Deep Density Estimation
10716: Advanced Machine Learning
Pradeep Ravikumar

1 Introduction

Consider the density estimation problem where we wish to estimate the density p of some distribution P, and where we are given samples $\{X_i\}_{i=1}^n$ drawn iid from that distribution. Suppose we wish to do parametric density estimation: we then start with a parametric class of densities $\{p_\theta\}_{\theta \in \Theta}$, and then estimate the density $p_{\widehat{\theta}}$, for some $\widehat{\theta} \in \Theta$, with the best fit to the data $\{X_i\}_{i=1}^n$. There are two technical facets to this: (a) how to specify a parametric family of densities, and (b) how to determine goodness of fit of any member of the family to data.

2 Multivariate Exponential Families (MEFs)

A very classical and popular class of distributions is the exponential family class:

$$p(X) = \exp(\langle \theta, \phi(X) \rangle + B(X) - A(\theta)),$$

which are specified by sufficient statistics $\phi : \mathcal{X} \to \mathbb{R}^d$, and a log base measure $B(X)$. $A(\theta)$ is the log-normalization constant, also known as the log-partition function:

$$A(\theta) = \int_{\mathcal{X}} \exp(\langle \theta, \phi(X) \rangle + B(X)) dX.$$

Most of the "named" distributions you have heard of are members of the exponential family class: normal, exponential, gamma, chi-squared, beta, Dirichlet, Bernoulli, Poisson, categorical, Wishart, inverse Wishart, and geometric, to name a few. Each of these make some specific choice of $\phi(X)$ and $B(X)$, usually depending also on the domain $\mathcal{X}$ (e.g. Poisson for count data). However all of these are examples of univariate exponential families where $\mathcal{X} \subseteq \mathbb{R}$. While this is well-defined for multivariate data, a key question is how do we specify the functions $\phi(X)$ and $B(X)$?

There is however one class of exponential family distributions where the multivariate counterpart is also equally popular: the multivariate categorical distribution. This is a popular class of distributions for discrete or categorical data, and is simply the probability table where the rows are the different configurations of all the discrete variables and with one probability column that has the corresponding probabilities. If the probabilities are all non-negative, this corresponds to an exponential family with indicator sufficient statistics. If

there are d variables, and each takes k values, then there are d^k possible configurations, so that there are too many sufficient statistics. A natural approach might be truncate the set of sufficient statistics to only have upto r-order interactions. Such approaches form the rich subject of categorical data analysis, and discrete graphical models.

The other popular multivariate exponential family is for continuous data and is the multivariate Gaussian distribution. One of the defining characteristics of the multivariate Gaussian is that it is the unique distribution where the conditional distributions of a variable conditioned on the other variables $P(X_i|X_{-i})$ are univariate Gaussian for any fixed values of X_{-i}. This observation thus naturally allows us to answer the question: how do we take any of the beloved univariate exponential family distributions (e.g. univariate Poisson) to a corresponding canonical multivariate distribution?

[Yang et al., 2015] showed the following. Suppose that for all $i \in [n]$:

$$P(X_i|X_{-i}) \propto \exp(\langle \theta_i(X_{-i}), \Phi_i(X_i) \rangle + B_i(X_i)).$$

Then the only joint distribution $P(X)$ that is consistent with these conditional distributions has the form:

$$P(X) \propto \exp(\sum_{S \subseteq [d]} \theta_S \prod_{i \in S} \Phi_i(X_i) + \sum_{i \in [d]} B_i(X_i)).$$

In other words, the set of sufficient statistics for the multivariate exponential family are specified by tensor products of the univariate exponential family sufficient statistics. (Exercise: verify that this holds for the multivariate Gaussian). [Yang et al., 2015, Inouye et al., 2017] develop parametric exponential family distributions for multivariate data using the above recipe, and further reduce the parameterization above via some deeper connections to probabilistic graphical models. These parametric classes could be enriched further via mixtures of such exponential family distributions. Nonetheless, these might not always be a good fit for data such as images, with low-level features.

3 Energy Based Models

Now instead of such "classical" parametric families, suppose we have a very expressive class of parametric functions $\{f_\theta\}$ (e.g. deep neural networks) that can approximate very complex functions very well. How do we use these for density estimation? One caveat to directly using these as a class of parametric densities is the contraint that the densities be non-negative, and integrate to one. One approach to enforce that is by parameterizing the logistic transform $\eta(x)$ instead, so that $p_\theta(x) = \frac{\exp(\eta_\theta(x))}{\int_{x \in \mathcal{X}} \exp(\eta_\theta(x)) dx}$ is non-negative and is normalizable by construction, with no further constraints on $\eta(x)$ (other than for identifiability such as that $\int_{x \in \mathcal{X}} \eta(x) = 0$, or $\eta(x_0) = 0$, for some $x_0 \in \mathcal{X}$). In some of the literature, these are referred to as "energy based models" where $\eta_\theta(x) = -E_\theta(x)$ is referred to as the negative energy, so that

higher energy is associated with lower probability (and vice versa), as a nod to statistical physics. Before we discuss approaches to train these models, it is worthwhile to briefly tour some of the classical "energy based models", which were not fully non-parameteric.

3.1 Early Energy Based Models

Hopfield Network/Ising Model

$$P(X) \propto \exp(\sum_{i,j} w_{ij} X_i X_j),$$

where $w_{ij} = w_{ji}$, and $X_i \in \{-1, +1\}$, $\forall i \in [n]$. Note that the most probable assignment to "neuron" X_i given X_{-i} is given the neuronal computation:

$$\widehat{X}_i = \text{sign}(\sum_{j \neq i} w_{ij} X_j),$$

and hence were one of the first neural networks, and that were also energy based models with energy $E_\theta(x) = -\sum_{i \neq j} w_{ij} x_i x_j$.

These could also be viewed as an instance of a categorical or discrete MEF/graphical model with binary variables and pairwise factors (i.e. allowing for terms with interactions of atmost two variables).

Boltzmann Machines The limited expressibility led to Boltzmann machines by Hinton et al. [1986], which were again over binary vectors, but also had hidden units Z (also binary, taking values in $\{-1, +1\}$) with joint distribution:

$$P(X, Z) \propto \exp(\sum_{i,j} w_{ij}^{XX} X_i X_j + \sum_{i,j} w_{ij}^{ZZ} Z_i Z_j + \sum_{i,j} w_{ij}^{XZ} X_i Z_j).$$

Even though this is also a graphical model with atmost pairwise factors, the hidden variables provide considerably more flexibility. The caveat however was that these were difficult to train, since the likelihood (or its gradient with respect to parameters) of the observed variables was not tractable, and required sampling based approximations.

Restricted Boltzmann Machines One simplification of the above was to only allow a bipartitite graph between the observed and hidden units, so that only $w^{XZ} \neq 0$: these were called restricted Boltzmann machines [Salakhutdinov et al., 2007]. These are a bit easier to train, and also have better semantics for the hidden units, since:

$$P(Z_j = 1|X) = \sigma(\sum_i W_{ij}^{XZ} X_i),$$

where $\sigma(\cdot)$ is the sigmoid function, so that the hidden units could be thought of as stochastic neural layer on top of the observed variables. (And accordingly were stacked to build deeper neural networks).

3.2 Approximations for Tractable Learning

The main caveat with energy based models is the normalization constant, which involves a multi-dimensional integral. For instance, the MLE estimate of the parameters would yield:

$$\inf_{\theta} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \eta_\theta(x_i) + \log \int_{x \in \mathcal{X}} \exp(\eta_\theta(x)) dx \right\},$$

which is in general intractable due to the multi-dimensional integral. There have been a wide series of approaches, from multiple communities (AI, non-parametric statistics, statistical physics, theoretical computer science), on approaches to address this intractability of the normalization factor. We only briefly tour these; discussing these as length would comprise its own course.

3.2.1 Sampling Based Approaches

Most of these are used to approximate the gradient of the learning objective. Computing the gradient of the MLE objective above:

$$g_\theta(x) = -\widehat{E} \nabla \eta_\theta(x) + \mathbb{E}_{P_\theta} \nabla \eta_\theta(x).$$

The second term is where the intractability comes from since it requires computing an expectation with respect to the intractable energy based model given the current parameters.

The ideal sampling based approach would be to use MCMC to sample from P_θ and use those to approximate the expectation with respect to P_θ. These however could very long to generate samples with some guarantees, and hence in practice, one might truncated set of MCMC steps, and use the resulting samples. Carreira-Perpinan and Hinton [2005] suggest using just a few MCMC steps, which they termed contrastive divergence.

3.2.2 Variational Likelihood Approximations

Variational surrogate likelihoods are another approach. Jeon and Lin [2006] for instance (in the context of more general non-parametric density estimation) suggest the following estimator:

$$\inf_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \exp(-\eta_\theta(x_i)) + \int_{x \in \mathcal{X}} \eta_\theta(x) \rho(x) dx \right\},$$

where $\rho(x)$ is any simpler known density with the same support as the true density $p(x)$. As they show, the M-estimator above is consistent, and moreover is much more tractable than the MLE. But overall, such surrogate likelihood approaches coupled with the logistic transform seem less popular, since their theoretical properties are less well-understood, and perhaps empirically they have not performed as well.

We will study more variational approximations in the sequel.

3.2.3 Contrastive Approaches

Gutmann and Hyvärinen [2010] propose to learn density ratios with respect to a known noise distribution instead of the density itself, which allows us to use contrastive or discriminative approaches to learn the models. Let Q be a given noise distribution (e.g. standard Gaussian). Suppose we learn a discriminant function $h : \mathcal{X} \to \mathbb{R}$ to contrast samples from the data distribution P vs samples from Q via the cross-entropy loss:

$$\min_{h} \mathbb{E}_{X \sim P} \ln h(X) + \mathbb{E}_{X \sim Q} \ln[1 - h(X)].$$

The optimal discriminant function can be seen to be the density ratio:

$$h(x) = \frac{P(x)}{P(x) + Q(x)},$$

which we can thus use to construct our density estimate:

$$\widehat{P}(x) = Q(x) \frac{\widehat{h}(x)}{1 - \widehat{h}(x)},$$

given the learnt discriminant function $\widehat{h}$.

We can also use this noise contrastive approach to fit an explicit parameterized energy based model $P_\theta(x) \propto \exp(\eta_\theta(x))$. Assuming the true data distribution also follows this energy based model for some θ^*, the optimal discriminant function would have the form:

$$h^*(x) = \frac{\exp(\eta_{\theta^*}(x) + c^*)}{\exp(\eta_{\theta^*}(x) + c^*) + Q(x)}.$$

We can thus fit a discriminant function from the following class of functions:

$$h(x) = \frac{\exp(\eta_\theta(x) + c)}{\exp(\eta_\theta(x) + c) + Q(x)}.$$

Gutmann and Hyvärinen [2010] show that doing so leads to a well-defined discriminant learning problems, that is, optimizing over the constant c that represents the log-normalization constant does not lead to an unbounded objective. As they show, there is a slight statistical inefficiency in using such a noise contrastive approach to learning energy based models. Liu et al. [2021] further show that when the noise distribution Q is far from the true data distribution P_{θ^*}, the landscape of the classification objective becomes very flat far away from the optimum (so that the classification objective itself could be very small, but the distance in parameter space is very large). They suggest that normalized gradients over an appropriately chosen surrogate classification objective could ameliorate some of these landscape challenges.

3.2.4 Score Matching

Consider the score function:

$$s(x) = \frac{\partial}{\partial x} \log p(x).$$

It can be seen that for an energy based model $p_\theta(x)$, the score simplifies to:

$$s(x) = \frac{\partial}{\partial x} \eta_\theta(x),$$

and in particular does not need to grapple with the log-partition function. Consider the "score matching" objective:

$$J(\theta) = \frac{1}{2} \mathbb{E} \| s_\theta(X) - s(X) \|^2,$$

that aims to match the score of the energy based model with respect to true score function.

Hyvärinen and Dayan [2005] show that the above can be re-written as:

$$J(\theta) = \mathbb{E} \sum_{j=1}^{d} \left(\frac{\partial}{\partial X_i} s_{\theta;i}(X) + \frac{1}{2} s_{\theta;i}^2(X) \right) + \text{const.},$$

so that one could estimate the above via:

$$\widehat{J}(\theta) = \widehat{\mathbb{E}} \sum_{j=1}^{d} \left(\frac{\partial}{\partial X_i} s_{\theta;i}(X) + \frac{1}{2} s_{\theta;i}^2(X) \right) + \text{const.},$$

given samples $\{x_i\}_{i \in [n]}$ drawn from P.

As score matching (and contrastive learning) finds wider usage, their statistical and optimization landscape caveats are increasingly being analyzed. Koehler et al. [2022] for instance show that for energy based models with a large isoperimetric constant (loosely: worst case over sets of ratio of probability mass of set boundary over probability mass of the set itself) score matching can be very inefficient.

4 Neural Generative Models

Over the last decade there have been a slew of alternative approaches that sidestep the energy model/logistic transform route, with its normalization difficulties altogether, and specify the random vector X as a *transformation* of some other latent variables Z with some known distribution. These transformations in general can be relatively unconstrained, so that we sidestep issues of normalizability. These transformations typically involve deep neural network based parametric functions, and hence are loosely called deep density estimators. Let us consider various classes of these generative models in the sequel.

5 Normalizing Flows

Suppose that we have a latent representation $Z \sim N(0, I)$, and that we have a deterministic transformation from Z to the data X as:

$$X = g_\theta(Z),$$

for some flexible parametric function g_θ. Suppose g_θ is invertible (which is a big if). Then by the change of variables formula:

$$p_{X;\theta}(x) = p_Z(g_\theta^{-1}(x))|\text{det}Jg^{-1}(x)|,$$

where $[Jh(x)]_{ij} = \partial h_i(x)/\partial x_j$, so that the density has a nice closed form expression. Thus, given samples $\{x_i\}_{i=1}^n$, we could thus directly solve for the MLE:

$$\inf_\theta \sum_{i=1}^n -\log p_{X;\theta}(x_i).$$

Note that these can be stacked, so that we could obtain a stacked transformation $Z_K = g_K \circ \ldots \circ g_1(Z_0)$, which in turn will have the log-density:

$$\log p_K(z_K) = \log p_0(z_0) - \sum_{k=1}^K \log |\text{det}Jg_k(z_k)|.$$

The random variables Z_k are called "flows," and the distributions P_k are called "normalizing flows" [Rezende and Mohamed, 2015].

Note that by the so-called reparameterization trick introduced earlier $E_{p_K}[h(X)] = E_{p_0}[h(g_K \circ \ldots \circ g_1(z_0))]$ which does not involve Jacobian calculations.

5.1 Invertible Maps

The key caveats with normalizing flows are two-fold: (a) the transformation g_θ has to be invertible, and (b) the density involves the Jacobian of the transformation, which could be expensive for general invertible maps.

Some simple classes of invertible transformations (which as noted above can be stacked to get "deep" flow tranforms) include:

$$g(z) = z + uh(w^T z + b),$$

which are invertible for specific settings of (h, u, w) e.g. $h = \tanh(\cdot)$ and $w^T u >= -1$ [Rezende and Mohamed, 2015].

An alternative approach, called NICE [Dinh et al., 2014], is to split $X = (X_1, X_2)$ as well as $Z = (Z_1, Z_2)$ into two blocks of variables with the blocked transform:

$$\begin{aligned} X_1 &= Z_1 \\ X_2 &= Z_2 + m(Z_1), \end{aligned}$$

for an arbitrary, potentially non-invertible function $m(\cdot)$. It can be seen that the transformation from Z to X is trivially invertible:

$$\begin{aligned} Z_1 &= X_1 \\ Z_2 &= X_2 - m(X_1). \end{aligned}$$

Moreover, the Jacobian of the transformation is triangular, so that its determinant is simply the product of diagonal entries, and hence easy to compute. A related triangular Jacobian transformation [Dinh et al., 2016] is given by:

$$\begin{aligned} X_1 &= Z_1 \\ X_2 &= Z_2 \odot \exp(m_1(Z_1)) + m_2(Z_1), \end{aligned}$$

which can again be trivially inverted for arbitrary $m_1(\cdot), m_2(\cdot)$, via:

$$\begin{aligned} Z_1 &= X_1 \\ Z_2 &= (X_2 - m_2(X_1)) \odot \exp(-m_1(X_1)). \end{aligned}$$

6 Autoregressive Flows

The simple triangular Jacobian examples had an implicit auto-regressive character: we could specify the joint distribution via the marginal distribution of a subset of variables X_1, and

the conditional distribution of the remaining subset X_2 conditioned on X_1. Autoregressive flows generalize this to allow for more general auto-regressive transformations.

In a so-called Masked Autoregressive Flow (MAF) [Papamakarios et al., 2017], this is given as:

$$X_i = \mu_i + Z_i \exp(\alpha_i),$$

where $Z_i \sim N(0,1)$, and $\mu_i = g_{\mu_i}(X_{<i})$, and $\alpha_i = g_{\sigma_i}(X_{<i})$, so that X is a transformation of the standard Gaussian vector Z, and where the transformation is specified in an auto-regressive manner. It can be seen that the inverse is easily computed:

$$Z_i = (X_i - \mu_i) \exp(-\alpha_i),$$

so that Z can be recovered given X, and that moreover the determinant of the Jacobian of the transformation $X = g(Z)$ is easily computed as $|\det J g^{-1}(x)| = \exp(-\sum_i \alpha_i)$. MAF can transform X to Z in one (parallelized) iteration: since the information to specify each Z_j is fully available in X, and we do not need to wait to compute $Z_{<j}$. But then we require p iterations to transform Z to X: since to specify X_j it does not just suffice to provide Z, but also $X_{<j}$.

A variant of MAF is Inverse Autoregressive Flow (IAF) [Kingma et al., 2016], where we have:

$$X_i = \mu_i + Z_i \exp(\alpha_i),$$

where $\mu_i = g_{\mu_i}(Z_{<i})$, and $\alpha_i = g_{\alpha_i}(Z_{<i})$. Its inverse is again given as:

$$Z_i = (X_i - \mu_i) \exp(-\alpha_i),$$

but note that in this case, IAF can transform Z to X in one vectorized iteration, but requires p iterations to transform X to Z. Thus the slight difference in choices between IAF and MAF can result in vastly different computational times for specific tasks. Note that transform X to Z is required for calculating the density of a point X, while transforming Z to X is required to generate new samples.

Stacking such auto-regressive transformations $X = g_K \circ \ldots \circ g_1(Z)$ is called an "autoregressive flow", as a special instance of normalizing flows.

6.1 General Auto-regressive Distributions

Classically, auto-regressive models were used to directly parameterize joint distributions (rather than simply transformations) via parameterizing conditional distributions specified by the standard chain rule

$$p_\theta(x) = \prod_{i=1}^{p} p_\theta(x_i | x_{<i}).$$

A classical approach to model $p_\theta(x_i|x_{<i})$ is to make a Markov assumption that $p_\theta(x_i|x_{<i}) = p_\theta(x_i|x_{i-1}, \ldots, x_{i-k})$ so that the conditional distribution of X_i conditioned on all previous variables only depends on the k most recent variables before X_i. Another approach is to use sequence model based recurrences, such as:

$$h_i = f_{\theta_h}(h_{i-1})$$
$$x_i = f_{\theta_x}(x_{i-1}, h_i),$$

for some parametric functions $f_{\theta;h}(\cdot)$, and $f_{\theta;x}(\cdot, \cdot)$. Such recurrence based sequence models, such as recurrent neural networks (RNNs), were popular parametric models for sequence based data where the sequence order is very naturally specified, for instance, via time. But they are far less popular when there is no such natural sequence order, in large part because the performance of such models is very sensitive to such ordering. Papamakarios et al. [2017], in their Figure 1, provide an example with two variables, where an auto-regressive model with the ordering (x_1, x_2) is able to model the data, but an auto-regressive model with the ordering (x_2, x_1) is not able to.

One approach to address this to use different orderings, and use an ensemble or mixture of the resulting distributions. Another approach is to use different orderings in each layer of an "autoregressive flow" $X = g_K \circ \ldots \circ g_1(Z)$, where we use a different ordering for each auto-regressive transformation g_i, for $i = 1, \ldots, K$. This provides another rationale for the use of auto-regressive flows, rather than sequence based auto-regressive recurrence models, in addition to other benefits of normalizing flows, such as the ease of computing the density, and sampling.

7 Generative Adversarial Networks (GANs)

Suppose we have a parametric transformation $X = g_\theta(Z)$ of some base distribution z. Provided the transformation g_θ is invertible as with normalizing flows, we can compute the density $p_\theta(x)$ of the random vector x, and consequently use MLE:

$$\widehat{\theta} \in \arg\inf_\theta \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i),$$

to estimate the parameters $\widehat{\theta}$ given samples $\{x_i\}_{i=1}^n$. There are two caveats here. The first is that this is not feasible when g_θ is not invertible, which would be the case for instance, for most modern architectures of deep neural networks. The second caveat is more subtle, and is due to the very nature of the MLE as minimizing the empirical variant of the KL divergence between the true data distribution P and the distribution P_θ over X with density p_θ:

$$\inf_\theta \mathrm{KL}(P, P_\theta).$$

Note that $KL(P, P_\theta) = \int p(x) \log p(x)/p_\theta(x) dx$, so that this would be large if there are P-likely regions where $p_\theta(x)$ is small and $p(x)$ is large: which encourages $p_\theta(x)$ to have support in the P-likely regions of the input space. But this does not ensure that $p_\theta(x)(x)$ be small where $p(x)$ is small: such a property would ne required to ensure that samples from $P_\theta(x)$ be P-realistic (i.e. do not have small density with respect to true data distribution P). How do we encourage the latter property? By simply minimizing $KL(P_\theta, P) = \int p_\theta(x)(x) \log p_\theta(x)(x)/p(x) dx$, which would be large if there are $P_\theta(x)$-likely regions where $p(x)$ is small and $p_\theta(x)$ is large. A caveat with $KL(P_\theta, P)$ on the other hand is practical: it is not decomposable, so that it is not clear how to optimize this given just samples $\{x_i\}_{i=1}^n$ from P. Combining both these asymmetric KL divergences yields the Jensen-Shannon divergence:

$$\text{JSD}(P, P_\theta(x)) = \frac{1}{2}\text{KL}\left(P\,, \frac{P + P_\theta(x)}{2}\right) + \frac{1}{2}\text{KL}\left(P_\theta(x)\,, \frac{P + P_\theta(x)}{2}\right),$$

which has the additional advantage of being symmetric in its arguments. This loss again is not decomposable, so that it is not clear how to optimize this given just samples $\{x_i\}_{i=1}^n$ from P. In a seminal paper, Goodfellow et al. [2014] showed that one can indeed minimize the Jensen-Shannon divergence given samples by considering a variational form using "generators" and "discriminators".

Suppose $D : \mathcal{X} \mapsto [0, 1]$ be a classifier (ideally probabilistic, but more generally with an output of classifier scores between 0 and 1). Given the parameterized density q_θ, and the density of the true data distribution p, consider the following variational form:

$$V(p, q_\theta, D) = \mathbb{E}_{x\sim p}[\log D(x)] + \mathbb{E}_{x\sim q_\theta}[\log(1 - D(x))].$$

Goodfellow et al. [2014] then showed the following useful result:

$$\max_D V(p, q_\theta, D) = -\log(4) + 2\text{JSD}(p, q_\theta),$$

so that

$$\arg\min_\theta \max_D V(p, q_\theta, D) = \arg\min_\theta \text{JSD}(p, q_\theta).$$

The interesting facet of $V(p, q_\theta, D)$ is that it is decomposable, so that it can be approximated well via samples (from both p as well as q_θ), thus facilitating learning the parameters of the density q_θ by minimizing the Jensen-Shannon divergence itself with respect to the true data distribution.

The variational objective $V(p, q_\theta, D)$ can also be motivated as specifying a min-max adversarial game between the "generative" density q_θ, and a discriminator D that aims to discriminate between samples from Q_θ and P, while the generator Q_θ aims to fool the discriminator D. Specifically, consider the following classification task, where $Y = 1$ indicates the true data distribution and $Y = 0$ indicates Q_θ, so that $X|(Y = 1) \sim P$, and $X|(Y = 0) \sim Q_\theta$. The

expected cross-entropy loss of a probabilistic discriminator $D : \mathcal{X} \mapsto [0, 1]$ is then given by

$$\begin{aligned}
&\mathbb{E}\left[Y \log D(X) + (1 - Y) \log(1 - D(X))\right] \\
&= \mathbb{E}[\log D(X)|Y = 1]P(Y = 1) + \mathbb{E}[\log(1 - D(X))|Y = 0]P(Y = 0) \\
&= 0.5 * \mathbb{E}_{X\sim P}[\log D(X)] + 0.5 * \mathbb{E}_{X\sim Q_\theta}[\log(1 - D(X))] \\
&= 0.5 * V(p, q_\theta, D)
\end{aligned}$$

using $P(Y = 1) = P(Y = 0) = 1/2$.

Thus, the variational objective $V(p, q_\theta, D)$ is simply twice the expected cross entropy loss of the discriminator $D(\cdot)$ in the classification task of discriminating between the true distribution P and the generative model Q_θ.

8 Variational Auto-Encoders

A key caveat to the above approaches is that they might not be able to capture the multimodality of the inputs given the latent representation if X is a deterministic function of Z. This prevents the latent variables from only representing "coarser" information about the input. One way to accomodate this is to allow X to be a stochastic function of Z.

A very natural approach [Kingma and Welling, 2013] along these lines is to have:

$$\begin{aligned}
Z &\sim N(0, I_d) \\
X|Z = z &\sim N(\mu_\theta(z), \sigma^2_\theta(z)I),
\end{aligned}$$

or alternatively:

$$X = \mu_\theta(Z) + \sigma_\theta(Z)\, W$$

where $Z, W \sim N(0, I)$. Thus, X has a well-defined density even when the mean and variance functions $\{\mu_\theta(z), \sigma_\theta(z)\}_{\theta\in\Theta}$ are relatively unconstrained, with no normalization terms. For instance, a highly expressive parametric family of choice for these mean and variance functions are deep neural networks. The density of X is then given as:

$$p(X; \theta) = \int_{z\in\mathbb{R}^d} p_N(x; \mu_\theta(z), \sigma^2_\theta(z)I)\, p_N(z; 0, I)dz,$$

where $p_N(\mu, \Sigma)$ is the multivariate Gaussian density. It can be seen that the density does not have an explicit tractable form. To fit the parameters θ to the data, we could maximize the likelihood of the observed data:

$$\max_\theta \frac{1}{n} \sum_{i=1}^{n} \log p(x_i; \theta),$$

but this has all of the difficulties of fitting a latent generative model.

As before, we could optimize a surrogate likelihood instead. In so-called variational inference, we compute parameterized lower bounds of the likelihood and optimize this lower bound instead. Thus, if $p_\theta(X) \geq g_{\theta;\gamma}(X)$, for $\gamma \in \Gamma$, then we solve for:

$$\max_{\theta \in \Theta, \gamma \in \Gamma} \frac{1}{n} \sum_{i=1}^{n} \log g_{\theta;\gamma}(x_i).$$

With the above latent variable model, we have the following classical variational bound, also called the Evidence Lower Bound or ELBO:

$$\begin{aligned}
\log p_\theta(x) &= \log \int_z p_\theta(x, z) dz \\
&= \log \int_z q_\phi(z|x) p_\theta(x|z) p(z)/q_\phi(z|x) \\
&\geq \int_z q_\phi(z|x) \log p(z)/q_\phi(z|x) + \int_z q_\phi(z|x) \log p_\theta(x|z) \\
&= L(\theta, \phi, x) := -\mathrm{KL}(q_\phi(z|x) \| p(z)) + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)],
\end{aligned}$$

so that instead of the maximizing the empirical expectation of $\log p_\theta(x)$, we maximize the empirical expectation of the lower bound $L(\theta, \phi, x)$ instead. We can moreover show that:

$$\log p_\theta(x) - L(\theta, \phi, x) = \mathrm{KL}(q_\phi(z|x) || p_\theta(z|x)),$$

so that the ELBO bound gets tighter as the variational approximation $q_\phi(z|x)$ gets closer to the intractable true posterior $p_\theta(z|x)$.

A natural flexible parameterization is simply:

$$q_\phi(z|x) = N(\mu_\phi(x), \sigma_\phi^2(x) I),$$

where again the mean and variance functions $\mu_\phi(x), \sigma_\phi(x)$, can again be parameterized by flexible families such as deep neural networks. Note that when taking an expectation $\mathbb{E}_{q_\phi(z|x)}[f(z)]$, we can "reparameterize" $z = h(x, w) := \mu_\phi(x) + \sigma_\phi(x) w$ in terms of $w \sim N(0, I)$, so that

$$\mathbb{E}_{q_\phi(z|x)}[f(z)] = \mathbb{E}_{w \sim N(0,I)}[f(h(w, x))],$$

which can be approximated via Monte Carlo samples of w, and no explicit density calculations of $q_\phi(z|x)$. This is called the "reparameterization trick".

The above approach is also called the Variational Auto-encoder, since it is reminiscent of auto-encoders that were used to learn compact representations of the input x. As an instance, suppose we wish to get a representation $z \in \mathbb{R}^d$ of the input $x \in \mathbb{R}^p$ for some $d < p$, via the following "encoder" model:

$$z = g(b + Wx),$$

which is then coupled with a "decoder" model:

$$\widehat{x} = g(c + Vz),$$

for some point-wise non-linearity $g(\cdot)$, and some vectors b, c, and matrices W, V. We could learn these parameters by minimizing the reconstruction error:

$$\inf_{\theta} \sum_{i} \|\widehat{x}_i - x_i\|.$$

Here the "encoder" transformation from x to z, as well as a "decoder" transformation from z to x are both deterministic, and hence this does not specify a density model for x. With the variational autoencoder, both these transformations are stochastic, and moreover, there was an explicit distribution imposed on the latent representations z, which thus specified a distribution over the inputs x. Additionally, the traditional auto-encoders, in order to learn a non-identity transformation (implemented via arbitrary encoders and decoders that are inverses of each other) either use a bottleneck (where dimensionality of encoding z is much smaller than that of x), or add noise to the input x and aim to predict the denoised input (with the idea that then the encoding z is forced to capture the salient information about x). The variational auto-encoder could thus be viewed as a principled Bayesian approach to *denoising auto-encoders*.

While in the original variational auto-encoder, $q_\phi(z|x)$ was set to be a Gaussian with parameterized mean and variance, one could also use other flexible parameterizations, including the invertible neural networks or normalizing flows to be discussed in the next section.

One could also use a stacked set of latent Gaussians as:

$$\begin{aligned} z_L &\sim N(0, I) \\ z_l | z_{l+1} &\sim N(\mu_l(z_{l+1}), \sigma_l^2(z_{l+1})I) \\ x | z_1 &\sim N(\mu_0(z_1), \sigma_0^2(z_1)I) \end{aligned}$$

in what are called Deep Latent Gaussian Models [Rezende et al., 2014], though these seem less popular, perhaps due to the added complexity.

9 Destructive Distribution Learning

So far we have considered a largely "constructive" learning approach where we learn a transformation $g_\theta(Z)$ of a random vector Z with known simple distribution (such as independent Gaussian) and fit the parameters so that the transformed distribution $P_{g_\theta(Z)}$ is as close to the true data distribution P_X as possible, for instance by solving for the following (population) objective:

$$\inf_{\theta} KL(P_X, P_{g_\theta(Z)}).$$

An alternative approach is to consider a "destructive" learning approach where we learn a transformation $h_\theta(X)$ of the data random vector X, and fit the parameters so that the transformed distribution $P_{h_\theta(X)}$ is as close to a random vector Z with known simple distribution (such as independent Gaussian). Such a transformation is called destructive learning since we aim to "destroy" the structure in X, reducing it to say an independent Gaussian distribution.

But while (imperfectly) transforming Z to X seems useful from a density estimation perspective, why would we want to (imperfectly) transform X to Z? There are two reasons to do so. The first is that of representation learning, as we discuss further below. The second is that we could also use it as density estimation procedure.

9.1 Representation Learning

Tranforming X to Z with known or simple distribution could be cast as "encoding" the data X into a representation that is "simple". A variant of this is Independent Component Analysis (ICA), where we (typically) only assume that Z is independent. This is not however sufficient to make the transformation identifiable. For instance, if Z_1 and Z_2 are independent, then so are component-wise transformations $f_1(Z_1)$ and $f_2(Z_2)$. Even if we restrict the distribution of the independent vector Z, the indeterminacy remains. Suppose we have a transformation $h : \mathcal{X} \mapsto [0,1]^d$ that maps X to a uniform random vector $Z \in [0,1]^d$. Then, given any measure-preserving automorphism $g : [0,1]^d \mapsto [0,1]^d$, it is clear that $g \circ h$ will be another solution to the ICA problem of transforming X to a uniform random vector.

So, such a destructive mapping, if it exists, is not unique. But what about existence of such a mapping? One can show this via the following constructive mapping.

Suppose we set $Z_1 = F_1(X_1) \sim \text{Unif}[0,1]$. And for $j = 2, \ldots, d$, denote $F_j(x; z_1, ..., z_j) = P(X_j \leq x | Z_1 = z_1, \ldots, Z_j = z_j)$ as the conditional CDF of X_j conditioned on j uniform random variables $\{Z_\ell\}_{\ell=1}^j$. Then set $Z_{j+1} = F_j(X_j; z_1, \ldots, z_j)$. It can be seen that $(Z_1, ..., Z_d) \sim \text{Unif}[0,1]^d$. This is simply the multivariate extension of the classical univariate CDF transformation result that for any real-valued random variable$V \in \mathbb{R}$, if F_V is its CDF, then $F_V(V) \sim \text{Unif}[0,1]$. Thus, stitching these conditional CDF transformations together, we get the mapping: $Z = h(X)$. The main caveat with this constructive mapping is that such conditional CDFs are difficult to estimate for multivariate data.

9.2 Density Estimation

The other reason we might want to learn an imperfect mapping $h_\theta(\cdot)$ from X to Z is that in the limit where we are able to truly convert X to Z with known distribution, then we can recover the density of X by the change of variable formula applied to the transformation

$h_\theta^{-1}(Z)$, so that

$$p_{h_\theta^{-1}(Z)}(x) = p_Z(h_\theta(x))|\det Jh_\theta(x)|,$$

as with normalizing flows. Since h_θ is an imperfect transformation of X to Z, similarly, h_θ^{-1} will be an imperfect transformation from Z to X, which is the case with constructive approaches such as normalizing flows as well. A more critical concern with the destructive learning approach is computational/practical.

Consider the objective:

$$\begin{aligned}&\inf_\theta \mathrm{KL}(P_{h_\theta(X)}, P_Z)\\ =&\inf_\theta \int_z p_{h_\theta(X)}(z) \log p_{h_\theta(X)}(z)/p_Z(z) dz.\end{aligned}$$

It can be seen that it is not clear how to optimize this objective given just samples $\{x_i\}_{i=1}^n$ drawn from P_X, since without access to the true density p_X, we might not be able to evaluate the transformed density $p_{h_\theta(X)}(z)$ (note that in the case of normalizing flows, we had access to the base density p_Z, and so could evaluate the density of transformations of this base density). But the following simple identity essentially notes reduces it to constructive learning:

Theorem 1 (Destructive-Constructive Identity)

$$KL(P_{h_\theta(X)}, P_Z) = KL(P_X, P_{h_\theta^{-1}(Z)}).$$

This theorem has (re-)appeared in many recent generative model papers; see for instance [Ballé et al., 2015, Papamakarios et al., 2017]. The proof just follows from some applications of the change of variables formula:

$$\begin{aligned}\mathrm{KL}(P_{h_\theta(X)}, P_Z) &= \int_z p_{h_\theta(X)}(z) \left(\log p_{h_\theta(X)}(z) - \log p_Z(z)\right) dz\\ &= \int_x p_X(x) \left|\det \frac{\partial x}{\partial z}\right| \left(\log p_X(x) \left|\det \frac{\partial x}{\partial z}\right| - \log p_Z(h_\theta(x))\right) \left|\det \frac{\partial z}{\partial x}\right| dx\\ &= \int_x p_X(x) \left(\log p_X(x) - \log p_Z(h_\theta(x)) \left|\det \frac{\partial z}{\partial x}\right|\right) dx\\ &= \int_x p_X(x) \left(\log p_X(x) - \log p_{h_\theta^{-1}(Z)}(x)\right) dx\\ &= \mathrm{KL}(P_X, P_{h_\theta^{-1}(Z)}),\end{aligned}$$

where $\frac{\partial x}{\partial z} = Jh_\theta^{-1}(h_\theta(x))$, and $\frac{\partial z}{\partial x} = Jh_\theta(x)$, and where we used the property of Jacobians that $\frac{\partial x}{\partial z} = (\frac{\partial z}{\partial x})^{-1}$, and the property of determinants that $\det(A^{-1}) = 1/\det(A)$.

Thus, destructive learning is equivalent to constructive learning with invertible transformations, so that it is not clear why we should not simply use constructive learning approaches

if we care about density estimation. One methodological advantage could be that we could use insights from other fields to obtain invertible "destructive" transformations. For instance [Ballé et al., 2015] suggest the following "divisive normal" transformation from X to Z:

$$U = H\,X$$
$$Z_i = \frac{U_i}{(\beta_i + \sum_{j=1}^{d} \gamma_{ij}|U_j|^{\alpha_{ij}})^{\epsilon_i}},$$

for some parameter matrices $H, \alpha, \gamma \in \mathbb{R}^{d\times d}$, and parameter vectors $\beta, \epsilon \in \mathbb{R}^d$ which they motivate from neuroscience considerations.

There are also particular classes of transformations where the solution of the destructive learning problem is easier. Consider the class of transformations of the following form:

$$\mathcal{H} = \{h : \mathcal{X} \subseteq \mathbb{R}^d \mapsto \mathbb{R}^d \,|\, h = \Psi(Ax),$$

where $A \in \mathbb{R}^{d\times d}$, and $\Psi(u) = (\Psi_1(u_1), \ldots, \Psi_d(u_d))$, where $\{\Psi_j\}_{j=1}^d$ are pointwise invertible univariate transformations. Thus, the transformations $h = \Psi(Ax)$ consist of a linear transformation, followed by coordinatewise transformations.

We then wish to solve for:

$$\inf_{\Psi, A} \mathrm{KL}(P_{h_{\Psi,A}(X)}, P_Z).$$

The solution (Ψ^*, A^*) to this can be characterized simply:

$$A^* = \arg\inf_A I(AX),$$
$$\Psi_j^* = \Phi^{-1} \odot F_{\langle A_j^*, X\rangle}(\langle A_j^*, X\rangle),$$

where $I(\cdot)$ is the mutual information of a random vector, and $\Phi(\cdot)$ is the standard Gaussian CDF. Minimizing $I(AX)$ over matrices A is essentially the linear ICA problem, which aims to find a linear transformation of X that reduces dependence among the transformed variables as much as possible. While Ψ_j^* is simply a univariate Gaussianization transform, which is a composition of the univariate CDF of the linear transformed variable, and an inverse of the standard Gaussian CDF. Both of these have practical if approximate implementations.

To see why the solution has such a nice closed form, let us first define the marginal KL divergence: marginal-$\mathrm{KL}(P_U, P_V) := \sum_{j=1}^d \mathrm{KL}(P_{U_j}, P_{V_j})$, as the sum of the KL divergences between the corresponding d marginal distributions. From some algebraic calculations, we can then write

$$\mathrm{KL}(P_{\Psi\circ AX}, P_Z) = \text{marginal-KL}(P_{\Psi\circ AX}, P_Z) + I(\Psi \circ AX).$$

But for invertible Ψ, $I(\Psi \circ AX) = I(AX)$, so that Ψ can be obtained by minimizing just the first term, which yields that it is the pointwise Gaussianization of the optimal linear

transformation of X. Given this optimal Ψ^*, the first term becomes zero, so that the objective then reduces to the second term which is precises the linear ICA objective $I(AX)$. Thus, one could simply solve for linear ICA to obtain the matrix A.

One can also use the decomposition above to suggest a linear ICA algorithm. Note that if we restrict A to be orthogonal, we then have that:

$$\mathrm{KL}(P_X, P_Z) = \mathrm{KL}(P_{AX}, P_Z) = \text{marginal-KL}(P_{AX}, P_Z) + I(AX).$$

Since the LHS does not depend on A, we then get that:

$$\inf_A I(AX) = \sup_A \text{marginal-KL}(P_{AX}, P_Z),$$

so that we aim to find a linear transformation A that makes the coordinates of AX be as non-Gaussian as possible. The overall optimal solution then seems very intuitive: A aims to find the directions under which the projection of X is most non-Gaussian. Ψ then marginally Gaussianizes these transformed variables.

Both the destructive transforms above might not seem that flexible. But one advantage of destructive transformations such as the above is that we can iterate over these, destroying a bit of X in every iteration. So, starting with $X_1 = X$, in iteration $t = 1, \ldots$, we find $h_t = \arg\inf_{h \in \mathcal{H}} \mathrm{KL}(P_{h(X_t)}, P_Z)$, and then destroy X_t as $X_{t+1} = h_t(X_t)$. A consistency property that would be good to have is that $\mathrm{KL}(P_{X_t}, P_Z) \to 0$. This was shown to indeed be the case with the Gaussianization transformation above [Chen and Gopinath, 2001].

One could also view the iterates above as greedily learnt stacked destructors: $X_{t+1} = h_t \circ \ldots \circ h_1(X)$. Given the equivalence in the beginning of the section, Inouye and Ravikumar [2018] thus suggested the following general algorithm for destructive learning:

$$g_t = \arg\inf_{g \in \mathcal{G}} \mathrm{KL}(P_{X_t}, P_{g(Z)}),$$

for some simple class of invertible functions $\mathcal{G}$, and then use $X_{t+1} = g_t^{-1}(X_t)$. This thus generalizes the Gaussianization transforms above to a much larger class of generative models, that simply fit a generative model over the current data iterate X_t, and then use this generative model to extract a destructive transform to further transform the data and iterate. Note that such a destructive iterative algorithm is much more computationally feasible than a constructive iterative algorithm that would aim to solve:

$$g_t = \arg\inf_{g \in \mathcal{G}} \mathrm{KL}(P_X, P_{g(Z_t)}),$$

where $Z_t = g_{t-1} \circ \ldots \circ g_1(Z)$, where the main computational concern is the computation of the densities of Z_t.

Inouye and Ravikumar [2018] also show that even if we simply solve for simple or shallow density estimation via:

$$Q_t = \arg\inf_{Q \in \mathcal{Q}} \mathrm{KL}(P_{X_t}, P_Q),$$

one can in most cases extract a destructive transform $h(\cdot)$ from P_Q such that $P_Q = P_{h(Z)}$. This further increases the ease of each destructive iteration: one performs shallow density estimation using their method of choice, extract the corresponding destructive transform, and use this to further transform the data, and iterate. So by a series of shallow density estimation procedures, we are able to fit a "deep" density destructively.

One simple approach to extract the invertible destructor $h(\cdot)$, for a given Q (so that $P_{h(Z)} \equiv P_Q$) is to use the conditional univariate CDF transformations discussed earlier. But in most cases such destructive transforms can be obtained even more simply. See [Inouye and Ravikumar, 2018] for examples with many common used shallow densities.

References

Eunho Yang, Pradeep Ravikumar, Genevera I Allen, and Zhandong Liu. Graphical models via univariate exponential family distributions. *The Journal of Machine Learning Research*, 16(1):3813–3847, 2015.

David I Inouye, Eunho Yang, Genevera I Allen, and Pradeep Ravikumar. A review of multivariate distributions for count data derived from the poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(3):e1398, 2017.

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

Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798, 2007.

Miguel A Carreira-Perpinan and Geoffrey Hinton. On contrastive divergence learning. In *International workshop on artificial intelligence and statistics*, pages 33–40. PMLR, 2005.

Yongho Jeon and Yi Lin. An effective method for high-dimensional log-density anova estimation, with application to nonparametric graphical model building. *Statistica Sinica*, pages 353–374, 2006.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

Bingbin Liu, Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. Analyzing and improving the optimization landscape of noise-contrastive estimation. *arXiv preprint arXiv:2110.11271*, 2021.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. *arXiv preprint arXiv:2210.00726*, 2022.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

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

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

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

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751, 2016.

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

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

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and variational inference in deep latent gaussian models. In *International Conference on Machine Learning*, volume 2, 2014.

Johannes Ballé, Valero Laparra, and Eero P Simoncelli. Density modeling of images using a generalized normalization transformation. *arXiv preprint arXiv:1511.06281*, 2015.

Scott Saobing Chen and Ramesh A Gopinath. Gaussianization. In *Advances in neural information processing systems*, pages 423–429, 2001.

David Inouye and Pradeep Ravikumar. Deep density destructors. In *International Conference on Machine Learning*, pages 2167–2175, 2018.