Journal of Machine Learning Research 26 (2025) 1-47 Submitted 11/23; Revised 7/25; Published 12/25

A statistical perspective on algorithm unrolling models for inverse problems

Yves Atchadé ATCHADE@BU.EDU
Department of Mathematics and Statistics
Boston University
Boston, MA 02215, USA

Xinru Liu XINRULIU@BU.EDU
Department of Mathematics and Statistics
Boston University
Boston, MA 02215, USA

Qiuyun Zhu QIUYUN-ZHU@UIOWA.EDU
Department of Statistics and Actuarial Science
The University of Iowa
Iowa City, IA 52242, USA

Editor: Maxim Raginsky

Abstract

We consider inverse problems where the forward model, that is the conditional distribution of the observation $\mathbf{y} \in \mathbb{R}^{d_y}$ given the latent variable of interest $\mathbf{x} \in \mathbb{R}^{d_x}$ is known, and access is given to a data set in which multiple instances of $(\mathbf{x}, \mathbf{y})$ are 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 properties of the gradient descent network (GDN), a well-known architecture driven by proximal gradient descent that epitomizes unrolling learning. Under some regularity conditions, we show that when $d_y \geq d_x$, the GDN estimator solves the inverse problem at a statistical rate faster than the nonparametric minimax rate achievable while ignoring the forward model. Furthermore, when the negative log-density of the latent variable $\mathbf{x}$ has a simple proximal operator, we show that GDN achieves the parametric rate $O(1/\sqrt{n})$. Furthermore, our results are explicit in the unrolling depth of the network and suggest that unrolling models are typically prone to overfitting as the unrolling depth increases, and careful tuning as function of the sample size is required for best performances. We provide several examples to illustrate these results.

Keywords: inverse problems, algorithm unrolling models, nonparametric regression, Bayesian deep learning, gradient descent networks, posterior contraction

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

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

©2025 Yves Atchadé, Xinru Liu and Qiuyun Zhu.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v26/23-1512.html.

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 and 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

$$\pi(\mathrm{d}\mathbf{x}|\mathbf{y}_i) \propto \exp\left(-\mathcal{R}(\mathbf{x}) - f(\mathbf{y}_i|\mathbf{x})\right)\mathrm{d}\mathbf{x}, \tag{5}$$

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

$$g(\mathbf{y}) \stackrel{\text{def}}{=} \operatorname*{Argmin}_{\mathbf{x}\in\mathbb{R}^{d_x}} \left[f(\mathbf{y}|\mathbf{x}) + \mathcal{R}(\mathbf{x})\right]. \tag{6}$$

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 $\pi(\cdot|\mathbf{y})$ is typically unknown, and the main goal (and the main challenge) 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 $\pi(\cdot|\mathbf{y})$.

Algorithm unrolling models learn to solve the inverse problem by fitting a regression of $\mathbf{x}$ on $\mathbf{y}$ using the dataset $\mathcal{D}$, bypassing the need to estimate μ. Once trained, the unrolling model can be used to solve new instances of the inverse problem much faster than alternative methods that focus on computing (6), or sampling from (5). The approach has become popular in computational imaging over the last decade (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 and 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 and LeCun (2010); Sreter and 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

We analyze in this work the generalization error of the descent neural network (GDN), one of the simplest deep learning unrolling models (Gregor and LeCun, 2010; Kamilov and Mansour, 2016). Specifically we consider the nonparametric regression model

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

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 GDN function obtained by unrolling D' times a parametrized proximal gradient descent algorithm for solving (6). We give precise definition below. The architecture thus makes explicit use of the forward map f. We develop a Bayesian framework for estimating (7) and we analyze the statistical performance of the resulting estimator of g, assuming that the function $\mathcal{R}$ is convex, but not necessarily differentiable. We are particularly interested in the statistical estimation rate of GDN compared to classical minimax estimation rate for estimating g, and the effect of the unrolling depth D' on performances. Our main contributions are as follow:

1. Under some regularity conditions, particularly assuming that the conditional distribution of $\mathbf{x}_i$ given $\mathbf{y}_i$ is approximately Gaussian with mean $g(\mathbf{y}_i)$, and ignoring logarithmic terms, we show that the GDN for estimating g achieves the statistical error rate

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

for some constant C_1 that depends on the input dimensions d_x, d_y (however the dependence can potentially be poor[1]). 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. Indeed, in the regime $d_y \geq d_x$, the forward model provides more information than unknown, which is adequately leverage by the GDN.

2. We show that the convergence rate of the estimator can be faster than the aforementioned rate. Indeed, when the proximal map of $\mathcal{R}$ is simple and can be well-approximated by a simple neural network function, we show that the GDN architecture achieves a faster statistical rate. For instance, if μ is as in (21) below, a common assumption in image restoration tasks, then ignoring log terms, our result shows that the GDN achieves the parametric rate $O(D'/\sqrt{n})$, for some dimension-dependent constant.

3. Importantly, our result also 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 behavior of the model as D' increases. The same overfitting behavior also appear empirically with the Neumann Network model of (Gilton et al., 2020), suggesting indeed a fundamental behavior. This result suggests that careful tuning of the unrolling depth as function of the sample size is needed for optimal performance.

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, highlighting the need to carefully control 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 and Klusowski (2018); Schmidt-Hieber (2020); Taheri et al. (2021); Ee et al. (2020)). Our framework is closer to Polson and 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

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))

computationally tractable using the sparse asynchronous SGLD algorithm of Atchade and Wang (2023).

Finally we contrast our nonparametric regression approach with the two-step approach proposed for instance by Chang et al. (2017), where the proximal operator of $\mathcal{R}$ is first estimated from the dataset $\mathcal{D}$, and g is then estimated by solving (6) using the estimated proximal operator obtained from the first step. With the rise of diffusion models for estimating marginal distributions and their scores (Ho et al., 2020; Song and Ermon, 2019), variations of this two-step approach have become very popular in recent years (see e.g. Wu et al. (2024) and the review papers Daras et al. (2024); Uehara et al. (2025)). However more research is needed on the comparative strengths and limitations of these two approaches.

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, if A is a square matrix, $\lambda_{\min}(A)$ (resp. $\lambda_{\max}(A)$) denotes the smallest (resp. largest) eigenvalue of A.

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 $W = (W_D, \ldots, W_1) \in \mathcal{W} \stackrel{\text{def}}{=} \mathbb{R}^{p_D \times p_{D-1}} \times \cdots \times \mathbb{R}^{p_1 \times p_0}$. By vectorization, we will view $\mathcal{W}$ as the Euclidean space $\mathbb{R}^q$ (where $q \stackrel{\text{def}}{=} \sum_{\ell=1}^D p_\ell p_{\ell-1}$), 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.

H1 *We have a data set* $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i),\ 1 \le i \le n\}$ *of i.i.d. samples generated according to (4) such that for* $1 \le i \le n$,

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

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* $\bar{\sigma} < \infty$.

Since the unrolling model reduces the inverse problem to a regression model, some assumption of the form H1 is needed for the approach to make sense. H1 requires the conditional distribution of $\mathbf{x}_i$ given $\mathbf{y}_i$ to satisfy a sufficient concentration properties around its mode $g(\mathbf{y}_i)$ which is a type of Bernstein-von Mises (BvM) phenomenon (van der Vaart, 1998; Johnstone, 2010). The BvM theorem is a cornerstone result of Bayesian statistics and gives conditions under which a posterior distribution becomes approximately Gaussian with a center at the maximum likelihood estimator (MLE), or the penalized MLE.

The BvM theorem is well-known to hold for low-dimensional regular models (see e.g. van der Vaart (1998)). In this regime, under some standard regularity conditions it is known that the prior becomes asymptotically irrelevant and the posterior distribution is approximately centered at the MLE, with inverse Fisher information as covariance matrix. For instance, suppose that the forward model is (2), and the negative log-prior in (4) is $\mathcal{R}(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_0)^\top C^{-1}(\mathbf{x} - \mathbf{x}_0)$, for some mean $\mathbf{x}_0$ and a symmetric positive definite matrix C. Then clearly the conditional distribution $\pi(\mathbf{x}|\mathbf{y}_i)$ is exactly Gaussian with mean $g(\mathbf{y}_i)$ and covariance matrix $(C^{-1} + A'A/v^2)^{-1}$, and H1 obvioulsly holds. Even if $\mathcal{R}$ is not quadratic, if we keep d_x fixed, and let $d_y \to \infty$, then the BvM theorem holds and states that if $\mathbf{x}_i$ is a draw from $\pi(\cdot|\mathbf{y}_i)$, then $\sqrt{d_y}(\mathbf{x}_i - g(\mathbf{y}_i))$ converges weakly to a mean-zero Gaussian distribution. We refer the reader to Schervish (1996) Theorem 7.89 for a statement. Hence for any given d_y large we can write $\mathbf{x}_i = g(\mathbf{y}_i) + \boldsymbol{\xi}_i$, where $\boldsymbol{\xi}_i$ is approximately a mean-zero Gaussian random variable, which is H1.

These kind of results can also be derived in the high-dimensional regime where d_x and d_y have roughly the same dimension, and even in infinite-dimensional spaces where $\mathbf{x}$ is a function (Johnstone (2010); Bickel and Kleijn (2012); Nickl (2017)). In these regimes, the prior typically does not vanish, and the posterior distribution concentrates around the corresponding penalyzed MLE. In these new developments, the sparse signal recovery is the most well-developed (Castillo et al., 2015). In many inverse problems the support of the marginal distribution μ lays in a much smaller (but unknown) submanifold of $\mathbb{R}^{d_x}$. This is the so-called manifold hypothesis widely accepted in machine learning (see e.g. Cayton et al. (2008); Fefferman et al. (2016); Whiteley et al. (2025) and the references therein). Under the manifold hypothesis we also expect the BvM phenomenon to hold. This is a topic of current research with several noteworthy results (Bhattacharya and Dunson, 2010; Tang and Yang, 2023; Berenfeld et al., 2024).

To summarize, assumption H1 is a fundamental prerequisite for the validity of unrolling models. The assumption captures a well-understood statistical phenomenon that is known to hold in many settings – but not all. However checking that the BvM theorem holds on specific models remains a technically challenging and on-going research problem in statistics that is beyond the scope of this work.

By the i.i.d. condition from H1 and non-randomness assumption on sub-Gaussian norm, we can deduce each $\boldsymbol{\xi}_i$ given $\mathbf{y}_i$ shares the same sub-Gaussian norm. Let $\bar{\varsigma}$ denote a non-random 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 $\bar{\varsigma} < \infty$ (see e.g. Theorem 3.1.1 of Vershynin (2018)).

2.1 Gradient descent networks

The gradient descent network is a nonparametric model $\{g_W,\ W \in \mathcal{W}\}$ where the function g_W leverages the structure of the function g as given in (6). Given a symmetric positive definite matrix $V \in \mathbb{R}^{d_x \times d_x}$, for $a \in \mathbb{R}^{d_x}$, we set $\|a\|_V \stackrel{\text{def}}{=} \sqrt{a^\top V a}$. We define the proximal map of $\mathcal{R}$ in the V-metric as

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

When $V = \mathbf{I}_{d_x}$ we write $\text{Prox}_{\mathcal{R}}(\cdot)$ instead of $\text{Prox}_{\mathcal{R}}^{\mathbf{I}_{d_x}}(\cdot)$. It can be shown that if $\mathcal{R}$ is convex then $\text{Prox}_{\mathcal{R}}^{V}$ is uniquely defined and satisfies

$$\text{Prox}_{\mathcal{R}}^{V}(\mathbf{x}) = V^{-\frac{1}{2}} \circ \text{Prox}_{\mathcal{R}\circ V^{-\frac{1}{2}}}\left(V^{\frac{1}{2}}\mathbf{x}\right),\ \mathbf{x} \in \mathbb{R}^{d_x},$$

where $f \circ g$ is the composition of f with g. For a proof of this statement, see e.g. (Becker et al., 2019). For an in-depth introduction to proximal maps and proximal algorithms, we refer the reader to (Bauschke and Combettes, 2011; Parikh and Boyd, 2014). For $\mathbf{x} \in \mathbb{R}^{d_x}$, $\mathbf{y} \in \mathbb{R}^{d_y}$, step-size $\gamma > 0$, and $j \geq 1$, we set

$$F_{\mathbf{y}}(\mathbf{x}) \stackrel{\text{def}}{=} \text{Prox}_{\gamma\mathcal{R}}^{V}\left(\mathbf{x} - \gamma V^{-1}\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})\right), \quad \text{and} \quad F_{\mathbf{y}}^{j}(\mathbf{x}) \stackrel{\text{def}}{=} \underbrace{F_{\mathbf{y}} \circ \cdots \circ F_{\mathbf{y}}}_{j \text{ times}}(\mathbf{x}).$$

When $V = \mathbf{I}_{d_x}$, $F_{\mathbf{y}}$ corresponds to the iteration map of the proximal gradient descent algorithm (Beck and Teboulle (2010); Parikh and Boyd (2014)). Under appropriate convexity assumption and for a well-selected step size γ, the proximal gradient descent algorithm is known to converge to the minimizer of the function $\mathbf{x} \mapsto f(\mathbf{y}|\mathbf{x}) + \mathcal{R}(\mathbf{x})$. However as a first-order method the convergence of the proximal gradient descent is typically slow. The matrix V is introduced to mitigate this issue, and faster convergence of the sequence $F_{\mathbf{y}}^{j}(\mathbf{x}^{(0)})$ as $j \to \infty$ is achievable with an appropriate choice of V, where $\mathbf{x}^{(0)}$ is the initial value. For instance if f is convex, and V is taken as $\nabla^{(2)} f$ then $F_{\mathbf{y}}$ corresponds to the iteration map of the proximal Newton algorithm, which has better convergence properties (Lee et al., 2014; Becker et al., 2019).

When it converges the proximal gradient algorithm gives the representation

$$g(\mathbf{y}) = \lim_{j\to\infty} F_{\mathbf{y}}^{j}(\mathbf{x}^{(0)}).$$

The gradient descent network is a deep neural network function class that mimic this representation of g. To describe the model we first introduce a generic feed-forward deep neural network function $H_W : \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$ that serves as an approximation for the proximal map $\mathrm{Prox}^V_{\gamma\mathcal{R}}$. Let $D > 0$ be the depth of the feed-forward deep neural 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 $1 \le \ell \le D$, let $\mathsf{a}_\ell : \mathbb{R}^{p_\ell} \to \mathbb{R}^{p_\ell}$ be activation functions that we assume Lipschitz: for all $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{p_\ell}$,

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

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

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

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

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

Remark 1 *Feed-forward deep neural network models are usually written with additional bias terms (that is, by defining $\Psi^{(\ell)}_B(\mathbf{z})$ 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 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

$$F_{\mathbf{y},W}(\mathbf{x}) \stackrel{\text{def}}{=} H_W\left(\mathbf{x} - \gamma V^{-1} \nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})\right).$$

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

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

for some initial value $\mathbf{x}^{(0)} \in \mathbb{R}^{d_x}$. The function g_W defines a deep neural network function 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). We note that in addition to W, the function g_W depends also on the step-size γ, the symmetric positive definite matrix V, the depth D', and the initial value $\mathbf{x}^{(0)}$. We assume these additional parameters to be fixed. However in practice one can choose to include some of them in the training.

If for some W, $H_W \approx \mathrm{Prox}^V_{\gamma\mathcal{R}}$, then we can expect $g_W \approx g$ for D' sufficiently large, by standard convex optimization theory. As a result, the function class $\{g_W,\ W \in \mathcal{W}\}$ comes with good skills in approximating g. We impose next the necessary assumptions for the intuition above to hold.

H 2 *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, $g(\mathbf{y})$ is uniquely defined.*

2. *We choose the symmetric positive definite matrix V, and the step size γ such that $0 < \gamma \leq 2\lambda_{\min}(V)/M$, and there exist $R_0 < \infty$, $\varrho_n \in [0,1)$ such that for all $k \geq 1$,*

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

Remark 2 *In many inverse problems the forward map $f(\mathbf{y}|\mathbf{x})$ is well-understood, and assumption H2-(1) can be directly checked. For instance in the important special case of the Gaussian linear forward map as given in (2), H2-(1) holds with $M = \lambda_{\max}(A^\top A)/v^2$. When solving inverse problems, the uniqueness of the inversion $g(\mathbf{y})$ is desirable. And with a good regularization $\mathcal{R}$ to constrain the inversion process, it is reasonable to expect the resulting $g(\mathbf{y})$ to be uniquely defined. The convexity assumption on $\mathcal{R}$ is imposed mainly for mathematical convenience, in order to work more easily with the proximal operators.*

Assumption H2-(2) relates to the nature of the optimization problem (6). It is well-known that the proximal gradient algorithm can solve many composite optimization problems at the linear rate provided that the initialization $\mathbf{x}^{(0)}$ is chosen close enough to $g(\mathbf{y})$ (local linear convergence). The local linear convergence of iterative algorithms is generally expected albeit challenging to establish. For instance with the Gaussian linear forward map as given in (2), and taking $\mathcal{R}(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$ (the lasso penalty), Tao et al. (2016) Theorem 5.3 shows that for $\lambda > 0$ well-chosen, $g(\mathbf{y})$ is uniquely defined, and for all k

$$\|F^k_{\mathbf{y}}(\mathbf{x}^{(0)}) - g(\mathbf{y})\|_2 \leq R_0 \varrho^k,$$

for some $\varrho \in (0,1)$ provided $\mathbf{x}^{(0)}$ is chosen close enough to $g(\mathbf{y})$. For more general proximal gradient descent algorithms, their local linear convergence has been recently obtained by Liang et al. (2017) under some regularity conditions.

We also impose the following assumption that models the approximation of the proximal map $\mathrm{Prox}^V_{\gamma\mathcal{R}}$ by deep neural network functions.

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 $R < \infty$,*

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

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 3 *Neural networks are known to be universal approximators, and there has been a flurry of research activities in recent years to derive precise estimates on the approximation error for various architectures of neural networks, using various norms and under different smoothness assumptions on the function of interest (Yarotsky, 2017; Schmidt-Hieber, 2020;*

DeVore et al., 2021; Lu et al., 2021; Belomestny et al., 2023). Since $\mathbf{x} \mapsto \mathrm{Prox}^V_{\gamma\mathcal{R}}(\mathbf{x})$ *is a Lipschitz map, we can always invoke these results (e.g. Schmidt-Hieber (2020)) to conclude that Assumption 3 holds with* $\beta_1 = \beta_2 = d_x$*. However better approximation is achievable if* $\mathrm{Prox}^V_{\gamma\mathcal{R}}$ *is a simple map. The condition in (13) is a technical condition needed in the 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 $\mathcal{W} \stackrel{\text{def}}{=} \mathbb{R}^{p_D \times p_{D-1}} \times \cdots \times \mathbb{R}^{p_1 \times p_0}$. As indicated at the end of the introduction, at times we shall view $\mathcal{W}$ as the Euclidean space $\mathbb{R}^q$, where

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

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 and 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 and Bhattacharyya (2018)). To that end, we introduce a sparsity structure parameter $\Lambda = (\Lambda_D, \ldots, \Lambda_1) \in \mathcal{S} \stackrel{\text{def}}{=} \{0,1\}^{p_D \times p_{D-1}} \times \cdots \times \{0,1\}^{p_1 \times p_0}$ 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

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

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 large, of order n. We refer the reader to (Atchade and Bhattacharyya (2018)) for further discussion. Using this prior distribution and the regression model (7), we consider the posterior distribution

on $\Theta \stackrel{\text{def}}{=} \mathcal{S} \times \mathcal{W}$ 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

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

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

Theorem 4 *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 $n \geq \sigma^2 \log(q)$, we can construct a function class $\{H_W,\ W \in \mathcal{W}\}$, such that at unrolling depth D' that satisfies*

$$D' \gtrsim \frac{\log(n)}{-\log(\varrho_n)}, \tag{18}$$

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

$$\Pi\left(\|g_{\Lambda\odot W} - g\|_n > M\bar{\sigma}\frac{(D')^{1+\frac{\beta_2}{2}}}{n^{\frac{1}{2+\beta_2}}} \mid \mathcal{D}\right) \leq \frac{12}{q}, \tag{19}$$

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. ■

Remark 5 *We give some general intuition on the requirement (18) on the unrolling depth D'. The condition appears by requiring the statistical and the approximation errors of the model $\{g_W,\ W \in \mathcal{W}\}$ to match. For $\mathbf{y} \in \mathbb{R}^{d_y}$, $W \in \mathcal{W}$, and with g_W as defined in (11), we can write the model error $g_W(\mathbf{y}) - g(\mathbf{y})$ as*

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

Given $\epsilon > 0$, under H3 we can find $W \in \mathcal{W}$ as described in H3 such that (12) holds. We then deduce (see Lemma 16) that $\|F_{\mathbf{y},W}^{D'}(\mathbf{x}^{(0)}) - F_{\mathbf{y}}^{D'}(\mathbf{x}^{(0)})\|_2 \leq D'\kappa\epsilon$, where $\kappa = \lambda_{\max}(V)/\lambda_{\min}(V)$. Under H2, $\|g(\mathbf{y}) - F_{\mathbf{y}}^{D'}(\mathbf{x}^{(0)})\|_2 \leq R_0\varrho_n^{D'}$. Hence the approximate error of the function $\{g_W,\ W \in \mathcal{W}\}$ at depth D' is

$$\|g(\mathbf{y}) - g_W(\mathbf{y})\|_2 \leq R_0\varrho_n^{D'} + D'\kappa\epsilon.$$

Given a sample size n, the statistical error of the model is of the order $\sqrt{\mathsf{D}/n}$, where D is some measure of complexity of the model (e.g. number of non-zero coefficients, or VC-dimension). Here it is enough to use the lower bound $\sqrt{\mathsf{D}/n} \geq \sqrt{1/n}$. The model performs at its best when the approximation error matches the statistical error. For that we require taking D' such

$$R_0 \varrho_n^{D'} \lesssim \frac{1}{\sqrt{n}}, \quad \textit{and} \quad \epsilon \asymp \frac{1}{D' \kappa \sqrt{n}},$$

which leads to (18) and the derived contraction rate in (19). The argument is developed more carefully in the proof of Theorem 4.

2.3 Further discussion

We make several remarks here. (a) Theorem 4 suggests that the convergence rate of the model deteriorates as D' increases, with an optimal choice

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

that depends on the sample size. 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. Since ϱ_n is typically not known this conclusion does not lead to a practical tuning rule, but instead gives a general guideline that unlike classical DL models, unrolling models require careful, and sample size dependent tuning of D' for optimal performance. (b) 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 (See Proposition 15). 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 4 that GDN achieves the convergence rate

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

Hence, Theorem 4 implies that in inverse problems where d_y is larger than d_x (that it in settings where the forward map is hyghly informative), 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, we caution that the constants C_1, C_2 in the rates posted above depend 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).

(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 (19) 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

$$\sup_{W \in \widetilde{W}^{(j)}} \left| n^{-1} \sum_{i=1}^{n} (g_W(\mathbf{y}_i) - g(\mathbf{y}_i))^2 - \|g_W - g\|_2^2 \right|, \tag{20}$$

in Lemma D.5. Because the sup in (20) is taken over well behaved sets $\widetilde{W}^{(j)}$, this uniform deviation can be controlled using standard tools as in Wainwright (2019) Chapter 14.

2.3.1 Application to sparse marginal distributions

It is known that deep learning models can sometimes 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 achieve faster rate in the estimation of compositional functions. We give a similar example below. We give another application of Theorem 4 where the posterior predictive function obtained from the GDN achieves the parametric rate. When dealing with images, several authors such as Beck and Teboulle (2010); Dong et al. (2011) have argued for the validity of the manifold hypothesis, where natural image data exhibit sparsity after linear transformation (such as difference operators, or wavelet transforms), and suggested modeling the marginal distribution μ as

$$\mu(\mathrm{d}\mathbf{x}) = \frac{1}{c_\mu} e^{-\mathcal{R}_0(B\mathbf{x})} \mathrm{d}\mathbf{x}, \tag{21}$$

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 and Wajs (2005)), we have

$$\mathrm{Prox}_{\gamma\mathcal{R}}(\mathbf{x}) = B^{-1} \mathrm{Prox}_{\gamma\mathcal{R}_0}(B\mathbf{x}). \tag{22}$$

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

$$\mathcal{R}_0(\mathbf{x}) = \lambda_1\|\mathbf{x}\|_1 + \frac{\lambda_2}{2}\|\mathbf{x}\|_2^2. \tag{23}$$

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 $\mathsf{ReLu}(t) \stackrel{\text{def}}{=} \max(t, 0)$. 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 4 yields the following.

Corollary 6 *Suppose that H1 holds with f as in (2), and suppose that μ is as in (21) with some orthogonal matrix B, and $\mathcal{R}_0$ as in (23). 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 $D' \gtrsim -\log(n)/\log(\varrho)$ 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 through both a simulation and a real-data deblurring task, and we we compare the performance of GDN with the Neumann networks (NMN) of Gilton et al. (2020). For all examples we approximately sample from the posterior distribution (17) using the Sparse Asynchronous Stochastic Gradient Langevin Dynamics (SA-SGLD) sampler introduced by Atchade and Wang (2023). The algorithm is a type of Gibbs sampler that alternates between an update of W given Λ using the SGLD of Welling and Teh (2011), and an update of Λ given W, using a variation of the asynchronous Gibbs sampler of De Sa et al. (2016). A github implementation is available at `https://github.com/xliu-522/inverse-problem`.

3.1 Illustration with a simulated data deblurring problem

Image deblurring is a classical inverse problem in computational imaging. To illustrate our approach, we begin with a simulated dataset $\mathcal{D}$ for experimental evaluation.

Data generation. We simulate a dataset of structured synthetic matrices $\mathbf{x} \in \mathbb{R}^{16\times 16}$, each constructed by combining structured signal components with noise[2]. Each matrix $\mathbf{x}$ is then blurred to produce the observation $\mathbf{y}$ using a Gaussian convolution kernel with variance 3 without restriction on the kernel range. We thus generate a dataset $\mathcal{D} = (\mathbf{x}_i, \mathbf{y}_i),\ 1 \leq i \leq n\}$ with $n \in \{5 \times 10^2, 5 \times 10^4\}$.

Model architecture. We construct H_W in (10) as a 3-layer ReLu-activated convolutional neural network[3]. The total number of parameters is $q = 18,881$. We explore both the GDN and NMN models at unrolling depth $D' = 4, 8, 12$, and 18, and compare their performance against a larger feedforward convolutional neural network (FNN) that does not incorporate the forward model. The FNN architecture consists of 6 layers with total number of parameters $q = 136,641$ [4]. All layers are padded to preserve the spatial dimensions

2. Each matrix is partitioned into 4 blocks: the upper left is a sparse diagonal matrix with non-zero entries sampled from $\mathrm{N}(20, 2)$ where $i - j = 4$; the lower-right block is a periodic diagonal matrix with non-zero entries sampled from $\mathrm{N}(-10, 0.5)$ where $i = j \bmod 6$; the upper-right block is a 10×10 matrix with *i.i.d.* entries sampled from $\mathrm{N}(10, 1)$; and the lower-left block is an 8×8 matrix with entries sampled from $\mathrm{N}(-10, 5)$.
3. The network consists of 3 convolutional layers with kernel sizes $3, 3, 1$ and number of filters $32, 64, 1$, respectively. Each layer, except the final one, is followed by a LayerNorm layer and a ReLu activation
4. The FNN includes 2 convolutional layers, 1 channel-wise fully connected layer, and 3 deconvolutional layers, with respective sizes $5, 3, 2, 4, 5, 3$ and filter numbers $32, 64, 64, 64, 32, 1$. Each layer except the final one is followed by a LayerNorm layer and a ReLu activation.

of the input. In the GDN network, we take V as a rank-k approximation of the Hessian matrix $A^\top A$, with $k = 200$. We set $\gamma = 0.01$ for the GDN model, and $\gamma = 0.1$ for NMN. We set $\mathbf{x}^{(0)} = \mathbf{0} \in \mathbb{R}^{16\times 16}$ for all models.

Training details: For the Bayesian prior, we set $\rho_0 = n, \rho_1 = 1$, and $\mathsf{u} = 10$. We choose $\sigma^2 = 0.01$ in (17), and run the SA-SAGLD with a constant step size of 5×10^{-5} for the FNN model and 5×10^{-7} for GDN and NMN models. The mini-batch size is fixed at 50 across all experiments. All MCMC samplers are implemented in PyTorch and executed on an NVIDIA Tesla V100 GPU system with 384 GB of GPU memory. Each sampler is run for 10^5 iterations.

Evaluation procedure and results. We compare both the GDN and the NMN at varying unrolling depth $D' \in \{4, 8, 12, 18, 22\}$. We compare the models using the mean square error

$$\mathrm{MSE}(\Lambda, W) = \frac{1}{500}\sum_{i=1}^{500} \|g_{\Lambda\odot W}(\mathbf{y}_i) - g(\mathbf{y}_i)\|_2^2,$$

computed over a test dataset $\{(\mathbf{x}_i, \mathbf{y}_i),\ 1 \le i \le 500\}$. For each posterior chain, we evaluate $\mathrm{MSE}(\Lambda, W)$ over the last 1000 MCMC samples (Λ, W). The boxplots in Figure 1 display the distribution of the mean square errors for each model. Figure 2 presents a representative reconstruction example from each model. The results are consistent with Theorem 4. Indeed we observe a U-shape dependence of the MSE as function of D' as predicted by our theorem. Furthermore, increasing the sample size n increases the value of D' at which performance is best, as suggested by our theory. We note a similar behavior for NMN, suggesting that our theory may hold more broadly for unrolling models beyond GND. We observe that both the GDN and NMN outperform the FNN, despite the FNN's more complex architecture. With a small sample size ($n = 500$) NMN consistently outperforms the GDN, which is consistent with the finding of Gilton et al. (2020). However the difference in performance between the two models vanishes at $n = 50k$.

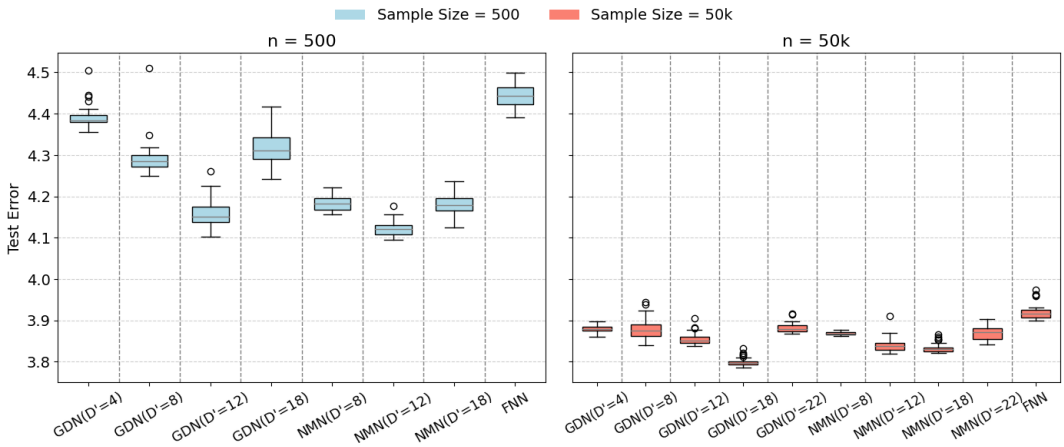

Figure 1: Test loss comparison between FNN, GDN ($D' = 4$), GDN ($D' = 8$), GDN ($D' = 12$), GDN ($D' = 18$), GDN ($D' = 22$), NMN ($D' = 8$), NMN ($D' = 12$), NMN ($D' = 18$), NMN ($D' = 22$) under different training sample size.

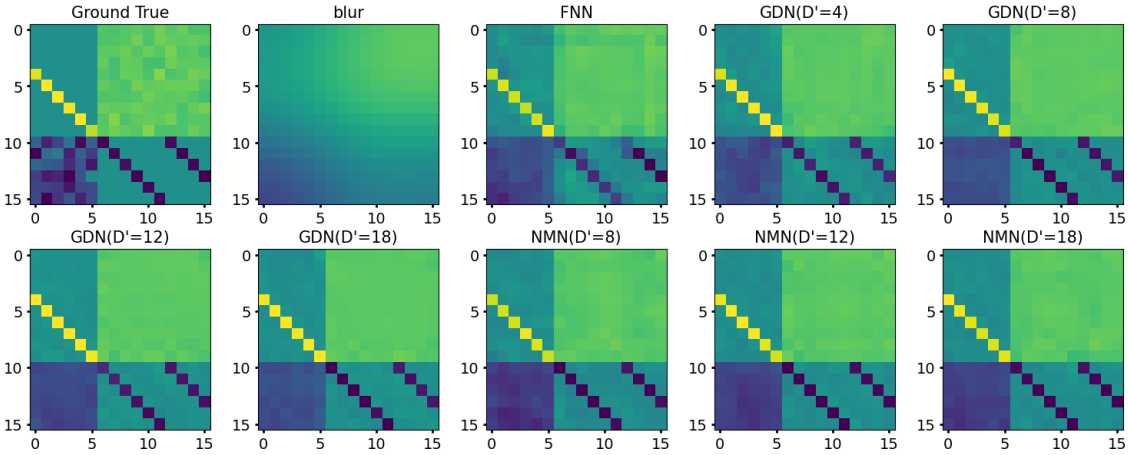

Figure 2: With sample size $n = 500$, a reconstruction example from FNN, GDN ($D' = 4$), GDN ($D' = 8$), GDN ($D' = 12$), GDN ($D' = 18$), NMN ($D' = 8$), NMN ($D' = 12$), NMN ($D' = 18$).

3.2 Illustration with CelebA dataset

We extend the last example to a real dataset, using 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 blurred images $\mathbf{y}_i$ through the linear forward model (1), where A is a Gaussian convolution matrix with variance 6.25, and where $v^2 = 0.01$ leading to a highly ill-conditioned inverse problem.

Model architecture. We construct the function H_W in (10) as a 3-layer ReLu-activated convolutional neural network[5]. We explore both the GDN and NMN models at unrolling depth $D' = 4, 6, 8, 12$. The total number of parameters is the same across all three cases and equal to $q = 592,897$. For comparison, we also evaluate a feedforward model with the same architecture as H_W. In the GDN network, we take V as a rank-k approximation of the Hessian matrix $A^\top A$, with $k = 10$. We set $\gamma = 0.001$ for the GDN model, and $\gamma = 0.01$ for NMN. We set $\mathbf{x}^{(0)} = \mathbf{0} \in \mathbb{R}^{16\times 16}$ across all models.

Training details. For the Bayesian prior, we set $\rho_0 = 10,000, \rho_1 = 1$, and $\mathsf{u} = 10$. The SA-SGLD sampler is run with a constant step size of 5×10^{-8} across all models using a mini-batch size of $B = 164$. All MCMC samplers are implemented in PyTorch and executed on an NVIDIA Tesla V100 GPU system with 384 GB of GPU memory. Each algorithm is run for $500,000$ iterations.

Evaluation procedure and results. Figure 3 shows the distributions of the test MSE computed from 500 posterior samples drawn after burn-in. We observe very similar behaviors as in the simulated data setting. The NMN models consistently outperform the GDN models at similar depth. However for both models, performance degrades at a deeper

5. The network uses convolutional layers with kernel sizes $3, 3, 1$, and filter sizes $256, 256, 1$, respectively. All layers are padded to preserve the input image dimensions.

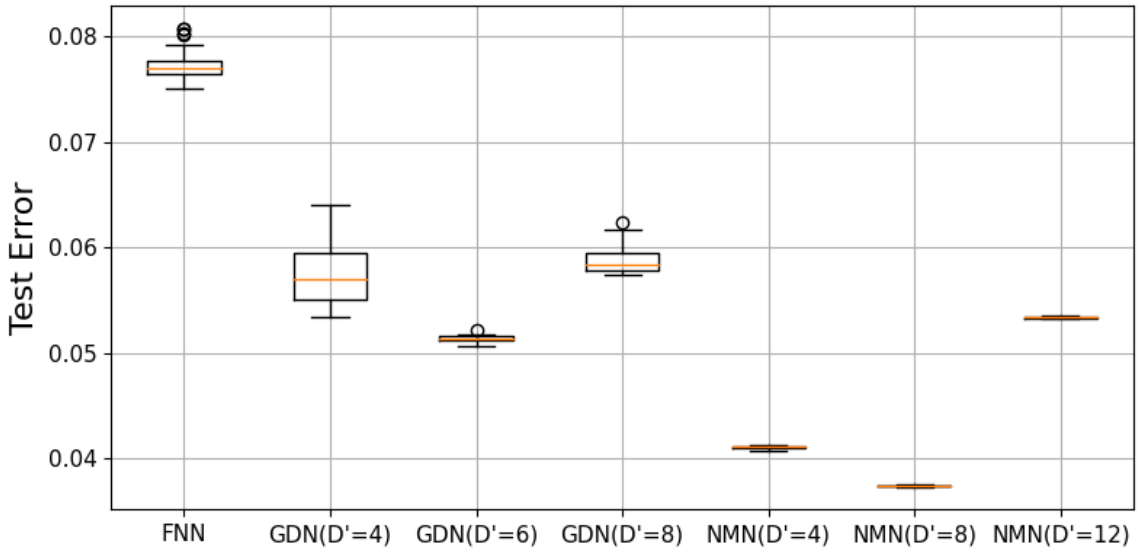

Figure 3: Test loss comparison between FNN, GDN (D'=4), GDN (D'=6), GDN (D'=8), NMN (D'=4), NMN(D'=8), NMN(D'=12)

unrolling depth ($D' = 8$ for GDN, and $D' = 12$ for NMN), again indicating overfitting. Figure 4 displays five examples of reconstructed images.

4. A general Bayesian posterior contraction result

Theorem 4 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 $\mathcal{W} \stackrel{\text{def}}{=} \mathbb{R}^{p_D \times p_{D-1}} \times \cdots \times \mathbb{R}^{p_1 \times p_0}$, 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 times we shall view $\mathcal{W}$ as the Euclidean space $\mathbb{R}^q$, with Euclidean norm denoted $\|\cdot\|_2$, where

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

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*

$$\|g_W(\mathbf{y}) - g_{W'}(\mathbf{y})\|_2 \leq L(\eta)\|W - W'\|_2. \tag{24}$$

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.

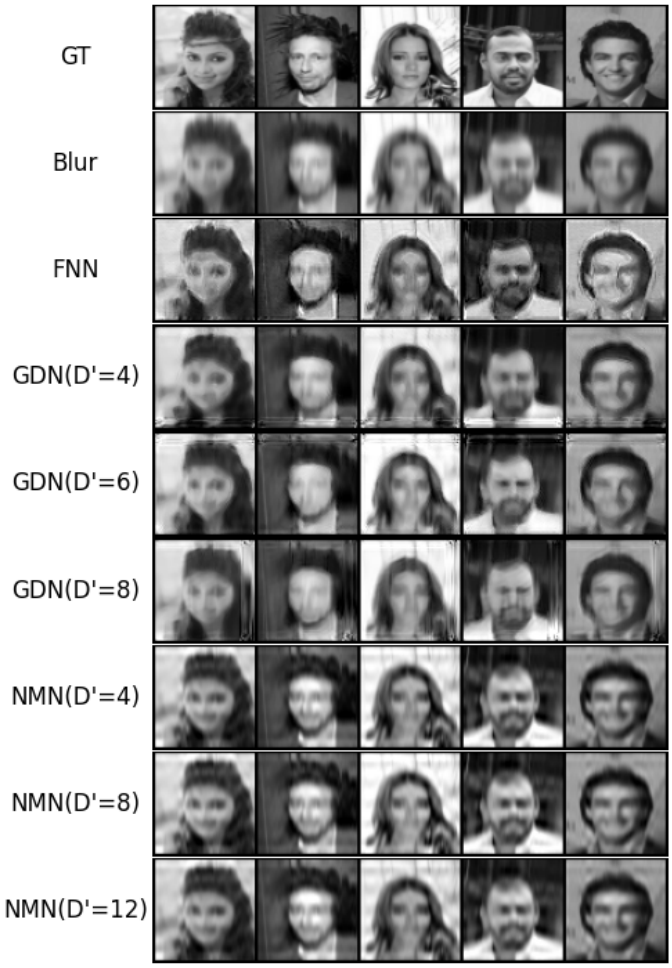

Figure 4: From top to bottom: Ground truth image, blurred image, FNN, GDN (D'=4), GDN (D'=6), GDN (D'=8), NMN (D'=4), NMN(D'=8), NMN(D'=12)

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*

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

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

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

where

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

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

$$\Pi\left(\|g_{\Lambda\odot W} - g\|_n > M\mathsf{r}\,|\mathcal{D}\right) \leq \frac{12}{q}, \tag{25}$$

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 and 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 and 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*

$$\mathcal{E} \subseteq \left\{ z \in \mathsf{Z} : \int_\Theta \frac{f_\theta(z)}{f(z)}\pi(\mathrm{d}\theta) \geq e^{-\beta},\ \int_{S^c} \frac{f_\theta(z)}{f(z)}\pi(\mathrm{d}\theta) \leq \rho \right. \\ \left. \text{and} \quad \sup_{\theta\in\Xi_j} [\log f_\theta(z) - \log f(z)] \leq -r_j \quad \text{for all} \quad j \geq 1 \right\}. \quad (26)$$

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

$$\Pi(B^c|z) \leq e^\beta \left(\rho + \sum_{j\geq 1} e^{-r_j} \right). \quad (27)$$

Proof Using the lower bound on the normalizing constant provided by the event (26), 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 (27) 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 (27) 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 one of the simplest 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. Our results also show 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 H1. We contend that the reduction of the inverse problem

to a regression problem requires the BvM phenomenon. Solving the inverse problem without H1 boils down to estimating the entire conditional distribution, not just its mode. Several recent works have proposed extensions of algorithm unrolling architectures for conditional density estimation (Ardizzone et al. (2019)). Extending our analysis to these conditional density models is a possible direction for future research.

An outstanding challenge of unrolling models is their computational and memory cost. 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 and Welling (2019)). Mitigating this memory cost and easing the implementation of algorithm unrolling architectures is another important methodological problem for future research.

In recent years the use of pre-trained generative models (particularly diffusion models) for solving inverse problems has attracted a copious literature (Daras et al., 2024; Wu et al., 2024; Uehara et al., 2025). Unrolling models may offer different approaches for leveraging pre-trained models. For instance, in this work we have assumed that the unrolling model is trained starting from a fixed initial solution $\mathbf{x}^{(0)}$. However if a generative model for $\mathbf{x}$ is available one could imagine a more realistic model that draws $\mathbf{x}^{(0)}$ from the generative model. Furthermore, it may be possible to leverage pre-trained score functions in the unrolling process.

References

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

Lynton Ardizzone, Jakob Kruse, Carsten Rother, and Ullrich Kothe. Analyzing inverse problems with invertible neural networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=rJed6j0cKX`.

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

Yves Atchade and Liwei Wang. A fast asynchronous Markov chain Monte Carlo sampler for sparse Bayesian inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1492–1516, 2023. ISSN 1369-7412. doi: 10.1093/jrsssb/qkad078.

Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *ArXiv*, abs/1607.06450, 2016.

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

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer Publishing Company, Incorporated, 1st edition, 2011. ISBN 1441994661.

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

Stephen Becker, Jalal Fadili, and Peter Ochs. On quasi-newton forward-backward splitting: Proximal calculus and convergence. *SIAM Journal on Optimization*, 29(4):2445–2481, 2019. doi: 10.1137/18M1167152.

Denis Belomestny, Alexey Naumov, Nikita Puchkin, and Sergey Samsonov. Simultaneous approximation of a smooth function and its derivatives by deep neural networks with piecewise-polynomial activations. *Neural Networks*, 161:242–253, 2023. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2023.01.035.

Clément Berenfeld, Paul Rosa, and Judith Rousseau. Estimating a density near an unknown manifold: A Bayesian nonparametric approach. *The Annals of Statistics*, 52(5):2081 – 2111, 2024. doi: 10.1214/24-AOS2423. URL `https://doi.org/10.1214/24-AOS2423`.

Abhishek Bhattacharya and David B. Dunson. Nonparametric bayesian density estimation on manifolds with applications to planar shapes. *Biometrika*, 97(4):851–865, 2010. ISSN 00063444, 14643510.

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

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

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

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

Ismaël Castillo, Johannes Schmidt-Hieber, and Aad van der Vaart. Bayesian linear regression with sparse priors. *Ann. Statist.*, 43(5):1986–2018, 10 2015. doi: 10.1214/15-AOS1334.

Lawrence Cayton et al. *Algorithms for manifold learning.* eScholarship, University of California, 2008.

J.H. Rick Chang, Chun-Liang Li, Barnabás Poczos, B.V.K. Vijaya Kumar, and Aswin C. Sankaranarayanan. One network to solve them all – solving linear inverse problems using deep projection models. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5889–5898, 2017.

Xiaohan Chen, Jialin Liu, Zhangyang Wang, and Wotao Yin. Theoretical linear convergence of unfolded ista and its practical weights and thresholds. In S. Bengio, H. Wallach,

H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yong Chun and Jeffrey A. Fessler. 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)*, pages 1–5, 2018. doi: 10.1109/IVMSPW.2018.8448694.

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

Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G. Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems, 2024. URL `https://arxiv.org/abs/2410.00083`.

Christopher De Sa, Kunle Olukotun, and Christopher Ré. Ensuring rapid mixing and low bias for asynchronous gibbs sampling. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1567–1576, 2016.

Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, 2021.

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

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

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

S.A. Geer, S. van de Geer, R. Gill, B.D. Ripley, S. Ross, B. Silverman, D. Williams, and M. Stein. *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`.

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

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Iain M Johnstone. High dimensional bernstein-von mises: simple examples. *Institute of Mathematical Statistics Collections*, 6:87, 2010.

Ulugbek S. Kamilov and Hassan Mansour. Learning optimal nonlinearities for iterative thresholding algorithms. *IEEE Signal Processing Letters*, 23(5):747–751, 2016. doi: 10.1109/LSP.2016.2548245.

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

Jason D. Lee, Yuekai Sun, and Michael A. Saunders. Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014. doi: 10.1137/130921428.

Yuelong Li, Mohammad Tofighi, Junyi Geng, Vishal Monga, and Yonina C. Eldar. 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.

Jingwei Liang, Mohamed-Jalal Fadili, and Gabriel Peyré. Activity identification and local linear convergence of forward-backward-type methods. *SIAM J. Optim.*, 27:408–437, 2017.

Risheng Liu, Shichao Cheng, Long Ma, Xin Fan, and Zhongxuan Luo. 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.

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

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021. doi: 10.1137/20M134695X.

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

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

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014. ISBN 1461346916.

Richard Nickl. 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.

Gregory Ongie, Ajil Jalal, Christopher Baraniuk, Alexandros Dimakis, and Rebecca Willett. 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.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014. ISSN 2167-3888. doi: 10.1561/2400000003. URL `http://dx.doi.org/10.1561/2400000003`.

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

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

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

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

M.J. Schervish. *Theory of Statistics*. Springer Series in Statistics. Springer New York, 1996. ISBN 9780387945460. URL `https://books.google.com/books?id=F9A9af4It10C`.

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

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

Nir Shlezinger, Jay Whang, Yonina C. Eldar, and Alexandros G. Dimakis. Model-based deep learning: Key approaches and design guidelines. In *2021 IEEE Data Science and Learning Workshop (DSLW)*, pages 1–6, 2021. doi: 10.1109/DSLW51110.2021.9523403.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

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

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

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

Mahsa Taheri, Fang Xie, and Johannes Lederer. Statistical guarantees for regularized neural networks. *Neural Networks*, 142:148–161, 2021.

Rong Tang and Yun Yang. Minimax rate of distribution estimation on unknown submanifolds under adversarial losses. *The Annals of Statistics*, 51(3):1282 – 1308, 2023. doi: 10.1214/23-AOS2291. URL https://doi.org/10.1214/23-AOS2291.

Shaozhe Tao, Daniel Boley, and Shuzhong Zhang. Local linear convergence of ista and fista on the lasso problem. *SIAM Journal on Optimization*, 26(1):313–336, 2016. doi: 10.1137/151004549.

Bahareh Tolooshams, Andrew Song, Simona Temereanca, and Demba Ba. Convolutional dictionary learning based auto-encoders for natural exponential-family distributions. In *International Conference on Machine Learning*, pages 9493–9503. PMLR, 2020.

Masatoshi Uehara, Yulai Zhao, Chenyu Wang, Xiner Li, Aviv Regev, Sergey Levine, and Tommaso Biancalani. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review, 2025. URL https://arxiv.org/abs/2501.09685.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, New York, 1998.

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

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

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 681–688, USA, 2011. Omnipress. ISBN 978-1-4503-0619-5.

Nick Whiteley, Annie Gray, and Patrick Rubin-Delanchy. Statistical exploration of the manifold hypothesis, 2025. URL https://arxiv.org/abs/2208.11665.

Zihui Wu, Yu Sun, Yifan Chen, Bingliang Zhang, Yisong Yue, and Katherine Bouman. Principled probabilistic imaging using diffusion models as plug-and-play priors. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Xq9HQf7VNV.

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

yan Yang, Jian Sun, Huibin Li, and Zongben Xu. Deep admm-net for compressive sensing mri. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

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

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3929–3938, 2017.

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

Acknowledgments

We are grateful to Demba Ba and Alexander Lin for insightful discussions on unrolling models. We acknowledge the support of the NSF grants DMS-2015485, DMS-2210664 and DMS-2515787, and the Faragher Fellowship from the University of Minnesota. The authors have no conflict of interest to declare.

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 $\mathcal{D} \stackrel{\text{def}}{=} (\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$. For $W \in \mathcal{W}$, we define

$$f_W(\mathcal{D}) \stackrel{\text{def}}{=} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n \|\mathbf{x}_i - g_W(\mathbf{y}_i)\|_2^2\right),$$
$$\text{and} \quad f_\star(\mathcal{D}) \stackrel{\text{def}}{=} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n \|\mathbf{x}_i - g(\mathbf{y}_i)\|_2^2\right). \quad (28)$$

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

$$\Pi(A|\mathcal{D}) = \frac{\int_A \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})}\Pi_0(\mathrm{d}\Lambda, \mathrm{d}W)}{\int_\Theta \frac{f_{\Lambda\odot W}(\mathcal{D})}{f_\star(\mathcal{D})}\Pi_0(\mathrm{d}\Lambda, \mathrm{d}W)}. \quad (29)$$

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

$$\log\left(\frac{f_W(\mathcal{D})}{f_\star(\mathcal{D})}\right) = \frac{1}{2\sigma^2}\sum_{i=1}^n \left(\|\mathbf{x}_i - g(\mathbf{y}_i)\|_2^2 - \|\mathbf{x}_i - g_W(\mathbf{y}_i)\|_2^2\right)$$
$$= -\frac{n}{2\sigma^2}\|g_W - g\|_n^2 - \frac{1}{\sigma^2}\sum_{i=1}^n \langle \mathbf{x}_i - g(\mathbf{y}_i), g(\mathbf{y}_i) - g_W(\mathbf{y}_i)\rangle. \quad (30)$$

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

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

and

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

We set

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

and

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

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

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

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

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

and consider the $\mathcal{E}$

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

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 $\Lambda \odot W \in \cup_{j \ge 1} \mathcal{W}_j(s, b_s)$. Let $\check\Pi_0$ be the distribution of $\Lambda \odot W$, when $(\Lambda, W) \sim \Pi_0$. Starting from (29), and following the same argument leading to (27), for $\mathcal{D} \in \mathcal{E}$, we have

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

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

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

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

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

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

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

We show in Lemma 12 below that

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

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

$$\frac{288}{\sqrt{n}}\int_{\frac{x^2}{32\bar{\zeta}}}^{x}\sqrt{\log\mathcal{N}\left(\epsilon,\mathcal{W}^{(x)}(s,b_s),\|\cdot\|_n\right)}\mathrm{d}\epsilon \leq \frac{x^2}{\bar{\sigma}}, \quad \text{for all} \quad x\geq\mathsf{r}, \tag{31}$$

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

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

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 $\|W-W'\|_n \stackrel{\text{def}}{=} \|g_W - g_{W'}\|_n$. 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 (31). First we use the majoration

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

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 $\sigma^2 \geq \bar{\sigma}^2$. For all integers $s \geq 1$, with $b_s \stackrel{\text{def}}{=} \sqrt{2(1+\mathsf{u})(1+s)\log(q)/\rho_1}$, 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

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

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

$$\Pi_0(\mathcal{A}(s)) \leq \frac{4}{q^{\mathsf{u}(1+s)}}. \tag{32}$$

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 (30) 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 $\|\boldsymbol{\xi}_i|\mathbf{y}_i\|_{\psi_2}\leq\bar{\sigma}$. Therefore,

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

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

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

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

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

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

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

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

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

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

$$\max_{1\le i\le n} \sup_{W\in\mathcal{N}(\eta)} \|g_{\Lambda_\star \odot W}(\mathbf{y}_i) - g_{W_\star}(\mathbf{y}_i)\|_2 \le \sigma\sqrt{\frac{s_\star \log(q)}{n}}. \tag{33}$$

Switching the sign and taking the conditional expectation in (30) 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 (33) 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 (30), we have

$$\log\left(\frac{f_\star(\mathcal{D})}{f_{\Lambda_\star\odot W}(\mathcal{D})}\right) - \mathbb{E}\left[\log\left(\frac{f_\star(\mathcal{D})}{f_{\Lambda_\star\odot W}(\mathcal{D})}\right) \mid \mathbf{y}_{1:n}\right] = \frac{1}{\sigma^2}\sum_{i=1}^n \langle \boldsymbol{\xi}_i, g(\mathbf{y}_i) - g_{\Lambda_\star\odot W}(\mathbf{y}_i)\rangle. \tag{34}$$

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

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

In the sequel, we set

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

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

$$\sup_{W \in \mathcal{N}(\eta)} \mathbb{P}\left(\mathcal{D} \notin \mathcal{E}_{\Lambda_\star \odot W} \mid \mathbf{y}_{1:n}\right) \leq \frac{2}{q^{s_\star}}. \tag{35}$$

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,

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

using (35). 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

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

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*

$$\frac{288}{\sqrt{n}}\int_{\frac{x^2}{16\bar{\sigma}}}^{x}\sqrt{\log\mathcal{N}(\epsilon,\mathcal{W}^{(x)},\|\cdot\|_n)}\mathrm{d}\epsilon \leq \frac{x^2}{\bar{\sigma}}, \tag{36}$$

where $\mathcal{W}^{(x)} \stackrel{\text{def}}{=} \{W \in \mathcal{W}_0,\ \|g_W - g\|_n \leq x\}$. Let f_W and $f_\star$ be as defined in (28). 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 $\widetilde{\mathcal{W}}^{(j)} \stackrel{\text{def}}{=} \{W \in \mathcal{W}_0 :\ jM\mathsf{r} < \|g_W - g\|_n \leq (j+1)M\mathsf{r}\}$.

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 $\boldsymbol{\xi}_i \stackrel{\text{def}}{=} \mathbf{x}_i - g(\mathbf{y}_i)$, and we define

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

Using (30) we can re-express the log-likelihood ratio as

$$\log\left(\frac{f_W(\mathcal{D})}{f_\star(\mathcal{D})}\right) = -\frac{n}{2\sigma^2}\|g_W - g\|_n^2 - nZ_n(g_W). \tag{37}$$

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

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

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

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

Therefore,

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

where

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

For each $j \ge 1$, we set

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

and each $\iota = 1, \ldots$, let $\mathcal{C}_j^{(\iota)} \stackrel{\text{def}}{=} \{g_{j,1}^{(\iota)}, \ldots, g_{j,N_{j,\iota}}^{(\iota)}\}$ 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 $\|g_W - g_{j,W}^{(\iota)}\|_n \le \mathsf{r}_{j+1}2^{-\iota}$. Let $\ell_j \ge 0$, be the smallest integer such that

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

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 (37) 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

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

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

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

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

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

$$\sum_{\iota=1}^{\ell_j}\eta_{j,\iota}\le 1. \tag{38}$$

Due to (38), we can use the sequence $\{\eta_{j,\iota},\ \iota=1,\ldots,\ell_j\}$ to say that

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

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

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

Therefore by Hoefdding's inequality,

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

By construction,

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

which gives

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

using (36). Therefore

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

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 (38), 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,

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

■

A.2 Proof of Theorem 4

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

$$f(\mathbf{x}) - g(\mathbf{x}) = \sum_{j=1}^{K} f_K \circ \cdots \circ f_j \left(g_{j-1} \circ \cdots \circ g_1(\mathbf{x})\right) - f_K \circ \cdots \circ f_{j+1} \circ g_j \left(g_{j-1} \circ \cdots \circ g_1(\mathbf{x})\right), \quad (39)$$

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 $H_W^{(0)}(\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{x}$, and for $1 \le \ell \le D$, define $H_W^{(\ell)}(\mathbf{x}) \stackrel{\text{def}}{=} \Psi_{W_\ell}^{(\ell)}(H_W^{(\ell-1)}(\mathbf{x}))$, 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,

$$\|H_W^{(\ell)}(\mathbf{x}_1) - H_W^{(\ell)}(\mathbf{x}_2)\|_2 \le \prod_{j=1}^{\ell} \|W_j\|_{\mathsf{op}} \|\mathbf{x}_1 - \mathbf{x}_2\|_2. \quad (40)$$

Similarly, for any $1 \le \ell \le D$, (39) 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,

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

Since the activation functions satisfy $\mathsf{a}_\ell(0) = 0$, we have the bound

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

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

$$\|H^{(\ell)}_W(\mathbf{x}) - H^{(\ell)}_{W'}(\mathbf{x})\|_2 \leq \|\mathbf{x}\|_2 \sum_{j=1}^{\ell} \|W_j - W'_j\|_{\mathsf{op}} \prod_{k=1}^{j-1} \|W'_k\|_{\mathsf{op}} \prod_{k=j+1}^{\ell} \|W_k\|_{\mathsf{op}}. \tag{41}$$

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

$$F_{\mathbf{y}}(\mathbf{x}) = \mathrm{Prox}^V_{\gamma\mathcal{R}}\left(\mathbf{x} - \gamma V^{-1}\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})\right), \quad \text{and} \quad F_{\mathbf{y},W}(\mathbf{x}) = H_W\left(\mathbf{x} - \gamma V^{-1}\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})\right).$$

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 $g_W(\mathbf{y}) = F^{D'}_{\mathbf{y},W}(\mathbf{x}^{(0)})$.

Lemma 14 *Under H2, and for $\gamma > 0$ taken such that $\gamma \leq 2\lambda_{\min}(V)/M$, the function $F_{\mathbf{y}}$ is non-expansive in the V-norm: for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{d_x}$,*

$$\|F_{\mathbf{y}}(\mathbf{x}) - F_{\mathbf{y}}(\mathbf{x}')\|_V \leq \|\mathbf{x} - \mathbf{x}'\|_V.$$

Proof Given $\mathbf{x}$, we set $\widehat{\mathbf{x}} \stackrel{\text{def}}{=} \mathbf{x} - \gamma V^{-1}\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})$, so that $F_{\mathbf{y}}(\mathbf{x}) = \mathrm{Prox}^V_{\gamma\mathcal{R}}(\widehat{\mathbf{x}})$. Lemma 3.2 of Becker et al. (2019)) gives the representation

$$\mathrm{Prox}^V_{\mathcal{R}}(\mathbf{x}) = V^{-\frac{1}{2}} \circ \mathrm{Prox}_{\mathcal{R}\circ V^{-\frac{1}{2}}}\left(V^{\frac{1}{2}}\mathbf{x}\right), \ \mathbf{x} \in \mathbb{R}^{d_x}.$$

Therefore,

$$\|F_{\mathbf{y}}(\mathbf{x}) - F_{\mathbf{y}}(\mathbf{x}')\|_V = \left\|\mathrm{Prox}^V_{\mathcal{R}}(\widehat{\mathbf{x}}) - \mathrm{Prox}^V_{\mathcal{R}}(\widehat{\mathbf{x}'})\right\|_V = \left\|\mathrm{Prox}_{\mathcal{R}\circ V^{-\frac{1}{2}}}(V^{\frac{1}{2}}\widehat{\mathbf{x}}) - \mathrm{Prox}_{\mathcal{R}\circ V^{-\frac{1}{2}}}(V^{\frac{1}{2}}\widehat{\mathbf{x}'})\right\|_2.$$

We assume in H2 that $\mathcal{R}$ is convex. Therefore $\mathcal{R} \circ V^{1/2}$ is convex, which implies that its proximal operator is non-expansive in the Euclidean norm (see e.g. Bauschke and Combettes (2011) Proposition 12.27). Hence

$$\|F_{\mathbf{y}}(\mathbf{x}) - F_{\mathbf{y}}(\mathbf{x}')\|_V \leq \|V^{\frac{1}{2}}\widehat{\mathbf{x}'} - V^{\frac{1}{2}}\widehat{\mathbf{x}}\|_2.$$

Now, since $\widehat{\mathbf{x}} \stackrel{\text{def}}{=} \mathbf{x} - \gamma V^{-1}\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})$, we have

$$\|V^{\frac{1}{2}}\widehat{\mathbf{x}'} - V^{\frac{1}{2}}\widehat{\mathbf{x}}\|_2^2 = \|\mathbf{x} - \mathbf{x}'\|_V^2 + \gamma^2\|\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x}) - \nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x}')\|_{V^{-1}}^2 - 2\gamma\left\langle \mathbf{x} - \mathbf{x}', \nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x}) - \nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x}')\right\rangle.$$

If $\mathbf{x} \mapsto \nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})$ is convex, M-smooth and μ-strongly convex, Theorem 2.1.12 of Nesterov (2014) implies that

$$\left\langle \mathbf{x} - \mathbf{x}', \nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x}) - \nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x}')\right\rangle \geq \frac{\mu M}{\mu + M}\|\mathbf{x} - \mathbf{x}'\|_2^2 + \frac{1}{\mu + M}\|\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x}) - \nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x}')\|_2^2.$$

Here we do not assume strong convexity, so $\mu = 0$. We can then conclude that

$$\|V^{\frac{1}{2}}\widehat{\mathbf{x}}' - V^{\frac{1}{2}}\widehat{\mathbf{x}}\|_2^2 \leq \|\mathbf{x} - \mathbf{x}'\|_V^2 + \left(\frac{\gamma^2}{\lambda_{\min}(V)} - \frac{2\gamma}{M}\right)\|\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x}) - \nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x}')\|_2^2 \leq \|\mathbf{x} - \mathbf{x}'\|_V^2,$$

provided that $0 < \gamma \leq 2\lambda_{\min}(V)/M$. Note that in the strongly convex case with $\mu > 0$, we obtain

$$\|V^{\frac{1}{2}}\widehat{\mathbf{x}}' - V^{\frac{1}{2}}\widehat{\mathbf{x}}\|_2^2 \leq \|\mathbf{x} - \mathbf{x}'\|_V^2 \left(1 - \frac{2\gamma\mu M}{\lambda_{\min}(V)(\mu + M)}\right),$$

which is a contraction for $0 < 2\gamma \leq \lambda_{\min}(V)/M$. ■

Proposition 15 *Assume H2-(1). Suppose also that for all* $\mathbf{x}$*, the function* $\mathbf{y} \mapsto \nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})$ *is* M*-Lipschitz, and the function* $\mathbf{x} \mapsto \nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})$ *is* μ*-strongly convex for all* $\mathbf{y}$*. Then for all* $\gamma > 0$ *such that* $2\gamma M/\lambda_{\min}(V) \leq \mu/(\mu + M)$*,* g *is Lipschitz.*

Proof Fix $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^{d_y}$, and set $\mathbf{x}_1 \stackrel{\text{def}}{=} g(\mathbf{y}_1)$, and $\mathbf{x}_2 \stackrel{\text{def}}{=} g(\mathbf{y}_2)$. It is well-known that under H2-(1), and for all $\mathbf{y} \in \mathbb{R}^{d_y}$, $g(\mathbf{y})$ is the unique solution of the equation

$$\text{Prox}_{\gamma\mathcal{R}}^{V}\left(\mathbf{x} - \gamma V^{-1}\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})\right) = \mathbf{x}.$$

For a proof of this statement see (Combettes and Wajs (2005) Proposition 3.1) with appropriate modification to account for V. Using this fixed point representation of $g(\mathbf{y})$, and the expression $\text{Prox}_{\mathcal{R}}^{V}(\mathbf{x}) = V^{-\frac{1}{2}} \circ \text{Prox}_{\mathcal{R}\circ V^{-\frac{1}{2}}}\left(V^{\frac{1}{2}}\mathbf{x}\right)$, we have

$$\mathbf{x}_i = V^{-1/2}\,\text{Prox}_{\gamma\mathcal{R}\circ V^{-1/2}}\left[V^{1/2}\left(\mathbf{x}_i - \gamma V^{-1}\nabla_{\mathbf{x}} f(\mathbf{y}_i|\mathbf{x}_i)\right)\right], \quad i \in \{1, 2\}.$$

The proximal operator being nonexpansive we deduce that

$$\|V^{1/2}\mathbf{x}_1 - V^{1/2}\mathbf{x}_2\|_2 \leq \|V^{1/2}\left(\mathbf{x}_1 - \gamma V^{-1}\nabla_{\mathbf{x}} f(\mathbf{y}_1|\mathbf{x}_1)\right) - V^{1/2}\left(\mathbf{x}_2 - \gamma V^{-1}\nabla_{\mathbf{x}} f(\mathbf{y}_2|\mathbf{x}_2)\right)\|_2.$$

Under the stated assumptions, for $0 < 2\gamma M/\lambda_{\min}(V) \leq \mu/(\mu + M)$, the function $\mathbf{x} \to V^{1/2}(\mathbf{x} - \gamma V^{-1}\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x}))$ is a contraction for all $\mathbf{y} \in \mathbb{R}^{d_y}$ (See Proposition 14). We use this to write

$$\|\mathbf{x}_1 - \mathbf{x}_2\|_V \leq \left(1 - \frac{2\gamma\mu M}{\lambda_{\min}(V)(\mu + M)}\right)^{1/2}\|\mathbf{x}_1 - \mathbf{x}_2\|_V + \gamma\|V^{-1/2}\left(\nabla_{\mathbf{x}} f(\mathbf{y}_1|\mathbf{x}_2) - \nabla_{\mathbf{x}} f(\mathbf{y}_2|\mathbf{x}_2)\right)\|_2.$$

It follows that

$$\|\mathbf{x}_1 - \mathbf{x}_2\|_V \leq \sqrt{\lambda_{\min}(V)}\left(1 + \frac{M}{\mu}\right)\|\mathbf{y}_1 - \mathbf{y}_2\|_2,$$

as claimed. ■

Lemma 16 *Assume H2 and H3. Let* $\kappa \stackrel{\text{def}}{=} \lambda_{\max}(V)/\lambda_{\min}(V)$. *Given* $\epsilon > 0$, *we can find* $W \in \mathcal{W}$ *as described in H3 such that*

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

Proof For any $\mathbf{y} \in \mathbb{R}^{d_y}$, and for all W, we can write

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

By Assumption H2, we have

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

For $\mathbf{y} \in \mathbb{R}^{d_y}$, let $G_{\gamma,\mathbf{y}}(\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{x} - \gamma V^{-1}\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x})$. Let H_W as constructed in H3. Since $F_{\mathbf{y}}$ is non-expansive as shown in Lemma 14, by the telescoping argument (39), we have

$$\begin{aligned}\left\| F^{D'}_{\mathbf{y}_i}(\mathbf{x}^{(0)}) - F^{D'}_{\mathbf{y}_i,W}(\mathbf{x}^{(0)}) \right\|_2 &\le \kappa \sum_{j=1}^{D'} \| F_{\mathbf{y}_i,W}\left(F^{j-1}_{\mathbf{y}_i,W}(\mathbf{x}^{(0)})\right) - F_{\mathbf{y}_i}\left(F^{j-1}_{\mathbf{y}_i,W}(\mathbf{x}^{(0)})\right) \|_2, \\ &= \kappa \sum_{j=1}^{D'} \| H_W \circ G_{\gamma,\mathbf{y}}\left(F^{j-1}_{\mathbf{y}_i,W}(\mathbf{x}^{(0)})\right) - \mathrm{Prox}^V_{\gamma\mathcal{R}} \circ G_{\gamma,\mathbf{y}}\left(F^{j-1}_{\mathbf{y}_i,W}(\mathbf{x}^{(0)})\right) \|_2.\end{aligned}$$

By assumption H3, $\|F^j_{\mathbf{y}_i,W}(\mathbf{x}^{(0)})\|_2 \le R_1$ for all $1 \le i \le n$, and all $j \ge 1$. Hence

$$\left\| G_{\gamma,\mathbf{y}}\left(F^{j-1}_{\mathbf{y}_i,W}(\mathbf{x}^{(0)})\right) \right\|_2 \le R_1' \stackrel{\text{def}}{=} \max_{1\le i\le n} \max_{\mathbf{x}\in\mathbb{R}^{d_x}:\, \|\mathbf{x}\|_2 \le R} \left\| G_{\gamma,\mathbf{y}_i}(\mathbf{x}) \right\|_2.$$

It follows that

$$\left\| F^{D'}_{\mathbf{y}_i}(\mathbf{x}^{(0)}) - F^{D'}_{\mathbf{y}_i,W}(\mathbf{x}^{(0)}) \right\|_2 \;\le\; D'\kappa \sup_{\mathbf{x}\in\mathbb{R}^{d_x},\, \|\mathbf{x}\|_2 \le R_1'} \|H_W(\mathbf{x}) - \mathrm{Prox}^V_{\gamma\mathcal{R}}(\mathbf{x})\|_2 \;\le\; D'\kappa\epsilon.$$

The result follows by taking the max over i. ■

Lemma 17 *Assume H2, H3, and let* $\{g_W,\ W \in \mathcal{W}\}$ *be as in (11). Let* $\kappa \stackrel{\text{def}}{=} \lambda_{\max}(V)/\lambda_{\min}(V)$. *For any* $\eta > 0$, *and any* $W, W' \in \mathcal{W}$, *such that* $\max(\|W\|_2, \|W'\|_2) \le \eta$, *we have*

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

with

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

and

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

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 $\mathbf{x} \mapsto G_{\gamma,\mathbf{y}}(\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{x} - \gamma V^{-1} \nabla f(\mathbf{y}|\mathbf{x})$ is non-expansive on $\mathbb{R}^{d_x}$ in the norm $\|\cdot\|_V$ (see the proof of Lemma 14). First, we apply (40) to obtain that for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^{d_x}$,

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

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

$$\begin{aligned}
&\|F_{\mathbf{y},W}(\mathbf{x}_1) - F_{\mathbf{y},W}(\mathbf{x}_2)\|_2 \\
&\qquad \leq \frac{\lambda_W}{\lambda_{\min}(V)} \|\mathbf{x}_1 - \mathbf{x}_2 - \gamma V^{-1} \left(\nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x}_1)\right) - \nabla_{\mathbf{x}} f(\mathbf{y}|\mathbf{x}_2)) \|_V \\
&\qquad\qquad \leq \frac{\lambda_W \lambda_{\max}(V)}{\lambda_{\min}(V)} \|\mathbf{x}_1 - \mathbf{x}_2\|_2. \quad (42)
\end{aligned}$$

We recall that $\kappa \stackrel{\text{def}}{=} \lambda_{\max}(V)/\lambda_{\min}(V)$. Setting $r_W \stackrel{\text{def}}{=} \kappa \lambda_W$, and using (42), and the telescoping identity (39), we obtain

$$\|g_W(\mathbf{y}) - g_{W'}(\mathbf{y})\|_2 \leq \sum_{j=0}^{D'} r_W^{D'-j} \left\| F_{\mathbf{y},W}\left(F_{\mathbf{y},W'}^{j-1}(\mathbf{x}^{(0)})\right) - F_{\mathbf{y},W'}\left(F_{\mathbf{y},W'}^{j-1}(\mathbf{x}^{(0)})\right) \right\|_2. \quad (43)$$

We apply (40)and the non-expansiveness of $G_{\gamma,\mathbf{y}}$ to write that for all $\mathbf{x} \in \mathbb{R}^{d_x}$

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

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

$$\|F_{\mathbf{y},W}^j(\mathbf{x}^{(0)})\|_2 \leq \left(\sum_{\ell=0}^{j} r_W^\ell\right) \left(\|\mathbf{x}^{(0)}\|_2 + \|G_{\gamma,\mathbf{y}}(\mathbf{x}^{(0)})\|_2\right) \leq C_n \sum_{\ell=0}^{j} r_W^\ell, \quad (44)$$

where

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

Setting

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

we then apply (41) to write that for all $\mathbf{x} \in \mathbb{R}^{d_x}$

$$\begin{aligned}
&\left\| F_{\mathbf{y},W}\left(F^{j-1}_{\mathbf{y},W'}(\mathbf{x}^{(0)})\right) - F_{\mathbf{y},W'}\left(F^{j-1}_{\mathbf{y},W'}(\mathbf{x}^{(0)})\right)\right\|_2 \\
&\qquad = \left\| H_W \circ G_{\gamma,\mathbf{y}}\left(F^{j-1}_{\mathbf{y},W'}(\mathbf{x}^{(0)})\right) - H_{W'} \circ G_{\gamma,\mathbf{y}}\left(F^{j-1}_{\mathbf{y},W'}(\mathbf{x}^{(0)})\right)\right\|_2 \\
&\qquad \leq \kappa\lambda_{W,W'} \left\| G_{\gamma,\mathbf{y}}\left(F^{j-1}_{\mathbf{y},W'}(\mathbf{x}^{(0)})\right) - G_{\gamma,\mathbf{y}}(\mathbf{x}^{(0)}) + G_{\gamma,\mathbf{y}}(\mathbf{x}^{(0)})\right\|_2 \\
&\qquad\qquad \leq 2C_n\kappa\lambda_{W,W'} \sum_{\ell=0}^{j} r^{\ell}_{W'}.
\end{aligned}$$

The last display together with (43) yields,

$$\max_{1\leq i\leq n} \|g_W(\mathbf{y}_i) - g_{W'}(\mathbf{y}_i)\|_2 \leq 2C_n\kappa\lambda_{W,W'} \sum_{j=1}^{D'}\sum_{\ell=0}^{j} r_W^{D'-j} r^{\ell}_{W'}. \tag{45}$$

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

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

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

$$\sum_{j=1}^{D'}\sum_{\ell=0}^{j} r_W^{D'-j} r^{\ell}_{W'} \leq \sum_{j=1}^{D'} j r_W^{D'-j} r^{j}_{W'} \leq (D')^2 \left(\kappa + \frac{\kappa\eta^2}{D}\right)^{DD'/2},$$

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

$$\max_{1\leq i\leq n} \|g_W(\mathbf{y}_i) - g_{W'}(\mathbf{y}_i)\|_2 \leq C_n\sqrt{D}(D')^2 \left(\kappa + \frac{\kappa\eta^2}{D}\right)^{DD'/2} \|W - W'\|_2. \tag{46}$$

The statement in the lemma follows by noting that

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

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

A.2.1 Proof of Theorem 4

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 16, 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 17, 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

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

Therefore the term L_s in Theorem 7 scales like

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

which gives,

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

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

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

■