## A. Annealing the Noise

In section(1.1.1) we discussed the common approach to first adding noise to a model $\mathbb{Q}$ in order to define a proper density and then using maximum likelihood to fit that 'noised model' to data.

We can use standard Woodberry identities to rewrite the expected log likelihood eq(11) as

$$-\frac{\theta_p^2}{\sigma^2} + \frac{\left(\theta_p^\mathsf{T}\theta_q\right)^2}{\sigma^2\left(\sigma^2+\theta_q^2\right)} - \log\left(1+\frac{\theta_q^2}{\sigma^2}\right) - D\log\sigma^2. \tag{65}$$

where $D = \dim(\theta_p)$.

Differentiating wrt $\theta_q$, we note that the optimal solution is given when

$$\theta_q = \gamma\theta_p, \tag{66}$$

for scalar $\gamma$. Plugging this form back into eq(65) we find that the optimum is obtained when

$$\theta_q = \sqrt{\frac{\theta_p^2-\sigma^2}{\theta_p^2}}\theta_p. \tag{67}$$

For finite Gaussian noise $\sigma^2 > 0$ the resulting estimator for the toy model in section(1.1.1) is therefore not consistent.

A natural question is what would happen if one uses a numerical optimisation of eq(65) but anneals the noise $\sigma^2$ to zero during the optimisation process? As $\sigma^2$ tends to zero, the expression eq(65) blows up. This means that a naive approach to annealing $\sigma^2$ towards zero whilst using a standard optimisation technique is unlikely to result in $\theta_q$ converging to $\theta_p$. However, if one considers removing the additive constant $D\log\sigma^2$ and multiplying the remaining objective by $\sigma^2$, the resulting quantity

$$\frac{\left(\theta_p^\mathsf{T}\theta_q\right)^2}{\left(\sigma^2+\theta_q^2\right)} - \sigma^2\log\left(1+\frac{\theta_q^2}{\sigma^2}\right), \tag{68}$$

is well-behaved as $\sigma^2 \to 0$, as plotted in figure(7).

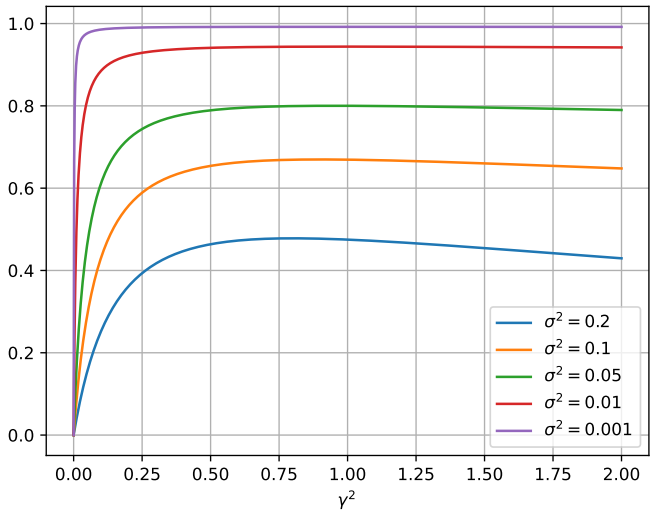


*Figure 7.* The (modified) expected log likelihood eq(68) when adding noise $\sigma^2$ to the model only and for unit length true data generating parameter $\theta_p^2 = 1$. The $x$-axis is the value $\gamma^2$ assuming that the optimal $\theta_q$ is of the form $\theta_q = \gamma\theta_p$. As we see, as $\sigma^2 \to 0$ the scaled objective becomes flat around the optimum point $\gamma^2 = 1$.

Nevertheless, the objective eq(68) becomes flat (with respect to $\theta_q$) around the optimum as $\sigma^2 \to 0$. In figure(7) we plot the scaling behaviour of the objective eq(68), assuming $\theta_q = \gamma\theta_p$, showing how it becomes flat with respect to $\gamma$ as $\sigma^2$ is annealed towards zero. This means that a standard first-order numerical optimisation approach, even for this modified objective, will result in a 'critical slowing down' phenomenon, leading to $\theta_q$ not updating. This might be fixable by taking the curvature of the objective into consideration.

However, addressing all the above issues requires an understanding of the small $\sigma^2$ behaviour of the original objective; dealing with arbitrarily large constants, arbitrarily large scaling and loss of curvature. In general, such insight is unlikely to be available for any given implicit generative model. Thus, we are doubtful that it will be possible to find an annealing schedule and associated general numerical optimisation procedure that will result in a consistent estimator.

## B. Noise Requirements for Discrete Distributions

Our main interest is to define a new divergence in situations where the original divergence $\mathrm{D}(p||q)$ is itself not defined. For discrete variables $x \in \{1, \ldots, n\}$, $y \in \{1, \ldots, n\}$, the spread $P_{ij} = p(y = i|x = j)$ must be a distribution; $\sum_i P_{ij} = 1$, $P_{ij} \geq 0$, and

$$\tilde{p}_i \equiv \sum_j P_{ij} p_j = \sum_j P_{ij} q_j \equiv \tilde{q}_i \qquad \forall i \tag{69}$$

$$\Rightarrow \qquad p_j = q_j \qquad \forall j, \tag{70}$$

which is equivalent to the requirement that the matrix $P$ is invertible. In addition, for the Spread Divergence to exist in the case of $f$-divergences, $\tilde{p}$ and $\tilde{q}$ must have the same support. This requirement is guaranteed if

$$\sum_j P_{ij} p_j > 0, \qquad \sum_j P_{ij} q_j > 0 \qquad \forall i, \tag{71}$$

which is satisfied if $P_{ij}>0$. Therefore, in general there is a space of spread distributions $p(y|x)$ that define a valid Spread Divergence in the discrete case.

## C. Validity of Stationary Spread $f$-Divergence

**Lemma 1.** *let $X$ and $Y$ be two random variables with Borel probability measure $\mathbb{P}_X$ and $\mathbb{P}_Y$. Let $K$ be an absolutely continuous random variable that is independent of $X$ and $Y$ and has density function $p_K(x)$. We define $\tilde{X}$ and $\tilde{Y}$ as*

$$\tilde{X} = X + K, \quad \tilde{Y} = Y + K,$$

*with distribution $\mathbb{P}_{\tilde{X}}$ and $\mathbb{P}_{\tilde{Y}}$. Then $\tilde{X}$ and $\tilde{Y}$ are absolutely continuous with density functions*

$$p_{\tilde{X}}(\tilde{x}) = \int_x p_K(\tilde{x} - x) d\mathbb{P}_X, \quad p_{\tilde{Y}}(\tilde{y}) = \int_y p_K(\tilde{y} - y) d\mathbb{P}_Y.$$

*Proof.* The proof can be found in Durrett (2019, Theorem 2.1.16). □

**Theorem 1.** *Let $X$ and $Y$ be two random variables[12] with Borel probability measure $\mathbb{P}_X$ and $\mathbb{P}_Y$. Let the stationary spread noise $K$ be an absolutely continuous random variable that is independent of $X$ and $Y$, and its density function $p_K(x)$ has support[13] $\mathbb{R}$. Using lemma(1) we define spreaded random variables $\tilde{X} = X + K$, $\tilde{Y} = Y + K$ with density functions $p_{\tilde{X}}$, $p_{\tilde{Y}}$. We then define the stationary spread $f$-divergence between $\mathbb{P}_X$ and $\mathbb{P}_Y$ as*

$$\tilde{D}_f(\mathbb{P}_X||\mathbb{P}_Y) \equiv D_f(p_{\tilde{X}}||p_{\tilde{Y}}) \equiv \int f\left(\frac{p_{\tilde{X}}(x)}{p_{\tilde{Y}}(x)}\right) p_{\tilde{Y}}(x) dx.$$

*Furthermore, denote the characteristic function[14] of the spread noise $K$ by $\phi_K$. Given $\phi_K \neq 0$ or $\phi_K > 0$ on at most a countable set, then the stationary spread $f$-divergence is a valid divergence with the properties*

$$\tilde{D}_f(\mathbb{P}_X||\mathbb{P}_Y) \geq 0, \quad \tilde{D}_f(\mathbb{P}_X||\mathbb{P}_Y) = 0 \Leftrightarrow \mathbb{P}_X = \mathbb{P}_Y.$$

[12] We don't require $X$ (or $Y$) to be absolutely continuous.

[13] The extension to higher dimensions is straightforward.

[14] When a distribution $\mathbb{P}_X$ allows a density function $p_X$, its characteristic function is equal to the Fourier transform of the density function: $\phi_X = \mathcal{F}\{p_X\}$, so the Fourier transform treatment used in the main text can be seen as a special case of the characteristic function treatment.

*Proof.* The proof contains the following two steps.

**First step:** We show that if $K$ is an absolutely continuous random variable and its density function $p_K$ has support $\mathbb{R}$, then $\tilde{\mathrm{D}}_f(\mathbb{P}_X||\mathbb{P}_Y) = 0 \Leftrightarrow \mathbb{P}_{\tilde{X}} = \mathbb{P}_{\tilde{Y}}$. By Lemma 1, we have $\tilde{X}$ and $\tilde{Y}$ are absolutely continuous and allow probability density functions $p_{\tilde{X}}$ and $p_{\tilde{Y}}$. Since $p_K$ has support $\mathbb{R}$, $p_{\tilde{X}}$ and $p_{\tilde{Y}}$ will also have support $\mathbb{R}$. The $f$-divergence between two absolutely continuous distributions with common support is equal to zero if and only if two distributions are equal (Csiszár, 1967; 1972). We have $\mathrm{D}_f(p_{\tilde{X}}||p_{\tilde{Y}}) = 0 \Leftrightarrow \mathbb{P}_{\tilde{X}} = \mathbb{P}_{\tilde{Y}}$. Therefore,

$$\tilde{\mathrm{D}}_f(\mathbb{P}_X||\mathbb{P}_Y) = 0 \Leftrightarrow \mathbb{P}_{\tilde{X}} = \mathbb{P}_{\tilde{Y}}.$$

**Second step:** We show that if the characteristic function of the spread noise $\phi_K \neq 0$ or $\phi_K = 0$ on at most a countable set then $\mathbb{P}_{\tilde{X}} = \mathbb{P}_{\tilde{Y}} \Leftrightarrow \mathbb{P}_X = \mathbb{P}_Y$.

The characteristic function of a probability measure $\mathbb{P}_X$ is defined as $\phi_X(w) = \int e^{iwx} d\mathbb{P}_X$. Since a probability measure is uniquely determined by its characteristic function (Kallenberg, 2006, Theorem 4.3), we have

$$\mathbb{P}_{\tilde{X}} = \mathbb{P}_{\tilde{Y}} \Leftrightarrow \phi_{\tilde{X}} = \phi_{\tilde{Y}}.$$

Using the fact that the characteristic function of the sum of two random variables is equal to the product of their characteristic functions (Durrett, 2019, Theorem 3.3.2), we can write

$$\phi_{\tilde{X}} = \phi_{\tilde{Y}} \Leftrightarrow \phi_X \phi_K = \phi_Y \phi_K.$$

When $\phi_K \neq 0$, we have $\phi_X \phi_K = \phi_Y \phi_K \Leftrightarrow \phi_X = \phi_Y$.

When $\phi_K = 0$ on at most a countable set $\mathcal{C}$, we show that $\phi_X \phi_K = \phi_Y \phi_K \Leftrightarrow \phi_X = \phi_Y$ still holds. We prove this by contradiction:

We assume there is a point $w_0 \in \mathcal{C}$ where $\phi_X(w_0) \neq \phi_Y(w_0)$. Without loss of generality, we assume $\phi_X(w_0) - \phi_Y(w_0) = \delta > 0$. For points $w_0 + h$ that are not in $\mathcal{C}$, we have $\phi_K(w_0 + h) \neq 0$, so $\phi_X \phi_K = \phi_Y \phi_K$ implies $\phi_X(w_0 + h) - \phi_Y(w_0 + h) = 0$. Since the characteristic function of a distribution is uniform continuous (Durrett, 2019, Theorem 3.3.1), we have $\delta = \phi_X(w_0 + h) - \phi_Y(w_0 + h) \to 0$ when $h \to 0$, which leads to a contradiction (since $\delta$ cannot be zero). Therefore, $\phi_X \phi_K = \phi_Y \phi_K \Leftrightarrow \phi_X = \phi_Y$.

By the uniqueness of the characteristic function (Kallenberg, 2006, Theorem 4.3), we have

$$\phi_X = \phi_Y \Leftrightarrow \mathbb{P}_X = \mathbb{P}_Y.$$

Using the results of the two steps, we can conclude

$$\tilde{\mathrm{D}}_f(\mathbb{P}_X||\mathbb{P}_Y) = 0 \Leftrightarrow \mathbb{P}_{\tilde{X}} = \mathbb{P}_{\tilde{Y}} \Leftrightarrow \mathbb{P}_X = \mathbb{P}_Y.$$

□

## D. Spread Noise Makes Distributions More Similar

The data processing inequality for $f$-divergences (Gerchinovitz et al., 2018) states that $\mathrm{D}_f(\tilde{p}(y)||\tilde{q}(y)) \leq \mathrm{D}_f(p(x)||q(x))$. For completeness, we provide here an elementary proof of this result. We consider the following joint distributions with densities

$$q(y, x) = p(y|x)q(x), \qquad p(y, x) = p(y|x)p(x), \tag{72}$$

whose marginals are the spreaded distributions

$$\tilde{p}(y) = \int p(y|x)p(x)dx, \qquad \tilde{q}(y) = \int p(y|x)q(x)dx. \tag{73}$$

The divergence between the two joint distributions is

$$\mathrm{D}_f(p(y, x)||q(y, x)) = \int q(y, x) f\left(\frac{p(y|x)p(x)}{p(y|x)q(x)}\right) dxdy = \mathrm{D}_f(p(x)||q(x)). \tag{74}$$

More generally, the $f$-divergence between two marginal distributions is no larger than the $f$-divergence between the joint (Zhang et al., 2018). To see this, consider

$$\mathrm{D}_f(p(u,v)||q(u,v)) = \int q(u) \int q(v|u) f\left(\frac{p(u,v)}{q(u,v)}\right) dydu \tag{75}$$

$$\geq \int q(u) f\left(\int q(v|u) \frac{p(u,v)}{q(v|u)q(u)} dv\right) du \tag{76}$$

$$= \int q(u) f\left(\frac{p(u)}{q(u)}\right) du = \mathrm{D}_f(p(u)||q(u))\,. \tag{77}$$

Hence,

$$\tilde{\mathrm{D}}_f(q(x)||p(x)) \equiv \mathrm{D}_f(\tilde{p}(y)||\tilde{q}(y)) \leq \mathrm{D}_f(p(y,x)||q(y,x)) = \mathrm{D}_f(p(x)||q(x))\,. \tag{78}$$

Intuitively, spreading two distributions increases their overlap, reducing the divergence. When the distributions $\mathbb{P}$ and $\mathbb{Q}$ are absolutely continuous and their densities $p$ and $q$ have the same support, the spread $f$-divergence is always a lower bound of $f$-divergence. When the densities do not have the same support or are not well defined, then $\mathrm{D}_f(\mathbb{P}||\mathbb{Q})$ is not well-defined.

## E. Mixture Divergence

We motivated the Spread Divergence between distribution $\mathbb{P}$ and $\mathbb{Q}$ by the requirement to produce a divergence that satisfying $\tilde{\mathrm{D}}(\mathbb{P}||\mathbb{Q}) = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$, where the original $\mathrm{D}(\mathbb{P}||\mathbb{Q})$ does not exist. We briefly discuss the case that $\mathbb{P}$ and $\mathbb{Q}$ are absolutely continuous but their density functions $p$ and $q$ have different supports, so $f$-divergence $\mathrm{D}_f(\mathbb{P}||\mathbb{Q}) = \mathrm{D}(p||q)$ is still not defined. For example, $\mathbb{P}$ and $\mathbb{Q}$ can be two uniform distributions with different supports. We mention here an alternative divergence that also can be used , namely a mixture divergence, and discuss why we focus on the Spread Divergence thereafter. Specifically, we can define a mixture model with density $\tilde{p}(x)$ of the original distribution and a 'noise' distribution with density function $n(x)$:

$$\tilde{p}(x) = \alpha p(x) + (1-\alpha) n(x) \tag{79}$$

for $0 < \alpha < 1$. Provided $n(x)$ is non-zero, then $\tilde{p}(x)$ has support everywhere. Similarly, we can define

$$\tilde{q}(x) = \alpha q(x) + (1-\alpha) n(x). \tag{80}$$

As with the Spread Divergence formulation presented previously, this will usually enable us to define a divergence $\mathrm{D}(\tilde{p}||\tilde{q})$ when $\text{supp}\,(p) \neq \text{supp}\,(q)$. Furthermore, provided the divergence between $\tilde{p}$ and $\tilde{q}$ is zero, then the two distributions $\tilde{p}$ and $\tilde{q}$ match, as do the original distributions $p$ and $q$ since

$$\tilde{p}(x) = \tilde{q}(x) \Leftrightarrow \alpha p(x) + (1-\alpha) n(x) = \alpha q(x) + (1-\alpha) n(x) \Leftrightarrow p(x) = q(x). \tag{81}$$

Therefore, creating a mixture model in this way also allows us to define a divergence between absolutely continuous distributions that otherwise would not have an appropriate divergence[15]. However, in contrast to the Spread Divergence formulation, we cannot use this approach for distributions that are not absolutely continuous, which for many applications of interest cannot be achieved. As a simple example, consider generalised densities $p(x) = \delta\,(x - \mu_p)$, $q(x) = \delta\,(x - \mu_q)$ with

$$\tilde{p}(x) = \alpha\delta\,(x - \mu_p) + (1-\alpha) n(x), \qquad \tilde{q}(x) = \alpha\delta\,(x - \mu_q) + (1-\alpha) n(x). \tag{82}$$

In this case, the divergence $\mathrm{D}(\tilde{p}(x)||\tilde{q}(x))$ is not defined since neither $\tilde{p}(x)$ nor $\tilde{q}(x)$ is a valid probability density. A similar issue arises in training implicit generative models in which a value cannot be feasibly computed for $\tilde{p}$ or $\tilde{q}$; see section(4.2). Hence, for implicit models in, we cannot feasibly assign a value to this mixture divergence. As such it appears to have only limited value in training continuous variable models.

One can combine the spread and the mixture approaches to produce a more general affine divergence

$$\tilde{p}(y) = \alpha \int p(y|x) p(x) dx + (1-\alpha) n(y), \tag{83}$$

[15] This approach is equivalent to the 'anti-freeze' method used in (Furmston & Barber, 2009), which was used to enable EM style training in deterministic transition Markov Decision Processes of discrete states - see also (Barber, 2012).

for spread $p(y|x)$ and (generalised) density $p(x)$. It follows for this case that $\mathrm{D}(\tilde{p}||\tilde{q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$; however, the benefit of the mixture noise over the spread noise is not clear. Our central interest in this work is to train implicit models and, as such, we focus interest only on the first 'spread' term $\int_x p(y|x)p(x)$ in eq(83) and leave the study of the potential additional benefits of including a mixture component $n(y)$ for future work.

## F. Statistical Properties of Maximum Likelihood Estimator

### F.1. Existence of Spread MLE

In some situations there may not exist a Maximum Likelihood Estimator (MLE) for $p(x|\theta)$, but there can exist a MLE for the spread model $p(y|\theta) = \int p(y|x)p(x|\theta)dx$. For example, suppose that $X \sim \mathcal{N}(\mu, \sigma^2)$ $(\mu, 0 < \sigma^2 < \infty)$. So $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$. Assume we only have one data point $x$. Then the log-likelihood function is $L(x; \theta) \propto -\log\sigma - \frac{1}{2\sigma^2}(x-\mu)^2$. Maximising with respect to $\mu$, we have $\mu = x$ and the log-likelihood becomes unbounded as $\sigma^2 \to 0$. In this sense, the MLE for $(\mu, \sigma^2)$ does not exist, see (Casella & Berger, 2002) for more discussions.

In contrast, we can check whether the MLE for $p(y|\theta)$ exists. We assume Gaussian spread noise with fixed variance $\sigma_f^2$. Since we only have one data point $x$, the spread data distribution becomes $p(y|x) = \mathcal{N}(y|x, \sigma_f^2)$, and the model is $p(y|\theta) = \mathcal{N}(y|\mu, \sigma^2 + \sigma_f^2)$. We can sample $N$ points from the spread model, so the spread log likelihood function is (neglecting constants) $L(y_1, \ldots, y_N; \theta) = -\frac{N}{2}\log(\sigma^2 + \sigma_f^2) - \frac{1}{2(\sigma^2+\sigma_f^2)}\sum_{i=1}^N (y_i - \mu)^2$. The MLE solution for $\mu$ is $\mu = \frac{1}{N}\sum_{i=1}^N y_i$; the MLE solution for $\sigma^2$ is $\sigma^2 = \frac{1}{N}\sum_i (y_i - \mu)^2 - \sigma_f^2$, which has bounded spread likelihood value. Note that in the limit of a large number of spread samples $N \to \infty$ , the MLE $\sigma^2 = \frac{1}{N}(y_i - \mu)^2 \to \sigma_f^2$ tends to 0. Throughout, however, the (scaled by $N$) log likelihood remains bounded.

### F.2. Consistency

Consistency of an estimator is an important property that guarantees the validity of the resulting estimate at convergence as the number of data points tends to infinity. In what follows, we outline the sufficient conditions for a consistent MLE estimator, before addressing the question of whether using spread MLE is also consistent and under what conditions.

#### F.2.1. Consistency for MLE

Sufficient conditions for the MLE being consistent and converging to the *global* maximum are given in (Wald, 1949). However, they are usually difficult to check even for some standard distributions. The sufficient conditions for MLE being consistent and converging to a *local* maxima are given in (Cramér, 1999) and are more straight forward to check:

C1. (Identifiable): $p(x|\theta_1) = p(x|\theta_2) \to \theta_1 = \theta_2$.

C2. The parameter space $\Theta$ is an open interval $(\underline{\theta}, \bar{\theta})$, $\Theta : -\infty \leq \underline{\theta} < \theta < \bar{\theta} \leq \infty$.

C3. $p(x|\theta)$ is continuous in $\theta$ and differentiable with respect to $\theta$ for all x.

C4. The set $A = \{x : p_\theta(x) > 0\}$ is independent of $\theta$.

Let $X_1, X_2, \ldots$ be $i.i.d$ with density $p(x|\theta_0)$ $(\theta \in \Theta)$ satisfying conditions C1–C4, then there exists a sequence $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ of local maxima of the likelihood function $L(\theta_0) = \prod_{i=1}^n p(x_i|\theta_0)$ which is consistent:

$$\hat{\theta} \xrightarrow{p} \theta_0 \quad \text{for all } \theta \in \Theta.$$

The proof can be found in (Lehmann, 2004) or (Cramér, 1999).

#### F.2.2. Consistency of spread MLE

We provide the necessary conditions for Spread MLE being consistent.

C1. (Identifiable): $p(x|\theta)$ is identifiable. From section(2.1) it follows immediately that $p(y|\theta_1) = p(y|\theta_2) \to p(x|\theta_1) = p(x|\theta_2) \to \theta_1 = \theta_2$, where the final implication follows from the assumption that $p(x|\theta)$ is identifiable. Hence if $p(x|\theta)$ is identifiable, so is $p(y|\theta)$.

C2. The parameter space $\Theta$ is an open interval $(\underline{\theta}, \bar{\theta})$, $\Theta : -\infty \leq \underline{\theta} < \theta < \bar{\theta} \leq \infty$. This condition is unchanged for $p(y|\theta)$.

C3. On $p(y|\theta)$, we require the same condition on $p(x|\theta)$ as in MLE; $p(y|\theta)$ is continuous in $\theta$ and differentiable with respect to $\theta$ for all $y$.

C4. For spread noise $p(y|x)$ who has full support on $\mathbb{R}^d$ (for example Gaussian noise), $p(y|\theta)$ is greater than zero everywhere and hence the original condition C4 is automatically guaranteed.

The conditions that guarantee consistency for spread MLE are weaker for the spread model $p(y|\theta)$ than for the standard model $p(x|\theta)$, since C4 is automatically satisfied. (Ferguson, 1982) gives an example for which MLE exists but is not consistent by violating condition C4, whereas spread MLE can be used to obtain a consistent estimator.

### F.3. Asymptotic Efficiency

A key desirable property of any estimator is that it is efficient. The Cramer-Rao bound places a lower bound on the variance of any unbiased estimator and an efficient estimator must reach this minimal value in the limit of a large amount of data. Under certain conditions (see below) the Maximum Likelihood Estimator attains this minimal variance value meaning that there is no better estimator possible (in the limit of a large amount of data). This is one of the reasons that the maximum likelihood is a cherished criterion.

#### F.3.1. Asymptotic Efficiency for MLE

Building upon conditions C1-C4, additional conditions on $p(x|\theta)$ are required to show MLE is asymptotical efficient:

C5. For all $x$ in its support, the density $p_\theta(x)$ is three times differentiable with respect to $\theta$ and the third derivative is continuous.

C6. The derivatives of the integral $\int p_\theta(x)dx$ respect to $\theta$ can be obtained by differentiating under the integral sign, that is: $\nabla_\theta \int p_\theta(x)dx = \int \partial_\theta p_\theta(x)dx$.

C7. There exists a positive number $c(\theta_0)$ and a function $M_{\theta_0}(x)$ such that

$$\left|\frac{\partial^3}{\partial\theta^3}\log p_\theta(x)\right| \leq M_{\theta_0}(x) \quad \text{for all } x \in A, |\theta - \theta_0| < c(\theta_0),$$

where $A$ is the support set of $x$ and $\mathbb{E}_{\theta_0}\left[M_{\theta_0}(x)\right] < \infty$.

Let $X_1, ..., X_n$ be $i.i.d$ with density $p_\theta(x)$ and satisfy conditions C1-C7, then any consistent sequence $\hat{\theta} = \hat{\theta}_n\left(X_1, ..., X_n\right)$ of roots of the likelihood equation satisfies

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, F(\theta_0)^{-1}\right), \tag{84}$$

where $F^{-1}(\theta_0)$ is the inverse of Fisher information matrix (also called Cramér-Rao Lower Bound, which is a lower bound on variance of any unbiased estimators ). The conditions and proof can be found in (Lehmann, 2004).

#### F.3.2. Asymptotic Efficiency for MLE

As with MLE above, we require further conditions on $p(y|\theta)$ for ensuring spread MLE is asymptotically efficient:

C5. On $p(y|\theta)$, we require the same condition as applied to $p(x|\theta)$ in the MLE case; for all $y$ in its support, the density $p_\theta(y)$ is three times differentiable with respect to $\theta$ and the third derivative is continuous.

C6. For spread noise $p(y|x)$, which has full support on $\mathbb{R}^d$ (for example Gaussian noise), the support of $y$ is independent of $\theta$. Leibniz's rule[16] allows us to differentiate under the integral: $\nabla_\theta \int p_\theta(y)dy = \int \partial_\theta p_\theta(y)dy$, so this condition is automatically satisfied.

[16]Leibniz's rule tells us: $\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} p(x,\theta)dx = \int_{a(\theta)}^{b(\theta)} \partial_\theta p(x,\theta)dx + p(b(\theta),\theta)\frac{d}{d\theta}b(\theta) - p(a(\theta),\theta)\frac{d}{d\theta}a(\theta)$, so if $a(\theta)$ and $b(\theta)$ are independent of $\theta$, then $\frac{d}{d\theta}\int_a^b p(x,\theta)dx = \int_a^b \partial_\theta p(x,\theta)dx$.

C7. On $p(y|\theta)$, we require the same condition as applied to $p(x|\theta)$ in the MLE case; There exist positive number $c(\theta_0)$ and a function $M_{\theta_0}(y)$ such that

$$\left|\frac{\partial^3}{\partial\theta^3}\log p_\theta(y)\right| \leq M_{\theta_0}(y) \quad \text{for all } y \in A, |\theta - \theta_0| < c(\theta_0),$$

where $A$ is the support set of $y$ and $\mathbb{E}_{\theta_0}\left[M_{\theta_0}(y)\right] < \infty$.

Thus the conditions that guarantee asymptotic efficiency for the spread model $p(y|\theta)$ are weaker than for the standard model $p(x|\theta)$, since C4 and C6 are automatically satisfied.

## G. Spread Divergence for Deterministic Deep Generative Models

Instead of minimising the likelihood, we train an implicit generative model by minimising the Spread Divergence

$$\min_\theta \mathrm{KL}(\tilde{p}(y)||\tilde{p}_\theta(y))\,. \tag{85}$$

For Gaussian noise with fixed diagonal noise $p(y|x) = N(y|x, \sigma^2 I_X)$, we can write

$$\tilde{p}(y) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{N}\left(y|x_n, \sigma^2 I_X\right). \tag{86}$$

and

$$\tilde{p}_\theta(y) = \int p(y|x)p_\theta(x)dx = \int \mathcal{N}\left(y|g_\theta(z), \sigma^2 I_X\right)p(z)dz = \int p_\theta(y|z)p(z)dz. \tag{87}$$

For the Spread Divergence with learned covariance Gaussian noise, which is discussed in section(3.1), we can write

$$p_\psi(y|x) = \mathcal{N}\left(y|x, \Sigma_\psi\right), \qquad \tilde{p}(y) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{N}\left(y|x_n, \Sigma_\psi\right) \tag{88}$$

and Spread Divergence with a learned injective function as discussed in section(3.2)

$$p_\psi(y|x) = \mathcal{N}\left(y|f_\psi(x), \sigma^2 I_X\right), \qquad \tilde{p}(y) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{N}\left(y|f_\psi(x), \sigma^2 I_X\right). \tag{89}$$

According to our general theory,

$$\min_\theta \mathrm{KL}(\tilde{p}(y)||\tilde{p}_\theta(y)) = 0 \quad \Leftrightarrow \quad p(x) = p_\theta(x). \tag{90}$$

Here

$$\mathrm{KL}(\tilde{p}(y)||\tilde{p}_\theta(y)) = -\int \tilde{p}(y)\log\tilde{p}_\theta(y)dy + const. \tag{91}$$

Typically, the integral over $y$ will be intractable and we resort to an unbiased sampled estimate (though see below for Gaussian $q$). Neglecting constants, the KL divergence estimator is

$$\frac{1}{NS}\sum_{n=1}^{N}\sum_{s=1}^{S}\log\tilde{p}_\theta(y_s^n), \tag{92}$$

where $y_s^n$ is a perturbed version of $x_n$. For example $y_s^n \sim \mathcal{N}\left(y_s^n|x_n, \sigma^2 I_X\right)$ for the fixed Gaussian noise case. In most cases of interest, with non-linear $g$, the distribution $\tilde{p}_\theta(y)$ is intractable. We therefore use the variational lower bound

$$\log\tilde{p}_\theta(y) \geq \int q_\phi(z|y)\left(-\log q_\phi(z|y) + \log\left(p_\theta(y|z)p(z)\right)\right)dz. \tag{93}$$

Parameterising the variational distribution as a Gaussian,

$$q_\phi(z|y) = \mathcal{N}\left(z|\mu_\phi(y), \Sigma_\phi(y)\right), \tag{94}$$

we can then use the reparameterization trick (Kingma & Welling, 2013) and write

$$\log \tilde{p}_\theta(y) \geq H(\Sigma_\phi(y)) + \mathbb{E}_{\mathcal{N}(\epsilon|0,I)}\left[\log p_\theta(y|z = h_\phi(y,\epsilon)) + \log p(z = h_\phi(y,\epsilon))\right], \tag{95}$$

where $h_\phi(y,\epsilon) = \mu_\phi(y) + C_\phi(y)\epsilon$ and $H(\Sigma_\phi(y))$ is the entropy of a Gaussian with covariance $\Sigma_\phi(y)$, where $C_\phi(y)$ is the Cholesky decomposition of $\Sigma_\phi(y)$. For fixed covariance Gaussian spread noise in $D$ dimensions, this is (ignoring the constant)

$$\log \tilde{p}_\theta(y) \geq H(\Sigma_\phi(y)) + \mathbb{E}_{\mathcal{N}(\epsilon|0,I)}\left[-\frac{1}{\left(2\sigma^2\right)^{D/2}}\left(y - g_\theta\left(h_\phi(y,\epsilon)\right)\right)^2 + \log p(z = h_\phi(y,\epsilon))\right]. \tag{96}$$

We can integrate the above equation over $y$ to give the bound (ignoring the constant)

$$\begin{aligned}\int \mathcal{N}\left(y|x,\sigma^2 I_X\right)\log \tilde{p}_\theta(y) \geq \mathbb{E}_{\mathcal{N}(y|x,\sigma^2 I_X)}\left[H(\Sigma_\phi(y)) + \mathbb{E}_{\mathcal{N}(\epsilon|0,I)}\left[\log p(z = h_\phi(y,\epsilon))\right]\right] \\ - \frac{1}{\left(2\sigma^2\right)^{D/2}}\mathbb{E}_{\mathcal{N}(\epsilon|0,I)}\left[\mathbb{E}_{\mathcal{N}(y|x,\sigma^2 I_X)}\left[\left(y - g_\theta\left(h_\phi(y,\epsilon)\right)\right)^2\right]\right],\end{aligned} \tag{97}$$

where

$$\begin{aligned}\mathbb{E}_{\mathcal{N}(y|x,\sigma^2 I_X)}\left[\left(y - g_\theta\left(h_\phi(y,\epsilon)\right)\right)^2\right] = \sigma^2 - 2\mathbb{E}_{\mathcal{N}(\epsilon_x|0,I_X)}\left[\epsilon_x g_\theta(h_\phi(x + \sigma\epsilon_x,\epsilon))\right] \\ + \mathbb{E}_{\mathcal{N}(\epsilon_x|0,I_X)}\left[\left(x - g_\theta(h_\phi(x + \sigma\epsilon_x,\epsilon))\right)^2\right].\end{aligned} \tag{98}$$

We notice that the second term is zero, so the final bound for the fixed Gaussian spread KL divergence is (ignoring the constant)

$$\begin{aligned}\int \mathcal{N}\left(y|x,\sigma^2 I_X\right)\log \tilde{p}_\theta(y) \geq \mathbb{E}_{\mathcal{N}(y|x,\sigma^2 I_X)}\left[H(\Sigma_\phi(y)) + \mathbb{E}_{\mathcal{N}(\epsilon|0,I)}\left[\log p(z = h_\phi(y,\epsilon))\right]\right] \\ - \frac{1}{\left(2\sigma^2\right)^{D/2}}\mathbb{E}_{\mathcal{N}(\epsilon_x|0,I_X)}\left[\mathbb{E}_{\mathcal{N}(\epsilon|0,I)}\left[\left(x - g_\theta(h_\phi(x + \sigma\epsilon_x,\epsilon))\right)^2\right]\right].\end{aligned} \tag{99}$$

By analogy, for spread KL divergence with learned variance, the bound is (ignoring the constant)

$$\begin{aligned}\int \mathcal{N}\left(y|x,\Sigma_\psi\right)\log \tilde{p}_\theta(y) \geq \mathbb{E}_{\mathcal{N}(y|x,\Sigma_\psi)}\left[H(\Sigma_\phi(y)) + \mathbb{E}_{\mathcal{N}(\epsilon|0,I)}\left[\log p(z = h_\phi(y,\epsilon))\right]\right] \\ - \mathbb{E}_{\mathcal{N}(\epsilon_x|0,\Sigma_\psi)}\left[\mathbb{E}_{\mathcal{N}(\epsilon|0,I)}\left[\left(x - g_\theta(h_\phi(x + S_\psi\epsilon_x,\epsilon))\right)^T \Sigma_\psi^{-1}\left(x - g_\theta(h_\phi(x + S_\psi\epsilon_x,\epsilon))\right)\right]\right],\end{aligned} \tag{100}$$

where $S_\psi$ is the cholesky decomposition of $\Sigma_\psi$.

For spread KL divergence with a learned injective function, the bound is (ignoring the constant)

$$\begin{aligned}\int \mathcal{N}\left(y|f_\psi(x),\sigma^2 I_X\right)\log \tilde{p}_\theta(y) \geq \mathbb{E}_{\mathcal{N}(y|x,\sigma^2 I_X)}\left[H(\Sigma_\phi(y)) + \mathbb{E}_{\mathcal{N}(\epsilon|0,I)}\left[\log p(z = h_\phi(y,\epsilon))\right]\right] \\ - \frac{1}{\left(2\sigma^2\right)^{D/2}}\mathbb{E}_{\mathcal{N}(\epsilon_x|0,I_X)}\left[\mathbb{E}_{\mathcal{N}(\epsilon|0,I)}\left[\left(f_\psi(x) - f_\psi(g_\theta(h_\phi(x + \sigma\epsilon_x,\epsilon)))\right)^2\right]\right].\end{aligned} \tag{101}$$

The overall procedure is therefore a straightforward modification of the standard VAE method (Kingma & Welling, 2013) with an additional sub-routine for learning the spread online to maximize the divergence:

1. Choose a noise distribution $p(y|x)$.

2. Choose a tractable family for the variational distribution, for example $q_\phi(z|y) = \mathcal{N}(z|\mu_\phi(y), \Sigma_\phi(y))$, and initialise $\phi$.
3. Sample a $y_n$ for each data point (if we're using $S = 1$ samples).
4. If learning the spread noise:
   (a) Draw samples $\epsilon$ to estimate $-\log \tilde{p}_\theta(y_n)$ according to the corresponding bound.
   (b) Do a gradient ascent step in $\psi$.
5. Draw samples $\epsilon$ to estimate $\log \tilde{p}_\theta(y_n)$ according to the corresponding bound.
6. Do a gradient ascent step in $(\theta, \phi)$.
7. Go to 3 and repeat until convergence.

## H. MNIST Experiment

We first scaled the MNIST data to lie in $[0, 1]$. We use Laplace spread noise with $\sigma = 0.3$ and Gaussian spread noise with $\sigma = 0.3$ for the $\delta$-VAE case. Both the encoder and the decoder networks contain 3 feed-forward layers, each layer has 400 units and use ReLu activation functions. The latent dimension is $Z = 64$. The variational inference network $q_\phi(z|y) = \mathcal{N}(z|\mu_\phi(y), \sigma^2_\phi I_Z)$ has a similar structure for the mean network $\mu_\phi(y)$. For fixed spread $\delta$-VAE , learning was done using the Adam (Kingma & Ba, 2014) optimizer with learning rate $5e^{-4}$ for 200 epochs. For $\delta$-VAE with learned spread (learned covariance), we interleave 2 covariance training epochs with 10 model training epochs (using the Adam optimizer with learning rate $5e^{-5}$).

## I. CelebA Experiment

We pre-processed CelebA images by first taking 140x140 centre crops and then resizing to 64x64. Pixel values were then rescaled to lie in $[0, 1]$. For the learned spread we use Gaussian noise with a learned injective function ResNet $f_\psi(\cdot) = I(\cdot) + g_\psi(\cdot)$, where $g_\psi(\cdot)$ is a one layer convolutional neural net with kernel size $3 \times 3$, with stride length 1. We use spectral normalization (Miyato et al., 2018) to satisfy the Lipschitz constraint. That is, we replace the weight matrix $w$ of the convolution kernel by $w_{SN}(w) := c \times w/\sigma(w)$, where $\sigma(w)$ is the spectral norm of $w$ and $c \in (0, 1)$. This guarantees that $f_\psi$ is invertible - see (Behrmann et al., 2018).

The encoder and decoder are 4-layer convolutional neural networks with batch norm (Ioffe & Szegedy, 2015). Both networks use a fully convolutional architecture with 5x5 convolutional filters with stride length 2 in both vertical and horizontal directions, except the last deconvolution layer where we use stride length 1. $\text{Conv}_k$ represents a convolution with $k$ filters and $\text{DeConv}_k$ represents a deconvolution with k filters, BN for the batch normalization (Ioffe & Szegedy, 2015), Relu for the rectified linear units, and $\text{FC}_k$ for the fully connected layer mapping to $\mathbb{R}^k$.

$$
\begin{aligned}
x \in \mathbb{R}^{64\times64\times3} &\to \text{injective} f(\cdot) \in \mathbf{R}^{64\times64\times3} \\
&\to \text{Conv}_{128} \to \text{BN} \to \text{Relu} \\
&\to \text{Conv}_{256} \to \text{BN} \to \text{Relu} \\
&\to \text{Conv}_{512} \to \text{BN} \to \text{Relu} \\
&\to \text{Conv}_{1024} \to \text{BN} \to \text{Relu} \to \text{FC}_{100}
\end{aligned}
$$

$$
\begin{aligned}
z \in \mathbb{R}^{100} &\to \text{FC}_{10\times10\times1024} \\
&\to \text{DeConv}_{512} \to \text{BN} \to \text{Relu} \\
&\to \text{DeConv}_{256} \to \text{BN} \to \text{Relu} \\
&\to \text{DeConv}_{128} \to \text{BN} \to \text{Relu} \to \text{DeConv}_3 \to sigmoid(\cdot) \\
&\to \text{injective} f(\cdot) \in \mathbb{R}^{64\times64\times3}
\end{aligned}
$$

We use batch size 100 and latent dimension $\dim(Z) = 100$ in all CelabA experiments. For the $\delta$-VAE with fixed spread, we use the fixed Gaussian noise with 0 mean and $(0.5)^2 I$ covariance. We train the model for 500 epochs using Adam optimizer with learning rate $1e^{-4}$. We decay the learning rate with scaling factor 0.9 every 100000 iterations.

For the $\delta$-VAE with learned spread we first train a $\delta$-VAE with fixed $f(x) = x$ and fixed Gaussian noise with 0 mean and $(0.5)^2 I$ diagonal covariance for 300 epochs. We decay the learning with scaling factor $0.9$ every $100000$ iterations. We start iterative training by doing one step inner maximisation over the Spread Divergence parameters $\psi$ using Adam optimizer with learning rate $1e^{-5}$ and one step minimization over the model parameter's $(\theta, \phi)$ using Adam optimizer for additional 200 epochs. We can share the first 300 epochs between the two models. When we sample form two models, we first sample from a $100$ dimensional standard Gaussian distribution $z \sim \mathcal{N}(0, I)$ and use the same latent code $z$ to get samples from both $\delta$-VAE with fixed and learned spread, so we can easily compare the sample quality between two models.

(a) Laplace with fixed covariance

(b) Gaussian with fixed covariance

(c) Gaussian with learned covariance

*Figure 8.* Samples from an implicit generative model trained using $\delta$-VAE with (a) Laplace noise with fixed covariance, (a) Gaussian noise with fixed covariance and (c) Gaussian noise with learned covariance.

(a) Fixed spread noise

(b) Learned spread noise

*Figure 9.* Samples from an implicit generative model trained using $\delta$-VAE with (a) fixed and (b) learned spread with injective mean transformation.