Analysis of high dimensional distributions with decoupled norm and direction

Jose Gallego
Mila and DIRO
Université de Montréal
gallegoj@mila.quebec

Ankit Vani
Mila and DIRO
Université de Montréal
ankit.vani@umontreal.ca

Max Schwarzer
Mila and DIRO
Université de Montréal
max.schwarzer@umontreal.ca

Sanae Lotfi
CERC
Polytechnique Montréal and CentraleSupélec
sanae.lotfi@umontreal.ca

Abstract

In this work we explore theoretical and practical properties of high dimensional distributions which arise from a decoupling of the sampling of the norm and the direction into independent random variables. We provide the first characterization of maximum entropy distributions in high dimensions with non-increasing norm densities. We use these finding as a prior in the context of generative models, and illustrate its advantages as data augmentation strategy.

1 Introduction

Modern machine learning methods rely heavily on probability distributions in very high dimensions: Bayesian neural networks [3] establish priors on the weights, GANs [5] involve a choice of a noise distribution, and VAEs [12] require parametric probabilistic encoders and decoders as well as a prior on the latent space. Multivariate Gaussian random variables are the most common choice for these distributions. This approach has several theoretical and practical justifications: these distributions are numerically stable; have analytic expressions for information theory and other probability theory quantities; are the maximum entropy continuous distribution with fixed variance; and play a crucial role in the Central Limit Theorem.

However, multivariate distributions with independent and identically distributed components exhibit a property in which most of the probability mass is concentrated away from the mean and around a hypersphere. This is due to the way in which volume behaves in high dimensions. Huszár [8] provides a nice overview of this behavior for high dimensional Gaussians, and points out that these distributions in high dimensions resemble *soap bubbles*, contrary to widespread intuitions regarding their behavior. This is an undesirable property in many tasks that require sampling, as it implies the existence of a mismatch between the properties of the mean of the distribution and those observed in samples. This effect is illustrated in Figure 1.

The maximum entropy property makes Gaussian distributions a good choice for priors over latent variables without imposing additional structure. However, in many applications in machine learning, we also introduce encoders for mapping data into the latent space, and the latent variables take the role of codes. From the point of view of coding theory, it is reasonable to expect that codes which have larger density should be sampled more frequently. This expectation is matched in the 1-dimensional case, but does not hold in higher dimensions because of the soap bubble effect mentioned previously. In short, we want to make sure that regions in code space in which we have prescribed a large probability *density* also have large probability *mass*.

Preprint. Under review.

The rest of this paper is structured as follows. In Section 2 we formally present our model along with our theoretical contributions. Section 3 displays the results of our experiments and comparisons to state-of-the-art methods. Finally, Sections 4 and 5 present a brief summary of related approaches and the conclusions of our work, respectively. We provide proofs for our theorems in Appendix A, along with details on our experimental settings.

The code for this project can be found at `https://github.com/jgalle29/soap_bubbles`.

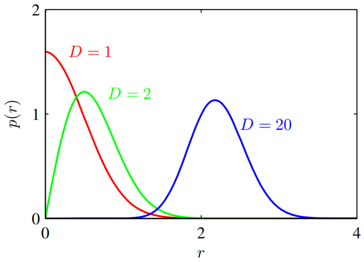

Figure 1: Density of the norm of a D-dimensional standard Gaussian. Taken from [2].

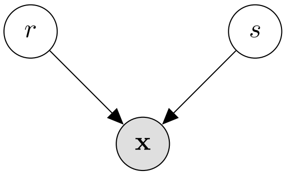

Figure 2: $r \sim p_r$, $\mathbf{s} \sim \mathcal{U}(\mathcal{S}^{n-1}(1))$ and $p(\mathbf{x}|r, \mathbf{s}) = \delta_{r\mathbf{s}}(\mathbf{x})$.

Notation

We follow standard notations for scalars (e.g. r, n) and vectors (e.g. $\mathbf{x}$, $\mathbf{s}$). $\mathcal{S}^{n-1}(r)$ represents the $(n-1)$-sphere embedded in $\mathbb{R}^n$, i.e., $\mathcal{S}^{n-1}(r) = \{\mathbf{x} \in \mathbb{R}^n \mid \|x\| = r\}$. $\mathcal{A}^{n-1}(r)$ denotes the surface area of $\mathcal{S}^{n-1}(r)$. For a compact d-dimensional manifold Ω on $\mathbb{R}^n$, $\mathcal{U}(\Omega)$ represents the uniform distribution over Ω. Let A be a set and $a \in A$; δ_a denotes a Dirac distribution at a. Unless noted otherwise, all norms refer to the ℓ_2 norm. $\Gamma(\cdot)$ and $\psi(\cdot)$ denote the gamma and digamma functions.

2 Theory

We propose to decouple the sampling process of a multidimensional point $\mathbf{x}$ into distributions over its norm and direction, independent of each other. This decomposition allows us to ensure that the norm of the samples follows the arbitrary univariate distribution p_r on $[0, \infty)$.

$$r \sim p_r(r) \qquad \mathbf{s} \sim \mathcal{U}(\mathcal{S}^{n-1}(1)) \qquad \mathbf{x} = r \cdot \mathbf{s} \tag{1}$$

Let us cast several common distributions in this framework in terms of the implicit assumption being made on the distribution of the norm. First, a Dirac distribution at the origin in $\mathbb{R}^n$ corresponds to a choice of p_r to be a δ_0 in $\mathbb{R}$. More generally, a uniform distribution over $\mathcal{S}^{n-1}(\hat{r})$ is induced by $p_r = \delta_{\hat{r}}$. Furthermore, a choice $r \sim \chi(n)$ recovers the multivariate Gaussian distribution, which suffers from the soap-bubbling effect for all $n > 1$. We highlight the fact that the cause for this lies on the increasing behavior of p_r around 0.

A closer look to the graphical model in Figure 2 reveals that, in general, the individual coordinates of the vector $\mathbf{x}$ are not independent random variables. In Appendix A we show how the specific case of $r \sim \chi(n)$ of the Gaussian model preserves the independence of the coordinates as random variables.

2.1 Properties of the proposed model

Theorem 1. *The density of a random variable $\mathbf{x} \in \mathbb{R}^n$ following Equation 1 is given by:*

$$p(\mathbf{x}) = \frac{p_r(\|\mathbf{x}\|)}{\mathcal{A}^{n-1}(\|\mathbf{x}\|)} \tag{2}$$

Furthermore, $\mathbb{E}[\mathbf{x}] = \mathbf{0}$ *and* $\mathbb{V}[\mathbf{x}] = \frac{\mathbb{E}[r^2]}{n}\mathbb{I}$.

We now provide an intuitive way to understand the content of the previous theorem. Consider the distribution of the norms, p_r. For each possible radius r, the model is taking the density $p_r(r)$ and *spreading* it uniformly over the sphere $\mathcal{S}^{n-1}(r)$. Note that this is implicitly *penalizing* vectors with high norm since in those regions, the density provided by p_r needs to be large in order to compensate for the increase in the surface area of a sphere of a larger radius.

Standard affine transformation can be applied to random variables following our decoupled parameterization in order to obtain distributions with arbitrary mean and covariance structure. For simplicity, we concentrate our theoretical developments on isotropic distributions centered at the origin.

Theorem 2. *The (differential) entropy of a random variable* $\mathbf{x} \in \mathbb{R}^n$ *distributed according to the graphical models displayed in Figure 2 is given by:*

$$\mathbb{H}[\mathbf{x}] = \mathbb{H}[r] + \mathbb{E}_r\left[\log\left(\mathcal{A}^{n-1}(r)\right)\right] = \mathbb{H}[r] + \log\left(\frac{2\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}\right) + (n-1)\mathbb{E}[\log(r)] \tag{3}$$

Theorem 3. *Let* p_r *and* q_r *be two distributions with support on* $\mathbb{R}_{\geq 0}$*. Let* $p(\mathbf{x})$ *and* $q(\mathbf{x})$ *be the distributions on* $\mathbb{R}^n$ *induced by* p_r *and* q_r*, respectively.*

$$\mathbb{KL}(p(\mathbf{x})\,||\,q(\mathbf{x})) = \mathbb{KL}(p_r(r)\,||\,q_r(r)). \tag{4}$$

2.2 Maximum Entropy Distribution

As we have mentioned previously, the choice of a Gaussian distribution as the *de facto* default in many machine learning applications leads to the apparently paradoxical soap-bubbling behavior. In this section we explore the theoretical consequences that arise when one imposes the constraint of a non-increasing density in the norm from a maximum entropy point of view. In other words, if we require the distribution p_r to have a non-increasing density, what is the class of such distributions p_r which minimize the amount of prior information built into the high-dimensional distribution $p(\mathbf{x})$?

To answer this question, we first make a brief detour and define the Generalized Gamma distribution and present some of its properties:

Definition 1. *The Generalized Gamma distribution with parameters* $\alpha, \beta > 0$*,* $\eta \neq 0$ *is denoted by* $\tilde{\Gamma}(\alpha, \beta, \eta)$*. It has support on* $\mathbb{R}_{\geq 0}$*. Its density function is given by:*

$$p(r; \alpha, \beta, \eta) = \frac{|\eta|}{\Gamma(\alpha)\beta^{\alpha\eta}}\, r^{\alpha\eta - 1} \exp\left[-\left(\frac{r}{\beta}\right)^{\eta}\right] \quad \textit{for } r \in \mathbb{R}_{\geq 0} \tag{5}$$

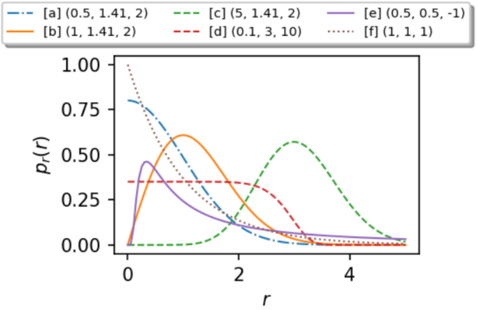

Figure 3: PDF of a Generalized Gamma distribution for several parameter configurations. Captions stand for parameters (α, β, η).

Distribution	$\tilde{\Gamma}(\alpha, \beta, \eta)$
$\chi(n)$	$(\frac{n}{2}, \sqrt{2}, 2)$
$\Gamma(\kappa, \theta)$	$(\kappa, \theta, 1)$
$\lvert$Gen-$\mathcal{N}\rvert(0, \beta, \eta)$	$(\frac{1}{\eta}, \beta, \eta)$
$(\Gamma(\kappa, \theta))^q$	$(\kappa, \theta^q, \frac{1}{q})$
Lévy(c)	$(\frac{1}{2}, \frac{c}{2}, -1)$

Table 1: Some distributions related to the Generalized Gamma.

Figure 3 shows the behavior of the density of the Generalized Gamma distribution for several configurations of the parameters. Along with Table 1 we see that we recover plenty of familiar distributions as special cases. We highlight some relevant examples.

First, lines [b] and [c] in Figure 3 correspond to $\chi(10)$ distribution, which exemplifies the soap-bubbling behavior mentioned earlier. Line [a] is a folded standard Gaussian distribution, i.e., the distribution of the absolute value of a standard normal. Whenever the product $\alpha\eta = 1$, we recover a folded generalized Gaussian distribution, in which the density is proportional to $\exp\left(-\left(r/\beta\right)^{\eta}\right)$.

Finally, line [e] shows that $\eta = -1$ and $\alpha = \frac{1}{2}$ induces a Lévy distribution. An important property of the Lévy distribution (shared with the Gaussian and Cauchy distributions) is that it is *stable*: the sum of two Lévy random variables follows again a Lévy distribution. This is particularly useful in the context of interpolation in generative models, since we want to ensure that the distribution of the codes used during the training of the generator $(\mathbf{z}, \mathbf{z}' \sim p(\mathbf{z}))$ and that of points along interpolation paths $(\theta\mathbf{z} + (1-\theta)\mathbf{z}')$ enjoy similar properties.

Lemma 1. *Let r be a random variable following a $\tilde{\Gamma}(\alpha, \beta, \eta)$ distribution and let $k \in \mathbb{N}$.*

$$\mathbb{E}[r^k] = \frac{\beta^k \Gamma\left(\alpha + \frac{k}{\eta}\right)}{\Gamma(\alpha)}, \quad \mathbb{E}[\log(r)] = \log(\beta) + \frac{\psi(\alpha)}{\eta}, \quad \mathbb{H}[r] = \left(\frac{1}{\eta} - \alpha\right)\psi(\alpha) + \log\left(\frac{\beta\Gamma(\alpha)}{\eta}\right) + \alpha. \tag{6}$$

Theorem 4 (Main Result)**.** *The distribution $p_r(r)$ on the norms which induces maximum entropy on the high-dimensional distribution $p(\mathbf{x})$ subject to the constraints that p_r is non-increasing and has a fixed second moment, is given by a Generalized Gamma distribution.*

We refer the interested reader to Appendix A for a proof of this result along with several insights on the nature of this result. Remarkably, although an analytic solution to express the parameters of the maximum entropy distribution seems elusive, our numerical experiments allow us to conjecture that this corresponds to a folded Generalized Gaussian distribution on the norm.

One of the consequences of our proof is that the value of η^* grows (approximately linearly) with the dimension of the space, n. This is relevant since the density of a $\tilde{\Gamma}(\frac{1}{\eta}, \beta, \eta)$ distribution converges point-wise to that of a $\mathcal{U}(0, \beta)$. Note that this is a uniform distribution over the norm, rather than a uniform distribution over a ball (in $\mathbf{x}$) of radius β. This implies that we draw samples heavily around the origin and decrease polynomially fast (with degree $1 - n$) as we move away from the origin. This behavior is shown by line [d] in Figure 3.

3 Experiments

We experimented with WGAN [1] with gradient penalty [6] with a Gaussian prior and our proposed prior, i.e., a maximum-entropy Generalized Gamma norm distribution. We report our results on MNIST [13] and cropped LFW faces [7, 16] datasets. Unless specified otherwise, ”our distribution" refers to the maximum entropy distribution with non-increasing p_r with identity covariance. In other words, $p_r(\|\mathbf{x}\|) = \tilde{\Gamma}(\alpha, \beta, \eta)$ where $\eta \approx 3.27n$, $\alpha = 1/\eta$, $\beta = \sqrt{n\Gamma(\alpha)/\Gamma(3\alpha)}$, and n is the dimension of the latent space. The Gamma distance model proposed by Kilcher et al. [10] replaces η by 2 and α by $1/2$.

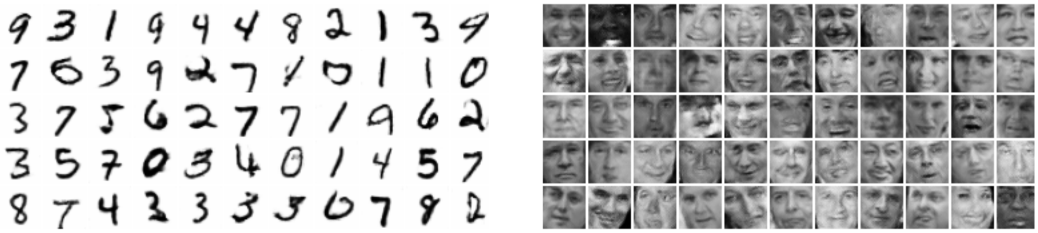

Figure 4: Samples generated by a WGAN trained using our proposed distribution on the latent space for MNIST (left) and LFW (right).

Figure 4 shows samples drawn according to our proposed distribution in the latent space, generated by a WGAN trained using the same distribution. The samples generated by latent codes away from zero are qualitatively comparable to those generated by WGAN trained with a Gaussian prior. However, Figure 5 shows that our model behaves better closer to the origin than models trained with a Gaussian prior. This is not unexpected: the Gaussian prior rarely provides the generator with codes that have low norm, and thus the generator network has not been trained in that region. Further experimental details are provided in Appendix B.

Interestingly, we observe that there exists a gap in the lower bound of the Wasserstein distance between the real data distribution and the generated distribution for models that are trained with a Gaussian prior versus our model, shown in Figure 6. We hypothesize this is due to the fact that our model places a large density our model, which comes from the normalization by the area of a sphere of a very small radius. Therefore, we often sample noise vector close to zero, at which the model is unable to generate samples with high quality since those are easily detected by the critic to be fake.

Another interesting application of our distribution is on data augmentation. Adding Gaussian noise to high-dimensional data (e.g. images) induces samples located on a sphere at a certain radius away

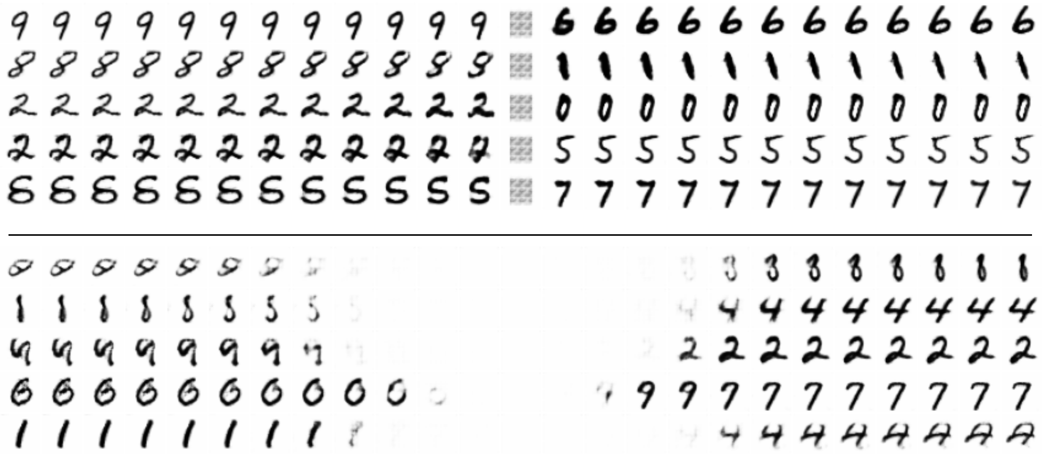

Figure 5: Interpolations through the origin for a WGAN trained with our proposal (top) and a Gaussian distribution (bottom) on a 100-dimensional latent space. The horizontal scale is comparable for both models.

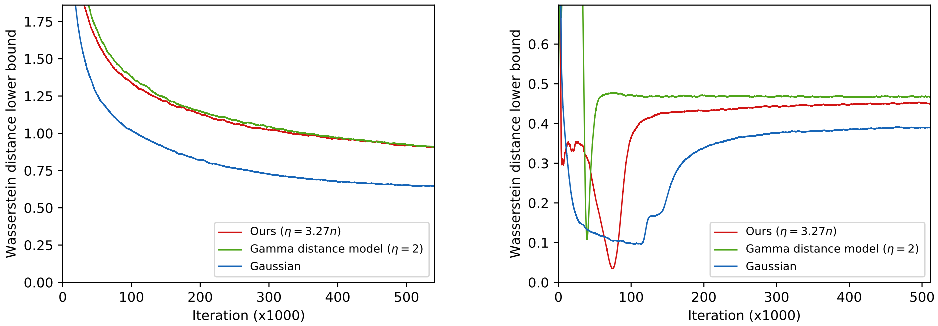

Figure 6: Lower bound on the Wasserstein distance estimated by the WGAN critic during training for MNIST (left) and LFW (right).

from the original observation at the center. Note that, as a result of this, the original datapoint will rarely be sampled. Figure 7 shows how using our distribution and the Gamma distance model compares with noising using Gaussian distributions. Recall that in high dimensions, our maximum entropy distribution converges point-wise to a uniform distribution over a certain ranges of norms. This produces a "uniform degradation" of the image across samples, compared to that of a Gaussian distribution, which adds perturbations of (roughly) constant magnitude.

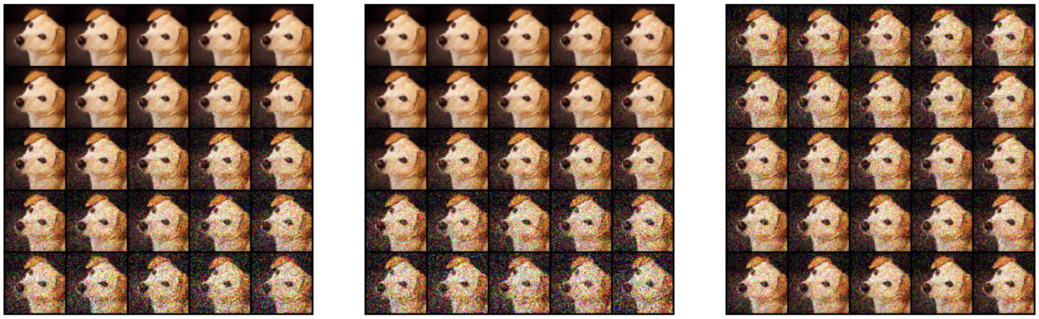

Figure 7: Stochastic perturbation of image data using our maximum entropy distribution (left), $r^2 \sim \Gamma(\frac{1}{2}, \theta)$ (center) (see Section 4) and isotropic Gaussian noise (right). All noise distributions have mean zero and identical variance. The samples are sorted by the norm of the noise vector in ascending order.

Having the ability to sample close to the zero vector as well as farther away ensures that we obtain both the image of interest as well as its noisy variations. Among other applications, such image data augmentation can prove useful in regularizing classification models (which can make them more robust to adversarial examples [17]) or training denoising autoencoders [19].

4 Related work

The idea of reparameterizing high-dimensional vectors into their magnitude and directional components has been explored previously. [10] choose a family of distributions $\{\Gamma(\frac{1}{2}, \theta)\}_{\theta>0}$ for the distribution of r^2. However, they do not perform a thorough analysis of the general reparametrization for high-dimensional distributions. This turns our to be a folded Gaussian distribution or, in the shape of a Generalized Gamma, $\tilde{\Gamma}(\frac{1}{2}, \theta, 2)$, which is a specific case of our maximum entropy model. However, this specific configuration only provides maximum entropy in the case $n = 1$. The authors call their resulting distribution the "Gamma distance model", and perform experiments aiming to provide more meaningful linear interpolations (through the origin) in the latent space.

[14] provide a similar analysis of the theory of interpolation in generative models using a Cauchy distribution as the distribution of the norms. This is another example of a stable distribution, similar to the Lévy distribution mentioned earlier. In contrast to the behavior of the Cauchy distribution, all the moments of our maximum entropy distribution are finite. Note that the fact that the moments of the Cauchy distribution are infinite, prevents it for being a potential candidate in our search for a maximum entropy distribution since our constraint for the second moment assumes a fixed *finite* value. It is an interesting future direction to explore whether the stability property of the Lévy distribution provides similar benefits in practice as those of the Cauchy distribution.

[4] propose to drop Gaussians altogether and commit entirely to a spherical latent space. They present a way to perform the reparameterization trick and find an expression for the Kullback-Leibler divergence between von Mises-Fisher distributions in several dimensions. We note that the uniform distribution over a sphere considered here is a special case of the family of functions we are examining, where p_r is a delta distribution at a given radius.

In the context of optimization, [15] show that such a decomposition for weights in deep networks improves conditioning and leads to faster and more stable training using stochastic gradient descent.

5 Conclusions

We considered a simple method to represent isotropic distributions as a decomposition of a distribution over norms and a distribution over directions. We further analyze popular distributions as a decomposition of norm and direction, and present theoretical properties of such a decomposition. We found the maximum entropy distribution with fixed variance and a non-increasing norm distribution p_r and learned that the norm is distributed according to a Generalized Gamma. The choice of Generalized Gamma for the norm distribution also generalizes previous work [10].

Our future work involves further analysis of this decomposition in the non-isotropic case. For instance, the proposed decoupling can be extended to distributions with arbitrary mean and covariance , which can be obtained via standard affine transformations of random vectors. While some of our theoretical properties apply trivially or with minor modifications, we found some other analyses to be more challenging. For example, Theorem 3 does not hold when p and q have different means and covariances, and a closed form solution has been elusive for a general norm distribution.

References

[1] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. *arXiv preprint arXiv:1701.07875*.

[2] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[3] Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight Uncertainty in Neural Networks. In *International Conference on Machine Learning*.

[4] Davidson, T. R., Falorsi, L., De Cao, N., Kipf, T., and Tomczak, J. M. (2018). Hyperspherical variational auto-encoders. In *Conference on Uncertainty in Artificial Intelligence*.

[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

[6] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5767–5777. Curran Associates, Inc.

[7] Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. (2008). Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

[8] Huszár, F. (2017). Gaussian Distributions are Soap Bubbles. `https://www.inference.vc/high-dimensional-gaussian-distributions-are-soap-bubble/`. [Online].

[9] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

[10] Kilcher, Y., Lucchi, A., and Hofmann, T. (2018). Semantic interpolation in implicit models. In *International Conference on Learning Representations*.

[11] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[12] Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*.

[13] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[14] Lesniak, D., Sieradzki, I., and Podolak, I. T. (2018). On latent distributions without finite mean in generative models. *arXiv:1806.01670*.

[15] Salimans, T. and Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 901–909.

[16] Sanderson, C. and Lovell, B. C. (2009). Multi-region probabilistic histograms for robust and scalable identity inference. In *International Conference on Biometrics*, pages 199–208. Springer.

[17] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations*.

[18] Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

[19] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408.

A Proofs of things

Theorem 1. *The density of a random variable $\mathbf{x} \in \mathbb{R}^n$ following Equation (1) is given by:*

$$p(\mathbf{x}) = \frac{p_r(\|\mathbf{x}\|)}{\mathcal{A}^{n-1}\left(\|\mathbf{x}\|\right)} \tag{7}$$

Furthermore, $\mathbb{E}[\mathbf{x}] = \mathbf{0}$ *and* $\mathbb{V}[\mathbf{x}] = \frac{\mathbb{E}[r^2]}{n}\mathbb{I}$.

Proof.

$$p(\mathbf{x}) = \int_{\mathbb{R}_{\geq 0}} p(\mathbf{x}|r)\, p_r(r)\, \mathrm{d}r = \int_{\mathbb{R}_{\geq 0}} \frac{p_r(r)}{\mathcal{A}^{n-1}\left(r\right)} \delta_{\|\mathbf{x}\|}(r)\, \mathrm{d}r = \frac{p_r(\|\mathbf{x}\|)}{\mathcal{A}^{n-1}\left(\|\mathbf{x}\|\right)}, \tag{8}$$

where the last equality follows from the property of the Dirac distribution as an evaluation functional: $\delta_\omega[\phi] = \phi[\omega]$.

The expectation of $\mathbf{x}$ vanishes since $\mathbf{x} \mapsto \mathbf{x}\, p(\mathbf{x})$ is an odd mapping. Finally, let $\mathbf{s} \sim \mathcal{U}(\mathcal{S}^{n-1}\left(r\right))$. It is easy to show that $\mathbb{V}[\mathbf{s}] = \frac{1}{n}\mathbb{I}$. Therefore,

$$\mathbb{V}[\mathbf{x}] = \mathbb{E}_r[\mathbb{V}[\mathbf{x}|r]] = \mathbb{E}_r[r^2\, \mathbb{V}[\mathbf{s}]] = \frac{\mathbb{E}[r^2]}{n}\, \mathbb{I}. \tag{9}$$

□

Theorem 2. *The (differential) entropy of a random variable $\mathbf{x} \in \mathbb{R}^n$ distributed according to Equation (1) is given by:*

$$\mathbb{H}[\mathbf{x}] = \mathbb{H}[r] + \mathbb{E}\left[\log\left(\mathcal{A}^{n-1}\left(r\right)\right)\right] = \mathbb{H}[r] + \log\left(\frac{2\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}\right) + (n-1)\mathbb{E}[\log r] \tag{10}$$

Proof.

$$\mathbb{H}[\mathbf{x}] = -\int_{\mathbb{R}^n} p(x) \log p(\mathbf{x})\, \mathrm{d}\mathbf{x} \tag{11a}$$

$$= \int_{r \in \mathbb{R}_{\geq 0}} \int_{\mathbf{s} \in \mathcal{S}^{n-1}(r)} \frac{p_r(r)}{\mathcal{A}^{n-1}\left(r\right)} \log\left(\frac{\mathcal{A}^{n-1}\left(r\right)}{p_r(r)}\right) \mathrm{d}\mathbf{s}\, \mathrm{d}r \tag{11b}$$

$$= \int_{\mathbb{R}_{\geq 0}} p_r(r) \log\left(\frac{\mathcal{A}^{n-1}\left(r\right)}{p_r(r)}\right) \mathrm{d}r \tag{11c}$$

$$= \mathbb{H}[r] + \mathbb{E}\left[\log\left(\mathcal{A}^{n-1}\left(r\right)\right)\right] \tag{11d}$$

$$= \mathbb{H}[r] + \log\left(\frac{2\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}\right) + (n-1)\mathbb{E}[\log r] \tag{11e}$$

□

Theorem 3. *Let p_r and q_r be two distributions with support on $\mathbb{R}_{\geq 0}$. Let $p(\mathbf{x})$ and $q(\mathbf{x})$ be the distributions on $\mathbb{R}^n$ induced by p_r and q_r, respectively.*

$$\mathbb{KL}(p(\mathbf{x}) \,||\, q(\mathbf{x})) = \mathbb{KL}(p_r(r) \,||\, q_r(r)). \tag{12}$$

Proof.

$$\begin{aligned}
\mathbb{KL}(p(\mathbf{x}) \,||\, q(\mathbf{x})) &= \int_{\mathbb{R}^n} p(\mathbf{x}) \log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) \mathrm{d}\mathbf{x} && \text{(13a)}\\
&= \int_{\mathbb{R}^n} \frac{p_r(\|\mathbf{x}\|)}{\mathcal{A}^{n-1}\left(\|\mathbf{x}\|\right)} \log\left(\frac{p_r(\|\mathbf{x}\|)}{\cancel{\mathcal{A}^{n-1}\left(\|\mathbf{x}\|\right)}} \frac{\cancel{\mathcal{A}^{n-1}\left(\|\mathbf{x}\|\right)}}{q_r(\|\mathbf{x}\|)}\right) \mathrm{d}\mathbf{x} && \text{(13b)}\\
&= \int_{r \in \mathbb{R}_{\geq 0}} \int_{\mathbf{s} \in \mathbb{S}^{n-1}(r)} \frac{p_r(r)}{\mathcal{A}^{n-1}\left(r\right)} \log\left(\frac{p_r(r)}{q_r(r)}\right) \mathrm{d}\mathbf{s}\, \mathrm{d}r && \text{(13c)}\\
&= \int_{r \in \mathbb{R}_{\geq 0}} p_r(r) \log\left(\frac{p_r(r)}{q_r(r)}\right) \mathrm{d}r && \text{(13d)}\\
&= \mathbb{KL}(p_r(r) \,||\, q_r(r)) && \text{(13e)}
\end{aligned}$$

□

Lemma 2. *Let r be a random variable following a $\tilde{\Gamma}(\alpha, \beta, \eta)$ distribution and let $k \in \mathbb{N}$.*

$$\mathbb{E}[r^k] = \frac{\beta^k \Gamma\left(\alpha + \frac{k}{\eta}\right)}{\Gamma(\alpha)}, \; \mathbb{E}[\log(r)] = \log(\beta) + \frac{\psi(\alpha)}{\eta}, \; \mathbb{H}[r] = \left(\frac{1}{\eta} - \alpha\right)\psi(\alpha) + \log\left(\frac{\beta\Gamma(\alpha)}{\eta}\right) + \alpha. \tag{14}$$

Theorem 4 (Main Result)**.** *The distribution $p_r(r)$ on the norms which induces maximum entropy on the high-dimensional distribution $p(\mathbf{x})$ subject to the constraints that p_r is non-increasing and has a fixed second moment, is given by a Generalized Gamma distribution.*

Proof. Let p_r be an arbitrary distribution on $\mathbb{R}_{\geq 0}$. Consider the following functional Lagrangian:

$$\mathcal{J}[p_r] = \int_{\mathbb{R}_{\geq 0}} p_r(r) \log\left(\frac{r^{n-1}}{p_r(r)}\right) + \lambda_0 \left(\int_{\mathbb{R}_{\geq 0}} p_r(r)\, \mathrm{d}\, r - 1\right) + \lambda_1 \left(\int_{\mathbb{R}_{\geq 0}} r^2 p_r(r)\, \mathrm{d}\, r - A\right) \tag{15}$$

Note that $\mathcal{J}[p_r]$ represents the entropy of the corresponding distribution induced by p_r on $\mathbb{R}^n$, offset by a constant $\log(\mathcal{A}^{n-1}\left(1\right))$, along with Lagrange multipliers ensuring the proper normalization of p_r, as well as the constraint on the second moment. This latter constraint is related to the fact that the variance of the induced distribution is $\frac{\mathbb{E}[r^2]}{n}\mathbb{I}$. Thus, in order to obtain a distribution $p(\mathbf{x})$ with variance $\sigma^2\mathbb{I}$, we set $\mathbb{E}[r^2] = A = n\sigma^2$.

If one extremizes this Lagrangian with respect to p_r without further restrictions, one arrives at a $\chi(n)$ distribution (multiplied by some adequate constant to satisfy the second moment constraint), corresponding to the familiar result on the Gaussian law being the continuous distribution of maximum entropy with a given variance. Algebraically,

$$\begin{aligned}
\frac{\delta \mathcal{J}[p_r]}{\delta p_r} &= -\log\left(p_r\right) - 1 + \log\left(r^{n-1}\right) + \lambda_0 - \lambda_1 r^2 = 0 && \text{(16a)}\\
p_r(r) &= e^{\lambda_0 - 1} r^{n-1} e^{-\lambda_1 r^2} \sim \text{constant} \cdot \chi(n) && \text{(16b)}
\end{aligned}$$

The factor r^{n-1} induces an increasing behavior in p_r around zero for all $n > 1$. This is precisely the *soap-bubbling* effect we would like to avoid. In the case $n = 1$, this factor vanishes, and thus this effect is not present in one dimension.

Taking a step back, we realize that the functional form of optima for $\mathcal{J}[p_r]$ corresponds to that of a $\tilde{\Gamma}(\alpha, \beta, \eta)$ distribution (whose parameters are yet to be determined). We use this parametric point of view to enforce the non-increasing behavior and the constraint on the second moment of p_r. Note that the parametric treatment provides proper normalization automatically. We restate the density of the Generalized Gamma distribution for convenience.

$$p(r; \alpha, \beta, \eta) = \frac{|\eta|}{\Gamma(\alpha)\beta^{\alpha\eta}}\; r^{\alpha\eta - 1} \exp\left[-\left(\frac{r}{\beta}\right)^{\eta}\right] \qquad \text{for } r \in \mathbb{R}_{\geq 0} \tag{17}$$

We now enforce the non-increasing behavior in p_r, by realizing that this is directly controlled by the product $\alpha\eta$. If we demand that $\alpha\eta \leq 1$, we obtain a density p_r which is non-increasing in r. Note how this is violated in the $\chi(n)$ distribution which corresponds to $\alpha = \frac{n}{2}$ and $\eta = 2$; and how the "magnitude" of the violation of this constraint increases with the dimension of the space.

In the following statements we use heavily the results obtained in Theorem 2 and Lemma 2 to satisfy the constraint on the second moment and to calculate the resulting entropy.

$$\mathbb{E}[r^2] = \frac{\beta^2 \Gamma\left(\alpha + \frac{2}{\eta}\right)}{\Gamma(\alpha)} = A \qquad \Rightarrow \qquad \beta = \sqrt{\frac{A\Gamma(\alpha)}{\Gamma\left(\alpha + \frac{2}{\eta}\right)}} \tag{18}$$

Replacing the value of β in the expression for the entropy for a Generalized Gamma distribution, we obtain:

$$\mathbb{H}[x] = \log\left(\frac{2\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}\right) + \mathbb{H}[r] + (n-1)\mathbb{E}[\log(r)] \tag{19a}$$

$$= \text{cst} + \alpha(1 - \psi(\alpha)) + n\frac{\psi(\alpha)}{\eta} + \frac{n}{2}\log\left(\frac{A\,\Gamma(\alpha)}{\Gamma\left(\alpha + \frac{2}{\eta}\right)}\right) + \log\left(\frac{\Gamma(\alpha)}{\eta}\right) \tag{19b}$$

$$\triangleq H(\alpha, \eta; n, A) \tag{19c}$$

Thus, we have reduced the problem of searching over all possible distributions with non-increasing density to a two-dimensional constrained optimization problem (with two equivalent formulations). We make explicit the dependence of the optimization problem at hand on the dimension of the space n, and the constraint on the raw second moment A.

$$\max_{\alpha\eta \leq 1} H(\alpha, \eta; n, A) \;\equiv\; \max_{\eta > 0,\, \alpha \leq \frac{1}{\eta}} H(\alpha, \eta; n, A) \tag{20}$$

Unfortunately, the nature of the functions involved in the optimization did not allow us to obtain an analytic solution for what the optimal α^* and η^* are, as a function of n and A. However, we performed numerical optimization jointly over α and η for several values of n and obtained several interesting findings.

For all of the experiments we performed, the constraint $\alpha\eta \leq 1$ was active at the optimum. This implies that our maximum entropy model corresponds to a specific sub-family among the Generalized Gamma distribution: $\tilde{\Gamma}\left(\frac{1}{\eta^*}, \beta, \eta^*\right)$, where β is set according to Equation 18. This is precisely the distribution of a folded Generalized Gaussian distribution.

In particular, the numerical optimization yields that for the case $n = 1$, $\eta^* = 2$ (and thus, since we have a tight constraint, $\alpha^* = \frac{1}{2}$). This is equivalent to a folded Gaussian distribution of adequate variance, consistent with the fact that 1-dimensional Gaussian distributions do not experience the soap-bubbling behavior.

Even though the optimization problem in Eq. (20) can be solved efficiently, we found that the optimal value for η^* (with $\sigma^2 = 1$, and thus $A = n$) follows a (surprisingly) linear behavior, illustrated in Figure 8. We used this approximation in our applications.

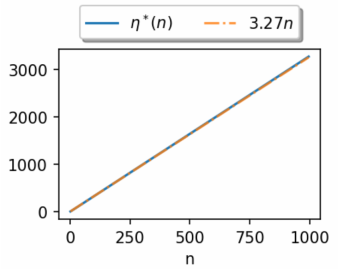

Figure 8: Approximation for the optimal configuration η^* as a function of the dimension.

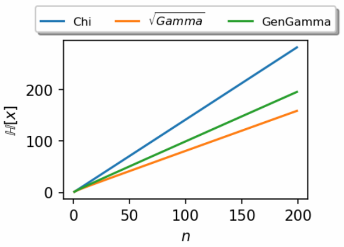

Figure 9: Differential entropy for several distributions of the norm.

Finally, Figure 9 shows the entropy of the high-dimensional distribution induced by several choices of the distribution of the norms. We use the $\chi(n)$ distribution (which corresponds to a Gaussian distribution in $\mathbf{x}$) as an upper-bound on the possible entropy that can be obtained in n dimensions, and compare to the $\sqrt{\Gamma\left(\frac{1}{2}, \theta\right)}$ model employed by Kilcher et al. [10]. Our findings clearly perform better (in terms of entropy) than current proposed solutions to the soap-bubbling phenomenon.

$\square$

B WGAN experimental details

We used a WGAN-GP [1, 6] for our experiments. In practice, we noticed that the gradient penalty converges early in the training, so the results reported are not affected by the gradient penalty except the very early steps.

Unless specified otherwise, “our distribution" refers to the maximum entropy distribution with non-increasing p_r with identity covariance. In other words, $p_r(\|\mathbf{x}\|) = \tilde{\Gamma}(\alpha, \beta, \eta)$ where $\eta \approx 3.27n$, $\alpha = 1/\eta$, $\beta = \sqrt{n\Gamma(\alpha)/\Gamma(3\alpha)}$, and n is the dimension of the latent space. The Gamma distance model proposed by Kilcher et al. [10] replaces η by 2 and α by $1/2$.

Our input images are greyscale and resized to 32×32, and the latent space has 100 dimensions. For both the generator and the discriminator, we used the Adam optimizer [11] with learning rate 0.0001, $\beta_1 = 0.5$, $\beta_2 = 0.9$. We updated the generator after every 5 discriminator training steps. We used a batch size of 64 during training.

Both the discriminator and the generator consist of four convolutional layers. The discriminator uses the leaky ReLU non-linearity with instance normalization [18], whereas the generator uses the ReLU non-linearity with batch normalization [9].