Implicit Maximum Likelihood Estimation

Ke Li **Jitendra Malik**
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA 94720
United States
{ke.li,malik}@eecs.berkeley.edu

Abstract

Implicit probabilistic models are models defined naturally in terms of a sampling procedure and often induces a likelihood function that cannot be expressed explicitly. We develop a simple method for estimating parameters in implicit models that does not require knowledge of the form of the likelihood function or any derived quantities, but can be shown to be equivalent to maximizing likelihood under some conditions. Our result holds in the non-asymptotic parametric setting, where both the capacity of the model and the number of data examples are finite. We also demonstrate encouraging experimental results.

1 Introduction

Probabilistic models are a cornerstone of machine learning and can be divided into two categories: *prescribed models* and *implicit models* (Diggle & Gratton, 1984; Mohamed & Lakshminarayanan, 2016). Prescribed models are defined by an explicit specification of the density, and so their unnormalized complete likelihood can be usually expressed in closed form. Implicit models, on the other hand, are defined most naturally in terms of a (simple) sampling procedure, like the following:

1. Sample $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$
2. Return $\mathbf{x} := T_\theta(\mathbf{z})$

where $T_\theta(\cdot)$ is a deterministic parameterized transformation like a neural net. Examples of the former include mixture of Gaussians (Everitt, 1985) and Boltzmann machines (Hinton & Sejnowski, 1986), and examples of the latter include generative adversarial nets (GANs) (Goodfellow et al., 2014; Gutmann et al., 2014) and generative moment matching nets (GMMNs) (Li et al., 2015; Dziugaite et al., 2015). Some models like variational autoencoders (Kingma & Welling, 2013; Rezende et al., 2014) belong to both categories.

Ideally, probabilistic models should be trained using the principle of maximum likelihood (Fisher, 1912; Edgeworth, 1908), which has a number of appealing properties: under mild regularity conditions, it is asymptotically consistent and efficient [1]. Unfortunately, doing so is often computationally challenging. For prescribed models, maximizing likelihood directly requires computing the partition function, which is intractable for all but the simplest models. However, many powerful techniques have been developed to attack this problem, including variational methods (Jordan et al., 1999), contrastive divergence (Hinton, 2002; Welling & Hinton, 2002), score matching (Hyvärinen, 2005) and pseudolikelihood maximization (Besag, 1975), among others.

Unfortunately, such techniques are not applicable for implicit models, as there is no term in the log-likelihood function that is in closed form; evaluating any term requires computing an intractable

[1] More justification for maximum likelihood from a practical perspective and responses to possible criticisms of maximum likelihood are found in the appendix.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

integral. As a result, implicit models must be trained using likelihood-free approaches, which do not require evaluating likelihood. Popular training objectives include adversarial loss used by GANs, and moment matching used by GMMNs. Unfortunately, these methods have limitations. For example, GANs suffer from a number of well-documented issues, such as mode dropping/collapse (Goodfellow et al., 2014; Arora & Zhang, 2017), vanishing gradients (Arjovsky & Bottou, 2017; Sinn & Rawat, 2017) and training instability (Goodfellow et al., 2014; Arora et al., 2017). Perhaps the most significant from a modelling perspective is mode dropping, which deprives the model designer of control over which data examples are modelled and which are not. Effectively, the model can choose which data examples it wants to model and disregard the rest, and so likelihood of the data could be zero under the learned model.

1.1 Our Contribution

In this paper, we present an alternative method for estimating parameters in implicit models. Even though the method is likelihood-free, it can be shown to be equivalent to maximizing likelihood under some conditions. Unlike prior methods, our result holds when the capacity of the model is finite and the number of data examples is finite.

Our method relies on the following observation: under a model distribution that maximizes the likelihood of the data, because likelihood is the product of densities evaluated at all data examples, the model density at each data example should be high. Suppose we don't observe the model distribution directly, and instead only observe independent and identically distributed (i.i.d.) samples drawn from the model. Because the density at data examples is high, then there would be more samples in the neighbourhood of each data example than elsewhere. Therefore, to maximize likelihood, we need to make this happen by adjusting parameters of the model so that each *data example* is close to some *sample*. Note that this is the *opposite* of what adversarial loss does – it ensures that each *sample* is close to some *data example*. Some data examples may not be chosen by any sample, resulting in mode dropping.

To make each data example close to some sample, the proposed method works by minimizing the distance from each data example to the nearest sample. This objective can sidestep the three common issues of existing methods: mode collapse, vanishing gradients and training instability. Modes are not dropped because the loss ensures each data example has a sample nearby at optimality; gradients do not vanish because the gradient of the distance between a data example and its nearest sample does not become zero unless they coincide; training is stable because the estimator is the solution to a simple minimization problem. By leveraging recent advances in fast nearest neighbour search algorithms (Li & Malik, 2016, 2017), this approach is able to scale to large, high-dimensional datasets.

2 Implicit Maximum Likelihood Estimator

2.1 Definition

We are given a set of n data examples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and some unknown parameterized probability distribution P_θ with density p_θ. We also have access to an oracle that allows us to draw independent and identically distributed (i.i.d.) samples from P_θ.

Let $\tilde{\mathbf{x}}_1^\theta, \ldots, \tilde{\mathbf{x}}_m^\theta$ be i.i.d. samples from P_θ, where $m \geq n$. The implicit maximum likelihood estimator $\hat{\theta}_{\mathrm{IMLE}}$ is defined as:

$$\hat{\theta}_{\mathrm{IMLE}} \coloneqq \arg\min_{\theta} \mathbb{E}_{\tilde{\mathbf{x}}_1^\theta, \ldots, \tilde{\mathbf{x}}_m^\theta} \left[\sum_{i=1}^{n} \min_{j \in [m]} \left\| \tilde{\mathbf{x}}_j^\theta - \mathbf{x}_i \right\|_2^2 \right]$$

2.2 Algorithm

We outline the proposed parameter estimation procedure in Algorithm 1. In each outer iteration, we draw m i.i.d. samples from the current model P_θ. We then randomly select a batch of examples from the dataset and find the nearest sample from each data example. We then run a standard iterative optimization algorithm, like stochastic gradient descent (SGD), to minimize a sample-based version of the Implicit Maximum Likelihood Estimator (IMLE) objective.

Algorithm 1 Implicit maximum likelihood estimation (IMLE) procedure

```
Require: The dataset $D = \{\mathbf{x}_i\}_{i=1}^n$ and a sampling mechanism for the implicit model $P_\theta$
  Initialize $\theta$ to a random vector
  for $k = 1$ to $K$ do
    Draw i.i.d. samples $\tilde{\mathbf{x}}_1^\theta, \ldots, \tilde{\mathbf{x}}_m^\theta$ from $P_\theta$
    Pick a random batch $S \subseteq \{1, \ldots, n\}$
    $\sigma(i) \leftarrow \arg\min_j \left\|\mathbf{x}_i - \tilde{\mathbf{x}}_j^\theta\right\|_2^2 \;\forall i \in S$
    for $l = 1$ to $L$ do
      Pick a random mini-batch $\tilde{S} \subseteq S$
      $\theta \leftarrow \theta - \eta \nabla_\theta \left( \frac{n}{|\tilde{S}|} \sum_{i \in \tilde{S}} \left\|\mathbf{x}_i - \tilde{\mathbf{x}}_{\sigma(i)}^\theta\right\|_2^2 \right)$
    end for
  end for
  return $\theta$
```

Because our algorithm needs to solve a nearest neighbour search problem in each outer iteration, the scalability of our method depends on our ability to find the nearest neighbours quickly. This was traditionally considered to be a hard problem, especially in high dimensions. However, this is no longer the case, due to recent advances in nearest neighbour search algorithms (Li & Malik, 2016, 2017), which avoid the curse of dimensionality in time complexity that often arises in nearest neighbour search.

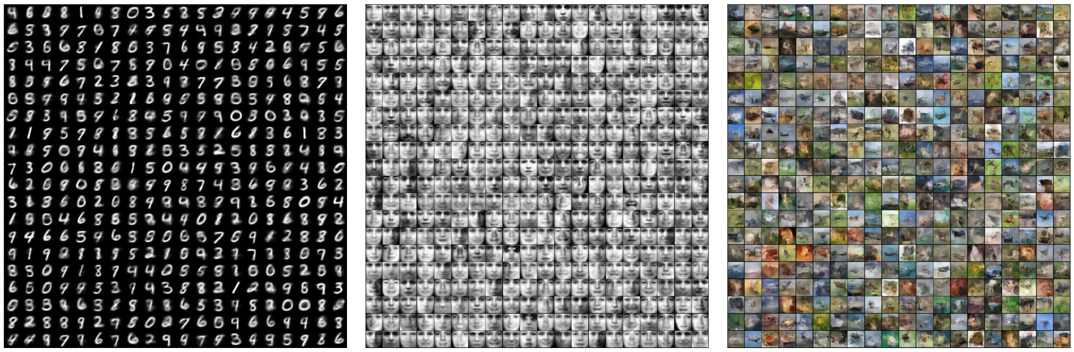
3 Equivalence to Maximum Likelihood

Below we illustrate the intuition behind why the proposed estimator is equivalent to maximum likelihood under some conditions. For simplicity, we will consider the special case where we only have a single data example $\mathbf{x}_1$ and a single sample $\tilde{\mathbf{x}}_1^\theta$. Consider the total density of P_θ inside a ball of radius of t centred at $\mathbf{x}_1$ as a function of t, a function that will be denoted as $\tilde{F}^\theta(t)$. If the density in the neighbourhood of $\mathbf{x}_1$ is high, then $\tilde{F}^\theta(t)$ would grow rapidly as t increases. If, on the other hand, the density in the neighbourhood of $\mathbf{x}_1$ is low, then $\tilde{F}^\theta(t)$ would grow slowly. So, maximizing likelihood is equivalent to making $\tilde{F}^\theta(t)$ grow as fast as possible. To this end, we can maximize the area under the function $\tilde{F}^\theta(t)$, or equivalently, minimize the area under the function $1 - \tilde{F}^\theta(t)$. Observe that $\tilde{F}^\theta(t)$ can be interpreted as the cumulative distribution function (CDF) of the Euclidean distance between $\mathbf{x}_1$ and $\tilde{\mathbf{x}}_1^\theta$, which is a random variable because $\tilde{\mathbf{x}}_1^\theta$ is random and will be denoted as $\tilde{R}^\theta$. Because $\tilde{R}^\theta$ is non-negative, recall that $\mathbb{E}\left[\tilde{R}^\theta\right] = \int_0^\infty \Pr\left(\tilde{R}^\theta > t\right) dt = \int_0^\infty \left(1 - \tilde{F}^\theta(t)\right) dt$, which is exactly the area under the function $1 - \tilde{F}^\theta(t)$. Therefore, we can maximize likelihood of a data example $\mathbf{x}_1$ by minimizing $\mathbb{E}\left[\tilde{R}^\theta\right]$, or in other words, minimizing the expected distance between the data example and a random sample. To extend this reasoning to the case with multiple data examples, we show in the appendix that if we have an objective function that is a summation, applying a monotonic transformation to each term and then reweighting appropriately preserves the optimizer under some conditions. The precise theoretical result and the proofs are in the appendix.

4 Experiments

We trained implicit probabilistic models using the proposed method on three standard benchmark datasets, MNIST, the Toronto Faces Dataset (TFD) and CIFAR-10. All models take the form of feedforward neural nets with isotropic Gaussian noise as input.

It is important to note that evaluation for implicit probabilistic models remains an open problem. Various evaluation metrics have been proposed, including estimated log-likelihood and visual assessment of sample quality. While recent literature has focused more on the latter and less on the former, it should be noted that they evaluate different properties – sample quality reflects precision, i.e.: how accurate the model samples are compared to the ground truth, whereas estimated log-likelihood focuses



(a) MNIST (b) TFD (c) CIFAR-10

Figure 1: Representative random samples from the model trained on (a) MNIST, (b) Toronto Faces Dataset and (c) CIFAR-10.

on recall, i.e.: how much of the diversity in the data distribution the model captures. Consequently, both are important metrics; one is not a replacement for the other.

Solely focusing on precision can be misleading – two models that achieve different levels of precision may simply be at different points on the same precision-recall curve, and so an improvement in precision may not mean better modelling performance if it comes at the expense of recall. Historically, because most generative models maximized likelihood or a lower bound on the likelihood, full recall was guaranteed, and so the only property that differed across models was precision. As a result, sample quality correlated with estimated log-likelihood and was a reliable indicator of modelling performance. However, with the advent of models that may drop modes, both precision and recall need to be measured.

We report the estimated log-likelihood in Table 2 and visualize randomly chosen samples in Figure 1. As shown in Figure 1, despite its simplicity, the proposed method is able to generate reasonably good samples for MNIST, TFD and CIFAR-10. A high estimated log-likelihood in Table 2 suggests that the model did not suffer from significant mode dropping. While our sample quality may not yet be state-of-the-art, it is important to remember that these results are obtained under the setting of full recall. The fact that our method is able to generate more plausible samples on CIFAR-10 than other methods at similar stages of development, such as the initial versions of GAN (Goodfellow et al., 2014) and PixelRNN (van den Oord et al., 2016), despite the minimal sophistication of our method and architecture, shows the promise of the approach.

Method	MNIST	TFD
DBN (Bengio et al., 2013)	138 ± 2	1909 ± 66
SCAE (Bengio et al., 2013)	121 ± 1.6	2110 ± 50
DGSN (Bengio et al., 2014)	214 ± 1.1	1890 ± 29
GAN (Goodfellow et al., 2014)	225 ± 2	2057 ± 26
GMMN (Li et al., 2015)	147 ± 2	2085 ± 25
IMLE (Proposed)	**257 ± 6**	**2139 ± 27**

Figure 2: Log-likelihood of the test data under the Gaussian Parzen window density estimated from samples generated by different methods.

5 Conclusion

We presented a simple method for parameter estimation for implicit probabilistic models, which works by minimizing the distance from each data example to the nearest sample. We showed that performing this estimator is equivalent to maximum likelihood under some conditions. The proposed method can capture the full diversity of the data and avoids common issues like mode collapse, vanishing gradients and training instability. The method combined with vanilla model architectures is able to achieve encouraging results on MNIST, TFD and CIFAR-10.

References

Arjovsky, Martin and Bottou, Léon. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

Arjovsky, Martin, Chintala, Soumith, and Bottou, Léon. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Arora, Sanjeev and Zhang, Yi. Do GANs actually learn the distribution? an empirical study. *arXiv preprint arXiv:1706.08224*, 2017.

Arora, Sanjeev, Ge, Rong, Liang, Yingyu, Ma, Tengyu, and Zhang, Yi. Generalization and equilibrium in generative adversarial nets (GANs). *arXiv preprint arXiv:1703.00573*, 2017.

Bengio, Yoshua, Mesnil, Grégoire, Dauphin, Yann, and Rifai, Salah. Better mixing via deep representations. In *ICML (1)*, pp. 552–560, 2013.

Bengio, Yoshua, Laufer, Eric, Alain, Guillaume, and Yosinski, Jason. Deep generative stochastic networks trainable by backprop. In *International Conference on Machine Learning*, pp. 226–234, 2014.

Besag, Julian. Statistical analysis of non-lattice data. *The statistician*, pp. 179–195, 1975.

Diggle, Peter J and Gratton, Richard J. Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 193–227, 1984.

Dziugaite, Gintare Karolina, Roy, Daniel M, and Ghahramani, Zoubin. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.

Edgeworth, Francis Ysidro. On the probable errors of frequency-constants. *Journal of the Royal Statistical Society*, 71(2):381–397, 1908.

Everitt, Brian S. *Mixture Distributions—I*. Wiley Online Library, 1985.

Fisher, Ronald A. On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41:155–160, 1912.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Grover, Aditya, Dhar, Manik, and Ermon, Stefano. Flow-GAN: Bridging implicit and prescribed learning in generative models. *arXiv preprint arXiv:1705.08868*, 2017.

Gutmann, Michael U, Dutta, Ritabrata, Kaski, Samuel, and Corander, Jukka. Likelihood-free inference via classification. *arXiv preprint arXiv:1407.4981*, 2014.

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

Hinton, Geoffrey E and Sejnowski, Terrence J. Learning and releaming in boltzmann machines. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1(282-317):2, 1986.

Huszár, Ferenc. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*, 2015.

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

Jordan, Michael I, Ghahramani, Zoubin, Jaakkola, Tommi S, and Saul, Lawrence K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Li, Ke and Malik, Jitendra. Fast k-nearest neighbour search via Dynamic Continuous Indexing. In *International Conference on Machine Learning*, pp. 671–679, 2016.

Li, Ke and Malik, Jitendra. Fast k-nearest neighbour search via Prioritized DCI. In *International Conference on Machine Learning*, pp. 2081–2090, 2017.

Li, Yujia, Swersky, Kevin, and Zemel, Rich. Generative moment matching networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 1718–1727, 2015.

Mohamed, Shakir and Lakshminarayanan, Balaji. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

Rezende, Danilo Jimenez, Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and variational inference in deep latent gaussian models. In *International Conference on Machine Learning*, 2014.

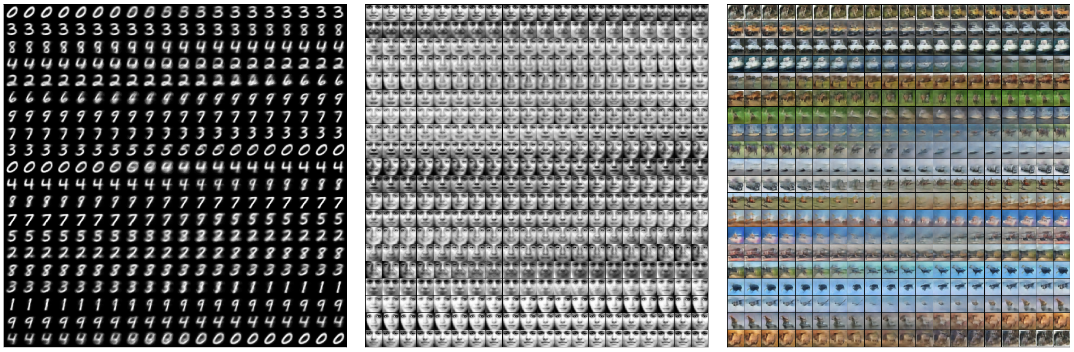
Sinn, Mathieu and Rawat, Ambrish. Non-parametric estimation of jensen-shannon divergence in generative adversarial network training. *arXiv preprint arXiv:1705.09199*, 2017.

Theis, Lucas, Oord, Aäron van den, and Bethge, Matthias. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

van den Oord, Aaron, Kalchbrenner, Nal, and Kavukcuoglu, Koray. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

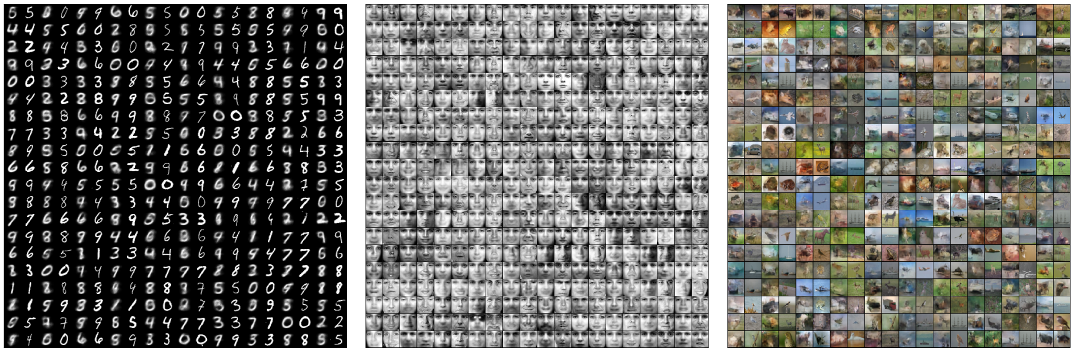
Welling, Max and Hinton, Geoffrey. A new learning algorithm for mean field boltzmann machines. *Artificial Neural Networks—ICANN 2002*, pp. 82–82, 2002.

6 Appendix I: Additional Results



(a) MNIST (b) TFD (c) CIFAR-10

Figure 3: Linear interpolation between samples in the latent variable space. The first image in every row is an independent sample; all other images are interpolated between the previous and the subsequent sample. Images along the path of interpolation are shown in the figure arranged from left to right, top to bottom. They also wrap around, so that images in the last row are interpolations between the last and first samples.



(a) MNIST (b) TFD (c) CIFAR-10

Figure 4: Comparison of samples and their nearest neighbours in the training set. Images in odd-numbered columns are samples; to the right of each sample is its nearest neighbour in the training set.

In Figure 4, we show samples and their nearest neighbours in the training set. Each sample is quite different from its nearest neighbour in the training set, suggesting that the model has not overfitted to examples in the training set. If anything, it seems to be somewhat underfitted, and further increasing the capacity of the model may improve sample quality.

Next, we visualize the learned manifold by walking along a geodesic on the manifold between pairs of samples. More concretely, we generate five samples, arrange them in arbitrary order, perform linear interpolation in latent variable space between adjacent pairs of samples, and generate an image from the interpolated latent variable. As shown in Figure 3, the images along the path of interpolation appear visually plausible and do not have noisy artifacts. In addition, the transition from one image to the next appears smooth, including for CIFAR-10, which contrasts with findings in the literature that suggest the transition between two natural images tends to be abrupt. This indicates that the support of the model distribution has not collapsed to a set of isolated points and that the proposed method is able to learn the geometry of the data manifold, even though it does not learn a distance metric explicitly.

Finally, we illustrate the evolution of samples as training progresses in Figure 5. As shown, the samples are initially blurry and become sharper over time. Importantly, sample quality consistently improves over time, which demonstrates the stability of training.

7 Appendix II: Discussion

7.1 Why Maximum Likelihood

There has been debate (Huszár, 2015) over whether maximizing likelihood of the data is the appropriate objective for the purposes of learning generative models. Recall that maximizing likelihood is equivalent to minimizing $D_{KL}\left(p_{\text{data}} \| p_\theta\right)$, where p_{data} denotes the empirical data distribution and p_θ denotes the model distribution. One proposed alternative is to minimize the reverse KL-divergence, $D_{KL}\left(p_\theta \| p_{\text{data}}\right)$, which is suggested (Huszár, 2015) to be better because it severely penalizes the model for generating an implausible sample, whereas the standard KL-divergence, $D_{KL}\left(p_{\text{data}} \| p_\theta\right)$, severely penalizes the model for assigning low density to a data example. As a result, when the model is underspecified, i.e. has less capacity than what's necessary to fit all the modes of the data distribution, minimizing $D_{KL}\left(p_\theta \| p_{\text{data}}\right)$ leads to a narrow model distribution that concentrates around a few modes, whereas minimizing $D_{KL}\left(p_{\text{data}} \| p_\theta\right)$ leads to a broad model distribution that hedges between modes. The success of GANs in generating good samples is often attributed to the former phenomenon (Arjovsky & Bottou, 2017).

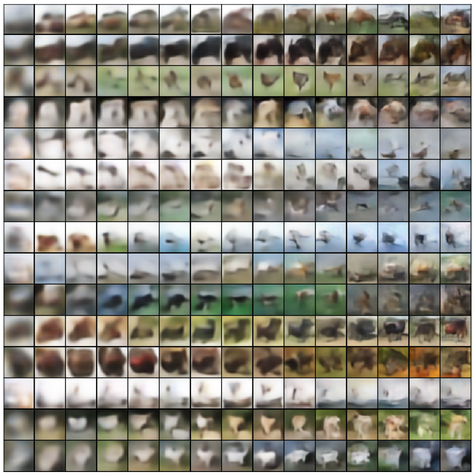


Figure 5: Samples corresponding to the same latent variable values at different points in time while training the model on CIFAR-10. Each row corresponds to a sample, and each column corresponds to a particular point in time.

This argument, however, relies on the assumption that we have access to an infinite number of samples from the true data distribution. In practice, however, this assumption rarely holds: if we had access to the true data distribution, then there is usually no need to fit a generative model, since we can simply draw samples from the true data distribution. What happens when we only have the empirical data distribution? Recall that $D_{KL}\left(p \| q\right)$ is defined and finite only if p is absolutely continuous w.r.t. q, i.e.: $q(x) = 0$ implies $p(x) = 0$ for all x. In other words, $D_{KL}\left(p \| q\right)$ is defined and finite only if the support of p is contained in the support of q. Now, consider the difference between $D_{KL}\left(p_{\text{data}} \| p_\theta\right)$ and $D_{KL}\left(p_\theta \| p_{\text{data}}\right)$: minimizing the former, which is equivalent to maximizing likelihood, ensures that the support of the model distribution contains all data examples, whereas minimizing the latter ensures that the support of the model distribution is contained in the support of the empirical data distribution, which is just the set of data examples. In other words, maximum likelihood disallows mode dropping, whereas minimizing reverse KL-divergence forces the model to assign zero density to unseen data examples and effectively prohibits generalization. Furthermore, maximum likelihood discourages the model from assigning low density to any data example, since doing so would make the likelihood, which is the product of the densities at each of the data examples, small.

From a modelling perspective, because maximum likelihood is guaranteed to preserve all modes, it can make use of all available training data and can therefore be used to train high-capacity models that have a large number of parameters. In contrast, using an objective that permits mode dropping allows the model to pick and choose which data examples it wants to model. As a result, if the goal is to train a high-capacity model that can learn the underlying data distribution, we would not be able to do so using such an objective because we have no control over which modes the model chooses to drop. Put another way, we can think about the model's performance along two axes: its ability to generate plausible samples (precision) and its ability to generate all modes of the data distribution (recall). A model that successfully learns the underlying distribution should score high along both axes. If mode dropping is allowed, then an improvement in precision may be achieved at the expense of lower recall and could represent a move to a different point on the same precision-recall curve. As a result, since sample quality is an indicator of precision, improvement in sample quality in this setting may not mean an improvement in density estimation performance. On the other hand, if mode dropping is disallowed, since full recall is always guaranteed, an improvement in precision is achieved without sacrificing recall and so implies an upwards shift in the precision-recall curve. In this case, an improvement in sample quality does signify an improvement in density estimation performance, which may explain sample quality historically was an important way to evaluate the performance of generative models, most of which maximized likelihood. With the advent of generative models that permit mode dropping, however, sample quality is no longer a reliable indicator of density estimation performance, since good sample quality can be trivially achieved by dropping all but a few modes. In this setting, sample quality can be misleading, since a model with low recall on a lower precision-recall curve can achieve a better precision than a model with high recall on a higher precision-recall curve. Since it is hard to distinguish whether an improvement in sample quality is due to a move along the same precision-recall curve or a real shift in the curve, an objective that disallows mode dropping is critical tool that researchers can use to develop better models, since they can be sure that an apparent improvement in sample quality is due to a shift in the precision-recall curve.

7.2 Frequently Asked Questions

In this section, we consider and address some possible criticisms of maximum likelihood and/or the proposed method.

7.2.1 Does Maximizing Likelihood Necessarily Lead to Poor Sample Quality?

It has been suggested (Huszár, 2015) that maximizing likelihood leads to poor sample quality because when the model is underspecified, it will try to cover all modes of the empirical data distribution and therefore assign high density to regions with few data examples. There is also empirical evidence (Grover et al., 2017) for a negative correlation between sample quality and log likelihood, suggesting an inherent trade-off between maximizing likelihood and achieving good sample quality. A popular solution is to minimize reverse KL-divergence instead, which trades off recall for precision. This is an imperfect solution, as the ultimate goal is to model all the modes *and* generate high-quality samples.

Note that this apparent trade-off exists that the model capacity is assumed to be fixed. We argue that a more promising approach would be to increase the capacity of the model, so that it is less underspecified. As the model capacity increases, avoiding mode dropping becomes more important, because otherwise there will not be enough training data to fit the larger number of parameters to. This is precisely a setting appropriate for maximum likelihood. As a result, it is possible that a combination of increasing the model capacity and maximum likelihood training can achieve good precision and recall simultaneously.

7.2.2 Would Minimizing Distance to the Nearest Samples Cause Overfitting?

When the model has infinite capacity, minimizing distance from data examples to their nearest samples will lead to a model distribution that memorizes data examples. The same is true if we maximize likelihood. Likewise, minimizing any divergence measure will lead to memorization of data examples, since the minimum divergence is zero and by definition, this can only happen if the model distribution is the same as the *empirical* data distribution, whose support is confined to the set of data examples. This implies that whenever we have a finite number of data examples, any method that learns a model with infinite capacity will memorize the data examples and will hence overfit.

To get around this, most methods learn a parametric model with finite capacity. In the parametric setting, the minimum divergence is not necessarily zero; the same is true for the minimum distance from data examples to their nearest samples. Therefore, the optimum of these objective functions is not necessarily a model distribution that memorizes data examples, and so overfitting will not necessarily occur.

7.2.3 Does Disjoint Support Break Maximum Likelihood?

Arjovsky et al. (2017) observes that the data distribution and the model distribution are supported on low-dimensional manifolds and so they are unlikely to have a non-negligible intersection. They point out $D_{KL}\left(p_{\text{data}}\,\|p_\theta\right)$ would be infinite in this case, or equivalently, the likelihood would be zero. While this does not invalidate the theoretical soundness of maximum likelihood, since the maximum of a non-negative function that is zero almost everywhere is still well-defined, it does cause a lot of practical issues for gradient-based learning, as the gradient is zero almost everywhere. This is believed to be one reason that models like variational autoencoders (Kingma & Welling, 2013; Rezende et al., 2014) use a Gaussian distribution with high variance for the conditional likelihood/observation model rather than a distribution close to the Dirac delta, so that the support of the model distribution is broadened to cover all the data examples (Arjovsky et al., 2017).

This issue does not affect our method, as our loss function is different from the log-likelihood function, even though their optima are the same (under some conditions). As the result, the gradients of our loss function are different from those of log-likelihood. When the supports of the data distribution and the model distribution do not overlap, each data example is likely far away from its nearest sample and so the gradient is large. Moreover, the farther the data examples are from the samples, the larger the gradient gets. Therefore, even when the gradient of log-likelihood can be tractably computed, there may be situations when the proposed method would work better than maximizing likelihood directly.

8 Appendix III: Theoretical Analysis

8.1 Key Result

We first state the key theoretical result that establishes equivalence between the proposed estimator and maximum likelihood:

Theorem 1. *Consider a set of observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, a parameterized family of distributions P_θ with probability density function (PDF) $p_\theta(\cdot)$ and a unique maximum likelihood solution θ^*. For any $m \geq 1$, let $\tilde{\mathbf{x}}_1^\theta, \ldots, \tilde{\mathbf{x}}_m^\theta \sim P_\theta$ be i.i.d. random variables and define $\tilde{r}^\theta := \left\|\tilde{\mathbf{x}}_1^\theta\right\|_2^2$, $R^\theta := \min_{j\in[m]} \left\|\tilde{\mathbf{x}}_j^\theta\right\|_2^2$ and $R_i^\theta := \min_{j\in[m]} \left\|\tilde{\mathbf{x}}_j^\theta - \mathbf{x}_i\right\|_2^2$. Let $F^\theta(\cdot)$ be the cumulative distribution function (CDF) of $\tilde{r}^\theta$ and $\Psi(z) := \min_\theta \left\{\mathbb{E}\left[R^\theta\right] | p_\theta(\mathbf{0}) = z\right\}$.*

If P_θ satisfies the following:

- *$p_\theta(\mathbf{x})$ is differentiable w.r.t. θ and continuous w.r.t. $\mathbf{x}$ everywhere.*
- *$\forall \theta, \mathbf{v}$, there exists θ' such that $p_\theta(\mathbf{x}) = p_{\theta'}(\mathbf{x}+\mathbf{v})\ \forall \mathbf{x}$.*
- *For any θ_1, θ_2, there exists θ_0 such that $F^{\theta_0}(t) \geq \max\left\{F^{\theta_1}(t), F^{\theta_2}(t)\right\}\ \forall t \geq 0$ and $p_{\theta_0}(\mathbf{0}) = \max\left\{p_{\theta_1}(\mathbf{0}), p_{\theta_2}(\mathbf{0})\right\}$.*
- *$\exists \tau > 0$ such that $\forall i \in [n]\ \forall \theta \notin B_{\theta^*}(\tau)$, $p_\theta(\mathbf{x}_i) < p_{\theta^*}(\mathbf{x}_i)$, where $B_{\theta^*}(\tau)$ denotes the ball centred at θ^* of radius τ.*
- *$\Psi(z)$ is differentiable everywhere.*
- *For all θ, if $\theta \neq \theta^*$, there exists $j \in [d]$ such that*
$$\left\langle \begin{pmatrix} \frac{\Psi'(p_\theta(\mathbf{x}_1))p_\theta(\mathbf{x}_1)}{\Psi'(p_{\theta^*}(\mathbf{x}_1))p_{\theta^*}(\mathbf{x}_1)} \\ \vdots \\ \frac{\Psi'(p_\theta(\mathbf{x}_n))p_\theta(\mathbf{x}_n)}{\Psi'(p_{\theta^*}(\mathbf{x}_n))p_{\theta^*}(\mathbf{x}_n)} \end{pmatrix}, \begin{pmatrix} \nabla_\theta \left(\log p_\theta(\mathbf{x}_1)\right)_j \\ \vdots \\ \nabla_\theta \left(\log p_\theta(\mathbf{x}_n)\right)_j \end{pmatrix} \right\rangle \neq 0.$$

Then,

$$\arg\min_\theta \sum_{i=1}^n \frac{\mathbb{E}\left[R_i^\theta\right]}{\Psi'(p_{\theta^*}(\mathbf{x}_i))p_{\theta^*}(\mathbf{x}_i)} = \arg\max_\theta \sum_{i=1}^n \log p_\theta(\mathbf{x}_i)$$

Furthermore, if $p_{\theta^*}(\mathbf{x}_1) = \cdots = p_{\theta^*}(\mathbf{x}_n)$, then,

$$\arg\min_\theta \sum_{i=1}^n \mathbb{E}\left[R_i^\theta\right] = \arg\max_\theta \sum_{i=1}^n \log p_\theta(\mathbf{x}_i)$$

Now, we examine the restrictiveness of each condition. The first condition is satisfied by nearly all analytic distributions. The second condition is satisfied by nearly all distributions that have an unrestricted location parameter, since one can simply shift the location parameter by $\mathbf{v}$. The third condition is satisfied by most distributions that have location and scale parameters, like a Gaussian distribution, since the scale can be made arbitrarily low and the location can be shifted so that the constraint on $p_\theta(\cdot)$ is satisfied. The fourth condition is satisfied by nearly all distributions, whose density eventually tends to zero as the distance from the optimal parameter setting tends to infinity. The fifth condition requires $\min_\theta \left\{\mathbb{E}\left[R^\theta\right] | p_\theta(\mathbf{0}) = z\right\}$ to change smoothly as $\mathbf{z}$ changes. The final condition requires the two n-dimensional vectors, one of which can be chosen from a set of d vectors, to be not exactly orthogonal. As a result, this condition is usually satisfied when d is large, i.e. when the model is richly parameterized.

There is one remaining difficulty in applying this theorem, which is that the quantity $1/\Psi'(p_{\theta^*}(\mathbf{x}_i))p_{\theta^*}(\mathbf{x}_i)$, which appears as an coefficient on each term in the proposed objective, is typically not known. If we consider a new objective that ignores the coefficients, i.e. $\sum_{i=1}^n \mathbb{E}\left[R_i^\theta\right]$, then minimizing this objective is equivalent to minimizing an upper bound on the ideal objective, $\sum_{i=1}^n \mathbb{E}\left[R_i^\theta\right] / \Psi'(p_{\theta^*}(\mathbf{x}_i))p_{\theta^*}(\mathbf{x}_i)$. The tightness of this bound depends on the difference between the highest and lowest likelihood assigned to individual data points at the optimum, i.e. the maximum likelihood estimate of the parameters. Such a model should not assign high likelihoods to some points and low likelihoods to others as long as it has reasonable capacity, since doing so would make the overall likelihood, which is the product of the likelihoods of individual data points, low. Therefore, the upper bound is usually reasonably tight.

8.2 Proofs

Before proving the main result, we first prove the following intermediate results:

Lemma 1. *Let $\Omega \subseteq \mathbb{R}^d$ and $V \subseteq \mathbb{R}$. For $i \in [N]$, let $f_i : \Omega \to V$ be differentiable on Ω and $\Phi : V \to \mathbb{R}$ be differentiable on V and strictly increasing. Assume $\arg\min_{\theta\in\Omega} \sum_{i=1}^N f_i(\theta)$ exists and is unique. Let $\theta^* := \arg\min_{\theta\in\Omega} \sum_{i=1}^N f_i(\theta)$ and $w_i := 1/\Phi'(f_i(\theta^*))$. If the following conditions hold:*

- *There is a bounded set $S \subseteq \Omega$ such that $\mathrm{bd}(S) \subseteq \Omega$, $\theta^* \in S$ and $\forall f_i,\ \forall\theta \in \Omega \setminus S,\ f_i(\theta) > f_i(\theta^*)$, where $\mathrm{bd}(S)$ denotes the boundary of S.*
- *For all $\theta \in \Omega$, if $\theta \neq \theta^*$, there exists $j \in [d]$ such that* $\left\langle \begin{pmatrix} w_1\Phi'(f_1(\theta)) \\ \vdots \\ w_n\Phi'(f_n(\theta)) \end{pmatrix}, \begin{pmatrix} \partial f_1/\partial\theta_j(\theta) \\ \vdots \\ \partial f_n/\partial\theta_j(\theta) \end{pmatrix} \right\rangle \neq 0.$

Then $\arg\min_{\mathbf{x}\in\mathbf{\Omega}} \sum_{i=1}^N w_i\Phi(f_i(\theta))$ exists and is unique. Furthermore, $\arg\min_{\theta\in\Omega} \sum_{i=1}^N w_i\Phi(f_i(\theta)) = \arg\min_{\theta\in\Omega} \sum_{i=1}^N f_i(\theta)$.

Proof. Let $S \subseteq \Omega$ be the bounded set such that $\mathrm{bd}(S) \subseteq \Omega$, $\theta^* \in S$ and $\forall f_i,\ \forall\theta \in \Omega \setminus S,\ f_i(\theta) > f_i(\theta^*)$. Consider the closure of $S := S \cup \mathrm{bd}(S)$, denoted as $\bar{S}$. Because $S \subseteq \Omega$ and $\mathrm{bd}(S) \subseteq \Omega$, $\bar{S} \subseteq \Omega$. Since S is bounded, $\bar{S}$ is bounded. Because $\bar{S} \subseteq \Omega \subseteq \mathbb{R}^d$ and is closed and bounded, it is compact.

Consider the function $\sum_{i=1}^N w_i\Phi(f_i(\cdot))$. By the differentiability of f_i's and Φ, $\sum_{i=1}^N w_i\Phi(f_i(\cdot))$ is differentiable on Ω and hence continuous on Ω. By the compactness of $\bar{S}$ and the continuity of $\sum_{i=1}^N w_i\Phi(f_i(\cdot))$ on $\bar{S} \subseteq \Omega$, Extreme Value Theorem applies, which implies that $\min_{\theta\in\bar{S}} \sum_{i=1}^N w_i\Phi(f_i(\theta))$ exists. Let $\tilde{\theta} \in \bar{S}$ be such that $\sum_{i=1}^N w_i\Phi(f_i(\tilde{\theta})) = \min_{\theta\in\bar{S}} \sum_{i=1}^N w_i\Phi(f_i(\theta))$.

By definition of S, $\forall f_i,\ \forall\theta \in \Omega \setminus S,\ f_i(\theta) > f_i(\theta^*)$, implying that $\Phi(f_i(\theta)) > \Phi(f_i(\theta^*))$ since Φ is strictly increasing. Because $\Phi'(\cdot) > 0$, $w_i > 0$ and so $\sum_{i=1}^N w_i\Phi(f_i(\theta)) > \sum_{i=1}^N w_i\Phi(f_i(\theta^*))\ \forall\theta \in \Omega \setminus S$. At the same time, since $\theta^* \in S \subset \bar{S}$, by definition of $\tilde{\theta}$, $\sum_{i=1}^N w_i\Phi(f_i(\tilde{\theta})) \leq \sum_{i=1}^N w_i\Phi(f_i(\theta^*))$. Combining these two facts yields $\sum_{i=1}^N w_i\Phi(f_i(\tilde{\theta})) \leq$

$\sum_{i=1}^N w_i\Phi(f_i(\theta^*)) < \sum_{i=1}^N w_i\Phi(f_i(\theta))\ \forall\theta \in \Omega \setminus S$. Since the inequality is strict, this implies that $\tilde{\theta} \notin \Omega \setminus S$, and so $\tilde{\theta} \in \bar{S} \setminus (\Omega \setminus S) \subseteq \Omega \setminus (\Omega \setminus S) = S$.

In addition, because $\tilde{\theta}$ is the minimizer of $\sum_{i=1}^N w_i\Phi(f_i(\cdot))$ on $\bar{S}$, $\sum_{i=1}^N w_i\Phi(f_i(\tilde{\theta})) \leq \sum_{i=1}^N w_i\Phi(f_i(\theta))\ \forall\theta \in \bar{S}$. So, $\sum_{i=1}^N w_i\Phi(f_i(\tilde{\theta})) \leq \sum_{i=1}^N w_i\Phi(f_i(\theta))\ \forall\theta \in \bar{S} \cup (\Omega \setminus S) \supseteq S \cup (\Omega \setminus S) = \Omega$. Hence, $\tilde{\theta}$ is a minimizer of $\sum_{i=1}^N w_i\Phi(f_i(\cdot))$ on Ω, and so $\min_{\theta\in\Omega}\sum_{i=1}^N w_i\Phi(f_i(\theta))$ exists. Because $\sum_{i=1}^N w_i\Phi(f_i(\cdot))$ is differentiable on Ω, $\tilde{\theta}$ must be a critical point of $\sum_{i=1}^N w_i\Phi(f_i(\cdot))$ on Ω.

On the other hand, since Φ is differentiable on V and $f_i(\theta) \in V$ for all $\theta \in \Omega$, $\Phi'(f_i(\theta))$ exists for all $\theta \in \Omega$. So,

$$\begin{aligned}\nabla\left(\sum_{i=1}^N w_i\Phi(f_i(\theta))\right) &= \sum_{i=1}^N w_i\nabla\left(\Phi(f_i(\theta))\right)\\ &= \sum_{i=1}^N w_i\Phi'(f_i(\theta))\nabla f_i(\theta)\\ &= \sum_{i=1}^N \frac{\Phi'(f_i(\theta))}{\Phi'(f_i(\theta^*))}\nabla f_i(\theta)\end{aligned}$$

At $\theta = \theta^*$,

$$\begin{aligned}\nabla\left(\sum_{i=1}^N w_i\Phi(f_i(\theta^*))\right) &= \sum_{i=1}^N \frac{\Phi'(f_i(\theta^*))}{\Phi'(f_i(\theta^*))}\nabla f_i(\theta^*)\\ &= \sum_{i=1}^N \nabla f_i(\theta^*)\end{aligned}$$

Since each f_i is differentiable on Ω, $\sum_{i=1}^N f_i$ is differentiable on Ω. Combining this with the fact that θ^* is the minimizer of $\sum_{i=1}^N f_i$ on Ω, it follows that $\nabla\left(\sum_{i=1}^N f_i(\theta^*)\right) = \sum_{i=1}^N \nabla f_i(\theta^*) = 0$. Hence, $\nabla\left(\sum_{i=1}^N w_i\Phi(f_i(\theta^*))\right) = 0$ and so θ^* is a critical point of $\sum_{i=1}^N w_i\Phi(f_i(\cdot))$.

Because $\forall\theta \in \Omega$, if $\theta \neq \theta^*$, $\exists j \in [d]$ such that $\left\langle \begin{pmatrix} w_1\Phi'(f_1(\theta)) \\ \vdots \\ w_n\Phi'(f_n(\theta)) \end{pmatrix}, \begin{pmatrix} \partial f_1/\partial\theta_j(\theta) \\ \vdots \\ \partial f_n/\partial\theta_j(\theta) \end{pmatrix} \right\rangle \neq 0$, $\sum_{i=1}^N w_i\Phi'(f_i(\theta))\nabla f_i(\theta) = \nabla\left(\sum_{i=1}^N w_i\Phi(f_i(\theta))\right) \neq 0$ for any $\theta \neq \theta^* \in \Omega$. Therefore, θ^* is the only critical point of $\sum_{i=1}^N w_i\Phi(f_i(\cdot))$ on Ω. Since $\tilde{\theta}$ is a critical point on Ω, we can conclude that $\theta^* = \tilde{\theta}$, and so θ^* is a minimizer of $\sum_{i=1}^N w_i\Phi(f_i(\cdot))$ on Ω. Since any other minimizer must be a critical point and θ^* is the only critical point, θ^* is the unique minimizer. So, $\arg\min_{\theta\in\Omega}\sum_{i=1}^N f_i(\theta) = \theta^* = \arg\min_{\theta\in\Omega}\sum_{i=1}^N w_i\Phi(f_i(\theta))$. □

Lemma 2. *Let P be a distribution on $\mathbb{R}^d$ whose density $p(\cdot)$ is continuous at a point $\mathbf{x}_0 \in \mathbb{R}^d$ and $\mathbf{x} \sim P$ be a random variable. Let $\tilde{r} := \|\mathbf{x} - \mathbf{x}_0\|_2$, $\kappa := \pi^{d/2}/\Gamma\left(\frac{d}{2}+1\right)$, where $\Gamma(\cdot)$ denotes the gamma function* [2]*, and $r := \kappa\tilde{r}^d$. Let $G(\cdot)$ denote the cumulative distribution function (CDF) of r and $\partial_+ G(\cdot)$ denote the one-sided derivative of G from the right. Then, $\partial_+ G(0) = p(\mathbf{x}_0)$.*

[2] The constant κ is the the ratio of the volume of a d-dimensional ball of radius $\tilde{r}$ to a d-dimensional cube of side length $\tilde{r}$.

Proof. By definition of $\partial_+ G(\cdot)$,

$$\begin{aligned}\partial_+ G(0) &= \lim_{h\to 0^+} \frac{G(h) - G(0)}{h} = \lim_{h\to 0^+} \frac{G(h)}{h} \\ &= \lim_{h\to 0^+} \frac{\Pr\left(r \le h\right)}{h} = \lim_{h\to 0^+} \frac{\Pr\left(\tilde{r} \le \sqrt[d]{h/\kappa}\right)}{h}\end{aligned}$$

If we define $\tilde{h} \coloneqq \sqrt[d]{h/\kappa}$, the above can be re-written as:

$$\partial_+ G(0) = \lim_{\tilde{h}\to 0^+} \frac{\Pr\left(\tilde{r} \le \tilde{h}\right)}{\kappa \tilde{h}^d} = \lim_{\tilde{h}\to 0^+} \frac{\int_{B_{\mathbf{x}_0}(\tilde{h})} p(\mathbf{u}) d\mathbf{u}}{\kappa \tilde{h}^d}$$

We want to show that $\lim_{\tilde{h}\to 0^+} \left(\int_{B_{\mathbf{x}_0}(\tilde{h})} p(\mathbf{u}) d\mathbf{u}\right) / \kappa \tilde{h}^d = p(\mathbf{x}_0)$. In other words, we want to show $\forall \epsilon > 0\ \exists \delta > 0$ such that $\forall \tilde{h} \in (0, \delta)$, $\left|\frac{\int_{B_{\mathbf{x}_0}(\tilde{h})} p(\mathbf{u}) d\mathbf{u}}{\kappa \tilde{h}^d} - p(\mathbf{x}_0)\right| < \epsilon$.

Let $\epsilon > 0$ be arbitrary.

Since $p(\cdot)$ is continuous at $\mathbf{x}_0$, by definition, $\forall \tilde{\epsilon} > 0\ \exists \tilde{\delta} > 0$ such that $\forall \mathbf{u} \in B_{\mathbf{x}_0}(\tilde{\delta})$, $|p(\mathbf{u}) - p(\mathbf{x}_0)| < \tilde{\epsilon}$. Let $\tilde{\delta} > 0$ be such that $\forall \mathbf{u} \in B_{\mathbf{x}_0}(\tilde{\delta})$, $p(\mathbf{x}_0) - \epsilon < p(\mathbf{u}) < p(\mathbf{x}_0) + \epsilon$. We choose $\delta = \tilde{\delta}$.

Let $0 < \tilde{h} < \delta$ be arbitrary. Since $p(\mathbf{x}_0) - \epsilon < p(\mathbf{u}) < p(\mathbf{x}_0) + \epsilon\ \forall \mathbf{u} \in B_{\mathbf{x}_0}(\tilde{\delta}) = B_{\mathbf{x}_0}(\delta) \supset B_{\mathbf{x}_0}(\tilde{h})$,

$$\begin{aligned}\int_{B_{\mathbf{x}_0}(\tilde{h})} p(\mathbf{u}) d\mathbf{u} &< \int_{B_{\mathbf{x}_0}(\tilde{h})} \left(p(\mathbf{x}_0) + \epsilon\right) d\mathbf{u} \\ &= \left(p(\mathbf{x}_0) + \epsilon\right) \int_{B_{\mathbf{x}_0}(\tilde{h})} d\mathbf{u}\end{aligned}$$

Observe that $\int_{B_{\mathbf{x}_0}(\tilde{h})} d\mathbf{u}$ is the volume of a d-dimensional ball of radius $\tilde{h}$, so $\int_{B_{\mathbf{x}_0}(\tilde{h})} d\mathbf{u} = \kappa \tilde{h}^d$. Thus, $\int_{B_{\mathbf{x}_0}(\tilde{h})} p(\mathbf{u}) d\mathbf{u} < \kappa \tilde{h}^d \left(p(\mathbf{x}_0) + \epsilon\right)$, implying that $\left(\int_{B_{\mathbf{x}_0}(\tilde{h})} p(\mathbf{u}) d\mathbf{u}\right) / \kappa \tilde{h}^d < p(\mathbf{x}_0) + \epsilon$. By similar reasoning, we conclude that $\left(\int_{B_{\mathbf{x}_0}(\tilde{h})} p(\mathbf{u}) d\mathbf{u}\right) / \kappa \tilde{h}^d > p(\mathbf{x}_0) - \epsilon$.

Hence,

$$\left|\frac{\int_{B_{\mathbf{x}_0}(\tilde{h})} p(\mathbf{u}) d\mathbf{u}}{\kappa \tilde{h}^d} - p(\mathbf{x}_0)\right| < \epsilon\ \forall \tilde{h} \in (0, \delta)$$

Therefore,

$$\partial_+ G(0) = \lim_{\tilde{h}\to 0^+} \frac{\int_{B_{\mathbf{x}_0}(\tilde{h})} p(\mathbf{u}) d\mathbf{u}}{\kappa \tilde{h}^d} = p(\mathbf{x}_0)$$

□

Lemma 3. *Let P_θ be a parameterized family of distributions on $\mathbb{R}^d$ with parameter θ and probability density function (PDF) $p_\theta(\cdot)$ that is continuous at a point $\mathbf{x}_i$. Consider a random variable $\tilde{\mathbf{x}}_1^\theta \sim P_\theta$ and define $\tilde{r}_i^\theta \coloneqq \left\|\tilde{\mathbf{x}}_1^\theta - \mathbf{x}_i\right\|_2^2$, whose cumulative distribution function (CDF) is denoted by $F_i^\theta(\cdot)$. Assume P_θ has the following property: for any θ_1, θ_2, there exists θ_0 such that $F_i^{\theta_0}(t) \ge \max\left\{F_i^{\theta_1}(t), F_i^{\theta_2}(t)\right\}\ \forall t \ge 0$ and $p_{\theta_0}(\mathbf{x}_i) = \max\left\{p_{\theta_1}(\mathbf{x}_i), p_{\theta_2}(\mathbf{x}_i)\right\}$. For any $m \ge 1$, let $\tilde{\mathbf{x}}_1^\theta, \dots, \tilde{\mathbf{x}}_m^\theta \sim P_\theta$ be i.i.d. random variables and define $R_i^\theta \coloneqq \min_{j\in[m]} \left\|\tilde{\mathbf{x}}_j^\theta - \mathbf{x}_i\right\|_2^2$. Then the function $\Psi_i : z \mapsto \min_\theta \left\{\mathbb{E}\left[R_i^\theta\right] | p_\theta(\mathbf{x}_i) = z\right\}$ is strictly decreasing.*

Proof. Let $r_i^\theta := \kappa\left(\tilde{r}_i^\theta\right)^{d/2} = \kappa\left\|\tilde{\mathbf{x}}_1^\theta - \mathbf{x}_i\right\|_2^d$ be a random variable and let $G_i^\theta(\cdot)$ be the CDF of r_i^θ. Since R_i^θ is nonnegative,

$$\begin{aligned}
\mathbb{E}\left[R_i^\theta\right] &= \int_0^\infty \Pr\left(R_i^\theta > t\right) dt \\
&= \int_0^\infty \left(\Pr\left(\left\|\tilde{\mathbf{x}}_1^\theta - \mathbf{x}_i\right\|_2^2 > t\right)\right)^m dt \\
&= \int_0^\infty \left(\Pr\left(\kappa\left\|\tilde{\mathbf{x}}_1^\theta - \mathbf{x}_i\right\|_2^d > \kappa t^{d/2}\right)\right)^m dt \\
&= \int_0^\infty \left(\Pr\left(r_i^\theta > \kappa t^{d/2}\right)\right)^m dt \\
&= \int_0^\infty \left(1 - G_i^\theta\left(\kappa t^{d/2}\right)\right)^m dt
\end{aligned}$$

Also, by Lemma 2, $p_\theta(\mathbf{x}_i) = \partial_+ G_i^\theta(0)$. Using these facts, we can rewrite $\min_\theta \left\{\mathbb{E}\left[R_i^\theta\right] \middle| p_\theta(\mathbf{x}_i) = z\right\}$ as $\min_\theta \left\{\int_0^\infty \left(1 - G_i^\theta\left(\kappa t^{d/2}\right)\right)^m dt \middle| \partial_+ G_i^\theta(0) = z\right\}$. By definition of Ψ_i, $\min_\theta \left\{\int_0^\infty \left(1 - G_i^\theta\left(\kappa t^{d/2}\right)\right)^m dt \middle| \partial_+ G_i^\theta(0) = z\right\}$ exists for all z. Let $\phi_i(z)$ be a value of θ that attains the minimum. Define $G_i^*(y,z) := G_i^{\phi_i(z)}(y)$. By definition, $\frac{\partial_+}{\partial y} G_i^*(0,z) = z$, where $\frac{\partial_+}{\partial y} G_i^*(y,z)$ denotes the one-sided partial derivative from the right w.r.t. y. Also, since $G_i^*(\cdot, z)$ is the CDF of a distribution of a non-negative random variable, $G_i^*(0,z) = 0$.

By definition of $\frac{\partial_+}{\partial y} G_i^*(0,z)$, $\forall \epsilon > 0\ \exists \delta > 0$ such that $\forall h \in (0,\delta)$, $\left|\frac{G_i^*(h,z) - G_i^*(0,z)}{h} - z\right| < \epsilon$.

Let $z' > z$. Let $\delta > 0$ be such that $\forall h \in (0,\delta)$, $\left|\frac{G_i^*(h,z) - G_i^*(0,z)}{h} - z\right| < \frac{z'-z}{2}$ and $\delta' > 0$ be such that $\forall h \in (0,\delta')$, $\left|\frac{G_i^*(h,z') - G_i^*(0,z')}{h} - z'\right| < \frac{z'-z}{2}$.

Consider $h \in (0, \min(\delta,\delta'))$. Then, $\frac{G_i^*(h,z) - G_i^*(0,z)}{h} = \frac{G_i^*(h,z)}{h} < z + \frac{z'-z}{2} = \frac{z+z'}{2}$ and $\frac{G_i^*(h,z') - G_i^*(0,z')}{h} = \frac{G_i^*(h,z')}{h} > z' - \frac{z'-z}{2} = \frac{z+z'}{2}$. So,

$$\frac{G_i^*(h,z)}{h} < \frac{z+z'}{2} < \frac{G_i^*(h,z')}{h}$$

Multiplying by h on both sides, we conclude that $G_i^*(h,z) < G_i^*(h,z')\ \forall h \in (0, \min(\delta,\delta'))$.

Let $\alpha := \sqrt[d]{\min(\delta,\delta')/\kappa}$. We can break $\int_0^\infty \left(1 - G_i^*\left(\kappa t^{d/2}, z\right)\right)^m dt$ into two terms:

$$\begin{aligned}
&\int_0^\infty \left(1 - G_i^*\left(\kappa t^{d/2}, z\right)\right)^m dt \\
=&\int_0^\alpha \left(1 - G_i^*\left(\kappa t^{d/2}, z\right)\right)^m dt + \int_\alpha^\infty \left(1 - G_i^*\left(\kappa t^{d/2}, z\right)\right)^m dt
\end{aligned}$$

We can also do the same for $\int_0^\infty \left(1 - G_i^*\left(\kappa t^{d/2}, z'\right)\right)^m dt$.

Because $G_i^*(h,z) < G_i^*(h,z')\ \forall h \in (0,\min(\delta,\delta'))$, $G_i^*(\kappa t^{d/2}, z) < G_i^*(\kappa t^{d/2}, z')\ \forall t \in (0,\alpha)$. It follows that $1 - G_i^*(\kappa t^{d/2}, z) > 1 - G_i^*(\kappa t^{d/2}, z')$ and $\left(1 - G_i^*\left(\kappa t^{d/2}, z\right)\right)^m > \left(1 - G_i^*(\kappa t^{d/2}, z')\right)^m\ \forall t \in (0,\alpha)$. So, $\int_0^\alpha \left(1 - G_i^*\left(\kappa t^{d/2}, z\right)\right)^m dt > \int_0^\alpha \left(1 - G_i^*\left(\kappa t^{d/2}, z'\right)\right)^m dt$.

We now consider the second term. First, observe that $F_i^\theta(t) = \Pr\left(\left\|\tilde{\mathbf{x}}_1^\theta - \mathbf{x}_i\right\|_2^2 \leq t\right) = \Pr\left(\kappa\left\|\tilde{\mathbf{x}}_1^\theta - \mathbf{x}_i\right\|_2^d \leq \kappa t^{d/2}\right) = G_i^\theta\left(\kappa t^{d/2}\right)$ for all $t \geq 0$. So, by the property of P_θ, for any θ_1, θ_2, there exists θ_0 such that $G_i^{\theta_0}(\kappa t^{d/2}) = F_i^{\theta_0}(t) \geq \max\left\{F_i^{\theta_1}(t), F_i^{\theta_2}(t)\right\} = \max\left\{G_i^{\theta_1}(\kappa t^{d/2}), G_i^{\theta_2}(\kappa t^{d/2})\right\}\ \forall t \geq 0$ and $\partial_+ G_i^{\theta_0}(0) = p_{\theta_0}(\mathbf{x}_i) = \max\left\{p_{\theta_1}(\mathbf{x}_i), p_{\theta_2}(\mathbf{x}_i)\right\} = \max\left\{\partial_+ G_i^{\theta_1}(0), \partial_+ G_i^{\theta_2}(0)\right\}$.

Take $\theta_1 = \phi_i(z)$ and $\theta_2 = \phi_i(z')$. Let θ_0 be such that $G_i^{\theta_0}(\kappa t^{d/2}) \geq \max\left\{G_i^{\theta_1}(\kappa t^{d/2}), G_i^{\theta_2}(\kappa t^{d/2})\right\}\ \ \forall t \geq 0$ and $\partial_+ G_i^{\theta_0}(0) = \max\left\{\partial_+ G_i^{\theta_1}(0), \partial_+ G_i^{\theta_2}(0)\right\}$. By definition of $\phi_i(\cdot)$, $\partial_+ G_i^{\theta_1}(0) = z$ and $\partial_+ G_i^{\theta_2}(0) = z'$. So, $\partial_+ G_i^{\theta_0}(0) = \max\{z, z'\} = z'$. Since $G_i^{\theta_0}(\kappa t^{d/2}) \geq G_i^{\theta_2}(\kappa t^{d/2})\ \ \forall t \geq 0$, $1 - G_i^{\theta_0}\left(\kappa t^{d/2}\right) \leq 1 - G_i^{\theta_2}\left(\kappa t^{d/2}\right)\ \ \forall t \geq 0$ and so $\int_0^\infty \left(1 - G_i^{\theta_0}\left(\kappa t^{d/2}\right)\right)^m dt \leq \int_0^\infty \left(1 - G_i^{\theta_2}\left(\kappa t^{d/2}\right)\right)^m dt$. On the other hand, because $\theta_2 = \phi_i(z')$ minimizes $\int_0^\infty \left(1 - G_i^{\theta}\left(\kappa t^{d/2}\right)\right)^m dt$ among all θ's such that $\partial_+ G_i^{\theta}(0) = z'$ and $\partial_+ G_i^{\theta_0}(0) = z'$, $\int_0^\infty \left(1 - G_i^{\theta_2}\left(\kappa t^{d/2}\right)\right)^m dt \leq \int_0^\infty \left(1 - G_i^{\theta_0}\left(\kappa t^{d/2}\right)\right)^m dt$. We can therefore conclude that $\int_0^\infty \left(1 - G_i^{\theta_0}\left(\kappa t^{d/2}\right)\right)^m dt = \int_0^\infty \left(1 - G_i^{\theta_2}\left(\kappa t^{d/2}\right)\right)^m dt$. Since $1 - G_i^{\theta_0}\left(\kappa t^{d/2}\right) \leq 1 - G_i^{\theta_2}\left(\kappa t^{d/2}\right)\ \ \forall t \geq 0$, the only situation where this can happen is when $G_i^{\theta_0}\left(\kappa t^{d/2}\right) = G_i^{\theta_2}\left(\kappa t^{d/2}\right)\ \forall t \geq 0$.

By definition of G_i^*, $G_i^*\left(\kappa t^{d/2}, z\right) = G_i^{\phi_i(z)}(\kappa t^{d/2}) = G_i^{\theta_1}(\kappa t^{d/2})$ and $G_i^*\left(\kappa t^{d/2}, z'\right) = G_i^{\phi_i(z')}(\kappa t^{d/2}) = G_i^{\theta_2}(\kappa t^{d/2}) = G_i^{\theta_0}\left(\kappa t^{d/2}\right)$. By definition of θ_0, $G_i^{\theta_0}\left(\kappa t^{d/2}\right) \geq G_i^{\theta_1}(\kappa t^{d/2})\ \forall t \geq 0$. So, $G_i^*\left(\kappa t^{d/2}, z'\right) = G_i^{\theta_2}(\kappa t^{d/2}) \geq G_i^{\theta_1}(\kappa t^{d/2}) = G_i^*\left(\kappa t^{d/2}, z\right)\ \ \forall t \geq 0$. Hence, $\int_\alpha^\infty \left(1 - G_i^*\left(\kappa t^{d/2}, z'\right)\right)^m dt \leq \int_\alpha^\infty \left(1 - G_i^*\left(\kappa t^{d/2}, z\right)\right)^m dt$.

Combining with the previous result that $\int_0^\alpha \left(1 - G_i^*\left(\kappa t^{d/2}, z'\right)\right)^m dt < \int_0^\alpha \left(1 - G_i^*\left(\kappa t^{d/2}, z\right)\right)^m dt$, it follows that:

$$\begin{aligned}
&\int_0^\infty \left(1 - G_i^*\left(\kappa t^{d/2}, z'\right)\right)^m dt \\
=&\int_0^\alpha \left(1 - G_i^*\left(\kappa t^{d/2}, z'\right)\right)^m dt + \int_\alpha^\infty \left(1 - G_i^*\left(\kappa t^{d/2}, z'\right)\right)^m dt \\
<&\int_0^\alpha \left(1 - G_i^*\left(\kappa t^{d/2}, z\right)\right)^m dt + \int_\alpha^\infty \left(1 - G_i^*\left(\kappa t^{d/2}, z'\right)\right)^m dt \\
\leq&\int_0^\alpha \left(1 - G_i^*\left(\kappa t^{d/2}, z\right)\right)^m dt + \int_\alpha^\infty \left(1 - G_i^*\left(\kappa t^{d/2}, z\right)\right)^m dt \\
=&\int_0^\infty \left(1 - G_i^*\left(\kappa t^{d/2}, z\right)\right)^m dt
\end{aligned}$$

By definition,

$$\begin{aligned}
&\int_0^\infty \left(1 - G_i^*\left(\kappa t^{d/2}, z\right)\right)^m dt \\
=&\int_0^\infty \left(1 - G_i^{\phi_i(z)}(\kappa t^{d/2})\right)^m dt \\
=&\ \min_\theta \left\{\int_0^\infty \left(1 - G_i^\theta\left(\kappa t^{d/2}\right)\right)^m dt \,\middle|\, \partial_+ G_i^\theta(0) = z\right\} \\
=&\ \min_\theta \left\{\mathbb{E}\left[R_i^\theta\right] | p_\theta(\mathbf{x}_i) = z\right\} \\
=&\ \Psi_i(z)
\end{aligned}$$

Similarly, $\int_0^\infty \left(1 - G_i^*\left(\kappa t^{d/2}, z'\right)\right)^m dt = \Psi_i(z')$. We can therefore conclude that $\Psi_i(z') < \Psi_i(z)$ whenever $z' > z$. □

We now prove the main result.

Theorem 1. *Consider a set of observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, a parameterized family of distributions P_θ with probability density function (PDF) $p_\theta(\cdot)$ and a unique maximum likelihood solution θ^*. For any $m \geq 1$, let $\tilde{\mathbf{x}}_1^\theta, \ldots, \tilde{\mathbf{x}}_m^\theta \sim P_\theta$ be i.i.d. random variables and define $\tilde{r}^\theta := \left\|\tilde{\mathbf{x}}_1^\theta\right\|_2^2$, $R^\theta := \min_{j\in[m]} \left\|\tilde{\mathbf{x}}_j^\theta\right\|_2^2$ and $R_i^\theta := \min_{j\in[m]} \left\|\tilde{\mathbf{x}}_j^\theta - \mathbf{x}_i\right\|_2^2$. Let $F^\theta(\cdot)$ be the cumulative distribution function (CDF) of $\tilde{r}^\theta$ and $\Psi(z) := \min_\theta \left\{\mathbb{E}\left[R^\theta\right] | p_\theta(\mathbf{0}) = z\right\}$.*

If P_θ satisfies the following:

- $p_\theta(\mathbf{x})$ *is differentiable w.r.t.* θ *and continuous w.r.t.* $\mathbf{x}$ *everywhere.*
- $\forall \theta, \mathbf{v}$, *there exists* θ' *such that* $p_\theta(\mathbf{x}) = p_{\theta'}(\mathbf{x}+\mathbf{v})\ \forall \mathbf{x}$.
- *For any* θ_1, θ_2, *there exists* θ_0 *such that* $F^{\theta_0}(t) \geq \max\left\{F^{\theta_1}(t), F^{\theta_2}(t)\right\}\ \forall t \geq 0$ *and* $p_{\theta_0}(\mathbf{0}) = \max\{p_{\theta_1}(\mathbf{0}), p_{\theta_2}(\mathbf{0})\}$.
- $\exists \tau > 0$ *such that* $\forall i \in [n]\ \forall \theta \notin B_{\theta^*}(\tau)$, $p_\theta(\mathbf{x}_i) < p_{\theta^*}(\mathbf{x}_i)$, *where* $B_{\theta^*}(\tau)$ *denotes the ball centred at* θ^* *of radius* τ.
- $\Psi(z)$ *is differentiable everywhere.*
- *For all* θ, *if* $\theta \neq \theta^*$, *there exists* $j \in [d]$ *such that* $\left\langle \begin{pmatrix} \frac{\Psi'(p_\theta(\mathbf{x}_1))p_\theta(\mathbf{x}_1)}{\Psi'(p_{\theta^*}(\mathbf{x}_1))p_{\theta^*}(\mathbf{x}_1)} \\ \vdots \\ \frac{\Psi'(p_\theta(\mathbf{x}_n))p_\theta(\mathbf{x}_n)}{\Psi'(p_{\theta^*}(\mathbf{x}_n))p_{\theta^*}(\mathbf{x}_n)} \end{pmatrix}, \begin{pmatrix} \nabla_\theta \left(\log p_\theta(\mathbf{x}_1)\right)_j \\ \vdots \\ \nabla_\theta \left(\log p_\theta(\mathbf{x}_n)\right)_j \end{pmatrix} \right\rangle \neq 0.$

Then,

$$\arg\min_\theta \sum_{i=1}^n \frac{\mathbb{E}\left[R_i^\theta\right]}{\Psi'(p_{\theta^*}(\mathbf{x}_i))p_{\theta^*}(\mathbf{x}_i)} = \arg\max_\theta \sum_{i=1}^n \log p_\theta(\mathbf{x}_i)$$

Furthermore, if $p_{\theta^*}(\mathbf{x}_1) = \cdots = p_{\theta^*}(\mathbf{x}_n)$, *then,*

$$\arg\min_\theta \sum_{i=1}^n \mathbb{E}\left[R_i^\theta\right] = \arg\max_\theta \sum_{i=1}^n \log p_\theta(\mathbf{x}_i)$$

Proof. Pick an arbitrary $i \in [n]$. We first prove a few basic facts.

By the second property of P_θ, $\forall \theta\ \exists \theta'$ such that $p_\theta(\mathbf{u}) = p_{\theta'}(\mathbf{u} - \mathbf{x}_i)\ \forall \mathbf{u}$. In particular, $p_\theta(\mathbf{x}_i) = p_{\theta'}(\mathbf{x}_i - \mathbf{x}_i) = p_{\theta'}(\mathbf{0})$. Let F_i^θ be as defined in Lemma 3.

$$\begin{aligned} F_i^\theta(t) &= \Pr\left(\tilde{r}_i^\theta \leq t\right) = \Pr\left(\left\|\tilde{\mathbf{x}}_1^\theta - \mathbf{x}_i\right\|_2 \leq \sqrt{t}\right) \\ &= \int_{B_{\mathbf{x}_i}(\sqrt{t})} p_\theta(\mathbf{u}) d\mathbf{u} = \int_{B_{\mathbf{x}_i}(\sqrt{t})} p_{\theta'}(\mathbf{u} - \mathbf{x}_i) d\mathbf{u} \\ &= \int_{B_{\mathbf{0}}(\sqrt{t})} p_{\theta'}(\mathbf{u}) d\mathbf{u} = \Pr\left(\tilde{r}^{\theta'} \leq t\right) = F^{\theta'}(t) \end{aligned}$$

Similarly, $\forall \theta'\ \exists \theta$ such that $p_{\theta'}(\mathbf{u}) = p_\theta(\mathbf{u} + \mathbf{x}_i)\ \forall \mathbf{u}$. In particular, $p_{\theta'}(\mathbf{0}) = p_\theta(\mathbf{0} + \mathbf{x}_i) = p_\theta(\mathbf{x}_i)$.

$$\begin{aligned} F^{\theta'}(t) &= \Pr\left(\tilde{r}^{\theta'} \leq t\right) = \int_{B_{\mathbf{0}}(\sqrt{t})} p_{\theta'}(\mathbf{u}) d\mathbf{u} \\ &= \int_{B_{\mathbf{0}}(\sqrt{t})} p_\theta(\mathbf{u} + \mathbf{x}_i) d\mathbf{u} = \int_{B_{\mathbf{x}_i}(\sqrt{t})} p_\theta(\mathbf{u}) d\mathbf{u} \\ &= \Pr\left(\left\|\tilde{\mathbf{x}}_1^\theta - \mathbf{x}_i\right\|_2 \leq \sqrt{t}\right) = \Pr\left(\tilde{r}_i^\theta \leq t\right) = F_i^\theta(t) \end{aligned}$$

Let θ_1, θ_2 be arbitrary. The facts above imply that there exist θ_1' and θ_2' such that $F_i^{\theta_1}(t) = F^{\theta_1'}(t)$, $F_i^{\theta_2}(t) = F^{\theta_2'}(t)$, $p_{\theta_1}(\mathbf{x}_i) = p_{\theta_1'}(\mathbf{0})$ and $p_{\theta_2}(\mathbf{x}_i) = p_{\theta_2'}(\mathbf{0})$.

By the third property of P_θ, let θ_0' be such that $F^{\theta_0'}(t) \geq \max\left\{F^{\theta_1'}(t), F^{\theta_2'}(t)\right\}\ \forall t \geq 0$ and $p_{\theta_0'}(\mathbf{0}) = \max\left\{p_{\theta_1'}(\mathbf{0}), p_{\theta_2'}(\mathbf{0})\right\}$. By the facts above, it follows that there exists θ_0 such that $F^{\theta_0'}(t) = F_i^{\theta_0}(t)$ and $p_{\theta_0'}(\mathbf{0}) = p_{\theta_0}(\mathbf{x}_i)$.

So, we can conclude that for any θ_1, θ_2, there exists θ_0 such that $F_i^{\theta_0}(t) \geq \max\left\{F_i^{\theta_1}(t), F_i^{\theta_2}(t)\right\}\ \forall t \geq 0$ and $p_{\theta_0}(\mathbf{x}_i) = \max\{p_{\theta_1}(\mathbf{x}_i), p_{\theta_2}(\mathbf{x}_i)\}$.

By Lemma 3, $\Psi_i(z) = \min_\theta \left\{\mathbb{E}\left[R_i^\theta\right] | p_\theta(\mathbf{x}_i) = z\right\}$ is strictly decreasing.

Consider any θ. By the facts above, there exists θ' such that $p_\theta(\mathbf{x}_i) = p_{\theta'}(\mathbf{0})$ and $F_i^\theta(t) = F^{\theta'}(t)\ \forall t$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[R_i^\theta\right] =& \int_0^\infty \Pr\left(R_i^\theta > t\right) dt \\
=& \int_0^\infty \left(\Pr\left(\left\|\tilde{\mathbf{x}}_1^\theta - \mathbf{x}_i\right\|_2^2 > t\right)\right)^m dt \\
=& \int_0^\infty \left(1 - F_i^\theta(t)\right)^m dt \\
=& \int_0^\infty \left(1 - F^{\theta'}(t)\right)^m dt \\
=& \int_0^\infty \Pr\left(R^{\theta'} > t\right) dt \\
=& \mathbb{E}\left[R^{\theta'}\right]
\end{aligned}$$

So, $\forall z$

$$\begin{aligned}
\Psi_i(z) =& \min_\theta \left\{\mathbb{E}\left[R_i^\theta\right] | p_\theta(\mathbf{x}_i) = z\right\} \\
=& \min_{\theta'} \left\{\mathbb{E}\left[R^{\theta'}\right] | p_{\theta'}(\mathbf{0}) = z\right\} \\
=& \Psi(z)
\end{aligned}$$

Because $\Psi_i(\cdot)$ is strictly decreasing, $\Psi(\cdot)$ is also strictly decreasing.

We would like to apply Lemma 1, with $f_i(\theta) = -\log p_\theta(\mathbf{x}_i)\ \forall i \in [n]$ and $\Phi(y) = \Psi(\exp(-y))$. By the first property of P_θ, $p_\theta(\cdot)$ is differentiable w.r.t. θ and so $f_i(\theta)$ is differentiable for all i. By the fifth property of P_θ, $\Psi(\cdot)$ is differentiable and so $\Phi(\cdot)$ is differentiable. Since $y \mapsto \exp(-y)$ is strictly decreasing and $\Psi(\cdot)$ is strictly decreasing, $\Phi(\cdot)$ is strictly increasing. Since there is a unique maximum likelihood solution θ^*, $\min_\theta \sum_{i=1}^n f_i(\theta) = \max_\theta \sum_{i=1}^n \log p_\theta(\mathbf{x}_i)$ exists and has a unique minimizer. By the fourth property of P_θ, the first condition of Lemma 1 is satisfied. By the sixth property of P_θ, the second condition of Lemma 1 is satisfied. Since all conditions are satisfied, we apply Lemma 1 and conclude that

$$\begin{aligned}
\min_\theta \sum_{i=1}^n w_i \Phi(f_i(\theta)) =& \min_\theta \sum_{i=1}^n w_i \Psi(p_\theta(\mathbf{x}_i)) \\
=& \min_\theta \sum_{i=1}^n w_i \Psi_i(p_\theta(\mathbf{x}_i)) \\
=& \min_\theta \sum_{i=1}^n \frac{\mathbb{E}\left[R_i^\theta\right]}{\Psi'(p_{\theta^*}(\mathbf{x}_i)) p_{\theta^*}(\mathbf{x}_i)}
\end{aligned}$$

exists and has a unique minimizer. Furthermore,

$$\begin{aligned}
\arg\min_\theta \sum_{i=1}^n \frac{\mathbb{E}\left[R_i^\theta\right]}{\Psi'(p_{\theta^*}(\mathbf{x}_i)) p_{\theta^*}(\mathbf{x}_i)} =& \arg\min_\theta \sum_{i=1}^n -\log p_\theta(\mathbf{x}_i) \\
=& \arg\max_\theta \sum_{i=1}^n \log p_\theta(\mathbf{x}_i)
\end{aligned}$$

If $p_\theta(\mathbf{x}_1) = \cdots p_\theta(\mathbf{x}_n)$, then $w_1 = \cdots = w_n$, and so $\arg\min_\theta \sum_{i=1}^n w_i \mathbb{E}\left[R_i^\theta\right] = \arg\min_\theta \sum_{i=1}^n \mathbb{E}\left[R_i^\theta\right] = \arg\max_\theta \sum_{i=1}^n \log p_\theta(\mathbf{x}_i)$. □