

Learning Conditionally Untangled Latent Spaces using Fixed Point Iteration

Victor Enescu
victor.enescu@lip6.fr

Hichem Sahbi
hichem.sahbi@lip6.fr

Sorbonne University
CNRS, LIP6
F-75005, Paris, France

1 Content

- Additional images and results [2](#).
- TSNE visualization in the latent space in section [3](#).
- Proposition justifying the KLD loss to separate the Gaussians in the latent space in section [4](#).
- Convergence of the fixed point iteration in section [5](#).
- Positive definiteness for the covariances, as well as NF^{GD} variant in section [6](#).
- Extra training details for the NF models in section [7](#).

2 Additional Images and results

2.1 ImageNet

2.1.1 ImageNet training details

As NFs are computationally very expensive due to the bijection, in order to train them on ImageNet, we use the latent space of a pretrained VAE [[12](#)], which downsamples RGB images $\mathbf{x}' \in \mathbb{R}^{256 \times 256 \times 3}$ into images of shape $\mathbf{x} \in \mathbb{R}^{32 \times 32 \times 4}$.

2.1.2 ImageNet transformers accuracy

We further evaluate the performance of our NF based augmentations on ImageNet dataset [[3](#)], by fine-tuning very powerful, pretrained transformers from the well known timm [[16](#)] library. We select the most performant transformer with less than 100M parameters for this experiment, and refer to this [page](#) for comparison. It is an EVA02 [[5](#), [6](#)] transformer originally ranked 9 (in the list), with 87.12M parameters, that was pretrained on ImageNet21k [[11](#)], and then fine-tuned on ImageNet. It takes images of size 448×448 pixels, and is called

"eva02_base_patch14_448. mim_in22k_ft_in22k_in1k" in timm. In accordance with the results shown in the web [link](#), we use ImageNet Real [2] labels for the validation, as they fix labeling mistakes present in the original validation set.

In table 1, we compare our transformer fine-tuned using NF^π against the best ones from timm. It can be noted that it reaches the second highest top 1 and top 5 accuracy score, and outperforms models with 1B parameters. It also outperforms models with the same number of parameters by more than 1 accuracy point.

model name	top1-acc	top5-acc	#of Parameters in Millions	image size expressed in X x X pixels
EVA02 LARGE [5, 13]	91.129	98.713	305.08	448
EVA02 base [5, 13] NF^π augmentations (ours)	91.082	98.854	87.12	448
EVA[6]	90.969	98.672	1,014.45	560
EVA02 base [5, 13]	90.896	98.802	87.12	448
CAFormer [17]	90.781	98.860	98.75	384
Beit [1, 4]	90.687	98.753	305.67	512
VOLO [18]	90.614	98.698	296.09	512
Swinv2 [9]	90.407	98.734	87.92	384
ViT [4]	90.211	98.702	86.86	384
CaiT [14]	90.051	98.495	271.22	384
DeiT [15]	89.891	98.602	86.88	384

Table 1: Comparison of NF based augmentations with the best models from timm. It can be noted the NF^π based augmentation (blue) achieve the second highest rank, in term of top1 accuracy and top5 accuracy.

2.1.3 Augmented Images on ImageNet

In this section, we show augmented images obtained with NF^π on ImageNet. The Image on the left is the original one, and the 3 images on the right are augmented using the proposed latent space, learned using NF^π . For the mixup, we follow the principal modes of the data.



Figure 1: Augmentation on class 21.



Figure 2: Augmentation on class 36.



Figure 3: Augmentation on class 52.



Figure 4: Augmentation on class 70.



Figure 5: Augmentation on class 89.



Figure 6: Augmentation on class 107.



Figure 7: Augmentation on class 117.



Figure 8: Augmentation on class 236.



Figure 9: Augmentation on class 258.

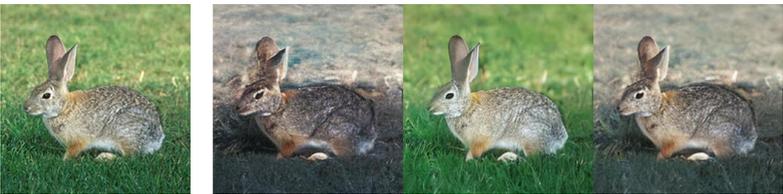


Figure 10: Augmentation on class 330.

2.1.4 Interpolated Images on ImageNet

In this section, we do linear interpolations between images from identical or different classes. We can observe a smooth transition, indicating a continuum in the latent space.



Figure 11: Linear interpolation on class 1.



Figure 12: Linear interpolation on class 24.



Figure 13: Linear interpolation on class 510.



Figure 14: Linear interpolation on class 511.



Figure 15: Linear interpolation on class 562.



Figure 16: Linear interpolation on class 930.



Figure 17: Linear interpolation between random classes.



Figure 18: Linear interpolation between random classes.



Figure 19: Linear interpolation between random classes.

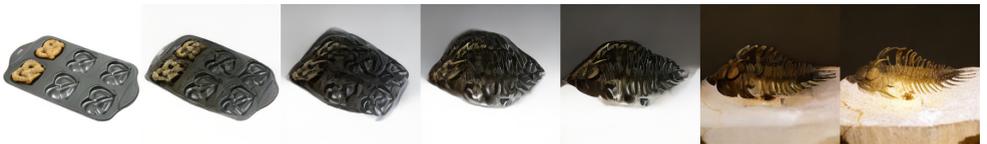


Figure 20: Linear interpolation between random classes.

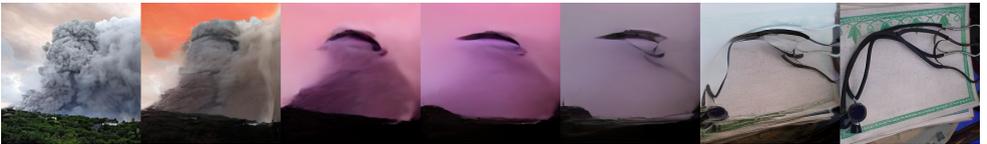


Figure 21: Linear interpolation between random classes.

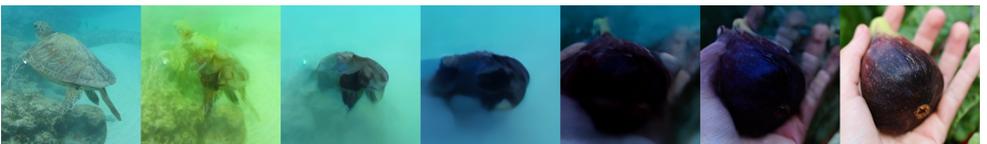


Figure 22: Linear interpolation between random classes.

2.1.5 Sampled Images on ImageNet

In this section, we show class conditional samples from ImageNet, filtered with a classifier.

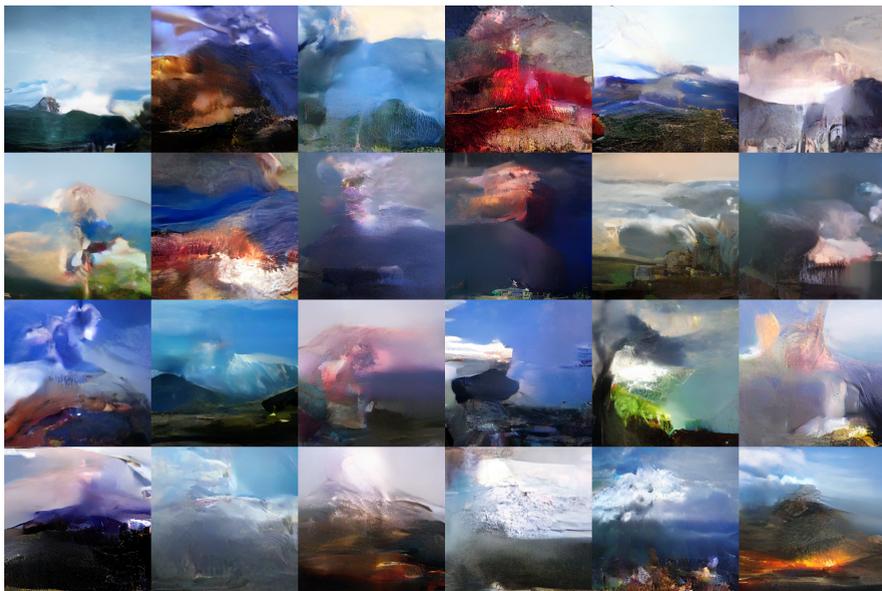


Figure 23: Samples from class 980 (Volcano).



Figure 24: Samples from class 414 (Backpack).

2.2 CIFAR

For CIFAR datasets, we wish to remind that image dimensions are smaller than for ImageNet, so the quality/resolution is not limited by the trained NFs, but by the image dimensions. Figure 25 and 26 show sampled images on respectively CIFAR10 and CIFAR100, and figure 27 shows augmented images on CIFAR10.

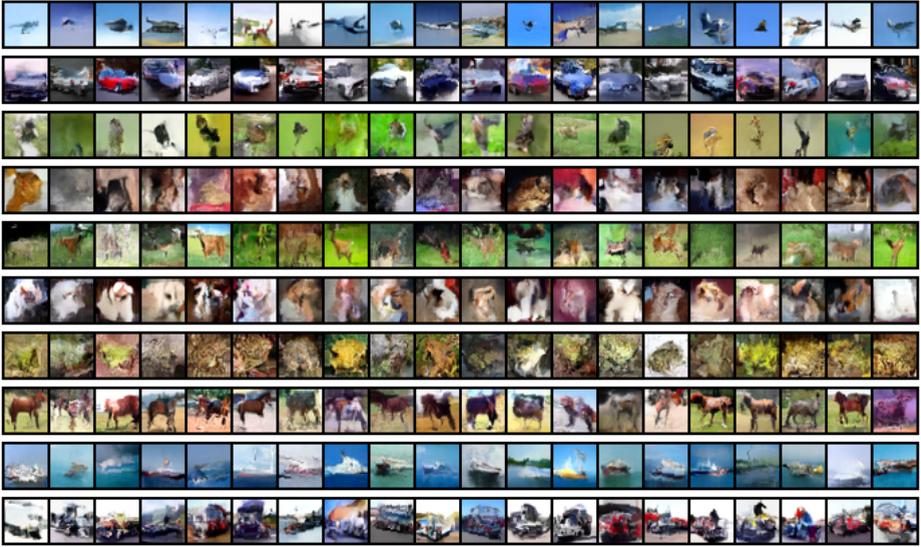


Figure 25: Samples on CIFAR10. From top to bottom, the classes are air, airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck.

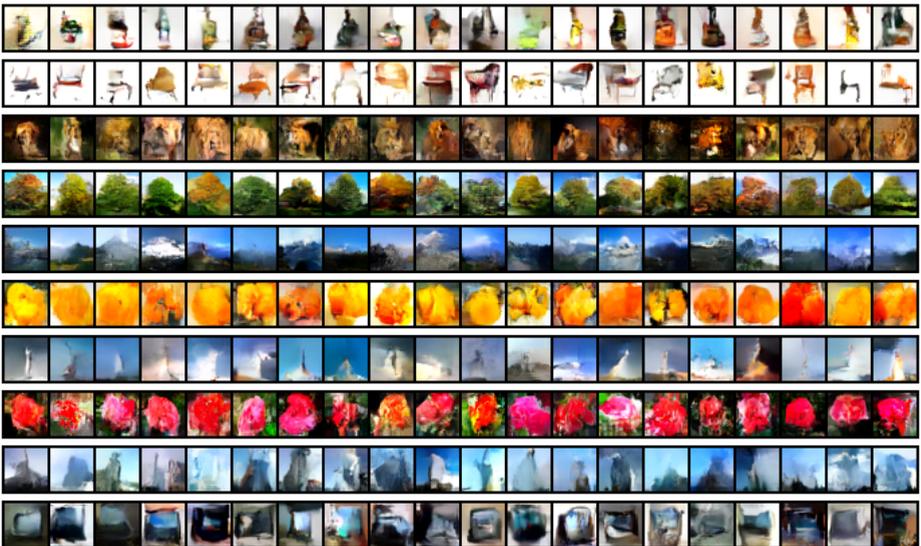


Figure 26: Samples on CIFAR100. From top to bottom, the classes are bottle, chair, lion, maple tree, mountain, orange, rocket, rose, skyscraper, television.



Figure 27: Example of augmented images on CIFAR10. The first row shows the original images, and the 3 below show 3 augmented versions obtained with the NF.

2.2.1 BPD results

In Table 2, we provide conditional bits per dimensions (BPD) results for different λ values on CIFAR10 and CIFAR100.

Dataset	CIFAR10			CIFAR100		
	100	1000	3000	10	100	750
BPD	3.38	3.54	3.73	3.47	3.72	5.29
NF-Acc	91.00	92.11	93.63	32.75	66.70	71.65

Table 2: Bits per dimensions (BPD) along with NF-Acc for different λ values on CIFAR10 and CIFAR100.

3 Latent space

Figures 28 29 30 show the TSNE latent spaces for different λ values on CIFAR100, as well as the variances of a single Gaussian. It can be noted that Gaussians are gradually separated as λ increases, and that variances decrease.

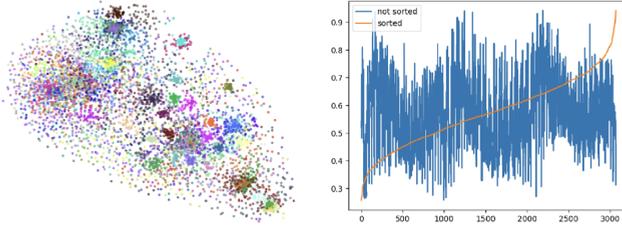


Figure 28: TSNE (left) and sorted variances of an arbitrary Gaussian taken from the dataset (right) for $\lambda = 10$ on CIFAR100. The x axis represents the dimension and the y axis the value of the variance.

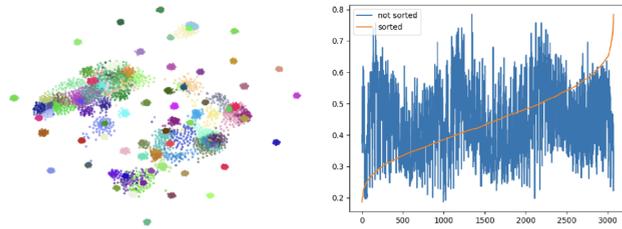


Figure 29: TSNE (left) and sorted variances of an arbitrary Gaussian taken from the dataset (right) for $\lambda = 25$ on CIFAR100. The x axis represents the dimension and the y axis the value of the variance.

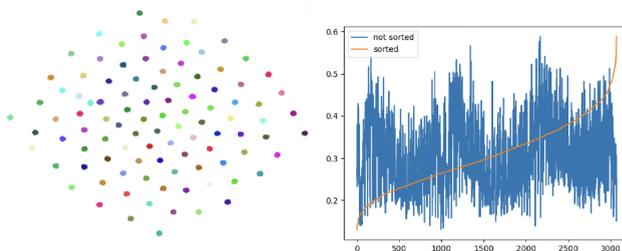


Figure 30: TSNE (left) and sorted variances of an arbitrary Gaussian taken from the dataset (right) for $\lambda = 75$ on CIFAR100. The x axis represents the dimension and the y axis the value of the variance.

4 Proposition cited in section 4.2 of the paper

Let Eq. 11 be the negative log-likelihood of Eq. 2.

$$\mathcal{L}_{NF}(\Theta, \Psi) = \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}} -\log \mathcal{N}(f(\mathbf{x}); \mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) - \log |\det \mathbf{J}_{f(\mathbf{x})}|. \quad (11)$$

Proposition 2 Let $P_{\mathbf{X}}(\cdot | \mathbf{y}_1)$, $P_{\mathbf{X}}(\cdot | \mathbf{y}_2)$ be two probability distributions defined on a compact $\mathcal{X} \subseteq \mathcal{B}^d$. The optimization of Eq. 11 leads to

$$\mathbb{E}[(P_{\mathbf{Z}}(\mathbf{Z} | \mathbf{y}_1) - P_{\mathbf{Z}}(\mathbf{Z} | \mathbf{y}_2))^2] \leq \mathbf{B}, \quad (12)$$

with

$$\mathbf{B} = \mathbb{E}[(P_{\mathbf{X}}(\mathbf{X} | \mathbf{y}_1) - P_{\mathbf{X}}(\mathbf{X} | \mathbf{y}_2))^2] \times \frac{\exp(1)}{\det(\Sigma_{\mathbf{y}^*})}, \quad (13)$$

being $\mathbf{y}^* = \arg \min_{\mathbf{y} \in \mathcal{Y}} \det(\Sigma_{\mathbf{y}})$ and \mathcal{B}^d a zero centered unit ball enclosing data¹ in \mathcal{X} .

Details of the proof are given below. More importantly, the bound in Eq. 13 shows that when optimizing Eq. 11 a good continuum is achieved, but that may result into highly confounded Gaussians, that can make label conditioning erroneous.

Proof 2 Using Eq. 2, one may write

$$\mathbb{E}[(P_{\mathbf{X}}(\mathbf{X} | \mathbf{y}_1) - P_{\mathbf{X}}(\mathbf{X} | \mathbf{y}_2))^2] = \mathbb{E}[(P_{\mathbf{X}}(f(\mathbf{X}) | \mathbf{y}_1) - P_{\mathbf{X}}(f(\mathbf{X}) | \mathbf{y}_2))^2 \cdot \det(\mathbf{J}_{f(\mathbf{X})})^2], \quad (14)$$

here the expectation is w.r.t. the marginal distribution of \mathbf{X} . The class of widely used flows² can be written using quasi-linear mapping as $\mathbf{Z} = f(\mathbf{X}) = \mathbf{W}_{\mathbf{X}}\mathbf{X} + \mathbf{c}_{\mathbf{X}}$, with $(\mathbf{Z} | \mathbf{Y}) \sim \mathcal{N}(\mu_{\mathbf{Y}}, \Sigma_{\mathbf{Y}})$. With this quasi-linear form, $\det(\mathbf{J}_{f(\mathbf{X})}) = \det(\mathbf{W}_{\mathbf{X}})$, and $\exists \mathbf{x}_0 \in \mathcal{X}$ s.t.

$$\frac{\mathbb{E}[(P_{\mathbf{Z}}(\mathbf{Z} | \mathbf{y}_1) - P_{\mathbf{Z}}(\mathbf{Z} | \mathbf{y}_2))^2]}{\mathbb{E}[(P_{\mathbf{X}}(\mathbf{X} | \mathbf{y}_1) - P_{\mathbf{X}}(\mathbf{X} | \mathbf{y}_2))^2]} \leq \frac{1}{\det(\mathbf{W}_{\mathbf{x}_0})^2}, \quad (15)$$

by plugging $f(\mathbf{X})$ in Eq. 11, \mathcal{L}_{NF} becomes

$$\mathbb{E} \left[\frac{1}{2} (\mathbf{W}_{\mathbf{X}}\mathbf{X} + \mathbf{c}_{\mathbf{X}} - \mu_{\mathbf{Y}})' \Sigma_{\mathbf{Y}}^{-1} (\mathbf{W}_{\mathbf{X}}\mathbf{X} + \mathbf{c}_{\mathbf{X}} - \mu_{\mathbf{Y}}) + \frac{1}{2} \log(\det(\Sigma_{\mathbf{Y}})) - \log |\det(\mathbf{W}_{\mathbf{X}})| \right]. \quad (16)$$

For $\mathbf{X} \simeq \mathbf{x}_0$, the stationary solution of \mathcal{L}_{NF} w.r.t. $\mathbf{W}_{\mathbf{x}_0}$ leads to

$$\Sigma_{\mathbf{y}_0}^{-1} (\mathbf{c}_{\mathbf{x}_0} + \mathbf{W}_{\mathbf{x}_0}\mathbf{x}_0 - \mu_{\mathbf{y}_0}) \mathbf{x}'_0 - \frac{\text{sign}(\det(\mathbf{W}_{\mathbf{x}_0}))}{|\det(\mathbf{W}_{\mathbf{x}_0})|} \cdot \text{adj}(\mathbf{W}_{\mathbf{x}_0})' = 0 \quad (17)$$

since

$$\frac{\text{sign}(\det(\mathbf{W}_{\mathbf{x}_0}))}{|\det(\mathbf{W}_{\mathbf{x}_0})|} \cdot \text{adj}(\mathbf{W}_{\mathbf{x}_0}) = \mathbf{W}_{\mathbf{x}_0}^{-1}, \quad (18)$$

and using Eq. 17

$$\mathbf{W}_{\mathbf{x}_0} (\mathbf{x}'_0 \mathbf{x}'_0 + \mathbf{W}_{\mathbf{x}_0}^{-1} (\mathbf{c}_{\mathbf{x}_0} - \mu_{\mathbf{y}_0}) \mathbf{x}'_0) \mathbf{W}'_{\mathbf{x}_0} = \Sigma_{\mathbf{y}_0}, \quad (19)$$

¹This is easily obtainable by rescaling the data in \mathcal{X} .

²including linear mapping, affine and additive coupling layers as well as their composition.

being $\mathbf{x}_0\mathbf{x}'_0$ the autocorrelation matrix of \mathbf{x}_0 . Let $I \in \mathbb{R}^{d \times d}$ be the identity matrix, since

$$\det((\mathbf{x}_0\mathbf{x}'_0 + \mathbf{W}_{\mathbf{x}_0}^{-1}(\mathbf{c}_{\mathbf{x}_0} - \mu_{\mathbf{y}_0})\mathbf{x}'_0) \leq \det(\mathbf{x}_0\mathbf{x}'_0 + I), \quad (20)$$

Eq. 19 leads to

$$\begin{aligned} \frac{1}{\det(\mathbf{W}_{\mathbf{x}_0})^2} &\leq \det(I + \mathbf{x}_0\mathbf{x}'_0) \det(\Sigma_{\mathbf{y}_0})^{-1} \\ &\leq \left(\frac{\text{tr}(\mathbf{x}_0\mathbf{x}'_0 + I)}{d} \right)^d \det(\Sigma_{\mathbf{y}_0})^{-1} \\ &= \left(\frac{\text{tr}(\mathbf{x}_0\mathbf{x}'_0) + \text{tr}(I)}{d} \right)^d \det(\Sigma_{\mathbf{y}_0})^{-1} \\ &= \left(\frac{\|\mathbf{x}_0\|_2^2 + d}{d} \right)^d \det(\Sigma_{\mathbf{y}_0})^{-1} \\ &\leq \left(\frac{1}{d} + 1 \right)^d \det(\Sigma_{\mathbf{y}_0})^{-1} \\ &\leq \lim_{d \rightarrow \infty} \left(\frac{1}{d} + 1 \right)^d \det(\Sigma_{\mathbf{y}_0})^{-1} \\ &\leq \exp(1) \det(\Sigma_{\mathbf{y}^*})^{-1}, \end{aligned}$$

which also results from $\mathbf{x}_0 \in \mathcal{B}^d$. By plugging this upper bound in Eq. 15, we complete the proof. ■

5 Convergence of fixed point iteration

Proposition 3 Let $\|\cdot\|_1$ denote the entrywise L_1 -norm. Provided that $\Sigma_{\mathbf{y}}^{(t)} = \Sigma_{\mathbf{y}'}^{(t)}, \forall \mathbf{y}, \mathbf{y}' \in \{1, \dots, K\}, \forall t \in \{1, \dots, T\}$, and provided that the following inequality holds,

$$\lambda < K(K-1) \left(\sum_{\mathbf{y}, \mathbf{y}' \neq \mathbf{y}} (K_{\mathbf{y}\mathbf{y}'} + K_{\mathbf{y}'\mathbf{y}}) \right)^{-1}, \quad (21)$$

the optimization problem (4) admits a solution $\bar{\mu}$ as the limit of $\mu^{(t)} = \{\mu_{\mathbf{y}}^{(t)}\}_{\mathbf{y}}$ with

$$\mu_{\mathbf{y}}^{(t)} = \frac{1}{N_{\mathbf{y}}} \sum_{i=1}^{N_{\mathbf{y}}} f(\mathbf{x}_i) + \frac{\lambda}{K(K-1)} \sum_{\mathbf{y}' \neq \mathbf{y}}^K \mathbf{G}_{\mathbf{y}\mathbf{y}'}^{(t-1)} (\mu_{\mathbf{y}}^{(t-1)} - \mu_{\mathbf{y}'}^{(t-1)}), \quad (22)$$

being

$$\mathbf{G}_{\mathbf{y}\mathbf{y}'}^{(t-1)} = \Sigma_{\mathbf{y}}^{(t-1)} (K_{\mathbf{y}\mathbf{y}'} (\Sigma_{\mathbf{y}'}^{(t-1)})^{-1} + K_{\mathbf{y}'\mathbf{y}} (\Sigma_{\mathbf{y}}^{(t-1)})^{-1}).$$

Furthermore, $\mu^{(t)}$ — with $t \in \{1, \dots, T\}$ — satisfy the convergence property

$$\|\mu^{(t)} - \bar{\mu}\|_1 \leq L^t \|\mu^{(0)} - \bar{\mu}\|_1,$$

with $\|\mu^{(t)} - \mu^{(t-1)}\|_1 = \sum_{\mathbf{y}} \|\mu_{\mathbf{y}}^{(t)} - \mu_{\mathbf{y}}^{(t-1)}\|_1$ and $L = \frac{\lambda}{K(K-1)} \sum_{\mathbf{y}, \mathbf{y}' \neq \mathbf{y}} (K_{\mathbf{y}\mathbf{y}'} + K_{\mathbf{y}'\mathbf{y}})$.

Proof 3 (Sketch of the Proof) The necessary condition of the fixed-point relation in Eq. 22 results from $\frac{\partial \mathcal{L}}{\partial \mu} = 0$ (details about derivatives are omitted in the proof). We will now prove that the function in Eq. 22 is L -Lipschitzian, with $L = \frac{\lambda}{K(K-1)} \sum_{\mathbf{y}, \mathbf{y}' \neq \mathbf{y}} (K_{\mathbf{y}\mathbf{y}'} + K_{\mathbf{y}'\mathbf{y}})$. Given two vectors $\mu_{\mathbf{y}}^{(t)}, \mu_{\mathbf{y}'}^{(t-1)}$, we have $\|\mu_{\mathbf{y}}^{(t)} - \mu_{\mathbf{y}}^{(t-1)}\|_1 = (*)$, with

$$\begin{aligned} (*) &= \frac{\lambda}{K(K-1)} \left\| \sum_{\mathbf{y}' \neq \mathbf{y}} \mathbf{G}_{\mathbf{y}\mathbf{y}'}^{(t-1)} (\mu_{\mathbf{y}}^{(t-1)} - \mu_{\mathbf{y}'}^{(t-1)}) - \mathbf{G}_{\mathbf{y}\mathbf{y}'}^{(t-2)} (\mu_{\mathbf{y}}^{(t-2)} - \mu_{\mathbf{y}'}^{(t-2)}) \right\|_1 \\ &= \frac{\lambda}{K(K-1)} \left\| \sum_{\mathbf{y}' \neq \mathbf{y}} \mathbf{G}_{\mathbf{y}\mathbf{y}'}^{(t-1)} (\mu_{\mathbf{y}}^{(t-1)} - \mu_{\mathbf{y}'}^{(t-1)}) - (\mathbf{G}_{\mathbf{y}\mathbf{y}'}^{(t-1)} - \mathbf{G}_{\mathbf{y}\mathbf{y}'}^{(t-2)}) (\mu_{\mathbf{y}}^{(t-2)} - \mu_{\mathbf{y}'}^{(t-2)}) \right\|_1 \\ &\leq \frac{\lambda}{K(K-1)} \sum_{\mathbf{y}' \neq \mathbf{y}} \|\mathbf{G}_{\mathbf{y}\mathbf{y}'}^{(t-1)}\|_1 \|\mu_{\mathbf{y}}^{(t-1)} - \mu_{\mathbf{y}'}^{(t-2)}\|_1 + \sum_{\mathbf{y}' \neq \mathbf{y}} \|\mathbf{G}_{\mathbf{y}\mathbf{y}'}^{(t-1)}\|_1 \|\mu_{\mathbf{y}'}^{(t-1)} - \mu_{\mathbf{y}'}^{(t-2)}\|_1 \\ &\leq \frac{\lambda}{K(K-1)} \left[\sum_{\mathbf{y}' \neq \mathbf{y}} \|\mathbf{G}_{\mathbf{y}\mathbf{y}'}^{(t-1)}\|_1 \right] \times \sum_{\mathbf{y}'} \|\mu_{\mathbf{y}}^{(t-1)} - \mu_{\mathbf{y}'}^{(t-2)}\|_1, \end{aligned} \quad (23)$$

hence,

$$\|\mu^{(t)} - \mu^{(t-1)}\|_1 \leq L \|\mu^{(t-1)} - \mu^{(t-2)}\|_1 \quad (24)$$

with $L = \frac{\lambda}{K(K-1)} \sum_{\mathbf{y}, \mathbf{y}' \neq \mathbf{y}} (K_{\mathbf{y}\mathbf{y}'} + K_{\mathbf{y}'\mathbf{y}})$.

Eq. 21 shows that when the KLD loss is used (i.e., $\lambda \neq 0$), different values of λ lead to stable (convergent) training. ■

6 Positive Definiteness of the Covariances

NF^{GD} baseline In order to constrain the covariance matrix to remain positive definite in the original implementation based on gradient descent (NF^{GD}), we consider a reparametrization in Eq. 4 — particularly the covariance matrices — as $\Sigma_{\mathbf{y}} = \psi(\hat{\Sigma}_{\mathbf{y}})$ for some $\hat{\Sigma}_{\mathbf{y}} \in \mathbb{R}^{d \times d}$ with ψ applied entrywise, and this allows free settings of $\{\hat{\Sigma}_{\mathbf{y}}\}_{\mathbf{y}}$ during optimization while guaranteeing the positive definiteness of $\Sigma_{\mathbf{y}}$. During backpropagation, the gradient of the loss \mathcal{L} (now w.r.t. $\hat{\Sigma}_{\mathbf{y}}$) is updated using the chain rule as

$$\frac{\partial \mathcal{L}}{\partial \text{vec}(\hat{\Sigma}_{\mathbf{y}})} = \mathbf{J}_{\psi} \cdot \frac{\partial \mathcal{L}}{\partial \text{vec}(\Sigma_{\mathbf{y}})}, \quad (25)$$

here $\text{vec}(\Sigma_{\mathbf{y}})$ is a columnwise vectorization of $\Sigma_{\mathbf{y}}$ and \mathbf{J}_{ψ} is a diagonal Jacobian whose i^{th} diagonal element equates $\psi'([\hat{\Sigma}_{\mathbf{y}}]_{ii})$. In practice, $\psi(\cdot) = a(1 + \exp\{-\beta(\cdot)\})^{-1} + c$ with a , b and c being positive values that respectively control the amplitude (scale) and the slope (smoothness) as well as the shift of the reparametrization ψ . Besides, $\frac{a+c}{c}$ controls the conditioning of the trained covariance matrices and thereby the shape of the underlying Gaussians in the latent space.

NF ^{π} (ours) concerning the fixed point iteration contribution (NF ^{π}), starting from equations 6 and 7, it can be shown that the term A (see equation 26) is symmetric positive definite (simply noted as PD), since (i) the initial covariances $\Sigma_{\mathbf{y}}^{(0)}$ are initialized to be PD, (ii) the inverses of PD matrices are PD, (iii) vector outer products $(f(\mathbf{x}_i) - \mu_{\mathbf{y}}^{(t-1)})(f(\mathbf{x}_i) - \mu_{\mathbf{y}}^{(t-1)})^{\top}$ are PD, (iv) the left and right product around a PD matrix with another PD matrix is also PD, and (v) the sum of PD matrices is PD. It should also be noted that the term $(\Sigma_{\mathbf{y}}^{(t-1)})^{-1}(1 + \frac{\lambda}{K(K-1)} \sum_{\mathbf{y}' \neq \mathbf{y}} (K_{\mathbf{y}\mathbf{y}'} - K_{\mathbf{y}'\mathbf{y}}))$ is PD despite the subtraction, as λ never reached values that were too high in the experiments.

Similarly for B , the expression is PD since: (i) the initial covariances are initialized to be PD, (ii) the inverses of PD matrices are PD, (iii) vector outer products are PD, (iv) the left and right product around a PD matrix with another PD matrix is also PD, and (v) the sum of PD matrices is PD.

$$\begin{aligned} A &= (\Sigma_{\mathbf{y}}^{(t-1)})^{-1} \left(1 + \frac{\lambda}{K(K-1)} \sum_{\mathbf{y}' \neq \mathbf{y}} (K_{\mathbf{y}\mathbf{y}'} - K_{\mathbf{y}'\mathbf{y}}) \right) \\ &\quad + \frac{\lambda}{K(K-1)} (\Sigma_{\mathbf{y}}^{(t-1)})^{-1} \sum_{\mathbf{y}' \neq \mathbf{y}} K_{\mathbf{y}'\mathbf{y}} \left[(\mu_{\mathbf{y}}^{(t-1)} - \mu_{\mathbf{y}'}^{(t-1)})(\mu_{\mathbf{y}}^{(t-1)} - \mu_{\mathbf{y}'}^{(t-1)})^{\top} + \Sigma_{\mathbf{y}'}^{(t-1)} \right] (\Sigma_{\mathbf{y}}^{(t-1)})^{-1} \\ B &= \frac{1}{N_{\mathbf{y}}} \sum_{i=1}^{N_{\mathbf{y}}} (f(\mathbf{x}_i) - \mu_{\mathbf{y}}^{(t-1)})(f(\mathbf{x}_i) - \mu_{\mathbf{y}}^{(t-1)})^{\top} + \frac{\lambda}{K(K-1)} \sum_{\mathbf{y}' \neq \mathbf{y}} K_{\mathbf{y}\mathbf{y}'} \Sigma_{\mathbf{y}}^{(t-1)} (\Sigma_{\mathbf{y}'}^{(t-1)})^{-1} \Sigma_{\mathbf{y}}^{(t-1)}. \end{aligned} \quad (26)$$

Finally, it can be shown the solution in equation 6 rewritten below, is positive definite, since the matrices A and B are positive definite, and the geometric mean of PD matrices is also PD [7, 10].

$$\Sigma_{\mathbf{y}}^{(t)} = A^{-1/2} (A^{1/2} B A^{1/2})^{1/2} A^{-1/2}, \quad (27)$$

7 Training Details

Dataset	bits	L	F	RBs per Level	Batch Size	Epochs	GPU hours	Params (M)
CIFAR10	8	3	8	[8, 4, 2]	512	500	14	42.9
CIFAR100	8	3	8	[8, 4, 2]	512	500	16	43.7
ImageNet	VAE latent space [12]	3	8	[8, 4, 2]	1792	80	350	47

Table 3: NF models are built using L levels, where each contains F steps of flows. Each step of flow is made of N_i residual blocks (RBs) [8] with 128 hidden channels, where $i = 1$ corresponds to the level index. Experiments were run on a single NVIDIA A100 gpu for CIFAR datasets and 7 for ImageNet, the GPU hours are the summed total hours.

7.1 Learning rate policies for NF^{GD}

Concerning the learning rate policies used for the baselines (NF^{GD}) in table 1, we have:

Linear : The learning rate linearly decreases from $3 \cdot 10^{-2}$ to $1 \cdot 10^{-3}$ over 100 epochs.

Sublinear v1 The learning rate geometrically decreases from 10^{-1} to 10^{-2} over 10 epochs, and then decreases from 10^{-2} to 10^{-3} over 90 epochs, for a total of 100 epochs.

Sublinear v2 Follows a schedule of the form

$\gamma_{current} = \gamma_{final} + (\gamma_{initial} - \gamma_{final}) * (1 - \exp(-\alpha * (epoch_{total} - epoch_{current})))$ where $\gamma_{initial}$ and γ_{final} are the initial and final learning rates, respectively equal to $5 \cdot 10^{-2}$ and 10^{-3} , $\alpha = 0.03$ and the total number of epochs $epoch_{total}$ is equal to 100.

Superlinear v1 The learning rate geometrically decreases from $2 \cdot 10^{-2}$ to $1.7 \cdot 10^{-2}$ over 90 epochs, and then geometrically decreases to 10^{-2} over 10 epochs, for a total of 100 epochs.

Superlinear v2 Follows a schedule of the form

$\gamma_{current} = \gamma_{initial} + (\gamma_{final} - \gamma_{initial}) * (1 - \exp(-\alpha * epoch_{current}))$ where $\gamma_{initial}$ and γ_{final} are the initial and final learning rates, respectively equal to $3 \cdot 10^{-2}$ and $3 \cdot 10^{-3}$, $\alpha = 0.03$, and the total number of epochs $epoch_{total}$ is equal to 100.

References

- [1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [2] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [5] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023.
- [6] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.
- [7] Luyining Gan and Sejong Kim. Revisit on spectral geometric mean. *Linear and Multilinear Algebra*, 72(6):944–955, 2024.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [9] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.
- [10] Wieslaw Pusz and S Lech Woronowicz. Functional calculus for sesquilinear forms and the purification map. *Reports on Mathematical Physics*, 8(2):159–170, 1975.
- [11] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lih Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.
- [12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [13] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

-
- [14] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021.
- [15] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European conference on computer vision*, pages 516–533. Springer, 2022.
- [16] Ross Wightman. Pytorch image models. <https://github.com/rwightman/pytorch-image-models>, 2019.
- [17] Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. Metaformer baselines for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [18] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):6575–6586, 2022.