Nonparametric Approaches for Training Deep Generative Networks

Kubra Cilingir, Brian Kulis
Department of Elec. and Comp. Engg.
Boston University
Boston, MA 02215, USA
{kubra,bkulis}@bu.edu

Manesh K. Singh
AI Innovation Lab
Verisk Analytics
Jersey City, NJ 08540, USA
maneesh.singh@verisk.com

Abstract

Algorithms that learn to generate data whose distributions match that of the training data, such as generative adversarial networks (GANs), have been a focus of much recent work in deep unsupervised learning. Unfortunately, GAN models have drawbacks, such as instable training due to the minmax optimization formulation and the issue of zero gradients. To address these problems, we explore and develop a new family of nonparametric objective functions and corresponding training algorithms to train a DNN generator that learn the probability distribution of the training data. Preliminary results presented in the paper demonstrate that the proposed approach converges faster and the trained models provide very good quality results even with a small number of iterations. Special cases of our formulation yield new algorithms for the Wasserstein and the MMD metrics. We also develop a new algorithm based on the Prokhorov metric between distributions, which we believe can provide promising results on certain kinds of data. We conjecture that the nonparametric approach for training DNNs can provide a viable alternative to the popular GAN formulations.

1 Introduction

Generative adversarial networks (GANs) have demonstrated an impressive ability to learn complex, implicit, data distributions, especially in the computer vision domain. Nonetheless, it is widely acknowledged that they are unstable and hard to train, prompting active and vigorous research to analyze and improve upon them. Salimans et al. (2016), Metz et al. (2016), Arora et al. (2017). Data distributions typically lie in low dimensional manifolds in a high dimensional embedding space. Fefferman et al. (2016). This causes difficulty for learning algorithms due to the phenomenon of vanishing gradients. Arjovsky & Bottou (2017). Another problem arises due to well-known difficulties in optimizing min-max formulations in adversarial settings. Goodfellow (2016).

We propose an alternative approach, based on nonparametric loss functions, to help mitigate both the above issues. We describe a general class of nonparametric loss functions and provide two instantiations, based on the Prokhorov and Wasserstein metrics (popularized by WGANs Arjovsky et al. (2017)) and provide novel learning algorithms based on them. We chose the Prokhorov metric since it metrizes weak convergence of measures and ensures the convergence of the two distributions regardless of their respective supports. Pollard (2012). We compared the proposed nonparametric approach to vanilla GANs Li et al. (2015) on the MNIST dataset. Preliminary results point to faster and more stable convergence, with very few iterations (as low as 20-30).

2 Approach

Recall the setup for GANs. Let \( \{x_{r}^{(i)}\}_{i=1}^{n} \) be samples from the real distribution (training data), and let \( \{x_{g_{\theta}}^{(i)}\}_{i=1}^{m} \) be samples produced by the generative model, \( g_{\theta} \) with instantiated parameters \( \theta \). \( g_{\theta} \) takes samples from a prior \( \{z^{(i)}\}_{i=1}^{m} \sim p_{z}(z) \) as input and deterministically transforms these samples to \( \{x_{g_{\theta}}^{(i)}\}_{i=1}^{m} \).
2.1 General Form of Nonparametric Loss Functions

A generic nonparametric loss function, $L$, can be represented as the following where $C \subseteq \mathbb{R}^{n \times m}$ is appropriately defined feasible region.

$$L(\theta) = \min_{\{a_{ij}\} \in C} \sum_{i,j} a_{ij} d(x^{(i)}_r, g_\theta(z^{(j)})) + \gamma E(a_{ij}) \quad (1)$$

The first term is a sum of weighted pairwise distances computed over all data pairs from the two empirical distributions. The distance $d$ can take various forms. The second term serves as a regularizer, typically entropic. The hyperparameter $\gamma$ is decreased as the training proceeds. Both Wasserstein and MMD metrics can be written in this form. We provide below nonparametric formulations for two choices of the distance metric $d$.

2.2 The Prokhorov Metric

The Prokhorov metric is the natural measure metrizing weak convergence in the space of probability measures. Minimizing the metric naturally yields a nonparametric formulation. Furthermore, the Prokhorov metric can be defined for distributions on continuous-, discrete- or nominal-valued random variable.

Let $\Omega$ be a complete, separable metric space endowed with a metric $d$. The Prokhorov metric can be defined for distributions on continuous-, discrete- or nominal-valued random variable. The Prokhorov metric is the natural measure metrizing weak convergence in the space of probability measures. We define an $\epsilon$-close neighborhood of a set $S$ as $S' = \{x \in \Omega | \inf_{\xi \in S} d(x, x') \leq \epsilon\}$, where $S$ is an element of Borel $\sigma$-algebra of $\Omega$, $B(\Omega)$.

$$\pi(p, q) := \inf \{\epsilon > 0 | p(S) \leq q(S') + \epsilon \text{ for any } S \in B(\Omega)\}.$$ However, the metric is not invariant under linear transformations. To address this, we define the metric by normalizing by the variance $\sigma^2$ to redefine the $\epsilon$-close neighborhood as $S^\epsilon = \{x \in \Omega | \inf_{\xi \in A} \frac{d(x, x')}{\sigma^2} \leq \delta\}$. This form of Prokhorov is intractable, hence, we utilize the following results to obtain a tractable upper bound instead.

**Theorem:** Let $f$ be a measurable mapping of the space $\Omega$ onto itself. Consider $S' = \{x \mid d(x, f(x)) < \delta\}$. If $p(S') > 1 - \epsilon$, where the metric $d$ is defined over $\Omega$, then $\pi(p, p \circ f) \leq \max(\epsilon, \delta)$. The theorem [Prokhorov, 1956] essentially provides an upper bound between two distributions, if the second distribution can be represented via a mapping, $f$, from the first distribution. The following Lemma utilizes the above to provide an upper bound on the distance measure between real and generated distributions.

**Lemma:** Let $\mathbb{1}$ be the indicator function. Assume real and generated data is $S_r = \{x_r^{(1)}, \ldots, x_r^{(N)}\}$ with $p_r(X_r) = \frac{1}{N} \sum \mathbb{1}_{x_r=x^{(i)}}$ and $S_g = \{x_g^{(1)}, \ldots, x_g^{(N)}\}$ with $P_g(X_g) = \frac{1}{N} \sum \mathbb{1}_{x_g=x^{(i)}}$. Let $f : S_r \cup S_g \to S_g$. Determine a new measure $\mu$ defined on $S_r \cup S_g$ such that, for every $\zeta > 0$ and for all $i \in 1, \ldots, N$, $\text{vol}(S_g) < \zeta$ and $\mu(S_g) < \zeta$, $\mu(x^{(i)}_r) > \frac{1-\zeta}{N}$. Note that $\mu(S_r \cup S_g) = 1$ as desired. Define $f$ as:

$$f(x) = \begin{cases} x_g^{(i)}, & \text{if } x = x_g^{(i)} \\ x_r^{(i)} & \text{if } x \in S^\delta(x_r^{(i)}) \\ x^{(i)} & \text{if } x \in S^\delta(x^{(i)}) \setminus S^\delta(x_r^{(i-1)}, \delta), i \neq 1 \end{cases}$$

for $S^\delta(X^{(i)}_r) := \delta$-close neighborhood of the set $\{x^{(1)}_r, \ldots, x^{(i)}_r\}$. Then if

$$\frac{1}{N} \sum \mathbb{1}_{d(x^{(i)}_r, f(x^{(i)}) < \delta)} > 1 - \epsilon, \pi(X_r, X_g) \leq \max(\delta, \epsilon)$$

It is easy to see that $\delta, \epsilon \in [0, 1]$. Since our target is to minimize $\max\{\delta, \epsilon\}$, we minimize the larger of the two at each step, applied iteratively. Rearranging the constraint from the lemma, we get

$$\epsilon > 1 - \frac{1}{N} \sum \mathbb{1}_{d(x^{(i)}_r, f(x^{(i)}) < \delta)}.$$ This means minimizing RHS minimizes $\epsilon$. We use this term as our loss in $\epsilon$ minimization step. Substituting $f$ into the constraint, the loss function becomes:

$$\min \frac{1}{N} - \frac{1}{N} \sum_{i=1}^N \min\{1, I_{p_r, \delta}(x^{(i)}_r) - \frac{1}{N} \sum_{i=1}^N \min\{1, I_{p_g, \delta}(g_\theta(z^{(i)}))\}\},$$

2
where \( I_{p,j}(.) = 1 \) if \( \delta - d(., y_j) \geq 0 \) and 0 otherwise. This is a special case of the general formulation (1). We minimize the loss in two alternating steps. For \( \delta \) minimization, we use a decaying function. Since decreasing \( \delta \) increases \( \epsilon \), we decrease until this point. Then, \( \epsilon \) is reduced in the next step. We approximate the indicator function using a sigmoid for a differential loss function.

### 2.3 Nonparametric Wasserstein

Using a similar approach, we can create a (novel) nonparametric formulation for the Wasserstein metric and utilize it to directly train the generator without using discriminator DNN, thus avoiding the minimax problem. The nonparametric Wasserstein loss (details omitted), \( \hat{W}(X_r, g_\theta(Z)) \) becomes:

\[
\min_{\alpha_{ij}} \sum_{i,j} \alpha_{ij} \| x_i^{(j)}, g_\theta(z_j^{(j)}) \| + \max_{\lambda_1} \lambda_1 (1 - \sum_{i,j} \alpha_{ij}) + \max_{\lambda_2} \lambda_2 (1 - \sum_{i,j} \alpha_{ij}) + \beta a_{ij} \log(a_{ij}) \quad (2)
\]

Figure 1: (Results on the MINST dataset: (left) vanilla GAN, (middle) nonparametric Prokhorov, (right) nonparametric Wasserstein. (top row) Convergence - nonparametric formulations lead to faster convergence and better stability (variance). Nonparametric Wasserstein performs the best. Note the difference in the y-axis scales. (bottom row). Exemplary samples generated from vanilla-GANs \( (100 \text{ iterations}) \) and the two nonparametric approaches after \( 30 \text{ iterations only} \).

### 3 Experiments

We compared the two proposed nonparametric formulations using the Prokhorov metric and the Wasserstein metric with a (vanilla) GAN formulation. We trained the same 5-layer MLP generator with a leaky-ReLU for all cases. The convergence plots and exemplary samples from the trained networks are shown in figure 1. It is clear that the proposed nonparametric formulations, particularly the nonparametric Wasserstein, achieve faster and more well-behaved convergence, and start producing reasonable quality images with smaller number of iterations (as small as 20-30 iterations). See the caption for more details.

### 4 Conclusion and Future Work

We demonstrate that using nonparametric loss functions is a viable option to the popular GAN framework with adversarial co-training along with a discriminator DNN. The proposed approach avoids optimizing minimax formulations and formally ensures non-zero loss gradients over the feasible space. Preliminary experiments provide encouraging results. Immediate future work will explore these formulations further as well their application to more complex distributions. We also seek to explore more memory efficient approaches (e.g. using kd-Trees or Fast Gauss Transforms).
REFERENCES


