Chapter 5

Generative adversarial networks

Contents

5.1 General principles

The goal of this chapter is to give insights and to provide theoretical guarantees for the so-called *generative adversarial network* (GAN) strategies in machine learning. Among others, applications include the generation of images, video, text, music. Neat examples of generators trained with the GAN strategy can be found here:

https://thisxdoesnotexist.com/

From a statistical point of view, we will see how it relates to density estimation. Before giving a general definition, we introduce the idea of GANs via a more standard approach: maximum likelihood.

5.1.1 From maximum likelihood to generative adversariality

Assume that we observe an i.i.d. n-sample $X_1, \dots, X_n$ over $\mathbb{R}^d$, generated from an unknown probability density $p^* : \mathbb{R}^d \to \mathbb{R}$. Consider the problem of estimating the density p^* via maximum likelihood within a class of densities $\mathscr{P} = \{p_\theta\}_{\theta \in \Theta}$, where $\Theta \subset \mathbb{R}^p$ is some parameter space. To picture things out, one can think of p^* as being $\mathscr{C}^\beta$-smooth, and $\mathscr{P}$ consisting of all the Fourier series with p coefficients,

parametrized by its coefficients in $\Theta = [-R, R]^p$. Here, provided it exists, the maximum (log-)likelihood estimator (MLE) writes as

$$\hat{p}_{MLE} = p_{\hat{\theta}_{MLE}} \in \operatorname*{argmax}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(X_i).$$

Informally, this estimator aims at estimating the limiting density[1]

$$\begin{aligned}\overline{p}_{MLE} &\in \operatorname*{argmax}_{\theta\in\Theta} \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(X_i) \\ &= \operatorname*{argmax}_{\theta\in\Theta} \mathbb{E}_{X\sim p^*}\left[\log p_\theta(X)\right] \text{ a.s.}\end{aligned}$$

From this formulation, we can link the MLE with the Kullback-Leibler divergence, an information-theoretic measure of discrepancy between measure, which the definition writes as follows.

Definition 5.1 (Kullback-Leibler divergence)**.** *For two probability distributions P and Q on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$, we define the* Kullback-Leibler *divergence between them as*

$$\mathrm{KL}(P\|Q) = \begin{cases} \mathbb{E}_{X\sim P}\left[\log\left(\mathrm{dP}/\mathrm{dQ}(\mathrm{X})\right)\right], & \text{if } P \ll Q, \\ +\infty, & \text{otherwise.}\end{cases}$$

Because

$$\mathrm{KL}(P\|Q) = \mathbb{E}_{X\sim Q}\left[\log\left(\frac{\mathrm{dP}}{\mathrm{dQ}}(X)\right)\frac{\mathrm{dP}}{\mathrm{dQ}}(X)\right]$$

and $\mathbb{R}_{\geq 0} \ni u \mapsto u\log u$ is strictly convex, KL is non-negative and $\mathrm{KL}(P\|Q) = 0$ if and only if $P = Q$. However, separator $\|$ is here to insist on the fact that KL is not symmetric. Given this formula, if p^* and all the p_θ's dominate each other, we can reformulate the above expression via

$$\begin{aligned}\operatorname*{argmax}_{\theta\in\Theta} \mathbb{E}_{X\sim p^*}\left[\log p_\theta(X)\right] &= \operatorname*{argmax}_{\theta\in\Theta} \mathbb{E}_{X\sim p^*}\left[\log p_\theta(X)\right] - \mathbb{E}_{X\sim p^*}\left[\log p^*(X)\right] \\ &= \operatorname*{argmax}_{\theta\in\Theta} -\mathbb{E}_{X\sim p^*}\left[\log\left(p^*(X)/p_\theta(X)\right)\right] \\ &= \operatorname*{argmin}_{\theta\in\Theta} \mathrm{KL}(p^*\|p_\theta),\end{aligned}$$

where we abused notation by identifying Lebesgue density functions with the associated probability distributions. As a result, $\hat{\theta}_{MLE}$ can be seen as an empirical minimizer of the Kullback-Leibler divergence to the target. Written this way, one directly sees the major drawbacks of the MLE. As a Kullback-Leibler approach — with KL not defined for all pair of distributions and not symmetric —, it is not intrinsic, and unstable with respect to model misspecification. To overcome this issue, one may symmetrize the KL divergence in such a way that it is defined for all pair of distributions. This symmetrized divergence is referred to as *Jensen-Shannon*.

Definition 5.2 (Jensen-Shannon divergence)**.** *Write $\mu = (P+Q)/2$. By $\mathrm{JS}(P,Q)$, we denote the* Jensen-Shannon *divergence*

$$\mathrm{JS}(P,Q) = \frac{1}{2}\mathrm{KL}\left(P\|\mu\right) + \frac{1}{2}\mathrm{KL}\left(Q\|\mu\right).$$

The map $(P,Q) \mapsto \sqrt{\mathrm{JS}(P,Q)}$ is a distance over the the of probability distributions over $\mathbb{R}^d$.

[1] We insist on the fact that this derivation is only informal and does need to be properly justified, for instance via the regularity of the model $(p_\theta)_{\theta\in\Theta}$.

Similarly as for KL, we may use JS to design an estimation strategy. Namely, the counterpart of $\overline{\theta}_{MLE}$ would naturally be defined as

$$\overline{\theta}_{\mathrm{JS}} \in \operatorname*{argmin}_{\theta\in\Theta} \mathrm{JS}(p^*, p_\theta). \tag{5.1}$$

From a purely theoretical point of view, the above expression can be made rigorous in very general models. However, it is not very practical since it explicitly involves density functions, and hence normalizing constants that can be hard to compute in high-dimensional settings or complicated models. Alternatively, GANs leverage the following dual formulation of JS.

Proposition 5.3 (Dual formulation of JS)**.** *The set of all* discriminator *functions is*

$$\mathscr{D}_\infty = \left\{D : \mathbb{R}^d \to [0,1] \text{ measurable}\right\}.$$

For all Borel probability distributions P, Q over $\mathbb{R}^d$,

$$\mathrm{JS}(P,Q) - \log(2) = \frac{1}{2} \max_{D\in\mathscr{D}_\infty} \left\{\mathbb{E}_{X\sim P}[\log D(X)] + \mathbb{E}_{Z\sim Q}[\log(1-D(Z))]\right\}.$$

Furthermore, if p and q denote respective density functions of P and Q with respect to $\mu = (P+Q)/2$, then the maximum on the right hand side is attained for

$$D^*_{(P,Q)}(x) = \frac{p(x)}{p(x)+q(x)}.$$

Proof. For all $D \in \mathscr{D}_\infty$,

$$\begin{aligned}&\frac{1}{2}\mathbb{E}_{X\sim P}[\log D(X)] + \frac{1}{2}\mathbb{E}_{Z\sim Q}[\log(1-D(Z))]\\ &\quad= \frac{1}{2}\int_{\mathbb{R}^d} p(x)\log(D(x)) + q(x)\log(1-D(x))\mu(\mathrm{d}x).\end{aligned}$$

But for all $(a,b) \in \mathbb{R}^2 \setminus \{(0,0)\}$, the real map $[0,1] \ni y \mapsto a\log y + b\log(1-y)$ achieves its maximum at $a/(a+b)$, so that

$$\begin{aligned}&\frac{1}{2}\sup_{D\in\mathscr{D}_\infty}\left\{\mathbb{E}_{X\sim P}[\log D(X)] + \frac{1}{2}\mathbb{E}_{Z\sim Q}[\log(1-D(Z))]\right\}\\ &\quad= \frac{1}{2}\int_{\mathbb{R}^d} p(x)\log\left(\frac{p(x)}{p(x)+q(x)}\right)\mu(\mathrm{d}x) + \frac{1}{2}\int_{\mathbb{R}^d} q(x)\log\left(1-\frac{p(x)}{p(x)+q(x)}\right)\mu(\mathrm{d}x)\\ &\quad= \frac{1}{2}\int_{\mathbb{R}^d} p(x)\log\left(\frac{p(x)}{p(x)+q(x)}\right)\mu(\mathrm{d}x) + \frac{1}{2}\int_{\mathbb{R}^d} q(x)\log\left(\frac{q(x)}{p(x)+q(x)}\right)\mu(\mathrm{d}x),\end{aligned}$$

and as $(p+q)/2 = 1$ since p and q are densities with respect to $\mu = (P+Q)/2$, the right hand side rewrites as

$$\frac{1}{2}\int_{\mathbb{R}^d} p(x)\log\left(\frac{p(x)}{2}\right)\mu(\mathrm{d}x) + \frac{1}{2}\int_{\mathbb{R}^d} q(x)\log\left(\frac{q(x)}{2}\right)\mu(\mathrm{d}x) = \mathrm{JS}(P,Q) - \log(2),$$

which concludes the proof. □

As a result, an equivalent min-max formulation of (5.1) writes as

$$\overline{\theta}_{\mathrm{JS}} \in \operatorname*{argmin}_{\theta\in\Theta} \max_{D\in\mathscr{D}_\infty} \mathbb{E}_{X\sim p^*}[\log D(X)] + \mathbb{E}_{Z\sim p_\theta}[\log(1-D(Z))], \tag{5.2}$$

which GANs will exploit through an empirical version of it.

5.1.2 Goodfellow generative adversarial (networks)

Note that in the above discussion, we have gone a step towards a density-free formulation of the problem, as (5.1) uses likelihood explicitly in JS, while (5.2) does not. In the same random variable-oriented spirit, GANs will only consider statistical models $\mathscr{P} = \mathscr{P}_\Theta$ composed of pushforward distribution of a simple-to-simulate distribution via maps called *generators*. That is, given a set of functions $\mathscr{G}$ from $\mathbb{R}^d$ to itself, we consider the variant

$$\overline{\theta}_{\mathrm{JS}} \in \operatorname*{argmin}_{g\in\mathscr{G}} \max_{D\in\mathscr{D}_\infty} \left\{\mathbb{E}_{X\sim P^*}[\log D(X)] + \mathbb{E}_{Y\sim Q_0}[\log(1-D(g(Y)))]\right\}, \tag{5.3}$$

where Q_0 usually stands for the uniform distribution over $[0,1]^d$, or the standard d-dimensional Gaussian.

Finally, to design a consistent empirical counterpart of (5.3), we allow for considering subset of discriminators $\mathscr{D} \subset \mathscr{D}_\infty$. We are now in position to define the so-called vanilla (or Goodfellow [GPAM+14]) *generative adversarial networks*

Definition 5.4 (Goodfellow Generative Adversarial Estimator)**.** *Let $X_1,\dots,X_n$ be a n-sample with distribution P^*, defined on some measurable space $(\mathsf{X},\mathscr{X})$. Pick:*

- *An* input distribution Q_0 *over a measurable space* $(\mathsf{Y},\mathscr{Y})$*;*
- *A* generator *class $\mathscr{G}$, composed of measurable functions $g:\mathsf{Y}\to\mathsf{X}$;*
- *A* discriminator *class $\mathscr{D}$, composed of measurable functions $D:\mathsf{X}\to[0,1]$.*

For all $g\in\mathscr{G}$, write P_g for the distribution of $g(Y)$ when $Y\sim Q_0$.

Given an independent n-sample $Y_1,\dots,Y_n$ with distribution Q_0, the Goodfellow generative adversarial distribution estimator *of P^* is the plugin $P_{\hat{g}}$, where*

$$\hat{g} \in \operatorname*{argmin}_{g\in\mathscr{G}} \max_{D\in\mathscr{D}} L_n(g,D),$$

where

$$L_n(g,D) = \frac{1}{2n}\sum_{i=1}^n \log D(X_i) + \frac{1}{2n}\sum_{j=1}^n \log\big(1-D(g(Y_j))\big). \tag{5.4}$$

The min-max formulation of GANs explains the naming "adversarial". It can be interpreted as a two-player game, opposing the generator and the discriminator. More precisely,

($\mathscr{G}$) The generator observes data $X_1,\dots,X_n$ only and does not know the true underlying distribution. It tries to fool the discriminator by producing "fake samples" $g(Y_1),\dots,g(Y_n)$ with $g\in\mathscr{G}$.

($\mathscr{D}$) Then, the discriminator is given the labelled data $X_1,\dots,X_n$ and $Y_1,\dots,Y_n$, and its goals is to classify them as best as it can with classifiers from the class $\mathscr{D}$. Here, $D(x)=0$ codes for "true data" and $D(x)=1$ for "fake data".

As suggested by the previous section, if $\mathscr{D}$ is properly chosen, $p_{\hat{g}}$ is expected to converge towards a minimizer of $\mathscr{G}\ni g\mapsto \mathrm{JS}(p^*,p_g)$. Furthermore, if Q_0 is well chosen and $\mathscr{G}$ rich enough, it shall hence converge towards p^* as n goes to infinity. See Section 5.3 for a more detailed discussion.

Remark 5.5.

- *In (5.4), the number of fake samples is equal to the number of real ones, but our analysis is also valid when the number of fake instances is greater than n.*
- *The estimator $\hat{g}$ is randomized, as it depends on external randomness arising from $Y_1,\dots,Y_n$.*

- *The above definition does not use neural networks at all, as generator and discriminator classes $\mathscr{G}$ and $\mathscr{D}$ are kept fully general. However, in practice, these classes are usually built with neural nets.*
- *As expected, GANs provide a somehow implicit statistical model: as it trains a* generator *to produce samples, this implicitly defines a probability distribution. Namely the distribution of samples that the network generates. Here, the model does not explicitly represent the probability distribution itself through its density. For instance, one cannot directly derive the probability assigned to a particular region of space given the candidate generator.*

5.2 A zoology of generative-adversarial strategies

5.2.1 f-divergences

The above min-max GAN approach can be applied for other objective functions than JS. The key property of JS is that it can be written as a supremum of a difference of expected values over a set of test (discriminator) functions. One of the most commonly used family of such divergences is called f-divergences.

Definition 5.6 (f-Divergence)**.** *Write $\mu = (P+Q)/2$. By $\mathrm{D}_f(P\|Q)$, we denote the* f-divergence,

$$\mathrm{D}_f(P\|Q) = \int_{\mathscr{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) \mu(\mathrm{d}x), \tag{5.5}$$

where the generator function $f : \mathbb{R}_+ \to \overline{\mathbb{R}}$ *is a strictly convex differentiable*[2] *function satisfying $f(1) = 0$.*

Here, the convexity of f and Jensen inequality ensure that $\mathrm{D}_f(P\|Q) \geq 0$, and $f(1) = 0$ that $\mathrm{D}_f(P\|Q) = 0$ if and only if $P = Q$. To obtain a variational formula similar to Proposition 5.3, we express f itself as a supremum. For this, consider the *Fenchel conjugate* f^* of f, also known as *convex conjugate*. It is defined as

$$f^*(t) = \sup_{u\in\mathbb{R}^d | f(u)<\infty} \{ut - f(u)\}. \tag{5.6}$$

One easily checks that f^* is convex[3] and lower-semi-continuous if f is lower-semi-continuous. Furthermore, if f is convex, then the pair (f, f^*) is dual to another in the sense that $f^{**} = f$. That is, we have

$$f(u) = \sup_{t\in\mathbb{R}^d | f^*(t)<\infty} \{tu - f^*(t)\}.$$

Therefore,

$$\mathrm{D}_f(P\|Q) = \int_{\mathbb{R}^d} q(x) \sup_{t\in\mathbb{R}^d | f^*(t)<\infty} \left\{ t\frac{p(x)}{q(x)} - f^*(t) \right\} \mu(\mathrm{d}x)$$

For all $x \in \mathbb{R}^d$, the supremum of $t \mapsto tp(x)/q(x) - f^*(t)$ is attained for $p(x)/q(x) - (f^*)'(t) = 0$. But since $f' \circ (f^*)'(t) = f' \circ (f^*)'(t) = t$ for all t, this is equivalent to $t = f'(p(x)/q(x))$. Hence, given any class $\mathscr{T}$ of measurable functions $T : \mathbb{R}^d \to \mathbb{R}$,

$$\begin{aligned}\mathrm{D}_f(P\|Q) &\geq \sup_{T\in\mathscr{T}} \left\{ \int_{\mathbb{R}^d} p(x)\, T(x)\, \mu(\mathrm{d}x) - \int_{\mathbb{R}^d} q(x)\, f^*(T(x))\, \mu(\mathrm{d}x) \right\} \\ &= \sup_{T\in\mathscr{T}} \left\{ \mathbb{E}_{X\sim P}[T(X)] - \mathbb{E}_{Z\sim Q}\left[f^*(T(Z))\right] \right\},\end{aligned}$$

[2] Generalizations of f-divergences to lower-semi-continuous (i.e. $\liminf_{x\to x_0} f(x) \geq f(x_0)$ for all x, i.e. sublevel sets $\{f \leq y\}$ are closed, i.e. infima of f on compact sets are attained) convex f is also standard. We omit this aspect for sake of simplicity.

[3] Even if f is not.

Discrepancy	Convex function $f(u)$	Conjugate $f^*(t)$	Optimal $T^*_{(P,Q)}(x)$
Total Variation	$\frac{1}{2}\lvert u-1\rvert$	$t\mathbb{1}_{[-1/2,1/2]}$	$\frac{1}{2}\text{sign}(\frac{p(x)}{q(x)}-1)$
Kullback-Leibler	$u\log u$	$\exp(t-1)$	$1+\log\frac{p(x)}{q(x)}$
Pearson χ^2	$(u-1)^2$	$t+t^2/4$	$2(\frac{p(x)}{q(x)}-1)$
Squared Hellinger	$\left(\sqrt{u}-1\right)^2$	$t/(1-t)$	$(\sqrt{\frac{p(x)}{q(x)}}-1)\cdot\sqrt{\frac{q(x)}{p(x)}}$
Jensen-Shannon	$u\log u-(u+1)\log\frac{1+u}{2}$	$-\log(2-\exp(t))$	$\log\frac{2p(x)}{p(x)+q(x)}$

Table 5.1: Adapted from [NCT16]. List of convex functions f, conjugate $f^*(t)$, and optimal $T^*_{(P,Q)}(x)$ (Proposition 5.7), for some classical f-divergences $\mathrm{D}_f(P\|Q)$ (Definition 5.6). Note that for the total variation distance, f is not differentiable everywhere. Though, the proof of Proposition 5.7 can be adapted using sub-differential arguments.

with equality when $\mathcal{T}$ contains the map

$$\widetilde{T}^*_{(P,Q)}(x) = f'\left(\frac{p(x)}{q(x)}\right).$$

This yields the following result.

Proposition 5.7 (Variational formulation of f-Divergences)**.** *Let $\mathcal{T}$ be a class of measurable functions $T:\mathbb{R}^d\to\mathbb{R}$ that contains the map $\widetilde{T}^*_{(P,Q)}(x)=f'\left(p(x)/q(x)\right)$. Then,*

$$\mathrm{D}_f(P\|Q)=\sup_{T\in\mathcal{T}}\left\{\mathbb{E}_{X\sim P}\left[T(X)\right]-\mathbb{E}_{Z\sim Q}\left[f^*(T(Z))\right]\right\}.$$

Some of the most standard f-divergences are displayed in Table 5.1, with their associated convex conjugate and discriminator function $T^*_{(P,Q)}$.

Remark 5.8. *To explicitly connect notation of Proposition 5.3 and Proposition 5.7, take $T=\log 2D$.*

Exercise 5.9 (f-GANs)**.** *Similarly to the Goodfellow GANs (Definition 5.4), design a f-generative adversarial density estimator with a f-divergence as an objective function.*

5.2.2 Integral probability metrics

Studying f-divergences in the previous section led to variational formulation of discrepancies between probability measures, taking the form

$$\sup_{T\in\mathcal{T}}\left\{\mathbb{E}_{X\sim P}\left[T(X)\right]-\mathbb{E}_{Z\sim Q}\left[f^*(T(Z))\right]\right\},$$

where $\mathcal{T}$ is a rich-enough class of functions $T:\mathbb{R}^d\to\mathbb{R}$ and f^* convex. Even for itself, this formula actually yields important discrepancies by only considering $f^*(t)=t$ but with specific classes $\mathcal{T}$. They are usually referred to as *integral probability metrics* (IPM).

Definition 5.10 (Integral Probability Metric)**.** *Let $\mathscr{D}_\infty$ be a class of measurable functions $D:\mathbb{R}^d\to\mathbb{R}$ called* discriminators. *Given two probability distributions P and Q on $(\mathbb{R}^d,\mathscr{B}(\mathbb{R}^d))$, $\mathrm{IPM}_{\mathscr{D}_\infty}(P,Q)$ is defined as*

$$\mathrm{IPM}_{\mathscr{D}_\infty}(P,Q)=\sup_{D\in\mathscr{D}_\infty}\left|\mathbb{E}_{X\sim P}[D(X)]-\mathbb{E}_{Z\sim Q}[D(Z)]\right|.$$

Discrepancy	Discriminator class $\mathscr{D}_\infty$	Comment
Total Variation	$\{D : \|D\|_\infty \le 1\}$	Also a f-divergence
Wasserstein 1	$\{D : \|D\|_{\mathrm{Lip}} \le 1\}$	Also a transport distance
Bounded Lipschitz	$\{D : \|D\|_\infty + \|D\|_{\mathrm{Lip}} \le 1\}$	
Kolmogorov	$\{\mathbb{1}_{(-\infty,t]}\}_{t\in\mathbb{R}}$	Only defined for $d = 1$

Table 5.2: List of classical integral probability metrics over $\mathbb{R}^d$. Here, $\|D\|_\infty = \sup_{x\in\mathbb{R}^d} |D(x)|$ and $\|D\|_{\mathrm{Lip}} = \sup_{x\neq y\in\mathbb{R}^d} |f(x) - f(y)|/\|x - y\|$.

The discrepancy $\mathrm{IPM}_{\mathscr{D}_\infty}$ clearly is non-negative, symmetric, and finite provided that $\mathscr{D}_\infty$ is not too wild. On the other hand, it satisfies separation if $\mathscr{D}_\infty$ is rich enough, hence making $\mathrm{IPM}_{\mathscr{D}_\infty}$ a metric over the space of distributions. See Table 5.2 for a few standard examples of integral probability metrics. As expected, the fact that $\mathrm{IPM}_{\mathscr{D}_\infty}$ writes as a supremum naturally yields generative-adversarial strategies.

Exercise 5.11 (IPM-GANs)**.** *Similarly to the Goodfellow GANs (Definition 5.4), design a generative adversarial density estimator based on an integral probability metric.*

Remark 5.12 (Wasserstein GANs)**.** *The most broadly used IPM-GANs correspond to* $\mathscr{D}_\infty = \{\|D\|_{\mathrm{Lip}} \le 1\}$ *and are usually referred to as* Wasserstein GANs.

5.3 A heuristic for optimality of GANs

In addition to providing random generators and being implementable in practice for high-dimensional data, generative adversarial strategies share the nice feature of mimicking optimal estimation schemes. Indeed, they can roughly be thought of as risk minimization, where the risk is an actual discrepancy between probability measures.

To see this, consider the pushforward distribution $P_{\hat{g}} := \hat{g}_\sharp Q_0$ of an f-GAN with generator class $\mathscr{G}$ and discriminator $\mathscr{D}$. That is, given sample $X_1, \dots, X_n$ generated from P^* and $Y_1, \dots, Y_n$ from Q_0, define

$$\hat{g} \in \operatorname*{argmin}_{g\in\mathscr{G}} \max_{D\in\mathscr{D}} L_n(g, D),$$

where for all $g \in \mathscr{G}$ and $D \in \mathscr{D}$,

$$L_n(g, D) := \frac{1}{n}\sum_{i=1}^n D(X_i) - \frac{1}{n}\sum_{j=1}^n f^*\big(D(g(Y_j))\big).$$

We also denote the integrated version of $L_n(g, D)$ by

$$L(g, D) := \mathbb{E}_{X\sim P}[D(X)] - \mathbb{E}_{Y\sim Q_0}\big[f^*\big(D(g(Y))\big)\big]$$

Heuristic. Going from integrated to sample quantities, we write

$$\begin{aligned}
\mathrm{D}_f(P_{\hat{g}}, P^*) &= \sup_{D\in\mathscr{D}_\infty} L(g, D) && \text{From Proposition 5.7,}\\
&\simeq \sup_{D\in\mathscr{D}} L(\hat{g}, D) && \text{If } \mathscr{D} \text{ is rich enough,}\\
&\simeq \sup_{D\in\mathscr{D}} L_n(\hat{g}, D) && \text{If } n \text{ is large enough,}\\
&= \inf_{g\in\mathscr{G}} \sup_{D\in\mathscr{D}} L_n(g, D) && \text{By definition of } \hat{g}.
\end{aligned}$$

Then unraveling this expression, we have

$$\begin{aligned}
\inf_{g\in\mathscr{G}} \sup_{D\in\mathscr{D}} L_n(g,D) &\simeq \inf_{g\in\mathscr{G}} \sup_{D\in\mathscr{D}} L(g,D) && \text{If } n \text{ is large enough,}\\
&\leq \inf_{g\in\mathscr{G}} \sup_{D\in\mathscr{D}_\infty} L(g,D) && \text{If } \mathscr{D}\subset\mathscr{D}_\infty,\\
&= \inf_{g\in\mathscr{G}} \mathrm{D}_f(P_g,P^*) && \text{From Proposition 5.7 again.}
\end{aligned}$$

See Remark 5.17 for a more precise heuristic taking into account bias terms. □

That is, in the limit $n\to\infty$ and under approximation assumptions on $\mathscr{D}$, the output distribution $P_{\hat{g}}$ performs approximately as good as the best distribution in the model $\{P_g\}_{g\in\mathscr{G}}$ for the distance D_f, in the sense that

$$\mathrm{D}_f(P_{\hat{g}},P^*) \lesssim \inf_{g\in\mathscr{G}} \mathrm{D}_f(P_g,P^*).$$

This property, called *oracle inequality*, will be made rigorous in Section 5.4.2 in the Jensen-Shannon case.

5.4 Non-asymptotic performance of Goodfellow GANs

This section, mostly borrowed from [BMN+21], studies the minimax convergence rates for density estimation with Goodfellow GANs. The main result (Exercise 5.21) relies on a general oracle bound for Goodfellow GANs (Theorem 5.16), which is applied to well-chosen classes $\mathscr{G}$ and $\mathscr{D}$ of neural networks (Theorem 5.19) for Hölder-smooth underlying densities (Assumption 5.13).

5.4.1 Notation

Pushforward density For all smooth one-to-one map $g:\mathsf{Y}\mapsto\mathsf{X}$, and random variable $Y\in\mathsf{Y}$ with density ϕ, a change of variable easily yields that the image $g(Y)$ of Y has density

$$p_g(x) = |\det[\nabla g(g^{-1}(x))]|^{-1}\phi(g^{-1}(x)), \quad x\in\mathsf{X}.$$

In what follows, we will only consider ϕ being the uniform density over compact Y with $\mathrm{Vol}(\mathsf{Y})=1$, so that we get the expression

$$p_g(x) = |\det[\nabla g(g^{-1}(x))]|^{-1}, \quad x\in\mathsf{X}.$$

Function smoothness and Hölder classes For all integer $s\geq 1$, the function class $C^s(\Omega)$ denotes the set of functions over the domain Ω which have bounded and continuous partial derivatives up to order s. More precisely,

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

where, for all multi-index $\boldsymbol{\gamma}=(\gamma_1,\dots,\gamma_d)\in\mathbb{N}_0^d$, the partial differential operator $D^{\boldsymbol{\gamma}}$ is defined as

$$D^{\boldsymbol{\gamma}} f_i = \frac{\partial^{|\boldsymbol{\gamma}|} f_i}{\partial x_1^{\gamma_1}\cdots\partial x_d^{\gamma_d}}, \quad i\in\{1,\dots,m\}, \text{ and } \|D^{\boldsymbol{\gamma}} f\|_{L^\infty(\Omega)} = \max_{1\leq i\leq m} \|D^{\boldsymbol{\gamma}} f_i\|_{L^\infty(\Omega)},$$

where $|\boldsymbol{\gamma}| := \sum_{i=1}^d \gamma_i$ is the *order* of $D^{\boldsymbol{\gamma}}$. For order one, we use the usual notation $\nabla f = (\partial f_i/\partial x_j)$ $i=1,\dots,m$, $j=1,\dots,d$.

Given a C^2 function $\varphi : \mathbb{R}^d \mapsto \mathbb{R}$, we write $\nabla^2\varphi(x) \in \mathbb{R}^{d\times d}$ for its Hessian at $x \in \Omega$. For a function $f : \Omega \to \mathbb{R}^m$ and any positive number $0 < \delta \le 1$, the *Hölder constant* of order δ is defined as

$$[f]_\delta := \max_{i\in\{1,\dots,m\}} \sup_{x\neq y\in\Omega} \frac{|f_i(x) - f_i(y)|}{\min\{1, \|x-y\|\}^\delta}\,. \tag{5.7}$$

Now, for all $\alpha > 0$, we can define the *Hölder ball* $\mathscr{H}^\alpha(\Omega, H)$. If we let $s = \lfloor\alpha\rfloor$ be the largest integer *strictly less* than α, $\mathscr{H}^\alpha(\Omega, H)$ contains all functions in $C^s(\Omega)$ with δ-Hölder-continuous, $\delta = \alpha - s > 0$, partial derivatives of order s. Formally,

$$\mathscr{H}^\alpha(\Omega, H) = \left\{f \in C^s(\Omega) : \quad \|f\|_{\mathscr{H}^\alpha} := \max\{\|f\|_{C^s}, \max_{|\boldsymbol{\gamma}|=s} [D^{\boldsymbol{\gamma}} f]_\delta\} \le H\right\}.$$

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

$$\left|\frac{\partial f_i(x)}{\partial x_j} - \frac{\partial f_i(y)}{\partial x_j}\right| \le \|f\|_{\mathscr{H}^{1+\beta}} \cdot \|x-y\|^{1\wedge\beta} \le H \cdot \|x-y\|^{1\wedge\beta},$$

for all $x, y \in \Omega$, since $\|f\|_{\mathscr{H}^{\beta_1}} \le \|f\|_{\mathscr{H}^{\beta_2}}$ for all $\beta_2 \ge \beta_1$. We will also write $f \in \mathscr{H}^\alpha(\Omega)$ if $f \in \mathscr{H}^\alpha(\Omega, H)$ for some $H < \infty$. We also introduce a class of Λ-regular functions $\mathscr{H}^\alpha_\Lambda(\Omega, H)$, $\Lambda > 1$:

$$\mathscr{H}^\alpha_\Lambda(\Omega, H) = \left\{f \in \mathscr{H}^\alpha(\Omega, H) : \Lambda^{-2}\mathrm{Id}_{d\times d} \preceq \nabla f^\top \nabla f(x) \preceq \Lambda^2 \mathrm{Id}_{d\times d} \text{ for all } x \in \Omega\right\}, \tag{5.8}$$

where for symmetric matrices $A, B \in \mathbb{R}^{d\times d}$ we write $A \preceq B$ if $u^\top(B - A)u \ge 0$ for all $u \in \mathbb{R}^d$.

5.4.2 A general oracle inequality

This section is devoted to the statement of a general oracle inequality for density estimation with Goodfellow GANs (Theorem 5.16). For this, let us list the assumptions to made to obtain it. We first require smoothness and specific form of the target density.

Assumption 5.13 (On p^*)**.** *There exist constants $\beta > 2$, $H^* > 0$ and $\Lambda > 1$ such that p^* is of the form*

$$p^*(x) := p_{g^*}(x) = |\det[\nabla g^*((g^*)^{-1}(x))]|^{-1}, \quad x \in \mathsf{X},$$

for some $g^* \in \mathscr{H}^{1+\beta}_\Lambda(\mathsf{Y}, H^*)$.

In particular, this implies that g^* is β-Hölder. Even more precisely, we have that p^* can actually be represented as the density of a random variable $g^*(Y)$, for some $g^* \mathscr{H}^{1+\beta}_\Lambda(\mathsf{Y}, H^*)$ and Y uniform over Y with $\mathrm{Vol}(\mathsf{Y}) = 1$.

Given this assumption, it is then natural to require that the generator class $\mathscr{G}$ is a subset of "the true space" $\mathscr{H}^2_\Lambda(\mathsf{Y}, H_{\mathscr{G}})$. In short, this means that the estimator $\hat{g}$ and its target g^* will live in the same regularity space. However, note that g^* needs not actually be in $\mathscr{G}$. Also, generators being forced to be well-conditioned diffeomorphisms will greatly simplify proof technicalities.

Assumption 5.14 (On $\mathscr{G}$)**.** *The latent random elements $Y_1, \dots, Y_n$ are uniformly distributed on a compact set Y with Vol$(\mathsf{Y}) = 1$. Moreover, there exist $H_{\mathscr{G}} > 0$ and $\Lambda > 1$ such that the class of generators fulfills*

$$\mathscr{G} = \mathscr{G}(\Lambda, H_{\mathscr{G}}) \subseteq \mathscr{H}^2_\Lambda(\mathsf{Y}, H_{\mathscr{G}})\,.$$

Finally, we require smoothness of the generator classes, in the following sense.

Assumption 5.15 (On $\mathscr{D}$). *There exist constants $H_{\mathscr{D}} > 0$ and $0 < D_{\min} \le D_{\max} < 1$ such that*

$$\mathscr{D} = \mathscr{D}(D_{\min}, D_{\max}, H_{\mathscr{D}}) \subseteq \mathscr{H}^1(\mathsf{X}, H_{\mathscr{D}}),$$

and each $D \in \mathscr{D}$ satisfies

$$D(x) \in [D_{\min}, D_{\max}] \subset (0,1) \quad \textit{for all } x \in \mathsf{X}.$$

To get insights on this assumption, recall that under Assumption 5.13, $p^* = p_{g^*}$ with g^* at least $C^{\beta+1}$. Furthermore, from Proposition 5.3, the optimal discriminator between p^* and a candidate generator $g \in \mathscr{G}$ is $D_g^* := p_{g^*}/(p_g + p_{g^*})$. Hence, for the adversary, this means that there is no need to seek for a discriminator that is not smooth, and that keeping its search space within $\mathscr{H}^1$ is sufficient. Furthermore, as $\hat{g}$ converges to g^* (in $\mathscr{H}^1$), $D_{\hat{g}}^*$ goes to $1/2$. As a result, only considering discriminators bounded away from 0 and 1 is also sufficient.

We are now in position to state the sharp oracle inequality for general classes $\mathscr{G}$ and D satisfying Assumptions 5.14 and 5.15.

Theorem 5.16 ([BMN+21, Theorem 1]). *Suppose that Assumptions 5.13, 5.14, and 5.15 hold. Let*

$$r_n^0(\mathscr{D}) = \log\big(\mathscr{N}(\mathscr{D}, \|\cdot\|_{L^\infty(\mathsf{X})}, 1/n)\big),$$

and

$$r_n^1(\mathscr{G}) = \log\big(\mathscr{N}(\mathscr{G}, \|\cdot\|_{\mathscr{H}^1(\mathsf{Y})}, 1/n)\big),$$

where $\mathscr{N}(\mathscr{G}, \|\cdot\|_{\mathscr{H}^1(\mathsf{Y})}, 1/n)$ and $\mathscr{N}(\mathscr{D}, \|\cdot\|_{L^\infty(\mathsf{X})}, 1/n)$ are $(1/n)$-covering numbers of the classes $\mathscr{G}$ and $\mathscr{D}$, respectively. Then for all $\delta \in (0, 1/4)$, with probability at least $1 - 4\delta$,

$$\begin{aligned} \mathrm{JS}(p_{\hat{g}}, p^*) - \Delta_{\mathscr{G}} - \Delta_{\mathscr{D}} \lesssim & \sqrt{\frac{\Delta_{\mathscr{G}}\left(r_n^1(\mathscr{G}) + \log(2/\delta)\right)}{n}} \\ & + \sqrt{\frac{\Delta_{\mathscr{D}}\left(r_n^1(\mathscr{G}) + r_n^0(\mathscr{D}) + \log(2/\delta)\right)}{n}} \\ & + \frac{\left(r_n^1(\mathscr{G}) + r_n^0(\mathscr{D}) + \log(2/\delta)\right)}{n}, \end{aligned} \tag{5.9}$$

where

$$\begin{aligned} \Delta_{\mathscr{G}} &:= \min_{g \in \mathscr{G}} \mathrm{JS}(p_g, p^*), \\ \Delta_{\mathscr{D}} &:= \max_{g \in \mathscr{G}} \big|\mathrm{JS}(p^*, p_g) - \max_{D \in \mathscr{D}} \big(L(g, D) + \log 2\big)\big|, \end{aligned}$$

and

$$D_g^*(x) := \frac{p^*(x)}{p^*(x) + p_g(x)}, \quad x \in \mathsf{X}. \tag{5.10}$$

Remark 5.17 (On Intrinsic Biases of GANs). • $\Delta_{\mathscr{G}}$ *is called* generator bias, *and $\Delta_{\mathscr{D}}$ is called* discriminator bias. *To see why, assume that you have access to an infinite sample (i.e. $n \to \infty$ with fixed $\mathscr{G}$ and $\mathscr{D}$, i.e. replace $L_n(g, D)$ by its integrated version $L(g, D)$), and consider*

$$\widetilde{g} \in \operatorname*{argmin}_{g \in \mathscr{G}} \max_{D \in \mathscr{D}} L(g, D).$$

Also consider the generator $\overline{g} \in \mathscr{G}$ for which $p_{\overline{g}}$ best approximates p^* in Jensen-Shannon distance. That is, $\mathrm{JS}(p^*, p_{\overline{g}}) := \min_{g\in\mathscr{G}} \mathrm{JS}(p^*, p_g) = \Delta_{\mathscr{G}}$. *Equivalently, from Proposition 5.3, we can write*

$$\overline{g} \in \operatorname*{argmin}_{g\in\mathscr{G}} \max_{\substack{D:\mathsf{X}\to[0,1]\\ \text{measurable}}} L(g,D).$$

Hence, since from Proposition 5.3, $L(g, D_g^*) = \mathrm{JS}(p^*, p_g) - \log(2)$, *we have*

$$\begin{aligned}
\mathrm{JS}(p^*, p_{\widetilde{g}}) &= \mathrm{JS}(p^*, p_{\overline{g}}) + \big(\mathrm{JS}(p^*, p_{\widetilde{g}}) - \mathrm{JS}(p^*, p_{\overline{g}})\big) \\
&= \Delta_{\mathscr{G}} + \Big(L(\widetilde{g}, D_{\widetilde{g}}^*) - L(\overline{g}, D_{\overline{g}}^*)\Big) \\
&= \Delta_{\mathscr{G}} + \underbrace{\big(L(\widetilde{g}, D_{\widetilde{g}}^*) - \max_{D\in\mathscr{D}} L(\widetilde{g}, D)\big)}_{B_1} + \underbrace{\big(\max_{D\in\mathscr{D}} L(\widetilde{g}, D) - L(\overline{g}, D_{\overline{g}}^*)\big)}_{B_2}.
\end{aligned}$$

But rewriting the first term of the right-hand side, we notice that since $\widetilde{g} \in \mathscr{G}$,

$$\begin{aligned}
B_1 &= \min_{D\in\mathscr{D}} \Big\{L(\widetilde{g}, D_{\widetilde{g}}^*) - L(\widetilde{g}, D)\Big\} \\
&\le \max_{g\in\mathscr{G}} \min_{D\in\mathscr{D}} \Big\{L(\widetilde{g}, D_{\widetilde{g}}^*) - L(\widetilde{g}, D)\Big\} \\
&=: \Delta_{\mathscr{D}}.
\end{aligned}$$

Finally, B_2 actually is non-positive, since $\mathscr{D}$ consists only of measurable maps $D : \mathsf{X} \to [0,1]$, and

$$B_2 = \min_{g\in\mathscr{G}} \max_{D\in\mathscr{D}} L(g,D) - \min_{g\in\mathscr{G}} \max_{\substack{D:\mathsf{X}\to[0,1]\\ \text{measurable}}} L(g,D) \le 0.$$

- *From the AM–GM inequality $\sqrt{ab} \le (a+b)/2$, (5.9) can be cast into a simpler form: it yields that with probability at least $1-\delta$,*

$$\mathrm{JS}(p_{\widehat{g}}, p^*) \lesssim \Delta_{\mathscr{G}} + \Delta_{\mathscr{D}} + \frac{\big(r_n^1(\mathscr{G}) + r_n^0(\mathscr{D}) + \log(1/\delta)\big)}{n}.$$

  However, (5.9) is written so as to exhibit the optimal bias terms $\Delta_{\mathscr{G}}$ and $\Delta_{\mathscr{G}}$, with no extra multiplicative constants possibly hidden using the $\lesssim$ sign.

Sketch of proof of Theorem 5.16. In what follows, we let

$$\overline{g} \in \operatorname*{argmin}_{g\in\mathscr{G}} \mathrm{JS}(p^*, p_g),$$

so that $\mathrm{JS}(p^*, p_{\overline{g}}) = \Delta_{\mathscr{G}} = L(\overline{g}, D_{\overline{g}}^*) + \log 2$. For short, we also let $F(g) := \mathrm{JS}(p^*, p_g) = L(g, D_g^*) + \log(2)$, and $F_n(g) := L_n(g, D_g^*) + \log(2)$ its empirical counterpart (see (5.4)), so that by definition of $\hat{g}$,

$$\hat{g} \in \operatorname*{argmin}_{g\in\mathscr{G}} \max_{D\in\mathscr{D}} L_n(g,D).$$

Given this notation, we write

$$\begin{aligned}
\mathrm{JS}(p^*, p_{\hat{g}}) - \Delta_{\mathscr{G}} &= F(\hat{g}) - F(\overline{g}) \\
&= \underbrace{\big(F(\hat{g}) - F_n(\hat{g})\big)}_{T_1} + \underbrace{\big(F_n(\hat{g}) - F_n(\overline{g})\big)}_{T_2} + \underbrace{\big(F(\overline{g}) - F_n(\overline{g})\big)}_{T_3}.
\end{aligned} \tag{5.11}$$

- Terms T_1 and T_3 both consist of a difference between $F(g)$ and its empirical counterpart $F_n(g)$, for $g \in \{\hat{g}, \overline{g}\} \subset \mathscr{G}$. We will hence use a union bound together with a covering number argument.
  - For the covering argument, elementary differential calculus yields that for all $g_0, g_1 \in \mathscr{G}$, $|F(g_0) - F(g_1)| \lesssim \|p_{g_0} - p_{g_1}\|_{L^\infty(\mathsf{X})} \lesssim \|g_0 - g_1\|_{\mathscr{H}^1(\mathsf{Y})}$, so that we need to cover $\mathscr{G}$ in $\mathscr{H}^1$ norm.
  - Then, for all *fixed* $g \in \mathscr{G}$, we derive a concentration inequality of $F_n(g)$ around its mean $\mathbb{E}[F_n(g)] = F(g)$ using Bernstein inequality. As a sum of independent variables, the variance of $F_n(g)$ writes as

$$\begin{aligned}\mathrm{Var}(F_n(g)) &= \frac{\mathrm{Var}\big(\log\big(D_g^*(X)\big)\big) + \mathrm{Var}\big(\log\big(1 - D_g^*(Y)\big)\big)}{4n} \\ &= \frac{\mathrm{Var}\big(\log\big(2D_g^*(X)\big)\big) + \mathrm{Var}\big(\log\big(2(1 - D_g^*(Y))\big)\big)}{4n}.\end{aligned}$$

    This last expression with an extra factor 2 in the log allows to recognize distance terms. Namely, by definition of D_g^*,

$$\begin{aligned}\mathrm{Var}\big(\log\big(2D_g^*(X)\big) &\le \mathbb{E}\big[\log^2\big(2D_g^*(X)\big)\big] \\ &= \int_{\mathsf{X}} \log^2\left(\frac{2p^*}{p^* + p_g}\right) p^* \\ &= \int_{\mathsf{X}} \log^2\left(1 + \frac{p^* - p_g}{p^* + p_g}\right) p^* \\ &\lesssim \int_{\mathsf{X}} (p^* - p_g)^2 p^* \\ &= \|p^* - p_g\|^2_{L^2(p^*)} \\ &\lesssim F(g) \\ &= \mathrm{JS}(p^*, p_g),\end{aligned}$$

    where the $\lesssim$ inequalities come from the fact that both p^* and p_g are bounded away from zero and infinity on X, which is a consequence of g^* and g being well-conditioned diffeomorphisms (Assumptions 5.13 and 5.14). Similarly, we have $\mathrm{Var}\big(\log\big(2(1 - D_g^*(Y))\big)\big) \lesssim F(g)$.

    Remark 5.18. *Note that here, up to constants, the variance of $F_n(g)$ is bounded by its expectation $F(g) = \mathrm{JS}(p^*, p_g)$. This phenomenon, called* variance localization *in stochastic process theory, yields that the closer g is from g^*, the better $F_n(g)$ concentrates around its mean. Technically speaking, it is this phenomenon that allows for stochastic terms of order $\sqrt{\Delta_{\mathscr{G}}/n}$ and $\sqrt{\Delta_{\mathscr{D}}/n}$ (with $\Delta_{\mathscr{G}}, \Delta_{\mathscr{D}} \to 0$) instead of only $\sqrt{1/n}$ in the final bound.*

    As a result, from Bernstein inequality, for $g = \overline{g}$, we obtain

$$T_3 \lesssim \sqrt{\frac{F(\overline{g})\log(1/\delta)}{n}} = \sqrt{\frac{\Delta_{\mathscr{G}}\log(1/\delta)}{n}}.$$

    For $g = \hat{g}$, applying Bernstein again together with the equality $F_n(\hat{g}) = F_n(\overline{g}) + T_2$ then allows to conclude[4] the study of T_1 and T_3, provided that we derive a bound on T_2.

- Now examining $T_2 = F_n(\hat{g}) - F_n(\overline{g})$, let us introduce

$$\widehat{D}_g \in \operatorname*{argmax}_{D \in \mathscr{D}} L_n(g, D)$$

[4] A few steps skipped here. See [BMN+21] for a fully detailed derivation.

for the (empirically) selected discriminator against fixed candidate generator $g \in \mathscr{G}$. Similarly as above, we decompose T_2 as

$$\begin{aligned} T_2 &= L_n(\hat{g}, D^*_{\hat{g}}) - L_n(\overline{g}, D^*_{\overline{g}}) \\ &= \underbrace{\big(L_n(\hat{g}, D^*_{\hat{g}}) - L_n(\hat{g}, \widehat{D}_{\hat{g}})\big)}_{T_{2,1}} + \underbrace{\big(L_n(\hat{g}, \widehat{D}_{\hat{g}}) - L_n(\overline{g}, \widehat{D}_{\overline{g}})\big)}_{T_{2,2}} + \underbrace{\big(L_n(\overline{g}, \widehat{D}_{\overline{g}}) - L_n(\overline{g}, D^*_{\overline{g}})\big)}_{T_{2,3}}. \end{aligned}$$

By definition of $\hat{g}$ and the fact that $\overline{g} \in \mathscr{G}$, we have $T_{2,2} \leq 0$. On the other hand, $T_{2,1}$ and $T_{2,3}$ both have the form $L_n(g, D^*_g) - L_n(g, \widehat{D}_g)$ for some $g \in \mathscr{G}$. Similarly as above, Bernstein inequality on the quantity $|L_n(g,D) - L(g,D) - L_n(g, D^*_g) + L(g, D^*_g)|$ for all $g \in \mathscr{G}$, and a covering argument allows to conclude. Although we do not detail this derivation, notice that here, the covering of $\mathscr{D}$ is only made in $L^\infty(\mathsf{X})$ norm, since $L(g, D_0) - L(g, D_1) \lesssim \|D_0 - D_1\|_{L^\infty(\mathsf{X})}$. Also note that the final bound naturally makes $\Delta_{\mathscr{D}}$ appear, since $T_{2,3}$ tends to essentially be negative, and $T_{2,1} = L_n(\hat{g}, D^*_{\hat{g}}) - L_n(\hat{g}, \widehat{D}_{\hat{g}})$ to concentrate around $L(\hat{g}, D^*_{\hat{g}}) - L(\hat{g}, \overline{D}_{\hat{g}})$ with $\overline{D}_{\hat{g}} \in \operatorname{argmax}_{D \in \mathscr{D}} L(\hat{g}, D)$ being the (integrated-)best discriminator of $\mathscr{D}$ between p^* and $p_{\hat{g}}$, hence yielding $T_{2,1} \leq \Delta_{\mathscr{D}}$ up to stochastic terms. □

5.4.3 Building generators with deep nets

We now build classes of neural networks $\mathscr{G}$ and $\mathscr{D}$ to be plugged in Theorem 5.16, in the case where the underlying distribution p^* is β-Hölder. These network classes should hence approximate Hölder classes of order β and $\beta+1$ in L^∞ and $\mathscr{H}^1$ norms respectively. As shown in Theorem 4.5, one could actually achieve precision ε with ReLU networks having constant depth and at most $O(\varepsilon^{-d/\beta})$ coefficients. However, these networks would not themselves be Hölder smooth for $\beta > 1$. Hence, such network classes cannot fulfill Assumptions 5.14 and 5.15.

Instead, we use ReQU activation function $\rho(u) = (u_+)^2$, which allow to realize any piecewise polynomial exactly. In particular, ReQU networks allow to realize *splines* exactly, which are designed to be smooth at points of "piece gluing".

Theorem 5.19 ([BMN+21, Proposition 3] Simplified)**.** *Let $\beta > 2$ and let $p, d, K \in \mathbb{N}$. Then, for all $f : [0,1]^d \to \mathbb{R}^p$, $f \in \mathscr{H}^\beta([0,1]^d, H)$, there exists a neural network $\Phi_f : [0,1]^d \to \mathbb{R}^p$ with ReQU activation function $\rho(u) = (u_+)^2$ such that:*

- *For all $\ell \in \{0, \dots, \lfloor\beta\rfloor\}$, $\big\|f - \Phi_f\big\|_{\mathscr{H}^\ell([0,1]^d)} \lesssim \dfrac{H}{K^{\beta-\ell}}$;*
- *Φ_f is such that $L(\Phi_f) \lesssim 1$, $\|\Phi_f\|_0 \lesssim p(K+\beta)^d$, and $W(\Phi_f) \lesssim p(K+\beta)^d$;*
- *$\Phi_f \in \mathscr{H}^{\lfloor\beta\rfloor}([0,1]^d, H_0)$ with $H_0 = H + C_{d,\beta,\lfloor\beta\rfloor} H / K^{\beta-\ell}$.*

Proof of Theorem 5.19. See [BMN+21, Proposition 3]. □

As a direct consequence, we obtain the following ready-to-use corollary.

Corollary 5.20. *Let $\beta > 2$. Then for sufficiently small $\varepsilon > 0$, there exists a class $\mathscr{G}$ of ReQU neural networks such that for all $\Phi \in \mathscr{G}$, $L(\Phi) = O(1)$, $\|\Phi\|_0 = O(\varepsilon^{-d/\beta})$, $W(\Phi) = O(\varepsilon^{-d/\beta})$, and such that*

$$\sup_{g^* \in \mathscr{H}^{\beta+1}_\Lambda(\mathsf{Y}, H^*)} \inf_{\Phi \in \mathscr{G}} \|\Phi - g^*\|_{\mathscr{H}^1(\mathsf{Y})} \leq \varepsilon.$$

Moreover, if Assumption 5.13 is fulfilled, there exists a class $\mathscr{D}$ of ReQU neural networks such that for all $\widetilde{\Phi} \in \mathscr{G}$, $L(\widetilde{\Phi}) \lesssim 1$, $\|\widetilde{\Phi}\|_0 \lesssim \varepsilon^{-d/\beta}$, $W(\widetilde{\Phi}) \lesssim \varepsilon^{-d/\beta}$, and such that

$$\sup_{\Phi \in \mathscr{G}} \inf_{\widetilde{\Phi} \in \mathscr{D}} \|D^*_\Phi - \widetilde{\Phi}\|_{L^\infty(\mathsf{X})} \leq \varepsilon.$$

Here the hidden constants in $\lesssim$ depend on d, $\lfloor\beta\rfloor$, and H^*.

5.4.4 Minimaxity of Goodfellow GANs

Building upon the two previous sections, we are finally able to derive the main result of Section 5.4.

Exercise 5.21 (Rate of convergence for Goodfellow GANs)**.** *Combine Theorem 5.16 and Corollary 5.20 to show that under Assumption 5.13, a suitable choice of neural network classes $\mathscr{G}$ and $\mathscr{D}$ yields a Goodfellow GAN strategy with associated density estimator $p_{\hat{g}}$ that satisfies*

$$\mathrm{JS}(p^*, p_{\hat{g}}) \lesssim \left(\frac{\log n}{n}\right)^{2\beta/(2\beta+d)} + \frac{\log(1/\delta)}{n}$$

with probability at least $1-\delta$. Comment on this rate of convergence.

Bibliography

[BMN+21] Denis Belomestny, Eric Moulines, Alexey Naumov, Nikita Puchkin, and Sergey Samsonov. Rates of convergence for density estimation with GANs. *arXiv e-prints*, page arXiv:2102.00199, January 2021.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[NCT16] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.