Domain Adaptation

Ronald Cruz, Yves Greatti

1 Introduction

Domain adaptation can be understood as learning a target distribution given an initial distribution and additionally few samples from the target distribution. A model is built from some fixed source domain, but needs to be deployed across one or more different target domains. For example, large-scale speech recognition systems need to work well across arbitrary speech, text processing systems trained on news often need to be applied to blogs or forum posts. Gene finders are trained on a particular organism, but often they need to to identify the genes of another organism or even group of organisms. This report presents two approaches to solve this problem. In doing so, we show the limitations for each strategy, then we evoke a way to combine both approaches in a zero-sum two-player game opening the field for further research.

2 Notation

This section formally defines the learning scenario and the needed terminology. We introduce the definition and notation which coincides with Cortes et al. (2014).

Let $\mathcal{X}$ denote the input space and $\mathcal{Y} \subset \mathbb{R}$ denote the output space. We define a *domain* as a pair formed by a distribution $\mathcal{X}$ and a labeling function mapping $\mathcal{X}$ to $\mathcal{Y}$. In this paper, the *source domain* is defined to be (Q, f_Q) and the *target domain* to be (P, f_P) where Q and P are distributions over $\mathcal{X}$ and f_Q and f_P are labeling functions for the source and target respectively.

The scenario in *domain adaptation* is as follows: The learner receives two samples. The first sample is a labeled sample of m points $\mathcal{S} = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ from the source domain where $x_1, \ldots, x_m$ are drawn i.i.d according to Q and $y_i = f_Q(x_i)$ for $i \in [1, m]$. The second is an unlabeled sample $\mathcal{T} = (x'_1, \ldots, x'_n) \in \mathcal{X}^n$ drawn i.i.d from the target distribution P. We denote by $\widehat{Q}$ and $\widehat{P}$ the empirical distribution corresponding to Q and P respectively. In practice, it is also possible to have a small amount labeled of samples $\mathcal{T}' = ((x''_1, y''_1), \ldots, (x''_s, y''_s)) \in (\mathcal{X} \times \mathcal{Y})^s$ from the target domain.

We consider a loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ jointly convex in its two arguments. For any two target labeling functions $h, h' : \mathcal{X} \to \mathcal{Y}$ and for any distribution D over $\mathcal{X}$, we denote by $\mathcal{L}_D(h, h')$ the expected loss of h and h': $\mathcal{L}_D(h, h') = \mathrm{E}_{x \sim D}[L(h(x), h'(x)]$. We also extend this for functions with finite support $\mathsf{q} : \mathcal{X} \to \mathbb{R}$ to be $\mathcal{L}_{\mathsf{q}}(h, h') = \sum_{x \in \mathcal{X}} \mathsf{q}(x) L(h(x), h'(x))$. The learning problem in domain adaptation is to select a hypothesis h out of a hypothesis set H that minimizes the expected loss $\mathcal{L}_P(h, f_P)$ with respect to the target domain.

3 Generalized Discrepancy Minimization

3.1 Preliminaries

In order to qualitatively determine when successful adaptation occurs, a *divergence*, or a measure of how different two distributions are is needed. In statistics, there are a large amount of divergences

such as the Kullback-Leibler Divergence, Hellinger distance, total variation, and many more. However, many of these notions of divergences only take into account the probability distribution. They do not take into consideration the structure of a learning problem.

The issue with the previously mentioned divergences is that they do not take into consideration the class of hypothesis set or the loss function. In the context of learning, these two concepts are crucial. Without them, learning becomes infeasible. As a remedy to this problem, Mansour et al. (2009) and Cortes and Mohri (2011) show that a key measure of the difference between two distributions is that of *discrepancy*.

Definition 1. *Given a hypothesis H, the discrepancy,* disc*, of two distributions P and Q over an input space $\mathcal{X}$ is defined as:*

$$\text{disc}(P, Q) = \max_{h, h' \in H} |\mathcal{L}_P(h', h) - \mathcal{L}_Q(h', h)| \tag{1}$$

This new notion of divergence measures the difference of losses of the hypothesis set over both the source and target distribution. It can be shown that the discrepancy provides a finer measure of divergence over other notions such as the total variation Medina (2015). Furthermore, the notion of discrepancy also benefits from a variety of learning guarantees and generalization bounds for kernel-based regularization algorithms given by Mansour et al. (2009), Cortes and Mohri (2011), Ben-David and Urner (2012). In particular, these guarantees show that a smaller empirical discrepancy, $\text{disc}(\hat{P}, \hat{Q})$ between two distributions P and Q guarantees small pointwise losses.

These bounds motivate *discrepancy minimization* algorithm for a general class of kernel-based algorithms. Given a positive semi-definite kernel K, the hypothesis returned by this algorithm is a solution to the following optimization problem

$$\min_{h \in \mathbb{H}} \lambda \|h\|_K^2 + \mathcal{L}_{\mathsf{q}_{\min}}(h, f_Q) \tag{2}$$

where $\| \cdot \|_K$ denotes the norm of the Hilbert space $\mathbb{H}$ induced by the kernel K and $\mathsf{q}_{\min} = \text{argmin}_{\mathsf{q} \in \mathcal{Q}} \text{disc}(\mathsf{q}, \widehat{P})$, where $\mathcal{Q}$ is the set of all probability distributions defined over the support of $\widehat{Q}$. The distribution $\mathsf{q}_{\min}$ can be interpreted as a constant reweighing of the losses from a sample $\mathcal{S}_\mathcal{X} = \{x_1, \ldots x_m\}$ to the set $[0, 1]$.

3.2 Main Idea

The main idea of the new algorithm presented in Cortes et al. (2014) is to first consider learning a probability distribution in the ideal scenario of supervised learning, where the target labels are known. Exploiting this information, an algorithm such as regularized empirical risk minimization over a Hilbert space $\mathbb{H}$ induced by a PSD kernel K returns a hypotheses h^* which is a solution of $\min_{h \in \mathbb{H}} F(h)$ where

$$F(h) = \lambda \|h\|_K^2 + \mathcal{L}_{\widehat{P}}(h, f_P) \tag{3}$$

where $\lambda \geq 0$ is a regularization parameter. In this scenario, h^* is the ideal solution.

In practice, it is often the case that the target labels f_P are not known and the learner only has access to the training sample $\mathcal{S}$. This brings us to the scenario of domain adaptation. Due to this lack of information, we relax the problem. Instead of having $\mathcal{L}_{\widehat{P}}(h, f_P)$ in our objective function, we instead try to learn a weighting function that is uniformly close to $\mathcal{L}_{\widehat{P}}(h, f_P)$ over $\mathcal{S}$. The idea is to learn, for any $h \in \mathbb{H}$, a reweighing function $\mathsf{Q}_h : \mathcal{S}_\mathcal{X} = \{x_1, \ldots, x_m\} \to \mathbb{R}$ such that $|\mathcal{L}_{\mathsf{Q}_h}(h, f_Q) - \mathcal{L}_{\widehat{P}}(h, f_P)|$ is minimized.

In finding such a reweighing function, we run into the same issue which is that we are attempting to minimize over an unknown f_P. Thus, the authors propose to relax the problem further and consider a nonempty convex surrogate hypothesis set $H'' \subseteq H$ that could contain the target labeling function f_P. Using this subset instead, we can define Q_h to be

$$\mathsf{Q}_h = \operatorname*{argmin}_{\mathsf{q}\in\mathcal{F}(\mathcal{S}_\mathcal{X},\mathbb{R})} \max_{h''\in H''} |\mathcal{L}_\mathsf{q}(h, f_Q) - \mathcal{L}_{\widehat{P}}(h, h'')| \tag{4}$$

where $\mathcal{F}(\mathcal{S}_\mathcal{X}, \mathbb{R})$ is the set of real valued functions defined over $\mathcal{S}_\mathcal{X}$. Therefore, under the two relaxations, we can reformulate the problem of finding $\min_{h\in\mathbb{H}} F(h)$ to $\min_{h\in\mathbb{H}} G(h)$ where

$$G(h) = \lambda\|h\|_K^2 + \mathcal{L}_{\mathsf{Q}_h}(h, f_Q) \tag{5}$$

Proposition 2. *let Q_h be defined by (5) for any $h \in \mathbb{H}$. The following identity holds for any $h \in \mathbb{H}$:*

$$\mathcal{L}_{\mathsf{Q}_h}(h, f_Q) = \frac{1}{2}\left(\max_{h''\in H''} \mathcal{L}_{\widehat{P}}(h, h'') + \min_{h''\in H''} \mathcal{L}_{\widehat{P}}(h, h'')\right)$$

Proof. For any $h \in \mathbb{H}$, the equation $\mathcal{L}_\mathsf{q}(h, f_Q) = l$ with $l \in \mathbb{R}$ admits a solution $\mathsf{q} \in \mathcal{F}(\mathcal{S}_\mathcal{X}, \mathbb{R})$. Thus, $\{\mathcal{L}_\mathsf{q}(h, f_Q) : \mathsf{q} \in \mathcal{F}(\mathcal{S}_\mathcal{X}, \mathbb{R})\} = \mathbb{R}$ and for any $h \in \mathbb{H}$, we can write:

$$\begin{aligned}
\mathcal{L}_{\mathsf{Q}_h}(h, f_Q) &= \operatorname*{argmin}_{l\in\{\mathcal{L}_\mathsf{q}(h,f_Q):\, \mathsf{q}\in\mathcal{F}(\mathcal{S}_\mathcal{X},\mathbb{R})\}} \max_{h''\in H''} |l - \mathcal{L}_{\widehat{P}}(h, h'') \\
&= \operatorname*{argmin}_{l\in\mathbb{R}} \max_{h''\in H''} |l - \mathcal{L}_{\widehat{P}}(h, h'')| \\
&= \operatorname*{argmin}_{l\in\mathbb{R}} \max\left\{\max_{h''\in H''} \mathcal{L}_{\widehat{P}}(h, h'') - l, l - \min_{h''\in H''} \mathcal{L}_{\widehat{P}}(h, h'')\right\} \\
&= \frac{1}{2}\left(\max_{h''\in H''} \mathcal{L}_{\widehat{P}}(h, h'') - l, l - \min_{h''\in H''} \mathcal{L}_{\widehat{P}}(h, h'')\right)
\end{aligned}$$

since the minimizing l is obtained by setting $\max_{h''\in H''} \mathcal{L}_{\widehat{P}}(h, h'') - l = l - \min_{h''\in H''} \mathcal{L}_{\widehat{P}}(h, h'')$. □

Using the above proposition, we the objective function (5) can be written as for all $h \in \mathbb{H}$:

$$G(h) = \lambda\|h\|_K^2 + \frac{1}{2}\left(\max_{h''\in H''} \mathcal{L}_{\widehat{P}}(h, h'') + \min_{h''\in H''} \mathcal{L}_{\widehat{P}}(h, h'')\right) \tag{6}$$

Since $\mathcal{L}_{\widehat{P}}$ is jointly convex by assumption, $h \to \max_{h''\in H''} \mathcal{L}_{\widehat{P}}(h, h'')$ is also convex because it is the pointwise supremum of convex functions. Similarly, since partial minimization preserves convexity $h \to \min_{h''\in H''} \mathcal{L}_{\widehat{P}}(h, h'')$ is also convex. Therefore, G is convex as it is a sum of convex functions, resulting in a convex optimization problem.

This algorithm differs from prior work in discrepancy minimization (2) in several aspects. Firstly, Q_h is not constrained to a probability distribution. Instead, it is allowed to cover a much richer space, $\mathbb{R}$, and does not have to sum to 1. Furthermore, a key difference is that this reweighing scheme is *hypothesis dependent*, rather than being constant for all hypothesis as it was in discrepancy minimization.

In particular, this means that for any hypothesis h, we can select Q_h such that $\mathcal{L}_{\mathsf{Q}_h}(h, f_Q) = \mathcal{L}_{\widehat{P}}(h, f_P)$ as the expected loss $\mathcal{L}_\mathsf{q}(h, f_Q)$ is linear in q. Thus, we can recover the ideal solution h^* by solving a linear program. This was not the case for discrepancy minimization, as in general $\mathcal{L}_{\mathsf{q}_{\min}}(h, f_Q) = \mathcal{L}_{\widehat{P}}(h, f_P)$ would not hold true for all h. However Q_h requires knowing the convex surrogate hypothesis set H'', the choice of which is key in practice. In the following section, we present the results on the existence of such a set of L_p losses and the learning guarantees.

4 Learning Guarantees

In this section, we assume that the loss function L is μ-admissible, meaning that there exists $\mu > 0$ such that

$$|L(h(x), y) - L(h'(x), y)| \leq \mu |h(x) - h'(x)|$$

holds for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $h, h' \in H$.

To have a point of comparison for this new algorithm, we present the existing pointwise guarantee for the discrepancy minimization algorithm given by (2). The guarantee is given in terms of the discrepancy and a term $\eta_H(f_P, f_Q)$ defined by

$$\eta_H(f_P, f_Q) = \min_{h_0 \in H} \left(\max_{x \in \text{supp}(\widehat{P})} |f_P(x) - h_0(x)| + \max_{x \in \text{supp}(\widehat{Q})} |f_Q(x) - h_0(x)| \right)$$

which measures the difference between the source and target labeling functions.

Theorem 3. *(Cortes and Mohri, 2011) Let* q *be an arbitrary distribution over* $\mathcal{S}_\mathcal{X}$ *and let* h^* *and* h_q *be the hypothesis minimizing* $\lambda \|h\|_K^2 + \mathcal{L}_{\widehat{P}}(h, f_P)$ *and* $\lambda \|h\|_K^2 + \mathcal{L}_\mathsf{q}(h, f_Q)$ *respectively. Then the following inequality holds:*

$$\lambda \|h^* - h_\mathsf{q}\|_K^2 \leq \mu \eta_H(f_P, f_Q) + \text{disc}(\widehat{P}, \mathsf{q})$$

The discrepancy minimization algorithm defined in (2) selects the distribution q which minimizes the term $\text{disc}(\widehat{P}, \mathsf{q})$. For the new algorithm, Cortes et al. presents similar pointwise guarantees in terms of the new reweighing function Q. In doing so, they present a new notion of generalized discrepancy, which we now define.

Let $\mathcal{A}(H)$ by the set of all functions $\mathsf{U} : h \to \mathsf{U}_h$ mapping H to $\mathcal{F}(\mathcal{S}_\mathcal{X}, \mathbb{R})$ such that for all $h \in H, h \to \mathcal{L}_{\mathsf{U}_h}(h, f_Q)$ is convex. Observe that $\mathcal{A}(H)$ contains all constant functions such that $\mathsf{U}_h = \mathsf{q}$ where q is a distribution over $\mathcal{S}_\mathcal{X}$ as well as the Q_h used in the algorithm.

Definition 4. For any $\mathsf{U} \in \mathcal{A}(H)$, the notion of generalized discrepancy between two distributions $\widehat{P}$ and U is denoted as $\text{DISC}(\widehat{P}, \mathsf{U})$ defined by

$$\text{DISC}(\widehat{P}, \mathsf{U}) = \max_{h \in H, h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{U}_h}(h, f_Q)|$$

The authors also present a measure of distance over a distribution between a target labeling function f_P and convex subset H'' given by

$$d_\infty^{\widehat{P}}(f_P, H'') = \min_{h_0 \in H''} \max_{x \in \text{supp}\, \widehat{P}} |h_0(x) - f_P(x)|$$

The following is a pointwise learning guarantee of the algorithm based on $d_\infty^{\widehat{P}}(f_P, H'')$ and generalized discrepancy.

Theorem 5. *Let* U *be an arbitrary element of* $\mathcal{A}(H)$ *and let* h^* *and* h_U *be the hypothesis minimizing* $\lambda \|h\|_K^2 + \mathcal{L}_{\widehat{P}}(h, f_P)$ *and* $\lambda \|_h\|_K^2 + \mathcal{L}_{\mathsf{U}_h}(h, f_Q)$ *respectively. Then the following inequality holds for any convex subset* $H'' \subseteq H$*:*

$$\lambda \|h^* - h_\mathsf{U}\|_K^2 \leq \mu d_\infty^{\widehat{P}}(f_P, H'') + \textit{DISC}(\widehat{P}, \mathsf{U})$$

Proof. Fix $\mathsf{U} \in \mathcal{A}(H)$ and let $G_{\widehat{P}}$ denote $h \to \mathcal{L}_{\widehat{P}}(h, f_P)$ and G_U denote $h \to \mathcal{L}_{\mathsf{U}_h}(h, f_Q)$. Since $h \to \lambda \|h\|_K^2 + G_{\widehat{P}}(h)$ is convex and differentiable and since h^* is a minimizer, the gradient is zero at h^*, that is $2\lambda h^* = -\nabla G_{\widehat{P}}(h^*)$. Similarly, since $h \to \lambda \|h\|_K^2 + G_\mathsf{U}(h)$ is convex, it admits a sub-differential at any $h \in \mathbb{H}$. Since h_U is a minimizer, its sub-differential at h_U must be 0. Then, there exists a sub-gradient $g_0 \in \partial G_\mathsf{U}(h_\mathsf{U})$ such that $2\lambda h_\mathsf{U} = -g_0$, where $\partial G_\mathsf{U}(h_\mathsf{U})$ denotes the sub-differential of G_U at h_U. Using these two inequalities, we can write

$$
\begin{aligned}
2\lambda\|h^* - h_{\mathsf{U}}\|_K^2 &= \langle h^* - h_{\mathsf{U}}, g_0 - \nabla G_{\widehat{P}}(h^*)\rangle \\
&= \langle g_0, h^* - h_{\mathsf{U}}\rangle - \langle \nabla G_{\widehat{P}}(h^*), h^* - h_{\mathsf{U}}\rangle \\
&\leq G_{\mathsf{U}}(h^*) - G_{\mathsf{U}}(h_{\mathsf{U}}) + G_{\widehat{P}}(h_{\mathsf{U}}) - G_{\widehat{P}}(h^*) \\
&= \mathcal{L}_{\widehat{P}}(h_{\mathsf{U}}, f_P) - \mathcal{L}_{\mathsf{U}_h}(h_{\mathsf{U}}, f_Q) + \mathcal{L}_{\mathsf{U}_h}(h^*, f_Q) - \mathcal{L}_{\mathsf{U}_h}(h^*, f_P) \\
&\leq 2 \max_{h \in H} |\mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\mathsf{U}_h}(h, f_Q)|
\end{aligned}
$$

where convexity of $G_{\widehat{P}}$ and the sub-gradient property of $g_0 \in \partial G_{\mathsf{U}}(h_{\mathsf{U}})$ was used. For any $h \in H$, using the μ-admissibility of the loss, we can upper bound the operand of the $\max$ operator as follows:

$$
\begin{aligned}
|\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{U}_h}(h, f_Q)| &\leq |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\widehat{P}}(h, h_0)| + |\mathcal{L}_{\widehat{P}}(h, h_0) - \mathcal{L}_{\mathsf{U}_h}(h, f_Q)| \\
&\leq \mu \operatorname*{E}_{x \sim \widehat{P}}[f_P(x) - h_0(x)] + \max_{h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{U}_h}(h, f_Q)| \\
&\leq \mu \max_{x \in \operatorname{supp}(\widehat{P})} |f_P(x) - h_0(x)| + \max_{h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{U}_h}(h, f_Q)|
\end{aligned}
$$

where h_0 is an arbitrary element of H. Since this bound holds for all $h_0 \in H''$, it follows immediately that

$$
\lambda\|h^* - h_{\mathsf{U}}\|_K^2 \leq \mu \min_{h_0 \in H''} \max_{x \in \operatorname{supp}(\widehat{P})} |f_P(x) - h_0(x)| + \max_{h \in H} \max_{h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{U}_h}(h, f_Q)|
$$

which concludes the proof. □

The algorithm presented by Cortes et al. focuses on finding a reweighing function $U \in \mathcal{A}(H)$ such that $\mathrm{DISC}(\widehat{P}, \mathsf{U})$ is minimized, similar to that of discrepancy minimization. However, this in itself does not necessarily show that the bound provided is better than theorem 3. In minimizing the generalized discrepancy, a natural question arises of whether there exists a convex set H'' such that the given bound is uniformly better than that of discrepancy minimization. The following theorem shows the existence of such a set for L_p losses. This result depends on the idea of local discrepancy which can be defined as

$$
\operatorname{disc}_{H''}(\widehat{P}, \mathsf{q}) = \max_{h \in H, h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{q}}(h, h'')|
$$

Local discrepancy is a finer measure than discrepancy which is defined as the pairwise maximum over a pair of hypothesis *both* in $H \supseteq H''$.

Theorem 6. *Let L be the L_p loss for some $p \geq 1$ and h_0^* be the minimizer in the definition of $\eta_H(f_p, f_Q) = \operatorname{argmin}_{h_0 \in H} \left(\max_{x \in \operatorname{supp} \widehat{P}} |f_Q(x) - h_0^*(x)| + \max_{x \in \operatorname{supp} \widehat{Q}} |f_Q(x) - h_0^*(x)| \right)$. Define $r \geq 0$ by $r = \max_{x \in \operatorname{supp} \widehat{Q}} |f_Q(x) - h_0^*(x)|$. Let q be a distribution over $\mathcal{S}_{\mathcal{X}}$ and let H'' be defined by $H'' = \{h'' \in H \mid \mathcal{L}_{\mathsf{q}}(h'', f_Q) \leq r^p\}$. Then $h_0^* \in H''$ and the following inequality holds:*

$$
\mu d_{\infty}^{\hat{P}}(f_P, H'') + \max_{h \in H, h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{q}}(h, f_Q)| \leq \mu \eta_H(f_P, f_Q) + \operatorname{disc}_{H''}(\widehat{P}, \mathsf{q}) \tag{7}
$$

Proof. The fact that $h_0 \in H''$ follows from

$$
\mathcal{L}_{\mathsf{q}}(h_0^*, f_Q) = \operatorname*{E}_{x \sim \mathsf{q}}[|h_0^*(x) - f_Q(x)|^p] \leq \max_{x \in \operatorname{supp} \widehat{Q}} |h_0^*(x) - f_Q(x)|^p \leq r^p
$$

Using the μ-admissibility of the loss function, for any $h, h'' \in H$, we have that $|\mathcal{L}_{\mathsf{q}}(h, h'') - \mathcal{L}_{\mathsf{q}}(h, f_Q)| \leq \mu |\mathcal{L}_{\mathsf{q}}(h'', f_Q)|^{\frac{1}{p}}$. Utilizing this, we have

$$
\begin{aligned}
&\max_{h\in H,h''\in H''}|\mathcal{L}_{\widehat{P}}(h,h'')-\mathcal{L}_{\mathsf{q}}(h,f_Q)| \\
&\qquad\leq \max_{h\in H,h''\in H'}|\mathcal{L}_{\widehat{P}}(h,h'')-\mathcal{L}_{\mathsf{q}}(h,h'')+\max_{h\in H,h''\in H''}|\mathcal{L}_{\mathsf{q}}(h,h'')-\mathcal{L}_{\mathsf{q}}(h,f_Q)| \\
&\qquad\leq \mathrm{disc}_{H''}(\widehat{P},\mathsf{q})+\max_{h''\in H''}|\mathcal{L}_{\mathsf{q}}(h,h'')-\mathcal{L}_{\mathsf{q}}(h,f_Q)| \\
&\qquad\leq \mathrm{disc}_{H''}(\widehat{P},\mathsf{q})+\mu r \\
&\qquad= \mathrm{disc}_{H''}(\widehat{P},\mathsf{q})+\mu\max_{x\in\operatorname{supp}\widehat{Q}}|f_Q(x)-h_0^*(x)|
\end{aligned}
$$

Using the inequality above, we can now write

$$
\begin{aligned}
&\mu d_\infty^{\hat{P}}(f_P,H'')+\max_{h\in H,h''\in H''}|\mathcal{L}_{\widehat{P}}(h,h'')-\mathcal{L}_{\mathsf{q}}(h,f_Q)| \\
&\quad\leq \mu\min_{h_0\in H''}\max_{x\in\operatorname{supp}\widehat{P}}|f_P(x)-h_0(x)|+\mathrm{disc}_{H''}(\widehat{P},\mathsf{q})+\mu\max_{x\in\operatorname{supp}\widehat{Q}}|f_Q(x)-h_0^*(x)| \\
&\quad\leq \mu\left(\max_{x\in\operatorname{supp}\widehat{P}}|f_P(x)-h_0^*(x)|+\mu\max_{x\in\operatorname{supp}\widehat{Q}}|f_Q(x)-h_0^*(x)|\right)+\mathrm{disc}_{H''}(\widehat{P},\mathsf{q}) \\
&\quad= \mu\min_{h_0\in H}\left(\max_{x\in\operatorname{supp}\widehat{P}}|f_P(x)-h_0(x)|+\max_{x\in\operatorname{supp}\widehat{Q}}|f_Q(x)-h_0(x)|\right)+\mathrm{disc}_{H''}(\widehat{P},\mathsf{q}) \\
&\quad= \mu\eta_H(f_P,f_Q)+\mathrm{disc}_{H''}(\widehat{P},\mathsf{q})
\end{aligned}
$$

concluding the proof. □

The theorem above shows that for a specific choice of H'' and L_p losses for $p\geq 1$, the algorithm by Corinna et al. has a more favorable guarantee than that of discrepancy minimization. The following theorem gives a pointwise guarantee for the solution returned by the algorithm.

Theorem 7. *Let h^* be a minimizer of $\lambda\|h\|_K^2+\mathcal{L}_{\widehat{P}}(h,f_P)$ and h_{Q} a minimizer of $\lambda\|h\|_K^2+\mathcal{L}_{\mathsf{Q}_h}(h,f_Q)$. Then the following holds for any convex set $H''\subseteq H$:*

$$
|L(h_Q(x),y)-L(h^*(x),y)|\leq \mu R\sqrt{\frac{\mu d_\infty^{\hat{P}}(f_P,H'')+DISC(\widehat{P},\mathsf{Q})}{\lambda}},\forall x\in\mathcal{X},y\in\mathcal{Y} \tag{8}
$$

where $R^2=\sup_{x\in\mathcal{X}}K(x,x)$.

Proof. By the μ-admissibility of the loss, reproducing property of $\mathbb{H}$, and the Cauchy-Schwarz inequality, the following holds for all $x\in\mathcal{X}$ and $y\in\mathcal{Y}$:

$$
\begin{aligned}
|L(h_{\mathsf{Q}}(x),y)-L(h^*(x),y)|\leq\mu|h'(x)-h^*(x)| &= |\langle h'-h^*,K(x,\cdot)\rangle_K| \\
&\leq \|h'-h^*\|_K\sqrt{K(x,x)} \\
&\leq R\|h'-h^*\|_K
\end{aligned}
$$

□

Rewriting the bound just presented gives us an upper bound in terms of the loss of our solution with the ideal solution for domain adaptation

$$
\mathcal{L}_P(h_{\mathsf{Q}},f_P)\leq\mathcal{L}_P(h^*,f_P)+\mu R\sqrt{\frac{\mu d_\infty^{\hat{P}}(f_P,H'')+\mathrm{DISC}(\widehat{P},\mathsf{Q})}{\lambda}}
$$

In practice, it is sometimes the case that there is access to a small amount of labeled target data $\mathcal{T}'$. The authors recommend using this data as a validation set to determine the parameter r of the convex subset H''. They also propose using the sample to enhance the learning algorithm by taking a weighted combination of the empirical training sample $\hat{Q}$ and the small target labeled sample $\mathcal{T}'$.

As shown above, for L_p losses and a specific choice of H'', this new algorithm benefits from stronger guarantees than that of discrepancy minimization. However, even though it was formulated as a convex optimization problem, the solution to this optimization problem is not trivial. The authors formulate for L_2 losses, solutions for specific convex sets H'' via an SDP, and an approximation algorithm given by a QP and a sampling technique. The formulation of both of these optimization problems can be solved using standard solvers for SDPs and QPs.

5 Generative Adversarial Models

5.1 Background

Analysis introduced in Goodfellow et al. (2014) shows that with a "large enough" sample size and deep enough nets, GANS (Generative Adversarial Nets) succeed in learning the target distribution. Further theoretical research by Arora et al. (2017) shows that even when capacity of the GANS and training set size are bounded, an approximate pure equilibrium exists. We start with an introduction to GAN and describe their properties. We then present two papers from Daskalakis et al. (2017) and Gulrajani et al. (2017). We also show the limitation of the Gradient Penalty used in the first paper and recent research made to address it. We also propose our own improvement of the Gradient Penalty algorithm and a more general approach related to the curvature of the loss function.

5.2 Introduction

In a generative model approach, the focus is to estimate the density of a data distribution $\mathbb{P}_\mathbf{r}$ which may or may not exist, with $\mathbb{P}_\theta$ a distribution of the parametrized density $(P_\theta)_{\theta\in\mathbb{R}^d}$. The classical approach of learning a probability density is to maximize the likehood on the data $\{x_i\}_{i=1}^m$:

$$\max_{\theta\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^{m} \log P_\theta\{x_i\}$$

This amounts in minimizing the KL divergence $KL(\mathbb{P}_\mathbf{r}||\mathbb{P}_\theta)$. However $\mathbb{P}_\theta$ has low dimensional support and it is unlikely that all $\mathbb{P}_\mathbf{r}$ lies within $\mathbb{P}_\theta$ support and in this scenario, the KL divergence is not defined. To address this issue, noise is added to the model distribution. The GAN training strategy consists then, in a zero-sum game between a generator deep neural network $G_\theta(.)$ and a discriminator deep neural network $D_w(.)$. Given a distribution of data points $Q \in \Delta(X)$, the discriminator $D_w(.)$ takes a sample either from Q or $G_\theta(.)$ and classified it as real or fake. The generator takes as input random noise $z \sim F$ and outputs a sample $G_\theta(z) \in X$. The goal of the generator is to output a sample close to the target distribution $\mathbb{P}_\mathbf{r}$. The discriminator deep net (or *critique*) is trained using samples from $\mathbb{P}_\mathbf{r}$ and from the generator trying to classify them as real or fake. As long the discriminator is successful at this task, it generates a feedback to the generator thus improving its distribution $\mathbb{P}_\theta$. Training is continued until the generator wins. Success for the generator is measured in how $\mathbb{P}_\theta$ is close to $\mathbb{P}_\mathbf{r}$ in some distance metric.

5.3 Distances and Generalization bounds

The GAN training consists of training the parameters θ and w so as to optimize the objective function:

$$L(\theta, w) = \min_\theta \max_w \mathop{\mathrm{E}}_{x\sim Q}[\phi(D_w(x))] - \mathop{\mathrm{E}}_{z\sim F}[\phi(D_w(G_\theta(z)))] \tag{9}$$

where ϕ is a measuring function which must be concave, the standard GAN training use $\phi(x) = \log$. However since the log function can cause problems, in practice the training often uses $\phi(x) =$

$\log(\delta + (1-\delta)x)$. Then the problem of the generator amounts in minimizing the Jensen-Shannon divergence between the real distribution and the generator distribution. Other measuring ϕ functions and other choice of discriminator class leads to different distance functions. Gulrajani et al. (2017) show that when $\phi(x) = x$, and the discriminator is chosen among all 1-Lipschitz functions, maxing out the discriminator consists for the generator to minimize the Wasserstein distance or earth-mover (EM) distance.

Definition 8. Earth-mover distance

$$W_1(\mu_s, \mu_t) = \left(\inf_{\gamma \in \Pi_{\mu_s, \mu_t}} \int_{\mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\| \, d\pi(\mathbf{x}, \mathbf{y}) \right)$$

$\gamma(x, y)$ represents the cost of transport from x to y. The Kantovorich-Rubinstein formulation of the optimal transport reads:

$$\inf_{\theta} \sup_{w} \mathop{\mathrm{E}}_{x \sim Q}[D_w(x)] - \mathop{\mathrm{E}}_{z \sim F}[D_w(G_\theta(z))]$$

where the discriminator covers all 1-Lipschitz, or K-Lipschitz with scaling factor K, functions of x. The Wasserstein distance compared to KL or JS distances gives better guarantee of continuity and differentiability and also is weaker in convergence (Arjovsky et al., 2017). However Wasserstein distance and likewise JS distance can lead to overfitting.

Lemma 9. *Let μ be the Gaussian distribution $\mathcal{N}(0, \frac{1}{d})$ and $\widehat{\mu}$ be the empirical distribution of μ with m samples. Then we have*

$$d_{JS}(\mu, \widehat{\mu}) = \log 2$$
$$d_W(\mu, \widehat{\mu}) \geq 1.1$$

To address this shortcoming Arora et al. (2017) propose a generalization of the previous distances called neural net distance $d_{\mathcal{F},\phi}$.

Definition 10. $\mathcal{F}$-distance Let $\mathcal{F}$ the class of functions from $\mathbb{R}^d$ to [0,1] such that if $f \in \mathcal{F}$, $1 - f \in \mathcal{F}$, ϕ a concave measuring function. Then the $\mathcal{F}$-divergence w.r.t. ϕ between two distributions μ and ν is defined as:

$$d_{\mathcal{F},\phi}(\mu, \nu) = \sup_{D \in \mathcal{F}} \mathop{\mathrm{E}}_{x \sim \mu}[\phi(D(x))] + \mathop{\mathrm{E}}_{x \sim \nu}[\phi(1 - D(x))] - 2\phi(1/2)$$

When $\phi(t) = t$ then the objection functions used in WGAN is equivalent to $\min_G d_{\mathcal{F}}(\mathbb{P}_{\mathbf{r}}, \mathbb{P}_G)$. In addition Arora et al. (2017) reveal that, if the discriminator has size p, then the training could be ϵ close to $\epsilon-$equilibrium when the target distribution has support only $\tilde{O}(p \log(p)/\epsilon^2$.

Theorem 11. *Let ϕ a measuring function in $[-\Delta, \Delta]$ and L_ϕ-Lipschitz , p the number of discriminators, and $\{G^{(1)}, \ldots, G^{(K)}\}$ K generators in the K iterations of the training, and assume $\log(K) \leq p$. Then there exist constant C_1,C_2 such that when $m \geq \frac{C_1 \log(C_2 p)}{\epsilon^2}$, with probability at least $1 - e^{-p}$, for all $t \in [K]$:*

$$|d_{\mathcal{F},\phi}(\mathcal{D}_{real}, \mathcal{D}_{G^{(t)}}) - d_{\mathcal{F},\phi}(\hat{\mathcal{D}}_{real}, \hat{\mathcal{D}}_{G^{(t)}})| \leq \epsilon$$

where $\hat{\mathcal{D}}_{real}$ is the empirical distribution with distribution $\mathcal{D}_{real}$ and $\hat{\mathcal{D}}_{G^{(t)}}$ is the empirical version of the generated distribution $\mathcal{D}_{G^{(t)}}$.

The same authors then propose a mixture of generators and discriminators in a traditional expert settings theory (exponential weighted algorithm) to improve the asymptotic convergence to equilibrium. They also raise the possibility that the neural net distance $d_{\mathcal{F},\phi}(\mu, \nu)$ can be small and μ, ν not very close.

5.4 Optimistic Gradient Descent

The classic WGAN approach to solve equation (9) is to run gradient descent (GD) for each player. And standard result shows that in the zero-sum game with $L(\theta, w)$ convex in θ, concave in w, on average ϵ-equilibrium is reached.

Proof. Using for example RWM algorithm

$$\frac{1}{T}(\sum_t p_t^T U q - \min_p \sum_t p_t^T U q) \leq \epsilon$$

$$\begin{aligned}\min_p \max_q p^T U q &\leq \frac{1}{T}\sum_t p_t^T U q \\ &= \frac{1}{T}\min_p \sum_t p_t^T U q + \epsilon \\ &\leq \max_q \min_p p^T U q + \epsilon\end{aligned}$$

□

Rakhlin and Sridharan (2012) proposed an alternative algorithm for solving zero-sum games, namely Optimistic Mirror Descent (OMD), that achieves faster convergence rates to equilibrium ϵ=$O(\frac{\log T}{T})$ instead of ϵ=$O(\frac{1}{T})$ for the average of parameters $\bar{\theta}$ and $\bar{w}$. The intuition of OMD is that first FTL algorithm is online GD and as a consequence of FTL if the learner knew in advance the next iteration then the GD step would lead to a constant regret coming from the regularization term.

Proof. Let $f_t(w) = \langle w, z_t \rangle$ for some vector z_t, and
$w_t = \text{argmin} \sum_{s=1}^{t-1} f_s(w) + \frac{1}{2\eta}\|w\|^2$. Let $G(x) = \langle x, z_t \rangle + \frac{1}{2\eta}\|x\|^2$. Then

$$\begin{aligned}\nabla G(w_t) = 0 \Rightarrow \sum_{s=1}^{t-1} z_s + \frac{1}{\eta} w_t &= 0 \\ \Rightarrow w_t &= -\eta \sum_{s=1}^{t-1} z_s \\ \Rightarrow w_t &= w_{t-1} - \eta z_{t-1} \\ \Rightarrow w_t &= w_{t-1} - \eta \nabla f_{t-1}(w_{t-1})\end{aligned}$$

□

OMD adds a predictor M_{t+1} on the next iteration of the gradient:

$$w_{t+1} = \underset{w}{\text{argmax}}\, \eta.(\sum_{s=1}^{t} \langle w, \nabla_{w,s} \rangle + \langle w, \mathbf{M}_{w,t+1} \rangle) - \|w\|_2^2$$
$$\theta_{t+1} = \underset{\theta}{\text{argmin}}\, \eta.(\sum_{s=1}^{t} \langle \theta, \nabla_{\theta,s} \rangle + \langle \theta, \mathbf{M}_{\theta,t+1} \rangle) + \|\theta\|_2^2$$

And the update rules become:

$$w_{t+1} = w_t + \eta.(\nabla_{w,t} + M_{w,t+1} - M_{w,t})$$
$$\theta_{t+1} = \theta_t - \eta.(\nabla_{\theta,t} + M_{\theta,t+1} - M_{\theta,t})$$

The predictor of the next iteration's gradient can be simply last iteration's gradient, or an average of a window of last gradients, or a discounted average of past gradients. In the case where the predictor is the last iteration gradient, the update rules for OMD are in the following simple form:

$$w_{t+1} = w_t + 2.\eta.\nabla_{w,t} - \eta \nabla_{w,t-1} \tag{10}$$
$$\theta_{t+1} = \theta_t - 2.\eta.\nabla_{\theta,t} + \eta \nabla_{\theta,t-1} \tag{11}$$

5.5 Stochastic Optimistic Gradient Descent

Daskalakis et al. (2017) propose in practice unbiased estimators of the true distributions Q and F by small batch of B samples. The OMD iteration has the variant:

$$\widehat{\nabla}_{w,t} = \frac{1}{|B|}\sum_{i\in B}(\nabla_w D_{w_t}(x_i) - \nabla_w D_{w_t}(G_{\theta_t}(z_i)))$$
$$\widehat{\nabla}_{\theta,t} = -\frac{1}{|B|}\sum_{i\in B}\nabla_\theta(D_{w_t}(G_{\theta_t}(z_i)))$$

5.6 Last Iterate convergence of Optimistic Gradient Descent

OMD converges to equilibrium in last-iterate convergence rather on average-iterate. Daskalakis et al. (2017) show different experimental results in which this property lead to more robust and stable solutions compared to GD or variants of GD which exhibit limit cycles. In equations (10) , OMD is similar to the $\min_p \max_q p^T Aq$ problem for some matrix A.

Theorem 12 (Last iterate convergence)**.** *In equations* (10) *, OMD converges in last iterate. Let* $x_0 \in \mathcal{R}(A)$*,* $y_0 \in \mathcal{R}(A^T)$*,* $\frac{1}{2}x_{-1} = x_0$*,* $\frac{1}{2}y_{-1} = y_0$*,* $\gamma = \max(\|(AA^T)^\dagger\|, \|(A^TA)^\dagger\|)$*. Suppose also* $\|A\| \leq 1$ *and a small enough constant* $\eta < \frac{1}{3\gamma^2}$*, with* $\Delta_t = \|A^Tx_t\|_2^2 + \|Ay_t\|_2^2$*, the OMD algorithm satisfies:*

$$\Delta_1 = \Delta_0 \geq \frac{1}{1+\eta^2}\Delta_2$$
$$\forall t \geq 3 : \Delta_t \leq (1 - (\frac{\eta}{\gamma})^2)\Delta_{t-1} + 16\eta^3\Delta_0$$

Last iterate convergence: In particular, as $t \to \infty$*, the last iterate of OMD is within* $\mathcal{O}(\gamma\sqrt{\eta.\Delta_0^0})$ *distance from the space of equilibrium points, where* $\sqrt{\Delta_0^0}$ *is the distance of* (x_0, y_0) *to the equilibrium space and where the distance is taken w.r.t* $\sqrt{x^TAA^Tx + y^TA^TAy}$*.*

5.7 Gradient Penalty

Arjovsky et al. (2017) enforce the k-Lipschitz constraint on the discriminator by clipping the weight within a compact $[-\Delta, \Delta]$. Gulrajani et al. (2017) demonstrate that this approach leads to learn extremely simple functions. Showing that a differentiable function is 1-Lipschitz if and only if the gradient norm of the critic's output w.r.t. its input is one, they propose to enforce the Lipschitz constraint with a penalty on the gradient norm using samples uniformly drawn from a straight line connecting points of $\mathbb{P}_\mathbf{r}$ and $\mathbb{P}_g$. The objective function for regularized WGAN then becomes:

$$L(\theta, w) = \mathop{\mathrm{E}}_{x\sim Q}[D_w(x)] - \mathop{\mathrm{E}}_{z\sim F}[D_w(G_\theta(z))] - \lambda \mathop{\mathrm{E}}_{\widehat{x}\sim Q_\epsilon}[(\|\nabla_x D_w(\widehat{x})\| - 1)^2]$$

where Q_ϵ is the distribution of a random vector $\epsilon x + (1-\epsilon)G(z)$, when $x \sim Q$, $z \sim F$.

Proposition 13. Let $\mathbb{P}_\mathbf{r}$ and $\mathbb{P}_g$ be the two distributions in a compact metric $\mathcal{X}$. Then there is a 1-Lipschitz function f* solution of the problem $\mathrm{argmax}_{\|f\|_L\leq 1} E_{y\sim\mathbb{P}_\mathbf{r}}[f(y)] - E_{y\sim\mathbb{P}_g}[f(x)]$.
Let π the optimizer of $W(\mathbb{P}_\mathbf{r}, \mathbb{P}_g) = \inf_{\pi\in\Pi((\mathbb{P}_\mathbf{r},\mathbb{P}_g)} \mathrm{E}_{(x,y)\sim\pi}[\|x-y\|]$.
Then if f* is differentiable, $\pi(x = y) = 0$, and $x_t = (1-t)x + ty, t \in [0,1]$ then $\mathbb{P}_{(x,y)\sim\pi}[\nabla f^*(x_t) = \frac{y-x_t}{\|y-x_t\|}] = 1$.

Corollary 14. f* has gradient norm 1 almost everywhere under $\mathbb{P}_\mathbf{r}$ and $\mathbb{P}_g$.

Proof. Since X is a compact space, there exists an optimal f^*. If π is an optimal coupling:

$$\mathbb{P}_{(x,y)\sim\pi}[f^*(y) - f^*(x) = \|y - x\|] = 1$$

Let (x,y) such that $f^*(y) - f^*(x)] = \|y - x\|$, and WLOG $y \neq x$.
Let $\psi(t) = f^*(x_t) - f^*(x)$ and $t, t' \in [0, 1]$

$$\begin{aligned} |\psi(t) - \psi(t')| &= \|f^*(x_t) - f^*(x_{t'}\| && (12)\\ &\leq \|x_t - x_{t'}\| \text{ by optimality of x and y} && (13)\\ &= |t - t'|\|x - y\| && (14)\end{aligned}$$

Therefore ψ is $\|x - y\|$-Lipschitz. Now

$$\begin{aligned}\psi(1) - \psi(0) &= \psi(1) - \psi(t) + \psi(t) - \psi(0)\\ &\leq (1-t)\|x-y\| + \psi(t) - \psi(0)\\ &\leq (1-t)\|x-y\| + t\|x-y\|\\ &= \|x-y\|\end{aligned}$$

However $|\psi(1) - \psi(0)| = \|f^*(y) - f^*(x)\| = \|y - x\|$ so the inequality (14) has to be a strict equality. And $\psi(t) - \psi(0) = t\|x-y\|$, since $\psi(0) = f^*(x) - f^*(x) = 0$ therefore $\psi(t) = t\|x-y\|$.

Let

$$\begin{aligned} v &= \frac{y - x_t}{\|y - x_t\|}\\ &= \frac{y - ((1-t)x + ty)}{\|y - ((1-t)x + ty)\|}\\ &= \frac{(1-t)(y-x)}{\|(1-t)(y-x)\|}\\ &= \frac{y-x}{\|y-x\|}\end{aligned}$$

Now we know that $\psi(t) = t\|x - y\| = f^*(x_t) - f^*(x)$ so $f^*(x_t) = f^*(x) + t\|x - y\|$. And the directional derivative in the direction v is

$$\begin{aligned}\nabla_v f^* &= \lim_{h\to 0} \frac{f^*(x_t + hv) - f^*(x_t)}{h}\\ &= \lim_{h\to 0} \frac{f^*(x + t(y-x) + h\frac{y-x}{\|y-x\|}) - f^*(x_t)}{h}\\ &= \lim_{h\to 0} \frac{f^*(x_{t+\frac{h}{\|y-x\|}}) - f^*(x_t)}{h}\\ &= \lim_{h\to 0} f^*(x) + (t + \frac{h}{\|y-x\|})\|x-y\| - (f^*(x) + t\|x-y\|)\\ &= \lim_{h\to 0} \frac{h}{h}\\ &= 1\end{aligned}$$

Using Pythagoras:

$$\begin{aligned}\|\nabla f^*(x_t)\|^2 &= \|\nabla f^*(x_t) - \langle \nabla f^*(x_t), v\rangle.v + \langle \nabla f^*(x_t), v\rangle.v\|^2 \\ &= \|\nabla f^*(x_t) - \langle \nabla f^*(x_t), v\rangle.v\|^2 + \langle \nabla f^*(x_t), v\rangle.v\|^2 \\ &= \|\nabla f^*(x_t) - \nabla_v f^*(x_t).v\|^2 + \|\nabla_v f^*(x_t)\|^2 \\ &= \|\nabla f^*(x_t) - v\|^2 + 1\end{aligned}$$

So $\|\nabla f^*(x_t)\| \geq 1$ but since f^* is 1-Lipschitz then $\|\nabla f^*(x_t)\| = 1$ and $\|\nabla f^*(x_t) - v\| = 0$ therefore $\nabla f^*(x_t) = v$ or $\nabla f^*(x_t) = \frac{y-x}{\|y-x\|}$. Since this happens with probability 1 under π therefore $\mathbb{P}_{(x,y)\sim\pi}[\nabla f^*(x_t) = \frac{y-x_t}{\|y-x_t\|}] = 1$ □

Gradient penalty is an improvement compared to weight clipping but yet can leave out significant points of the underlying manifold that supports the real distribution. Wei et al. (2018) propose to draw a pair of data points near the manifold by perturbing each sampled data point x twice and use a Lipschitz constant to bound the difference between the discriminator's output to the perturbed data points x', x''.

Let $D : \mathcal{X} \Rightarrow \mathcal{Y}$ a K-Lipschitz discriminator then for all $x', x'' \in \mathcal{X}$:

$$\|D(x'), D(x'')\|_2 \leq K.\|x' - x''\|_2$$

The consistency regularization term CT is defined as

$$CT_{x',x''} = \mathop{\mathrm{E}}_{x\sim\mathbb{P}_\mathbf{r}}[\max(0, \|D(x'), D(x'')\|_2 + 0.1 \cdot \|D\_(x'), D\_(x'')\|_2 - K')]$$

where $D(x')$ is the output of the discriminator given the input x and after applying a dropout to the hidden layers of discriminator, and $D\_(.)$ is the second-to-last layer of the discriminator.

The modified objective function is:

$$L = \mathop{\mathrm{E}}_{z\sim\mathbb{P}_\mathbf{z}}[D(G(z))] - \mathop{\mathrm{E}}_{x\sim\mathbb{P}_\mathbf{r}}[D(x)] + \lambda_1 GP_{\widehat{x}} + \lambda_2 CT_{x',x''}$$

Experimental results show that CT-GAN compared to GP-GAN is less prone to overfitting, generates images with better inception score.

5.7.1 Proposed Improvement of the Gradient Penalty

In many real cases, $\mathbb{P}_\mathbf{r}$ rests in a low dimensional manifold (see Arjovsky and Bottou (2017)). Therefore it is expected that $\mathbb{P}_g$ is also in a low dimensional manifold. When they have disjoint supports, GAN can converge to ϵ-equilibrium in which an almost perfect discriminator can separate real and fake samples.

Instead of sampling along a line connecting G (Generator) and Discriminator (D), we propose to use Flaxman gradient descent analysis, sampling the gradient in a sphere centered around the data point x. The gradient is estimated using a unit sphere around x:

$$\nabla\widehat{f}(x) = \frac{d}{\delta}\mathop{\mathrm{E}}_{x\sim U(\mathbb{S}_1)}[f(x + \delta z]$$

The penalty PEN becomes $\lambda\,\mathrm{E}(\max(1 - \|\nabla\widehat{f}(x)\|, 0))$ and the cost function is:

$$L = \mathop{\mathrm{E}}_{z\sim\mathbb{P}_\mathbf{z}}[D(G(z))] - \mathop{\mathrm{E}}_{x\sim\mathbb{P}_\mathbf{r}}[D(x)] + \lambda PEN_x$$

The modified Adam Gradient Penalty algorithm becomes:

Require: The gradient penalty coefficient λ, the number of critic iterations per generator iteration n_{critic}, the batch size m, Adam hyperparameters α, β_1,β_2.
Require: initial critic parameters w_0, initial generator parameters θ_0.
while θ has not converged **do**
  for $t = 1, \cdots, n_{critic}$ **do**
    for $i = 1, \cdots, m$ **do**
      Sample real data $\boldsymbol{x} \sim \mathbb{P}_r$, random vector $\boldsymbol{z} \sim \mathbb{P}_z$, and a random number $\epsilon \sim U[0,1]$
      $\tilde{x} \leftarrow G(\boldsymbol{z})$
      $u \sim U(\mathbb{S}_1)$
      $x^* \leftarrow x + \delta u$
      $\hat{x} \leftarrow \epsilon \boldsymbol{x^*} + (1-\epsilon)\tilde{x}$
      $L^{(i)} \leftarrow D_w(\tilde{x}) - D_w(x) + \lambda(\|\nabla_{\hat{x}} D_w(\hat{x})\|_2 - 1)^2$
    end for
    $w \leftarrow Adam(\nabla_w \frac{1}{m} \sum_{i=1}^{m} L^{(i)}, w, \alpha, \beta_1, \beta_2)$
  end for
  Sample a batch of latent variables $\{z^{(i)}\}_{i=1}^{m} \sim \mathbb{P}_z$
  $\theta \leftarrow Adam(\nabla_\theta \frac{1}{m} \sum_{i=1}^{m} -D_w(G(\boldsymbol{z})), \theta, \alpha, \beta_1, \beta_2)$
end while

5.7.2 Curvature Exploitation

One open question which is not solved in the two papers concerns OMD behavior in the minimax game of GAN which is not convex-concave. One approach for solving non-convex minimization problem is to use the negative curvature of the objective function. Let $(\lambda_\theta, v_\theta)$ be the minimum eigenvalue of $\nabla^2_\theta f(z)$ with its associated eigenvector, and (λ_w, v_w) be the maximum eigenvalue of $\nabla^2_w f(z)$ with its associated eigenvector.

Define

$$v_z^{(-)} = \begin{cases} \frac{\lambda_\theta}{2\rho_\theta} \operatorname{sign}(v_\theta^T \nabla_\theta f(z)) v_\theta, & \text{if } \lambda_\theta < 0. \\ 0, & \text{otherwise.} \end{cases}$$

$$v_z^{(+)} = \begin{cases} \frac{\lambda_w}{2\rho_w} \operatorname{sign}(v_w^T \nabla_w f(z)) v_w, & \text{if } \lambda_w > 0. \\ 0, & \text{otherwise.} \end{cases}$$

$$v_z = (v_z^{(-)}, v_z^{(+)})$$

v_z is the extreme curvature vector in the direction of z.

The new updates rules for the gradient steps are:

$$\begin{cases} \theta_{t+1} & = \theta_t + v_{z_t}^{(-)} - \eta . \nabla_{\theta_t} f(\theta, w) \\ w_{t+1} & = w_t + v_{z_t}^{(+)} + \eta . \nabla_{w_t} f(\theta, w) \end{cases}$$

Then with smoothness conditions, when gradient and Hessian of $f(z) = f(\theta, w)$ are Lipschitz w.r.t to (θ, w), it is shown that the Gradient dynamics converge to a locally optimal saddle point (local stationary point).

However storing and computing the hessian in high dimensions is very costly. The most efficient approach for obtaining the eigenvalues and eigenvectors of $\nabla^2_\theta f(z)$ is power iterations on $I - \beta \nabla^2_\theta f(z)$ as

$$v_{t+1} = (I - \beta \nabla^2_\theta f(z)) v_t$$

where v_{t+1} is normalized after every iteration and $\beta > 0$ is chosen such that $I - \beta \nabla^2_\theta f(z) \geq 0$. This method is implemented through a Hessian-product and can be as efficient as gradient evaluations (Pearlmutter, 1994).

Putting all together yields the algorithm 1:

Algorithm 1 Curvature Exploitation for the saddle point problem

Require: Smooth f, (θ, w), β, ρ_θ, ρ_w

function MINIMUMEIGENVALUE(function: f , parameters: x)
  $v_0 \leftarrow$ random vector $\in \mathbb{R}^{|x|}$with unit length.
  for $i = 1, \cdots, k$ **do**
    $v_i \leftarrow (I - \beta \nabla^2_{\mathbf{x}} f)\, v_{i-1}$
    $v_i \leftarrow \frac{v_i}{\|v_i\|}$
  end for
  $v \leftarrow -v_k \,.\, \text{sign}\,(v_k^T \nabla_x f))$
  $\lambda \leftarrow v^T \nabla^2_x f v$
  if $\lambda \geq 0$ **then**
    $\lambda \leftarrow 0$
    $v \leftarrow 0$
  end if
return v, λ
end function
for $t = 1, \cdots, T$ **do**
  $v_\theta, \lambda_\theta \leftarrow$ MINIMUMEIGENVALUE($f(\theta_t, w_t)$,θ_t)
  $v_w, -\lambda_w \leftarrow$ MINIMUMEIGENVALUE($-f(\theta_t, w_t)$,w_t)
  $\theta_t \leftarrow \theta_t + \frac{\lambda_\theta}{2\rho_\theta} - \eta \nabla_\theta f(\theta_t, w_t)$
  $w_t \leftarrow w_t + \frac{\lambda_w}{2\rho_w} + \eta \nabla_w f(\theta_t, w_t)$
end for

In practice GAN training is done by splitting the initial loss function in two individual optimization problems:

$$\max_w \mathop{\mathrm{E}}_{x \sim P_r}[D_w(x)] + \mathop{\mathrm{E}}_{z \sim P_z}[\log\left(1 - D_w(G_\theta(z))\right)]$$
$$\min_\theta - \mathop{\mathrm{E}}_{z \sim P_z}[\log\left(1 - D_w(G_\theta(z))\right)]$$

The algorithm 1 is then applied to each equation in turn.

6 Conclusion And Future Directions

We have analyzed two different strategies for Domain Adaption. They are different but they also show common grounds, both require convexity of the loss function (for GAN the concave condition can be reformulated as convex), and the loss function must be K-Lipschitz. They both aim at solving a minimax problem, for GDM it is reducing the discrepancy between distributions and for a GAN the goal is to reach the equilibrium of a zero-sum game. Combining both approaches in a single theory will allow to enjoy the power of generative models using a distance which takes into consideration the structure of the hypothesis set of the learning problem. It is a challenging project, one may want to tackle in the future.

References

M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks, 2017.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan, 2017.

S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and equilibrium in generative adversarial nets (gans). *CoRR*, abs/1703.00573, 2017. URL http://arxiv.org/abs/1703.00573.

S. Ben-David and R. Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, ALT'12, pages 139–153, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-34105-2. doi: 10.1007/978-3-642-34106-9_14. URL http://dx.doi.org/10.1007/978-3-642-34106-9_14.

C. Cortes and M. Mohri. Domain adaptation in regression. In J. Kivinen, C. Szepesvári, E. Ukkonen, and T. Zeugmann, editors, *Algorithmic Learning Theory*, pages 308–323, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-24412-4.

C. Cortes, M. Mohri, and A. M. Medina. Adaptation algorithm and theory based on generalized discrepancy. *CoRR*, abs/1405.1503, 2014. URL http://arxiv.org/abs/1405.1503.

C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training gans with optimism. *CoRR*, abs/1711.00141, 2017. URL http://arxiv.org/abs/1711.00141.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017. URL http://arxiv.org/abs/1704.00028.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *CoRR*, abs/0902.3430, 2009. URL http://arxiv.org/abs/0902.3430.

A. M. Medina. Learning theory and algorithms for auctioning and adaptation problems. 2015.

B. A. Pearlmutter. Fast exact multiplication by the hessian. *Neural Computation*, 6:147–160, 1994.

A. Rakhlin and K. Sridharan. Online learning with predictable sequences, 2012.

X. Wei, B. Gong, Z. Liu, W. Lu, and L. Wang. Improving the improved training of wasserstein gans: A consistency term and its dual effect. *CoRR*, abs/1803.01541, 2018. URL http://arxiv.org/abs/1803.01541.