Appendix A

A1. Error analysis In importance sampling, given a point $\mathbf{x}$, we sample $\mathbf{z}_i$ from an importance distribution $q(\mathbf{z})$ instead of base density $p(\mathbf{z})$ which can be represented by

$$\int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \int p(\mathbf{x}|\mathbf{z})w(\mathbf{z})q(\mathbf{z})d\mathbf{z} = \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[p(\mathbf{x}|\mathbf{z})w(\mathbf{z})] \approx \frac{1}{N}\sum_{i}^{N} p(\mathbf{x}|\mathbf{z}_i^q)w(\mathbf{z}_i^q) \quad (1)$$

where N is the sample size, $w(\mathbf{z}) = \frac{p(\mathbf{z})}{q(\mathbf{z})}$ is the importance weight function, $\{\mathbf{z}_i^q\}_{i=1}^N$ are $i.i.d$ samples from $q(\mathbf{z})$. We propose to set $q(\mathbf{z})$ to be a Student's t distribution with the center at $\tilde{\mathbf{z}} = H(\mathbf{x})$

To simplify, we denote target density value as μ_x given a point $\mathbf{x}$. Then we have $\mu_x = \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[p(\mathbf{x}|\mathbf{z})w(\mathbf{z})]$ according to (1). The corresponding variance $\sigma_x^2 = \mathrm{Var}_{\mathbf{z}\sim \mathrm{q}(\mathbf{z})}[\mathrm{p}(\mathbf{x}|\mathbf{z})\mathrm{w}(\mathbf{z})]$ can be represented as

$$\begin{aligned}
\sigma_x^2 &= \mathrm{Var}_{\mathbf{z}\sim \mathrm{q}(\mathbf{z})}[\mathrm{p}(\mathbf{x}|\mathbf{z})\mathrm{w}(\mathbf{z})] = \mathbb{E}_{\mathbf{z}\sim \mathrm{q}(\mathbf{z})}[\mathrm{p}^2(\mathbf{x}|\mathbf{z})\mathrm{w}^2(\mathbf{z})] - (\mathbb{E}_{\mathbf{z}\sim \mathrm{q}(\mathbf{z})}[\mathrm{p}(\mathbf{x}|\mathbf{z})\mathrm{w}(\mathbf{z}])^2 \\
&= \int p^2(\mathbf{x}|\mathbf{z})w^2(\mathbf{z})q(\mathbf{z})d\mathbf{z} - \mu_x^2 = \int \frac{p^2(\mathbf{x}|\mathbf{z})p^2(\mathbf{z})}{q(\mathbf{z})}d\mathbf{z} - \mu_x^2 \\
&= \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[(\frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) - \mu_x q(\mathbf{z})}{q(\mathbf{z})})^2] = \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[(p(\mathbf{x}|\mathbf{z})w(\mathbf{z}) - \mu_x)^2]
\end{aligned} \quad (2)$$

Then we analyze the expectation and variance of $\hat{\mu}_q = \frac{1}{N}\sum_i^N p(\mathbf{x}|\mathbf{z}_i^q)w(\mathbf{z}_i^q)$ (right hand side of (1)). As $\{\mathbf{z}_i^q\}_{i=1}^N$ are $i.i.d$ samples from $q(\mathbf{z})$, we have $\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\hat{\mu}_q] = \mu_x$ and $\mathrm{Var}_{\mathbf{z}\sim \mathrm{q}(\mathbf{z})}[\hat{\mu}_\mathrm{q}] = \frac{1}{\mathrm{N}}\sigma_\mathrm{x}^2$. It is obvious that $\hat{\mu}_q$ is a unbiased estimate of μ_x(target density at $\mathbf{x}$). Further, we give a natural variance estimate as $\hat{\sigma}_x = \frac{1}{N}\sum_i^N (p(\mathbf{x}|\mathbf{z}_i^q)w(\mathbf{z}_i^q) - \hat{\mu}_q)^2$. According to central limit theorem (CLT), we can give an appriximate 99% confidence interval for μ_x(target density at $\mathbf{x}$) as $\hat{\mu}_q \pm 2.58\hat{\sigma}_x/\sqrt{N}$.

A2. An illustrative example We used importance sampling to get numeric result of $\int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$. One key problem is to choose an appropriate importance distribution $q(\mathbf{z})$. In Roundtrip model, we chose $q(\mathbf{z})$ as student's t distribution with center at $H(\mathbf{x})$. $p(\mathbf{x}|\mathbf{z})$ always takes optimal maximum value at $\tilde{\mathbf{z}} = H(\mathbf{x})$ as

$$p(\mathbf{x}|\tilde{\mathbf{z}}) = (\frac{1}{\sqrt{2\pi}\sigma})^n e^{-\frac{\|\mathbf{x}-G(\tilde{\mathbf{z}})\|_2^2}{2\sigma^2}} = (\frac{1}{\sqrt{2\pi}\sigma})^n e^{-\frac{\|\mathbf{x}-G(H(\mathbf{x}))\|_2^2}{2\sigma^2}} \quad (3)$$

It is easy to prove that minimizing roundtrip loss $\rho(\mathbf{x}, G(H(\mathbf{x})))$ is equivalent to maximizing $p(\mathbf{x}|\tilde{\mathbf{z}})$.

To make the importance sampling strategy more understandable, we illustrated an example based on the simulation study here. We take the *Involute* simulation case in Section 3.3 for an example, we visualize $p(\mathbf{z})$, $p(\mathbf{x}|\mathbf{z})$ and $q(\mathbf{z})$ at the first dimension focusing on the density at the point **x=**(3,3) (Figure S1). At this point the true density is 4.33×10^{-3}. The confidence interval analyzed in section A1 is $(4.38\pm0.07)\times10^{-3}$ which covers the truth density well. However, if we set $q(\mathbf{z})$ as a stantard normal distribution. The confidence interval will be $(3.77\pm0.27)\times10^{-3}$ which fails to cover the ground truth and has a much larger interval length.

As $p(\mathbf{x}|\mathbf{z})$ typicaly decays much faster than $p(\mathbf{z})$, we chose $q(\mathbf{z})$ in which the center is close to the center of $p(\mathbf{x}|\mathbf{z})$ as much as possible. To sum up, in the density estimation framework using the importance sampling strategy, $G(\mathbf{z})$ network was used for genenating samples while $H(\mathbf{x})$ network was used for determining the center of importance distribution $q(\mathbf{z})$.

Appendix B

B1. The validity of constructed multivariate Gaussian We first prove $(\mathbf{I} + \lambda\mathbf{A})$ is positive definite. Denote the Jacobian matrix as $\mathbf{J} = \nabla G(\mathbf{z})$, then $\mathbf{A} = \mathbf{J}^T\mathbf{J}$. For any $\mathbf{y} \in \mathbb{R}^m \backslash \{\mathbf{0}\}$.

$$\begin{aligned}
\mathbf{y}^T(\mathbf{I} + \lambda\mathbf{A})\mathbf{y} &= \|\mathbf{y}\|_2^2 + \lambda\mathbf{y}^T\mathbf{A}\mathbf{y} \\
&= \|\mathbf{y}\|_2^2 + \lambda\mathbf{y}^T\mathbf{J}^T\mathbf{J}\mathbf{y} \\
&= \|\mathbf{y}\|_2^2 + \lambda\|\mathbf{J}\mathbf{y}\|_2^2 > \mathbf{0}
\end{aligned} \quad (4)$$

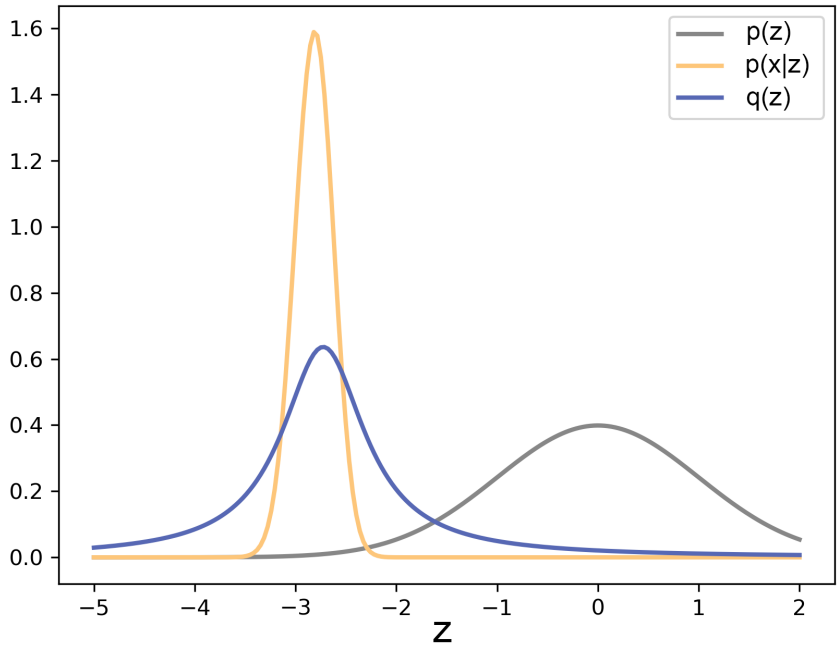

Figure S1. Distribution of $p(\mathbf{z})$, $p(\mathbf{x}|\mathbf{z})$ and $q(\mathbf{z})$ for estimating density at point **x**=(3,3).

So $(\mathbf{I}+\lambda\mathbf{A})$ is positive definite and all the eigenvalues $(\lambda_1, ..., \lambda_m)$ are positive. Then $\mathbf{\Sigma} = (\mathbf{I}+\lambda\mathbf{A})^{-1}$ is positive definite as all the eigenvalues $(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_m})$ are also positive. This property makes sure that the constructed covariance matrix is valid.

B2. *Change of variable rule* as a special case We first rephrase the density of **x** in equation (11) as the following

$$\begin{aligned} p(\mathbf{x}) &= (\frac{1}{\sqrt{2\pi}})^n \sigma^{-n} \sqrt{\det(\mathrm{inv}(\mathbf{I}+\sigma^{-2}\mathbf{A}))} e^{-\frac{c_2(\mathbf{x})}{2}} \\ &= (\frac{1}{\sqrt{2\pi}})^n \sigma^{-n} \sqrt{\det(\sigma^2 \mathrm{inv}(\mathbf{A}+\sigma^2\mathbf{I}))} e^{-\frac{c_2(\mathbf{x})}{2}} \\ &= (\frac{1}{\sqrt{2\pi}})^n \sigma^{-n} \sqrt{\sigma^{2m}\det(\mathrm{inv}(\mathbf{A}+\sigma^2\mathbf{I}))} e^{-\frac{c_2(\mathbf{x})}{2}} \\ &= (\frac{1}{\sqrt{2\pi}})^n \sigma^{m-n} \sqrt{\det(\mathrm{inv}(\mathbf{A}+\sigma^2\mathbf{I}))} e^{-\frac{c_2(\mathbf{x})}{2}} \end{aligned} \tag{5}$$

When $m=n$ and $H(\cdot) = G^{-1}(\cdot)$,then we have $x - G(\tilde{\mathbf{z}}) = \mathbf{x} - G(H(\mathbf{x})) = \mathbf{x} - G(G^{-1}(\mathbf{x})) = \mathbf{0}$, $\mathbf{b} = \nabla G^T(\tilde{\mathbf{z}})(\mathbf{x} - G(\tilde{\mathbf{z}})) = \mathbf{0}$, $\boldsymbol{\mu} = \mathbf{\Sigma}(\lambda\mathbf{b} - \tilde{\mathbf{z}}) = -\mathbf{\Sigma}\tilde{\mathbf{z}}$ and $c_2(\mathbf{x}) = \|\tilde{\mathbf{z}}\|_2^2 - \sigma^2\tilde{\mathbf{z}}^T(\mathbf{A}+\sigma^2\mathbf{I})^{-1}\tilde{\mathbf{z}}$

Finally, we take the limit of $\sigma \to 0$, we have $\lim_{\sigma\to 0} c_2(\mathbf{x}) = \|\tilde{\mathbf{z}}\|_2^2$ and

$$\begin{aligned} \lim_{\sigma\to 0} \sqrt{\det(\mathrm{inv}(\mathbf{A}+\sigma^2\mathbf{I}))} &= \sqrt{\det(\mathrm{inv}(\mathbf{A}))} = \sqrt{\det(\mathrm{inv}(\mathbf{J}^{\mathrm{T}}\mathbf{J}))} \\ &= \sqrt{\det(\mathbf{J}^{-\mathrm{T}}\mathbf{J}^{-1})} = |\det(\mathbf{J}^{-1})| = |\det(\frac{\partial \mathrm{G}(\tilde{\mathbf{z}})}{\partial \tilde{\mathbf{z}}^{\mathrm{T}}})|^{-1} \end{aligned} \tag{6}$$

So when m=n and $H(\cdot) = G^{-1}(\cdot)$, then $\lim_{\sigma\to 0} p(\mathbf{x}) = (\frac{1}{\sqrt{2\pi}})^n e^{-\frac{\|\tilde{\mathbf{z}}\|_2^2}{2}} |\det(\frac{\partial \mathrm{G}(\tilde{\mathbf{z}})}{\partial \tilde{\mathbf{z}}^{\mathrm{T}}})|^{-1} = \mathrm{p}(\tilde{\mathbf{z}})|\det(\frac{\partial \mathrm{G}(\tilde{\mathbf{z}})}{\partial \tilde{\mathbf{z}}^{\mathrm{T}}})|^{-1}$.

So we proved that under the three conditions (1) $m = n$, 2) $H(\cdot) = G^{-1}(\cdot)$ (3) $\sigma \to 0$, the proposed Laplace approximation is degraded into *Change of variable rule*. Our Laplace approximation

approach can be considered as an extension of *Change of variable rule* which requires equal dimension in base density and target density. More importantly, if the dimension in target density (n) is extremely large, traditional neural density estimators need to calculate the determinant and inverse of a $n \times n$ Jacobian matrix, which may be computational expensive. The Laplace approximation approach is flexible in setting the dimension of base density (m). The computation only relates to the determinant and inverse of a $m \times m$ matrix $\mathbf{A}$ which can largely reduce the computational complexity.

Appendix C

we took the case (a) independent Gaussian mixture for a further study by increasing the dimension up to 10 (containing 3^{10} modes). The Spearman correlation between estimated density and true density of the test set is calculated and shown in Figure S2. The kernel density estimator (KDE) performs comparable or even better when the dimension is less than 5. But the performance of KDE decreases sharply when the dimension is larger than 5. Our Roundtrip model with importance sampling (Roundtrip-IS) strategy can always achieve a better performance than other neural density estimators at different dimensions. We also note that the performance of Roundtrip model with closed-form solution (Roundtrip-CF) outperforms MADE but not as good as MAF and Real NVP.

Although we provid theoretical guarantees on the approximation solution, the success of Roundtrip-CF requires that the high order terms in equation (6) is negligible, which will introduce additional bias when estimating density. So we reported all results of density estimation using the more robust Roundtrip-IS model (default setting) as the result of importance sampling is unbiased.

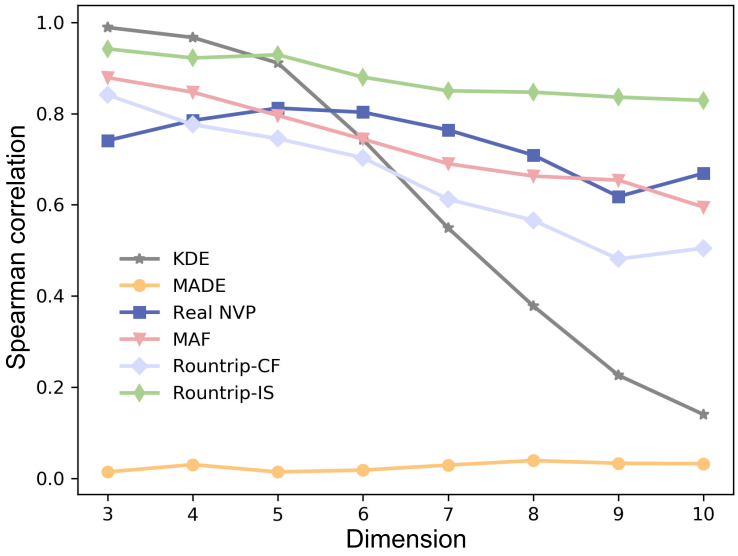

Figure S2. Performance of different density estimators at different dimensions.

Appendix D

UCI and image datasets We provided detailed descriptions about the data description and preprocessing of all datasets that were used in our study.

AreM. The Activity Recognition system based on Multisensor data fusion (AReM) [3] dataset contains temporal data from a Wireless Sensor Network worn by an actor performing the activities: bending, cycling, lying down, sitting, standing, walking. The time-domain features including 3 mean values and 3 standard deviations were collected from the multisensor system during a period

Table S1. Dimension and sample size of UCI/Image datasets

Dataset	Domain	Dim($\mathbf{z}$)	Dim($\mathbf{x}$)	Sample size		
				Train	Validation	Test
AReM	Social science	3	6	34215	3801	4223
CASP	Chemistry	5	9	37042	4115	4573
HEPMASS	Physics	8	21	315123	35013	174987
BANK	Finance	8	17	36621	4069	4521
YPMSD	Audio	20	90	417430	46381	51534
MNIST	Image	100	784	50000	10000	10000
CIFAR-10	Image	100	3072	45000	5000	10000

of time. Although it is time-series data but we treat it as if each example was drawn from an iid distribution from the target distribution. Then raw data was first applied a feature scaling through a min-max normalization and then randomly split into 90% training set and 10% test. Note that for neural density estimators, 10% of the training set will be kept for validation.

CASP. The CASP dataset contains the physicochemical properties of the protein tertiary structure. Each example denotes an individual residue which has 9 features, including total surface area, non-polar exposed area, fractional area of exposed non-polar residue, fractional area of exposed non-polar part of residue, molecular mass weighted exposed area, euclidian distance, secondary structure penalty and spacial distribution constraints (N.K Value). The same data normalization and split were used as AreM dataset.

HEPMASS. HEPMASS [1] dataset describes the particle collisions signatures of exotic particles in high energy physics. We preprocessed this dataset following the same strategy as [4]. Examples from the "1000" dataset were collected where the particle mass is 1000 and five features were removed due to too many reoccurring values.

BANK. BANK dataset [2] is related to a marketing campaign of a Portuguese banking institution where the goal is to predict whether the client will subscribe a deposit. The label encoding was used for discrete features in the raw data with values between 0 and `n_classes`. Then a uniform noise of $(-0.2, 0.2)$ was added to each feature. At last, the same data normalization and split were used as AreM dataset.

YPMSD. YPMSD (`http://millionsongdataset.com/`) is a dataset that contains the audio features of songs from different years ranging from 1922 to 2011. Each song has 90 features which relate to 12 timbre average and 78 timbre covariance. The same data normalization and split were used as AreM dataset.

The descriptions of the five UCI datasets and the two image datasets (MNIST and CIFAR-10), including feature dimension and sample size, were summarized in Table S1.

ODDS datasets *Shuttle*. Shuttle (`http://odds.cs.stonybrook.edu/shuttle-dataset/`) dataset contains 9 numerical features. The smallest five classes, i.e. 2, 3, 5, 6, 7 are combined to form the outliers class, while class 1 forms the inlier class. Data for class 4 is discarded. All inlier and outlier data were first mixed together and then randomly split into 90% training set and 10% test set. For neural density estimators, 10% of the training set were kept for validation.

Mammography Mammography (`http://odds.cs.stonybrook.edu/mammography-dataset/`) dataset describes the characteristics of 260 calcifications. The minority class of calcification is considered as an outlier class and the non-calcification class as inliers. The same data split strategy was used for Shuttle dataset.

ForestCover ForestCover (`http://odds.cs.stonybrook.edu/forestcovercovertype-dataset/`) dataset is used in predicting forest cover type from cartographic variables. Outlier detection dataset is created using only 10 quantitative attributes. Instances from class 2 are considered as normal points and instances from class 4 are anomalies. The same data split strategy was used for Shuttle dataset.

The descriptions of the three ODDS datasets are summarized in Table S2.

Table S2. Dimension and sample size of ODDS datasets

Dataset	Dim($\mathbf{z}$)	Dim($\mathbf{x}$)	Outliers(%)	Sample size		
				Train	Validation	Test
Shuttle	3	9	7	39770	4418	4909
Mammograph	3	6	2.32	9059	1006	1118
ForestCover	4	10	0.9	231700	25744	28604

References

[1] Pierre Baldi, Kyle Cranmer, Taylor Faucett, Peter Sadowski, and Daniel Whiteson. Parameterized machine learning for high-energy physics. *arXiv preprint arXiv:1601.07913*, 2016.

[2] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[3] Filippo Palumbo, Claudio Gallicchio, Rita Pucci, and Alessio Micheli. Human activity recognition using multisensor data fusion based on reservoir computing. *Journal of Ambient Intelligence and Smart Environments*, 8(2):87–107, 2016.

[4] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.