Dimension reduction via score ratio matching

Ricardo Baptista[*]
California Institute of Technology
Pasadena, CA 91125 USA
rsb@caltech.edu

Michael C. Brennan[*] **Youssef Marzouk**
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
{mcbrenn,ymarz}@mit.edu

Abstract

We propose a method to detect a low-dimensional subspace where a non-Gaussian target distribution departs from a known reference distribution (e.g., a standard Gaussian). We identify this subspace from gradients of the log-ratio between the target and reference densities, which we call the *score ratio*. Given only samples from the target distribution, we estimate these gradients via score ratio matching, with a tailored parameterization and a regularization method that expose the low-dimensional structure we seek. We show that our approach outperforms standard score matching for dimension reduction of in-class distributions, and that several benchmark UCI datasets in fact exhibit this type of low dimensionality.

1 Introduction and motivation

Dimension reduction methods are ubiquitous in large-scale data science, statistical modeling, and machine learning. The computational burdens of many common analyses and algorithms may scale poorly with the dimension of the problem, and accurately capturing complex dependence structure in high-dimensional problems may require massive sample sizes. Many such tasks involve characterizing an unknown target distribution π given only a representative set of samples $\{x_i\}_{i=1}^n \sim \pi$. Both generative modeling and density estimation are examples of these tasks [9, 15].

When one has access to the (unnormalized) target density and its gradients, dimension reduction methods are well-explored [19, 5, 20, 1, 2]. One such method, certified dimension reduction [20], uses gradients of the log-density to expose a low-dimensional subspace where the target distribution departs most strongly from a reference distribution. In the context of Bayesian inference, this structure has been used to accelerate MCMC sampling methods [3, 4] or flow-based variational approximations [2], with error guarantees. Here, the target is the posterior distribution and a natural choice for the reference is the prior distribution.

In this work, we develop an analogous dimension reduction method for when one is given a set of samples $\{x_i\}_{i=1}^n \sim \pi$, but the target density is unavailable. We propose an algorithm for uncovering this low-dimensional structure based on score *ratio* matching (§3), rather than direct score matching. Our algorithm employs a training objective, a network parameterization, and a regularization penalty all tailored to our dimension reduction goal. We demonstrate that our score ratio matching method better reveals low-dimensional structure compared to standard score matching, and that several common benchmark datasets in fact exhibit this kind of low-dimensional structure (§4).

2 Background

Score matching overview Score matching has recently appeared as a powerful framework with applications to generative modeling [16, 17, 10, 6] and Bayesian inference [21, 13]. The core task

[*] Authors contributed equally to this work.

Score-based Method Workshop at Conference on Neural Information Processing Systems (NeurIPS 2022).

is approximating the *score function* $\nabla_x \log \pi(x)$ with a neural network $s_\theta(x)$, referred to as the score network [17]. We direct readers to [16, 17] for an overview of score matching, especially the derivation of the objective function for learning the score, network training strategies, and its application to generative modeling using Langevin sampling.

Low-dimensional subspace hypothesis We begin by defining a class of target distributions that depart from a known reference distribution only in a few directions.

Definition 1. *Let ρ be a chosen reference density on $\mathbb{R}^d$. Given a unitary matrix $U \in \mathbb{R}^{d\times d}$ and an integer $r \leq d$, let $\mathcal{D}_r(U)$ be the set of distributions with densities of the form*

$$\pi_r(x) \propto f(U_r^\top x)\rho(x),$$

for some $f\colon \mathbb{R}^r \to \mathbb{R}_{>0}$, where $U_r \in \mathbb{R}^{d\times r}$ contains the first r columns of U.

Such structure can be exploited in several ways. Let $U = [U_r\ U_\perp]$ be unitary and denote $x_r = U_r^\top x$ and $x_\perp = U_\perp^\top x$. Then, the target distribution rotated into the basis of U decomposes as

$$\pi(x_r, x_\perp) \propto f(x_r)\rho_r(x_r)\rho_{\perp|r}(x_\perp|x_r) = \nu_r(x_r)\rho_{\perp|r}(x_\perp|x_r)$$

where $\nu_r(x_r) := f(x_r)\rho_r(x_r)$ and $\rho_{\perp|r}$ is known (and can be evaluated and simulated). Hence, generative modeling and density estimation tasks are reduced to learning the r-dimensional distribution ν_r using projected samples $\{U_r^\top x_i\}_{i=1}^n$. Leveraging such structure requires identifying *both* the basis U and the minimal reduced dimension r needed to accurately approximate π within the set $\mathcal{D}_r(U)$.

Methods that exploit this structure are popular in Bayesian inference, where π is the posterior distribution. There it is typical to take ρ to be the prior distribution; f is then an approximation to the likelihood function. This assumption is natural in situations where we expect the data to only be partially informative of the parameter. Such structure has been leveraged in both MCMC methods [5, 3] and variational inference [2].

In this work, we take ρ to be the standard Gaussian density $\mathcal{N}(0, I_d)$ and identify a subspace where π departs from ρ. We note that our results can be generalized to other reference distributions; see Appendix A. A similar structure was exploited for diffusion models in [11] by hand-selecting the low-dimensional subspace. Here, we find this subspace by measuring the error incurred from approximating π with some $\pi_r \in \mathcal{D}_r(U)$ for a given basis U and reduced dimension r. The following results provide a strategy to find a suitable U and r based on an upper bound for the KL divergence from π to its closest approximation within $\mathcal{D}_r(U)$.

Proposition 1 (Modified Proposition 2.12 of [20])**.** *Let ρ be the standard Gaussian density, and let*

$$H = \mathbb{E}_\pi \left[\nabla_x \log\left(\frac{\pi(x)}{\rho(x)}\right) \nabla_x \log\left(\frac{\pi(x)}{\rho(x)}\right)^\top \right] \tag{1}$$

be the so-called diagnostic matrix of size $d \times d$. Then,

1. *For any $r \leq d$ and unitary matrix $U \in \mathbb{R}^{d\times d}$ there exists $\pi_r \in \mathcal{D}_r(U)$ such that*

$$\mathcal{D}_{\mathrm{KL}}(\pi||\pi_r) \leq \frac{1}{2}\operatorname{tr}(U_r U_r^\top H) =: E_r(U). \tag{2}$$

2. *Let $(\lambda_i, u_i) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^d$ be the i-th eigenpair of the eigenvalue problem $Hu_i = \lambda_i u_i$, $\lambda_1 \geq \cdots \geq \lambda_d$ and take $U = [u_1, \ldots, u_d]$ to be the matrix containing the eigenvectors of H. Then, $E_r(U)$ is minimized for any $r \leq d$, and there exists $\pi_r \in \mathcal{D}_r(U)$ such that*

$$\mathcal{D}_{\mathrm{KL}}(\pi||\pi_r) \leq \frac{1}{2}(\lambda_{r+1} + \cdots + \lambda_d). \tag{3}$$

The results of Proposition 1 have several practical implications. First, given the ability to compute the *diagnostic matrix* H, one obtains an upper bound on the error (in KL divergence) induced by approximating π with a distribution in the class $\mathcal{D}_r(U)$, for any choice of U. Second, a natural choice for U is the eigenbasis of H, and one may choose r based on the decay of the eigenvalues of H; that is, given a KL error tolerance $\varepsilon > 0$, one can pick r so that $\frac{1}{2}(\lambda_{r+1} + \cdots + \lambda_d) < \varepsilon$. Indeed, if the rank of H is r, then $\pi \in \mathcal{D}_r(U)$ and thus the marginal of $x_\perp$ must be standard Gaussian.

3 Score ratio matching

We now describe how to approximate the score ratio function $\nabla_x \log(\pi(x)/\rho(x))$ using *score ratio matching*, enabling us to perform analogous dimension reduction given only samples x_i from π.

A naïve strategy would be to directly approximate the score of π and use it to compute the score ratio as $\nabla_x \log(\pi(x)/\rho(x)) = \nabla_x \log \pi(x) - \nabla_x \log \rho(x)$. Instead, we take a different approach that leverages the (possible) low-dimensional structure of π in Definition 1. We approximate the score ratio *directly* using a *score ratio network* $s_\theta \colon \mathbb{R}^d \to \mathbb{R}^d$ by minimizing an objective function that does not require access to the score of the target density π. This is made possible by the following proposition, whose proof is in Appendix B.

Proposition 2. *Let s_θ be differentiable. Then we have the following equivalence of objectives:*

$$\frac{1}{2}\mathbb{E}_\pi \left\| s_\theta(x) - \nabla_x \log\left(\frac{\pi(x)}{\rho(x)}\right)\right\|_2^2 = \mathbb{E}_\pi\left[\frac{1}{2}s_\theta(x)^\top s_\theta(x) + \operatorname{tr}(\nabla_x s_\theta(x)) + \nabla_x \log \rho(x)^\top s_\theta(x)\right] + C$$

where C is a constant that only depends on the densities π and ρ.

In practice, we replace the expectation above with the empirical sum over the dataset to obtain a key term of our optimization objective,

$$J(s_\theta) \coloneqq \sum_{i=1}^{n} \frac{1}{2}s_\theta(x_i)^\top s_\theta(x_i) + \operatorname{tr}(\nabla_x s_\theta(x_i)) + \nabla_x \log \rho(x_i)^\top s_\theta(x_i).$$

Under our hypothesis on the target density in Definition 1, we expect the score ratio, rather than the score itself, to be well approximated by a *ridge function* [14], i.e., a function that is constant for $x \in \operatorname{Im}(U_\perp)$. Next, we describe a parameterization for $s_\theta(x)$ and a regularization method that are tailored to learning this low-dimensional structure.

Score-ratio network parameterization and regularization For $\pi_r \in \mathcal{D}_r(U)$, the score ratio takes the specific form

$$\nabla_x \log\left(\frac{\pi_r(x)}{\rho(x)}\right) = U_r \nabla \log f(U_r^\top x).$$

We see that the range of the score ratio lies within the subspace spanned by U_r. We encode this observation into our score ratio network in two ways. First, we parameterize the network as

$$s_\theta(x) = V \tilde{s}_\theta(V^\top x)$$

where $V \in \mathbb{R}^{d\times d}$, is the first and last layer's weight matrix and $\tilde{s}_\theta \colon \mathbb{R}^d \to \mathbb{R}^d$ is a typical score network as described in [17]. This parameterization enforces that if V converges to a low (effective) rank during optimization, the range of $s_\theta(x)$ is restricted accordingly.

We also use a regularization technique that helps *reveal* low-dimensional structure when it is present. If Definition 1 holds, then we expect V to have (numerical) rank r. As r is unknown, we penalize the nuclear norm of $VV^\top$, as commonly used for low-rank matrix estimation [8]. This leads to the final objective function

$$F(s_\theta) = J(V\tilde{s}_\theta \circ V^\top) + \lambda \|VV^\top\|_*$$

where $\|\cdot\|_*$ is the nuclear norm.

Algorithm 1 Estimate low-dimensional subspace U_r

```
1: Input: Target data $\{x_i\}_{i=1}^n \sim \pi$, and user tolerance $\varepsilon > 0$
2: Center the mean and scale data by the Cholesky factor of the empirical precision matrix.
3: Solve $\min_{s_\theta} F(s_\theta)$ to obtain the score-ratio approximation $s_\theta(x)$.
4: Estimate the diagnostic matrix $\widehat{H} = \frac{1}{n}\sum_{i=1}^n s_\theta(x_i)s_\theta(x_i)^\top$.
5: Compute the eigenpairs of $\widehat{H}$, $(\lambda_i, u_i) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^d$.
6: Set $U = [u_1 \ \dots \ u_n]$ and pick $r$ so that $\widehat{E}_r(U) = \frac{1}{2}(\lambda_{r+1} + \cdots + \lambda_d) < \varepsilon$
```

4 Numerical results

We now present several numerical experiments showing that: (i) our score ratio method more accurately captures the relevant low-dimensional subspace in a toy problem where this structure is known to be present, and (ii) this structure is present in several datasets from the UCI repository [7]. Details on the score network parameterization and training procedure are in Appendix C.

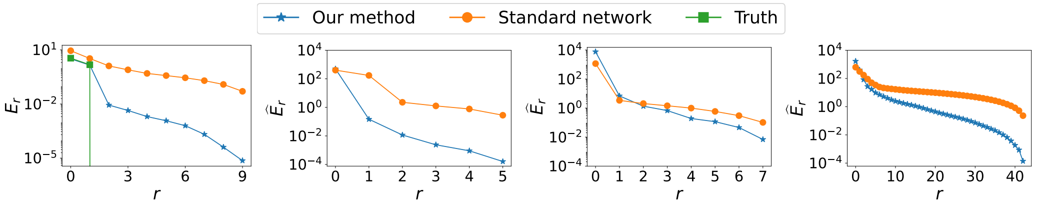
Embedded banana distribution Consider the following "embedded banana" distribution, where the data-generating process is defined by

$$y_1 \sim \mathcal{N}(0,1), \quad y_2 \sim \mathcal{N}(y_1^2, 1), \quad y_{3:10} \sim \mathcal{N}(0, I), \tag{4}$$

and $x = Ry$, where $R \in \mathbb{R}^{10\times 10}$ is a random rotation matrix that is sampled by computing the QR factorization of a matrix with standard Gaussian entries. In this case, we have $\pi \in \mathcal{D}_{r=2}(R)$. Hence, we expect our algorithm to find the subspace spanned by the two leading columns of R.

In this example, we compute the score ratio analytically and define a consistent estimator for the true diagnostic matrix H. In Figure 1a we plot the error bound $E_r(U) = \frac{1}{2}\operatorname{tr}(U_r U_r^\top H)$ for three different bases U: (1) the eigenbasis of the true diagnostic matrix; (2) the eigenbasis of the diagnostic matrix computed with our score ratio approximation; and (3) the eigenbasis of the diagnostic matrix computed with a standard score approximation (as described at the beginning of §3). For our method, we see that $E_r(U)$ sharply drops at $r = 2$ to less than 10^{-2}. We also see that our method yields considerably lower errors at each r compared to standard score matching. For a visual representation of the results, Appendix C shows a scatter plot of additional held-out samples from π (which were not used during training) and the samples rotated into our discovered basis U when taking $d = 3$.

UCI datasets We now report results for several datasets from the UCI repository [7], commonly used as benchmarks for density estimation: POWER, GAS, and MINIBOONE. Since we take the reference distribution to be standard normal, before applying our algorithm we whiten the data; see Appendix A. For these datasets we do not have access to an analytic score function, and thus can only report the estimated KL error bounds, $\widehat{E}_r(U)$, for each example. In Figure 1b–d, we see that low-dimensional structure seems to be present in each dataset, via the rapid decay in the error bound for small r. For example, an $r = 1$ dimensional subspace for the POWER dataset yields an approximation error on the order of 10^{-1}.



(a) Banana, $d = 10$ (b) POWER, $d = 6$ (c) GAS, $d = 8$ (d) MINIBOONE, $d = 43$

Figure 1: Upper bounds on the KL divergence as a function of the subspace dimension r for the embedded banana distribution, and the POWER, GAS, and MINIBOONE datasets. For the embedded banana distribution, we plot the error bound computed with the true diagnostic matrix, E_r, as an analytic score ratio is available. For the UCI datasets, we report the estimated error bound $\widehat{E}_r$.

5 Discussion

We have proposed a dimension reduction methodology based on score matching. Our framework identifies a subspace that best captures the departure of the data-generating distribution from a reference distribution. While such methods are well studied in the context of Bayesian inference, our approach brings the benefits of dimension reduction with error guarantees to settings where only samples are available. To aid in finding low dimensional structure, we introduced a network parameterization that exploits the score ratio's gradient structure, coupled with a low-rank matrix recovery technique. Future work will utilize the proposed framework to accelerate and improve the accuracy of density estimation and generative modeling by exploiting the discovered low-dimensional structure. We also plan to investigate how the intrinsic dimension of the problem affects the number of samples needed to learn the score ratio and its hyperparameters.

Acknowledgments and Disclosure of Funding

This work was supported in part by the DOE Office of Advanced Scientific Computing Research (ASCR) M2dt MMICC center.

References

[1] Ricardo Baptista, Youssef Marzouk, and Olivier Zahm. “Gradient-based data and parameter dimension reduction for Bayesian models: an information theoretic perspective”. In: *arXiv preprint arXiv:2207.08670* (2022).

[2] Michael Brennan et al. “Greedy inference with structure-exploiting lazy maps”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 8330–8342.

[3] Tiangang Cui, Kody JH Law, and Youssef Marzouk. “Dimension-independent likelihood-informed MCMC”. In: *Journal of Computational Physics* 304 (2016), pp. 109–137.

[4] Tiangang Cui and Xin T Tong. “A unified performance analysis of likelihood-informed subspace methods”. In: *Bernoulli* 28.4 (2022), pp. 2788–2815.

[5] Tiangang Cui et al. “Likelihood-informed dimension reduction for nonlinear inverse problems”. In: *Inverse Problems* 30.11 (2014), p. 114015.

[6] Valentin De Bortoli et al. “Diffusion Schrödinger bridge with applications to score-based generative modeling”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 17695–17709.

[7] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: `http://archive.ics.uci.edu/ml`.

[8] Maryam Fazel. “Matrix rank minimization with applications”. PhD thesis. PhD thesis, Stanford University, 2002.

[9] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[11] Bowen Jing et al. “Subspace diffusion generative models”. In: *arXiv preprint arXiv:2205.01490* (2022).

[12] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[13] Lorenzo Pacchiardi and Ritabrata Dutta. “Score Matched Neural Exponential Families for Likelihood-Free Inference.” In: *J. Mach. Learn. Res.* 23 (2022), pp. 38–1.

[14] Allan Pinkus. *Ridge functions*. Vol. 205. Cambridge University Press, 2015.

[15] Lars Ruthotto and Eldad Haber. “An introduction to deep generative modeling”. In: *GAMM-Mitteilungen* 44.2 (2021), e202100008.

[16] Yang Song and Stefano Ermon. “Generative modeling by estimating gradients of the data distribution”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[17] Yang Song et al. “Score-based generative modeling through stochastic differential equations”. In: *arXiv preprint arXiv:2011.13456* (2020).

[18] Yang Song et al. “Sliced score matching: A scalable approach to density and score estimation”. In: *Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 574–584.

[19] Alessio Spantini et al. “Optimal Low-rank Approximations of Bayesian Linear Inverse Problems”. In: *SIAM Journal on Scientific Computing* 37.6 (2015), A2451–A2487.

[20] Olivier Zahm et al. “Certified dimension reduction in nonlinear Bayesian inverse problems”. In: *Mathematics of Computation* 91.336 (2022), pp. 1789–1835.

[21] Cheng Zhang, Babak Shahbaba, and Hongkai Zhao. “Variational Hamiltonian Monte Carlo via score matching”. In: *Bayesian Analysis* 13.2 (2018), pp. 485–506.

A Choice of reference distribution and data pre-processing

In this work we choose the reference distribution ρ to be a standard Gaussian, though many other choices are possible. The results in Proposition 1 depend on ρ satisfying a log-Sobolev inequality, which allows for uniform, multivariate Gaussian, and mixture-of-Gaussian reference distributions, among others; see [20] for more details and examples.

In practice, ρ should be taken to be comparable to the samples in location and scale. To implement the algorithm presented in Section 3, it is only required that the score of the reference be known and easily computable.

Since here the reference distribution chosen here is standard normal, we whiten the data before applying our algorithm to datasets from the UCI repository, i.e., we standardize the data to have zero mean and identity covariance matrix. Specifically, we shift the data by its sample mean and transform it by the square root of a sample estimate of the precision matrix. This ensures that the data can be meaningfully compared with the standard normal.

B Proof of Proposition 2

Let

$$J^*(\theta) = \frac{1}{2}\mathbb{E}_\pi \left\| s_\theta(x) - \nabla_x \log\left(\frac{\pi(x)}{\rho(x)}\right)\right\|_2^2 .$$

We expand the squared norm to obtain

$$J^*(s_\theta) = \frac{1}{2}\mathbb{E}_\pi \left[s_\theta(x)^\top s_\theta(x) + \nabla_x \log\left(\frac{\pi(x)}{\rho(x)}\right)^\top \nabla_x \log\left(\frac{\pi(x)}{\rho(x)}\right) - 2 s_\theta(x)^\top \nabla_x \log\left(\frac{\pi(x)}{\rho(x)}\right)\right].$$

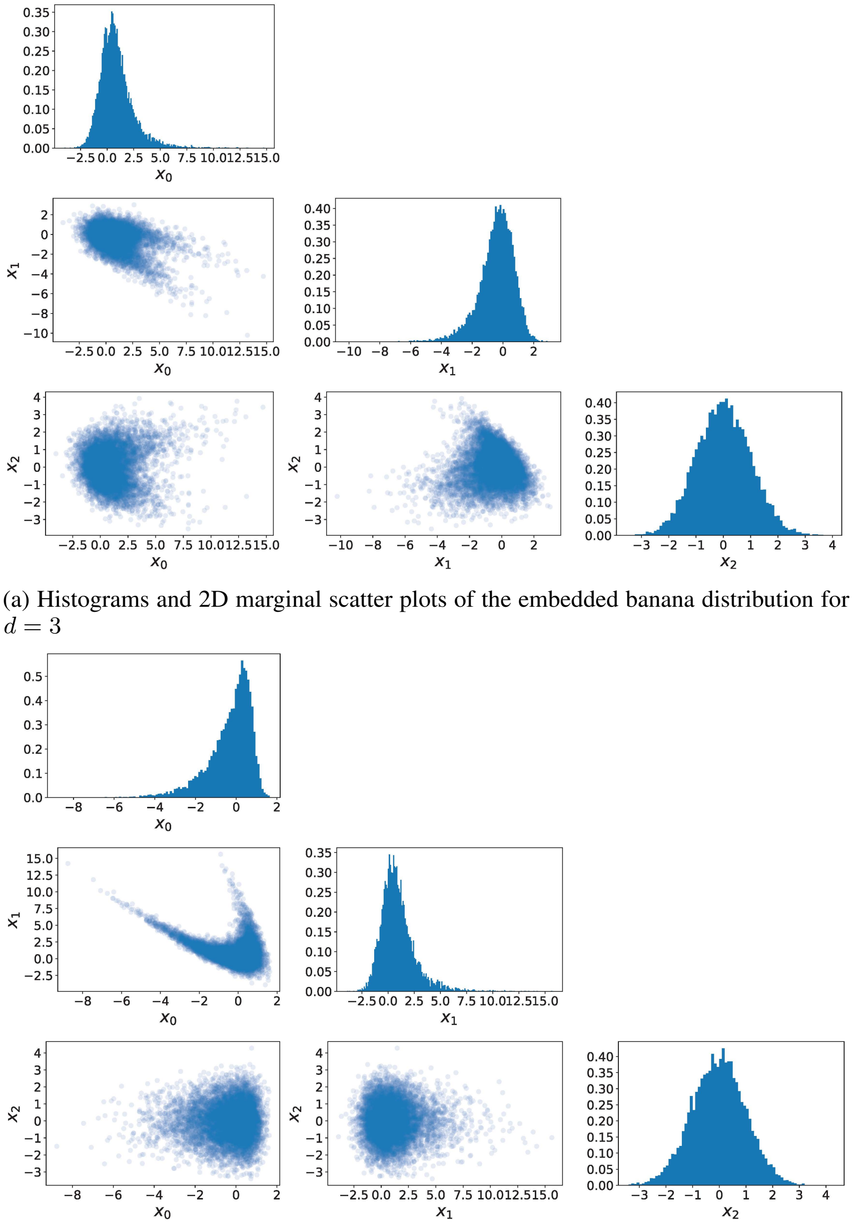
Note the second term, $\nabla_x \log\left(\frac{\pi(x)}{\rho(x)}\right)^\top \nabla_x \log\left(\frac{\pi(x)}{\rho(x)}\right)$, does not depend on the network parameters, and thus need not be included in our optimization objective. The following steps rewrites the third term into quantities we can evaluate:

$$\begin{aligned}
\mathbb{E}_\pi \left[s_\theta(x)^\top \nabla_x \log\left(\frac{\pi(x)}{\rho(x)}\right)\right] &= \mathbb{E}_\pi \left[s_\theta(x)^\top \nabla_x \left(\frac{\pi(x)}{\rho(x)}\right) \frac{\rho(x)}{\pi(x)}\right] \\
&= \int \pi(x) s_\theta(x)^\top \nabla_x \left(\frac{\pi(x)}{\rho(x)}\right) \frac{\rho(x)}{\pi(x)} \,\mathrm{d}x \\
&= \int \rho(x) s_\theta(x)^\top \nabla_x \left(\frac{\pi(x)}{\rho(x)}\right) \,\mathrm{d}x \\
&= \int \mathrm{tr}(\nabla_x(\rho(x) s_\theta(x))^\top \left(\frac{\pi(x)}{\rho(x)}\right) \,\mathrm{d}x \\
&= \int \pi(x) \left[\frac{\nabla_x \rho(x)}{\rho(x)} s_\theta(x) + \mathrm{tr}(\nabla_x s_\theta(x))\right] \\
&= \mathbb{E}_\pi \nabla_x \log \rho(x)^\top s_\theta(x) + \mathrm{tr}(\nabla_x s_\theta(x)).
\end{aligned}$$

This leads to the final result that

$$\begin{aligned}
J^*(s_\theta) = \mathbb{E}_\pi &\left[\frac{1}{2} s_\theta(x)^\top s_\theta(x) + \mathrm{tr}(\nabla_x s_\theta(x)) + \nabla_x \log \rho(x)^\top s_\theta(x)\right] \\
&+ \mathbb{E}_\pi \left[\nabla_x \log\left(\frac{\pi(x)}{\rho(x)}\right)^\top \nabla_x \log\left(\frac{\pi(x)}{\rho(x)}\right)\right].
\end{aligned}$$

C Implementation details and additional numerical results



(a) Histograms and 2D marginal scatter plots of the embedded banana distribution for $d = 3$

(b) Histograms and 2D marginal scatter plots of the embedded banana distribution in the basis U learned by our score-ratio matching method. Non-Gaussianity has been concentrated in the first two directions, and the third direction is now essentially independent of the first two.

Figure 2: Histograms and scatter plots of held-out samples from the embedded banana distribution before (a) and after (b) rotation by the learned basis U.

For each numerical example, we used 10^4 training samples. We use the Adam [12] optimizer with learning rate 5×10^{-3} and batch size 1000 to train the network for 500 epochs. We take the nuclear norm regularization parameter to be $\lambda = 0.8/d$. As discussed in [18], directly evaluating the trace operator in the objective function F is prohibitively expensive for even moderate dimensions d, and so we also make use of the sliced-score matching method proposed in that work with 100 projections. The network $\tilde{s}_\theta$ had 1 fully connected hidden layer for the embedded banana example, and 2 hidden layers for the UCI datasets with ReLU activation functions and width 128.