Whitening Convergence Rate of Affine Coupling Flows

Anonymous Author(s)
Affiliation
Address
email

Abstract

Affine coupling flows (RealNVPs) are a popular normalizing flow architecture that works surprisingly well in practice. This calls for theoretical understanding. Existing work shows that such flows *weakly* converge to arbitrary data distributions [1]. However, they make no statement about the stricter convergence criterion used in practice, the maximum likelihood loss. For the first time, we make a quantitative statement about this kind of convergence: We prove that affine coupling flows perform whitening of the data distribution (i.e. diagonalize the covariance matrix) and derive corresponding convergence bounds that show a linear convergence rate in the depth of the flow. Numerical experiments demonstrate the implications of our theory and point at open questions.

1 Introduction

Normalizing flows [2, 3] are among the most promising approaches to *generative* machine learning and have already demonstrated convincing performance in a wide variety of practical applications, ranging from image analysis [4, 5, 6, 7, 8] to astrophysics [9], mechanical engineering [10], computational biology [11] and medicine [12]. As the name suggests, normalizing flows represent complex data distributions as bijective transformations (also known as flows or *push-forwards*) of standard normal or other well-understood distributions.

In this paper, we focus on a theoretical underpinning of affine coupling flows (RealNVPs [5]), a particularly effective realization of normalizing flows in terms of invertible neural networks. All of the above mentioned applications are actually implemented using variants of RealNVP. Their central building blocks are *affine coupling layers*, which decompose the space into two subspaces called *active* and *passive* subspace. Only the active dimensions are transformed conditioned on the passive dimensions, using affine mappings that are computationally easy to invert. In order to vary the assignment of dimensions to the active and passive subspaces, coupling layers are combined with preceding orthonormal transformation layers into *affine coupling blocks*. These blocks are arranged into deep networks such that the orthonormal transformations are sampled uniformly at random from the orthogonal matrices and the affine layers are trained with the maximum likelihood objective, see Equation (2). Upon convergence of the training, the sequence of coupling blocks gradually transforms the probability density that generated the given training data, into a standard normal distribution (of the same dimensionality) and vice versa.

Since the resulting normalizing flows deviate significantly from *optimal* transport flows [13] and the bulk of the mathematical literature is focusing on optimal transport, an analysis tailored to the RealNVP architecture is lacking. In a landmark paper, [1] proved that sufficiently large affine coupling flows weakly converge to arbitrary data densities. The notion of weak convergence is critical here, as *it does not imply convergence in maximum likelihood* [14, Remark 3]. Maximum likelihood (or, equivalently, the Kullback-Leibler divergence) is the loss that is actually used in practice. It can be used for gradient descent and it guarantees not only convergence in samples ("$x \sim q(x) \to x \sim p(x)$") but also in density estimates ("$q(x) \to p(x)$"). It is strong in the sense

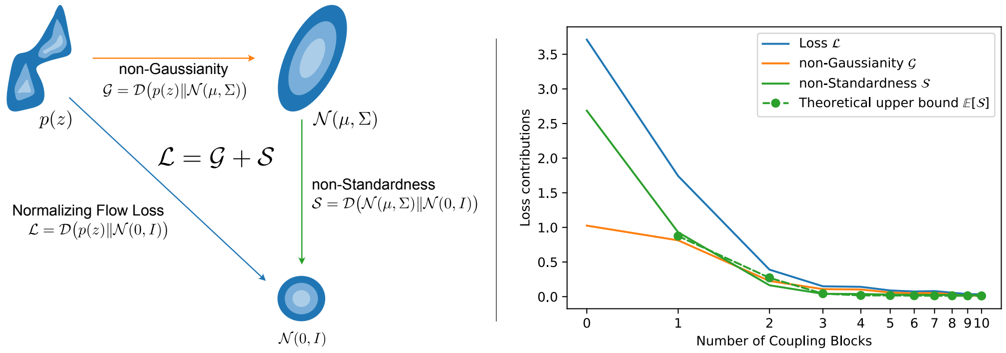


Figure 1: *(Left)* The Maximum Likelihood Loss $\mathcal{L}$ (blue) can be split into the *non-Gaussianity* $\mathcal{G}$ (orange) [17] and the *non-Standardness* $\mathcal{S}$ (green) of the latent code $z = f_\theta(x)$: $\mathcal{L} = \mathcal{G} + \mathcal{S}$ (Proposition 1). For the latter, we give explicit guarantees as one more coupling block is added in Theorems 1 and 2 and show a global convergence rate in Theorem 3. *(Right)* Typical fit of EMNIST digits by a standard affine coupling flow for various depths. Our theory (Theorem 1) upper bounds the average $\mathcal{S}$ for $L + 1$ coupling blocks given a trained model with L coupling blocks (dotted green). We observe in this experiment that our bound is predictive for how much end-to-end training reduces $\mathcal{S}$.

that the square root of the KL-divergence upper bounds (up to a factor 2) the total variation metric, and hence also the Wasserstein metric if the underlying space is bounded [15]. Moreover, convergence under the KL-divergence implies weak convergence which is fundamental for robust statistics [16].

We take a first step towards showing that affine coupling blocks also converge in terms of maximum likelihood. To the best of our knowledge, our paper presents for the first time a quantitative convergence analysis of normalizing flows realized by RealNVPs based on this strong notion of convergence.

Specifically, we make the following contributions towards this goal:

- We utilize that the loss of a normalizing flow can be decomposed into two parts (Figure 1): The divergence to the nearest Gaussian (*non-Gaussianity*) plus the divergence of that Gaussian to the standard normal (*non-Standardness*).
- The contribution of a single coupling layer on the non-Standardness is analyzed in terms of matrix operations (Schur complement and scaling).
- Explicit bounds for the non-Standardness after a single coupling block in expectation over all orthonormal transformations are derived.
- We use these results to prove that a sequence of coupling blocks whitens the data covariance and to derive linear convergence rates for this process.

We confirm our theoretical findings experimentally and identify directions for further improvement.

2 Related work

Analyzing which distributions affine coupling flows can approximate is an active area of research. [1] showed it is possible to approximate arbitrary invertible functions using single-active dimension affine couplings. From this, they conclude that such affine coupling flows *weakly* converge to any probability density. [14] require only three affine coupling blocks to show convergence to arbitrary densities in Wasserstein distance. Both universality results, however, require that the couplings become ill-conditioned (i.e. the learnt functions become increasingly discontinuous as the error decreases, whereas in practice one observes that functions remain smooth). Also, they consider only a finite subspace of the data space. Even more importantly, the convergence criterion employed in their proofs (weak convergence resp. convergence under Wasserstein metric) is critical: [14] remarked that those criteria do not imply convergence in the loss that is employed in practice, the Kullback-Leibler divergence (equivalent to maximum likelihood loss). An arbitrarily small distance in any of the above metrics can even result in an infinite KL divergence. In contrast to previous work on affine coupling flows, we work directly on the KL divergence. We decompose the KL in two contributions and show the flow's convergence for one of the parts.

Regarding when ill-conditioned flows need to arise to fit a distribution, [18] showed that well-conditioned affine couplings can approximate log-concave padded distributions, again in terms of Wasserstein distance. Lipschitz flows on the other hand cannot model arbitrary tail behavior, but this can be fixed by adapting the latent distribution [19].

For other kinds of normalizing flows, similar universality results exist: [20] for weak convergence of sum-of-squares polynomial coupling flows, extended by [21] to the stronger total variation convergence, which also does not imply convergence in KL divergence in general; [22] for zero-padded affine coupling flows in weak convergence; [23, 24] for Neural ODEs in invertible function approximation and weak convergence.

Closely related to our work, [14, Theorem 2] shows that 48 linear affine coupling blocks are sufficient to represent any invertible linear function $Ax + b$ with $\det(A) > 0$. This also allows mapping any Gaussian distribution $\mathcal{N}(m, \Sigma)$ to the standard normal $\mathcal{N}(0, I)$. We put this statement into context in terms of the KL divergence: The loss is exactly composed of the divergence to the nearest Gaussian and of that Gaussian to the standard normal. We then make strong statements about the convergence of the latter, concluding that for typical flows a smaller number of layers is required for accurate approximation than predicted by [14].

3 Affine coupling flow fundamentals

The idea of normalizing flows is to learn an invertible function $f_\theta(x)$ that maps data from some unknown distribution $p(x)$ given by samples to *latent variables* $z = f_\theta(x)$ so that z follow a simple distribution, typically the standard normal. f_θ then yields a model for the data distribution via the change of variables formula (e.g. [5]):

$$q(x) = \mathcal{N}(f_\theta(x); 0, I) |\det J|, \tag{1}$$

where $J = \nabla f_\theta(x)$ is the Jacobian of $f_\theta(x)$. We can train a normalizing flow via the maximum likelihood loss, which is equivalent to minimizing the Kullback-Leibler divergence between the distribution of the latent code $q(z)$, as given by $z = f_\theta(x)$ when $x \sim p(x)$, and the standard normal:

$$L = \mathcal{D}_{\text{KL}}(q(z) || \mathcal{N}(0, I)). \tag{2}$$

The invertible architecture that makes up f_θ has to (1) be computationally easy to invert, (2) be able to represent complex transformations, and (3) have a tractable Jacobian determinant [9]. Building such an architecture is an active area of research, see e.g. [2]. In this work, we focus on a typical variant of affine coupling flows, based on the simple and popular RealNVP architecture [5]. It is a deep architecture that consists of several blocks, each containing a rotation layer and an affine coupling layer:

$$f_{\text{block}}(x) = (f_{\text{cpl}} \circ f_{\text{rot}})(x). \tag{3}$$

The affine coupling f_{cpl} splits an incoming vector in two parts along the coordinate axis: The first part, which we call *passive*, is transformed by a global affine transformation. The second part, which we call *active*, is modified as a function of the passive dimensions:

$$f_{\text{cpl}}(x_0) = f_{\text{cpl}} \begin{pmatrix} p_0 \\ a_0 \end{pmatrix} = \begin{pmatrix} r \odot p_0 + u \\ s(p_0) \odot a_0 + t(p_0) \end{pmatrix} =: \begin{pmatrix} p_1 \\ a_1 \end{pmatrix}. \tag{4}$$

Here, $s : \mathbb{R}^{D/2} \to \mathbb{R}_+^{D/2}$ and $t : \mathbb{R}^{D/2} \to \mathbb{R}^{D/2}$ are arbitrary functions, that are typically represented by neural networks, and $r \in \mathbb{R}_+^{D/2}$ and $u \in \mathbb{R}^{D/2}$ are vector-valued parameters. $\odot$ denotes the element-wise product. As the passive dimensions are transformed by f_{cpl} in a simple manner, the above function is trivial to invert: Call $x_1 = (p_1; a_1)$ the output of the layer, then solving for $x_0 = (p_0; a_0)$ is trivial as $p_0 = (p_1 - u) \oslash r$ ($\oslash$ is element-wise division). Use p_0 to evaluate $s(p_0), t(p_0)$ and recover $a_0 = (a_1 - t(p_0)) \oslash s(p_0)$.

If we were to concatenate several RealNVP layers, the entire network would never change the passive dimensions apart from an element-wise affine transformation. Here, the rotation layers $f_{\text{rot}}(x) = Qx$

come into play. They multiply an orthogonal matrix Q to the data, changing which subspaces are passive resp. active. This matrix is typically fixed at random at initialization and then left unchanged during training. A rotation layer is trivial to invert by transposing the matrix: $f_{\text{rot}}^{-1}(x) = Q^{\mathrm{T}}x$.

Note that the affine transformation of the passive dimensions is a modification from the original formulation of RealNVP in [5]. However, it adds very little computational overhead while simplifying our theoretical analysis greatly. Also, it is used in practice like in the popular INN framework `FrEIA`.

4 Affine coupling layers as whitening transformation

The central mathematical question we answer in this work is the following: How can a deep affine coupling flow *whiten* the data? As the latent distribution is a standard normal, whitening is a necessary condition for the flow to converge. This is a direct property of the loss:

Proposition 1 (Pythagorean Identity, Proof in Appendix B.1)**.** *Given data with distribution $p(x)$ with mean m and covariance Σ. Then, the Kullback-Leibler divergence to a standard normal distribution decomposes as follows:*

$$\mathcal{D}_{KL}(p(x)||\mathcal{N}(0,I)) = \underbrace{\mathcal{D}_{KL}(p(x)||\mathcal{N}(m,\Sigma))}_{\text{non-Gaussianity } \mathcal{G}(p)} + \underbrace{\mathcal{D}_{KL}(\mathcal{N}(m,\Sigma)||\mathcal{N}(0,I))}_{\text{non-Standardness } \mathcal{S}(p)} \tag{5}$$

and the non-Standardness again decomposes:

$$\mathcal{S}(p) = \underbrace{\mathcal{D}_{KL}(\mathcal{N}(m,\Sigma)||\mathcal{N}(m,\mathrm{Diag}(\Sigma)))}_{\text{Correlation } \mathcal{C}(p)} + \underbrace{\mathcal{D}_{KL}(\mathcal{N}(m,\mathrm{Diag}(\Sigma))||\mathcal{N}(0,I))}_{\text{Diagonal non-Standardness}}. \tag{6}$$

This splits the transport from the data distribution to the latent standard normal into three parts: (1) From the data to the nearest Gaussian distribution $\mathcal{N}(m,\Sigma)$, measured by $\mathcal{G}$. (2) From that nearest Gaussian to the corresponding uncorrelated Gaussian $\mathcal{N}(m,\mathrm{Diag}(\Sigma))$, measured $\mathcal{C}$. (3) From the uncorrelated Gaussian to standard normal.

We do not make explicit use of the fact that the *non-Standardness* can again be decomposed, but we show it nevertheless to relate our result to the literature: The Pythagorean identity $\mathcal{D}_{\mathrm{KL}}(p(x)||\mathcal{N}(m,\mathrm{Diag}(\Sigma))) = \mathcal{G}(p) + \mathcal{C}(p)$ has been shown before by [17, Section 2.3]. Both their and our result are specific applications of the general [25, Theorem 3.8] from information geometry. Our proof is given in Appendix B.1.

Proposition 1 is visualized in Figure 1. In an experiment, we fit a set of affine coupling flows of increasing depths to the EMNIST digit dataset [26] using maximum likelihood loss and measure the capability of each flow in decreasing $\mathcal{G}$ and $\mathcal{S}$. The form of the non-Standardness $\mathcal{S}$ is given by the well-known the KL-divergence between the involved normal distributions, see Appendix B.1; it is invariant under rotations Q:

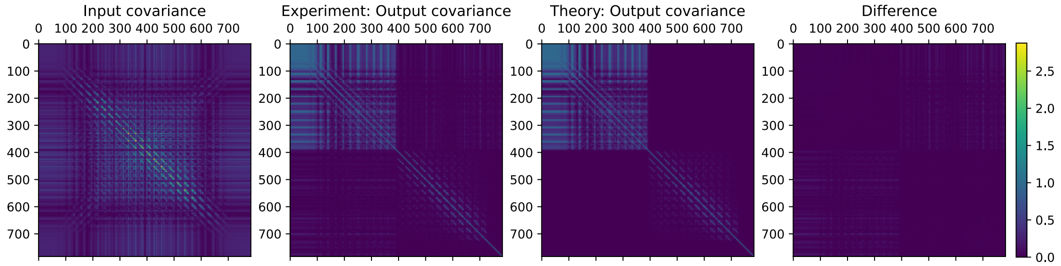
$$\mathcal{S}(m,\Sigma) := \mathcal{S}(p) = \frac{1}{2}(\|m\|^2 + \operatorname{tr}\Sigma - D - \log\det\Sigma)) = \mathcal{S}(Qm, Q\Sigma Q^{\mathrm{T}}). \tag{7}$$

By writing $\mathcal{S}(m,\Sigma)$, we emphasize the sole dependence on the first two moments.

The non-Standardness $\mathcal{S}$ will be our central measure on how far the covariance and mean have approached the target standard normal in the latent space. We give explicit loss guarantees for $\mathcal{S}$ for a single RealNVP block in Theorems 1 and 2 and imply a linear convergence rate for a deep network in Theorem 3. In practice, $\mathcal{G}$ converges similar to $\mathcal{S}$ (see Figure 1). While our results in the present paper provide quantitative bounds for $\mathcal{S}$, we merely guarantee that $\mathcal{G}$ does not increase from block to block and leave a quantitative analysis for future work.

Deep Normalizing Flows are typically trained end-to-end, i.e. all parameters are randomly initialized and the entire stack of blocks is trained at the same time. In this work, our ansatz is to consider how much an isolated coupling block can reduce the non-Standardness $\mathcal{S}(m,\Sigma)$. Then, we combine the effect of many isolated blocks, disregarding potential further improvements to $\mathcal{S}$ due to joint, cooperative learning of all blocks. This greatly simplifies the theoretical analysis of the network, but still results in strong statements. It is not a restriction on the model: Any function that is achieved in block-wise training could also be the solution of end-to-end training.

Our first result shows which mean m_1 and covariance Σ_1 a single coupling layer f_{cpl} produces to minimize $\mathcal{S}(m_1,\Sigma_1)$ given data with mean m and covariance Σ, rotated by Q:


Input covariance

Experiment: Output covariance

Theory: Output covariance

Difference

Figure 2: **How a single coupling layer can whiten the covariance** at the example of the EMNIST digits covariance matrix *(first panel)*. The covariance after a single layer trained experimentally to minimize non-Standardness $\mathcal{S}(m_1, \Sigma_1)$ *(second panel)*, which matches closely the prediction of Proposition 2 *(third panel)*. The difference between theory and experiment vanishes *(last panel)*.

Proposition 2 (Proof in Appendix B.2)**.** *Given D-dimensional data with mean m and covariance Σ and a rotation matrix Q. Split the covariance of the rotated data into four blocks, corresponding to the passive and active dimensions of the coupling layer:*

$$Q \Sigma Q^{\mathrm{T}} = \Sigma_0 = \begin{pmatrix} \Sigma_{0,pp} & \Sigma_{0,pa} \\ \Sigma_{0,ap} & \Sigma_{0,aa} \end{pmatrix} \tag{8}$$

Then, the moments m_1, Σ_1 minimizing Equation (7) that can be reached by f_{cpl} are:

$$m_1 = 0, \qquad \Sigma_1 = \begin{pmatrix} M(\Sigma_{0,pp}) & 0 \\ 0 & M(\Sigma_{0,aa} - \Sigma_{0,ap}\Sigma_{0,pp}^{-1}\Sigma_{0,pa}) \end{pmatrix}. \tag{9}$$

At this minimum of $\mathcal{S}$, $\mathcal{G}$ does not increase.

The function M takes a matrix A and rescales the diagonal to 1 as follows. It is a well-known operation in numerics called Diagonal scaling or Jacobi preconditioning:

$$M(A)_{ij} = \sqrt{A_{ii} A_{jj}}^{-1} A_{ij} = (\mathrm{Diag}(A)^{-1/2} A \, \mathrm{Diag}(A)^{-1/2})_{ij}. \tag{10}$$

Proposition 2 shows that the main change the coupling block can enact on the covariance is making the passive and active dimensions uncorrelated. This is achieved by the Schur complement $\Sigma_{0,aa} - \Sigma_{0,ap}\Sigma_{0,pp}^{-1}\Sigma_{0,pa}$. It coincides with the covariance of a Gaussian that is conditioned on the passive subspace. Afterwards, the diagonal is rescaled to one, matching the standard deviations of all dimensions with the desired latent code. The result also ensures that any improvement in $\mathcal{S}$ is not borrowed from the loss by increasing $\mathcal{G}$. In practice, we observe that both $\mathcal{S}$ and $\mathcal{G}$ can decrease by the action of a single layer (see Figure 1).

The proof is based on a more general result how a single layer maximally reduces the Maximum Likelihood Loss for arbitrary data [13]. Here, we apply their result to the non-Standardness $\mathcal{S}$. Details can be found in Appendix B.2.

Figure 2 shows an experiment in which a single layer was trained to bring the covariance of EMNIST digits [26] as close to the identity as possible. The experimental result coincides with the prediction by Proposition 2. Due to the finite batch-size, a small difference between theory and experiment remains.

5 Explicit convergence rate

In Section 4, we showed how a single affine coupling layer acts on the first two moments of a given data distribution to whiten it. We now explicitly demonstrate how much progress this means in terms of the non-Standardness $\mathcal{S}(m_1, \Sigma_1)$, averaged over rotations Q, (Theorems 1 and 2) and show the consequences for multiple blocks (Theorem 3).

5.1 Single affine coupling block guarantees

Proposition 2 allows the computation of the minimum non-Standardness after a single coupling block given its rotation Q, by evaluating $\mathcal{S}(m_1, \Sigma_1)$. In fact, if we were to choose Q such that the data is rotated so that principal components lie on the axes (i.e. obtain Q using PCA), a single coupling block suffices to reduce the covariance to the identity: $\Sigma_0 = Q\Sigma Q^{\mathrm{T}}$ would be a diagonal matrix and $\Sigma_1 = I$. This is not the case in practice, where this optimal orientation has zero probability: Q is chosen uniformly at random before training from all orthogonal matrices. One could argue that one should whiten the data before passing it to the affine coupling flow, reducing $\mathcal{S}$ to zero from the start. However, any change in the architecture could possibly alter the performance of the network with regard to reducing the non-Gaussianity $\mathcal{G}$. Also, our work shows that affine coupling flows are already well-equipped to bring the non-Standardness to zero without such modifications. To properly describe the achievable non-Standardness $\mathcal{S}$, **we formulate all guarantees as expectations over the rotation** Q, corresponding to the loss averaged over training runs.

We make two mild assumptions on our data that are part of usual data-preprocessing, when the mean is subtracted from the data and all data points are divided by the scalar $\sqrt{\operatorname{tr}\Sigma / D}$ (not to be confused with diagonal preconditioning, which acts dimension-wise).

Assumption 1. *The data* $p(x)$ *is centered:* $\mathbb{E}_{x\sim p(x)}[x] = 0$.

Assumption 2. *The covariance is normalized:* $\operatorname{tr}\Sigma = D$.

The assumptions simplify the non-Standardness in Equation (7), which now only depends on the determinant of Σ:

$$\mathcal{S}(\Sigma) = -\tfrac{1}{2}\log\det\Sigma == -\tfrac{1}{2}\log\det\Sigma_0 = \mathcal{S}(\Sigma_0) \tag{11}$$

for arbitrary rotation Q. We aim to compute the average non-Standardness after a single block $\mathbb{E}_{Q\in p(Q)}[\mathcal{S}(\Sigma_1(Q))]$. For any Q, $\mathcal{S}(\Sigma_1)$ is again given by the determinant of the covariance $\Sigma_1(Q)$ as Assumptions 1 and 2 remain fulfilled: By Proposition 2 $m_1 = 0$ and the diagonal preconditioning M ensures that the trace of Σ_1 is D. We write $\det(\Sigma_1)$ via M_a and M_p, the diagonal matrices that make up the diagonal preconditioning in Equation (10), and use the Schur determinantal formula for the determinant of block matrices: $\det(\Sigma_{0,pp})\det(\Sigma_{0,aa} - \Sigma_{0,ap}\Sigma_{0,pp}^{-1}\Sigma_{0,pa}) = \det(\Sigma_0) = \det(\Sigma)$ [27]. We thus get $\det(\Sigma_1) = \det(M_p\Sigma_{0,pp}M_p)\det(M_a(\Sigma_{0,aa} - \Sigma_{0,ap}\Sigma_{0,pp}^{-1}\Sigma_{0,pa})M_a) = \det(M_p^2)\det(M_a^2)\det(\Sigma)$. Inserting this into Equation (11), we find:

$$\mathcal{S}(\Sigma_1) = -\tfrac{1}{2}(\log\det\Sigma + \log\det M_p^2 + \log\det M_a^2) \leq \mathcal{S}(\Sigma_0) = \mathcal{S}(\Sigma). \tag{12}$$

The inequality $\mathcal{S}(\Sigma_1) \leq \mathcal{S}(\Sigma_0)$ holds because $\Sigma_1 = \Sigma_0$ is an admissible solution of the coupling layer optimization, but Σ_1 as given by Proposition 2 is a minimizer of $\mathcal{S}(\Sigma_1)$.

We average this quantity over training runs, i.e. over rotations Q:

$$\mathbb{E}_{Q\sim p(Q)}[\mathcal{S}(\Sigma_1)] = -\tfrac{1}{2}\big(\log\det\Sigma + \mathbb{E}_{Q\sim p(Q)}[\log\det M_p^2] + \mathbb{E}_{Q\sim p(Q)}[\log\det M_a^2]\big). \tag{13}$$

The main difficulty lies in the computation of $\mathbb{E}_{Q\sim p(Q)}[\log\det M_a^2]$. Here, we contribute the two strong statements Theorems 1 and 2 below.

5.1.1 Precise guarantee

The first result relies on projected orbital measures as developed by [28]. This theory describes the eigenvalues of submatrices of matrices in a random basis. We require such a result for integrating over $p(Q)$ in $\mathbb{E}_{Q\sim p(Q)}[\log\det M_a^2]$. In contrast to typical choices of $p(Q)$, the theory to this date only covers data rotated by unitary matrices.[1] To comply with [28], we make the following two additional assumptions:

Assumption 3. *The distribution of rotations is the Haar measure over* unitary *matrices* $U(D)$.

Assumption 4. *The eigenvalues of the covariance matrix* Σ *are distinct:* $\lambda_i \neq \lambda_j$ *for* $i \neq j$.

One could think that the step from orthogonal to unitary rotations takes us far away from the scenario we want to consider. We will later observe empirically that the difference between averaging over

[1] The only result known to us would yield predictions for $D = 2$ [29], whereas we are interested in the case of large D.

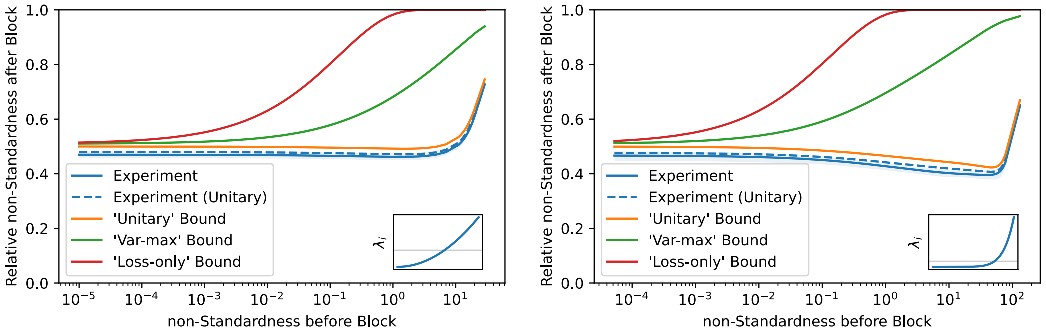


Figure 3: **Comparison between predicted non-Standardness and experiment** for 48-dimensional parametrized eigenvalue spectra *(insets)*, varied over a parameter which controls the spread of the spectrum and thus changes $\mathcal{S}$. The experimental average over *orthogonal* rotations matrices (blue, shaded by Interquartile Range IQR) is closely matched by the experimental average over *unitary* matrices (dotted blue). The prediction by Theorem 1 is a close upper bound that closely matches the experimental behavior (orange). The predictions by Theorem 2 are less precise, but converge to the same value as the precise bound for covariances close to the identitiy: 'Var-max' is Equation (16a) (green) and 'Loss-only' is Equation (16b) (red). More details and examples in Appendix A.3.

unitary and orthogonal matrices is negligible. Technically, the data remains real at the optimum if we apply the inverse of the rotation layer after the transformation (see Appendix B.3). We will write $\mathbb{E}_{Q\sim U(D)}[\,\cdot\,]$ to denote expectations over unitary matrices. Assumption 4 is typically satisfied when working with real data that are in 'general position'.

We are now ready to **compute the average training performance** of a single affine coupling block:

Theorem 1 (Proof in Appendix B.3)**.** *Given D-dimensional data with covariance Σ with eigenvalues $\lambda_1, \dots \lambda_D$. Assume that Assumptions 1 to 4 hold. Then, after a single affine coupling block, the expected non-Standardness is bounded from above:*

$$\mathbb{E}_{Q\in U(D)}[\mathcal{S}(\Sigma_1(Q))] < \mathcal{S}(\Sigma) + \tfrac{D}{2}\log\left((-1)^{\frac{D}{2}+1}\sum_{i=1}^{D}\lambda_i^{1-\frac{D}{2}}\log(\lambda_i)R(\lambda_i^{-1};\lambda_{\neq i}^{-1})e_{\frac{D}{2}-1}(\lambda_{\neq i}^{-1})\right). \tag{14}$$

Here, $\lambda_{\neq i} := \{\lambda_1, \dots, \lambda_{i-1}, \lambda_{i+1}, \dots, \lambda_D\}$ and R, e_K are given by:

$$R(a;\{b_i\}_{i=1}^N) = \prod_{i=1}^{N}\frac{1}{a-b_i} \quad \textit{and} \quad e_K(\{b_i\}_{i=1}^N) = \sum_{0<i_1<\dots<i_K\leq N} b_{i_1}\cdots b_{i_K}. \tag{15}$$

Inequality (14) sharply bounds the expected non-Standardness that can be achieved by a single affine block. The only approximation made is an inequality which comes close to equality as the dimension D increases due to the concentration of the corresponding probability distribution.

Figure 3 shows an experiment confirming Theorem 1. We start with covariance matrices using parametrized eigenvalue spectra. On each, we first apply a single affine coupling block with random Q and train the coupling that maximally reduces $\mathcal{S}$ (Proposition 2). Then we iteratively append 32 additional blocks in the same manner, building a flow of that depth. We average the resulting empirical ratio $\mathcal{S}(\Sigma_1)/\mathcal{S}(\Sigma)$ over several *orthogonal* orientations Q of the rotation layer for each input covariance matrix. Then, we compare this to (i) experimentally averaging over *unitary* rotations and (ii) to the prediction by Theorem 1 and confirm that it is a valid and close upper bound. Details for replication and more examples can be found in Appendix A.3.

The proof explicitly integrates $\mathbb{E}[M_a^2]$ using [28] (see Appendix B.3). Numerically evaluating Equation (14) can be hard even for small D as the summands scale as $\mathcal{O}(\exp(D))$, but the overall sum scales as $\mathcal{O}(D)$. High values cancel due to R alternating in sign, and one requires arbitrary-precision floating point software to evaluate Equation (14).

5.1.2 Interpretable guarantee

The guarantee in Theorem 1 yields useful predictions, but it does not lend itself to further analysis: How does the bound behave over several coupling blocks? What is the behavior for varying dimension D? Also, Assumption 3 restricts formal reasoning as we are interested in averaging over orthogonal and not unitary rotations. Our second single-block guarantee depends only on simple metrics of the covariance. Moreover, we drop Assumptions 3 and 4, averaging over orthogonal, not unitary, Q:

Theorem 2 (Proof in Appendix B.4)**.** *Given D-dimensional data fulfilling Assumptions 1 and 2 with covariance Σ with eigenvalues $\lambda_1, \ldots \lambda_D$. Then, after a single affine coupling block, the expected loss can be bounded from above:*

$$\begin{aligned}\mathbb{E}_{Q\in O(D)}[\mathcal{S}(\Sigma_1(Q))] &\le \mathcal{S}(\Sigma) + \tfrac{D}{4}\log\left(1 - \frac{D^2}{2(D-1)(D+2)}\frac{\mathrm{Var}[\lambda]}{\lambda_{\max}}\right) && \text{(16a)}\\ &\le \mathcal{S}(\Sigma) + \tfrac{D}{4}\log\left(1 - \frac{D^2}{(D-1)(D+2)}\frac{1-\sqrt{1-g^D}}{1+\sqrt{1-g^D}}(1-g)\right) && \text{(16b)}\\ &< \mathcal{S}(\Sigma). && \text{(16c)}\end{aligned}$$

Here, g is the geometric mean of the eigenvalues: $g = \prod_{i=1}^{D}\lambda_i^{1/D} = \exp(-2\mathcal{S}(\Sigma)/D) < 1$ which is a bijection of $\mathcal{S}(\Sigma)$.

These two new bounds on the average achievable non-Standardness $\mathcal{S}$ after a single block are also depicted in Figure 3. They make useful predictions, but are less precise than Theorem 1. The second bound will be especially useful in what follows because it only depends on the non-Standardness before the block $\mathcal{S}(\Sigma)$.

The full proof is given in Appendix B.4. It relies on the integration of monomials of entries of random orthogonal matrices as described by [30] and the arithmetic mean-geometric mean inequality by [31].

The first bound suggests an important property of the non-Standardness convergence of an affine coupling flow in terms of dimension: The performance only marginally depends on the dimension. To see this, divide Equation (16a) by D to obtain a statement about the non-Standardness per dimension $\mathcal{S}/D$. Then take several data sets with different dimension but same spectrum characteristics (i.e. same geometric mean, variance and maximum of covariance eigenvalues). The guarantee is then approximately constant in D (it varies slightly with $D^2/(D^2+D-2)$, which is always close to 1).

5.2 Deep network guarantee

The previous Section 5.1 was concerned with determining how much a *single* RealNVP block can typically contribute towards reducing the $\mathcal{S}$ to zero. Now, we extend this result to compute the expected non-Standardness after a *deep* affine coupling network as an explicit function of the number of blocks. We again treat the rotation layer of each block as a random variable, as it is randomly determined before training.

We find that the **convergence rate** of the covariance to the identity is (at least) **linear**:

Theorem 3 (Proof in Appendix B.5)**.** *Given D-dimensional data fulfilling Assumptions 1 and 2 with covariance Σ. Then, after L affine coupling blocks, the expected loss is smaller than:*

$$\mathbb{E}_{Q_1,\ldots,Q_L\in O(D)}[\mathcal{S}(\Sigma_L)] \le \gamma\big(\mathcal{S}(\Sigma)\big)^L \mathcal{S}(\Sigma), \tag{17}$$

where the convergence rate depends on the non-Standardness before training:

$$\gamma(\mathcal{S}) = 1 + \tfrac{1}{4\mathcal{S}/D}\log\left(1 - \frac{D^2}{(D-1)(D+2)}\frac{1-\sqrt{1-g(\mathcal{S})^D}}{1+\sqrt{1-g(\mathcal{S})^D}}\big(1-g(\mathcal{S})\big)\right) < 1. \tag{18}$$

The non-Standardness decreases at least exponentially fast in the number of blocks. The convergence rate that holds for a deep network is computed using the non-Standardness of the input data $\mathcal{S}(\Sigma)$. This rate comes from Equation (16b). The proof uses that $\gamma(\mathcal{S})$ improves from block to block as $\mathcal{S}$ decreases (see Appendix B.5). Again, $g(\mathcal{S}) = \exp(-2\mathcal{S}/D) < 1$ is the geometric mean of eigenvalues of Σ, which increases from block to block.

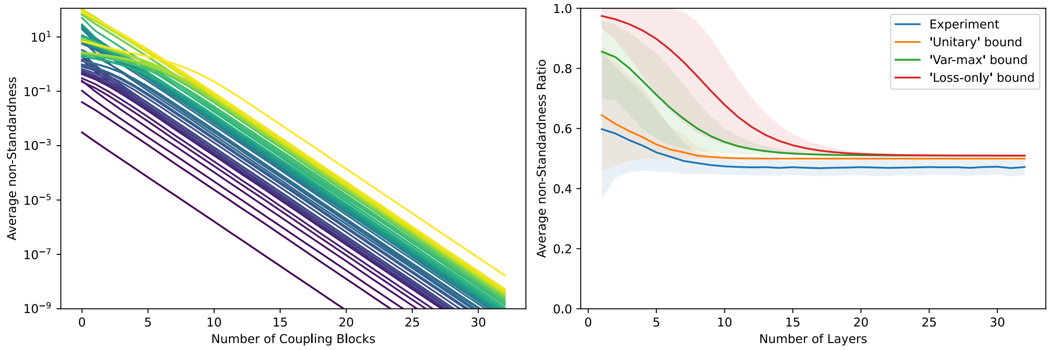


Figure 4: **Deep network convergence of covariance on toy dataset.** *(Left)* Each line shows the experimental convergence of $\mathcal{S}$ via the repeated application of Proposition 2, averaged over 32 runs with different rotations Q. *(Right)* The empirical convergence rate (blue), i.e. the ratio of $\mathcal{S}$ before and after a block, is correctly bounded from above by our predictions in Theorem 1 (orange), and the bounds in Theorem 2: Equation (16a) (green) and Equation (16b) (red). The solid lines show the ratio (bounds) averaged over the toy dataset and rotations, the shade is the IQR. The experiment suggests that a convergence rate like Theorem 3 can also be derived for the remaining bounds.

Figure 4 shows the convergence of the non-Standardness to zero in an experiment. We build a toy dataset of various covariances where we aim to capture a plethora of possible cases (see Appendix A.4). We apply a single affine coupling block with random Q and the coupling that maximally reduces $\mathcal{S}$ via Proposition 2. We iterate adding such blocks 32 times, building a flow of that depth. The resulting convergence of $\mathcal{S}$ as a function of depth is averaged over 32 runs with different rotations. The measured curve confirms Theorem 3. We find that the rate γ in Equation (18) is correct, but several experiments show even faster convergence in practice. Indeed, the experiments suggest that dividing all upper bounds for $\mathbb{E}[\mathcal{S}(\Sigma_1)]$ in Theorems 1 and 2 by $\mathcal{S}(\Sigma)$ also bounds the non-Standardness ratio for subsequent blocks. Formally, we conjecture that $\mathbb{E}[\mathcal{S}(\Sigma_L)]/\mathcal{S}(\Sigma) \leq (B/\mathcal{S}(\Sigma))^L$ where B is the rhs. of Equations (14) and (16a) (Theorem 3 shows exactly this for Equation (16b)). We leave a proof or falsification of this conjecture open to future work.

The experiment also suggests that all bounds agree after a few blocks, leaving a small gap to the experiment. We can explicitly compute this limit value of $\gamma(\mathcal{S})$ by taking $\mathcal{S} \to 0$:

$$\gamma(\mathcal{S}) \xrightarrow{\mathcal{S}\to 0} \tfrac{D(D+2)-4}{2(D-1)(D+2)} \in \big[1/2, 5/9\big]. \tag{19}$$

The two experimental observations together with this limit value suggest the heuristic that **a single additional coupling block typically reduces the non-Standardness $\mathcal{S}$ by a factor of approximately 50%** if previous blocks are left unchanged, and possibly faster if cooperations between blocks are taken into account.

6 Conclusion

To the best of our knowledge, this is the first work on affine coupling flows that provides a quantitative convergence analysis in terms of the KL divergence. Specifically, a minimal convergence rate is established at which affine flows whiten the covariance of the input data under this strong measure of discrepancy of probability distributions. Splitting the loss into the non-Gaussianity $\mathcal{G}$ and the non-Standardness $\mathcal{S}$, we show that this whitening is a necessary condition for the flow to converge and give explicit guarantees. Our derivations suggest the rule of thumb that $\mathcal{S}$ can typically be reduced by about 50% per coupling block.

Our central idea was to separate out the contribution a single isolated block can make to reduce the loss, arguing that end-to-end training can only outperform the concatenation of isolated blocks.

Having separated the tasks an affine coupling flow has to solve, and having explained how the non-Standardness $\mathcal{S}$ can be reduced to zero, we hope that explaining the convergence under non-Gaussianity $\mathcal{G}$ comes within reach. In practice, it converges similarly fast as $\mathcal{S}$ (see Figure 1).

References

[1] Takeshi Teshima, Isao Ishikawa, Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama. Coupling-based Invertible Neural Networks Are Universal Diffeomorphism Approximators. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3362–3373. Curran Associates, Inc., 2020.

[2] Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, November 2021.

[3] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. *J. Machine Learning Research*, 22:1–64, April 2021. arXiv: 1912.02762.

[4] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[5] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[6] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.

[7] Radek Mackowiak, Lynton Ardizzone, Ullrich Kothe, and Carsten Rother. Generative classifiers as a basis for trustworthy image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2971–2981, 2021.

[8] Lynton Ardizzone, Radek Mackowiak, Carsten Rother, and Ullrich Köthe. Training Normalizing Flows with the Information Bottleneck for Competitive Generative Classification. *Advances in Neural Information Processing Systems*, 33, 2020. arXiv: 2001.06448.

[9] Lynton Ardizzone, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Analyzing Inverse Problems with Invertible Neural Networks. In *International Conference on Learning Representations*, 2018.

[10] Pablo Noever-Castelos, Lynton Ardizzone, and Claudio Balzani. Model updating of wind turbine blade cross sections with invertible neural networks. *Wind Energy*, 25(3):573–599, March 2022.

[11] Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019. Publisher: American Association for the Advancement of Science.

[12] Tim J. Adler, Lynton Ardizzone, Anant Vemuri, Leonardo Ayala, Janek Gröhl, Thomas Kirchner, Sebastian Wirkert, Jakob Kruse, Carsten Rother, Ullrich Köthe, and Lena Maier-Hein. Uncertainty-aware performance assessment of optical imaging modalities with invertible neural networks. *International Journal of Computer Assisted Radiology and Surgery*, 14(6):997–1007, June 2019.

[13] Felix Draxler, Jonathan Schwarz, Christoph Schnörr, and Ullrich Köthe. Characterizing the Role of a Single Coupling Layer in Affine Normalizing Flows. In Zeynep Akata, Andreas Geiger, and Torsten Sattler, editors, *Pattern Recognition*, volume 12544, pages 1–14. Springer International Publishing, Cham, 2021. Series Title: Lecture Notes in Computer Science.

[14] Frederic Koehler, Viraj Mehta, and Andrej Risteski. Representational aspects of depth and conditioning in normalizing flows. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5628–5636. PMLR, July 2021.

[15] A. L. Gibbs and F. E. Su. On Choosing and Bounding Probability Metrics. *Int. Statistical Review*, 70(3):419–435, 2002.

[16] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. John Wiley & Sons, Inc., 2nd edition, 2009.

[17] Jean-Francois Cardoso. Dependence, Correlation and Gaussianity in Independent Component Analysis. *J. Machine Learning Research*, 4:1177–1203, 2003.

[18] Holden Lee, Chirag Pabbaraju, Anish Sevekari, and Andrej Risteski. Universal Approximation for Log-concave Distributions using Well-conditioned Normalizing Flows. *arXiv:2107.02951 [cs, stat]*, July 2021. arXiv: 2107.02951.

[19] Priyank Jaini, Ivan Kobyzev, Yaoliang Yu, and Marcus Brubaker. Tails of Lipschitz Triangular Flows. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4673–4681. PMLR, July 2020.

[20] Priyank Jaini, Kira A. Selby, and Yaoliang Yu. Sum-of-Squares Polynomial Flow. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3009–3018. PMLR, June 2019.

[21] Isao Ishikawa, Takeshi Teshima, Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama. Universal approximation property of invertible neural networks, 2022.

[22] Chin-Wei Huang, Laurent Dinh, and Aaron Courville. Augmented Normalizing Flows: Bridging the Gap Between Generative Flows and Latent Variable Models. *arXiv:2002.07101 [cs, stat]*, February 2020. arXiv: 2002.07101.

[23] Han Zhang, Xi Gao, Jacob Unterman, and Tom Arodz. Approximation Capabilities of Neural ODEs and Invertible Residual Networks. *arXiv:1907.12998 [cs, stat]*, February 2020. arXiv: 1907.12998.

[24] Takeshi Teshima, Koichi Tojo, Masahiro Ikeda, Isao Ishikawa, and Kenta Oono. Universal Approximation Property of Neural Ordinary Differential Equations. *arXiv:2012.02414 [cs, math, stat]*, December 2020. arXiv: 2012.02414.

[25] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, Rhode Island, April 2007.

[26] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: an extension of MNIST to handwritten letters. *arXiv:1702.05373 [cs]*, March 2017. arXiv: 1702.05373.

[27] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[28] Grigori Olshanski. Projections of orbital measures, Gelfand-Tsetlin polytopes, and splines. *arXiv:1302.7116 [math]*, February 2013. arXiv: 1302.7116.

[29] Jacques Faraut. Rayleigh theorem, projection of orbital measures and spline functions. *Advances in Pure and Applied Mathematics*, 6(4), January 2015.

[30] T. Gorin. Integrals of monomials over the orthogonal group. *Journal of Mathematical Physics*, 43(6):3342–3351, June 2002.

[31] D. I. Cartwright and M. J. Field. A refinement of the arithmetic mean-geometric mean inequality. *Proceedings of the American Mathematical Society*, 71(1):36–38, 1978.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] see Section 6.

    (c) Did you discuss any potential negative societal impacts of your work? [Yes] Generative modeling, which this paper aims to improve, can be used in harmful ways to generate Deepfakes for disinformation.
    (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
    (a) Did you state the full set of assumptions of all theoretical results? [Yes]
    (b) Did you include complete proofs of all theoretical results? [Yes] Full proofs are in Appendix B.
3. If you ran experiments...
    (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Code will be made available upon publication.
    (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix A.
    (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Except for Figure 1 which shows a single training run per block, but a different seed is used for each initialization.
    (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix A.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
    (a) If your work uses existing assets, did you cite the creators? [Yes]
    (b) Did you mention the license of the assets? [N/A]
    (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
    (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
    (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
    (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
    (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]