Rates of convergence for density estimation with GANs

Denis Belomestny, 1,3 Eric Moulines, 1,4 Alexey Naumov, 1 Nikita Puchkin, 1,2 and Sergey Samsonov 1

1HSE University, Russia, e-mail: anaumov@hse.ru; npuchkin@hse.ru; svsamsonov@hse.ru
2Institute for Information Transmission Problems RAS, Russia
3Duisburg-Essen University, Germany, e-mail: denis.belomestny@uni-due.de
4Ecole Polytechnique, France, e-mail: eric.moulines@polytechnique.edu

Abstract: We undertake a precise study of the non-asymptotic properties of vanilla generative adversarial networks (GANs) and derive theoretical guarantees in the problem of estimating an unknown $d$-dimensional density $p^*$ under a proper choice of the class of generators and discriminators. We prove that the resulting density estimate converges to $p^*$ in terms of Jensen-Shannon (JS) divergence at the rate $(\log n/n)^{2/\beta} (2\beta+d)$ where $n$ is the sample size and $\beta$ determines the smoothness of $p^*$. This is the first result in the literature on density estimation using vanilla GANs with JS rates faster than $n^{-1/2}$ in the regime $\beta > d/2$.

Keywords and phrases: generative model, excess risk bound, Jensen-Shannon risk, smoothness class.

1. Introduction

A generative adversarial network (GAN) is a minimax game between a generator $g$ whose goal is to generate fake samples that are close to the real data and a discriminator $D$ whose goal is to distinguish between the real and fake samples. From the perspective of statistics, GANs can be viewed as an unsupervised method to learn target data distributions. The main strand of research on GANs deals with empirical insights and basic mathematical properties. Recently researchers started to analyze the GAN problem from the statistical perspectives [Biau et al., 2020b,a, Liang, 2018, Singh et al., 2018, Luise et al., 2020, Uppal et al., 2019] as well as optimization and algorithmic viewpoints [Liang and Stokes, 2019, Kodali et al., 2017, Pfau and Vinyals, 2016, Nie and Patel, 2020, Nagarajan and Kolter, 2017, Genevay et al., 2018, 2019].

The following minimax problem is the original GAN problem, also called vanilla GAN, introduced in Goodfellow et al. [2014]

$$\min_{g \in \mathcal{G}} \max_{D \in \mathcal{D}} \mathbb{E}[\log D(X)] + \mathbb{E}[\log(1 - D(g(Y)))].$$

Here $Y$ denotes the generator’s input (latent variable), $X$ represents the random vector for the real data with unknown distribution $P^*$, $\mathcal{G}$ and $\mathcal{D}$ represent the classes of generators and discriminators, respectively. Implementation of this minimax game using deep neural network classes $\mathcal{G}$
and \( D \) lead to the state-of-the-art generative model for many different tasks. To shed light on the probabilistic meaning of vanilla GAN, \cite{Goodfellow2014} shows that given an unconstrained discriminator \( D \), that is, if \( D \) contains all possible functions, the minimax problem (1) will reduce to

\[
\min_{g \in \mathcal{G}} \text{JS}(P_{g(Y)}, P^*),
\]

where JS stands for the Jensen-Shannon (JS) divergence. The optimization problem (2) can be interpreted as finding the closest generative model to the data distribution \( P^* \) where distance is measured by the JS divergence. Various GAN formulations were later proposed by changing the divergence measure in (2): f-GAN \cite{Nowozin2016} generalizes vanilla GAN by minimizing a general f-divergence; Wasserstein GAN (WGAN) \cite{Arjovsky2017} considers the first-order Wasserstein (Kantorovich) distance (\( W_1 \) distance); MMD-GAN \cite{Dziugaite2015} considers the maximum mean discrepancy; energy-based GAN \cite{Zhao2016} minimizes the total variation distance as discussed in \cite{Arjovsky2017}; Quadratic GAN \cite{Feizi2020} finds the distribution minimizing the second-order Wasserstein (Kantorovich) distance.

**Model** In this paper, we focus on the following setup. Suppose that \( X_1, \ldots, X_n \) are independent identically distributed (i.i.d.) random vectors in \( \mathbb{R}^d \) drawn from the distribution \( P^* \) with a Lebesgue density \( p^* \) supported on a compact set \( X \subset \mathbb{R}^d \). Let \( Y_1, \ldots, Y_n \) be i.i.d. latent variables with a density \( \phi \) supported on a compact set \( Y \subset \mathbb{R}^d \). Given a family \( \mathcal{G} \) of invertible transformations \( g : Y \to X \) and a family \( \mathcal{D} \) of discriminators \( D : X \to (0, 1) \), we consider the empirical counterpart of the optimization problem (1):

\[
\hat{g} \in \arg\min_{g \in \mathcal{G}} \max_{D \in \mathcal{D}} L_n(g, D),
\]

where

\[
L_n(g, D) := \frac{1}{2n} \sum_{i=1}^{n} \log D(X_i) + \frac{1}{2n} \sum_{j=1}^{n} \log (1 - D(g(Y_j)))
\]

is the empirical version of the functional

\[
L(g, D) := \frac{1}{2} \int \log D(x)p^*(x) \, dx + \frac{1}{2} \int \log (1 - D(z))p_g(z) \, dz.
\]

In (4), we assume that the number of fake samples is equal to the number of real ones but our analysis is also valid for the case when the number of fake instances is larger than \( n \). In (5), \( p_g \) denotes the density of \( g(Y) \). The change-of-variables formula implies that

\[
p_g(x) = |\det[\nabla g(g^{-1}(x))]|^{-1} \phi(g^{-1}(x)),
\]

provided that \( \nabla g \) is non-degenerate. Our goal is to obtain a bound on the discrepancy between the estimate \( p_{\hat{g}} \) and the true distribution \( p^* \) in terms of the JS divergence. Thus our work can be seen as a continuation of the line of research initiated by \cite{Biau2020a} on theoretical properties of vanilla GANs. Note that for WGANs the convergence rates were thoroughly studied in the literature. For example, \cite{Schreuder2020} derived risk bounds where the generative model.
is based on a $\beta$ times differentiable transformation of the unit hypercube of dimension $d$. The obtained rates (in $W_1$ distance) are of order $n^{-\beta/d} \vee n^{-1/2}$. In Liang [2018], the author obtained the minimax rates $n^{-(\alpha+\beta)/(2\beta+d)} \vee n^{-1/2}$. As opposed to WGANs, the rates of convergence for vanilla GANs are not yet fully understood.

Contributions

Our contributions can be summarised as follows.

- Under the assumption that $G$ and $D$ are bounded subsets of the Hölder classes $H^{1+\beta}(Y)$ and $H^{\alpha}(X)$, respectively, for some $\alpha, \beta > 0$, with polynomially growing covering numbers, we prove the bound

$$
\mathbb{E} \text{JS}(p_{\hat{g}}, p^*) - \Delta_G - \Delta_D \lesssim \sqrt{\frac{\Delta_G \log(n)}{n}} + \sqrt{\frac{\Delta_D \log(n)}{n}} + \frac{\log(n)}{n},
$$

(7)

with

$$
\Delta_G := \min_{g \in G} \text{JS}(p_g, p^*), \quad \Delta_D := \max_{g \in G} \min_{D \in D} [L(g, D^*_g) - L(g, D)]
$$

(8)

and

$$
D^*_g(x) := \frac{p^*(x)}{p^*(x) + p_g(x)}, \quad x \in X.
$$

(9)

Here the notation $f(n) \lesssim g(n)$ means that for all positive integers $n$ $f(n) \leq Cg(n)$ for some constant $C$ not depending on $n$. It was shown in Goodfellow et al. [2014] that, for any $g \in G$ and any measurable $D : X \to (0, 1)$,

$$
L(g, D^*_g) - L(g, D) = \int_X K(D^*_g(x), D(x)) \frac{p_g(x) + p^*(x)}{2} \, dx,
$$

(10)

where, for any $u, v \in (0, 1)$,

$$
K(u, v) = u \log \frac{u}{v} + (1 - u) \log \frac{1 - u}{1 - v} \geq 0.
$$

(11)

Hence, $D^*_g$ is the optimal discriminator, that is $D^*_g = \arg\max_D L(g, D)$ for any $g \in G$. The bound (7) is a sharp oracle inequality. This bound significantly improves upon the existing results obtained in the literature for vanilla GANs. For example, Bian et al. [2020a] obtained an upper bound of the form (7) with the right-hand side of order $O(n^{-1/2})$ without $\Delta_G$ and $\Delta_D$ under the square root. The absence of these terms makes the obtained bound too rough, especially if $\Delta_G$ and $\Delta_D$ are small.

- Using the bound (7) and deep neural networks families for $G$ and $D$, we derive the convergence rates of the JS divergence between $p^*$ and $p_{\hat{g}}$ to zero:

$$
\mathbb{E} \text{JS}(p_{\hat{g}}, p^*) \lesssim \left( \frac{\log n}{n} \right)^{\frac{2\beta}{2\beta+d}}
$$

(12)

in the case when the true density $p^*$ is the density of the random variable $g^*(Y)$ for a smooth invertible transform $g^*$ not necessary belonging to $G$. The convergence rates (12) match the well-known minimax bounds of density estimation in $L_2(X)$ (note that $p^* \in H^\beta(X)$.
if $g^* \in \mathcal{H}^{1+\beta}(\mathcal{Y})$, see Tsybakov [2008], Section 1.2, and McDonald [2017]. To the best of our knowledge, this is the first result in the literature giving convergence rates of the density estimate $p_\hat{g}$ in the JS divergence for vanilla GANs faster than $n^{-1/2}$ for the case $\beta > d/2$. In this context, let us also mention a fully nonparametric setting studied recently in Asatryan et al. [2020] where both classes $\mathcal{D}$ and $\mathcal{G}$ are assumed to be closed subsets of $C^\alpha([0,1]^d)$. Asatryan et al. proved the consistency of the corresponding density estimator in JS divergence.

**Notations** Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. For two probability measures $P \ll Q$ on a measurable space $(\Omega, \mathcal{B}(\Omega))$ with Lebesgue densities $p$ and $q$, respectively, we define the Kullback-Leibler divergence between $P$ and $Q$ as $\text{KL}(P, Q) := \int p(x) \log \left( \frac{p(x)}{q(x)} \right) dx$. By $\text{JS}(P, Q) = \text{JS}(p, q)$, we denote the Jensen-Shannon divergence $\text{JS}(P, Q) := (\text{KL}(P, (P + Q)/2) + \text{KL}(Q, (P + Q)/2))/2$. For a function $f : \Omega \rightarrow \mathbb{R}^d$ we set

$$\|f\|_\infty := \sup_{x \in \Omega} |f(x)|,$$

$$\|f\|_{L_2(\Omega)} := \left\{ \int_{\Omega} |f(x)|^2 \, dx \right\}^{1/2},$$

and

$$\|f\|_{L_2(p)} := \left\{ \int_{\Omega} |f(x)|^2 p(x) \, dx \right\}^{1/2}.$$

For any $s \in \mathbb{N}$, the function space $C^s(\Omega)$ consists of those functions over the domain $\Omega$ which have partial derivatives up to order $s$ in $\Omega$, and these derivatives are moreover bounded and continuous in $\Omega$. Formally,

$$C^s(\Omega) = \{ f : \Omega \rightarrow \mathbb{R}^m : \| f \|_{C^s} := \max_{|\gamma| \leq s} \| D^\gamma f \|_\infty < \infty \},$$

where, for any multi-index $\gamma = (\gamma_1, \ldots, \gamma_d) \in \mathbb{N}_0^d$, the partial differential operator $D^\gamma$ is given by $D^\gamma f_i := \partial^{\gamma_i} f_i/\partial x_{i_1}^{\gamma_{i_1}} \cdots \partial x_{i_d}^{\gamma_d}, i = 1, \ldots, m$, and $\| D^\gamma f \|_\infty := \max_{i = 1, \ldots, m} \| D^\gamma f_i \|_\infty$. Here we have written $|\gamma| = \sum_{i=1}^d \gamma_i$ for the order of $D^\gamma$. For the matrix of first derivatives, we use the usual notation $\nabla f = (\partial f_i/\partial x_j)\ i = 1, \ldots, m, j = 1, \ldots, d$. For a function $f : \Omega \rightarrow \mathbb{R}^m$ and any positive number $0 < \delta \leq 1$, the Hölder constant of order $\delta$ is given by

$$[f]_\delta := \max_{i = 1, \ldots, m} \sup_{x \neq y \in \Omega} \frac{|f_i(x) - f_i(y)|}{\min(1, \| x - y \|_\delta)}.$$

Now, for any $\alpha > 0$, we can define the Hölder ball $\mathcal{H}^{s(\Omega, H)}$. If we let $s = \lfloor \alpha \rfloor$ be the largest integer strictly less than $\alpha$, it contains those functions in $C^s(\Omega)$ which have $\delta$-Hölder-continuous, $\delta = \alpha - s > 0$, partial derivatives of order $s$. Formally,

$$\mathcal{H}^{s(\Omega, H)} = \{ f \in C^s(\Omega) : \| f \|_{\mathcal{H}^s} := \max\{ \| f \|_{C^s}, \max_{|\gamma| = s} \| D^\gamma f \|_s \} \leq H \}.$$

Note that if $f \in \mathcal{H}^{1+\beta}(\Omega, H)$ for some $\beta > 0$, then it holds for $i = 1, \ldots, m, j = 1, \ldots, d$,

$$\left| \frac{\partial f_i(x)}{\partial x_j} - \frac{\partial f_i(y)}{\partial x_j} \right| \leq \| f \|_{\mathcal{H}^{1+\beta}} \cdot \| x - y \|^{1+\beta} \leq H \cdot \| x - y \|^{1+\beta}, \quad x, y \in \Omega.$$
for any $x, y \in \Omega$, since $\|f\|_{H^\beta_1} \leq \|f\|_{H^\beta_2}$ for any $\beta_2 \geq \beta_1$. We will also write $f \in H^\beta(\Omega)$ if $f \in H^\beta(\Omega, H)$ for some $H < \infty$.

To give a formal definition of the neural network, we first fix an activation function $\sigma : \mathbb{R} \to \mathbb{R}$. For a bias vector $v = (v_1, \ldots, v_p) \in \mathbb{R}^p$, we define the shifted activation function $\sigma_v : \mathbb{R}^p \to \mathbb{R}^p$ as

$$
\sigma_v(x) = (\sigma(x_1 - v_1), \ldots, \sigma(x_p - v_p)), x = (x_1, \ldots, x_p) \in \mathbb{R}^p.
$$

A neural network of depth $L$ (with $L$ hidden layers) is then a function of the form

$$
f : \mathbb{R}^p_0 \to \mathbb{R}^{p_{L+1}}, \quad f(x) = W_L \circ \sigma_{v_L} \circ W_{L-1} \circ \sigma_{v_{L-1}} \circ \cdots \circ W_0 \circ x,
$$

where $W_i \in \mathbb{R}^{p_{i+1} \times p_i}$ is a weight matrix and $v_i \in \mathbb{R}^{p_i}$ is a bias vector. In our paper, we consider ReLU activation function (rectified linear unit), which is defined as

$$
\sigma(x) = x_+ = x \vee 0
$$

and ReQU (rectified quadratic unit), defined as

$$
\sigma^2(x) = (x \vee 0)^2.
$$

2. Assumptions

Throughout this paper, we assume the following conditions on $p^*, \mathcal{D}$ and $\mathcal{G}$. Fix some some real numbers $\alpha > 0, \beta > 0$.

**Assumption A$\phi$.** There exist constants $H_\phi > 0$ and $\Phi > 1$ such that $\phi \in H^\beta(\mathbb{Y}, H_\phi)$ and

$$
\Phi^{-1} \leq \phi(y) \leq \Phi, \quad y \in \mathbb{Y}.
$$

**Assumption A$p^*$.** There exist constants $H^* > 0$ and $\Lambda > 1$ such that $p^*$ is of the form (6) with $g^* \in H^{1+\beta}(\mathbb{Y}, H^*)$ and

$$
\Lambda^{-2}I_{d \times d} \preceq \nabla g^*(y)^\top \nabla g^*(y) \preceq \Lambda^2 I_{d \times d}, \quad y \in \mathbb{Y}.
$$

**Assumption A$G$.** There exist constants $H_G > 0$ and $\Lambda > 1$ such that $\mathcal{G} \subseteq H^{1+\beta}(\mathbb{Y}, H_G)$ and

$$
\Lambda^{-2}I_{d \times d} \preceq \nabla g(y)^\top \nabla g(y) \preceq \Lambda^2 I_{d \times d}, \quad y \in \mathbb{Y}
$$

for all $g \in \mathcal{G}$. In addition, there exist constants $A_G, B_G, \gamma, \varepsilon_0 > 0$ such that the covering number $N(\mathcal{G}, \| \cdot \|_{H^{1+\beta}(\mathbb{Y})}, \varepsilon)$ satisfies

$$
N(\mathcal{G}, \| \cdot \|_{H^{1+\beta}(\mathbb{Y})}, \varepsilon) \leq A_G \left( \frac{B_G}{\varepsilon} \right)^\gamma, \quad \varepsilon \in (0, \varepsilon_0).
$$

(16)
The conditions (14) and (15) ensure well-posedness of the densities \( p_g \) and \( p_g^* \) defined using the change-of-variables formula (6) and cannot be avoided in the problem of density estimation. Note that we can always assume that \( A \) is the same in (14) and (15). This does not restrict the generality since we can always take the maximum of these two constants if they are different. Also, let us remark that one can only require the bound (14) if we restrict ourselves to a subset of \( G \) with good approximation properties with respect to \( g^* \). In particular, if we can find a sequence of generators \( g_n \) from \( H^{1+\beta}(\mathcal{Y}) \) that converge to \( g^* \) in \( H^{1+\beta}(\mathcal{Y}) \) as \( n \to \infty \), then the inequality (15) holds automatically for all \( n > n_0 \) with \( A \) replaced by \( A + \delta \) for arbitrary small \( \delta > 0 \) and \( n_0 = n_0(\delta) \). This would suffice to derive the rates (12) using, for example, deep neural networks with a smooth activation function (see the proof of Theorem 2 for further details). Finally, we impose the following condition on the class \( D \).

**Assumption AD.** There exists \( H_D > 0 \) such that \( D \subseteq \mathcal{H}^{\alpha}(X, H_D) \). Moreover suppose that \( D(x) \in [D_{\min}, D_{\max}] \subseteq [0, 1] \) for all \( x \in X, D \in D \). In addition, there exist constants \( A_D, B_D, \eta, \varepsilon_0 > 0 \) such that the covering number \( \mathcal{N}(D, \| \cdot \|_\infty, \varepsilon) \) satisfies

\[
\mathcal{N}(D, \| \cdot \|_\infty, \varepsilon) \leq A_D \left( \frac{B_D}{\varepsilon} \right)^\eta, \quad \varepsilon \in (0, \varepsilon_0).
\]

(17)

The requirement that all functions from \( D \) are bounded away from 0 and 1 is needed for the \( \log D \) and \( \log(1-D) \) to be well defined. Similar conditions appear in the literature for aggregation with Kullback-Leibler loss (for instance, in Polzehl and Spokoiny [2006], Belomestny and Spokoiny [2007], Rigollet [2012], Butucea et al. [2017]).

**Remark 1.** Without loss of generality we can assume that \( D^*_g \in [D_{\min}, D_{\max}] \), since under assumptions \( A\Phi, A p^* \) and \( A G \),

\[
p_{\min} \leq p_g(x) \leq p_{\max} \quad \text{and} \quad p_{\min} \leq p^*(x) \leq p_{\max}, \quad \text{for all} \ x \in X, \ g \in G
\]

with \( p_{\min} = (\Phi A^d)^{-1}, p_{\max} = \Phi A^d \) (see Lemma 3 for the details). Hence we can take \( 0 < D_{\min} \leq p_{\min}/(p_{\min} + p_{\max}), 1 > D_{\max} \geq p_{\max}/(p_{\min} + p_{\max}) \).

### 3. Main results

In this section, we formulate the main results of our paper.

**Theorem 1.** Let \( \alpha, \beta > 0 \) and assume \( A\Phi, A p^*, \text{AD} \) and \( AG \). Then for any \( \delta \in (0, 1/4) \) with probability at least \( 1 - 4\delta \),

\[
\text{JS}(p_\tilde{g}, p^*) - \Delta D - \Delta G \leq \sqrt{\frac{\Delta G}{n}\varphi(n, \gamma, \delta)} + \sqrt{\frac{\Delta D}{n}\psi(n, \gamma, \eta, \delta)} + \frac{\varphi^2(n, \gamma, \delta) + \psi^2(n, \gamma, \eta, \delta)}{n},
\]

where

\[
\varphi^2(n, \gamma, \delta) = \gamma \log n + \log (1/\delta)
\]
and

\[ \psi^2(n, \gamma, \eta, \delta) = (\gamma + \eta) \log n + \log(1/\delta). \]

Here \( \lesssim \) stands for inequality up to a constant depending on the parameters \( \alpha, \beta, \Phi, \Lambda, A_G, A_D, B_G \) and \( B_D \), see proof of Theorem 1 (Section 4) for precise dependence.

The theoretical properties of vanilla GANs were studied in Biau et al. [2020a] and Asatryan et al. [2020]. The closest result to our Theorem 1 in the literature is Theorem 4.1 from Biau et al. [2020a], so let us focus on the comparison of these two results. Let us first elaborate on the conditions of the theorems. In Biau et al. [2020a], the authors assume that the generators are parametrized by a parameter \( \theta \in \mathbb{R}^\gamma \) (i.e. \( \mathcal{G} = \{g_\theta : \theta \in \Theta \subset \mathbb{R}^\gamma\} \)) and require the maps \( \theta \mapsto g_\theta \) and \( (\theta, x) \mapsto p_{g_\theta}(x) \) to be in \( C^1(\Theta) \). Similarly, they parametrize discriminators by \( \nu \in \mathbb{R}^\eta \) (i.e. \( \mathcal{D}_\nu = \{D_\nu : \nu \in V \subset \mathbb{R}^\eta\} \)). Unfortunately, the authors do not provide any examples of when these assumptions are fulfilled. However, the role of these requirements in the proof of Theorem 4.1 in Biau et al. [2020a] is clear: they yield that the covering numbers of the classes \( \mathcal{D} \) and \( \{p_\theta : \theta \in \mathcal{G}\} \) with respect to the norm \( \| \cdot \|_\infty \) have a polynomial behaviour. Hence, Assumption AG and Assumption AD can be considered as an alternative to the group of assumptions \( (H_{reg}) \) introduced in Biau et al. [2020a]. Besides, Biau et al. [2020a] requires that \( D^*_\nu \) is bounded away from 0 and 1 for all \( \theta \in \Theta \). Under the aforementioned assumptions, Biau et al. [2020a] established the following theoretical guarantees on the excess JS risk of the vanilla GAN estimate \( p_\tilde{\varphi} \):

\[
\mathbb{E} \text{JS}(p_\tilde{\varphi}, p^*) - \Delta_\mathcal{G} - \Delta_\mathcal{D} \lesssim \sqrt{\frac{\gamma + \eta}{n}} .
\]

(18)

The proof of this result relies on the chaining technique. In addition, one can use McDiarmid’s inequality (see, e.g. [Boucheron et al., 2004, Corollary 4]) to get large deviation bounds on \( \text{JS}(p_\tilde{\varphi}, p^*) \) of the form

\[
\text{JS}(p_\tilde{\varphi}, p^*) - \Delta_\mathcal{G} - \Delta_\mathcal{D} \lesssim \sqrt{\frac{\gamma + \eta}{n}} + \sqrt{\frac{\log(1/\delta)}{n}},
\]

which holds with probability at least \( 1 - \delta \). In our case, Theorem 1 yields

\[
\text{JS}(p_\tilde{\varphi}, p^*) - \Delta_\mathcal{G} - \Delta_\mathcal{D} \lesssim \sqrt{\frac{\gamma \Delta_\mathcal{G} \log(n/\delta)}{n}} + \sqrt{\frac{\eta \Delta_\mathcal{D} \log(n/\delta)}{n}} + \frac{\log(n/\delta)}{n}.
\]

If the classes \( \mathcal{G} \) and \( \mathcal{D} \) are poor and cannot approximate \( g^* \) and \( \mathcal{D}^*_g \), \( g \in \mathcal{G} \), respectively, with high accuracy, then \( \Delta_\mathcal{G} \) and \( \Delta_\mathcal{D} \) are of order 1, and our result shows no improvements over Biau et al. [2020a]. In practice, one uses the classes of deep neural networks for \( \mathcal{G} \) and \( \mathcal{D} \) with good approximation quality. Recent results from the theory of approximation for deep neural networks (see e.g. Schmidt-Hieber [2020]) suggest that \( \Delta_\mathcal{G} \) and \( \Delta_\mathcal{D} \) tend to 0 (with polynomial decay) as the sample size \( n \) tends to \( \infty \), provided that the parameters of the neural networks (width, depth, and the number of non-zero weights) are chosen carefully. In this case, the result of Theorem 1 substantially improves the bound in [Biau et al., 2020a, Theorem 4.1] and leads to the following statement.

**Theorem 2.** Let \( \alpha > 0 \) and \( \beta \geq 1 \). Assume \( A \mathbf{p}^* \) and \( A \Phi \). There is a family of multi-layer neural networks \( \mathcal{D} \) with ReLU activation function and a family of multi-layer neural networks \( \mathcal{G} \)
with activation functions ReLU and ReQU satisfying Assumptions AD and AG, respectively such that, for any $\delta \in (0, 1)$ with probability at least $1 - \delta$, it holds that

$$\text{JS}(\hat{p}, p^*) \lesssim \left(\frac{\log(n/\delta)}{n}\right)^{\frac{2\beta}{4\beta + d}},$$

where $\lesssim$ means inequality up to a constant not depending on $n$.

Under Assumption AG and Assumption Ap*, the densities $\hat{p}$ and $p^*$ are bounded away from 0 and $\infty$. Lemma 10 implies that $\text{JS}(\hat{p}, p^*)$ is equivalent to $\|\hat{p} - p^*\|_{L_2(p^*)}^2$. It is known that kernel density estimates and projection estimates achieve the minimax rate of convergence $O(n^{-2\beta/(2\beta + d)})$ in the problem of density estimation when the target density is bounded away from 0. Then Theorem 2 claims that, under some additional technical assumptions, GAN achieves the same rate of convergence up to a logarithmic factor.

It is worth mentioning that, under the conditions of Theorem 2, the bound (18) in [Biau et al., 2020a, Theorem 4.1] would imply

$$\mathbb{E} \text{JS}(\hat{p}, p^*) \lesssim \left(\frac{\log n}{n}\right)^{-2\beta/(4\beta + d)},$$

leading to suboptimal rates of convergence. Hence, the result of Theorem 2 does not follow from Biau et al. [2020a]. This simple example illustrates the significance of improvements in Theorem 1 over [Biau et al., 2020a, Theorem 4.1].

The same concerns the bounds in [Asatryan et al., 2020, Theorem 5.4]. Similarly to Biau et al. [2020a], the authors use the chaining technique to prove that $\text{JS}(\hat{p}, p^*) \to 0$ as $n \to \infty$ almost surely in the case $\beta > d/2$ (see Theorem 5.4).

4. Proofs of the main results

Proof of Theorem 1. We begin with studying the excess risk

$$\text{JS}(\hat{p}, p^*) - \inf_{g \in G} \text{JS}(p, p^*) \equiv \text{JS}(\hat{p}, p^*) - \text{JS}(\hat{g}, p^*),$$

where we have introduced $\hat{g} \in \arg\min_{g \in G} \text{JS}(p, p^*)$. To simplify notations, let us denote $F(g) = \text{JS}(p, p^*)$, $g \in G$. Observe that $F(g)$ may be rewritten as $F(g) = L(g, D_g^*) + \log 2$, where $L(g, D)$ and $D_g^*$ are defined in (5) and (9), respectively. This representation allows us to introduce an empirical counterpart of $F(g)$ as $F_n(g) = L_n(g, D_n^*) + \log 2$, where $L_n$ is given in (4). The special structure of $D_g^*$ together with assumptions Ap* and AG allows us to obtain high-probability bounds on $|L_n(g, D_n^*) - L(g, D_n^*)|$ (check Lemma 8 for more details). We rewrite (20) as follows:

$$F(\hat{g}) - F(\hat{g}) = \underbrace{(F(\hat{g}) - F_n(\hat{g}))}_{T_1} + \underbrace{(F_n(\hat{g}) - F_n(\hat{g}))}_{T_2} + \underbrace{(F_n(\hat{g}) - F_n(\hat{g}))}_{T_3}.$$

Step 1 To control $T_1$ and $T_3$ in (21) we apply high-probability bounds on $|F(g) - F_n(g)|$. We need the following result.
Proposition 1. Grant the assumptions of Theorem 1 and fix \( \varepsilon \in (0, \varepsilon_0) \). Let \( \mathcal{G}_\varepsilon \) be a minimal \( \varepsilon \)-net of the set \( \mathcal{G} \) with respect to \( \| \cdot \|_{\mathcal{H}^{\beta+1}(\gamma)} \). Then, for any \( \delta \in (0,1) \), with probability at least \( 1 - \delta \) for any \( g \in \mathcal{G} \),

\[
|F_n(g) - F(g)| \leq C_{36a} \varepsilon^{\beta \wedge 1} + C_{36b} \varepsilon^{\beta \wedge 1} \sqrt{\frac{\log(2|\mathcal{G}_\varepsilon|/\delta)}{n}} + C_{36c} \varepsilon^{\beta \wedge 1} \sqrt{\frac{\log(2|\mathcal{G}_\varepsilon|/\delta)}{n}} + C_{36d} \varepsilon^{\beta \wedge 1} \sqrt{\frac{\log(2|\mathcal{G}_\varepsilon|/\delta)}{n}},
\]

where the constants \( C_{36a}, C_{36b}, C_{36c} \) and \( C_{36d} \) are given in (36).

The proof of Proposition 1 is provided in Appendix A.1. Now, we bound \( |\mathcal{G}_\varepsilon| \) using (17) and optimize the right-hand side of (22) over \( \varepsilon \). Taking \( \varepsilon = n^{-1/(\beta \wedge 1)} \), we obtain

\[
|F_n(g) - F(g)| \leq \sqrt{\frac{F(g) \wedge F_n(g)}{n}} \varphi(n, \gamma, \delta) + \frac{\varphi^2(n, \gamma, \delta)}{n}
\]

with probability at least \( 1 - \delta \), where \( \varphi(n, \gamma, \delta) \) is defined as

\[
\varphi^2(n, \gamma, \delta) = C_{24a} \gamma \log n + C_{24b} \log (1/\delta) + C_{24c},
\]

with

\[
C_{24a} = \frac{C_{36b} + C_{36c}^2 + C_{36d}}{\beta \wedge 1},
\]

\[
C_{24b} = C_{36b} + C_{36c}^2 + C_{36d},
\]

\[
C_{24c} = C_{36a} + \frac{C_{36b}^2}{n} + (C_{36b} + C_{36c}^2 + C_{36d}) \log (2A\gamma).
\]

Now (21), (23) and \( F_n(\bar{g}) \leq F_n(\overline{\gamma}) \) imply that, with probability at least \( 1 - \delta \),

\[
F(\bar{g}) - F(\overline{\gamma}) \leq T_2 + 2\varphi(n, \gamma, \delta) \sqrt{\frac{F_n(\overline{\gamma})}{n}} + \frac{2\varphi^2(n, \gamma, \delta)}{n}.
\]

In order to bound \( F_n(\overline{\gamma}) \), we apply again Proposition 1. It holds with probability at least \( 1 - \delta \), that

\[
F_n(\overline{\gamma}) \leq F(\overline{\gamma}) + \frac{\sqrt{F(\overline{\gamma})}}{n} \varphi(n, \gamma, \delta) + \frac{\varphi^2(n, \gamma, \delta)}{n} \leq \frac{3F(\overline{\gamma})}{2} + \frac{3\varphi^2(n, \gamma, \delta)}{2n},
\]

where (a) follows from the Young's inequality.

Step 2  The remaining term to bound is the right-hand side of (25) is

\[
T_2 = F_n(\bar{g}) - F_n(\overline{\gamma}) = \sum_{D \in D} L_n(g, D) - L_n(\overline{\gamma}, D).
\]

Let us introduce \( \hat{D}_g \in \arg \max_{\bar{D} \in D} L_n(g, D) \) for \( g \in \mathcal{G} \). Then we can represent

\[
L_n(\bar{g}, D) - L_n(\overline{\gamma}, D) = \left( L_n(\hat{g}, \hat{D}_g) - L_n(\overline{\gamma}, \hat{D}_g) \right) + \left( L_n(\bar{g}, D) - L_n(\hat{g}, \hat{D}_g) \right)
\]

\[
+ \left( L_n(\overline{\gamma}, D) - L_n(\overline{\gamma}, D) \right).
\]

Note that \( L_n(\hat{g}, \hat{D}_g) - L_n(\overline{\gamma}, \hat{D}_g) \leq 0 \) due to the definition of \( \hat{g} \). To control the differences \( L_n(\bar{g}, D) - L_n(\hat{g}, D) \) and \( L_n(\overline{\gamma}, D) - L_n(\overline{\gamma}, D) \), we apply the following proposition.
Proposition 2. Grant the assumptions of Theorem 1 and fix \( \varepsilon \in (0, \varepsilon_0) \). Let \( G_\varepsilon \) and \( D_\varepsilon \) be minimal \( \varepsilon \)-nets of the sets \( G \) and \( D \) with respect to \( \| \cdot \|_{\mathcal{H}^{\alpha+1}(\mathcal{X})} \) and \( \| \cdot \|_\infty \), respectively. Then, for any \( 0 < \delta < 1 \), with probability at least \( 1 - \delta \) for any \( g \in G \) and \( D \in D \),

\[
\left| L_n(g, D) - L(g, D) - L_n(g, D^*_g) + L(g, D^*_g) \right| \lesssim C_{40a} \varepsilon^{1+\alpha+\beta} + C_{40b} \left( \frac{\log(2|G_\varepsilon||D_\varepsilon|/\delta)}{n} \right) + C_{40c} \varepsilon^{1+\alpha+\beta} + C_{40d} \frac{\log(2|G_\varepsilon||D_\varepsilon|/\delta)}{n},
\]

where the constants \( C_{40a}, C_{40b}, C_{40c} \) and \( C_{40d} \) are given in (40).

One can find the proof of Proposition 2 in Appendix A.2. Bounding \( |G_\varepsilon| \) and \( |D_\varepsilon| \) using (17) and taking \( \varepsilon = n^{-1/(\alpha+\beta+1)} \), we obtain from Proposition 2 that

\[
L_n(g, D) - L_n(g, D^*_g) \leq L(g, D) - L(g, D^*_g) + \sqrt{\frac{\log(2|G_\varepsilon||D_\varepsilon|/\delta)}{n}} \psi(n, \gamma, \eta, \delta) + \frac{\psi^2(n, \gamma, \eta, \delta)}{n}
\]  

(28)

with probability at least \( 1 - \delta \), where we have defined

\[ \psi^2(n, \gamma, \eta, \delta) = C_{29a} \gamma + C_{29b} \log n + C_{29c} \log (1/\delta) + C_{29d}, \]

with

\[ C_{29a} = \frac{C_{40b}^2 + C_{40c} + C_{40d}}{\alpha + \beta + 1} + \log (B_\varepsilon \lor B_D), \]  

(29a)

\[ C_{29b} = C_{40b}^2 + C_{40c} + C_{40d}, \]  

(29b)

\[ C_{29c} = C_{40a} + \frac{C_{40c}}{n} + (C_{40b}^2 + C_{40c} + C_{40d}) \log (2A_\varepsilon A_D). \]  

(29c)

Let us also introduce, for \( g \in G \), \( \overline{D}_g \in \arg\max_{D \in D} L(g, D) \). Note that, in general, \( \overline{D}_g \neq D^*_g \).

Then, using (28), we have

\[
L_n(\hat{g}, D^*_g) - L_n(\hat{g}, \overline{D}_g) \leq L_n(\hat{g}, D^*_g) - L_n(\hat{g}, \overline{D}_g) + \frac{\sqrt{\log(2|G_\varepsilon||D_\varepsilon|/\delta)}}{n} \psi(n, \gamma, \eta, \delta) + \frac{\psi^2(n, \gamma, \eta, \delta)}{n}
\]  

(30)

with probability at least \( 1 - \delta \). Finally, for \( L_n(\overline{g}, \overline{D}_g) - L_n(\overline{g}, D^*_g) \), with probability at least \( 1 - \delta \), it holds that

\[
L_n(\overline{g}, \overline{D}_g) - L_n(\overline{g}, D^*_g) \leq -\left( L(\overline{g}, D^*_g) - L(\overline{g}, \overline{D}_g) \right) + \frac{\sqrt{\log(2|G_\varepsilon||D_\varepsilon|/\delta)}}{n} \psi(n, \gamma, \eta, \delta) + \frac{\psi^2(n, \gamma, \eta, \delta)}{n}.
\]
Maximizing the right-hand side of the previous inequality over $L_n(\overline{\mathcal{g}}, D_{\mathcal{g}}^*) - L_n(\overline{\mathcal{g}}, D_{\mathcal{g}}^*)$ yields

$$L_n(\overline{\mathcal{g}}, D_{\mathcal{g}}^*) - L_n(\overline{\mathcal{g}}, D_{\mathcal{g}}^*) \leq \frac{5\psi^2(n, \gamma, \eta, \delta)}{4n}$$

(31)

with probability at least $1 - \delta$. Combining (27), (30), and (31), we obtain the bound on $T_2$:

$$T_2 \leq \Delta_D + \sqrt{\frac{\Delta_D}{n}} \psi(n, \gamma, \eta, \delta) + \frac{9\psi^2(n, \gamma, \eta, \delta)}{4n}. \quad (32)$$

**Step 3.** Combining (25), (26), (32), and recalling that $F(\hat{g}) = \text{JS}(p_{\hat{g}}, p^*)$, $F(\overline{\mathcal{g}}) = \Delta_{\mathcal{g}}$, we obtain that with probability at least $1 - 4\delta$,

$$\text{JS}(p_{\hat{g}}, p^*) - \Delta_D - \Delta_{\mathcal{g}} \leq \sqrt{\frac{\Delta_{\mathcal{g}}}{n}} \phi(n, \gamma, \delta) + \sqrt{\frac{\Delta_D}{n}} \psi(n, \gamma, \eta, \delta) + \frac{\phi^2(n, \gamma, \delta) + \psi^2(n, \gamma, \eta, \delta)}{n}. \quad \square$$

**Proof of Theorem 2.** First, we describe the class of generators $\mathcal{G}$. We start with approximation of the function $g^*$ together with its derivative $\nabla g^*$ via multivariate splines. Let $\pi_K = \{A_{i_1, \ldots, i_d} : -K + 1 \leq i_j \leq K, 1 \leq j \leq d, i_j \in \mathbb{Z}, j \in \mathbb{Z}\}$ be a dyadic partition of $[-R, R]^d \supseteq Y$ into disjoint cubes

$$A_{i_1, \ldots, i_d} = \prod_{j=1}^d \left[ \frac{(i_j - 1)R}{K}, \frac{i_j R}{K} \right].$$

Theorem 7 in Oswald [1990] implies that there are $(2K)^d$ polynomials $P_1, \ldots, P_{(2K)^d}$ on $\mathbb{Y}$ each of degree less than $1 + \beta$ such that the spline $S_K(y) := \sum_{k=1}^{(2K)^d} P_k(y) \mathbb{1}(y \in A_k)$ is in $\mathcal{H}^2(\mathbb{Y}, H_K)$ for some $H_K > 0$ and it holds that

$$\|g^* - S_K\|_{\mathcal{H}^2(\mathbb{Y})} \lesssim K^{-\beta} H^*.$$

Hence by taking $K > K_0(A, H^*)$, where $K_0(A)$ is a constant depending on $A$ and $H^*$, we get

$$0.5A^{-2}I_{d \times d} \leq \nabla S_K(y)^\top \nabla S_K(y) \leq 2A^2 I_{d \times d}$$

for all $y \in \mathbb{Y}$. Similarly to the proof of Lemma 4 one can show that

$$\|p_{S_K} - p^*\|_\infty \leq L_2\|S_K - g^*\|_{\mathcal{H}^2(\mathbb{Y})} \lesssim L_2K^{-\beta} H^*.$$

By Lemma 1,

$$F(S_K) \leq \frac{\|p_{S_K} - p^*\|_{L_2(p^*)}^2}{8p_{\min}^2} \lesssim R^d p_{\min}^{-2} (L_2 H^*)^2 K^{-2\beta}.$$

Next, we show that the functions $S_K$ can be represented with neural networks with ReQU and ReLU activation functions. It is known that the functions

$$1, y_j, \ldots, y_j^{1+\beta}, (y_j - (1 - K)R/K)^{1+\beta}, \ldots, (y_j - (K - 1)R/K)^{1+\beta}$$

for all $\beta$ of degree less than $1$
form a basis in the space $S_{1+\beta,Y_j}$ of univariate splines of degree less than $1 + \beta$. Here $y_j$ stands for the $j$-th component of $y$. Then the space $S_{1+\beta} = \bigotimes_{j=1}^d S_{1+\beta,Y_j}$ of tensor product polynomial splines contains $S_K$. We construct a neural network to represent $S_K$ in the following way. On the first layer, we use ReLU activation function and $(2K)^d$ nodes to get

$$(y_j - (1 - K)R/K)_+, \ldots, (y_j - (K - 1)R/K)_+, \quad j = 1, \ldots, d.$$ 

After that, we apply [Li et al., 2019, Theorem 3.1] to ensure that there exists a neural network with $d[\log_2(1 + \beta)] + d$ layers, $O \left( \left( \frac{1 + [\beta]}{d} + d \right) \right)$ ReQU activation functions and $O \left( \left( \frac{1 + [\beta]}{d} + d \right) \right)$ non-zero weights (in both cases the hidden constant behind $O$ does not depend on $\beta$ and $d$) which approximates polynomial $\prod_{j=1}^d (y_j - i_j/R)_{+}^{k_j}$ with no error. Using the fact that $\left( \frac{1 + [\beta]}{d} + d \right) \leq (\epsilon(1 + [\beta]) + d)/(1 + [\beta])^{1+\beta}$, we conclude that there exists a neural network with $d[\log_2(1 + \beta)] + d + 1$ layers and $O ((2K)^d(\epsilon(1 + [\beta]) + d)/(1 + [\beta]))^{1+\beta}$ non-zero weights such that the spline $S_K$ can be approximated with no error. Thus, the splines $S_K$ form a subclass $G$ of the class of neural networks with $d[\log_2(1 + \beta)] + d + 1$ layers and $O ((2K)^d(\epsilon(1 + [\beta]) + d)/(1 + [\beta]))^{1+\beta}$ non-zero weights. The covering number of $G$, by the construction, is equal to the covering number in the space of splines, i.e.

$$\mathcal{N}(G, \| \cdot \|_{H^{\beta+1}(Y)}, \varepsilon) \leq A_G \left( \frac{B_G}{\varepsilon} \right)^{(2K)^d(\epsilon(1 + [\beta]) + d)/(1 + [\beta])^{1+\beta}}$$

for some proper constants $A_G$ and $B_G$.

Finally, we describe the class of discriminators. Let us apply approximation and covering number results for feed-forward ReLU neural networks from Schmidt-Hieber [2020] (Theorem 5 and Lemma 5, respectively). If $D$ is a class of neural networks with ReLU activation functions such that they have

- depth at most $L = 8 + (5 + C_{\beta,d,H_D})(1 + [\log_2(d \lor \beta)])$, where

$$C_{\beta,d,H_D} = \log_2(1 + 2H_D) + \log_2(1 + d^2 + \beta^2) + d \log_2 6 + (1 + \beta/d) \log_2 N,$$

- no more than $s = 141(d + \beta + 1)^{3+d}(C_{\beta,d,H_D} + 6)N$ non-zero weights,

- $d$ neurons on the first layer, 1 neuron on the last layer, and $6(d + [\beta])N$ neurons on other layers

then

$$\Delta_D \lesssim \max_{g \in G} \min_{D \in D} \| D_g - D \|_{L_\infty} \lesssim H_D^2 \beta^\beta N^{-2\beta/d},$$

and

$$\log \mathcal{N}(D, \| \cdot \|_{L_\infty}, \varepsilon) \lesssim (s + 1) \left( \log \frac{2(L + 1)(d + 1)}{\varepsilon} + L \log(6dN + 6[\beta]N + 1) \right).$$

Then Theorem 1 implies that with probability at least $1 - \delta$

$$JS(p_\hat{g}, p^*) \lesssim K^{-2\beta} + \frac{(2K)^d(\epsilon(1 + [\beta]) + d)/(1 + [\beta])^{1+\beta}}{n} \log(K/\delta)$$

$$+ N^{-2\beta/d} + \frac{(d + \beta + 1)^{3+d}N \log((d + \beta)N/\delta)}{n}.$$
Choosing
\[ K \asymp \left( \frac{\log(n/\delta)}{n} \right)^{-1/(2\beta+d)} \quad \text{and} \quad N \asymp \left( \frac{\log(n/\delta)}{n} \right)^{-d/(2\beta+d)}, \]
we obtain that, with probability at least \( 1 - \delta \),
\[ \text{JS}(\hat{p}_g, p^*) \lesssim \left( \frac{\log(n/\delta)}{n} \right)^{2\beta/(2\beta+d)}. \]

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, Proceedings of Machine Learning Research, volume 70, pages 214–223. PMLR, 2017.

Hayk Asatryan, Hanno Gottschalk, Marieke Lippert, and Matthias Rottmann. A convenient infinite dimensional framework for generative adversarial learning, 2020.

Denis Belomestny and Vladimir Spokoiny. Spatial aggregation of local likelihood estimates with applications to classification. Ann. Statist., 35(5):2287–2311, 2007. ISSN 0090-5364. URL https://doi.org/10.1214/009053607000000271.

Gérard Biau, Benoît Cadre, Maxime Sangnier, and Ugo Tanielian. Some theoretical properties of GANs. Annals of Statistics, 48(3):1539–1566, 2020a.

Gérard Biau, Maxime Sangnier, and Ugo Tanielian. Some theoretical insights into Wasserstein GANs. arXiv preprint arXiv:2006.02682, 2020b.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration Inequalities, pages 208–240. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-28650-9. URL https://doi.org/10.1007/978-3-540-28650-9_9.

Cristina Butucea, Jean-François Delmas, Anne Dutfoy, and Richard Fischer. Optimal exponential bounds for aggregation of estimators for the Kullback-Leibler loss. Electron. J. Stat., 11(1): 2258–2294, 2017. ISSN 1935-7524. URL https://doi.org/10.1214/17-EJS1269.

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

Soheil Feizi, Farzan Farnia, Tony Ginart, and David Tse. Understanding GANs in the LQG setting: Formulation, generalization and stability. IEEE Journal on Selected Areas in Information Theory, 2020.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning Generative Models with Sinkhorn Divergences. In International Conference on Artificial Intelligence and Statistics, pages 1608–1617, 2018.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 1574–1583. PMLR, 2019.
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.

Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of GANs. arXiv preprint arXiv:1705.07215, 2017.

Bo Li, Shanshan Tang, and Haijun Yu. Better approximations of high dimensional smooth functions by deep neural networks with rectified power units. Communications in Computational Physics, 27(2):379–411, 2019. ISSN 1991-7120. URL http://global-sci.org/intro/article_detail/cicp/13451.html.

Tengyuan Liang. On how well generative adversarial networks learn densities: Nonparametric and parametric results. arXiv preprint arXiv:1811.03179, 2018.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 907–915, 2019.

Giulia Luise, Massimiliano Pontil, and Carlo Ciliberto. Generalization properties of optimal transport GANs with latent distribution learning. arXiv preprint arXiv:2007.14641, 2020.

Daniel McDonald. Minimax density estimation for growing dimension. In Artificial Intelligence and Statistics, pages 194–203, 2017.

Vaishnavh Nagarajan and J Zico Kolter. Gradient descent GAN optimization is locally stable. In Advances in neural information processing systems, pages 5585–5595, 2017.

Weili Nie and Ankit B Patel. Towards a better understanding and regularization of GAN training dynamics. In Uncertainty in Artificial Intelligence, pages 281–291. PMLR, 2020.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In Advances in neural information processing systems, pages 271–279, 2016.

Peter Oswald. On the degree of nonlinear spline approximation in besov-sobolev spaces. Journal of approximation theory, 61(2):131–157, 1990.

David Pfau and Oriol Vinyals. Connecting generative adversarial networks and actor-critic methods. arXiv preprint arXiv:1610.01945, 2016.

Jörg Polzehl and Vladimir Spokoiny. Propagation-separation approach for local likelihood estimation. Probab. Theory Related Fields, 135(3):335–362, 2006. ISSN 0178-8051. URL https://doi.org/10.1007/s00440-005-0464-1.

Philippe Rigollet. Kullback-Leibler aggregation and misspecified generalized linear models. Ann. Statist., 40(2):639–665, 2012. ISSN 0090-5364. URL https://doi.org/10.1214/11-AOS961.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. Ann. Statist., 48(4):1875–1897, 08 2020. . URL https://doi.org/10.1214/19-AOS1875.

Nicolas Schreuder, Victor-Emmanuel Brunel, and Arnak Dalalyan. Statistical guarantees for generative models without domination, 2020.

Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabás Póczos.
Nonparametric density estimation with adversarial losses. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pages 10246–10257. Curran Associates Inc., 2018.
Alexandre B Tsybakov. Introduction to nonparametric estimation. Springer Science & Business Media, 2008.
A. Uppal, S. Singh, and B. Poczos. Nonparametric density estimation: Convergence rates for GANs under Besov IPM losses. In Advances in Neural Information Processing Systems 32, pages 9089–9100. Curran Associates, Inc., 2019.
Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. arXiv preprint arXiv:1609.03126, 2016.

Appendix A: Technical results

Recall that $L_n(g,D)$ is an empirical counterpart of

$$L(g,D) = \frac{1}{2} \int \log D(x)p^*(x)dx + \frac{1}{2} \int \log(1 - D(z))p_g(z)dz,$$

and, for any $g \in G$ any $D \in D$, it holds that

$$L(g,D_g^*) - L(g,D) = \int K(D_g^*(x),D(x)) \frac{p_g(x) + p^*(x)}{2} dx,$$

where for any $u,v \in (0,1)$

$$K(u,v) = u \log \frac{u}{v} + (1 - u) \log \frac{1 - u}{1 - v}.$$

We also introduce the quantity

$$\Delta_n(g,D) = L_n(g,D) - L(g,D).$$

A.1. Proof of Proposition 1

**Proposition 1.** Assume $A\phi$, $AG$, $AD$, and $Ap^*$. Fix $\varepsilon \in (0,1)$ and let $G_\varepsilon$ be a minimal $\varepsilon$-net of the set $G$ with respect to $\| \cdot \|_{\mathcal{H}^{\beta+1}(Y)}$. Then for any $\delta \in (0,1/|G_\varepsilon|)$, with probability at least $1 - \delta|G_\varepsilon|$ for any $g \in G$,\n
$$|\Delta_n(g,D_g^*)| \leq C_{36\alpha} \varepsilon^{\beta+1} + C_{36b} \varepsilon^{\beta+1} \sqrt{\frac{\log(2/\delta)}{n}} + C_{36c} \sqrt{\frac{(F(g) \wedge F_n(g)) \log(2/\delta)}{n}}$$

$$+ C_{36d} \frac{\log(2/\delta)}{n},$$

where the constants are given by
Due to Lemma 3, we get

\[
L_1 = A^{d+|\beta \wedge 1|} \left( H_\delta + \Phi \delta^d d^{2d+|\beta \wedge 1|} H_\delta \right),
\]

\[
L_2 = \Phi d^{2d+|\beta \wedge 1|} A^{d+1} (1 + H_\delta A \sqrt d) + d H_\delta A^{d+1}.
\]

Here \(p_{\text{min}}\) and \(p_{\text{max}}\) are defined in Lemma 3.

**Proof.** For arbitrary \(g \in \mathcal{G}\) let \(g_\varepsilon\) be an element of \(\mathcal{G}_\varepsilon\), such that \(\|g - g_\varepsilon\|_{H^{\beta+1}(\mathcal{Y})} \leq \varepsilon\). Using Lemma 4, Lemma 6, and Lemma 9,

\[
|\Delta_n(g, D_{g_\varepsilon}^*)| \leq |\Delta_n(g, D_g^*) - \Delta_n(g_\varepsilon, D_g^*)| + |\Delta_n(g_\varepsilon, D_g^*)|
\]

\[
\leq \frac{4\|D_g^* - D_g^*\|_\infty}{(1 - D_{\text{max}}) \wedge D_{\text{min}}} + 4L_1 \|g - g_\varepsilon\|_{\beta+1} \frac{\|g - g_\varepsilon\|_{\beta+1}}{(1 - D_{\text{max}}) p_{\text{min}}} + |\Delta_n(g_\varepsilon, D_g^*)|
\]

\[
\leq \frac{4(L_2 + L_1) \varepsilon^{\beta \wedge 1}}{p_{\text{min}}((1 - D_{\text{max}}) \wedge D_{\text{min}})} + |\Delta_n(g_\varepsilon, D_g^*)|.
\]

Due to Lemma 8 and the union bound, it holds with probability at least \(1 - \delta|\mathcal{G}_\varepsilon|\) that

\[
|\Delta_n(g_\varepsilon, D_g^*)| \leq \sqrt{\frac{8 \log^2 (1 + p_{\text{max}}/p_{\text{min}}) \log(2/\delta)}{np_{\text{min}}}} \left( \|p^* - p_\varepsilon\|_{L_2(p^*)} + \|p^* - p_\varepsilon\|_{L_2(p_\varepsilon)} \right) + C_D \log(2/\delta)
\]

(37)

with the constant \(C_D\) defined in (47). Using Minkowski’s inequality and Lemma 4,

\[
\|p^* - p_\varepsilon\|_{L_2(p^*)} \leq \|p^* - p_g\|_{L_2(p^*)} + \|p_g - p_\varepsilon\|_{\infty} \leq \|p^* - p_g\|_{L_2(p^*)} + L_2 \varepsilon^{\beta \wedge 1},
\]

and, similarly,

\[
\|p^* - p_\varepsilon\|_{L_2(p_\varepsilon)} \leq \|p^* - p_g\|_{L_2(p_\varepsilon)} + \|p_g \circ g - p^* \circ g_\varepsilon\|_{L_2(\phi)} + \|p_g \circ g - p_\varepsilon \circ g_\varepsilon\|_{L_2(\phi)} \leq \|p^* - p_g\|_{L_2(p_\varepsilon)} + (2L_1 + L_2) \varepsilon^{\beta \wedge 1}.
\]

Applying Lemma 10, we get

\[
\|p^* - p_g\|_{L_2(p^*)} \leq p_{\text{max}} \sqrt{F(g)} , \quad \|p^* - p_g\|_{L_2(p_\varepsilon)} \leq p_{\text{max}} \sqrt{F(g)}.
\]

Hence, from (37) with the union bound, it follows that with probability at least \(1 - |\mathcal{G}_\varepsilon|\delta\),

\[
|\Delta_n(g, D_g^*)| \leq \frac{4(L_2 + L_1) \varepsilon^{\beta \wedge 1}}{p_{\text{min}}((1 - D_{\text{max}}) \wedge D_{\text{min}})} + \frac{4(L_2 + L_1) \log (1 + p_{\text{max}}/p_{\text{min}}) \varepsilon^{\beta \wedge 1}}{p_{\text{min}}} \sqrt{\frac{2 \log(2/\delta)}{n}} + \frac{16p_{\text{max}} \log (1 + p_{\text{max}}/p_{\text{min}})}{p_{\text{min}}} \sqrt{\frac{F(g) \log(2/\delta)}{n}} + \frac{C_D \log(2/\delta)}{n}.
\]

(38)
Let us denote \( a = \sqrt{F(g)} \), \( b = \sqrt{F_n(g)} \),

\[
\begin{align*}
C_1(n) &= \frac{16p_{\text{max}} \log \left( 1 + \frac{p_{\text{max}}}{p_{\text{min}}} \right) \sqrt{\log(2/\delta)}}{p_{\text{min}}} \\
C_2(n) &= \frac{4(L_2 + L_1)\varepsilon^{\beta_1}}{p_{\text{min}} \left( (1 - D_{\text{max}}) \land D_{\text{min}} \right)} + \frac{C_D \log(2/\delta)}{n} \\
&\quad + \frac{4(L_2 + L_1) \log \left( 1 + \frac{p_{\text{max}}}{p_{\text{min}}} \right) \varepsilon^{\beta_1}}{p_{\text{min}}} \sqrt{\frac{2\log(2/\delta)}{n}},
\end{align*}
\]

we rewrite (38) as

\[
|a^2 - b^2| \leq C_1(n)a + C_2(n).
\]  

(39)

Solving the quadratic inequality yields

\[
a \leq \frac{C_1(n) + \sqrt{C_1(n)^2 + 4(b^2 + C_2(n))}}{2} \leq b + C_1(n) + \sqrt{C_2(n)}.
\]

Hence, (39) with Cauchy-Schwartz inequality imply

\[
|a^2 - b^2| \leq C_1(n)(b \land a) + \frac{3}{2}C_2(n) + \frac{3}{2}C_1(n)^2.
\]

Substituting for \( a, b, C_1(n) \) and \( C_2(n) \) yields the statement of lemma.

\[\Box\]

**A.2. Proof of Proposition 2**

**Proposition 2.** Assume \( \Phi, \, A\Phi^*, \, AG, \) and \( A\Phi^* \). Fix \( \varepsilon \in (0, \varepsilon_0) \). Let \( \mathcal{G}_\varepsilon \) and \( \mathcal{D}_\varepsilon \) be two finite \( \varepsilon \)-nets of the sets \( \mathcal{G} \) and \( \mathcal{D} \) with respect to \( \| \cdot \|_{H^{\beta_1 + 1}(\mathcal{G})} \) and \( \| \cdot \|_{\infty}, \) respectively. Then, for any \( \delta \in (0, 1/(|\mathcal{G}_\varepsilon| |\mathcal{D}_\varepsilon|)) \), with probability at least \( 1 - \delta |\mathcal{G}_\varepsilon| |\mathcal{D}_\varepsilon| \) for any \( g \in \mathcal{G} \) and \( D \in \mathcal{D} \),

\[
|\Delta_n(g, D) - \Delta_n(g, D^*)| \lesssim C_{40a} \varepsilon^{1 + \alpha \wedge \beta} + C_{40b} \sqrt{\frac{(L(g, D^*) - L(g, D)) \log(2/\delta)}{n}} \\
+ C_{40c} \varepsilon^{1 + \alpha \wedge \beta} \sqrt{\frac{\log(2/\delta)}{n}} + C_{40d} \frac{\log(2/\delta)}{n},
\]

where the constants are given by

\[
\begin{align*}
C_{40a} &= \frac{L_1 + L_2 + p_{\text{min}} + dH_p p_{\text{min}}}{((1 - D_{\text{max}}) \land D_{\text{min}}) p_{\text{min}}} \quad \text{(40a)} \\
C_{40b} &= \frac{1}{\sqrt{p_{\text{min}} ((D_{\text{min}} \land (1 - D_{\text{max}})))}} \quad \text{(40b)} \\
C_{40c} &= \frac{p_{\text{min}} + dH_p p_{\text{min}} + L_2}{((1 - D_{\text{max}}) \land D_{\text{min}}) p_{\text{min}}} \quad \text{(40c)} \\
C_{40d} &= \log \left( \frac{D_{\text{max}} \lor (1 - D_{\text{min}})}{D_{\text{min}} \lor (1 - D_{\text{max}})} \right) \quad \text{(40d)} \\
L_1 &= A^{d + (\beta_1)} \left( H_\phi + \Phi A^{2d} d^{2+d/2} H_\phi \right) \quad \text{(40e)} \\
L_2 &= \Phi d^{2+d/2} A^{3d} (1 + H_G A^{d/2}) + dH_\phi A^{d+1} \quad \text{(40f)}
\end{align*}
\]

Here \( p_{\text{min}} \) and \( p_{\text{max}} \) are defined in Lemma 3.
Proof. Fix any $g \in \mathcal{G}$ and $D \in \mathcal{D}$. Let $g_\varepsilon$ and $D_\varepsilon$ be the elements of $\mathcal{G}_\varepsilon$ and $\mathcal{D}_\varepsilon$, respectively, such that $\|g_\varepsilon - g\|_{H^\beta+1} \leq \varepsilon$ and $\|D_\varepsilon - D\|_{\infty} \leq \varepsilon$. Due to the triangle inequality,

$$\left| \Delta_n(g, D) - \Delta_n(g, D_\varepsilon) \right| \leq \left| \Delta_n(g_\varepsilon, D_\varepsilon) - \Delta_n(g_\varepsilon, D_\varepsilon^*) \right| + \left| \Delta_n(g_\varepsilon, D_\varepsilon^*) - \Delta_n(g_\varepsilon, D_\varepsilon) \right|$$

Lemma 5 implies that

$$\left| \Delta_n(g_\varepsilon, D_\varepsilon) - \Delta_n(g_\varepsilon, D) \right| \leq \frac{4 \|D - D_\varepsilon\|_{\infty}}{(1 - D_{\max}) \wedge D_{\min}} + \frac{2dH_D \|g_\varepsilon - g\|_{H^\beta+1}}{1 - D_{\max}}.$$  

Lemma 6 and Lemma 9 yield

$$\left| \Delta_n(g_\varepsilon, D_\varepsilon^*) - \Delta_n(g_\varepsilon, D_\varepsilon) \right| \leq \frac{4 \|D_\varepsilon^* - D_\varepsilon\|_{\infty}}{(1 - D_{\max}) \wedge D_{\min}} + \frac{4L_1 \|g - g_\varepsilon\|_{H^\beta+1}}{(1 - D_{\max})p_{\min}}$$

$$\leq \frac{4 (L_1 + L_2) \|g - g_\varepsilon\|_{H^\beta+1}}{(1 - D_{\max}) \wedge D_{\min}) p_{\min}}.$$  

Furthermore, due to Lemma 7 and the union bound, with probability at least $1 - |\mathcal{G}_\varepsilon|\mathcal{D}_\varepsilon\delta$, it holds that

$$\left| \Delta_n(g_\varepsilon, D_\varepsilon) - \Delta_n(g_\varepsilon, D_\varepsilon^*) \right| \leq \sqrt{2 \|D_\varepsilon^* - D_\varepsilon\|_{L_2(p^*)}^2 \log(2/\delta)} \frac{2 \log(D_{\max}/D_{\min}) \log(2/\delta)}{3n} + \frac{2 \log((1 - D_{\min})/(1 - D_{\max}) \log(2/\delta))}{3n}$$

simultaneously for all $g_\varepsilon \in \mathcal{G}_\varepsilon$, $D_\varepsilon \in \mathcal{D}_\varepsilon$. Next, we have

$$\|D_\varepsilon - D_\varepsilon^*\|_{L_2(p^*)} \leq \|D - D_\varepsilon\|_{L_2(p^*)} + \|D - D_\varepsilon\|_{L_2(p^*)} + \|D_\varepsilon^* - D_\varepsilon^*\|_{L_2(p^*)},$$

where, by Lemma 4,

$$\|D_\varepsilon^* - D_\varepsilon^*\|_{L_2(p^*)} \leq \left\| \frac{p^*}{p^* + p}\right\|_{\infty} \leq \frac{L_2 \|g - g_\varepsilon\|_{H^\beta+1}}{p_{\min}}.$$  

Hence,

$$\|D_\varepsilon^* - D_\varepsilon^*\|_{L_2(p^*)} \leq \|D - D_\varepsilon^*\|_{L_2(p^*)} + \varepsilon + \frac{L_2 \varepsilon \beta \Lambda \varepsilon}{p_{\min}}.$$  

Similarly,

$$\|D_\varepsilon - D_\varepsilon^*\|_{L_2(p_\varepsilon)} \leq \|D - D_\varepsilon\|_{L_2(p_\varepsilon)} + \|D - D_\varepsilon\|_{L_2(p_\varepsilon)} + \|D_\varepsilon^* - D_\varepsilon^*\|_{L_2(p_\varepsilon)}$$

$$\leq \|D - D_\varepsilon\|_{L_2(p_\varepsilon)} + \varepsilon + \frac{L_2 \varepsilon \beta \Lambda \varepsilon}{p_{\min}}.$$  

To bound $\|D - D^*_g\|_{L_2(p_g)}$, note that

$$
\|D - D^*_g\|_{L_2(p_g)} = \|D \circ g - D^*_g \circ g\|_{L_2(\phi)} \\
\leq \|D \circ g - D^*_g \circ g\|_{L_2(\phi)} + \|D \circ g - D \circ g\|_{L_2(\phi)} + \|D^*_g \circ g - D^*_g \circ g\|_{L_2(\phi)} \\
= \|D - D^*_g\|_{L_2(p_g)} + \|D \circ g - D \circ g\|_\infty + \|D^*_g \circ g - D^*_g \circ g\|_\infty,$$

where, due to Assumption AD,

$$
\|D \circ g - D \circ g\|_\infty \leq d_{X\epsilon}^{\alpha_1}
$$

and

$$
\left|\frac{p^*(g(y))}{p^*(g(y)) + p_d(g(y))} - \frac{p^*(g_x(y))}{p^*(g_x(y)) + p_d(g_x(y))}\right| \\
\leq \sup_{y \in Y} \left|\frac{p^*(g(y))p_d(g_x(y)) + p_d(g(y))p^*(g_x(y))}{(p^*(g(y)) + p_d(g(y))(p^*(g_x(y)) + p_d(g_x(y)))}\right| \\
\leq \frac{2L_1\epsilon^{\beta_1}}{p_{\min}},
$$

where the last inequality is due to Lemma 2. Finally, note that, due to Lemma 1,

$$
\|D - D^*_g\|_{L_2(p_g)}^2 \leq \kappa_{\min} \mathbb{E}_K(D^*_g(X), D(X))
$$

and

$$
\|D - D^*_g\|_{L_2(p_g)}^2 \leq \kappa_{\min} \mathbb{E}_K(D^*_g(g(Y)), D(g(Y))).
$$

The equality

$$
\frac{1}{2} \mathbb{E}_K(D^*_g(X), D(X)) + \frac{1}{2} \mathbb{E}_K(D^*_g(g(Y)), D(g(Y))) = L(g, D^*_g) - L(g, D)
$$

complete the proof. 

\[\square\]

### A.3. Auxiliary results

**Lemma 1.** For any two functions $D_1, D_2 : X \rightarrow [D_{\min}, D_{\max}], 0 < D_{\min} \leq D_{\max} < 1$, and any $x \in X$, it holds that

$$
\kappa_{\min} (D_1(x) - D_2(x))^2 \leq K(D_1(x), D_2(x)) \leq \kappa_{\max} (D_1(x) - D_2(x))^2,
$$

where

$$
\kappa_{\min} = 8 \left(D_{\min}(1 - D_{\min}) \wedge D_{\max}(1 - D_{\max})\right),
$$

$$
\kappa_{\max} = \frac{1}{8 \left(D_{\min}(1 - D_{\min}) \wedge D_{\max}(1 - D_{\max})\right)^2}.
$$

**Proof.** Fix arbitrary $x \in X$ and denote

$$
\nu_1 = \log \frac{D_1(x)}{1 - D_1(x)}, \quad \nu_2 = \log \frac{D_2(x)}{1 - D_2(x)}.
$$
Let $\Psi(\nu) = \log(1 + e^\nu)$. Then $\Psi(\nu_1) = -\log(1 - D_1(x))$, $\Psi(\nu_2) = -\log(1 - D_2(x))$, and

$$K(D_1(x), D_2(x)) = D_1(x) \left( \log \frac{D_1(x)}{1 - D_1(x)} - \log \frac{D_2(x)}{1 - D_2(x)} \right) + \log \frac{1 - D_1(x)}{1 - D_2(x)}$$

$$= \frac{e^{\nu_1}}{1 + e^{\nu_1}} (\nu_1 - \nu_2) - \Psi(\nu_1) + \Psi(\nu_2) = \Psi'(\nu_1) (\nu_1 - \nu_2) - \Psi(\nu_1) + \Psi(\nu_2).$$

The Lagrange theorem implies that, for some $\xi$ between $\nu_1$ and $\nu_2$,

$$K(D_1(x), D_2(x)) = \frac{\Psi''(\xi)}{2} (\nu_1 - \nu_2)^2.$$

On the other hand, there exists $\zeta \in (\nu_1, \nu_2)$ (or $\zeta \in (\nu_2, \nu_1)$ if $\nu_2 < \nu_1$) such that

$$D_1(x) - D_2(x) = \Psi'(\nu_1) - \Psi'(\nu_2) = \Psi''(\zeta) (\nu_1 - \nu_2).$$

Thus,

$$K(D_1(x), D_2(x)) = \frac{\Psi''(\xi)}{2\Psi''(\zeta)^2} (D_1(x) - D_2(x))^2. \quad (41)$$

For any $\nu$ from the interval $(\nu_1, \nu_2)$ (or $(\nu_2, \nu_1)$), it holds that

$$\Psi''(\nu) = \frac{e^\nu}{(1 + e^\nu)^2} = \frac{e^\nu}{1 + e^\nu} \left( 1 - \frac{e^\nu}{1 + e^\nu} \right) \in \left[ D_{\min}(1 - D_{\min}) \land D_{\max}(1 - D_{\max}), \frac{1}{4} \right],$$

where we used the fact that the function $f(x) = x(1 - x)$ is concave and achieves its minimum at one of the boundary points of the segment. Substituting $\Psi''(\xi)$ and $\Psi''(\zeta)$ in (41) by their upper and lower bounds, we obtain

$$x_{\min} (D_1(x) - D_2(x))^2 \leq K(D_1(x), D_2(x)) \leq x_{\max} (D_1(x) - D_2(x))^2,$$

where

$$x_{\min} = 8 (D_{\min}(1 - D_{\min}) \land D_{\max}(1 - D_{\max})), \quad x_{\max} = \frac{1}{8 (D_{\min}(1 - D_{\min}) \land D_{\max}(1 - D_{\max}))^2}.$$

\[\square\]

**Lemma 2.** Under Assumption $AG$ and Assumption $A\phi$, for any $x_1, x_2 \in \mathcal{X}$ and $g \in \mathcal{G}$ it holds that

$$|p_g(x_1) - p_g(x_2)| \leq \mathcal{L}_1 \|x_1 - x_2\|^{\beta + 1}.$$

with

$$\mathcal{L}_1 = A^{d + (\beta + 1)} \left( H + \Phi A^{2d} d^{d + 2} H_{\phi} \right). \quad (42)$$

Moreover, under Assumption $A\phi^*$ and Assumption $A\phi$, it holds that

$$|p^*(x_1) - p^*(x_2)| \leq \mathcal{L}_1 \|x_1 - x_2\|^{\beta + 1}$$

for any $x_1, x_2 \in \mathcal{X}$ with the same constant $\mathcal{L}_1$. 
Proof. Let \( y_1 = g^{-1}(x_1), y_2 = g^{-1}(x_2) \). Then \( \|y_1 - y_2\| \leq A\|x_1 - x_2\| \) and
\[
|p_g(x_1) - p_g(x_2)| = |\det[\nabla g(y_1)]^{-1}\phi(y_1) - \det[\nabla g(y_2)]^{-1}\phi(y_2)|
\leq |\det[\nabla g(y_1)]^{-1}(\phi(y_1) - \phi(y_2))| + |\det[\nabla g(y_1)]^{-1} - \det[\nabla g(y_2)]^{-1}| \phi(y_2)
\leq A^d |\phi(y_1) - \phi(y_2)| + \Phi |\det[\nabla g(y_1)]^{-1} - \det[\nabla g(y_2)]^{-1}|.
\]
The last inequality follows from the fact that, due to (AG),
\[
\det[\nabla g(y_1)]^{-1} = \sqrt{\det([\nabla g(y_1)]^{-1}[\nabla g(y_1)]^{-1})} \leq \sqrt{\det(A^2 I_{d \times d})} = A^d.
\]
Since \( \phi \in \mathcal{H}^{\beta \Lambda^1}(X, H_{\phi}) \), we have
\[
|\phi(y_1) - \phi(y_2)| \leq H_{\phi}\|y_1 - y_2\|^{\beta \Lambda^1} \leq H_{\phi} A^{\beta \Lambda^1}\|x_1 - x_2\|^{\beta \Lambda^1}.
\]
It remains to bound \( |\det[\nabla g(y_1)]^{-1} - \det[\nabla g(y_2)]^{-1}| \). Note that
\[
|\det[\nabla g(y_1)]^{-1} - \det[\nabla g(y_2)]^{-1}|
\leq |\det[\nabla g(y_1)]^{-2} |\det \nabla g(y_1) - \det \nabla g(y_2)|
\leq A^{2d} |\det \nabla g(y_1) - \det \nabla g(y_2)|.
\]
Next, since for any \( d \times d \) matrices \( A \) and \( B \) it holds that
\[
|\det A - \det B| \leq \|A - B\|_F \frac{\|A\|_F^d - \|B\|_F^d}{\|A\|_F - \|B\|_F} \leq d \max\{\|A\|_F^d, \|B\|_F^d\} \|A - B\|_F
\]
and for all \( y \in \mathcal{Y} \)
\[
\|\nabla g(y)\|_F^2 = \text{Tr} (\nabla g(y)^T \nabla g(y)) \leq A^2 d,
\]
we obtain
\[
|\det \nabla g(y_1) - \det \nabla g(y_2)| \leq A^d d^{1+d/2} \|\nabla g(y_1) - \nabla g(y_2)\|_F
\leq A^d d^{2+d/2} H_G \|y_1 - y_2\|^\beta \Lambda^1
\leq A^{d+(\beta \Lambda^1)} d^{2+d/2} H_G \|x_1 - x_2\|^{\beta \Lambda^1}.
\]
Hence, for all \( x_1, x_2 \in X \),
\[
|p_g(x_1) - p_g(x_2)| \leq A^{d+(\beta \Lambda^1)} \left( H_{\phi} + \Phi A^2 d^{2+d/2} H_G \right) \|x_1 - x_2\|^{\beta \Lambda^1},
\]
and the first claim of the lemma follows. The proof of the inequality
\[
|p^*(x_1) - p^*(x_2)| \leq A^{d+(\beta \Lambda^1)} \left( H_{\phi} + \Phi A^2 d^{2+d/2} H_G \right) \|x_1 - x_2\|^{\beta \Lambda^1}
\]
is absolutely similar. \( \square \)

**Lemma 3.** Assume AG and A\( \phi \). Then, for any \( x \in X \) and any \( g \in \mathcal{G} \), it holds that
\[
p_{\text{min}} \leq p_g(x) \leq p_{\text{max}},
\]
where
\[ p_{\min} = (\Phi A^d)^{-1}, \quad p_{\max} = \Phi A^d. \] (43)

Similarly, under Assumption \( Ap^* \) and Assumption \( A\phi \), the inequality
\[ p_{\min} \leq p^*(x) \leq p_{\max} \]
holds for all \( x \in X \).

**Proof.** Fix any \( g \in G \) and \( x \in X \). Assumption \( AG \) implies that
\[ A^{-d} = \sqrt{\det(A^{-2I_{d \times d}})} \leq \sqrt{\det(\nabla g(g^{-1}(x))\nabla g(g^{-1}(x)))} \leq \sqrt{\det(A^{2I_{d \times d}})} = A^d. \]
Then the equality \( p_g(x) = |\det \nabla g(g^{-1}(x))| \phi(g^{-1}(x)) \) and the inequality \( \Phi^{-1} \leq \phi(g^{-1}(x)) \leq \Phi \) yield
\[ (\Phi A^d)^{-1} \leq p_g(x) \leq \Phi A^d. \]
Similarly, Assumption \( Ap^* \) implies \( \Phi^{-1} A^{-d} \leq p^*(x) \leq A^d \) for all \( x \in X \). \( \square \)

**Lemma 4.** Under Assumption \( AG \) and Assumption \( A\phi \), for any \( f, g \in G \) it holds that
\[ \|p_g - p_f\|_\infty \leq L_2\|f - g\|_{\mathbb{M}^1}^{\beta A^d}. \]
with
\[ L_2 = \Phi d^{2+4/2}A^{d+1}(1 + H_g A\sqrt{d}) + dH_\phi A^{d+1}. \] (44)

**Proof.** Due to assumption \( AG \), \( \nabla g(y) \) is non-degenerate for all \( g \in G \) and all \( y \in Y \). Thus, \( \det \nabla g \) does not change its sign, and without loss of generality, we can assume that \( \det \nabla f(f^{-1}(x)) \) and \( \det \nabla g(g^{-1}(x)) \) are positive for all \( x \in X \). We have
\[ |p_f(x) - p_g(x)| = |(\det \nabla f(f^{-1}(x)))^{-1}\phi(f^{-1}(x)) - (\det \nabla g(g^{-1}(x)))^{-1}\phi(g^{-1}(x))| \]
\[ \leq |(\det \nabla f(f^{-1}(x)))^{-1}\phi(f^{-1}(x)) - (\det \nabla g(g^{-1}(x)))^{-1}\phi(f^{-1}(x))| \]
\[ + |(\det \nabla g(g^{-1}(x)))^{-1}\phi(f^{-1}(x)) - (\det \nabla g(g^{-1}(x)))^{-1}\phi(f^{-1}(x))| \]
\[ \leq \Phi|\det(\nabla f(f^{-1}(x)))^{-1} - \det(\nabla g(g^{-1}(x)))^{-1}| + H_\phi A^d\|f^{-1}(x) - g^{-1}(x)\|^{\beta A^d}. \]

Here we used the fact that, due to Assumption \( AG \),
\[ \det(\nabla g(g^{-1}(x)))^{-1} = \sqrt{\det(\nabla g(g^{-1}(x)))^{-1}} = \sqrt{\det(A^{2I_{d \times d}})} = A^d. \]

Let \( u = f^{-1}(x) \in Y \). Then \( x = f(u) \) and
\[ \|f^{-1}(x) - g^{-1}(x)\| = \|f^{-1}(f(u)) - g^{-1}(f(u))\| = \|g^{-1}(g(u)) - g^{-1}(f(u))\| \]
\[ \leq \Lambda\|f(u) - g(u)\| \leq \Lambda \sqrt{d}\|f - g\|_\infty \]
by the mean value theorem for vector valued functions. Furthermore, we have
\[ |\det(\nabla f(f^{-1}(x)))^{-1} - \det(\nabla g(g^{-1}(x)))^{-1}| \]
\[ \leq \min \{ \det(\nabla f(f^{-1}(x))), \det(\nabla g(g^{-1}(x)))\}^{-2} |\det(\nabla f(f^{-1}(x)) - \det(\nabla g(g^{-1}(x))| \]
\[ \leq \Lambda^d |\det(\nabla f(f^{-1}(x)) - \det(\nabla g(g^{-1}(x))|} \]
Next, since for any \(d \times d\) matrices \(A\) and \(B\) it holds that
\[
|\det A - \det B| \leq \|A - B\|_F \frac{\|A\|_F \|B\|_F}{\|A\|_F - \|B\|_F} \leq d \max\{\|A\|_F, \|B\|_F\} \|A - B\|_F
\]
and
\[
\|\nabla f(f^{-1}(x))\|_F^2 = \text{Tr} (\nabla f(f^{-1}(x))^T \nabla f(f^{-1}(x))) \leq \Lambda^2 d,
\]
\[
\|\nabla g(g^{-1}(x))\|_F^2 = \text{Tr} (\nabla g(g^{-1}(x))^T \nabla g(g^{-1}(x))) \leq \Lambda^2 d,
\]
we obtain
\[
|\det \nabla f(f^{-1}(x)) - \det \nabla g(g^{-1}(x))| \leq A^d d^{3 + d/2} \|\nabla f(f^{-1}(x)) - \nabla g(g^{-1}(x))\|_F
\]
\[
\leq A^d d^{3 + d/2} \|\nabla f(f^{-1}(x)) - \nabla g(f^{-1}(x))\|_F
\]
\[
+ A^d d^{3 + d/2} \|\nabla g(f^{-1}(x)) - \nabla g(g^{-1}(x))\|_F
\]
\[
\leq A^d d^{2 + d/2} \|f - g\|_{H^{1+\beta}(\mathcal{Y})}
\]
\[
+ A^d d^{2 + d/2} \|g_\beta\|_{f^{-1}(x) - g^{-1}(x)} \beta^1
\]
\[
\leq A^d d^{2 + d/2} (1 + H_\beta A \sqrt{d}) \|f - g\|_{H^{1+\beta}(\mathcal{Y})}.
\]

Hence
\[
\|p_f - p_g\|_\infty \leq \left( \Phi d^{2 + d/2} A^d (1 + H_\beta A \sqrt{d}) + H_\beta A^{d+\sqrt{d}} \right) \|f - g\|_{H^{1+\beta}(\mathcal{Y})}.
\]

**Lemma 5.** Let \(D_1, D_2 : X \to [D_{\min}, D_{\max}], 0 < D_{\min} \leq D_{\max} < 1\). Assume additionally that \(D_1 \in \mathcal{H}_0(X, H_D)\) or \(D_2 \in \mathcal{H}_0(X, H_D)\) for some \(\alpha > 0\). Then for any \(g_1, g_2 \in \mathcal{G}\) it holds
\[
|\Delta_n(g_1, D_1) - \Delta_n(g_2, D_2)| \leq \frac{4 \|D_1 - D_2\|_\infty}{(1 - D_{\max}) \wedge D_{\min}} + \frac{2dH_D \|g_1 - g_2\|_{2}^{\alpha \wedge 1}}{1 - D_{\max}}.
\]

**Proof.** Assume without loss of generality that \(D_2 \in \mathcal{H}_0(X, H_D)\). By the definition of \(\Delta_n(g, D)\),
\[
|\Delta_n(g_1, D_1) - \Delta_n(g_2, D_2)| \leq |L_n(g_1, D_1) - L_n(g_2, D_2)| + |L(g_1, D_1) - L(g_2, D_2)|.
\]

To estimate the first term, we note that
\[
L_n(g_1, D_1) - L_n(g_2, D_2) =: R_{1,n} + R_{2,n},
\]
where we have introduced the quantities
\[
R_{1,n} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{D_1(X_i)}{D_2(X_i)}
\]
and
\[
R_{2,n} = \frac{1}{n} \sum_{i=1}^{n} \log (1 - D_1(g_1(Y_i))) - \frac{1}{n} \sum_{i=1}^{n} \log (1 - D_2(g_2(Y_i)))
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} \log \left( \frac{1 - D_1(g_1(Y_i))}{1 - D_2(g_1(Y_i))} \right) - \frac{1}{n} \sum_{i=1}^{n} \log \left( \frac{1 - D_2(g_2(Y_i))}{1 - D_2(g_1(Y_i))} \right).
\]
Due to the Lagrange theorem,
\[ |R_{1,n}| \leq \frac{\|D_1 - D_2\|_\infty}{D_{\min}} \]
and
\[ |R_{2,n}| \leq \frac{\|D_1 - D_2\|_\infty}{1 - D_{\max}} + \frac{dH_D \|g_1 - g_2\|_\infty^{\alpha^2_1}}{1 - D_{\max}}. \]
As a result,
\[ |L_n(g_1, D_1) - L_n(g_2, D_2)| \leq \frac{2\|D_1 - D_2\|_\infty}{(1 - D_{\max}) \wedge D_{\min}} + \frac{dH_D \|g_1 - g_2\|_\infty^{\alpha^2_1}}{1 - D_{\max}}. \]
Similarly,
\[ |L(g_1, D_1) - L(g_2, D_2)| \leq \frac{2\|D_1 - D_2\|_\infty}{(1 - D_{\max}) \wedge D_{\min}} + \frac{dH_D \|g_1 - g_2\|_\infty^{\alpha^2_1}}{1 - D_{\max}}, \]
and the statement follows from (45).

**Lemma 6.** Under assumptions \(A\Phi, AG, AD,\) and \(Ap^*\), for any \(g_1, g_2 \in \mathcal{G}\) it holds
\[ |\Delta_n(g_1, D^{*^1}_{g_1}) - \Delta_n(g_2, D^{*^1}_{g_2})| \leq \frac{4\|D^{*^1}_{g_1} - D^{*^1}_{g_2}\|_\infty}{(1 - D_{\max}) \wedge D_{\min}} + \frac{4L_1\|g_1 - g_2\|_\infty^{\alpha_1^2}}{(1 - D_{\max})p_{\min}}, \]
with the constant \(L_1\) defined in (42).

**Proof.** Following the lines of Lemma 5, it is enough to bound
\[ L_n(g_1, D^{*^1}_{g_1}) - L_n(g_2, D^{*^1}_{g_2}) =: R_{1,n} + R_{2,n}, \]
where
\[ R_{1,n} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{D^{*^1}_{g_1}(X_i)}{D^{*^1}_{g_2}(X_i)} \]
and
\[ R_{2,n} = \frac{1}{n} \sum_{i=1}^{n} \log \left(1 - D^{*^1}_{g_1}(g_1(Y_i))\right) - \frac{1}{n} \sum_{i=1}^{n} \log \left(1 - D^{*^1}_{g_2}(g_1(Y_i))\right) = \frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{1 - D^{*^1}_{g_1}(g_1(Y_i))}{1 - D^{*^1}_{g_2}(g_1(Y_i))}\right) - \frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{1 - D^{*^1}_{g_2}(g_2(Y_i))}{1 - D^{*^1}_{g_2}(g_1(Y_i))}\right). \] (46)

Similarly to Lemma 5, we have
\[ |\log D^{*^1}_{g_1}(X_i) - \log D^{*^1}_{g_2}(X_i)| \leq \frac{\|D^{*^1}_{g_1} - D^{*^1}_{g_2}\|_\infty}{D_{\min}} \]
and
\[ |\log(1 - D^{*^1}_{g_1}(g_1(Y_i))) - \log(1 - D^{*^1}_{g_2}(g_1(Y_i)))| \leq \frac{\|D^{*^1}_{g_1} - D^{*^1}_{g_2}\|_\infty}{1 - D_{\max}} \]
for all \(i\) from 1 to \(n\). To bound the second term in the right-hand side of (46), note that, for any \(i \in \{1, \ldots, n\}\), it holds that
\[ \left|\log \frac{1 - D^{*^1}_{g_1}(g_1(Y_i))}{1 - D^{*^1}_{g_2}(g_2(Y_i))}\right| \leq \frac{|D^{*^1}_{g_1}(g_1(Y_i)) - D^{*^1}_{g_2}(g_2(Y_i))|}{1 - D_{\max}}. \]
and

\[ |D_{y_1}^*(g_1(Y_i)) - D_{y_2}^*(g_2(Y_i))| = \left| \frac{p^*(g_1(Y_i))}{p^*(g_1(Y_i)) + p_{y_2}(g_1(Y_i))} - \frac{p^*(g_2(Y_i))}{p^*(g_2(Y_i)) + p_{y_2}(g_2(Y_i))} \right| \]

\[ \leq \frac{p^*(g_1(Y_i))|p_{y_2}(g_1(Y_i)) - p_{y_2}(g_2(Y_i))| + p_{y_2}(g_1(Y_i))|p^*(g_1(Y_i)) - p^*(g_2(Y_i))|}{(p^*(g_1(Y_i)) + p_{y_2}(g_1(Y_i)))(p^*(g_2(Y_i)) + p_{y_2}(g_2(Y_i)))} \]

\[ \leq \frac{2L_1\|g_1 - g_2\|_\infty^{\beta + 1}}{\beta_{\min}}. \]

Lemma 7. For any functions \(D_1, D_2 : X \to [D_{\min}, D_{\max}], 0 < D_{\min} \leq D_{\max} < 1\), and any \(g \in \mathcal{G}\), with probability at least \(1 - \delta\)

\[ |\Delta_n(g, D_1) - \Delta_n(g, D_2)| \leq \sqrt{\frac{2\|D_1 - D_2\|_{L_2(p^*)}^2 \log(2/\delta)}{n D_{\min}^2}} + \frac{2 \log(D_{\max}/D_{\min}) \log(2/\delta)}{3n} \]

\[ + \sqrt{\frac{2\|D_1 - D_2\|_{L_2(p_*)}^2 \log(2/\delta)}{n(1 - D_{\max})^2}} + \frac{2 \log((1 - D_{\min})/(1 - D_{\max})) \log(2/\delta)}{3n}. \]

Proof. Denote

\[ \mathcal{E}_{1,n} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{D_1(X_i)}{D_2(X_i)} - E \left[ \frac{1}{n} \sum_{i=1}^{n} \log \frac{D_1(X_i)}{D_2(X_i)} \right] \]

\[ \mathcal{E}_{2,n} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{1 - D_1(g(Y_i))}{1 - D_2(g(Y_i))} - E \left[ \frac{1}{n} \sum_{i=1}^{n} \log \frac{1 - D_1(g(Y_i))}{1 - D_2(g(Y_i))} \right] \]

The Bernstein inequality yields for any fixed \(g \in \mathcal{G}\) and \(D \in \mathcal{D}\),

\[ |\mathcal{E}_{1,n}| \leq \sqrt{\frac{2E \log^2 \frac{D_1(X)}{D_2(X)} \log(2/\delta)}{n}} + \frac{2 \log(D_{\max}/D_{\min}) \log(2/\delta)}{3n} \]

\[ |\mathcal{E}_{2,n}| \leq \sqrt{\frac{2E \log^2 \frac{1 - D_1(g(Y))}{1 - D_2(g(Y))} \log(2/\delta)}{n}} + \frac{2 \log((1 - D_{\min})/(1 - D_{\max})) \log(2/\delta)}{3n} \]

with probability at least \(1 - \delta\). The assertion of the lemma follows from the inequalities

\[ |\log D_1(X) - \log D_2(X)| \leq \frac{|D_1(X) - D_2(X)|}{D_{\min}}, \]

\[ |\log(1 - D_1(X)) - \log(1 - D_2(X))| \leq \frac{|D_1(X) - D_2(X)|}{1 - D_{\max}} \]

Lemma 8. Under Assumptions \(Ap^*\), \(AG\), and \(A\phi\), for any fixed \(g \in \mathcal{G}\), with probability at least \(1 - \delta\)

\[ |\Delta_n(g, D_g^*)| \leq \frac{2 \log \left(1 + \frac{p_{\max}}{p_{\min}}\right)}{p_{\min}} \sqrt{\frac{2 \log(2/\delta)}{n}} \left( \|p^* - p_g\|_{L_2(p^*)} + \|p^* - p_g\|_{L_2(p_g)} \right) + \frac{C_D \log(2/\delta)}{n}. \]
where \( p_{\text{min}} \) is given in (43) and constant \( C_D \) is given by
\[
C_D = \frac{2}{3} \left( \log \left( \frac{1}{2D_{\text{min}}} \vee 2D_{\text{max}} \right) + \log \left( \frac{1}{2(1 - D_{\text{max}})} \vee 2(1 - D_{\text{min}}) \right) \right)
\]

(47)

Proof. Denote
\[
\mathcal{E}_{1,n} = \frac{1}{2n} \sum_{i=1}^{n} \log \left\{ 2D_g^*(X_i) \right\} - \mathbb{E} \left[ \frac{1}{2n} \sum_{i=1}^{n} \log \left\{ 2D_g^*(X_i) \right\} \right],
\]
\[
\mathcal{E}_{2,n} = \frac{1}{2n} \sum_{i=1}^{n} \log \left\{ 2(1 - D_g^*(g(Y_i))) \right\} - \mathbb{E} \left[ \frac{1}{2n} \sum_{i=1}^{n} \log \left\{ 2(1 - D_g^*(g(Y_i))) \right\} \right]
\]

Applying Bernstein’s inequality, it holds with probability at least \( 1 - \delta \) that
\[
|\mathcal{E}_{1,n}| \leq \sqrt{\frac{2\mathbb{E}[\log^2 2D_g^*(X)] \log(2/\delta)}{n} + \frac{2(\log 1/(2D_{\text{min}}) \vee \log 2D_{\text{max}}) \log(2/\delta)}{3n}},
\]
\[
|\mathcal{E}_{2,n}| \leq \sqrt{\frac{2\mathbb{E}[\log^2 2(1 - D_g^*(g(Y)))] \log(2/\delta)}{n} + \frac{2(\log 1/(2(1 - D_{\text{max}})) \vee \log 2(1 - D_{\text{min}})) \log(2/\delta)}{3n}}.
\]

To conclude, note that
\[
\mathbb{E}[\log^2 \{ 2D_g^*(X) \}] = \int \log^2 \frac{2p^*(x)}{p^*(x) + p_g(x)} p^*(x) \, dx
\]
\[
\leq 4 \log^2 \left( 1 + \frac{p_{\text{max}}}{p_{\text{min}}} \right) \int \frac{(p^*(x) - p_g(x))^2}{(p^*(x) + p_g(x))^2} p^*(x) \, dx
\]
\[
\leq \frac{4 \log^2 \left( 1 + \frac{p_{\text{max}}}{p_{\text{min}}} \right)}{p_{\text{min}}} \int (p^*(x) - p_g(x))^2 p^*(x) \, dx
\]
\[
= \frac{4 \log^2 \left( 1 + \frac{p_{\text{max}}}{p_{\text{min}}} \right)}{p_{\text{min}}} \|p^* - p_g\|_{L_2(p^*)}^2,
\]
and, similarly,
\[
\mathbb{E}[\log^2 \{ 2(1 - D_g^*(g(Y))) \}] = \int \log^2 \frac{2p_g(x)}{p^*(x) + p_g(x)} p_g(x) \, dx
\]
\[
\leq \frac{4 \log^2 \left( 1 + \frac{p_{\text{max}}}{p_{\text{min}}} \right)}{p_{\text{min}}} \|p^* - p_g\|_{L_2(p_g)}^2.
\]

\[
\square
\]

Lemma 9. Under Assumptions AG and A\( \Phi \), for any \( g_1, g_2 \in \mathcal{G} \), it holds that
\[
\left\| D_{g_1}^* - D_{g_2}^* \right\|_{\infty} \leq \frac{L_2}{p_{\text{min}}} \| g_1 - g_2 \|_{\mathcal{H}^{\beta + 1}(\mathcal{Y})}^\beta \| 1 \|_{\mathcal{H}^{\beta + 1}(\mathcal{Y})},
\]
where \( p_{\text{min}} \) and \( L_2 \) are defined in Lemma 4.

Proof. Note that, due to Assumption AG, \( p_{\text{min}} \leq p_{g_1} \leq p_{\text{max}} \) and \( p_{\text{min}} \leq p_{g_2} \leq p_{\text{max}} \). Using the expression (9) for the optimal discriminator \( D_{g}^* \), we obtain
\[
\left\| D_{g_1}^* - D_{g_2}^* \right\|_{\infty} = \left\| \frac{p_{g_1}(x)(g_1(x) - g_2(x))}{(p_{g_1}(x) + g_1(x))(p_{g_2}(x) + g_2(x))} \right\|_{\infty} \leq \frac{\| g_1 - g_2 \|_{\mathcal{H}^{\beta + 1}(\mathcal{Y})} \| 1 \|_{\mathcal{H}^{\beta + 1}(\mathcal{Y})}}{p_{\text{min}}} \left\| \frac{p_{g_1}(x)}{p_{g_2}(x)} \right\|_{\infty} \leq \frac{L_2}{p_{\text{min}}} \| g_1 - g_2 \|_{\mathcal{H}^{\beta + 1}(\mathcal{Y})}^\beta \| 1 \|_{\mathcal{H}^{\beta + 1}(\mathcal{Y})},
\]
where the last inequality is due to Lemma 4.

**Lemma 10.** Under Assumptions $AG$ and $A\phi$, for any $g \in \mathcal{G}$, it holds that

$$\frac{\|p_g - p^*\|_{L_2(p^*)}^2}{8p_{\text{max}}^2} \leq F(g) \leq \frac{\|p_g - p^*\|_{L_2(p^*)}^2}{8p_{\text{min}}^2},$$

and

$$\frac{\|p_g - p^*\|_{L_2(p_g)}^2}{8p_{\text{max}}^2} \leq F(g) \leq \frac{\|p_g - p^*\|_{L_2(p_g)}^2}{8p_{\text{min}}^2},$$

where $p_{\text{min}}$, $p_{\text{max}}$ are defined in (43).

**Proof.** Consider $f : [p_{\text{min}}, p_{\text{max}}] \rightarrow \mathbb{R}$,

$$f(u) = \frac{u}{2} \log \frac{2u}{u_0 + u} + \frac{u_0}{2} \log \frac{2u_0}{u_0 + u},$$

where $u_0 \in [p_{\text{min}}, p_{\text{max}}]$ is fixed. Direct calculations yield

$$f'(u) = \frac{1}{2} \log \frac{2u}{u_0 + u}, \quad f''(u) = \frac{u_0}{2u(u_0 + u)} \in \left[\frac{u_0}{4p_{\text{max}}^2}, \frac{u_0}{4p_{\text{min}}^2}\right].$$

Then, taking into account that $f(u_0) = f'(u_0) = 0$ and using Taylor’s expansion, we obtain

$$\frac{u_0(u - u_0)^2}{8p_{\text{max}}^2} \leq f(u) \leq \frac{u_0(u - u_0)^2}{8p_{\text{min}}^2}.$$

Now we apply this inequality to

$$F(g) = \frac{1}{2} \int \left( p_g(x) \log \frac{2p_g(x)}{p_g(x) + p^*(x)} + p^*(x) \log \frac{2p^*(x)}{p_g(x) + p^*(x)} \right) dx$$

with $u = p_g(x)$ and $u_0 = p^*(x)$. We obtain

$$\frac{\|p_g - p^*\|_{L_2(p^*)}^2}{8p_{\text{max}}^2} \leq F(g) \leq \frac{\|p_g - p^*\|_{L_2(p^*)}^2}{8p_{\text{min}}^2}.$$

Similarly, if we take $u = p^*(x)$ and $u_0 = p_g(x)$ then

$$\frac{\|p_g - p^*\|_{L_2(p_g)}^2}{8p_{\text{max}}^2} \leq F(g) \leq \frac{\|p_g - p^*\|_{L_2(p_g)}^2}{8p_{\text{min}}^2}.$$