3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–9

Bijective-Contrastive Estimation

Jae Hyun Lim* JAE.HYUN.LIM@UMONTREAL.CA
Mila, Université de Montréal

Chin-Wei Huang* CHIN-WEI.HUANG@UMONTREAL.CA
Mila, Université de Montréal

Aaron Courville AARON.COURVILLE@UMONTREAL.CA
Mila, Université de Montréal & CIFAR fellow

Christopher Pal CHRISTOPHER.PAL@POLYMTL.CA
Mila, Polytechnique Montréal & Canada CIFAR AI Chair

Abstract

In this work, we propose *Bijective-Contrastive Estimation* (BCE), a classification-based learning criterion for energy-based models. We generate a collection of contrasting distributions using bijections, and solve all the classification problems between the original data distribution and the distributions induced by the bijections using a classifier parameterized by an energy model. We show that if the classification objective is minimized, the energy function will uniquely recover the data density up to a normalizing constant. This has the benefit of not having to explicitly specify a contrasting distribution, like noise contrastive estimation. Experimentally, we demonstrate that the proposed method works well on 2D synthetic datasets. We discuss the difficulty in high dimensional cases, and propose potential directions to explore for future work.

1. Introduction

Training energy-based models (EBM) via maximum likelihood estimation (MLE) is difficult. This is because likelihood evaluation under the density of the EBM requires computing the log-partition function, which involves an intractable integral. For gradient-based optimization, one often needs to resort to *Markov chain Monte Carlo* (MCMC) methods in order to approximately sample from the energy model to estimate the gradients (Hinton, 2002). However, MCMC methods are notoriously computationally expensive and hard to tune, which prevents EBMs from being widely adopted among the practitioners. This computational bottleneck has therefore motivated the design of alternative training objectives of EBMs, such as score matching (SM, Hyvärinen (2005)), denoising score matching (DSM, Vincent (2011)), noise-contrastive estimation (NCE, Gutmann and Hyvärinen (2010)), and Stein discrepancy minimization (Liu et al., 2016; Grathwohl et al., 2020).

In our attempt to devise a new training criterion for EBMs, we draw inspiration from the recent development of flow-based methods. Notably, Dinh et al. (2019); Nielsen et al. (2020) demonstrate that the inverse model of a surjection map recovers the internal structure of the data distribution to a certain degree. In order to reconstruct the data distribution with

* Equal contributions

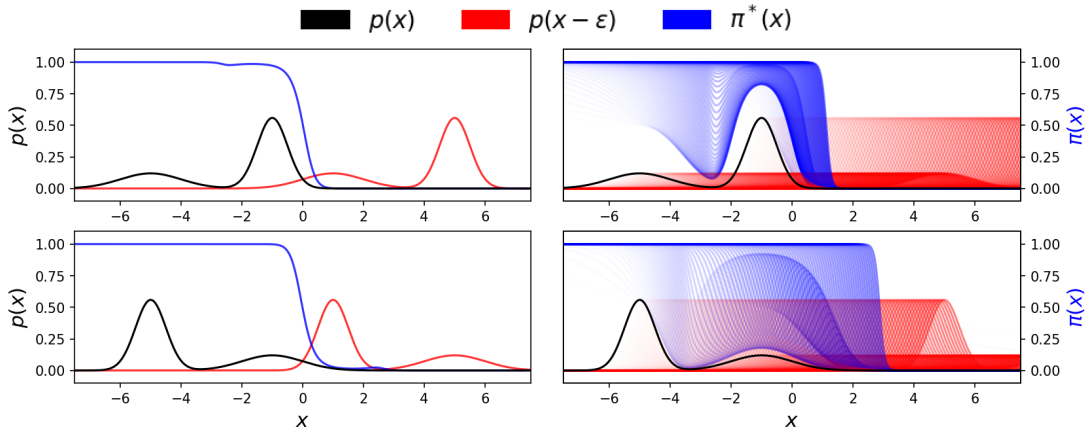

Figure 1: Examples of binary classification problems generated by bijections. Left: while two data densities are different (black solid lines in top and bottom-left), we have very similar Bayes optimal classifiers to differentiate $p(x)$ and $p(x-\epsilon)$. Right: a collection of optimal classifiers is fully governed by corresponding pdfs.

a higher precision, a chain of bijection and surjection maps needs to be composed, resulting in solving multiple classification problems in sequel.

Motivated by this, we propose to use bijections to generate a collection of classification problems and solve them by parameterizing the classifiers using a shared energy function. We refer to the corresponding parameter estimation principle as *Bijective-Contrastive Estimation* (BCE). As the EBM is trained to contrast the distributions induced by these bijections, it will learn to recover the ratio of the corresponding density functions. By jointly solving a sufficiently large collection of classification problems with shared parameters, the EBM is guaranteed to recover the data distribution's density up to a normalizing constant.

2. Bijective-Contrastive Estimation (BCE)

2.1. Generate Contrasting Distributions via Bijections

In this section, we first show how to generate a binary classification problem using a bijection and demonstrate how to solve the classification problem with an energy model. We then conjecture that while using the energy model to solve a single classification problem will not be enough to recover the data density, we can do so by jointly solving sufficiently many classification problems.

Let x be a random variable following the density $p(x)$, and define $x' = x + \epsilon$, where ϵ is a constant. Then the density of x' is equal to $p(x' - \epsilon)$; see Figure 1. We consider a binary classification problem between these two distributions, which admits an optimal solution $\pi^*(x) = \frac{p(x)}{p(x)+p(x-\epsilon)}$ (i.e. the blue curve). This implies we can plug in a parametric density, or an unnormalized density using an energy model in place of $p(x)$ as well as $p(x-\epsilon)$, and hope to recover $p(x)$ by training the classifier. Nevertheless, we can't recover $p(x)$ solely from π^*. This is because different density functions $p(x)$ can possibly generate very similar optimal solution. For example, the left hand side of Figure 1 shows two possible $p(x)$'s that

result in almost indistinguishable π^*'s for the same ϵ. While the optimal classifier generated by a single bijection may not be distinguishable for different density functions, we conjecture that the family of optimal classifiers generated by a sufficiently large collection of bijections will be uniquely dependent on the density function. Consider the same classification task above, but now with multiple ϵ's. In Figure 1 (right), we plot all the $p(x-\epsilon)$'s (as well as the respective optimal classifiers π^*) corresponding to different values of ϵ; the transparency of the curves reflects the magnitude of the ϵ value. Unlike the single classification case, we can now easily tell the two families of π^*'s apart.

This reveals that with an increasing number of ϵ's, there will be less available densities that match all of the optimal classifiers of the generated problems. With a sufficiently large number of classification problems, we hope to rule out all but one density function – the true density of the data distribution. This would imply that we can learn the data density by parameterizing the classifiers as a function of the energy and the bijections, and jointly solving the classification problems. In the next section, we formalize and generalize these examples, and propose a new classification-based learning principle for EBMs.

2.2. Bijective-Contrastive Estimation

Let $x \in \mathbb{R}^d \sim p_{\texttt{data}}(x)$ be a random variable representing the data, and $t : \mathbb{R}^d \to \mathbb{R}^d$ be a bijection. The probability density of $x_t := t(x)$ can be conveniently expressed as $p(x_t) = p_{\texttt{data}}(t^{-1}(x_t))|J_{t^{-1}}(x_t)|$ by the change-of-variable formula, where $J_f(x)$ denotes the Jacobian matrix of f wrt x. Similarly to the examples in the previous section, let us first consider a classification problem to differentiate between x and x_t for fixed t when an observation is drawn from $p_{\texttt{data}}(x)$ or $p(x_t)$ with equal probability. Let $\pi(x)$ be a probabilistic classifier $\pi : \mathbb{R}^d \to [0, 1]$, solving the following objective:

$$\mathcal{L}(\pi; t) \coloneqq -\mathop{\mathbb{E}}_{x}[\log \pi(x)] - \mathop{\mathbb{E}}_{x_t}[\log(1 - \pi(x_t))] = -\mathop{\mathbb{E}}_{x}[\log \pi(x) + \log(1 - \pi(t(x)))]. \tag{1}$$

The objective has a functional minimizer

$$\pi^*(x) = \frac{p_{\texttt{data}}(x)}{p_{\texttt{data}}(x) + p_{\texttt{data}}(t^{-1}(x))|J_{t^{-1}}(x)|}.$$

That is, $\mathcal{L}(\pi^*; t) \leq \mathcal{L}(\pi; t)$ for any π. Let $p(x; f) = e^{-f(x)}/Z(f)$, where f is the energy model and $Z(f)$ is the normalizing constant. We can parameterize the classifier as

$$\pi(x; f) \coloneqq \frac{p(x; f)}{p(x; f) + p(t^{-1}(x); f)|J_{t^{-1}}(x)|} = \frac{e^{-f(x)}}{e^{-f(x)} + e^{-f(t^{-1}(x))}|J_{t^{-1}}(x)|}. \tag{2}$$

Plugging $\pi(x; f)$ into Equation (1), we can rewrite $\mathcal{L}$ as a loss functional of f

$$\mathcal{L}(f; t) = -\mathop{\mathbb{E}}_{x}\left[\log \frac{e^{-f(x)}}{e^{-f(x)} + e^{-f(t^{-1}(x))}|J_{t^{-1}}(x)|} + \log \frac{e^{-f(x)}}{e^{-f(t(x))}|J_t(x)| + e^{-f(x)}}\right]. \tag{3}$$

As discussed in the previous section, we conjecture that f^* satisfying $e^{-f^*(x)} \propto p_{\texttt{data}}(x)$ will be the only minimizer of $\mathcal{L}(f; t)$ for a sufficiently large collection of t. Motivated from this, we propose to minimize an expected loss over a distribution of bijections:

$$\mathcal{L}_{\texttt{BCE-b}}(f) = -\mathop{\mathbb{E}}_{t}[\mathcal{L}(f; t)]. \tag{4}$$

We refer to learning energy functions by minimizing Equation (4) (expectation taken over $p_{\texttt{data}}$ or an empirical distribution) as **Bijective-Contrastive Estimation (BCE)**. We call the binary classification form of Equation (4) *binary BCE objective*. Let $p(t)$[1] denote the density of the bijection t. The following theorem provides a positive answer to our conjecture: under some mild assumption on $p(t)$ and the form of t, the minimizer f^* of the BCE loss is unique up to a normalizing constant and $e^{-f^*(x)}$ is proportional to the data density.

Theorem 1 (Uniqueness of BCE-based estimator) *Let $\mathcal{F}$ be the set of measurable functions. For any $x \in \text{supp}(p_{\texttt{data}}(x))$, define $g_x(t) = t^{-1}(x)$. If for any $x \in \text{supp}(p_{\texttt{data}}(x))$ and for any $p(t)$-almost sure set M, $g_x(M)^c$ has Lebesgue measure zero, then*

$$p_{\texttt{data}}(x) \propto e^{-f^*(x)} \text{ almost everywhere} \iff \mathcal{L}_{\texttt{BCE-b}}(f^*) \leq \mathcal{L}_{\texttt{BCE-b}}(f) \quad \forall f \in \mathcal{F}. \tag{5}$$

Corollary 2 (Additive BCE) *If $t(x) \coloneqq x + \epsilon$ where $\epsilon \in \mathbb{R}^d$ is a random variable with a density function $p(\epsilon)$ that is non-zero everywhere, then the minimizer f^* of the BCE objective in Equation (4) satisfies $p_{\texttt{data}}(x) \propto e^{-f^*(x)}$ almost everywhere.*

The proofs are deferred to Appendix A. Theorem 1 shows that BCE is powerful enough to recover $p_{\texttt{data}}$, up to a normalizing constant. Moreover, according to Corollary 2, this holds true for very simple classes of bijections and distributions; for example, additive noise with a standard normal distribution.

2.3. Variants of objective functions

For additive bijections, we observe that t and t^{-1} have the same density if $p(\epsilon)$ is standard normal. In this paper, if the distribution of bijections satisfies $p(t) = p(t^{-1})$, we informally refer to this property as *inverse-symmetry*. Inverse-symmetry allows us to simplify Equation (4) to (neglecting a multiplier of 2)

$$\mathcal{L}_{\texttt{BCE-b}}(f) = -\mathop{\mathbb{E}}_{x,t}\left[\log \frac{e^{-f(x)}}{e^{-f(x)} + e^{-f(t(x))}|J_t(x)|}\right]. \tag{6}$$

In addition to additions, random permutations and matrix-vector product of data with orthogonal matrices are also simple bijections that are inverse-symmetric.

Inverse-symmetry also largely simplifies the computation of the multiclass classification loss, which would normally have quadratically many terms, to be

$$\mathcal{L}_{\texttt{BCE-m}}(f) = -\mathop{\mathbb{E}}_{x,t_1,\ldots,t_n}\left[\log \frac{e^{-f(x)}}{e^{-f(x)} + \sum_{i=1}^{n} e^{-f(t_i(x))}|J_{t_i}(x)|}\right], \tag{7}$$

where $t_1, \ldots, t_n \sim p(t)$ (neglecting a multiplier of n). We find that training with multiclass BCE tends to be more stable as it has smaller gradient variance.

1. In this article, we assume that the bijections have real-valued representations. For convenience, we abuse notations and refer to $p(t)$ as a density function on the real-valued representations of t's. When we write $y = t(x)$, we refer to the functional form of t that maps x to y.

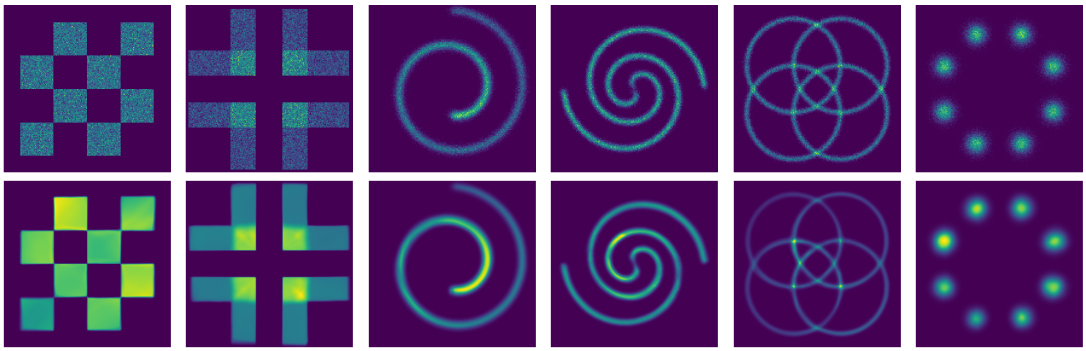

Figure 2: Density estimation on 2D synthetic datasets. First row: 2D histogram of synthetic datasets. Second row: unnormalized density functions learned by BCE.

2.4. Relation to noise contrastive estimation

The most closely related work to ours is noise-contrastive estimation (NCE, Gutmann and Hyvärinen (2010)), which uses a tractable noise distribution $q(x)$ and learns the energy function via the following objective,

$$\mathcal{L}_{\texttt{NCE}}(f) = -\mathop{\mathbb{E}}_{x\sim p_{\texttt{data}}}\left[\log \frac{e^{-f(x)}}{e^{-f(x)}+q(x)}\right] - \mathop{\mathbb{E}}_{x\sim q}\left[\log \frac{q(x)}{e^{-f(x)}+q(x)}\right].$$

If $q(x)$ is defined to be nonzero wherever $p_{\texttt{data}}(x)$ is nonzero, then $e^{-f^*(x)} = p_{\texttt{data}}(x)$ if f^* minimizes the NCE objective[2]. While NCE contrasts against a pre-defined noise distribution q, BCE uses bijections to generate many contrasting distributions implicitly.

3. Experiment on 2D synthetic datasets

In order to demonstrate that by minimizing the BCE loss, we can learn the correct data density, we run density estimation experiments on six 2D synthetic datasets. We use binary BCE loss as in Equation (6), additive bijections for the perturbation, and standard normal distribution for the additive bijections. We use a ReLU network with three 1,000-unit hidden layers, and train it using the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9, \beta_2 = 0.999$. We run 50K iterations. Mini-batch size and learning rate are set to 256 and 0.0001, respectively. The results are presented in Figure 2. We see that the energy-based model trained by BCE successfully learns the densities of a variety of 2D synthetic datasets.

4. Difficulties in modeling high-dimensional data

In preliminary experiments, we observe that the proposed method did not work well on high dimensional data. On the MNIST dataset, we train the EBM via BCE and evaluate the quality of samples generated by running Langevin dynamics (Grenander and Miller, 1994), but the generated samples all fail to resemble the true data samples. We conjecture that this is a result of using the cross entropy loss, which causes two types of vanishing gradients

2. In NCE, the normalizing constant is commonly absorbed in $f(x)$ as a parameter.

problems: the loss does not provide meaningful learning signals when $t(x)$ is either too far from or too close to the data manifold. The first case occurs when we use a simple family of bijections for BCE. For instance, adding an unstructured noise can easily push a highly structured data point off the data manifolds. Hence the classifiers can achieve near-perfect accuracy while the learned unnormalized density is nowhere near the true one. The gradients wrt the energy can also vanish when $p(t(x))$ is too close to the data density, *e.g.* when t is near identity.

Similarly to BCE, the first type of vanishing gradient problem also prevents NCE from scaling up, occurring when $q(x)$ is too far from $p_{\texttt{data}}(x)$. Recently, Rhodes et al. (2020) tackle this problem by generating a sequence of NCE problems, each of which contrasts two consecutive intermediate distributions interpolated between the data and a noise distribution. When the two distributions are close enough, one can potentially mitigate the vanishing gradient problem. In addition, generative adversarial networks (GANs, Goodfellow et al. (2014)) also suffer from similar vanishing gradient problems, which can be addressed by regularizing the discriminator (Arjovsky et al., 2017; Roth et al., 2017). These techniques can potentially be applied to BCE for modelling high-dimensional data.

5. Conclusion

In this paper, we propose a new classification-based EBM training method, called *Bijective-Contrastive Estimation* (BCE). We prove that the data density (up to a normalizing constant) is the only minimizer of the BCE loss. In the experiment, we show that the energy trained with BCE accurately models the 2D synthetic data distributions. We discuss the difficulty of BCE in modeling high-dimensional data due to vanishing gradients, and suggest future directions for addressing this issue.

6. Acknowledgments

We thank Samsung for funding this work.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Laurent Dinh, Jascha Sohl-Dickstein, Razvan Pascanu, and Hugo Larochelle. A rad approach to deep mixture models. *arXiv preprint arXiv:1903.07714*, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, and Richard Zemel. Learning the stein discrepancy for training and evaluating energy-based models without sampling. In *ICML*, 2020.

Ulf Grenander and Michael I Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(4):549–581, 1994.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *ICML*, 2016.

Didrik Nielsen, Priyank Jaini, Emiel Hoogeboom, Ole Winther, and Max Welling. Survae flows: Surjections to bridge the gap between vaes and flows. *arXiv preprint arXiv:2007.02731*, 2020.

Benjamin Rhodes, Kai Xu, and Michael U Gutmann. Telescoping density-ratio estimation. *arXiv preprint arXiv:2006.12204*, 2020.

Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *NIPS*, 2017.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Appendix A. Proofs

Proof of Theorem 1

Step 1 In order to prove Theorem 1, we first introduce another optimization problem equivalent to minimizing $\mathcal{L}_{\texttt{BCE-b}}(f)$ in Equation (4). Consider the following data generation process.

$$x = \begin{cases} \tilde{x}, & \text{if } y = 1, \\ t(\tilde{x}), & \text{otherwise.} \end{cases},$$

where $\tilde{x} \sim p_{\texttt{data}}(\tilde{x})$, $t \sim p(t)$, and $y \sim \text{Bernoulli}\left(\frac{1}{2}\right)$. Then the likelihood of x given y and t is $p(x|y=1,t) = p_{\texttt{data}}(x)$ or $p(x|y=0,t) = p_{\texttt{data}}(t^{-1}(x))|J_{t^{-1}}(x)|$ by the change of variable. Similarly, we define the "model" likelihood functions

$$q(x|y=1,t;f) \coloneqq \frac{e^{-f(x)}}{Z(f)} \quad \text{and} \quad q(x|y=0,t;f) \coloneqq \frac{e^{-f(t^{-1}(x))}|J_{t^{-1}}(x)|}{Z(f)}. \tag{8}$$

Then, the minimization of the BCE objective in Equation (4) is equivalent to minimizing the expected KL-divergence between the true posterior $p(y|x,t)$ and the model posterior $q(y|x,t)$:

$$\arg\min_f \mathcal{L}_{\texttt{BCE-b}}(f) = \arg\min_f \mathop{\mathbb{E}}_{x,t}\left[D_{KL}\left(p(y|x,t)\|q(y|x,t;f)\right)\right], \tag{9}$$

where

$$p(y|x,t) = \frac{p(x|y,t)}{p_{\texttt{data}}(x) + p_{\texttt{data}}(t^{-1}(x))|J_{t^{-1}}(x)|}$$

$$\text{and} \quad q(y|x,t;f) = \frac{q(x|y,t;f)}{\underbrace{\frac{e^{-f(x)}}{Z(f)}}_{=\,q(x|y=1,t;f)} + \underbrace{\frac{e^{-f(t^{-1}(x))}|J_{t^{-1}}(x)|}{Z(f)}}_{=\,q(x|y=0,t;f)}}.$$

The equality holds since $\mathcal{L}_{\texttt{BCE-b}}(f)$ is the cross entropy term in the expected KL and the entropy term of $p(y|x,t)$ does not depend on f.

Step 2 Due to Equation (9), proving Theorem 1 is equivalent to showing that if g_x maps any $p(t)$-almost sure set to a Lebesgue-almost everywhere set, then

$$p_{\texttt{data}}(x) \propto e^{-f^*(x)} \quad \text{almost everywhere} \iff \mathop{\mathbb{E}}_{x,t}\left[D_{KL}\left(p(y|x,t)\|q(y|x,t;f^*)\right)\right] = 0. \tag{10}$$

Proving from the LHS to the RHS is straightforward. Hence, we only show the other direction. Assume the RHS is true, since the integrand is non-negative, for almost every x, we have

$$\mathop{\mathbb{E}}_{t\sim p(t|x)}\left[D_{KL}\left(p(y|x,t)\|q(y|x,t;f)\right)\right] = 0$$

which then implies for almost all t (depending on x), the KL divergence is 0. KL divergence is 0 if and only if the distributions are identical, which means $p(y|x,t) = q(y|x,t;f)$.

That is, turning to the case where $y = 1$, we have that for almost every x (wrt the marginal probability $p(x) = \int p(x,t)dt$),

$$\frac{p_{\texttt{data}}(x)}{p_{\texttt{data}}(x) + p_{\texttt{data}}(t^{-1}(x))|J_{t^{-1}}(x)|} = \frac{e^{-f(x)}}{e^{-f(x)} + e^{-f(t^{-1}(x))}|J_{t^{-1}}(x)|}. \tag{11}$$

for almost all t (wrt the conditional probability $p(t|x)$). We can pick x so that $p_{\texttt{data}}(x) > 0$, which will then imply

$$\frac{p_{\texttt{data}}(t^{-1}(x))}{p_{\texttt{data}}(x)} = \frac{e^{-f(t^{-1}(x))}}{e^{-f(x)}}. \tag{12}$$

Since this holds true for $p(t|x)$-almost all t, $p_{\texttt{data}}(t^{-1}(x)) \propto e^{-f(t^{-1}(x))}$ almost surely. Finally, since $p(t|x)$ has the same support as $p(t)$, a $p(t|x)$-almost sure set is a $p(t)$-almost sure set. Then by the assumption on g_x, we have $p_{\texttt{data}}(x) \propto e^{-f(x)}$ almost everywhere on $\mathbb{R}^d$. ■

Proof of Corollary 2 For any fixed $x \in \text{supp}(p_{\texttt{data}}(x))$, we define $g_x(\epsilon) = x - \epsilon$. Since $p(\epsilon)$ is non-zero everywhere, Lebesgue measure λ is absolutely continuous wrt the probability measure ν corresponding to the density function $p(\epsilon)$. As Lebesgue measure is translation invariant and g_x is invertible, for any ν-almost sure set M, $\nu(M^c) = 0 \Rightarrow \lambda(M^c) = 0 \Rightarrow \lambda(g_x(M)^c) = 0$. Then the rest follows by an application of Theorem 1. ■