beta} = 1$. We truncate the number of data points for the nearest neighbors search case, since this gave CUDA out of memory errors in high cases.

It is a known fact that there is a volume element equivalence

$$dV_g = \sqrt{\det(g)}\lambda_{2d} = |\det(h)|\lambda_{2d}. \tag{92}$$

In particular, using the (1,1)-form, it is known Lemmon (2022) (see Zheng (2017), Mandolesi (2025) for other relevant literature)

$$\omega_p^n = d!\frac{i^d}{2^d}\zeta^1 \wedge \overline{\zeta}^1 \wedge \ldots \wedge \zeta^d \wedge \overline{\zeta}^d, \qquad \zeta^i \wedge \overline{\zeta}^i = -2i\chi^i \wedge v^i \tag{93}$$

we get $\omega_p^d = d!\mathrm{Vol}_g$, but since $\omega^n = d!\det(h)(i/2)^d dz^1 \wedge d\overline{z}^1 \wedge \ldots$, we have the result. We will also use the fact that complex integration has an equivalent form over $\mathbb{R}^d$

$$\int_{\mathbb{C}^d} \Gamma d\lambda = \int_{\mathbb{R}^{2d}} \Gamma(x+iy)dxdy. \tag{94}$$

The sampling procedure with respect to the Fisher metric of 22 is given by

$$\det \bigoplus_{\alpha\overline{\beta}} \partial_\alpha \partial_{\overline{\beta}} \mathrm{KL}(p(x|z) \parallel p(x|z'))\Big|_{z'=z} \Big/ \int_{\mathbb{C}^d} \det \bigoplus_{\alpha\overline{\beta}} \partial_\alpha \partial_{\overline{\beta}} \mathrm{KL}(p(x|z) \parallel p(x|z'))\Big|_{z'=z} d\lambda. \tag{95}$$

This is exactly equation 29 with our Fisher metric.

The sampling of our proposed metric is

$$\det \bigoplus_{\alpha\overline{\beta}} \mathbb{E}_w[\partial_\alpha \partial_{\overline{\beta}} \Psi] + \frac{1}{\rho^2}\mathrm{Cov}(\partial_\alpha \Psi, \partial_{\overline{\beta}} \Psi) \Bigg/ \int_{\mathbb{C}^d} \det \bigoplus_{\alpha\overline{\beta}} \Big[\mathbb{E}_w[\partial_\alpha \partial_{\overline{\beta}} \Psi] + \frac{1}{\rho^2}\mathrm{Cov}(\partial_\alpha \Psi, \partial_{\overline{\beta}} \Psi)\Big] d\lambda. \tag{96}$$

More specifically, we found the greatest empirical success in computing the above, and drawing samples from a multinomial distribution with the above pdf according to the above without replacement. Generally, the number of metric evaluations will exceed the number of points drawn.

6 Conclusions

We have developed two Kähler potentials whose complex Hessian $\partial\overline{\partial}$ is the decoded Fisher information metric, one exact and one approximate. Our approximate metric remains a valid Kähler potential K, and it is efficient because the computational burden is displaced from automatic differentiation to an expectation. This expectation is also more efficient than using a nearest neighbor algorithm. The efficiency matters because constructing a metric of the latent space and sampling from this metric is nontrivial. We have shown our method outperforms a baseline on the MNIST dataset. Our work is primarily interesting because ML is largely unexplored from the complex geometry perspective.

7 Acknowledgments

I gratefully acknowledge financial support from Purdue University Department of Mathematics.

References

Clément Chadebec and Stéphanie Allassonnière. A geometric perspective on variational autoencoders, 2022. URL https://arxiv.org/abs/2209.07370.

Clément Chadebec, Clément Mantoux, and Stéphanie Allassonnière. Geometry-aware hamiltonian variational auto-encoder, 2020. URL https://arxiv.org/abs/2010.11518.

S.L. Collier. Fisher information for a complex gaussian random variable: beamforming applications for wave propagation in a random medium. *IEEE Transactions on Signal Processing*, 53(11):4236–4248, 2005. doi: 10.1109/TSP.2005.857046.

Rodrigo de Azevedo. Why is log-of-sum-of-exponentials $f(x) = \log\left(\sum_{i=1}^{n} e^{x_i}\right)$ a convex function?, 2017. URL https://math.stackexchange.com/questions/2418554/why-is-log-of-sum-of-exponentials-fx-log-left-sum-i-1n-e-x-i-right. Mathematics StackExchange, asked 6 Sep 2017, modified 11 Mar 2021.

Garoe Dorta, Sara Vicente, Lourdes Agapito, Neill D. F. Campbell, and Ivor Simpson. Training vaes under structured residuals, 2018. URL https://arxiv.org/abs/1804.01050.

Andrés F. Duque, Sacha Morin, Guy Wolf, and Kevin R. Moon. Geometry regularized autoencoders. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(6):7381–7394, June 2023. URL https://doi.org/10.1109/TPAMI.2022.3222104.

Robin K. S. Hankin. The complex multivariate gaussian distribution. *The R Journal*, 7(1):73–80, 2015. ISSN 2073-4859. URL https://journal.r-project.org/articles/RJ-2015-006/RJ-2015-006.pdf.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Eq-vae: Equivariance regularized latent space for improved generative image modeling, 2025. URL https://arxiv.org/abs/2502.09509.

Wonjun Lee, Riley C. W. O'Neill, Dongmian Zou, Jeff Calder, and Gilad Lerman. Geometry-preserving encoder/decoder in latent generative models, 2025. URL https://arxiv.org/abs/2501.09876.

Johnny Lemmon. Volume form of a kähler manifold. Mathematics StackExchange question and answer, 2022. URL https://math.stackexchange.com/questions/4428928/volume-form-of-a-khler-manifold.

Alexander Lobashev, Dmitry Guskov, Maria Larchenko, and Mikhail Tamm. Hessian geometry of latent space in generative models, 2025. URL https://arxiv.org/abs/2506.10632.

Ryan Lopez and Paul J. Atzberger. Gd-vaes: Geometric dynamic variational autoencoders for learning nonlinear dynamics and dimension reductions, 2025. URL https://arxiv.org/abs/2206.05183.

André L. G. Mandolesi. Hermitian geometry of complex multivectors, determinants and orientations, 2025. URL https://arxiv.org/abs/2403.17022.

Alessandro Palma, Sergei Rybakov, Leon Hetzel, Stephan Günnemann, and Fabian J. Theis. Enforcing latent euclidean geometry in single-cell vaes for manifold interpolation, 2025. URL https://arxiv.org/abs/2507.11789.

Kaare Brandt Petersen and Michael Syskind Pedersen. The matrix cookbook. Technical Report Version: November 15, 2012, Technical University of Denmark, 2012. URL https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf.

Todd Rowland. Kähler form., 2025. URL https://mathworld.wolfram.com/KaehlerForm.html. From *MathWorld*—A Wolfram Resource.

Cynthia Rudin. Gaussian mixture models and expectation maximization. Course notes, Duke University, 2015. URL https://users.cs.duke.edu/~cynthia/CourseNotes/GMMEMNotes.pdf.

Yuying Xie, Thomas Arildsen, and Zheng-Hua Tan. Complex recurrent variational autoencoder with application to speech enhancement, 2023. URL https://arxiv.org/abs/2204.02195.

Wei Zeng, Xuebin Chang, Jianghao Su, Xiang Gu, Jian Sun, and Zongben Xu. Gmaplatent: Geometric mapping in latent space, 2025. URL https://arxiv.org/abs/2503.23407.

Tao Zheng. A parabolic monge–ampère type equation of gauduchon metrics. *International Mathematics Research Notices*, 2019(17):5497–5538, October 2017. ISSN 1687-0247. doi: 10.1093/imrn/rnx285. URL http://dx.doi.org/10.1093/imrn/rnx285.

Xingdong Zuo. Show that fisher information matrix is the second order gradient of kl divergence, 2017. URL https://math.stackexchange.com/questions/2239040/show-that-fisher-information-matrix-is-the-second-order-gradient-of-kl-divergenc. Mathematics Stack Exchange.

Çigdem Çelik. A survey of plurisubharmonic functions. Master's thesis, Sabancı University, January 2015. URL https://research.sabanciuniv.edu/id/eprint/32334/1/CigdemCelik_10064646.pdf. M.S. Thesis.

A Our Kähler potential and the Fisher metric analytic equivalence

In the primary section of our work earlier, we took

$$x \sim \mathcal{CN}(x; \mu_i, \Sigma_i). \tag{97}$$

In particular, our decoded output is closed to a true output μ_i that is previously sampled. Also, in that section, the weights our organized so that a particular, or multiple, i dominate, thus the log of the sum of the exponentials is heavily skewed in favor of certain i. In practice, the weights with the softmax mean that this work is analogous to a single exponential in the sum.

In this section, we will treat $\mu(z)$ as the sampled image, thus we will treat $\mu(z) \approx x$. Under sufficient closeness, it is true still true that $\mu(z) \approx \mu_i, \Sigma(z) \approx \Sigma_i$. With this framework, we can reconcile the differences between these two settings, one discrete and one continuous.

Let us take

$$p(x|z) = \mathcal{CN}(\mu(z), \Sigma(z)), \tag{98}$$

which is consistent with Kingma and Welling (2022) Dorta et al. (2018), and let us define

$$\Psi(z; x) = -\rho^2 \log p(x|z) = \rho^2 \big(\log\det\Sigma + (x-\mu)^\dagger \Sigma^{-1} (x-\mu) \big) + \text{constant}. \tag{99}$$

Let $v := x - \mu(z)$. Since $\mu(z)$ is holomorphic (Σ is not holomorphic), we will treat $z, \overline{z}$ as independent, so

$$\partial_\alpha \Psi = \rho^2 \Big(\mathrm{Tr}(\Sigma^{-1}\partial_\alpha \Sigma) - v^\dagger \Sigma^{-1} \partial_\alpha \mu - (\partial_\alpha \mu)^\dagger \Sigma^{-1} v - v^\dagger \Sigma^{-1} (\partial_\alpha \Sigma) \Sigma^{-1} v \Big), \tag{100}$$

and similarly for $\partial_{\overline{\beta}} \Psi$.

For the complex Gaussian, we have

$$\mathbb{E}[v] = 0, \qquad \mathbb{E}[vv^\dagger] = \Sigma, \qquad \mathrm{Cov}(v^\dagger A v,\, v^\dagger B v) = \mathrm{Tr}(A\Sigma B\Sigma), \tag{101}$$

for any fixed matrices A, B. Thus, we can compute the Hessian. Observe two terms are annihilated by the holomorphic independence, so

$$\partial_{\overline{\beta}} \partial_\alpha \Psi = \rho^2 \Big[\partial_{\overline{\beta}} \Big(\mathrm{Tr}(\Sigma^{-1} \partial_\alpha \Sigma) \Big) - \partial_{\overline{\beta}} \Big(v^\dagger \Sigma^{-1} (\partial_\alpha \Sigma) \Sigma^{-1} v \Big) \Big]. \tag{102}$$

and notice

$$\partial_{\overline{\beta}} \Big(\mathrm{Tr}(\Sigma^{-1} \partial_\alpha \Sigma) \Big) = \mathrm{Tr} \Big((\partial_{\overline{\beta}} \Sigma^{-1}) \partial_\alpha \Sigma + \Sigma^{-1} \partial_{\overline{\beta}} \partial_\alpha \Sigma \Big), \tag{103}$$

$$\partial_{\overline{\beta}} \Big(v^\dagger \Sigma^{-1} (\partial_\alpha \Sigma) \Sigma^{-1} v \Big) = v^\dagger (\partial_{\overline{\beta}} \Sigma^{-1}) (\partial_\alpha \Sigma) \Sigma^{-1} v + v^\dagger \Sigma^{-1} (\partial_\alpha \Sigma)(\partial_{\overline{\beta}} \Sigma^{-1}) v + v^\dagger \Sigma^{-1} (\partial_{\overline{\beta}} \partial_\alpha \Sigma) \Sigma^{-1} v. \tag{104}$$

We will use the identities

$$\partial_\beta \Sigma^{-1} = -\Sigma^{-1} (\partial_\beta \Sigma) \Sigma^{-1}, \quad \mathbb{E}[\, v^\dagger A v \,] = \mathrm{Tr}(A\Sigma), \quad \mathbb{E}[v^\dagger \Sigma^{-1} \partial_\gamma \mu] = 0. \tag{105}$$

thus

$$\mathbb{E}[\partial_\alpha \partial_{\overline{\beta}} \Psi] = \rho^2 \mathrm{Tr}\big(\Sigma^{-1} (\partial_\alpha \Sigma) \Sigma^{-1} (\partial_{\overline{\beta}} \Sigma) \big). \tag{106}$$

Also, and since $\partial_{\overline{\beta}} \mu^\dagger = (\partial_\beta \mu)^\dagger$

$$\mathrm{Cov}(\partial_\alpha \Psi, \partial_{\overline{\beta}} \Psi) \tag{107}$$

$$= \mathbb{E}[\partial_\alpha \Psi \partial_{\overline{\beta}} \Psi] - \underbrace{\mathbb{E}[\partial_\alpha \Psi] \mathbb{E}[\partial_{\overline{\beta}} \Psi]}_{=0} = \mathbb{E}[\partial_\alpha \Psi \partial_{\overline{\beta}} \Psi] \tag{108}$$

$$= \mathbb{E} \Bigg[\rho^2 \Big(\mathrm{Tr}(\Sigma^{-1} \partial_\alpha \Sigma) - v^\dagger \Sigma^{-1} \partial_\alpha \mu - (\partial_\alpha \mu)^\dagger \Sigma^{-1} v - v^\dagger \Sigma^{-1} (\partial_\alpha \Sigma) \Sigma^{-1} v \Big) \tag{109}$$

$$\times \quad \rho^2\Big(\mathrm{Tr}(\Sigma^{-1}\partial_{\overline{\beta}}\Sigma) - v^\dagger\Sigma^{-1}\partial_{\overline{\beta}}\mu - (\partial_{\overline{\beta}}\mu)^\dagger\Sigma^{-1}v - v^\dagger\Sigma^{-1}(\partial_{\overline{\beta}}\Sigma)\Sigma^{-1}v\Big)\Bigg] \tag{110}$$

$$= \rho^4\Bigg(\mathbb{E}\Big[(v^\dagger\Sigma^{-1}\partial_\alpha\mu)\left((\partial_{\overline{\beta}}\mu)^\dagger\Sigma^{-1}v\right)\Big] \tag{111}$$

$$+ \mathbb{E}\Bigg[\Big(\mathrm{Tr}(\Sigma^{-1}\partial_\alpha\Sigma) - v^\dagger\Sigma^{-1}(\partial_\alpha\Sigma)\Sigma^{-1}v\Big)\Big(\mathrm{Tr}(\Sigma^{-1}\partial_{\overline{\beta}}\Sigma) - v^\dagger\Sigma^{-1}(\partial_{\overline{\beta}}\Sigma)\Sigma^{-1}v\Big)\Bigg]\Bigg) \tag{112}$$

$$= \rho^4\Bigg((\partial_\alpha\mu)^\dagger\Sigma^{-1}(\partial_\beta\mu) + \mathrm{Tr}(\Sigma^{-1}\partial_\alpha\Sigma)\mathrm{Tr}(\Sigma^{-1}\partial_{\overline{\beta}}\Sigma) - \mathrm{Tr}(\Sigma^{-1}\partial_\alpha\Sigma)\mathbb{E}\Big[v^\dagger\Sigma^{-1}(\partial_{\overline{\beta}}\Sigma)\Sigma^{-1}v\Big] \tag{113}$$

$$- \mathrm{Tr}(\Sigma^{-1}\partial_{\overline{\beta}}\Sigma)\mathbb{E}\Big[v^\dagger\Sigma^{-1}(\partial_\alpha\Sigma)\Sigma^{-1}v\Big)\Big] + \mathbb{E}\Big[v^\dagger\Sigma^{-1}(\partial_\alpha\Sigma)\Sigma^{-1}vv^\dagger\Sigma^{-1}(\partial_{\overline{\beta}}\Sigma)\Sigma^{-1}v\Big)\Big]\Bigg) \tag{114}$$

$$= \rho^4\Bigg((\partial_\alpha\mu)^\dagger\Sigma^{-1}(\partial_\beta\mu) - \mathrm{Tr}(\Sigma^{-1}\partial_\alpha\Sigma)\mathrm{Tr}(\Sigma^{-1}\partial_{\overline{\beta}}\Sigma) + \mathbb{E}\Big[v^\dagger\Sigma^{-1}(\partial_\alpha\Sigma)\Sigma^{-1}vv^\dagger\Sigma^{-1}(\partial_{\overline{\beta}}\Sigma)\Sigma^{-1}v\Big)\Big]\Bigg) \tag{115}$$

$$= \rho^4\Bigg((\partial_\alpha\mu)^\dagger\Sigma^{-1}(\partial_\beta\mu) - \mathrm{Tr}(\Sigma^{-1}\partial_\alpha\Sigma)\mathrm{Tr}(\Sigma^{-1}\partial_{\overline{\beta}}\Sigma) \tag{116}$$

$$+ \mathrm{Tr}(\Sigma^{-1}(\partial_\alpha\Sigma))\mathrm{Tr}(\Sigma^{-1}(\partial_{\overline{\beta}}\Sigma)) + \mathrm{Tr}\Big(\Sigma^{-1}\partial_{\alpha\Sigma}\Sigma^{-1}\partial_{\overline{\beta}}\Sigma\Big)\Bigg) \tag{117}$$

$$= \rho^4\Bigg((\partial_\alpha\mu)^\dagger\Sigma^{-1}(\partial_\beta\mu) + \mathrm{Tr}\Big(\Sigma^{-1}(\partial_\alpha\Sigma)\Sigma^{-1}(\partial_{\overline{\beta}}\Sigma)\Big)\Bigg). \tag{118}$$

We have used Petersen and Pedersen (2012)

$$\mathbb{E}[v^\dagger Avv^\dagger Bv] = \mathrm{Tr}(A\Sigma)\mathrm{Tr}(B\Sigma) + \mathrm{Tr}(A\Sigma B\Sigma). \tag{119}$$

Thus,

$$\mathbb{E}[\partial_\alpha\partial_{\overline{\beta}}\Psi] + \frac{1}{\rho^2}\,\mathrm{Cov}(\partial_\alpha\Psi, \partial_{\overline{\beta}}\Psi) = \rho^2\Big((\partial_\alpha\mu)^\dagger\Sigma^{-1}(\partial_\beta\mu) + \mathrm{Tr}\big(\Sigma^{-1}(\partial_\alpha\Sigma)\Sigma^{-1}(\partial_{\overline{\beta}}\Sigma)\big)\Big) \tag{120}$$

exactly.

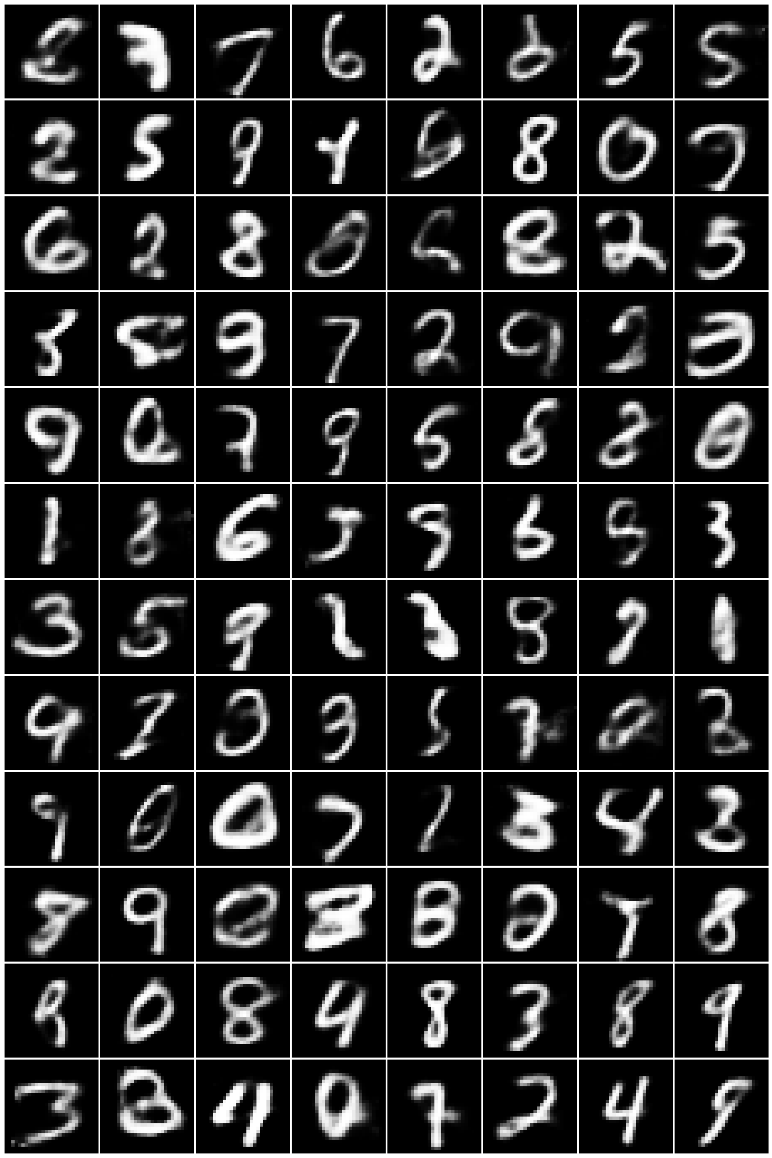

Figure 4: Uncurated generative samples from a Kähler-VAE with our proposed sampling procedure of 96 on the MNIST dataset.