A STATISTICAL PERSPECTIVE ON ALGORITHM UNROLLING MODELS FOR INVERSE PROBLEMS

YVES ATCHADÉ, XINRU LIU, AND QIUYUN ZHU

(April 2023)

Abstract. We consider inverse problems where the conditional distribution of the observation $\mathbf{y}$ given the latent variable of interest $\mathbf{x}$ (also known as the forward model) is known, and we have access to a data set in which multiple instances of $\mathbf{x}$ and $\mathbf{y}$ are both observed. In this context, algorithm unrolling has become a very popular approach for designing state-of-the-art deep neural network architectures that effectively exploit the forward model. We analyze the statistical complexity of the gradient descent network (GDN), an algorithm unrolling architecture driven by proximal gradient descent. We show that the unrolling depth needed for the optimal statistical performance of GDNs is of order $\log(n)/\log(\varrho_n^{-1})$, where n is the sample size, and ϱ_n is the convergence rate of the corresponding gradient descent algorithm. We also show that when the negative log-density of the latent variable $\mathbf{x}$ has a simple proximal operator, then a GDN unrolled at depth D' can solve the inverse problem at the parametric rate $O(D'/\sqrt{n})$. Our results thus also suggest that algorithm unrolling models are prone to overfitting as the unrolling depth D' increases. We provide several examples to illustrate these results.

1. Introduction

Inverse problems are common problems in science and engineering where one seeks information on a latent variable of interest, given some related observation. We consider an inverse problem with a latent quantity of interest $\mathbf{x} \in \mathbb{R}^{d_x}$ that is related to the observed variable $\mathbf{y} \in \mathbb{R}^{d_y}$ through the so-called forward statistical model

$$\mathbf{y} \mid \mathbf{x} \sim e^{-f(\mathbf{y}|\mathbf{x})}\mathrm{d}\mathbf{y}, \tag{1}$$

2010 *Mathematics Subject Classification.* 62F15, 62Jxx.

Key words and phrases. Inverse problems, algorithm unrolling models, nonparametric regression, Bayesian deep learning, gradient descent networks, Posterior contraction.

Y. Atchadé: Boston University, 111 Cummington Mall, Boston 02215 MA, United States. *E-mail address:* atchade@bu.edu.

X. Liu: Boston University, 111 Cummington Mall, Boston 02215 MA, United States. *E-mail address:* xinruliu@bu.edu.

Q. Zhu: University of Minnesota, 224 Church St SE, Minneapolis 55455 MN, United States. *E-mail address:* qzhu@umn.edu.

for some function $f(\cdot|\mathbf{x}) :\ \mathbb{R}^{d_y} \to \mathbb{R}$. Throughout the paper, unless otherwise stated, all model densities are defined with respect to the corresponding Lebesgue measure. Although the function f is unknown in general, we focus in this work on inverse problems for which the forward model is well-understood and f is known. This is the case with many inverse problems in imaging. An important special case in the applications is the Gaussian linear model corresponding (up to an additive constant that we ignore) to

$$f(\mathbf{y}|\mathbf{x}) = \frac{1}{2v^2}\|\mathbf{y} - A\mathbf{x}\|_2^2, \tag{2}$$

with known parameters $v > 0$, and $A \in \mathbb{R}^{d_y \times d_x}$. When the inverse problem is ill-posed, additional knowledge is fundamental for good recovery of $\mathbf{x}$. For example in the linear regression model (2), it is well-known that without any additional assumption, the minimax optimal rate in the estimation of $\mathbf{x}$ is of order $\sqrt{d_x/d_y}$. However this rate can be improved if $\mathbf{x}$ is known to possess some additional features such as smoothness or sparsity. A Bayesian perspective is particularly simple. If μ_0 denotes a prior distribution that encodes the information available on $\mathbf{x}$, then $\mathbf{x}$ is inferred using its posterior distribution

$$\pi_{\mu_0}(\mathrm{d}\mathbf{x}|\mathbf{y}) \propto \mu_0(\mathrm{d}\mathbf{x})e^{-f(\mathbf{y}|\mathbf{x})}. \tag{3}$$

Inverse problems have a long history in statistics and applied mathematics, and the posterior distribution in (3) as well as related penalized estimators are the backbone of rigorous inference Bissantz et al. (2007); Stuart (2010); Knapik et al. (2011); Blanchard & Mücke (2018); Rastogi et al. (2020). When valid information are available on $\mathbf{x}$ and appropriately encoded in μ_0, the posterior distribution π_{μ_0} can enjoy better statistical properties than say, the minimizer of $\mathbf{x} \mapsto f(\mathbf{y}|\mathbf{x})$. However, finding such good prior distributions is often very challenging in many applications.

1.1. **Learning to solve inverse problems.**

In a growing number of settings, particularly in image restoration tasks, researchers have access to datasets in which the latent variable $\mathbf{x}$ and the related observation $\mathbf{y}$ are both observed. Indeed such datasets can often be simulated in settings where f is known. Hence, suppose that we have a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i),\ 1 \le i \le n\}$ of i.i.d. samples, such that for $1 \le i \le n$,

$$\mathbf{x}_i \sim \mu, \quad \text{and} \quad \mathbf{y}_i \mid \mathbf{x}_i \sim e^{-f(\mathbf{y}|\mathbf{x}_i)}\mathrm{d}\mathbf{y}, \text{ and where } \mu(\mathrm{d}\mathbf{x}) = \frac{1}{c_\mu}e^{-\mathcal{R}(\mathbf{x})}\mathrm{d}\mathbf{x}, \tag{4}$$

for some function $\mathcal{R} : \mathbb{R}^{d_x} \to \mathbb{R}$, and a normalizing constant c_μ. Hence under (4), μ is the marginal distribution of the latent variables. The conditional distribution of $\mathbf{x}_i$ given $\mathbf{y}_i$ is then given by

$$\pi(\mathrm{d}\mathbf{x}|\mathbf{y}_i) \propto \exp\left(-\mathcal{R}(\mathbf{x}) - f(\mathbf{y}_i|\mathbf{x})\right)\mathrm{d}\mathbf{x},$$

and its modal value is given by the function $g:\ \mathbb{R}^{d_y} \to \mathbb{R}^{d_x}$ with

$$g(\mathbf{y}) \overset{\mathrm{def}}{=} \operatorname*{\mathsf{Argmin}}_{\mathbf{x}\in\mathbb{R}^{d_x}} \left[f(\mathbf{y}|\mathbf{x}) + \mathcal{R}(\mathbf{x})\right]. \tag{5}$$

We will assume below that $g(\mathbf{y})$ is uniquely defined. We stress again that the distribution μ in (4) is not a prior distribution of $\mathbf{x}$ as selected by the researcher, but the actual marginal distribution of $\mathbf{x}$ unknown to the researcher. Hence g and $\pi(\cdot|\mathbf{y})$ are typically unknown. In fact, one of the key challenges in inverse problems is building a prior distribution μ_0 that is as close as possible to μ so that the resulting posterior distribution as given in (3) approximates well the corresponding conditional distribution.

In keeping with the assumption that $\mathbf{x}$ possesses additional structures, in many inverse problems the support of the marginal distribution μ lays in a much smaller (but unknown) subspace of $\mathbb{R}^{d_x}$. As a result of such marginal distribution concentration, it is often the case that the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ is also tightly concentrated around $g(\mathbf{y})$, in the sense that

$$\mathbf{x}_i = g(\mathbf{y}_i) + \boldsymbol{\xi}_i,\ \text{where } \mathbb{E}(\boldsymbol{\xi}_i \mid \mathbf{y}_i) \approx 0,\ 1 \le i \le n. \tag{6}$$

The representation (6) makes clear that in such settings where we have an informative (but unknown) marginal distribution, and given a dataset $\mathcal{D}$, one can learn the function g by regressing $\mathbf{x}$ on $\mathbf{y}$. In other words, we can learn to solve directly the inverse problem by regression using the dataset $\mathcal{D}$. The approach has become popular in computational imaging (Burger et al. (2012); Xie et al. (2012); Lucas et al. (2018); Yang et al. (2016); Ravishankar et al. (2017); Aggarwal et al. (2017); Chun & Fessler (2018); Zhang et al. (2017); Liu et al. (2019); Li et al. (2020)). A remarkable contribution of this literature is a number of specific deep neural network architectures generally called algorithm unrolling networks that leverage the structure of the forward model (Gregor & LeCun (2010); Sreter & Giryes (2018); Sulam et al. (2020); Tolooshams et al. (2020)), see also the reviews (Ongie et al. (2020); Shlezinger et al. (2021); Monga et al. (2021)).

However a fundamental question that has not been addressed in the literature so far is how well one can estimate the function g using these unrolling-based deep neural network architectures.

1.2. **Main contributions.**

To address this problem, and assuming that the data generating process (6) holds, we consider the nonparametric regression model

$$\mathbf{x}_i = g_W(\mathbf{y}_i) + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{7}$$

with regression errors $\boldsymbol{\epsilon}_i \overset{i.i.d.}{\sim} \mathbf{N}(0, \sigma^2 I_{d_x})$, for some positive variance parameter σ^2 taken as known for simplicity, and for a function class $\{g_W,\ W \in \mathcal{W}\}$, where $g_W : \mathbb{R}^{d_y} \to \mathbb{R}^{d_x}$ is a gradient descent neural network (GDN) function obtained by unrolling D' times a parametrized proximal gradient descent algorithm for solving (5). We give precise definition below. The architecture thus makes explicit use of the forward map f. We develop a sparse Bayesian framework for estimating (7) using a spike-and-slab prior distribution on the parameter W, and we analyze the statistical performance of the resulting estimator of g. We focus on the setting where the function $\mathcal{R}$ is convex, but not necessarily differentiable. Under some additional regularity conditions, and ignoring logarithmic terms, we show that GDN for estimating g achieves the statistical error rate

$$C_1(D') \times n^{-\frac{1}{2+d_x}},$$

for some constant $C_1(D')$ that depends on the unrolling depth D', but also on the input dimensions d_x, d_y. Keeping dimensions and the number of unrolling D' fixed, the result implies for example that when $d_y \geq d_x$, the GDN architecture, by making explicit use of the forward model, achieves a better rate than the minimax rate $C_2 n^{-\frac{1}{2+d_y}}$ for estimating $g : \mathbb{R}^{d_y} \to \mathbb{R}^{d_x}$ viewed as a Lipschitz function. We note however that these constants $C_1(D'), C_2$ can depend poorly[1] on the dimensions d_x, d_y.

The convergence rate of the estimator can be faster than the aforementioned rate. Indeed, we also show that when the proximal map of $\mathcal{R}$ is simple and can be well-approximated by a simple neural network function, then the GDN architecture achieves a faster statistical rate. For instance, if μ is as in (20) below, a common assumption in image restoration tasks, then ignoring log terms, our result shows that the GDN achieves the parametric rate $C_3 \times D'/\sqrt{n}$, for some dimension-dependent constant C_3. Importantly, our result thus suggests that the statistical performance of a GDN unrolled at depth D' deteriorates as D' increases, implying an overfitting phenomenon. Although we do not have a matching lower bound theory to confirm this overfitting phenomenon, we have performed extensive numerical experiments that all show an overfitting of the model as D' increases.

One of the practical challenges in building a GDN is the lack of theoretical guidelines in the choice of the depth D'. An offshoot of our theoretical analysis is the derivation that the best performance of a GDN is achieved by scaling the network depth as $D' \sim \log(n)/\log(\varrho_n^{-1})$, where ϱ_n is the convergence rate of the proximal gradient algorithm for solving (5).

[1] The poor dependence on the input dimensions is not specific to our work, and is rooted in the current state of knowledge in deep learning approximation theory (see e.g. Yarotsky (2017))

1.3. **Related work.** Most of the existing theoretical results on algorithm unrolling have studied the approximation capability of the resulting function class in the linear case. For instance Chen et al. (2018) studied the capability of the GDN function class to recover directly the signal $\mathbf{x}$ in the linear model (2). Gilton et al. (2020) proposed a novel unrolling architecture based on the Neumann series identity, and studied its approximation capability in the noiseless version of the linear model (2). To the best of our knowledge, our work is the first to analyze the statistical properties of algorithm unrolling in a way that accounts for both its approximation capability *and* its complexity.

Several prior works have also considered the statistical complexity of other deep learning models using a similar nonparametric regression setting where regularization is explicitly introduced to control model complexity Barron & Klusowski (2018); Schmidt-Hieber (2020); Taheri et al. (2021); Ee et al. (2020). Our framework is closer to Polson & Ročková (2018) and employs a Bayesian approach. However none of these results can be directly applied to algorithm unrolling architectures. Another unique feature of our framework that is worth emphasizing is that it produces posterior distributions that are computationally tractable using the sparse asynchronous SGLD of (AtchadÃ© & Wang (2021)).

Finally we contrast our nonparametric regression approach with the two-step approach proposed for instance by Chang et al. (2017), where the proximity operator of $\mathcal{R}$ is first estimated from the dataset $\mathcal{D}$, and g is then estimated by solving (5) using the estimated proximal operator obtained from the first step. The strategy seems statistically sub-optimal because the estimation of the proximal operator requires estimating the density of μ in general, which is a fundamentally more difficult statistical problem (Samworth (2018)). However it is conceivable that an adaptive density estimation method may exist that achieve better rates for densities with simple proximal maps, thus matching the approach developed here. More research is needed on this topic.

1.4. **Outline of the paper.** The remainder of the paper is organized as follows. We close this introduction with some general notations. The main results are described in Section 2. The results are obtained using a more general Bayesian posterior contraction result of independent interest that we described in Section 4. Some supporting numerical illustrations are presented in Section 3. All the proofs are postponed to Appendix A.

1.5. **Notations.** We define the sub-Gaussian norm of a probability measure ν on $\mathbb{R}^d$ with expected value m as the smallest constant c for which the following holds

$$\int_{\mathbb{R}^d} e^{\langle u, z-m\rangle}\nu(\mathrm{d}z) \leq e^{\frac{c^2\|u\|_2^2}{2}}, \quad \text{for all } u \in \mathbb{R}^d.$$

If Z is a random variable with distribution ν, we write $\|Z\|_{\psi_2}$ to denote the sub-Gaussian norm of ν. We note that this definition applies also to conditional densities, and we write $\|Z|X\|_{\psi_2}$ to denote the sub-Gaussian norm of the conditional distribution of Z given X.

Throughout the paper the notation $a \lesssim b$ means that $a \leq cb$, for some constant c that does not depend on the sample size n.

1.5.1. *Vectorization.* Let $\{h_W,\ W \in \mathcal{W}\}$ denote a generic deep neural network class of function where $h_W:\ \mathbb{R}^{p_0} \to \mathbb{R}^{p_D}$, with parameter $W = (W_D, \ldots, W_1) \in \mathcal{W} \overset{\text{def}}{=} \mathbb{R}^{p_D \times p_{D-1}} \times \cdots \times \mathbb{R}^{p_1 \times p_0}$. By vectorization, we will view $\mathcal{W}$ as the Euclidean space $\mathbb{R}^q$ (where $q \overset{\text{def}}{=} \sum_{\ell=1}^{D} p_\ell p_{\ell-1}$), and we will use a generic notation $\|\cdot\|_2$ to denote its Euclidean norm. Similarly, we will write $\|W\|_0$ (resp. $\|W\|_\infty$) to denote the number of non-zeros components of W (resp. the largest absolute value of the components of W). For any $1 \leq \ell \leq D$, we will similarly view W_ℓ as a vector element of $\mathbb{R}^{p_\ell p_{\ell-1}}$, and define similarly $\|W_\ell\|_2$, $\|W_\ell\|_0$ and $\|W_\ell\|_\infty$. Hence, in what follows, for a matrix M, $\|M\|_2$ will always denote the Frobenius norm of M, not its spectral norm. We will write the spectral norm as $\|\cdot\|_{\mathsf{op}}$.

2. Learning to solve inverse problems

Summarizing the introductory discussion on the data generating process, we make the following assumption.

H 1. *We have a data set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i),\ 1 \leq i \leq n\}$ of i.i.d. samples generated according to (4) such that for $1 \leq i \leq n$,*

$$\mathbf{x}_i = g(\mathbf{y}_i) + \boldsymbol{\xi}_i, \quad \textit{where} \quad \mathbb{E}(\boldsymbol{\xi}_i \mid \mathbf{y}_i) = 0,$$

for some independent error terms $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)$. Furthermore we assume that each $\boldsymbol{\xi}_i$ is a conditionally sub-Gaussian random vector given $\mathbf{y}_i$, with a non-random sub-Gaussian norm $\sigma_i < \infty$.

Remark 1. Assumption H1 formalizes the discussion in the introduction on the concentration of the conditional distribution of $\mathbf{x}_i$ given $\mathbf{y}_i$. As expanded upon in the introduction, this assumption is conceptually justified in seetings where the latent variable $\mathbf{x}$ has additional structures, and the marginal distribution of $\mathbf{x}$ is concentrated on a low-dimensional subset of $\mathbb{R}^{d_x}$. Checking H1 is similar to establishing a

Bernstein-von Mises theorem for the conditional distribution of $\mathbf{x}_i$ given $\mathbf{y}_i$, a challenging problem that is beyond the scope of this work (Vaart (1998); Bickel & Kleijn (2012); Nickl (2017); Nickl & Sohl (2019)).

Let ς_i denote the conditional sub-Gaussian norm of $\|\boldsymbol{\xi}_i\|_2$ given $\mathbf{y}_i$. The conditional sub-Gaussian assumption on $\boldsymbol{\xi}_i$ imposed in Assumption 1 implies that $\varsigma_i < \infty$ (see e.g. Theorem 3.1.1 of Vershynin (2018)). Throughout we set

$$\bar{\sigma} \overset{\text{def}}{=} \max_{1\leq i\leq n} \sigma_i, \quad \text{and} \quad \bar{\varsigma} \overset{\text{def}}{=} \max_{1\leq i\leq n} \varsigma_i.$$

2.1. **Gradient descent networks.** We consider the nonparametric regression (7), where $\{g_W,\ W \in \mathcal{W}\}$ is a gradient descent network (GDN) function class that we now define. First we introduce a generic feed-forward deep neural network function $H_W:\ \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$. Let $D > 0$ be the depth of the network. Let $(p_D, \ldots, p_0)$ be a sequence of integers representing the sizes of the layers of the network, with $p_0 = d_x$, and $p_D = d_x$. For $1 \leq \ell \leq D$, let $\mathsf{a}_\ell:\ \mathbb{R}^{p_\ell} \to \mathbb{R}^{p_\ell}$ be activation functions that we assume Lipschitz: for all $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{p_\ell}$,

$$\mathsf{a}_\ell(\mathbf{0}) = \mathbf{0}, \quad \text{and} \quad \|\mathsf{a}_\ell(\mathbf{z}_1) - \mathsf{a}_\ell(\mathbf{z}_2)\|_2 \leq \|\mathbf{z}_1 - \mathbf{z}_2\|_2. \tag{8}$$

For $B \in \mathbb{R}^{p_\ell \times p_{\ell-1}}$, we set

$$\Psi_B^{(\ell)}(\mathbf{z}) \overset{\text{def}}{=} \mathsf{a}_\ell(B\mathbf{z}), \quad \mathbf{z} \in \mathbb{R}^{p_{\ell-1}}. \tag{9}$$

With parameter $W = (W_D, \ldots, W_1)$, where $W_\ell \in \mathbb{R}^{p_\ell \times p_{\ell-1}}$, we consider the function $H_W : \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$ defined as

$$H_W(\mathbf{x}) = \Psi_{W_D}^{(D)} \circ \cdots \circ \Psi_{W_1}^{(1)}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^{d_x}, \tag{10}$$

where $f \circ g$ is the composition of f with g.

Remark 2. Feed-forward deep neural network models are usually written with additional bias terms (that is, by defining $\Psi_B^{(\ell)}(\mathbf{z})$ as $\mathsf{a}_\ell(B\mathbf{z}+\mathbf{b})$). However our formulation incurs no loss of generality, since these bias parameters can always be subsumed into the matrix B, by appropriately enlarging B and adding an intercept to the input.

Given $\gamma > 0$, we use the function H_W to approximate the proximal map of $\gamma\mathcal{R}$ (where $\mathcal{R}$ is as in (4)), defined as

$$\operatorname{Prox}^{\gamma\mathcal{R}}(\mathbf{x}) \overset{\text{def}}{=} \underset{\mathbf{u}\in\mathbb{R}^{d_x}}{\mathsf{Argmin}} \left[\gamma\mathcal{R}(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2\right].$$

Given a step-size $\gamma > 0$, $W \in \mathcal{W}$, and $\mathbf{y} \in \mathbb{R}^{d_y}$ we thus define the function $F_{\mathbf{y},W}$: $\mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$ by

$$F_{\mathbf{y},W}(\mathbf{x}) \overset{\text{def}}{=} H_W\left(\mathbf{x} - \gamma\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})\right).$$

Given $D' \geq 1$ (the depth of the network), we consider the function g_W defined as

$$g_W(\mathbf{y}) \overset{\text{def}}{=} \underbrace{F_{\mathbf{y},W} \circ \cdots \circ F_{\mathbf{y},W}}_{D' \text{ times}}(\mathbf{x}^{(0)}), \tag{11}$$

for some initial value $\mathbf{x}^{(0)} \in \mathbb{R}^{d_x}$. We note that in addition to W, the function g_W depends also on the step-size γ, the depth D', and the initial value $\mathbf{x}^{(0)}$. The network architecture in (11) is the so-called (proximal) gradient descent network (GDN), and belong to the class of so-called algorithm unrolling (or unfolding) deep learning models, where a statistical model is built by iterating an optimization algorithm. Many variations have been proposed in the literature based on various other optimization schemes (we refer the reader to the references in the introduction).

For $\mathbf{x} \in \mathbb{R}^{d_x}$, and $\mathbf{y} \in \mathbb{R}^{d_y}$, we set

$$F_{\mathbf{y}}(\mathbf{x}) \overset{\text{def}}{=} \text{Prox}^{\gamma\mathcal{R}}\left(\mathbf{x} - \gamma\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})\right).$$

Looking at the definition of (11), it is clear that under appropriate convexity assumptions and for well-selected step size γ, the convergence of $F_{\mathbf{y}}^j(\mathbf{x})$ toward $g(\mathbf{y})$ is guaranteed, where h^j denotes the composition of h, j times. Therefore, if for some W, $H_W \approx \text{Prox}^{\gamma\mathcal{R}}$, then we can expect $g_W \approx g$ for D' sufficiently large, by standard convex optimization theory. As a result, the function class $\{g_W,\ W \in \mathcal{W}\}$ typically has good skills in approximating g. We impose next the necessary assumptions for the intuition above to hold.

H2. *(1) The function $\mathcal{R} :\ \mathbb{R}^{d_x} \to \mathbb{R}$ is convex. There exists M such that for all $\mathbf{y} \in \mathbb{R}^{d_y}$, the function $\mathbf{x} \mapsto f(\mathbf{y}|\mathbf{x})$ is convex, differentiable, and with a M-Lipschitz gradient. Furthermore, the step-size γ satisfies $0 < \gamma \leq M^{-1}$, and for all $\mathbf{y} \in \mathbb{R}^{d_y}$, $g(\mathbf{y})$ is uniquely defined.*

(2) There exist $R_0 < \infty$, $\varrho_n \in [0, 1)$ such that for all $k \geq 1$,

$$\max_{1 \leq i \leq n} \|F_{\mathbf{y}_i}^k(\mathbf{x}^{(0)}) - g(\mathbf{y}_i)\|_2 \leq R_0 \varrho_n^k.$$

Remark 3. Assumption H2-(1) is a standard set up for proximal gradient descent (Parikh & Boyd (2013)). Assumption H2-(2) is stronger and imposes a linear convergence rate. It is well-known to hold for strongly convex problems. For problems where a local linear convergence holds, H2-(2) can also be shown to follow from H2-(1) when $\mathbf{x}^{(0)}$ is close enough to the solution. For instance it is known that such local linear convergence of proximal gradient descent holds for the lasso problem (Tao et al. (2016)). The main challenge to go beyond the linear rate is the fact that sublinear rates are typically expressed in terms of the function value, not in terms of the parameter as needed here.

We also impose the following assumption that models the approximation of the proximal map $\mathrm{Prox}^{\gamma\mathcal{R}}$.

H3. *There exist $\beta_1, \beta_2 \geq 0$, such that for all $\epsilon \in (0,1)$ we can construct a feed-forward deep neural network H_W, as in (10), with depth $1 \leq D \leq D_0 \log(\sqrt{d_x}/\epsilon)$, maximum layer size no larger than $N_0 \left(\sqrt{d_x}/\epsilon\right)^{\beta_1}$, maximum parameter absolute value $\|W\|_\infty$ no larger than 1, and maximum sparsity $\|W\|_0$ no larger than $s_0 \left(\sqrt{d_x}/\epsilon\right)^{\beta_2}$, for constants D_0, N_0, s_0 that do not depend on ϵ such that for all $R < \infty$,*

$$\sup_{\mathbf{x}\colon \|\mathbf{x}\|_2 \leq R} \left\|H_W(\mathbf{x}) - \mathrm{Prox}^{\gamma\mathcal{R}}(\mathbf{x})\right\|_2 \leq \epsilon. \tag{12}$$

Furthermore, there exists $R_1 < \infty$ such that with the constructed network H_W,

$$\max_{j\geq 1} \max_{1\leq i\leq n} \|F^j_{\mathbf{y}_i,W}(\mathbf{x}^{(0)})\|_2 \leq R_1. \tag{13}$$

Remark 4. Since $\mathbf{x} \mapsto \mathrm{Prox}^{\gamma\mathcal{R}}(\mathbf{x})$ is a Lipschitz map, we can always invoke classical deep learning approximation theory for smooth functions (see e.g. Schmidt-Hieber (2020); DeVore et al. (2021)) to conclude that Assumption 3 holds with $\beta_1 = \beta_2 = d_x$. However better approximation is possible if $\mathrm{Prox}^{\gamma\mathcal{R}}$ is a simple map.

The condition in (13) is a technical assumption that simplifies the mathematical analysis. It can be automatically enforced by adding a layer-normalization layer in H_W (Ba et al. (2016)).

2.2. **Bayesian inference using spike-and-slab priors.** We consider the problem of fitting model (7), where $\{g_W,\ W \in \mathcal{W}\}$ is the GDN function class constructed in (11). The parameter space is $\mathcal{W} \overset{\text{def}}{=} \mathbb{R}^{p_D\times p_{D-1}} \times \cdots \times \mathbb{R}^{p_1\times p_0}$. As indicated at the end of the introduction, at times we shall view $\mathcal{W}$ as the Euclidean space $\mathbb{R}^q$, where

$$q \overset{\text{def}}{=} \sum_{\ell=1}^{D} (p_\ell \times p_{\ell-1}).$$

Our initial motivation in this work comes from inverse problems in remote sensing. It was therefore important for us to analyze a statistical procedure that can be implemented in practice. An important shortcoming of the current statistical theory of deep learning models under sparsity constraints (Barron & Klusowski (2018); Schmidt-Hieber (2020); Taheri et al. (2021); Ee et al. (2020)) is the lack of computational tractability of the resulting estimators. To address this issue we propose to fit the model $\{g_W,\ W \in \mathcal{W}\}$ in a Bayesian framework using a spike and slab prior (Atchade & Bhattacharyya (2018)). To that end, we introduce a sparsity structure parameter $\Lambda = (\Lambda_D, \ldots, \Lambda_1) \in \mathcal{S} \overset{\text{def}}{=} \{0,1\}^{p_D\times p_{D-1}} \times \cdots \times \{0,1\}^{p_1\times p_0}$ that encodes the support of W. We assume that Λ has a prior distribution given by

$$\Pi_0(\Lambda) \propto \left(\frac{1}{q}\right)^{(\mathsf{u}+1)\|\Lambda\|_0}, \quad \Lambda \in \mathcal{S}, \tag{14}$$

for some parameter $\mathsf{u} \geq 1$. This prior corresponds to the assumption that the entries of Λ are independent Bernoulli random variables $\mathbf{Ber}((1+q^{\mathsf{u}+1})^{-1})$. Given Λ we assume that the entries of W are conditionally independent with joint density

$$\Pi_0(W|\Lambda) \quad = \quad \prod_{\ell=1}^{D} \prod_{(i,k):\, \Lambda_{\ell,i,k}=1} \sqrt{\frac{\rho_1}{2\pi}} e^{-\frac{\rho_1}{2} W_{\ell,ik}^2} \prod_{(i,k):\, \Lambda_{\ell,k}=0} \sqrt{\frac{\rho_0}{2\pi}} e^{-\frac{\rho_0}{2} W_{\ell,i,k}^2}, \tag{15}$$

for some parameters $0 < \rho_1 < \rho_0$. Throughout the paper, and without further notice we set

$$\rho_1 = 1. \tag{16}$$

The variance parameter ρ_0 can be chosen fairly arbitrarily. However, in order to ease MCMC sampling from the resulting posterior distribution it is crucial to choose ρ_0 small, of order $1/n$. We refer the reader to (Atchade & Bhattacharyya (2018)) for further discussion. Using this prior distribution and the regression model (7), we consider the posterior distribution on $\Theta \overset{\text{def}}{=} \mathcal{S} \times \mathcal{W}$ with density given by

$$\Pi(\Lambda, W \mid \mathcal{D}) \propto \Pi_0(\Lambda, W) \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \|\mathbf{x}_i - g_{W\odot\Lambda}(\mathbf{y}_i)\|_2^2\right), \tag{17}$$

where $W \odot \Lambda$ denotes the component-wise product of W and Λ. To use this posterior distribution we draw sample $(\Lambda, W) \sim \Pi(\cdot|\mathcal{D})$, and use $g_{\Lambda\odot W}$ as inversion map. Since Λ is typically sparse under Π, $g_{\Lambda\odot W}$ is a sparse GDN. For $h : \mathbb{R}^{d_y} \to \mathbb{R}^{d_x}$, we set

$$\|h\|_n \overset{\text{def}}{=} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \|h(\mathbf{y}_i)\|_2^2}.$$

Our goal is to derive a bound on $\|g_{\Lambda\odot W} - g\|_n$, when $(\Lambda, W) \sim \Pi(\cdot|\mathcal{D})$.

Theorem 5. *Assume H1-H3. Consider the nonparametric regression (7) for estimating g, where the function class $\{g_W,\ W \in \mathcal{W}\}$ is as defined in (11), and the regression variance parameter σ satisfies $\sigma \geq \bar{\sigma}$. Then for all q large enough, and $n \geq \sigma^2 \log(p)$, we can construct a function class $\{H_W,\ W \in \mathcal{W}\}$, such that at unrolling depth D' that satisfies*

$$D' \gtrsim \frac{\log(n)}{-\log(\varrho_n)},$$

the posterior distribution $\Pi(\cdot|\mathcal{D})$ *in (17) satisfies*

$$\Pi\left(\|g_{\Lambda\odot W} - g\|_n > M\bar{\sigma}\frac{(D')^{1+\frac{\beta_2}{2}}}{n^{\frac{1}{2+\beta_2}}} \mid \mathcal{D}\right) \leq \frac{12}{q}, \tag{18}$$

with probability at least $1 - e^{-c_1 n} - \frac{c_1}{q}$, *for some absolute constant* c_1, *and a constant* $M \lesssim (\log(q))^{1/(2+\beta_2)} \log(n)^{3/2}$.

Proof. See Section A.2. □

We make several remarks here. (a) In contrast to common practice where D' is often chosen on an ad-hoc manner, Theorem 5 recommends carefully scaling the depth parameter D' as

$$D' \sim -\log(n)/\log(\varrho_n),$$

for optimal performance. (b) The expression of the rate in (18) suggests that the statistical performance of GDN unrolled at depth D' deteriorates as D' increases, implying an overfitting phenomenon. Although we do not have a matching lower bound theory to confirm this overfitting phenomenon, we have performed several numerical experiments that all show an overfitting of the model as D' increases. (c) Algorithm unrolling allows researchers to build deep neural network architectures that exploit the structure of the problem. Are those architecture provably better than off-the-shelves architectures that do not make use of the forward problem? Our results shed some light on this question. In the setting of H2, the function g of interest is at best Lipschitz[2]. Therefore the minimax rate in the estimation of g in a nonparametric regression setting without further knowledge on the structure of the problem is

$$C_2 n^{-\frac{1}{2+d_y}}.$$

We can invoke classical deep learning approximation theory (see e.g. Yarotsky (2017); Schmidt-Hieber (2020); DeVore et al. (2021)) to conclude that H3 holds with $\beta_1 = \beta_2 = d_x$. In that case, up to log-terms, we deduce from Theorem 5 that GDN achieves the convergence rate

$$C_1 n^{-\frac{1}{2+d_x}}.$$

Hence, Theorem 5 implies that in inverse problems where d_y is larger than d_x, the unrolling framework has a better convergence rate than the minimax rate of estimating g from the data $\mathcal{D}$ in a nonparametric regression. However Theorem 5 has some limitations. Firstly, the constants C_1, C_2 in the rates posted above depend

[2] Indeed, it can be easily shown that if for all $\mathbf{y} \in \mathsf{Y}$, $\mathbf{x} \mapsto f(\mathbf{y}|\mathbf{x})$ is strongly convex with strong convexity parameter $\underline{\kappa}$, and $\mathbf{x} \mapsto \nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})$ is $\bar{\kappa}$ Lipschitz then $\mathbf{y} \mapsto g(\mathbf{y})$ is L_g-Lipschitz with $L_g \leq 2\bar{\kappa}/\underline{\kappa}$

on d_x and d_y in ways that are poorly understood. This comes from the scalings of constants in current deep neural network approximation theory Yarotsky (2017); Schmidt-Hieber (2020). Another limitation of current minimax rates is the fact that deep learning models can often adapt to additional properties of the function of interest and converge much faster than the theoretical minimax rate. For instance Schmidt-Hieber (2020) shows that FNN models achieves faster rate in the estimation of compositional functions. We give a similar example below.

(d) The use of the empirical norm $\|u\|_n = \sqrt{\sum_{i=1}^n u(\mathbf{y}_i)^2}$ instead of the L^2 population norm of $\mathbf{y}$ in (18) is another limitation of our result, although this is a fairly common practice in nonparametric estimation, and does not fundamentally change the resulting contraction rate. More technically, working in the L^2 norm amounts to the additional control of the term

$$\sup_{W\in\widetilde{W}^{(j)}} \left| n^{-1}\sum_{i=1}^n (g_W(\mathbf{y}_i) - g(\mathbf{y}_i))^2 - \|g_W - g\|_2^2 \right|, \tag{19}$$

in Lemma D.5. Because the sup in (19) is taken over well behaved sets $\widetilde{W}^{(j)}$, this uniform deviation can be controlled using standard tools as in Wainwright (2019) Chapter 14, but would require additional assumptions on the marginal distribution of $\mathbf{y}$ that we wish to avoid making.

2.2.1. *Application to sparse marginal distributions.* We give another application of Theorem 5 where the posterior predictive function obtained from the GDN achieves the parametric rate. When dealing with images, several authors such as Beck & Teboulle (2010); Dong et al. (2011) have argued that natural image data are often sparse after linear transformation (such as a difference operators, or wavelet transforms), and suggested modeling the marginal distribution μ as

$$\mu(\mathrm{d}\mathbf{x}) = \frac{1}{c_\mu} e^{-\mathcal{R}_0(B\mathbf{x})}\mathrm{d}\mathbf{x}, \tag{20}$$

for some simple sparsity inducing function $\mathcal{R}_0$, and a non-singular matrix $B \in \mathbb{R}^{d_x\times d_x}$. In other words, $\mathcal{R}(\mathbf{x}) = \mathcal{R}_0(B\mathbf{x})$. A common choice is $\mathcal{R}_0(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$ or $\mathcal{R}_0(\mathbf{x}) = \lambda\|\mathbf{x}\|_1 + \lambda_2\|\mathbf{x}\|_2/2$, for parameters $\lambda, \lambda_1, \lambda_2 \geq 0$. If B is an orthogonal matrix, and $\mathrm{Prox}^{\gamma\mathcal{R}_0}$ denotes the proximal operator of $\mathcal{R}_0$, then by proximal calculus (see e.g. Lemma 2.8 of Combettes & Wajs (2005)), we have

$$\mathrm{Prox}^{\gamma\mathcal{R}}(\mathbf{x}) = B^{-1}\,\mathrm{Prox}^{\gamma\mathcal{R}_0}(B\mathbf{x})\,. \tag{21}$$

For example, given $\lambda_1 > 0, \lambda_2 \geq 0$, suppose that $\mathcal{R}_0$ is the elastic-net regularization prior of Zou & Hastie (2005) given by

$$\mathcal{R}_0(\mathbf{x}) = \lambda_1\|\mathbf{x}\|_1 + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2. \tag{22}$$

Then the proximal of $\gamma\mathcal{R}_0$ is $\mathrm{Prox}^{\gamma\mathcal{R}_0}(\mathbf{x}) = (\mathsf{s}_\gamma(x_1), \cdots \mathsf{s}_\gamma(x_{d_x}))^{\mathsf{T}}$, where

$$\mathsf{s}_\gamma(x) = \mathsf{ReLu}\left(\frac{x - \gamma\lambda_1}{1 + \gamma\lambda_2}\right) - \mathsf{ReLu}\left(\frac{-x - \gamma\lambda_1}{1 + \gamma\lambda_2}\right),$$

and where $\mathsf{ReLu}(t) \overset{\text{def}}{=} \max(t, 0)$. Therefore, $\mathrm{Prox}^{\gamma\mathcal{R}_0}(\mathbf{x})$ can be represented exactly using a 2-layer ReLu neural network with layer sizes $(d_x, 2d_x, d_x)$, and $\mathrm{Prox}^{\gamma\mathcal{R}}(\mathbf{x})$ can be represented exactly using a 4-layer ReLu neural network with layer sizes $(d_x, d_x, 2d_x, d_x, d_x)$. Hence, H3 holds with depth $D = 4$, $\beta_1 = \beta_2 = 0$. Furthermore, since $\mathcal{R}$ is strongly convex, if we focus on the linear regression model and take the forward model as in (2), then H2 holds. Hence Theorem 5 yields the following.

Corollary 6. *Suppose that H1 holds with f as in (2), and suppose that μ is as in (20) with some orthogonal matrix B, and $\mathcal{R}_0$ as in (22). Suppose also that $\sigma \geq \bar{\sigma}$. Then we can construct a deep learning function class $\{H_W,\ W \in \mathcal{W}\}$, with depth $D = 4$, such that at unrolling depth $D' \gtrsim -\log(n)/\log(\varrho_n)$ the posterior distribution $\Pi(\cdot|\mathcal{D})$ in (17) satisfies*

$$\Pi\left(\|g_{\Lambda\odot W} - g\|_n \geq \frac{M\bar{\sigma}D'}{\sqrt{n}} |\ \mathcal{D}\right) \leq \frac{12}{q},$$

with probability at least $1 - \frac{c_1}{q} - e^{-c_1 n}$, for some absolute constant c_1, where M depends on some log terms that we ignore.

3. Numerical illustration

We illustrate our theoretical results with a toy example, a simulation and a real data deblurring problem. For all examples we draw samples from the posterior distribution (17) using the Sparse Asynchronous Stochastic Gradient Langevin Dynamics (SA-SGLD) sampler of Atchade & Wang (202X), an approximate MCMC sampler designed for posterior distributions of the form (17), that employs asynchronicity for fast sampling. For more details on the approximate correctness of the sampler we refer the reader to Atchade & Wang (202X). Computationally the SA-SGLD sampler is implemented at the cost of 2 back-propagation through the GDN per MCMC iteration.

3.1. Learning the Elastic Net regression map.

In this section, we illustrate our theoretical results with the example of learning the elastic-net regularization to solve a linear regression model.

Data generation: Data generation: We generate a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i),\ 1 \leq i \leq n\}$ where $\mathbf{x}_i \overset{i.i.d.}{\sim} \mu$, and $\mathbf{y}_i|\mathbf{x}_i \sim \mathbf{N}(A\mathbf{x}_i, v^2\mathbf{I}_{d_y})$. The entries of the matrix A are generated independently from the standard normal distribution, and we set

$v^2 = 0.001$. We choose $\mu(\mathrm{d}\mathbf{x}) \propto e^{-\mathcal{R}_0(B\mathbf{x})}\mathrm{d}\mathbf{x}$ as in (20), where $\mathcal{R}_0$ is the elastic net density as in (22), and $B = \mathbf{I}_{d_x}$. We set $n = 200$, $d_x = d_y = 100$, and $\lambda_1 = 1$, $\lambda_2 = 1$.

Model architecture. We specify H_W as a FNN with depth $D = 2$, and layer $(p_0, p_1, p_2) = (d_x, 2d_x, d_x)$. To prevent overfitting, we specify the models to only learn the parameters that connect the nodes between layers in each dimension of $\mathbf{x}_i$, which reduces the number of model parameters to $q = 7 * d_x = 700$. We consider several values of the unrolling depth D': $D' = 1$ (GDN1), $D' = 5$ (GDN2), $D' = 10$ (GDN3), and $D' = 20$ (GDN4) for comparison[3].

Training details: For the Bayesian prior we choose $\rho_0 = n$, $\rho_1 = 1$, and $\mathsf{u} = 1$. We choose $\sigma = 0.001$ in (17), and run the SA-SGLD with a constant step-size $2 * 10^{-6}$ for GDN. The mini-batch size is set to 100. The initial value $\mathbf{x}^{(0)}$ of GDN is set to 0. The MCMC sampler is implemented in Pytorch on a high-performance computer with a Nvidia Tesla V100 GPU. We run the sampler for 10^4 iterations.

Evaluation procedure: For the comparison, we generate 1000 test samples and evaluate the prediction errors of the resulting estimator $g_{\Lambda\odot W}$, where $(\Lambda, W) \sim \Pi(\cdot|\mathcal{D})$. The performance of GDN are compared to the performance of g, which in this problem is easily calculated by proximal gradient descent. We do this comparison by computing the error

$$e(\Lambda, W) = \frac{1}{1000}\sum_{i=1}^{1000} \|g_{\Lambda\odot W}(\mathbf{y}_i) - g(\mathbf{y}_i)\|_2,$$

where the average is taking over the test sample. The boxplots in Figure 1 show the distributions of the errors obtained by taking 500 samples of (Λ, W) along the MCMC sampler. In this toy example, $D' = 10$ (GDN3) yields the best results and as predicted by our theory, GDN deteriorates as the unrolling depth increases.

3.2. **Illustration with a simulated data deblurring problem.** Image deblurring is a common inverse problem in computational imaging. Here for illustration, we start with a simulated dataset $\mathcal{D}$ for experimental purposes.

Data generation. We generate $n = 500$ images of size 16×16, that we then blurred using a Gaussian blurring convolution kernel with variance 3 without restriction on the kernel range[4].

[3] The step-size γ of GDN is taken as $\gamma_1 = \frac{2v^2}{\lambda_{\max}(A'A)}$ a, where $\lambda_{\max}(A'A)$ is the largest eigenvalue of $A'A$.

[4] Each image is 16×16 matrix partitioned into 4 blocks where the upper left and lower right are 8×8 diagonal matrices with diagonal elements sampled from $N(20, 0.5)$ and $N(-10, 0.1)$ respectively, where the upper right is a 8×8 matrix with entries sampled from $N(10, 0.1)$, and the lower left a 8×8 matrix with entries sampled from $N(-10, 5)$.

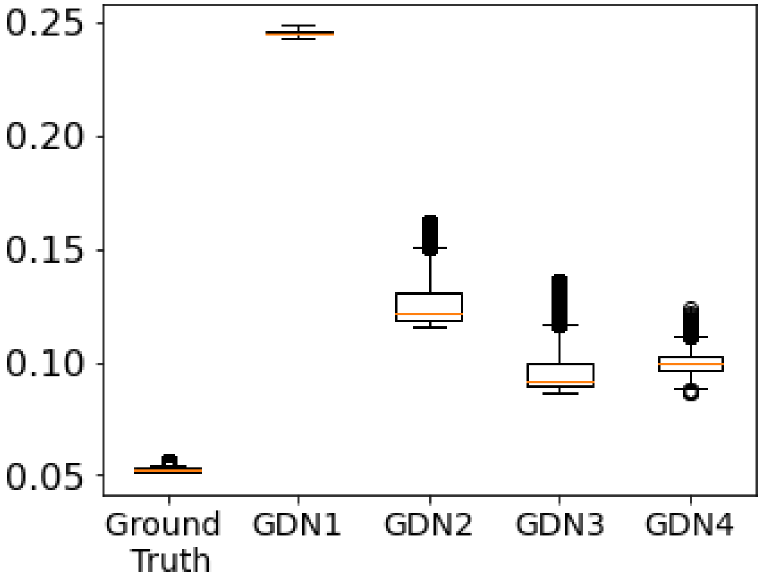

FIGURE 1. Distributions of prediction errors on test samples: GDN1 (D'=1), GDN2 (D'=5) and GDN3 (D'=10), GDN4 (D'=20).

Model architecture. We construct H_W in (10) as a 3-layer relu-convolutional neural network[5]. The total number of parameters is $q = 18,881$. We evaluate the GDN model at unrolling depth $D' = 2$ (GDN1), $D' = 4$ (GDN2), $D' = 12$ (GDN3) and $D' = 24$ (GDN4), and we do a comparison with two feedforward convolutional neural network (FNN) that do not make use of the forward problem. The first FNN has the same architecture as H_W (FNN1), while the second is a 6-layer FNN with total number of parameter $q = 136,641$ (FNN2)[6]. All the layers are padded to keep the image size constant.

Training details: For the Bayesian prior, we use $\rho_0 = n, \rho_1 = 1$, and $\mathsf{u} = 8000$. We choose $\sigma^2 = 0.01$ in (17), and run the SA-SAGLD with a constant step-size 2×10^{-8} for both FNN and GDN. The mini-batch size is set to 50 in both cases. The MCMC sampler are implemented on a Nvidia Tesla V100 GPU system with 384 GB GPU memory running Pytorch. We run both samplers for 10^4 iterations.

[5] consisting of 3 convolutional layers with respective sizes $3, 3, 1$, respective number of filters $32, 64, 1$. Each layer except the last layer is followed by a LayerNorm layer and a ReLu layer

[6] The 6-layer network consists of 2 convolutional layers, 1 channel-wise fully connected layer, 3 deconvolutional layers with respective sizes $5, 3, 2, 4, 5, 3$, respective number of filters $32, 64, 64, 64, 32, 1$. Each layer except the last layer is followed by a LayerNorm layer and a ReLu layer.

Evaluation procedure. We generate 500 test samples to evaluate the prediction errors of the six models. The boxplots in Figure 2 show the distribution of the mean square error (same as 3.1) of the last 2000 samples of (Λ, W) along the MCMC sampler of each model. Figure 3 shows an example of reconstruction from FNN1, FNN2, GDN1, and GDN3. We observe that GDN outperforms FNN1, and can achieve similar performance as FNN2 when the unrolling depth is appropriately selected (not too small, nor too large). The experiment again confirms the importance of scaling appropriately the unrolling depth as highlighted in our theoretical results.

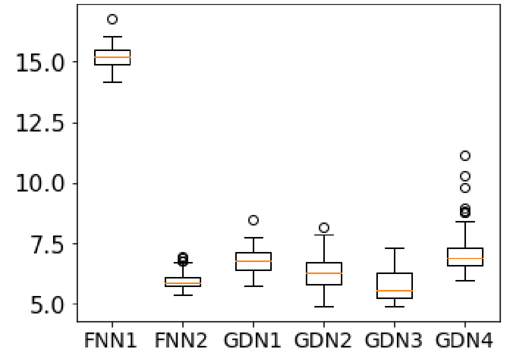

FIGURE 2. Test loss comparison between FNN1 (3 conv), FNN2 (3 conv + 1 cfc + 3 deconv), GDN1 ($D' = 2$), GDN2 ($D' = 4$), GDN3 ($D' = 12$) and GDN4 ($D' = 24$)

3.3. **Illustration with CelebA dataset.** We extend the last example to the deblurring of CelebA images Liu et al. (2015).

Data generation. We randomly select $20,000$ images from the celebA dataset that we resize to 64×64. We generate the corresponding observed measurements $\mathbf{y}_i$ through the linear forward model (1), where A is a Gaussian blurring convolution matrix with variance 6.25, and where $v^2 = 0.01$ leading to a highly ill-conditioned inverse problem.

Model architecture. We take H_W in (10) as a 3-layer relu-convolutional neural network[7]. The depth of the GDN is either $D' = 4$ (GDN1), $D' = 12$ (GDN2), or $D' = 24$ (GDN3). The total number of parameters in the same in all three cases and equal to $q = 267,777$. We compare this model with a feedforward architecture that

[7] with kernel sizes $4, 4, 2$, and the number of filters that we take here as $256, 256, 1$. All padded to keep image size constant

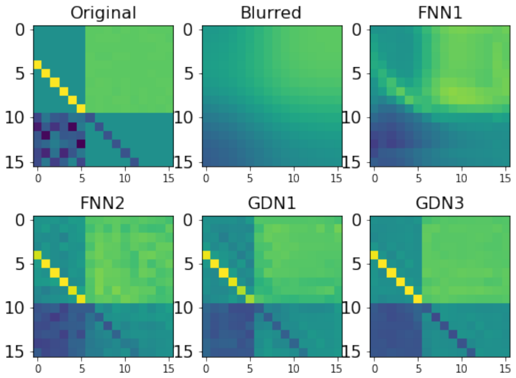

Figure 3. Reconstruction result from FNN1 (upper right), FNN2 (lower left), GDN1 $D' = 2$ (lower middle) and GDN3 $D' = 12$ (lower right)

doe not make use of the forward model[8]. The total number of parameter of the FNN is $q = 2,271,297$.

Training details: In the MCMC, the mini-batch size taken as $B = 164$, and the step-size is taken as $\gamma = 10^{-9}$ for FNN, GDN2, and GDN3, and $\gamma = 10^{-8}$ for GDN1. We run the algorithms for $80,0000$ iterations using a high-performance computier with a Nvidia Tesla V100 GPU running Matlab 2022a.

Evaluation procedure. Figure 4 shows the distributions of the test error from 500 drawn from the MCMC sampler after burn-in. We see again that at appropriate depth GDN matches FNN. However we see a decrease in performance at deeper depth $D' = 24$, which again suggests overfitting. Figure 5 shows three examples of reconstructed images.

4. A general Bayesian posterior contraction result

Theorem 5 is derived as special cases of a more general result of independent interest that we establish in this section. We consider again the regression model (7), where $\{g_W,\ W \in \mathcal{W}\}$ is some arbitrary deep neural network function class. We assume that the parameter space is $\mathcal{W} \overset{\text{def}}{=} \mathbb{R}^{p_D \times p_{D-1}} \times \cdots \times \mathbb{R}^{p_1 \times p_0}$, for some depth $D \geq 1$, and layer dimensions $p_0, p_1, \ldots, p_D \geq 1$. As indicated at the end of the introduction, at

[8] With 3 convolutional layers (with size $4, 4, 2$, filter number $256, 256, 32$, and strides $1, 2, 1$) followed by 4 corresponding deconvolutional layers (with kernel size $2, 4, 4, 2$, filter number $32, 256, 256, 1$, and strides $1, 2, 1, 1$).

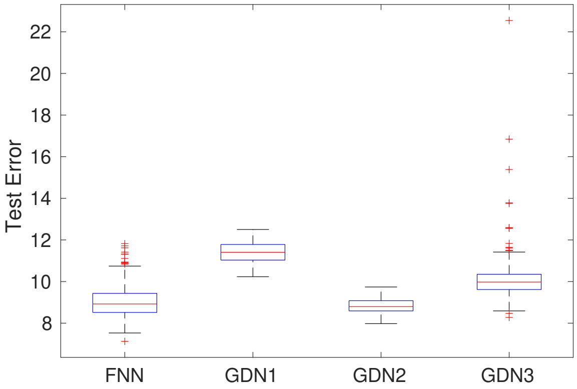

FIGURE 4. Test loss comparison between FNN (3 conv + 4 deconv), GDN1 (D'=4), GDN2 (D'=12) and GDN3 (D'=24)

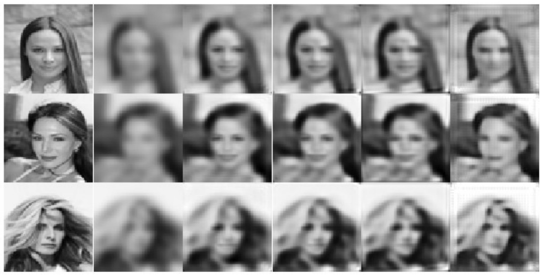

FIGURE 5. Some random examples of reconstructions. From left to right: CelebA image, blurred image, GDN1, GDN2, GDN3, and FNN

times we shall view $\mathcal{W}$ as the Euclidean space $\mathbb{R}^q$, with Euclidean norm denoted $\|\cdot\|_2$, where

$$q \overset{\text{def}}{=} \sum_{\ell=1}^{D} (p_\ell \times p_{\ell-1}).$$

We make the following local Lipschitz assumption on the function class.

H 4. *For all* $0 < \eta < \infty$*, there exists* $L(\eta) \geq 1$ *such that for all* $W, W' \in \mathcal{W}$ *that satisfy* $\max(\|W\|_2, \|W'\|_2) \leq \eta$*, and for all* $\mathbf{y} \in \mathsf{Y}$*, we have*

$$\|g_W(\mathbf{y}) - g_{W'}(\mathbf{y})\|_2 \leq L(\eta)\|W - W'\|_2. \tag{23}$$

The constant $L(\eta)$ is a local Lipschitz constant of the function $W \mapsto g_W(\mathbf{y})$. Controlling appropriately these local Lipschitz constants is a major theoretical challenges in dealing with deep neural networks.

Theorem 7. *Suppose that the dataset $\mathcal{D}$ is generated as in H1, and consider the nonparametric regression (7) for some function class $\{g_W,\ W \in \mathcal{W}\}$ that satisfies H4, and the corresponding posterior distribution (17). Suppose that the regression variance parameter σ satisfies $\sigma \geq \bar{\sigma}$. Let $\varpi_\star \geq 0$, $s_\star \geq 1$, be such that*

$$\min\ \{\|g_W - g\|_\infty,\ W \in \mathcal{W}\ \text{ s.t. }\ \|W\|_0 \leq s_\star,\ \ \|W\|_\infty \leq 1\} \leq \varpi_\star,$$

and set $\mathsf{L}_\star \stackrel{\text{def}}{=} L(2s_\star^{1/2})$, where the function L is as in H4. Define

$$s \stackrel{\text{def}}{=} \left(1 + \frac{\log(\mathsf{L}_\star\sqrt{n})}{\mathsf{u}\log(q)}\right) s_\star + \frac{4n\varpi_\star^2}{\sigma^2\mathsf{u}\log(q)}, \quad \text{and} \quad \mathsf{r} \stackrel{\text{def}}{=} \bar{\sigma}\sqrt{\frac{s\log(q) + s\log(\mathsf{L}_s)}{n}},$$

where

$$\mathsf{L}_s \stackrel{\text{def}}{=} L(s^{1/2}b_s), \quad \text{with} \quad b_s \stackrel{\text{def}}{=} \sqrt{2(1+\mathsf{u})(1+s)\log(q)}.$$

Then for all q large enough, and $n \geq \sigma^2\log(q)$, we can find a constant $M^2 \geq \mathsf{u}\max((\sigma/\bar{\sigma})^2, 1)$, and absolute constant c_1 such that

$$\Pi\left(\|g_{\Lambda\odot W} - g\|_n > M\mathsf{r}\,|\mathcal{D}\right) \leq \frac{12}{q}, \tag{24}$$

with probability at least $1 - e^{-c_1 n} - \frac{c_1}{q}$.

Proof. See Section A.1. □

Remark 8. Theorem 7 applies well beyond the GDN of interest in this work. For any function class $\{g_W,\ W \in \mathcal{W}\}$ trained under the proposed sparse spike-and-slab prior, one can read off the posterior contraction rate of $\Pi(\cdot|\mathcal{D})$ from Theorem 7. The rate is driven by the local Lipschitz constant $L(\eta)$ of the function class, and the relationship between $(s_\star, \beta_\star)$ and $\varpi_\star$, which captures the approximation capability of the function class.

4.1. **Sketch of the proof of Theorem 7.**

To improve readability we give here a high-level description of the proof of Theorem 7. Several approaches have been developed in the literature to study the contraction of posterior distributions. Here we follow an approach due to Shen & Wasserman (2001). The merit of their approach is that it makes a direct connection between the contraction properties of the posterior distribution and the properties of the corresponding log-likelihood empirical process.

Let f, $\{f_\theta,\ \theta \in \Theta\}$ be a family of densities on a measurable space Z equipped with a reference sigma-finite measure that we write as $\mathrm{d}z$. All densities considered on the sample space Z are defined with respect to $\mathrm{d}z$. The parameter space Θ is some arbitrary measurable space. Let π be a prior probability measure on Θ. We consider the posterior distribution of θ given by

$$\Pi(A|z) = \frac{\int_A f_\theta(z)\pi(\mathrm{d}\theta)}{\int_\Theta f_\theta(z)\pi(\mathrm{d}\theta)}, \quad A \text{ meas.}, \quad z \in \mathsf{Z}.$$

The next lemma is a generalization of Shen & Wasserman (2001), and summarizes the main arguments used in the proof of Theorem 7.

Lemma 9. *Let S, B and $\{\Xi_k,\ k \geq 1\}$ be measurable subsets of Θ, such that $S \cap B^c \subseteq \cup_{k\geq 1}\Xi_k$. Let $\beta > 0$, $\rho \geq 0$ and $\{r_j,\ j \geq 1\}$ a sequence of positive numbers. Let $\mathcal{E}$ be any subset of Z such that*

$$\mathcal{E} \subseteq \left\{ z \in \mathsf{Z} :\ \int_\Theta \frac{f_\theta(z)}{f(z)}\pi(\mathrm{d}\theta) \geq e^{-\beta},\ \ \int_{S^c} \frac{f_\theta(z)}{f(z)}\pi(\mathrm{d}\theta) \leq \rho \right.$$
$$\left. \textit{and} \quad \sup_{\theta\in\Xi_j}\ [\log f_\theta(z) - \log f(z)] \leq -r_j \quad \textit{for all} \quad j \geq 1 \right\}. \tag{25}$$

Then for all $z \in \mathcal{E}$, we have

$$\Pi(B^c|z) \leq e^\beta \left(\rho + \sum_{j\geq 1} e^{-r_j} \right). \tag{26}$$

Proof. Using the lower bound on the normalizing constant provided by the event (25), for $z \in \mathcal{E}$, we have

$$\Pi(B^c|z) = \frac{\int_{B^c} \frac{f_\theta(z)}{f(z)}\pi(\mathrm{d}\theta)}{\int_\Theta \frac{f_\theta(z)}{f(z)}\pi(\mathrm{d}\theta)} \leq e^\beta \left(\int_{S^c} \frac{f_\theta(z)}{f(z)}\pi(\mathrm{d}\theta) + \int_{S\cap B^c} \frac{f_\theta(z)}{f(z)}\pi(\mathrm{d}\theta) \right)$$
$$\leq e^\beta \left(\rho + \int_{S\cap B^c} \frac{f_\theta(z)}{f(z)}\pi(\mathrm{d}\theta) \right).$$

Furthermore, for $z \in \mathcal{E}$, the last integral in the last display satisfies

$$\int_{S\cap B^c} \frac{f_\theta(z)}{f(z)}\pi(\mathrm{d}\theta) \leq \sum_{j\geq 1} \int_{\Xi_j} \exp\left(\log f_\theta(z) - \log f(z)\right)\pi(\mathrm{d}\theta) \leq \sum_{j\geq 1} e^{-r_j}\pi(\Xi_j).$$

Equation (26) follows by collecting the terms. □

Remark 10. From the lemma we are left with the problem of finding $\rho, \beta, \{r_j,\ j \geq 1\}$ such that the right hand size of Equation (26) is small and $\mathbb{P}(Z \notin \mathcal{E})$ is small.

5. Concluding remarks

There is a need for a deeper theoretical understanding of deep learning models. We have focused here on a class of algorithm unrolling models for inverse problems. And we have shown that for convex inverse problems and under a concentration of measure assumption, GDN can recover the inverse map at optimal rate, provided that the unrolling depth is appropriately tuned. These findings are confirmed in our numerical example. Our results also suggest that algorithm unrolling models are prone to overfitting as the unrolling depth D' increases. The theoretical results are obtained as special cases of a more general posterior contraction result for Bayesian deep learning.

One natural question is whether our analysis extends beyond the concentration of measure assumption in Assumption 1. Without the content of H1, a more sensible approach would be to estimate the entire conditional distribution, not just its mean. Several recent works have proposed to extend algorithm unrolling architectures for conditional density estimation in inverse problems (Ardizzone et al. (2019)). Extending our analysis to these conditional density models is an important direction for future research.

Another outstanding challenge not addressed in this work is the computational and memory cost of implementing algorithm unrolling models. Our results suggest that fairly deep (but not too deep) networks are typically needed for optimal performance of GDNs. In practice, the gradient of the loss with respect to W in (17) is typically computed by back-propagation through the entire network of depth $D \times D'$, at a memory cost of order $O(D \times D')$. This often puts severe limitations on the unrolling depth that can be considered (Putzky & Welling (2019)). Mitigating this memory cost and easing the implementation of algorithm unrolling architectures (for instance by developing specialized back-propagation algorithms) is another important problem for future research.

References

Aggarwal, H., Mani, M., and Jacob, M. Modl: Model based deep learning architecture for inverse problems. *IEEE Transactions on Medical Imaging*, PP, 12 2017. doi: 10.1109/TMI.2018.2865356.

Ardizzone, L., Kruse, J., Rother, C., and Kothe, U. Analyzing inverse problems with invertible neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJed6j0cKX.

Atchade, Y. and Bhattacharyya, A. An approach to large-scale quasi-bayesian inference with spike-and-slab priors, 2018. URL https://arxiv.org/abs/1803.10282.

Atchade, Y. and Wang, L. A fast asynchronous mcmc sampler for sparse bayesian inference. *JRSS-B (to appear)*, 202X.

AtchadÃ©, Y. and Wang, L. A fast asynchronous mcmc sampler for sparse bayesian inference, 2021.

Ba, J., Kiros, J. R., and Hinton, G. E. Layer normalization. *ArXiv*, abs/1607.06450, 2016.

Barron, A. R. and Klusowski, J. M. Approximation and estimation for high-dimensional deep learning networks, 2018.

Beck, A. and Teboulle, M. Gradient-based algorithms with applications to signal-recovery problems. In *Convex optimization in signal processing and communications*, pp. 42–88. Cambridge Univ. Press, Cambridge, 2010.

Bickel, P. J. and Kleijn, B. J. K. The semiparametric Bernstein-von Mises theorem. *The Annals of Statistics*, 40(1):206–237, 2012.

Bissantz, N., Hohage, T., Munk, A., and Ruymgaart, F. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM Journal on Numerical Analysis*, 45(6):2610–2636, 2007.

Blanchard, G. and Mücke, N. Optimal rates for regularization of statistical inverse learning problems. *Found. Comput. Math.*, 18(4):971–1013, 2018.

Burger, H. C., Schuler, C. J., and Harmeling, S. Image denoising: Can plain neural networks compete with bm3d? In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2392–2399, 2012. doi: 10.1109/CVPR.2012.6247952.

Chang, J. R., Li, C.-L., Poczos, B., Vijaya Kumar, B., and Sankaranarayanan, A. C. One network to solve them all – solving linear inverse problems using deep projection models. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5889–5898, 2017.

Chen, X., Liu, J., Wang, Z., and Yin, W. Theoretical linear convergence of unfolded ista and its practical weights and thresholds. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31, 2018.

Chun, Y. and Fessler, J. A. Deep bcd-net using identical encoding-decoding cnn structures for iterative image recovery. In *2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pp. 1–5, 2018. doi: 10.1109/IVMSPW.2018.8448694.

Combettes, P. and Wajs, V. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation*, 4(4):1168–1200, 2005.

DeVore, R., Hanin, B., and Petrova, G. Neural network approximation. *Acta Numerica*, 30:327–444, 2021.

Dong, W., Zhang, L., Shi, G., and Wu, X. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, 2011.

Ee, W., Ma, C., and Wang, Q. Rademacher complexity and the generalization error of residual networks. *Communications in Mathematical Sciences*, 18:1755–1774, 01 2020.

Geer, S., van de Geer, S., Gill, R., Ripley, B., Ross, S., Silverman, B., Williams, D., and Stein, M. *Empirical Processes in M-Estimation.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000. ISBN 9780521650021. URL `https://books.google.com/books?id=2DYoMRz_0YEC`.

Gilton, D., Ongie, G., and Willett, R. M. Neumann networks for linear inverse problems in imaging. *IEEE Transactions on Computational Imaging*, 6:328–343, 2020.

Gregor, K. and LeCun, Y. Learning fast approximations of sparse coding. In *Proceedings of the 27th international conference on international conference on machine learning*, pp. 399–406, 2010.

Knapik, B., Vaart, A., and Zanten, J. Bayesian inverse problems with gaussian priors. *The Annals of Statistics*, 39, 03 2011.

Li, Y., Tofighi, M., Geng, J., Monga, V., and Eldar, Y. C. Efficient and interpretable deep blind image deblurring via algorithm unrolling. *IEEE Transactions on Computational Imaging*, 6:666–681, 2020. doi: 10.1109/TCI.2020.2964202.

Liu, R., Cheng, S., Ma, L., Fan, X., and Luo, Z. Deep proximal unrolling: Algorithmic framework, convergence analysis and applications. *IEEE Transactions on Image Processing*, 28(10):5013–5026, 2019. doi: 10.1109/TIP.2019.2913536.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Lucas, A., Iliadis, M., Molina, R., and Katsaggelos, A. K. Using deep neural networks for inverse problems in imaging: Beyond analytical methods. *IEEE Signal Processing Magazine*, 35(1):20–36, 2018.

Monga, V., Li, Y., and Eldar, Y. C. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2): 18–44, 2021.

Nickl, R. Bernstein - von mises theorems for statistical inverse problems i: Schrödinger equation. *Journal of the European Mathematical Society*, 22, 07 2017. doi: 10.4171/JEMS/975.

Nickl, R. and Sohl, J. Bernstein - von mises theorems for statistical inverse problems ii: Compound poisson processes. *Electronic Journal of Statistics*, 13:3513 – 3571, 2019.

Ongie, G., Jalal, A., Baraniuk, C., Dimakis, A., and Willett, R. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, PP:1–1, 05 2020. doi: 10.1109/JSAIT.2020.2991563.

Parikh, N. and Boyd, S. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

Polson, N. G. and Ročková, V. Posterior concentration for sparse deep learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 930–941. Curran Associates, Inc., 2018.

Putzky, P. and Welling, M. *Invert to Learn to Invert.* Curran Associates Inc., Red Hook, NY, USA, 2019.

Rastogi, A., Blanchard, G., and Mathé, P. Convergence analysis of Tikhonov regularization for non-linear statistical inverse problems. *Electronic Journal of Statistics*, 14(2):2798 – 2841, 2020.

Ravishankar, S., Chun, I. Y., and Fessler, J. A. Physics-driven deep training of dictionary-based algorithms for mr image reconstruction. In *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pp. 1859–1863, 2017. doi: 10.1109/ACSSC.2017.8335685.

Samworth, R. J. Recent Progress in Log-Concave Density Estimation. *Statistical Science*, 33(4):493 – 509, 2018.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with relu activation function. *Annals of Statistics*, 48, 08 2020. doi: 10.1214/19-AOS1875.

Shen, X. and Wasserman, L. Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687 – 714, 2001.

Shlezinger, N., Whang, J., Eldar, Y. C., and Dimakis, A. G. Model-based deep learning: Key approaches and design guidelines. In *2021 IEEE Data Science and Learning Workshop (DSLW)*, pp. 1–6, 2021. doi: 10.1109/DSLW51110.2021.9523403.

Sreter, H. and Giryes, R. Learned convolutional sparse coding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2191–2195. IEEE, 2018.

Stuart, A. M. Inverse problems: A bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

Sulam, J., Aberdam, A., Beck, A., and Elad, M. On multi-layer basis pursuit, efficient algorithms and convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):1968–1980, 2020.

Taheri, M., Xie, F., and Lederer, J. Statistical guarantees for regularized neural networks. *Neural Networks*, 142:148–161, 2021.

Tao, S., Boley, D., and Zhang, S. Local linear convergence of ista and fista on the lasso problem. *SIAM Journal on Optimization*, 26(1):313–336, 2016.

Tolooshams, B., Song, A., Temereanca, S., and Ba, D. Convolutional dictionary learning based auto-encoders for natural exponential-family distributions. In *International Conference on Machine Learning*, pp. 9493–9503. PMLR, 2020.

Vaart, A. W. v. d. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Xie, J., Xu, L., and Chen, E. Image denoising and inpainting with deep neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Yang, y., Sun, J., Li, H., and Xu, Z. Deep admm-net for compressive sensing mri. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Yarotsky, D. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

Zhang, K., Zuo, W., Gu, S., and Zhang, L. Learning deep cnn denoiser prior for image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3929–3938, 2017.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2): 301–320, 2005.

Appendix A. Proofs

A.1. **Proof of Theorem 7.**

Proof. We follow the same general steps outlined above in Lemma 9. We recall that the dataset is $\mathcal{D} \overset{\text{def}}{=} (\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n)$. For $W \in \mathcal{W}$, we define

$$f_W(\mathcal{D}) \overset{\text{def}}{=} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} \|\mathbf{x}_i - g_W(\mathbf{y}_i)\|_2^2\right),$$

$$\text{and} \quad f_\star(\mathcal{D}) \overset{\text{def}}{=} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} \|\mathbf{x}_i - g(\mathbf{y}_i)\|_2^2\right). \tag{27}$$

We recall that $\Theta = \mathcal{W}\times\mathcal{S}$. For any measurable set $A \subset \Theta$, we can write the posterior probability $\Pi(A|\mathcal{D})$ as

$$\Pi(A|\mathcal{D}) = \frac{\int_A \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})}\Pi_0(\mathrm{d}\Lambda, \mathrm{d}W)}{\int_\Theta \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})}\Pi_0(\mathrm{d}\Lambda, \mathrm{d}W)}. \tag{28}$$

We will repeatedly use the following observation. For $W \in \mathcal{W}$, we have

$$\begin{aligned}\log\left(\frac{f_W(\mathcal{D})}{f_\star(\mathcal{D})}\right) &= \frac{1}{2\sigma^2}\sum_{i=1}^n \left(\|\mathbf{x}_i - g(\mathbf{y}_i)\|_2^2 - \|\mathbf{x}_i - g_W(\mathbf{y}_i)\|_2^2\right)\\ &= -\frac{n}{2\sigma^2}\|g_W - g\|_n^2 - \frac{1}{\sigma^2}\sum_{i=1}^n \left\langle \mathbf{x}_i - g(\mathbf{y}_i), g(\mathbf{y}_i) - g_W(\mathbf{y}_i)\right\rangle. \end{aligned}\tag{29}$$

Given $s_0 \geq 1$, $\beta_0 \geq 0$, we set

$$\Theta(s_0, \beta_0) \overset{\text{def}}{=} \left\{(\Lambda, W) \in \Theta:\ \|\Lambda\|_0 \leq s_0, \ \ \text{and } \|\Lambda\odot W\|_\infty \leq \beta_0\right\},$$

and

$$\mathcal{W}(s_0, \beta_0) \overset{\text{def}}{=} \left\{W \in \mathcal{W}:\ \|W\|_0 \leq s_0, \quad \|W\|_\infty \leq \beta_0\right\}.$$

We set

$$s \overset{\text{def}}{=} \frac{1}{\mathsf{u}} + \left(1 + \frac{5}{\mathsf{u}} + \frac{\log(\mathsf{L}_\star\sqrt{n})}{\mathsf{u}\log(q)}\right)s_\star + \frac{4n\varpi_\star^2}{\sigma^2\mathsf{u}\log(q)}, \quad \text{and} \quad \mathsf{r} \overset{\text{def}}{=} \bar{\sigma}\sqrt{\frac{s\log(q\mathsf{L}_s)}{n}},$$

and

$$\bar{\alpha} \overset{\text{def}}{=} \mathsf{u}s - 1,$$

where $\mathsf{L}_\star \overset{\text{def}}{=} L(2s_\star^{1/2})$, $\mathsf{L}_s \overset{\text{def}}{=} L(2s^{1/2}b_s)$, and $b_s \overset{\text{def}}{=} \sqrt{2\rho_1^{-1}(1+\mathsf{u})(s+1)\log(q)}$, and where L is as in Assumption 4. Fix $M \geq 2$. For $j \geq 1$ we also set

$$\mathcal{W}_j(s_0, \beta_0) \overset{\text{def}}{=} \{W \in \mathcal{W}(s_0, \beta_0):\ j(M\mathsf{r}) < \|g_W - g\|_n \leq (j+1)M\mathsf{r}\}.$$

We shall apply the same idea as in Lemma 9. Specifically, let

$$B \overset{\text{def}}{=} \left\{(\Lambda, W) \in \Theta:\ \|g_{\Lambda\odot W} - g\|_n \leq M\mathsf{r}\right\},$$

and consider the $\mathcal{E}$

$$\mathcal{E} = \left\{\mathcal{D}:\ \int_\Theta \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})}\Pi_0(\mathrm{d}\Lambda, \mathrm{d}W) > \frac{1}{4q^{\bar{\alpha}}},\ \int_{\mathcal{A}(s)} \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})}\Pi_0(\mathrm{d}\Lambda, \mathrm{d}W) \leq \frac{1}{q^{\mathsf{u}s}}\right.$$
$$\left.\text{and } \sup_{W\in\mathcal{W}_j(s,b_s)} \left[\log f_W(\mathcal{D}) - \log f_\star(\mathcal{D})\right] \leq -\frac{n(jM\mathsf{r})^2}{8\sigma^2}, \ \text{ for all } \ j \geq 1\right\},$$

where $\mathcal{A}(s)$ denotes the complement of $\Theta(s, b_s)$. We note if $(\Lambda, W) \in B^c \cap \Theta(s, b_s)$, then $\Lambda\odot W \in \cup_{j\geq1}\mathcal{W}_j(s, b_s)$. Let $\check{\Pi}_0$ be the distribution of $\Lambda\odot W$, when $(\Lambda, W) \sim \Pi_0$.

Starting from (28), and following the same argument leading to (26), for $\mathcal{D} \in \mathcal{E}$, we have

$$\begin{aligned}
\Pi(B^c|\mathcal{D}) &\leq 4q^{\bar{\alpha}} \int_{B^c} \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})} \Pi_0(\mathrm{d}\Lambda, \mathrm{d}W) \\
&\leq 4q^{\bar{\alpha}} \left(\int_{\mathcal{A}(s)} \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})} \Pi_0(\mathrm{d}\Lambda, \mathrm{d}W) + \int_{B^c \cap \Theta(s,b_s)} \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})} \Pi_0(\mathrm{d}\Lambda, \mathrm{d}W) \right) \\
&\leq 4e^{\bar{\alpha}\log(q)} \left(\frac{1}{q^{\mathsf{u}s}} + \int_{B^c \cap \Theta(s,b_s)} \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})} \Pi_0(\mathrm{d}\Lambda, \mathrm{d}W) \right) \\
&\leq 4e^{\bar{\alpha}\log(q)} \left(\frac{1}{q^{\mathsf{u}s}} + \sum_{j\geq 1} \int_{\mathcal{W}_j(s,b_s)} \frac{f_W(\mathcal{D})}{f_\star(\mathcal{D})} \check{\Pi}_0(\mathrm{d}W) \right) \\
&\leq 4e^{\bar{\alpha}\log(q)} \left(e^{-\mathsf{u}s\log(q)} + \sum_{j\geq 1} e^{-\frac{n(jM\mathsf{r})^2}{8\sigma^2}} \right) \\
&\leq 4e^{\bar{\alpha}\log(q)} \left(e^{-\mathsf{u}s\log(q)} + 2e^{-\frac{n(M\mathsf{r})^2}{8\sigma^2}} \right).
\end{aligned}$$

By the definition of s and r above, we have $\mathsf{u}s = \bar{\alpha} + 1$, and

$$n(M\mathsf{r})^2 \geq M^2\bar{\sigma}^2 s\log(q) = M^2\bar{\sigma}^2 \left(\frac{1+\bar{\alpha}}{\mathsf{u}}\right) \log(q) \geq 8\sigma^2(1+\bar{\alpha})\log(q),$$

by taking $M^2 \geq 8\mathsf{u}(\sigma^2/\bar{\sigma}^2)$. Hence for $\mathcal{D} \in \mathcal{E}$,

$$\Pi(B^c|\mathcal{D}) \leq \frac{12}{q}.$$

This implies that with probability at least $\mathbb{P}(\mathcal{D} \in \mathcal{E})$, we have

$$\Pi(B^c|\mathcal{D}) \leq \frac{12}{q}.$$

We show in Lemma 12 below that

$$\mathbb{P}\left[\int_\Theta \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})} \Pi_0(\mathrm{d}\Lambda, \mathrm{d}W) \leq \frac{1}{4q^{\bar{\alpha}}} \mid \mathbf{y}_{1:n} \right] \leq \frac{4}{q^{s_\star}},$$

and we show in Lemma 11 below that

$$\mathbb{P}\left[\int_{\mathcal{A}} \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})} \Pi_0(\mathrm{d}\Lambda, \mathrm{d}W) > \frac{1}{q^{\mathsf{u}s}} \mid \mathbf{y}_{1:n} \right] \leq \frac{3}{q^{\mathsf{u}}}.$$

It follows that

$$\mathbb{P}(\mathcal{D} \notin \mathcal{E} \mid \mathbf{y}_{1:n}) \leq \frac{4}{q^{s_\star}} + \frac{3}{q^{\mathsf{u}}} \\ + \mathbb{P}\left[\bigcup_{j\geq 1}\left\{\sup_{W\in\mathcal{W}_j(s,b_s)} [\log f_W(\mathcal{D}) - \log f_\star(\mathcal{D})] > -\frac{n(jM\mathsf{r})^2}{8\sigma^2}\right\} \mid \mathbf{y}_{1:n}\right].$$

By Lemma 13 applied with $\mathcal{W}_0 = \mathcal{W}(s, b_s)$, the rightmost term in the last display is bounded from above by $e^{-c_0 n} + 4e^{-n(M\mathsf{r})^2/(c_0\bar{\sigma}^2)}$, for some absolute constant c_0 provided that the term r defined above satisfies

$$\frac{288}{\sqrt{n}}\int_{\frac{x^2}{32\bar{\varsigma}}}^{x}\sqrt{\log\mathcal{N}\left(\epsilon, \mathcal{W}^{(x)}(s,b_s), \|\cdot\|_n\right)}\mathrm{d}\epsilon \leq \frac{x^2}{\bar{\sigma}}, \quad \text{for all} \quad x \geq \mathsf{r}, \tag{30}$$

where for $s_0 \geq 1$, $x \geq 0$, $\beta_0 \geq 0$ we define

$$\mathcal{W}^{(x)}(s_0, \beta_0) \overset{\text{def}}{=} \{W \in \mathcal{W}(s_0, \beta_0),\ \|g_W - g\|_n \leq x\},$$

and given $\epsilon > 0$, and $A \subset \mathcal{W}$, $\mathcal{N}(\epsilon, A, \|\cdot\|_n)$ denotes the cardinality of a smallest ϵ-cover of A in the pseudo-metric $\|\cdot\|_n$ defined as $\|W - W'\|_n \overset{\text{def}}{=} \|g_W - g_{W'}\|_n$. We therefore reach the conclusion that with probability at least $1 - e^{-c_0 n} - c_1/q$,

$$\Pi(B^c|\mathcal{D}) \leq \frac{12}{q}.$$

for some absolute constants c_0, c_1. It remains to check (30). First we use the majoration

$$\int_{\frac{x^2}{32\bar{\varsigma}}}^{x}\sqrt{\log\mathcal{N}\left(\epsilon, \mathcal{W}^{(x)}(s,b_s), \|\cdot\|_n\right)}\mathrm{d}\epsilon \leq x\sqrt{\log\mathcal{N}\left(\frac{x^2}{32\bar{\varsigma}}, \mathcal{W}(s,b_s), \|\cdot\|_n\right)}.$$

We recall that our notation $\|W\|_2$ denotes the Euclidean norm of the vectorized parameter W. For $W \in \mathcal{W}(s, b_s)$, $\|W\|_2 \leq s^{1/2}b_s$. Hence, assumption H3, and the definition of $\mathsf{L}_s = L(s^{1/2}b_s)$ implies that for all $W, W' \in \mathcal{W}(s, b_s)$, we have

$$\|W - W'\|_n = \|g_W - g_{W'}\|_n \leq \mathsf{L}_s\|W - W'\|_2.$$

Therefore, we can use the metric entropy of the s-sparse ball of $\mathbb{R}^q$ with radius $s^{1/2}b_s/\mathsf{L}_s$ with respect to the Euclidean norm to get

$$\mathcal{N}(\epsilon, \mathcal{W}(s,b_s), \|\cdot\|_n) \leq q^s\left(1 + \frac{2s^{1/2}b_s\mathsf{L}_s}{\epsilon}\right)^s.$$

Hence

$$\frac{288}{\sqrt{n}}\int_{\frac{x^2}{32\bar{\varsigma}}}^{x}\sqrt{\log\mathcal{N}(\epsilon, \mathcal{W}^{(x)}(s,b_s), \|\cdot\|_n)}\mathrm{d}\epsilon \leq 288x\sqrt{\frac{s\log(q)}{n} + \frac{s\log\left(1 + \frac{64\bar{\varsigma}s^{1/2}b_s\mathsf{L}_s}{x^2}\right)}{n}}.$$

We can insist to search for $x \geq \sqrt{128\bar{\varsigma}/n}$, and conclude that the right hand side of the last display is always upper bounded by

$$288x\sqrt{\frac{s\log(q)}{n} + \frac{s\log\left(1+\frac{ns^{1/2}b_s\mathsf{L}_s}{2}\right)}{n}} \leq c_0 x\sqrt{\frac{s\log(q\mathsf{L}_s)}{n}},$$

for some absolute constant c_0. The right hand side of the last display is upper bounded by $\frac{x^2}{\bar{\sigma}}$ for all

$$x \geq c_0\bar{\sigma}\sqrt{\frac{s\log(q\mathsf{L}_s)}{n}},$$

hence the theorem, after moving the constant c_0 into M. □

Lemma 11. *Assume H1, and suppose that $\sigma^2 \geq \max_i \sigma_i^2$. For all integers $s \geq 1$, with $b_s \overset{\text{def}}{=} \sqrt{2(1+\mathsf{u})(1+s)\log(q)/\rho_1}$, we have*

$$\mathbb{P}\left[\int_{\mathcal{A}(s)} \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})}\Pi_0(\mathrm{d}\Lambda, \mathrm{d}W) > \frac{1}{q^{\mathsf{u}s}} \mid \mathbf{y}_{1:n}\right] \leq \frac{4}{q^{\mathsf{u}}},$$

where $\mathcal{A}(s)$ denotes the complement of the set $\Theta(s, b_s)$ where

$$\Theta(s,b) \overset{\text{def}}{=} \left\{(\Lambda, W) \in \Theta:\ \|\Lambda\|_0 \leq s, \quad \textit{and}\ \|\Lambda \odot W\|_\infty \leq b\right\}.$$

Proof. Since $\mathcal{A}(s)$ is the complement of the set $\Theta(s, b_s)$, we can write

$$\Pi_0(\mathcal{A}(s)) = \Pi_0(\|\Lambda\|_0 > s) + \sum_{\Lambda:\ \|\Lambda\|_0 \leq s} \Pi_0(\Lambda) \times \Pi_0(\|\Lambda \odot W\|_\infty > b_s | \Lambda).$$

If $(\Lambda, W) \sim \Pi_0$, then Λ is an ensemble of iid random variables drawn from the Bernoulli distribution with success probability $(1+q^{\mathsf{u}+1})^{-1}$. Hence

$$\Pi_0(\|\Lambda\|_0 > s) \leq \sum_{k>s}\binom{q}{k}\left(\frac{1}{1+q^{\mathsf{u}+1}}\right)^k\left(\frac{q^{\mathsf{u}+1}}{1+q^{\mathsf{u}+1}}\right)^{q-k}$$
$$\leq \sum_{k>s}\binom{q}{k}\left(\frac{1}{q^{\mathsf{u}+1}}\right)^k \leq 2\left(\frac{1}{q^{\mathsf{u}}}\right)^{s+1},$$

where we use $\binom{q}{k} \leq q^k$, and $q^{\mathsf{u}} \geq 2$. Given $\Lambda_k = 1$, $W_k \sim \mathbf{N}(0, \rho_1^{-1})$. Therefore, $\mathbb{P}(|W_k| > t) \leq 2e^{-\rho_1 t^2/2}$ for all $t \geq 0$. Hence by union bound, for $\|\Lambda\|_0 \leq s$, we obtain

$$\Pi_0\left(\|\Lambda \odot W\|_\infty > b_s \mid \Lambda\right) \leq 2e^{-\rho_1 b_s^2/2 + \log(s)} \leq \frac{2}{q^{\mathsf{u}(1+s)}}.$$

We conclude that

$$\Pi_0(\mathcal{A}(s)) \leq \frac{4}{q^{\mathsf{u}(1+s)}}. \tag{31}$$

Now, by Markov's inequality, and Fubini's theorem, we have

$$\mathbb{P}\left[\int_{\mathcal{A}(s)} \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})}\Pi_0(\mathrm{d}\Lambda,\mathrm{d}W) > \frac{1}{q^{\mathsf{u}s}} \mid \mathbf{y}_{1:n}\right] \\ \leq q^{\mathsf{u}s}\int_{\mathcal{A}(s)} \mathbb{E}\left[\frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})} \mid \mathbf{y}_{1:n}\right]\Pi_0(\mathrm{d}\Lambda,\mathrm{d}W),$$

and from (29) we have

$$\mathbb{E}\left[\frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})} \mid \mathbf{y}_{1:n}\right] = e^{-\frac{n}{2\sigma^2}\|g_{\Lambda\odot W}-g\|_n^2}\mathbb{E}\left[e^{-\frac{1}{\sigma^2}\sum_{i=1}^n\langle\boldsymbol{\xi}_i, g(\mathbf{y}_i)-g_{\Lambda\odot W}(\mathbf{y}_i)\rangle} \mid \mathbf{y}_{1:n}\right].$$

We have assumed in H1 that $\mathbb{E}(\boldsymbol{\xi}_i|\mathbf{y}_i) = 0$, and $\|\boldsymbol{\xi}_i|\mathbf{y}_i\|_{\psi_2} \leq \sigma_i$. Therefore,

$$\mathbb{E}\left[e^{-\frac{1}{\sigma^2}\sum_{i=1}^n\langle\boldsymbol{\xi}_i, g(\mathbf{y}_i)-g_{\Lambda\odot W}(\mathbf{y}_i)\rangle} \mid \mathbf{y}_{1:n}\right] \leq e^{\frac{1}{\sigma^2}\sum_{i=1}^n \frac{\sigma_i^2 d_i^2}{2\sigma^2}},$$

where d_i is a short for $\|g(\mathbf{y}_i) - g_{\Lambda\odot W}(\mathbf{y}_i)\|_2$. We conclude that

$$\mathbb{E}\left[\frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})} \mid \mathbf{y}_{1:n}\right] \leq \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n\left[\left(1-\frac{\sigma_i^2}{\sigma^2}\right)d_i^2\right]\right).$$

And we easily check that for $\sigma^2 \geq \bar{\sigma}^2$, the right hand size of the last display is bounded from above by 1. We conclude that

$$\mathbb{P}\left[\int_{\mathcal{A}(s)} \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})}\Pi_0(\mathrm{d}\Lambda,\mathrm{d}W) > \frac{1}{q^{\mathsf{u}s}} \mid \mathbf{y}_{1:n}\right] \leq q^{\mathsf{u}s}\Pi_0(\mathcal{A}(s)) \leq \frac{4}{q^{\mathsf{u}}}.$$

□

The next result lower bounds the normalizing constant of $\Pi(\cdot|\mathcal{D})$.

Lemma 12. *Under the assumption of Theorem 7 it holds,*

$$\mathbb{P}\left[\int_{\Theta} \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})}\Pi_0(\mathrm{d}\Lambda,\mathrm{d}W) \leq \frac{1}{4q^{\bar{\alpha}}} \mid \mathbf{y}_{1:n}\right] \leq \frac{4}{q^{s_\star}},$$

where

$$\bar{\alpha} \overset{\text{def}}{=} \left(\mathsf{u} + 5 + \frac{\log(\mathsf{L}_\star\sqrt{n})}{\log(q)}\right)s_\star + \frac{4n\varpi_\star^2}{\sigma^2\log(q)}.$$

Proof. By the assumption of Theorem 7, we can find $W_\star$ with $\|W_\star\|_0 \leq s_\star$, $\|W_\star\|_\infty \leq 1$, such that $\|g_{W_\star} - g\|_n \leq \varpi_\star$. Let $\Lambda_\star$ denote the sparsity support of $W_\star$. With $\mathsf{L}_\star = L(2s_\star^{1/2})$, we set

$$\eta \overset{\text{def}}{=} 1 \wedge \frac{\sigma}{\mathsf{L}_\star}\sqrt{\frac{\log(q)}{n}}, \quad \text{and} \quad \mathcal{N}(\eta) \overset{\text{def}}{=} \{W \in \mathcal{W} : \quad \|W\odot\Lambda_\star - W_\star\|_\infty \leq \eta\}.$$

We see that $\|W_\star\|_2 \leq s_\star^{1/2}$, and for $W \in \mathcal{N}(\eta)$,

$$\|W\odot\Lambda_\star\|_2 \leq \|W_\star\|_2 + \|W_\star - W\odot\Lambda_\star\|_2 \leq s_\star^{1/2} + s_\star^{1/2}\eta \leq 2s_\star^{1/2}.$$

Therefore, by H4 applied with $\eta = 2s_\star^{1/2}$, for all $W \in \mathcal{N}(\eta)$, we have

$$\max_{1\le i\le n} \|g_{\Lambda_\star \odot W}(\mathbf{y}_i) - g_{W_\star}(\mathbf{y}_i)\|_2 \le \mathsf{L}_\star \|\Lambda_\star \odot W - W_\star\|_2 \le \mathsf{L}_\star \sqrt{s_\star}\eta \le \sigma\sqrt{\frac{s_\star \log(q)}{n}}.$$

Hence

$$\max_{1\le i\le n} \sup_{W\in\mathcal{N}(\eta)} \|g_{\Lambda_\star \odot W}(\mathbf{y}_i) - g_{W_\star}(\mathbf{y}_i)\|_2 \le \sigma\sqrt{\frac{s_\star \log(q)}{n}}. \tag{32}$$

Switching the sign and taking the conditional expectation in (29) using $\mathbb{E}(\mathbf{x}_i|\mathbf{y}_i) = g(\mathbf{y}_i)$, yields

$$\mathbb{E}\left[\log\left(\frac{f_\star(\mathcal{D})}{f_{\Lambda_\star\odot W}(\mathcal{D})}\right) \mid \mathbf{y}_{1:n}\right] = \frac{1}{2\sigma^2}\sum_{i=1}^n \|g_{\Lambda_\star\odot W}(\mathbf{y}_i) - g(\mathbf{y}_i)\|_2^2,$$

and we conclude using (32) and the definition of $\varpi_\star$ and $\mathcal{W}_\star$ in Theorem 7 that

$$\begin{aligned}\sup_{W\in\mathcal{N}(\eta)} \mathbb{E}&\left[\log\left(\frac{f_\star(\mathcal{D})}{f_{\Lambda_\star\odot W}(\mathcal{D})}\right) \mid \mathbf{y}_{1:n}\right] \\ &\le \frac{n\varpi_\star^2}{\sigma^2} + \frac{1}{\sigma^2}\sum_{i=1}^n \sup_{W\in\mathcal{N}(\eta)} \|g_{\Lambda_\star\odot W}(\mathbf{y}_i) - g_{W_\star}(\mathbf{y}_i)\|_2^2 \\ &\le \frac{n\varpi_\star^2}{\sigma^2} + s_\star \log(q).\end{aligned}$$

Going back to (29), we have

$$\log\left(\frac{f_\star(\mathcal{D})}{f_{\Lambda_\star\odot W}(\mathcal{D})}\right) - \mathbb{E}\left[\log\left(\frac{f_\star(\mathcal{D})}{f_{\Lambda_\star\odot W}(\mathcal{D})}\right) \mid \mathbf{y}_{1:n}\right] = \frac{1}{\sigma^2}\sum_{i=1}^n \left\langle \boldsymbol{\xi}_i, g(\mathbf{y}_i) - g_{\Lambda_\star\odot W}(\mathbf{y}_i)\right\rangle. \tag{33}$$

We use the notation $\|Z\|_{\psi_2}$ to denote the sub-Gaussian norm of the conditional law of the random variable Z given $\mathbf{y}_{1:n}$. By conditional independence of the error terms $\boldsymbol{\xi}_i$, for all $W \in \mathcal{N}(\eta)$, we have

$$\begin{aligned}\Bigg\|\log\left(\frac{f_\star(\mathcal{D})}{f_{\Lambda_\star\odot W}(\mathcal{D})}\right) &- \mathbb{E}\left[\log\left(\frac{f_\star(\mathcal{D})}{f_{\Lambda_\star\odot W}(\mathcal{D})}\right) \mid \mathbf{y}_{1:n}\right]\Bigg\|_{\psi_2}^2 \\ &\le \frac{1}{\sigma^4}\sum_{i=1}^n \|\langle \boldsymbol{\xi}_i, g(\mathbf{y}_i) - g_{\Lambda_\star\odot W}(\mathbf{y}_i)\rangle\|_{\psi_2}^2 \\ &= \frac{1}{\sigma^4}\sum_{i=1}^n \sigma_i^2 \|g(\mathbf{y}_i) - g_{\Lambda_\star\odot W}(\mathbf{y}_i)\|_2^2 \le \frac{2n\varpi_\star^2}{\sigma^2} + 2s_\star\log(q).\end{aligned}$$

In the sequel, we set

$$a \overset{\text{def}}{=} 2\left(\frac{n\varpi_\star^2}{\sigma^2} + s_\star \log(q)\right).$$

Then by Hoeffding's inequality, for all $W \in \mathcal{N}(\eta)$, we have

$$\mathbb{P}\left[\left|\log\left(\frac{f_\star(\mathcal{D})}{f_{\Lambda_\star\odot W}(\mathcal{D})}\right) - \mathbb{E}\left[\log\left(\frac{f_\star(\mathcal{D})}{f_{\Lambda_\star\odot W}(\mathcal{D})}\right) \mid \mathbf{y}_{1:n}\right]\right| > a \mid \mathbf{y}_{1:n}\right] \leq 2e^{-a/2} \leq \frac{2}{q^{s_\star}}.$$

We can rewrite this statement in the following equivalent form. For $W \in \mathcal{W}$, define

$$\mathcal{E}_W \stackrel{\text{def}}{=} \left\{\mathcal{D} : \left|\log\left(\frac{f_\star(\mathcal{D})}{f_W(\mathcal{D})}\right) - \mathbb{E}\left[\log\left(\frac{f_\star(\mathcal{D})}{f_W(\mathcal{D})}\right) \mid \mathbf{y}_{1:n}\right]\right| \leq a\right\}.$$

We have

$$\sup_{W\in\mathcal{N}(\eta)} \mathbb{P}\left(\mathcal{D} \notin \mathcal{E}_{\Lambda_\star\odot W} \mid \mathbf{y}_{1:n}\right) \leq \frac{2}{q^{s_\star}}. \tag{34}$$

Using these observations, we have

$$\begin{aligned}\int_\Theta \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})}\Pi_0(\mathrm{d}\Lambda, \mathrm{d}W) \geq \int_{\Lambda_\star\times\mathcal{N}(\eta)} e^{-\mathbb{E}\left[\log\left(\frac{f_\star(\mathcal{D})}{f_{\Lambda\odot W}(\mathcal{D})}\right) \mid \mathbf{y}_{1:n}\right]} \\ \times \exp\left(-\left[\log\left(\frac{f_\star(\mathcal{D})}{f_{\Lambda\odot W}(\mathcal{D})}\right) - \mathbb{E}\left[\log\left(\frac{f_\star(\mathcal{D})}{f_{\Lambda\odot W}(\mathcal{D})}\right) \mid \mathbf{y}_{1:n}\right]\right]\right)\mathbf{1}_{\mathcal{E}_{\Lambda\odot W}}(\mathcal{D})\Pi_0(\mathrm{d}\Lambda, \mathrm{d}W) \\ \geq e^{-2a}\int_{\Lambda_\star\times\mathcal{N}(\eta)} \mathbf{1}_{\mathcal{E}_{\Lambda\odot W}}(\mathcal{D})\Pi_0(\mathrm{d}\Lambda, \mathrm{d}W) \\ = e^{-2a}\left(\Pi_0(\Lambda_\star\times\mathcal{N}(\eta)) - \int_{\Lambda_\star\times\mathcal{N}(\eta)} \mathbf{1}_{\mathcal{E}^c_{\Lambda\odot W}}(\mathcal{D})\Pi_0(\mathrm{d}\Lambda, \mathrm{d}W)\right).\end{aligned}$$

Therefore, by Chebyshev's inequality,

$$\begin{aligned}\mathbb{P}\left[\int_\Theta \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})}\Pi_0(\mathrm{d}\Lambda, \mathrm{d}W) \leq e^{-2a}\Pi_0(\Lambda_\star\times\mathcal{N}(\eta))/2 \mid \mathbf{y}_{1:n}\right] \\ \leq \mathbb{P}\left[\int_{\Lambda_\star\times\mathcal{N}(\eta)} \mathbf{1}_{\mathcal{E}^c_{\Lambda\odot W}}(\mathcal{D})\Pi_0(\mathrm{d}\Lambda, \mathrm{d}W) \geq \frac{1}{2}\Pi_0(\Lambda_\star\times\mathcal{N}(\eta)) \mid \mathbf{y}_{1:n}\right] \\ \leq \frac{2}{\Pi_0(\Lambda_\star\times\mathcal{N}(\eta))}\int_{\Lambda_\star\times\mathcal{N}(\eta)} \mathbb{P}\left(\mathcal{D} \notin \mathcal{E}_{\Lambda_\star\odot W} \mid \mathbf{y}_{1:n}\right)\Pi_0(\mathrm{d}\Lambda, \mathrm{d}W) \\ \leq 2\sup_{W\in\mathcal{N}(\eta)} \mathbb{P}_\star\left(\mathcal{D} \notin \mathcal{E}_{\Lambda_\star\odot W} \mid \mathbf{y}_{1:n}\right) \leq \frac{4}{q^{s_\star}},\end{aligned}$$

using (34). To conclude the proof it remains only to lower bound $\Pi_0(\Lambda_\star \times \mathcal{N}(\eta))$. Since $\log(1-x) \geq -2x$ for all $0 \leq x \leq 1/2$, for $q^{\mathsf{u}} \geq 2/\log(2)$, we have

$$\begin{aligned}\Pi_0(\Lambda_\star) &= \left(\frac{1}{1+q^{\mathsf{u}+1}}\right)^{\|\Lambda_\star\|_0}\left(1-\frac{1}{1+q^{\mathsf{u}+1}}\right)^{q-\|\Lambda_\star\|_0} \\ &= \left(\frac{1}{q^{\mathsf{u}+1}}\right)^{\|\Lambda_\star\|_0}\exp\left(q\log\left(1-\frac{1}{1+q^{\mathsf{u}+1}}\right)\right) \\ &\geq \left(\frac{1}{q^{\mathsf{u}+1}}\right)^{\|\Lambda_\star\|_0}\exp\left(-\frac{2q}{1+q^{\mathsf{u}+1}}\right) \geq \frac{1}{2}\left(\frac{1}{q^{\mathsf{u}+1}}\right)^{\|\Lambda_\star\|_0} \geq \frac{1}{2}\left(\frac{1}{q^{\mathsf{u}+1}}\right)^{s_\star}.\end{aligned}$$

If $U \sim \mathbf{N}(0,\rho_1)$, then $P(|U-a| \leq t) \geq P(|a| \leq U \leq |a|+t)$ for all $t \geq 0$. We use this inequality to deduce that

$$\begin{aligned}\Pi_0(\mathcal{N}(\eta) \mid \Lambda_\star) \geq (\Phi(\sqrt{\rho_1}(1+\eta)) - \Phi(\sqrt{\rho_1}))^{\|\Lambda_\star\|_0} &\geq (c_0\sqrt{\rho_1}\eta)^{s_\star} \\ &\geq \left(\frac{c_0\sigma}{\mathsf{L}_\star}\sqrt{\frac{\rho_1\log(q)}{n}}\right)^{s_\star} \geq \left(\frac{1}{\mathsf{L}_\star\sqrt{n}}\right)^{s_\star},\end{aligned}$$

for some absolute constant c_0 (c_0 can be taken as $e^{-2}/\sqrt{2\pi}$, since $\rho_1 = 1$), where Φ is the cdf of the standard normal distribution. The last inequality in the last display uses the assumption that $n \geq \sigma^2\log(p)$, and $c_0^2\sigma^2\log(q) \geq 1$. We conclude that

$$\begin{aligned}&e^{-2a}\Pi_0(\Lambda_\star \times \mathcal{N}(\eta)) \\ &\quad\geq \frac{1}{2}\exp\left(-\frac{4n\varpi_\star^2}{\sigma^2} - 4s_\star\log(q) - (\mathsf{u}+1)s_\star\log(q) - s_\star\log\left(\mathsf{L}_\star\sqrt{n}\right)\right), \\ &\qquad\qquad \geq \frac{1}{2}\exp\left(-\frac{4n\varpi_\star^2}{\sigma^2} - (\mathsf{u}+5)s_\star\log(q) - s_\star\log(\mathsf{L}_\star\sqrt{n})\right).\end{aligned}$$

Hence the result. □

Lemma 13. *Suppose that the dataset $\mathcal{D}$ is generated as in H1, and consider the nonparametric regression (7) for some function class $\{g_W,\ W \in \mathcal{W} \subseteq \mathbb{R}^q\}$. Let $\mathcal{W}_0$ be some subset of $\mathcal{W}$. Suppose that we can find $\mathsf{r} > 0$ such that for all $x \geq \mathsf{r}$, it holds*

$$\frac{288}{\sqrt{n}}\int_{\frac{x^2}{16\bar{\varsigma}}}^{x}\sqrt{\log\mathcal{N}(\epsilon, \mathcal{W}^{(x)}, \|\cdot\|_n)}\mathrm{d}\epsilon \leq \frac{x^2}{\bar{\sigma}}, \tag{35}$$

where $\mathcal{W}^{(x)} \overset{\text{def}}{=} \{W \in \mathcal{W}_0,\ \|g_W - g\|_n \leq x\}$. Let f_W and $f_\star$ be as defined in (27). Then there exists an absolute constant c_0 such that for all $M \geq 1$, such that $n(M\mathsf{r})^2 \geq c_0\bar{\sigma}^2$, it holds

$$\mathbb{P}\left[\bigcup_{j\geq 1}\left\{\sup_{W\in\widetilde{\mathcal{W}}^{(j)}}\ \log\left(\frac{f_W(\mathcal{D})}{f_\star(\mathcal{D})}\right) > -\frac{n(jM\mathsf{r})^2}{8\sigma^2}\right\} \mid \mathbf{y}_{1:n}\right] \leq e^{-c_0 n} + 4e^{-\frac{nM^2\mathsf{r}^2}{c_0\bar{\sigma}^2}},$$

where $\widetilde{\mathcal{W}}^{(j)} \overset{\text{def}}{=} \{W \in \mathcal{W}_0 :\ jM\mathsf{r} < \|g_W - g\|_n \leq (j+1)M\mathsf{r}\}$.

Proof. We proceed as in Lemma 3.2 of Geer et al. (2000). Throughout the proof, all expectations and probability are conditional given $\mathbf{y}_{1:n}$. However to ease notation we omit the conditioning. With M and r as in the statement, and for each integer j, we set $\mathsf{r}_j = M\mathsf{r}j$. We recall the definition of the error terms $\boldsymbol{\xi}_i \overset{\text{def}}{=} \mathbf{x}_i - g(\mathbf{y}_i)$, and we define

$$Z_n(g_W) \overset{\text{def}}{=} \frac{1}{n\sigma^2}\sum_{i=1}^n \langle \boldsymbol{\xi}_i, g(\mathbf{y}_i) - g_W(\mathbf{y}_i)\rangle\,, \quad W \in \mathcal{W}.$$

Using (29) we can re-express the log-likelihood ratio as

$$\log\left(\frac{f_W(\mathcal{D})}{f_\star(\mathcal{D})}\right) = -\frac{n}{2\sigma^2}\|g_W - g\|_n^2 - nZ_n(g_W). \tag{36}$$

Let ς_i denote the sub-Gaussian norm of $\|\boldsymbol{\xi}_i\|_2$,

$$\bar{\varsigma} \overset{\text{def}}{=} \max_{1\leq i\leq n} \varsigma_i.$$

The sub-Gaussian assumption on $\boldsymbol{\xi}_i$ implies that $\varsigma_i < \infty$ (see e.g. Theorem 6.3.2 of Vershynin (2018)), and that $\|\boldsymbol{\xi}_i\|_2^2$ is sub-exponential, with sub-exponential norm ς_i^2. We note also that $\mathbb{E}(\|\boldsymbol{\xi}_i\|_2^2) \leq 2\varsigma_i^2$. Therefore, by Bernstein inequality (see e.g. Theorem 2.8.1 of Vershynin (2018)),

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \|\boldsymbol{\xi}_i\|_2^2 > 3\bar{\varsigma}^2\right) \leq \mathbb{P}\left(\sum_{i=1}^n \|\boldsymbol{\xi}_i\|_2^2 - \mathbb{E}(\|\boldsymbol{\xi}_i\|_2^2) > n\bar{\varsigma}^2\right) \leq e^{-c_0 n},$$

for some absolute constant c_0. To make use of this bound, we define

$$\mathcal{F}_0 \overset{\text{def}}{=} \left\{\mathcal{D} :\ \sum_{i=1}^n \|\boldsymbol{\xi}_i\|_2^2 \leq 3n\bar{\varsigma}^2\right\}.$$

Therefore,

$$\mathbb{P}\left[\bigcup_{j\geq 1}\left\{\sup_{W\in\widetilde{\mathcal{W}}^{(j)}} \log\left(\frac{f_W(\mathcal{D})}{f_\star(\mathcal{D})}\right) > -\frac{n\mathsf{r}_j^2}{8\sigma^2}\right\}\right] \leq e^{-c_0 n} + \sum_{j\geq 1}\mathbb{P}\left[\mathcal{F}_j\right],$$

where

$$\mathcal{F}_j \overset{\text{def}}{=} \mathcal{F}_0 \bigcap \left\{\sup_{W\in\widetilde{\mathcal{W}}^{(j)}} \log\left(\frac{f_W(\mathcal{D})}{f_\star(\mathcal{D})}\right) > -\frac{n\mathsf{r}_j^2}{8\sigma^2}\right\}.$$

For each $j \geq 1$, we set

$$\mathcal{W}^{(j)} \overset{\text{def}}{=} \{W \in \mathcal{W}_0 :\ \|g_W - g\|_n \leq \mathsf{r}_{j+1}\},$$

and each $\iota = 1, \ldots$, let $\mathcal{C}_j^{(\iota)} \overset{\text{def}}{=} \{g_{j,1}^{(\iota)}, \ldots, g_{j,N_{j,\iota}}^{(\iota)}\}$ be a $(\mathsf{r}_{j+1}2^{-\iota})$-covering of $\mathcal{W}^{(j)}$. For $\iota = 0$, we set $\mathcal{C}_j^{(0)} = \{g\}$. The definition implies that for any $W \in \mathcal{W}^{(j)}$, we can find

$g_{j,W}^{(\iota)} \in \mathcal{C}_j^{(\iota)}$ such that $\|g_W - g_{j,W}^{(\iota)}\|_n \leq \mathsf{r}_{j+1}2^{-\iota}$. Let $\ell_j \geq 0$, be the smallest integer such that

$$\frac{\mathsf{r}_{j+1}}{2^{\ell_j}} \leq \frac{\mathsf{r}_j^2}{16\bar{\varsigma}}.$$

We consider separately the cases $\ell_j = 0$ and $\ell_j > 0$.

`Suppose` $\ell_j = 0$. In that case for any $W \in \mathcal{W}^{(j)}$, $\|g_W - g\|_n \leq \mathsf{r}_{j+1} \leq \mathsf{r}_j^2/(16\bar{\varsigma})$. Therefore, on the event $\mathcal{F}_0$, we have

$$\sup_{W\in\mathcal{W}_j} |Z_n(g_W)| \leq \frac{1}{n\sigma^2}\sum_{i=1}^n \|\boldsymbol{\xi}_i\|_2\|g_W(\mathbf{y}_i) - g(\mathbf{y}_i)\|_2 \leq \frac{\sqrt{3\bar{\varsigma}^2}}{\sigma^2}\|g_W - g\|_n$$
$$\leq \frac{\sqrt{3\bar{\varsigma}^2}}{\sigma^2}\frac{\mathsf{r}_j^2}{16\bar{\varsigma}} \leq \frac{\mathsf{r}_j^2}{8\sigma^2}.$$

Taking this conclusion to (36) implies that on $\mathcal{F}_0$,

$$\sup_{W\in\widetilde{\mathcal{W}}^{(j)}} \log\left(\frac{f_W(\mathcal{D})}{f_\star(\mathcal{D})}\right) \leq -\frac{n\mathsf{r}_j^2}{2\sigma^2} + n\sup_{W\in\mathcal{W}_j}|Z_n(g_W)| \leq -\frac{n\mathsf{r}_j^2}{4\sigma^2}.$$

Hence, when $\ell_j = 0$, $\mathbb{P}(\mathcal{F}_j) = 0$.

`Suppose` $\ell_j > 0$. Similarly, on the event $\mathcal{F}_0$, we have

$$\left|Z_n(g_W) - Z_n(g_{j,W}^{(\ell_j)})\right| \leq \frac{1}{n\sigma^2}\sum_{i=1}^n \|\boldsymbol{\xi}_i\|_2\|g_{j,W}^{(\ell_j)}(\mathbf{y}_i) - g_W(\mathbf{y}_i)\|_2$$
$$\leq \frac{\sqrt{3\bar{\varsigma}^2}}{\sigma^2}\|g_{j,W}^{(\ell_j)} - g_W\|_n \leq \frac{\sqrt{3\bar{\varsigma}^2}}{\sigma^2}\frac{\mathsf{r}_{j+1}}{2^{\ell_j}} \leq \frac{\sqrt{3\bar{\varsigma}^2}}{\sigma^2}\frac{\mathsf{r}_j^2}{16\bar{\varsigma}} \leq \frac{\mathsf{r}_j^2}{8\sigma^2}.$$

This implies that on $\mathcal{F}_0$,

$$\sup_{W\in\widetilde{\mathcal{W}}^{(j)}} \log\left(\frac{f_W(\mathcal{D})}{f_\star(\mathcal{D})}\right) \leq -\frac{n\mathsf{r}_j^2}{2\sigma^2} + n\sup_{W\in\mathcal{W}^{(j)}}\left|Z_n(g_W) - Z_n(g_{j,W}^{(\ell_j)})\right| + n\sup_{W\in\mathcal{W}^{(j)}}\left|Z_n(g_{j,W}^{(\ell_j)})\right|$$
$$\leq -\frac{3n\mathsf{r}_j^2}{8\sigma^2} + n\sup_{W\in\mathcal{W}^{(j)}}\left|Z_n(g_{j,W}^{(\ell_j)})\right|.$$

Hence

$$\mathbb{P}(\mathcal{F}_j) \leq \mathbb{P}\left[\sup_{W\in\mathcal{W}^{(j)}}\left|Z_n(g_{j,W}^{(\ell_j)})\right| > \frac{\mathsf{r}_j^2}{4\sigma^2}\right].$$

To bound this latter term we introduce

$$\delta_j \overset{\text{def}}{=} \int_{\frac{\mathsf{r}_{j+1}^2}{64\bar{\varsigma}}}^{\mathsf{r}_{j+1}} \sqrt{\log\mathcal{N}\left(\epsilon, \mathcal{W}^{(j)}, \|\cdot\|_n\right)}\mathrm{d}\epsilon,$$

$$\text{and}\quad \eta_{j,\iota} \overset{\text{def}}{=} \max\left(\frac{1}{6}\frac{\iota^{1/2}}{2^\iota}, \frac{\sqrt{\log N_{j,\iota}}}{4\delta_j}\frac{\mathsf{r}_{j+1}}{2^\iota}\right),\quad \iota = 1,\ldots,\ell_j.$$

and we write $g_{j,W}^{(\ell_j)}$ as a telescoping sum

$$g_{j,W}^{(\ell_j)} - g = \sum_{\iota=1}^{\ell_j} g_{j,W}^{(\iota)} - g_{j,W}^{(\iota-1)},$$

so that

$$\sup_{W\in\mathcal{W}^{(j)}} \left| Z_n(g_{j,W}^{(\ell_j)}) \right| \leq \sum_{\iota=1}^{\ell_j} \sup_{W\in\mathcal{W}^{(j)}} \left| \frac{1}{n\sigma^2} \sum_{i=1}^n \left\langle \boldsymbol{\xi}_i, g_{j,W}^{(\iota-1)}(\mathbf{y}_i) - g_{j,W}^{(\iota)}(\mathbf{y}_i) \right\rangle \right|.$$

We show below that the sequence $\{\eta_{j,\iota},\ \iota = 1,\ldots,\ell_j\}$ introduced above satisfies

$$\sum_{\iota=1}^{\ell_j} \eta_{j,\iota} \leq 1. \tag{37}$$

Due to (37), we can use the sequence $\{\eta_{j,\iota},\ \iota = 1,\ldots,\ell_j\}$ to say that

$$\mathbb{P}\left[\sup_{W\in\mathcal{W}^{(j)}} \left| Z_n(g_{j,W}^{(\ell_j)}) \right| > \frac{\mathsf{r}_j^2}{4\sigma^2} \right] \\ \leq \sum_{\iota=1}^{\ell_j} \mathbb{P}\left[\sup_{W\in\mathcal{W}^{(j)}} \left| \frac{1}{n\sigma^2} \sum_{i=1}^n \left\langle \boldsymbol{\xi}_i, g_{j,W}^{(\iota-1)}(\mathbf{y}_i) - g_{j,W}^{(\iota)}(\mathbf{y}_i) \right\rangle \right| > \frac{\eta_{j,\iota}\mathsf{r}_j^2}{4\sigma^2} \right].$$

The supremum on the right-hand side of the last display is in fact a max over a finite set of cardinality at most $N_{j,\iota-1} \times N_{j,\iota} \leq N_{j,\iota}^2$, and for $W \in \mathcal{W}^{(j)}$,

$$\frac{1}{n^2\sigma^4} \sum_{i=1}^n \sigma_i^2 \| g_{j,W}^{(\iota-1)}(\mathbf{y}_i) - g_{j,W}^{(\iota)}(\mathbf{y}_i) \|_2^2 \\ \leq \frac{2\max_i \sigma_i^2}{n^2\sigma^4} \left(n\|g_W - g_{j,W}^{(\iota)}\|_n^2 + n\|g_W - g_{j,W}^{(\iota-1)}\|_n^2 \right) \leq \frac{10}{n} \max_i \left(\frac{\sigma_i^2}{\sigma^4} \right) \frac{\mathsf{r}_{j+1}^2}{2^{2\iota}}.$$

Therefore by Hoefdding's inequality,

$$\mathbb{P}\left[\sup_{W\in\mathcal{W}^{(j)}} \left| Z_n(g_{j,W}^{(\ell_j)}) \right| > \frac{\mathsf{r}_j^2}{4\sigma^2} \right] \leq \sum_{\iota=1}^{\ell_j} \exp\left(2\log N_{j,\iota} - \frac{n2^{2\iota}\eta_{j,\iota}^2\mathsf{r}_j^4}{(20\times 16)\bar{\sigma}^2\mathsf{r}_{j+1}^2} \right).$$

By construction,

$$\frac{2^{2\iota}\eta_{j,\iota}^2}{\mathsf{r}_{j+1}^2} \geq \frac{\log N_{j\iota}}{16\delta_j^2},$$

which gives

$$\frac{n2^{2\iota}\eta_{j,\iota}^2\mathsf{r}_j^4}{(20\times 16)\bar{\sigma}^2\mathsf{r}_{j+1}^2} \geq \frac{n\mathsf{r}_j^4}{(20\times 32^2)\bar{\sigma}^2\delta_j^2} \times (4\log N_{j\iota}) \geq 4\log N_{j\iota},$$

using (35). Therefore

$$2\log N_{j,\iota} - \frac{n2^{2\iota}\eta_{j,\iota}^2 \mathsf{r}_j^4}{(20\times 16)\bar{\sigma}^2\mathsf{r}_{j+1}^2} \leq -\frac{n2^{2\iota}\eta_{j,\iota}^2 \mathsf{r}_j^4}{(20\times 32)\bar{\sigma}^2\mathsf{r}_{j+1}^2} \leq -\frac{n\mathsf{r}_j^2\iota}{(80\times 36\times 32)\bar{\sigma}^2},$$

where the last inequality uses the fact that $2^{2\iota}\eta_{j,\iota}^2 \geq \iota/36$. It follows that

$$\mathbb{P}\left[\sup_{W\in\mathcal{W}^{(j)}} \left|Z_n(g_{j,W}^{(\ell_j)})\right| > \frac{\mathsf{r}_j^2}{4\sigma^2}\right] \leq \sum_{\iota=1}^{\ell_j} \exp\left(-\frac{n\mathsf{r}_j^2\iota}{c_0\bar{\sigma}^2}\right) \leq 2\exp\left(-\frac{n\mathsf{r}_j^2}{c_0\bar{\sigma}^2}\right),$$

since $n\mathsf{r}_j^2 \geq c_0\bar{\sigma}^2\log(2)$, for some constant c_0 that can be taken as $c_0 = 80\times 36\times 32$. In conclusion,

$$\mathbb{P}\left[\bigcup_{j\geq 1}\left\{\sup_{W\in\widetilde{\mathcal{W}}^{(j)}} \log\left(\frac{f_W(\mathcal{D})}{f_\star(\mathcal{D})}\right) > -\frac{n\mathsf{r}_j^2}{8\sigma^2}\right\}\right]$$
$$\leq e^{-c_0 n} + 2\sum_{j\geq 1}\exp\left(-\frac{n\mathsf{r}_j^2}{c_0\bar{\sigma}^2}\right) \leq e^{-c_0 n} + 4\exp\left(-\frac{nM\mathsf{r}^2}{c_0\bar{\sigma}^2}\right).$$

To check (37), we note

$$\sum_{\iota=1}^{\ell_j}\eta_{j,\iota} \leq \frac{1}{6}\sum_{\iota=1}^{\ell_j}\frac{\iota^{1/2}}{2^\iota} + \frac{1}{4\delta_j}\sum_{\iota=1}^{\ell_j}\frac{\mathsf{r}_{j+1}}{2^\iota}\sqrt{\log N_{j,\iota}}.$$

The function $h(x) = x^{1/2}2^{-x} = x^{\alpha-1}e^{-\beta x}$, with $\alpha = 3/2$, $\beta = \log(2)$ is decreasing for $x\geq 1$. Hence

$$\sum_{\iota\geq 1}\frac{\iota^{1/2}}{2^\iota} = \frac{1}{2} + \sum_{\iota\geq 2}h(\iota) \leq \frac{1}{2} + \sum_{k\geq 2}\int_{k-1}^{k}h(x)\mathrm{d}x \leq \frac{1}{2} + \int_1^\infty x^{\alpha-1}e^{\beta x}\mathrm{d}x \leq 3.$$

Whereas,

$$\sum_{\iota=1}^{\ell_j}\frac{\mathsf{r}_{j+1}}{2^\iota}\sqrt{\log N_{j,\iota}} = \sum_{\iota=1}^{\ell_j} 2\int_{\frac{\mathsf{r}_{j+1}}{2^{\iota+1}}}^{\frac{\mathsf{r}_{j+1}}{2^\iota}} \sqrt{\log\mathcal{N}\left(\frac{\mathsf{r}_{j+1}}{2^\iota}, \mathcal{W}^{(j)}, \|\cdot\|_n\right)}\mathrm{d}\epsilon$$
$$\leq 2\int_{\frac{\mathsf{r}_{j+1}}{2^{\ell_j+1}}}^{\frac{\mathsf{r}_{j+1}}{2}} \sqrt{\log\mathcal{N}\left(\epsilon, \mathcal{W}^{(j)}, \|\cdot\|_n\right)}\mathrm{d}\epsilon$$
$$\leq 2\int_{\frac{\mathsf{r}_{j+1}^2}{64\bar{\varsigma}}}^{\mathsf{r}_{j+1}} \sqrt{\log\mathcal{N}\left(\epsilon, \mathcal{W}^{(j)}, \|\cdot\|_n\right)}\mathrm{d}\epsilon = 2\delta_j.$$

□

A.2. **Proof of Theorem 5.** We apply Theorem 7. The argument has two main steps. First, we show that the function g can be well approximated by elements of the function class $\{g_W,\ W \in \mathcal{W}\}$ constructed in (11), and secondly we show that the functions g_W are locally Lipschitz and we estimate the local Lipschitz constant. Both steps rely on a well-known telescoping argument that we outline first (see e.g. Proposition 6 of Taheri et al. (2021)). Given two functions $f = f_K \circ \cdots \circ f_1$, and $g = g_K \circ \cdots \circ g_1$, we write $f - g$ as a telescoping sum

$$f(\mathbf{x})-g(\mathbf{x}) = \sum_{j=1}^{K} f_K\circ\cdots\circ f_j\left(g_{j-1}\circ\cdots\circ g_1(\mathbf{x})\right)-f_K\circ\cdots\circ f_{j+1}\circ g_j\left(g_{j-1}\circ\cdots\circ g_1(\mathbf{x})\right), \tag{38}$$

with the convention that for $j = 1$, $g_{j-1} \circ \cdots \circ g_1$ is the identity map, and for $j = K$, $f_K \circ \cdots \circ f_{j+1}$ is the identity map. A bound on $\|f(\mathbf{x}) - g(\mathbf{x})\|$ can then be derived using the Lipschitz and boundedness properties of the functions f_j, g_j.

Specifically, define $H_W^{(0)}(\mathbf{x}) \overset{\text{def}}{=} \mathbf{x}$, and for $1 \le \ell \le D$, define $H_W^{(\ell)}(\mathbf{x}) \overset{\text{def}}{=} \Psi_{W_\ell}^{(\ell)}(H_W^{(\ell-1)}(\mathbf{x}))$, so that $H_W(\mathbf{x}) = H_W^{(D)}(\mathbf{x})$. We recall that

$$\Psi_M^{(\ell)}(\mathbf{x}) = \mathsf{a}_\ell(M\mathbf{x}),$$

where the activation functions $\mathsf{a}_\ell :\ \mathbb{R}^{p_\ell} \to \mathbb{R}^{p_\ell}$ are Lipschitz with constant 1. Then for $1 \le \ell \le D$, and all $W, W' \in \mathcal{W}$, $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^{d_x}$, by the Lipschitz property of the activation functions a_ℓ, we have

$$\begin{aligned}\|H_W^{(\ell)}(\mathbf{x}_1) - H_W^{(\ell)}(\mathbf{x}_2)\|_2 &\le \|W_\ell H_W^{(\ell-1)}(\mathbf{x}_1) - W_\ell H_W^{(\ell-1)}(\mathbf{x}_2)\|_2 \\ &\le \|W_\ell\|_{\mathsf{op}}\|H_W^{(\ell-1)}(\mathbf{x}_1) - H_W^{(\ell-1)}(\mathbf{x}_2)\|_2,\end{aligned}$$

where $\|\cdot\|_{\mathsf{op}}$ denotes the operator norm. Iterating this yields,

$$\|H_W^{(\ell)}(\mathbf{x}_1) - H_W^{(\ell)}(\mathbf{x}_2)\|_2 \le \prod_{j=1}^{\ell} \|W_j\|_{\mathsf{op}}\|\mathbf{x}_1 - \mathbf{x}_2\|_2. \tag{39}$$

Similarly, for any $1 \le \ell \le D$, (38) gives

$$\begin{aligned}H_W^{(\ell)}(\mathbf{x}) - H_{W'}^{(\ell)}(\mathbf{x}) = \sum_{j=1}^{\ell} &\Psi_{W_\ell}^{(\ell)} \circ \cdots \circ \Psi_{W_{j+1}}^{(j+1)} \circ \Psi_{W_j}^{(j)} \circ \left(\Psi_{W'_{j-1}}^{(j-1)} \circ \cdots \circ \Psi_{W'_1}^{(1)}(\mathbf{x})\right) \\ &- \Psi_{W_\ell}^{(\ell)} \circ \cdots \circ \Psi_{W_{j+1}}^{(j+1)} \circ \Psi_{W'_j}^{(j)} \left(\Psi_{W'_{j-1}}^{(j-1)} \circ \cdots \circ \Psi_{W'_1}^{(1)}(\mathbf{x})\right).\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left\|H_W^{(\ell)}(\mathbf{x}) - H_{W'}^{(\ell)}(\mathbf{x})\right\|_2 \\
&\le \sum_{j=1}^{\ell} \prod_{k=j+1}^{\ell} \|W_k\|_{\mathsf{op}} \left\|\Psi_{W_j}^{(j)}\left(\Psi_{W'_{j-1}}^{(j-1)} \circ \cdots \circ \Psi_{W'_1}^{(1)}(\mathbf{x})\right) - \Psi_{W'_j}^{(j)}\left(\Psi_{W'_{j-1}}^{(j-1)} \circ \cdots \circ \Psi_{W'_1}^{(1)}(\mathbf{x})\right)\right\|_2 \\
&\qquad\qquad \le \sum_{j=1}^{\ell} \prod_{k=j+1}^{\ell} \|W_k\|_{\mathsf{op}} \|W_j - W'_j\|_{\mathsf{op}} \left\|\Psi_{W'_{j-1}}^{(j-1)} \circ \cdots \circ \Psi_{W'_1}^{(1)}(\mathbf{x})\right\|_2 .
\end{aligned}$$

Since $\Psi_M^{(j)}(\mathbf{0}) = \mathbf{0}$, we have the bound

$$\left\|\Psi_{W'_{j-1}}^{(j-1)} \circ \cdots \circ \Psi_{W'_1}^{(1)}(\mathbf{x})\right\|_2 \le \|\mathbf{x}\|_2 \prod_{k=1}^{j-1} \|W'_j\|_{\mathsf{op}}.$$

In conclusion, for all $1 \le \ell \le D$, $W, W' \in \mathcal{W}$, and for all $\mathbf{x} \in \mathbb{R}^{d_x}$, we have

$$\|H_W^{(\ell)}(\mathbf{x}) - H_{W'}^{(\ell)}(\mathbf{x})\|_2 \le \|\mathbf{x}\|_2 \sum_{j=1}^{\ell} \|W_j - W'_j\|_{\mathsf{op}} \prod_{k=1}^{j-1} \|W'_k\|_{\mathsf{op}} \prod_{k=j+1}^{\ell} \|W_k\|_{\mathsf{op}}. \tag{40}$$

For $\mathbf{x} \in \mathbb{R}^{d_x}$, $\mathbf{y} \in \mathsf{Y}$, we define

$$F_{\mathbf{y}}(\mathbf{x}) \stackrel{\text{def}}{=} \operatorname{Prox}^{\gamma\mathcal{R}}\left(\mathbf{x} - \gamma\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})\right), \quad \text{and} \quad F_{\mathbf{y},W}(\mathbf{x}) \stackrel{\text{def}}{=} H_W\left(\mathbf{x} - \gamma\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})\right).$$

We use the notation h^k to denote the function h composed k times with the convention that h^0 is the identity map. Hence $g_W(\mathbf{y}) = F_{\mathbf{y},W}^{D'}(\mathbf{x}^{(0)})$. H2 implies that $F_{\mathbf{y}}$ is non-expansive.

Lemma 14. *Assume H2. Given $\epsilon > 0$, we can find $W \in \mathcal{W}$ as described in H2 such that*

$$\max_{1\le i\le n} \|g_W(\mathbf{y}_i) - g(\mathbf{y}_i)\|_2 \le R_0\varrho_n^{D'} + D'\epsilon.$$

Proof. For any $\mathbf{y} \in \mathbb{R}^{d_y}$, we can write

$$g(\mathbf{y}) - g_W(\mathbf{y}) = g(\mathbf{y}) - F_{\mathbf{y},W}^{D'}(\mathbf{x}^{(0)}) = g(\mathbf{y}) - F_{\mathbf{y}}^{D'}(\mathbf{x}^{(0)}) + F_{\mathbf{y}}^{D'}(\mathbf{x}^{(0)}) - F_{\mathbf{y},W}^{D'}(\mathbf{x}^{(0)}).$$

By Assumption 2, we have

$$\max_{1\le i\le n} \left\|g(\mathbf{y}_i) - F_{\mathbf{y}_i}^{D'}(\mathbf{x}^{(0)})\right\|_2 \le R_0\varrho_n^{D'}.$$

For $\mathbf{y} \in \mathbb{R}^{d_y}$, let $G_{\gamma,\mathbf{y}}(\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{x} - \gamma\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})$, and

$$R'_1 \stackrel{\text{def}}{=} \max_{1\le i\le n} \max_{\mathbf{x}\in\mathbb{R}^{d_x}:\, \|\mathbf{x}\|_2\le R} \|G_{\gamma,\mathbf{y}_i}(\mathbf{x})\|_2 .$$

Since $F_{\mathbf{y}}$ is non-expansive, by the telescoping argument (38), we have

$$\left\|F_{\mathbf{y}_i}^{D'}(\mathbf{x}^{(0)}) - F_{\mathbf{y}_i,W}^{D'}(\mathbf{x}^{(0)})\right\|_2 \leq \sum_{j=1}^{D'} \|F_{\mathbf{y}_i,W}\left(F_{\mathbf{y}_i,W}^{j-1}(\mathbf{x}^{(0)})\right) - F_{\mathbf{y}_i}\left(F_{\mathbf{y}_i,W}^{j-1}(\mathbf{x}^{(0)})\right)\|_2,$$
$$\leq D' \sup_{\mathbf{x}\in\mathbb{R}^{d_x},\ \|\mathbf{x}\|_2\leq R_1'} \|H_W(\mathbf{x}) - \mathrm{Prox}^{\gamma\mathcal{R}}(\mathbf{x})\|_2 \leq D'\epsilon.$$

The result follows by taking the max over i. □

Lemma 15. *Assume H2, H3, and let $\{g_W,\ W\in\mathcal{W}\}$ be as in (11). For any $\eta>0$, and any $W,W'\in\mathcal{W}$, such that $\max(\|W\|_2,\|W'\|_2)\leq\eta$, we have*

$$\max_{1\leq i\leq n} \|g_W(\mathbf{y}_i) - g_{W'}(\mathbf{y}_i)\|_2 \leq L(\eta)\|W-W'\|_2,$$

with

$$L(\eta) \overset{\text{def}}{=} C_n\left(e^4 + \frac{\eta^2}{D}\right)^{DD'},$$

and

$$C_n \overset{\text{def}}{=} \|\mathbf{x}^{(0)}\|_2 + \max_{1\leq i\leq n}\|\mathbf{x}^{(0)} - \gamma\nabla f(\mathbf{y}_i|\mathbf{x}^{(0)})\|_2.$$

Proof. We recall that the convexity of $\mathbf{x}\mapsto f(\mathbf{y}|\mathbf{x})$ and the choice of the step-size assumed in H2 imply that the function $\mathbf{x}\mapsto\mathbf{x}-\gamma\nabla_{\mathbf{x}}f(\mathbf{y}|\mathbf{x})$ is non-expansive on $\mathbb{R}^{d_x}$. First, we apply (39) to obtain that for all $\mathbf{x}_1,\mathbf{x}_2\in\mathbb{R}^{d_x}$,

$$\|H_W(\mathbf{x}_1) - H_W(\mathbf{x}_2)\|_2 \leq \lambda_W\|\mathbf{x}_1-\mathbf{x}_2\|_2, \quad \text{where} \quad \lambda_W \overset{\text{def}}{=} \prod_{\ell=1}^{D}\|W_\ell\|_{\mathsf{op}}\vee 1.$$

It follows that for all $\mathbf{y}\in\mathsf{Y}$, $\mathbf{x}_1,\mathbf{x}_2\in\mathbb{R}^{d_x}$,

$$\|F_{\mathbf{y},W}(\mathbf{x}_1) - F_{\mathbf{y},W}(\mathbf{x}_2)\|_2$$
$$\leq \lambda_W\|\mathbf{x}_1-\mathbf{x}_2-\gamma\left(\nabla_{\mathbf{x}}f(\mathbf{y}|\mathbf{x}_1)) - \nabla_{\mathbf{x}}f(\mathbf{y}|\mathbf{x}_2)\right)\|_2 \leq \lambda_W\|\mathbf{x}_1-\mathbf{x}_2\|_2. \quad (41)$$

Using (41), and the telescoping identity (38), we obtain

$$\|g_W(\mathbf{y}) - g_{W'}(\mathbf{y})\|_2 \leq \sum_{j=1}^{D'}\lambda_W^{D'-j}\left\|F_{\mathbf{y},W}\left(F_{\mathbf{y},W'}^{j-1}(\mathbf{x}^{(0)})\right) - F_{\mathbf{y},W'}\left(F_{\mathbf{y},W'}^{j-1}(\mathbf{x}^{(0)})\right)\right\|_2. \quad (42)$$

We set $G_{\gamma,\mathbf{y}}(\mathbf{x}) \overset{\text{def}}{=} \mathbf{x}-\gamma\nabla f(\mathbf{y}|\mathbf{x})$. We apply (39) with $\mathbf{x}_2=\mathbf{0}$ and the non-expansiveness of $G_{\gamma,\mathbf{y}}$ to write that for all $\mathbf{x}\in\mathbb{R}^{d_x}$

$$\|F_{\mathbf{y},W}(\mathbf{x})\|_2 = \|H_W\left(G_{\gamma,\mathbf{y}}(\mathbf{x})\right)\|_2 \leq \lambda_W\|G_{\gamma,\mathbf{y}}(\mathbf{x}) - G_{\gamma,\mathbf{y}}(\mathbf{x}^{(0)}) + G_{\gamma,\mathbf{y}}(\mathbf{x}^{(0)})\|_2$$
$$\leq \lambda_W\left(\|\mathbf{x}\|_2 + \|\mathbf{x}^{(0)}\|_2 + \|G_{\gamma,\mathbf{y}}(\mathbf{x}^{(0)})\|_2\right).$$

By iterating this inequality we obtain that for all for all $\mathbf{x} \in \mathbb{R}^{d_x}$

$$\|F^j_{\mathbf{y},W}(\mathbf{x}^{(0)})\|_2 \leq \left(\sum_{\ell=1}^{j} \lambda_W^\ell\right) \left(\|\mathbf{x}^{(0)}\|_2 + \|G_{\gamma,\mathbf{y}}(\mathbf{x}^{(0)})\|_2\right) \leq C_n \sum_{\ell=1}^{j} \lambda_W^\ell, \tag{43}$$

where

$$C_n \overset{\text{def}}{=} \|\mathbf{x}^{(0)}\|_2 + \max_{1\leq i\leq n} \|\mathbf{x}^{(0)} - \gamma \nabla f(\mathbf{y}_i|\mathbf{x}^{(0)})\|_2.$$

Setting

$$\lambda_{W,W'} \overset{\text{def}}{=} \sum_{j=1}^{D} \|W_j - W'_j\|_{\mathsf{op}} \prod_{k=1}^{j-1} \|W'_k\|_{\mathsf{op}} \prod_{k=j+1}^{D} \|W_k\|_{\mathsf{op}},$$

we then apply (40) to write that for all $\mathbf{x} \in \mathbb{R}^{d_x}$

$$\begin{aligned}
&\left\|F_{\mathbf{y},W}\left(F^{j-1}_{\mathbf{y},W'}(\mathbf{x}^{(0)})\right) - F_{\mathbf{y},W'}\left(F^{j-1}_{\mathbf{y},W'}(\mathbf{x}^{(0)})\right)\right\|_2 \\
&\qquad = \left\|H_W \circ G_{\gamma,\mathbf{y}}\left(F^{j-1}_{\mathbf{y},W'}(\mathbf{x}^{(0)})\right) - H_{W'} \circ G_{\gamma,\mathbf{y}}\left(F^{j-1}_{\mathbf{y},W'}(\mathbf{x}^{(0)})\right)\right\|_2 \\
&\qquad \leq \lambda_{W,W'} \left\|G_{\gamma,\mathbf{y}}\left(F^{j-1}_{\mathbf{y},W'}(\mathbf{x}^{(0)})\right) - G_{\gamma,\mathbf{y}}(\mathbf{x}^{(0)}) + G_{\gamma,\mathbf{y}}(\mathbf{x}^{(0)})\right\|_2 \\
&\qquad \leq \lambda_{W,W'} \left(\|F^{j-1}_{\mathbf{y},W'}(\mathbf{x}^{(0)})\|_2 + \|\mathbf{x}^{(0)}\|_2 + \|G_{\gamma,\mathbf{y}}(\mathbf{x}^{(0)}\|_2\right) \\
&\qquad \leq C_n \lambda_{W,W'} \sum_{\ell=0}^{j-1} \lambda_{W'}^\ell.
\end{aligned}$$

The last display together with (42) yields,

$$\max_{1\leq i\leq n} \|g_W(\mathbf{y}_i) - g_{W'}(\mathbf{y}_i)\|_2 \leq C_n \lambda_{W,W'} \sum_{j=1}^{D'} \sum_{\ell=0}^{j-1} \lambda_W^{D'-j} \lambda_{W'}^\ell. \tag{44}$$

Since the geometric mean is never larger than the arithmetic mean, we have

$$\lambda_W \overset{\text{def}}{=} \prod_{j=1}^{D} 1 \vee \|W_j\|_{\mathsf{op}} \leq \left(\frac{1}{D}\sum_{j=1}^{D} 1 \vee \|W_j\|_{\mathsf{op}}^2\right)^{D/2} \leq \left(1 + \frac{\|W\|_2^2}{D}\right)^{D/2}.$$

Therefore, for $\max(\|W\|_2, \|W'\|_2) \leq \eta$,

$$\sum_{j=1}^{D'} \sum_{\ell=0}^{j-1} \lambda_W^{D'-j} \lambda_{W'}^\ell \leq \sum_{j=1}^{D'} j \lambda_W^{D'-j} \lambda_{W'}^{j-1} \leq (D')^2 \left(1 + \frac{\eta^2}{D}\right)^{D(D'-1)/2},$$

and similarly,

$$\lambda_{W,W'} \leq \sqrt{D}\left(1 + \frac{2\eta^2}{D}\right)^{D/2} \|W - W'\|_2.$$

Hence, we conclude that

$$\max_{1\leq i\leq n}\|g_W(\mathbf{y}_i)-g_{W'}(\mathbf{y}_i)\|_2\leq C_n\sqrt{D}(D')^2\left(1+\frac{2\eta^2}{D}\right)^{DD'/2}\|W-W'\|_2. \tag{45}$$

The statement in the lemma follows by noting that

$$\sqrt{D}(D')^2\left(1+\frac{2\eta^2}{D}\right)^{DD'/2}\leq\sqrt{D}(D')^2\left(e^4+\frac{2\eta^2}{D}\right)^{DD'/2}\leq\left(e^4+\frac{2\eta^2}{D}\right)^{DD'},$$

using the fact that $A^{x/2}\geq x$ for all $x\geq 1$, and $A\geq e$, and $A^{x/2}\geq x^2$ for all $x\geq 1$, and $A\geq e^4$. □

A.2.1. *Proof of Theorem 5.*

Proof. We recall the notation $a\lesssim b$ means that $a\leq cb$, for some constant c that does not depend on the sample size n. Fix

$$\varpi_\star=\log(n)\sqrt{d_x}\left(\frac{\log(q)}{n}\right)^{\frac{1}{2+\beta_2}},\quad\text{and}\quad\frac{\log\left(\frac{2R_0}{\varpi_\star}\right)}{-\log(\rho)}\leq D'\leq n.$$

By Assumption 3, and Lemma 14, by taking a deep neural network function H_W, with depth $D=D_0\log(2D'\sqrt{d_x}/\varpi_\star)$, layer size $(p_0,\ldots,p_D)$ all at most $N_0(2D'\sqrt{d_x}/\varpi_\star)^{\beta_1}$, and $W\in\mathcal{W}$ with sparsity at most $s_\star=s_0(2D'\sqrt{d_x}/\varpi_\star)^{\beta_2}$, and we achieve

$$\begin{aligned}\max_{1\leq i\leq n}\|g_W(\mathbf{y}_i)-g(\mathbf{y}_i)\|_2&\leq R_0\rho^{D'}\\&\quad+D'\sup_{\mathbf{x}:\,\|\mathbf{x}\|_2\leq R_0'}\|H_W(\mathbf{x})-\mathrm{Prox}^{\gamma\mathcal{R}}(\mathbf{x})\|_2\\&\leq R_0\frac{\varpi_\star}{2R_0}+D'\frac{\varpi_\star}{2D'}=\varpi_\star.\end{aligned}$$

Then by Lemma 15, the term $\mathsf{L}_\star$ in Theorem 7 scales like

$$\begin{aligned}\mathsf{L}_\star&\simeq\left(e^4+\frac{s_\star}{D}\right)^{DD'}s_\star^{1/2}\\&\lesssim\left(e^4+\frac{s_\star}{D}\right)^{DD'}\left(1+\frac{s_\star}{D}\right)^{D/2}\\&\simeq\left(e^4+\frac{s_\star}{D}\right)^{D(D'+1/2)}\\&\lesssim\left(e^4+q\right)^{D(D'+1/2)}\end{aligned}$$

It follows that

$$\frac{\log\mathsf{L}_\star}{\log(q)}\lesssim DD'\lesssim D'\log(n),\quad\text{and}\quad s_\star=s_0\left(\frac{2D'\sqrt{d_x}}{\varpi_\star}\right)^{\beta_2}\lesssim\left(\frac{D'}{\log(n)}\right)^{\beta_2}\left(\frac{n}{\log(q)}\right)^{\frac{\beta_2}{2+\beta_2}},$$

and the term s in Theorem 7 is of order

$$\begin{aligned} s \lesssim & s_\star \left(\frac{\log(n)}{\log(q)} + \frac{\log(\mathsf{L}_\star)}{\log(q)} \right) + \frac{n \varpi_\star^2}{\log(q)} \\ \lesssim & \left[\big(1 + D' \log(n)\big) \left(\frac{D'}{\log(n)} \right)^{\beta_2} + (\log(n))^2 \right] \left(\frac{n}{\log(q)} \right)^{\frac{\beta_2}{2+\beta_2}} \\ \lesssim & \left(D'\right)^{1+\beta_2} (\log(n))^2 \left(\frac{n}{\log(q)} \right)^{\frac{\beta_2}{2+\beta_2}} . \end{aligned}$$

Therefore the term L_s in Theorem 7 scales like

$$\mathsf{L}_s = L(s^{1/2} b_s) \lesssim \left(e^4 + \frac{s b_s^2}{D} \right)^{D(D'+1/2)},$$

which gives,

$$\log(\mathsf{L}_s) \lesssim D D' \log(q) \lesssim D' \log(n) \log(q).$$

Noting that

$$\frac{s}{n} \lesssim \left(D'\right)^{1+\beta_2} \left(\frac{\log(q)}{n} \right)^{\frac{2}{2+\beta_2}} \frac{(\log(n))^2}{\log(q)},$$

we deduce that the conclusion of Theorem 7 holds with a rate

$$\mathsf{r} = \bar{\sigma} \sqrt{\frac{s \log(q) + s \log(\mathsf{L}_s)}{n}} \lesssim \bar{\sigma} \left(D'\right)^{1+\beta_2/2} (\log(n))^{3/2} \left(\frac{\log(q)}{n} \right)^{\frac{1}{2+\beta_2}} .$$

□