Journal of Machine Learning Research 1 (2017) 1-26 Submitted 11/17; Published TBA

A Probably-Approximately-Correct Algorithm for Learning a $C^{1,1}(\mathbb{R}^d)$ Function from Noisy Samples

Adam Gustafson ADAM.MARC.GUSTAFSON@GMAIL.COM
Department of Statistics, University of Washington, Seattle, WA 98195-4322 USA

Matthew Hirn MHIRN@MSU.EDU
Department of Mathematics and Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI, 48824 USA

Kitty Mohammed KITTYMOHAMMED1985@GMAIL.COM
Department of Statistics, University of Washington, Seattle, WA 98195-4322 USA

Hariharan Narayanan HARIHARAN.NARAYANAN@TIFR.RES.IN
School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai, Maharashtra 400 005 India

Jason Xu JQXU@UCLA.EDU
Department of Biomathematics, University of California, Los Angeles, CA 90095-1766 USA

Editor: Kevin Murphy and Bernhard Schölkopf

Abstract

One means of fitting functions to high-dimensional data is by providing smoothness constraints. Recently, the following smooth function approximation problem was proposed by Herbert-Voss, Hirn, and McCollum (2017): given a finite set $E \subset \mathbb{R}^d$ and a function $f : E \to \mathbb{R}$, interpolate the given information with a function $\hat{f} \in C^{1,1}(\mathbb{R}^d)$ (the class of first-order differentiable functions with Lipschitz gradients) such that $\hat{f}(a) = f(a)$ for all $a \in E$, and the value of $\mathrm{Lip}(\nabla \hat{f})$ is minimal. An algorithm is provided that constructs such an approximating function $\hat{f}$ and estimates the optimal Lipschitz constant $\mathrm{Lip}(\nabla \hat{f})$ in the noiseless setting.

We address statistical aspects of reconstructing the approximating function $\hat{f}$ given samples from noisy data. We observe independent and identically distributed samples $y(a) = f(a) + e_a$ for $a \in E$, where e_a is a noise term and the set $E \subset \mathbb{R}^d$ is fixed and known. We obtain uniform bounds relating the empirical risk and true risk over the class $\mathcal{F} = \left\{ f \in C^{1,1}(\mathbb{R}^d) \;\middle|\; \mathrm{Lip}(\nabla f) \leqslant \widetilde{M} \right\}$, where the quantity $\widetilde{M}$ grows with the number of samples at a rate governed by the metric entropy of the class $C^{1,1}(\mathbb{R}^d)$. Finally, we provide an implementation using Vaidya's algorithm, supporting our results via numerical experiments on simulated data.

Keywords: sample complexity, function interpolation, nonparametric regression

1. Introduction

Regression tasks are prevalant throughout statistical learning theory and machine learning. Given n samples in $E \subset \mathbb{R}^d$ and corresponding values $\mathcal{Y} = \{y(a)\}_{a \in E} \subset \mathbb{R}$, a regression function $f : \mathbb{R}^d \to \mathbb{R}$ learns a model for the data $(E, \mathcal{Y})$ that best generalizes to new points $x \notin E$. Absent any prior information on x, the best regression function $\hat{f}$, as measured by

the squared loss, is obtained by minimizing the ℓ^2 empirical risk over a specified function class $\mathcal{F}$,

$$\hat{f} = \arg\inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{a \in E} |f(a) - y(a)|^2,$$

subject to a regularization penalty. If $\mathcal{F}$ is equipped with a norm or semi-norm $\|\cdot\|_{\mathcal{F}}$, then the regularized risk can take either the form

$$\hat{f} = \arg\inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{a \in E} |f(a) - y(a)|^2 + \lambda \cdot \Omega(\|f\|_{\mathcal{F}}), \tag{1}$$

or

$$\hat{f} = \arg\inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{a \in E} |f(a) - y(a)|^2 \quad \text{subject to} \quad \|f\|_{\mathcal{F}} \leqslant M, \tag{2}$$

where λ and M are hyper-parameters, and $\Omega : [0, \infty) \to \mathbb{R}$ is a monotonically increasing function.

In either case, the quality of $\hat{f}$ is primarily determined by the functional class $\mathcal{F}$. Due to theoretical and computational considerations, $\mathcal{F}$ is often taken to be the span of a suitably defined dictionary of functions, or a reproducing kernel Hilbert space (RKHS) with an appropriate kernel. For example, when $\Omega(\|f\|_{\mathcal{F}}) = \frac{1}{2}\|f\|_{\mathcal{F}}^2$ and $\mathcal{F}$ is an RKHS, equation (1) leads to the popular kernel ridge regression scheme, which has a closed form solution that is simple to compute.

When $\mathcal{F} = \text{span}(\{\phi_k\}_k)$, the smoothness of $\hat{f}$ is determined by the dictionary $\{\phi_k\}_k$, or if $\mathcal{F}$ is a reproducing kernel Hilbert space, the regularity of $\hat{f}$ is determined by the kernel. An alternate approach that does not require choice of dictionary or kernel is to specify the smoothness of $\hat{f}$ directly, by taking $\mathcal{F} = C^m(\mathbb{R}^d)$ or $\mathcal{F} = C^{m-1,1}(\mathbb{R}^d)$. However, the computational complexity of minimizing the regularized risk over these spaces is generally prohibitive. An exception is the space $C^{0,1}(\mathbb{R}^d)$, which consists of functions f with finite Lipschitz constant, and for which several regression algorithms exist (von Luxburg and Bousquet, 2004; Beliakov, 2006; Gottlieb et al., 2013; Kyng et al., 2015).

In recent work, Herbert-Voss, Hirn, and McCollum (2017) provide an efficient algorithm for computing the interpolant $\hat{f} \in C^{1,1}(\mathbb{R}^d)$ that, given noiseless data $(E, \mathcal{Y})$, minimizes the Lipschitz constant of the gradient. In this paper we extend the methods of Herbert-Voss et al. (2017) to regularized risk optimizations of the form (2). In particular, we consider the noisy scenario in which the function to be reconstructed is not measured precisely on a finite subset, but instead is measured with some uncertainty.

An outline of this paper is as follows. In Section 1.1, we introduce the function interpolation problem considered by Herbert-Voss et al. (2017), and summarize the solution in the noiseless case in Section 1.2. Next, we consider the setting where the function is measured under uncertainty, and derive uniform sample complexity bounds on our estimator in Section 2.1. The resulting optimization problem can be solved using an algorithm due to Vaidya (1996); we provide details on computing the solution to the regularized risk in Sections 2.2 and 2.3. We implement the estimator and present reconstruction results on simulated data examples in Section 3, supporting our theoretical contributions, and close with a discussion.

1.1 Noiseless Function Interpolation Problem

Here we summarize the function approximation problem considered by Herbert-Voss et al. (2017). First, recall that the Lipschitz constant of an arbitrary function $g : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\mathrm{Lip}(g) := \sup_{x,y\in\mathbb{R}^d,\, x\neq y} \frac{|g(x) - g(y)|}{|x - y|},$$

where $|\cdot|$ denotes the standard Euclidean norm. Second, denote the gradient of such an arbitrary g as $\nabla g = (\frac{\partial g}{\partial x_1}, \ldots, \frac{\partial g}{\partial x_d})$. Finally, let $C^{m,1}(\mathbb{R}^d)$ be the class of m-times continuously differentiable functions whose derivatives have a finite Lipschitz constant. In the function approximation problem, we are given a finite set of points $E \subset \mathbb{R}^d$ such that $|E| = n$, and a function $f : E \to \mathbb{R}$ specified on E. The function approximation problem as stated by Herbert-Voss et al. (2017) is to compute an interpolating function $\hat{f} \in C^{1,1}(\mathbb{R}^d)$ such that

1. $\hat{f}(a) = f(a)$ for all $a \in E$.
2. Among all such interpolating functions satisfying condition (1), the value of $\mathrm{Lip}(\nabla \hat{f})$ is minimal.

The question of whether one can even reconstruct such an interpolating function was answered by Whitney (1934). Presume that we also have access to the gradients of f on E, and denote them $\{D_a f\}_{a\in E}$. In the case of $C^{1,1}(\mathbb{R}^d)$, the polynomials are defined by the specified function and gradient information:

$$P_a(x) = f(a) + D_a f \cdot (x - a), \quad a \in E,\ x \in \mathbb{R}^d.$$

Letting $\mathcal{P}$ denote the space of first-order polynomials (i.e., affine functions), the map $P : \mathbb{R}^d \to \mathcal{P}, a \mapsto P_a$ is known as a *1-field*. For any $f \in C^{1,1}(\mathbb{R}^d)$, the first order Taylor expansions of f are elements of $\mathcal{P}$, and are known as *jets* (Fefferman and Klartag, 2009), defined as:

$$J_a f(x) := f(a) + \nabla f(a) \cdot (x - a), \quad a, x \in \mathbb{R}^d.$$

Whitney's Extension Theorem for $C^{1,1}(\mathbb{R}^d)$ may be stated as follows:

Theorem 1 (Whitney's Extension Theorem for $C^{1,1}(\mathbb{R}^d)$) *Let $E \subset \mathbb{R}^d$ be closed and let $P : E \to \mathcal{P}$ be a 1-field with domain E. If there exists a constant $M < \infty$ such that*

(W_0) $|P_a(a) - P_b(a)| \leqslant M|a - b|^2$ for all $a, b \in E$, and

(W_1) $|\frac{\partial P_a}{\partial x_i}(a) - \frac{\partial P_b}{\partial x_i}(a)| \leqslant M|a - b|$ for all $a, b \in E$, and $i \in \{1, \ldots, d\}$,

then there exists an extension $f \in C^{1,1}(\mathbb{R}^d)$ such that $J_a f = P_a$ for all $a \in E$.

Given a finite set E as in the function interpolation problem, these conditions are automatically satisfied. However, this theorem does not provide a solution for the minimal Lipschitz constant of ∇f. Le Gruyer (2009) provides a solution to both problems, which we discuss next.

1.2 Minimal Value of Lip(∇f)

Herbert-Voss et al. (2017) define the following norm for when the first-order polynomials are known

$$\|P\|_{C^{1,1}(E)} := \inf\left\{\mathrm{Lip}(\nabla f) \;\middle|\; J_a f = P_a \text{ for all } a \in E\right\}, \tag{3}$$

and similarly define

$$\|f\|_{C^{1,1}(E)} := \inf\left\{\mathrm{Lip}(\nabla f) \;\middle|\; f(a) = f(a) \text{ for all } a \in E\right\}, \tag{4}$$

when the gradients $\{D_a f\}_{a\in E}$ are unknown. Presuming we are given the 1-field $P : E \to \mathcal{P}$, $a \mapsto P_a$, Le Gruyer (2009) defines the functional Γ^1 as:

$$\Gamma^1(P; E) = 2 \sup_{x\in\mathbb{R}^d} \left(\max_{a,b\in E,\, a\neq b} \frac{P_a(x) - P_b(x)}{|a-x|^2 + |b-x|^2} \right). \tag{5}$$

Given only functions f, Le Gruyer (2009) also defines the functional Γ^1 in terms of f as

$$\Gamma^1(f; E) = \inf\left\{\Gamma^1(P; E) \;\middle|\; P_a(a) = f(a) \text{ for all } a \in E\right\},$$

The following theorem is proven by Le Gruyer (2009), which shows that (5) and its equivalent formulation in (6) provides a solution for (3):

Theorem 2 (Le Gruyer) *Given a set $E \subset \mathbb{R}^d$ and a 1-field $P : E \to \mathcal{P}$,*

$$\Gamma^1(P; E) = \|P\|_{C^{1,1}(E)}.$$

An equivalent formulation of (5) which is amenable to implementation is as follows. Consider the following functionals mapping $E \times E \to [0, \infty]$:

$$\begin{aligned} A(P; a, b) &= \frac{(P_a(a) - P_b(a)) + (P_a(b) - P_b(b))}{|a-b|^2} \\ &= \frac{2(f(a) - f(b)) + (D_a f - D_b f)\cdot(b-a)}{|a-b|^2}, \\ B(P; a, b) &= \frac{|\nabla P_a(a) - \nabla P_b(a)|}{|a-b|} \\ &= \frac{|D_a f - D_b f|}{|a-b|}. \end{aligned}$$

Proposition 2.2 of Le Gruyer (2009) states that

$$\Gamma^1(P; E) = \max_{a\neq b\in E} \sqrt{A(P; a, b)^2 + B(P; a, b)^2} + |A(P; a, b)|, \tag{6}$$

whence a naive implementation allows $\Gamma^1(P; E)$ to be found in $O(n^2)$ computations. Inspired by Fefferman and Klartag (2009), Herbert-Voss et al. (2017) also construct algorithms which will solve for the order of magnitude of $\|P\|_{C^{1,1}(E)}$ in $O(n \log n)$ time, but we omit the details here. Additionally, as a consequence of the proof of Proposition 2.2 of Le Gruyer (2009), equation (5) may alternatively be written as

$$\Gamma^1(P; E) = 2 \max_{a,b\in E:\, a\neq b} \sup_{x\in B^d\left(\frac{a+b}{2}, \frac{|a-b|}{2}\right)} \frac{P_a(x) - P_b(x)}{|a-x|^2 + |b-x|^2}, \tag{7}$$

where $B^d(z,r)$ denotes the closed d-dimensional Euclidean ball centered at z with radius r.

Recall that the gradients $\{D_a f\}_{a\in E}$ are typically not known in applications. As a corollary, we have the following convex optimization problem for finding (4), and the minimizing 1-field provides the gradients $\{D_a f(a)\}_{a\in E}$.

Corollary 3 *Given a set $E \subset \mathbb{R}^d$ and a function $f : E \to \mathbb{R}$,*

$$\Gamma^1(f;E) = \|f\|_{C^{1,1}(E)}.$$

Recall that $P_a(x) = f(a) + D_a f \cdot (x-a)$ and we set $P_a(a) = f(a)$. The set $E \subset \mathbb{R}^d$ and the values $\{f(a)\}_{a\in E}$ are fixed, so the optimization problem is to solve for the gradients $\{D_a f\}_{a\in E}$ that minimize $\Gamma^1(P;E)$.

2. $C^{1,1}(E)$ Regression

In statistical applications where $f(a)$ is observed with uncertainty, one often assumes that we observe $\{y(a)\}_{a\in E}$, where $y(a) = f(a) + e(a)$, and $e(a)$ is assumed to be independent and identically distributed Gaussian noise for each $a \in E$. Since both the function values and the gradients $\{f(a), D_a f\}_{a\in E}$ are unknown, we minimize an empirical squared error loss over the $k := (d+1)n$ variables defining the 1-field. Given a bound on the $C^{1,1}$-norm of the unknown 1-field, regression entails solving an optimization problem of the form

$$\begin{aligned} \min_{P} \quad & \frac{1}{n}\sum_{a\in E}(y(a) - P_a(a))^2 \\ \text{s.t.} \quad & \|P\|_{C^{1,1}(E)} \leqslant M. \end{aligned} \tag{8}$$

This is a convex optimization problem: the objective function of the empirical squared error loss in (8) is convex, as is the constraint set since it is a ball specified by a seminorm. This section proceeds as follows: we begin by analyzing the sample complexity of the function class. These risk bounds establish almost sure convergence of the empirical risk minimizer, and guides the choice of M. Given M, we next appeal to Vaidya's algorithm to solve the resulting optimization problem (8). This allows us to apply efficient implementations of Wells' construction for finding the optimal interpolating function.

2.1 Sample Complexity and Empirical Risk Minimization

The constant $M > 0$ will be chosen via sample complexity arguments. To this end, we derive uniform risk bounds for classes of continuous functions $f : B \to \mathbb{R}$, where throughout this section B is the unit ball in $\mathbb{R}^d$. The function classes of interest are defined in terms of $C^{1,1}$-norm balls as

$$\mathcal{F}_{\widetilde{M}} = \left\{ f \;\middle|\; \|f\|_{C^{1,1}} \leqslant \widetilde{M} \right\},$$

where we are using the norm

$$\|f\|_{C^{1,1}} = \max\left\{ \sup_{x\in B} |f|\,,\ \sup_{x\in B} |\nabla f|\,,\ \sup_{x\in B} \mathrm{Lip}(\nabla f) \right\}.$$

We observe an i.i.d. sample

$$S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

drawn from a probability distribution $\mathcal{P} = S_X \times S_{Y|X}$ supported on $X \times Y \subset B \times \mathbb{R}$ under the assumption

$$S_{Y|X} \sim \mathcal{N}(f^*(X), \sigma^2), \quad \text{where } \|f^*\|_{C^{1,1}} = M^*.$$

(We only assume S_X is supported on the unit ball). Since we are in the regression setting, we use squared error loss

$$L(f(x), y) = (f(x) - y)^2 .$$

The true risk is defined as the expectation of L over $\mathcal{P}$:

$$R(f) = \mathbb{E}_{\mathcal{P}} \left[L(f(x), y) \right],$$

and the empirical risk is the expectation over S_S, the empirical distribution on the sample S:

$$\widehat{R}(f) = \mathbb{E}_{S_S} \left[L(f(x), y) \right].$$

In order for the empirical risk minimization procedure to converge to a minimizer of the true risk, we need to bound

$$\sup_{f \in \mathcal{F}_{\widetilde{M}}} \left| \widehat{R}(f) - R(f) \right|$$

with high probability. The most natural way to do so is by expanding the risk and appealing to entropy methods (i.e., covering number bounds) and standard concentration results. Recall that the covering number $N(\eta, \mathcal{G}, \|\cdot\|)$ is the minimum number of norm balls of radius η needed to cover a function class $\mathcal{G}$. We briefly discuss how this is useful toward deriving uniform bounds.

Given a class $\mathcal{G}$ of bounded functions $g : B \to \mathbb{R}$, an i.i.d. sample $T = \{z_1, \dots z_n\}$ drawn from a probability distribution $\mathcal{Q}$ supported on B, and a vector of i.i.d. Rademacher random variables $\sigma = (\sigma_1, \dots, \sigma_n)$, the following holds for $0 < \delta < 1$:

$$\mathbb{P}\left[\sup_{g \in \mathcal{G}} |\mathbb{E}_{\mathcal{Q}_T} g - \mathbb{E}_{\mathcal{Q}} g| < 2\mathcal{R}_n(\mathcal{G}) + \sqrt{\frac{2\log(2/\delta)}{n}} \right] > 1 - \delta,$$

where $\mathcal{Q}_T$ is the empirical distribution on T and

$$\mathcal{R}_n(\mathcal{G}) = \mathbb{E}_\sigma \frac{1}{n} \left[\sup_{g \in \mathcal{G}} \left(\sum_{i=1}^{n} \sigma_i g(x_i) \right) \right]$$

is the Rademacher average conditional on the sample. Sridharan and Srebro (2010) show that

$$\mathcal{R}_n(\mathcal{G}) \leqslant \inf_{\gamma \geqslant 0} \left\{ 4\gamma + 12 \int_{\gamma}^{\sup_{g \in \mathcal{G}} \|g\|_\infty} \sqrt{\frac{\log N(\eta, \mathcal{G}, \|\cdot\|_{\mathcal{L}_2(\mathcal{Q}_T)})}{n}} d\eta \right\},$$

where the right-hand side is a modified version of Dudley's entropy integral.

Since we are interested in bounding the risk, we use Lemma 5 to relate the Rademacher complexity of the loss class and the original class. We provide a proof based on three results due to Bartlett and Mendelson (2003); these require familiarity with McDiarmid's inequality, stated below in Lemma 4.

Lemma 4 (McDiarmid's inequality, McDiarmid, 1989) *Let $X_1, \ldots, X_n$ be independent random variables that take values in a set A. Suppose the function $f : A^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_n,x_i' \in A} \left| f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \right| \leqslant c_i$$

for every $1 \leqslant i \leqslant n$. Then, for $t > 0$,

$$\mathbb{P}\left[|f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n)| \geqslant t \right] \leqslant 2e^{-2t^2/\sum_{i=1}^n c_i^2}.$$

Lemma 5 *Let $\mathcal{F}_{\widetilde{M}}$ be a class of functions $f : B \to \mathbb{R}$ with $\sup_{f\in\mathcal{F}} |f| \leqslant \widetilde{M}$. Let $L : [-\widetilde{M}, \widetilde{M}] \times Y \to \mathbb{R}$ be a bounded loss function with Lipschitz constant L_L and $0 \leqslant L \leqslant L_{max}$. Then, the following is true for $0 < \delta < 1$:*

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}_{\widetilde{M}}} \left| R(f) - \widehat{R}(f) \right| < 4L_L \mathcal{R}_n(\mathcal{F}_{\widetilde{M}}) + 7L_{max}\sqrt{\frac{\log(8/\delta)}{2n}} \right] > 1 - \delta.$$

Proof In this lemma, we begin by adapting Theorem 8 of Bartlett and Mendelson (2003) to find a bound on the risk that depends on a probabilistic term plus the expectation of the Rademacher average of the class of loss functions. We follow the proof of Lemma 4 of Bousquet et al. (2004) for guidance. We apply the two-sided form of McDiarmid's inequality as we want bounds on the absolute value of $R(f) - \widehat{R}(f)$, and appeal to Theorems 11 and 12 of Bartlett and Mendelson (2003) to relate the expected Rademacher average of the loss class to the empirical Rademacher average of $\mathcal{F}_{\widetilde{M}}$.

Let $\widetilde{L} \circ \mathcal{F}_{\widetilde{M}}$ be the class of functions consisting of $\{(x, y) \mapsto L(f(x), y) - L(0, y)\}$. If $h \in \widetilde{L} \circ \mathcal{F}_{\widetilde{M}}$, then $-L_{max} \leqslant h \leqslant L_{max}$. For any $f \in \mathcal{F}_{\widetilde{M}}$, the triangle inequality shows that

$$\left| R(f) - \widehat{R}(f) \right| \leqslant \sup_{h\in\widetilde{L}\circ\mathcal{F}_{\widetilde{M}}} \left| \mathbb{E}h - \widehat{\mathbb{E}}_n h \right| + \left| \mathbb{E}L(0, y) - \widehat{\mathbb{E}}_n L(0, y) \right|.$$

McDiarmid's inequality yields more favorable expressions for both terms on the right-hand side as follows. The most that $\widehat{\mathbb{E}}_n L(0, y)$ can change by altering one sample is L_{max}/n. Since $\mathbb{E}\widehat{\mathbb{E}}_n L(0, y) = \mathbb{E}L(0, y)$, we have, with probability $1 - \delta/4$,

$$\left| \mathbb{E}L(0, y) - \widehat{\mathbb{E}}_n L(0, y) \right| \leqslant \sqrt{\frac{L_{max}^2 \log 8/\delta}{2n}}.$$

The most that $\sup_{h\in\widetilde{L}\circ\mathcal{F}_{\widetilde{M}}} \left| \mathbb{E}h - \widehat{\mathbb{E}}_n h \right|$ can change with an alteration of one sample is $2L_{max}/n$. Therefore, with probability $1 - \delta/4$,

$$\left| \sup_{h\in\widetilde{L}\circ\mathcal{F}_{\widetilde{M}}} \left| \mathbb{E}h - \widehat{\mathbb{E}}_n h \right| - \mathbb{E} \sup_{h\in\widetilde{L}\circ\mathcal{F}_{\widetilde{M}}} \left| \mathbb{E}h - \widehat{\mathbb{E}}_n h \right| \right| \leqslant \sqrt{\frac{4L_{max}^2 \log 8/\delta}{2n}}.$$

Now,

$$\mathbb{E} \sup_{h\in\widetilde{L}\circ\mathcal{F}_{\widetilde{M}}} \left| \mathbb{E}h - \widehat{\mathbb{E}}_n h \right| \leqslant \max \left\{ \mathbb{E} \sup_{h\in\widetilde{L}\circ\mathcal{F}_{\widetilde{M}}} \left(\mathbb{E}h - \widehat{\mathbb{E}}_n h \right), \mathbb{E} \sup_{h\in\widetilde{L}\circ\mathcal{F}_{\widetilde{M}}} \left(\widehat{\mathbb{E}}_n h - \mathbb{E}h \right) \right\}.$$

Let $S' := \{(x_1', y_1'), \ldots, (x_n', y_n')\}$ be a sample with the same distribution as S. Conditioning on the original sample,

$$\begin{aligned}
\mathbb{E} \sup_{h \in \widetilde{L} \circ \mathcal{F}_{\widetilde{M}}} \left(\mathbb{E}h - \widehat{\mathbb{E}}_n h \right) &= \mathbb{E} \sup_{h \in \widetilde{L} \circ \mathcal{F}_{\widetilde{M}}} \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n h(x_i', y_i') - \widehat{\mathbb{E}}_n h \middle| S \right] \\
&\leqslant \mathbb{E} \sup_{h \in \widetilde{L} \circ \mathcal{F}_{\widetilde{M}}} \left(\frac{1}{n} \sum_{i=1}^n h(x_i', y_i') - \widehat{\mathbb{E}}_n h \right) \\
&= \mathbb{E} \sup_{h \in \widetilde{L} \circ \mathcal{F}_{\widetilde{M}}} \frac{1}{n} \sum_{i=1}^n \sigma_i \left(h(x_i', y_i') - h(x_i, y_i) \right) \\
&\leqslant \mathbb{E} \sup_{h \in \widetilde{L} \circ \mathcal{F}_{\widetilde{M}}} \frac{1}{n} \sum_{i=1}^n \sigma_i h(x_i', y_i') + \mathbb{E} \sup_{h \in \widetilde{L} \circ \mathcal{F}_{\widetilde{M}}} \frac{1}{n} \sum_{i=1}^n -\sigma_i h(x_i, y_i) \\
&= 2\mathbb{E}\mathcal{R}_n(\widetilde{L} \circ \mathcal{F}_{\widetilde{M}}).
\end{aligned}$$

The second line follows by applying Jensen's inequality to sup, which is convex. Note the preceding argument is symmetric in $\mathbb{E}h$ and $\hat{E}_n h$. Therefore, $\mathbb{E} \sup_{h \in \widetilde{L} \circ \mathcal{F}_{\widetilde{M}}} \left| \mathbb{E}h - \widehat{\mathbb{E}}_n h \right|$ has the same upper bound and, with probability $1 - \delta/2$,

$$\left| R(f) - \hat{R}(f) \right| \leqslant 2\mathbb{E}\mathcal{R}_n(\widetilde{L} \circ \mathcal{F}_{\widetilde{M}}) + 3L_{max} \sqrt{\frac{\log(8/\delta)}{2n}}.$$

Theorem 11 of Bartlett and Mendelson (2003) uses McDiarmid's inequality to bound the difference between the empirical and expected Rademacher averages, but assumes that we are interested in the Rademacher complexity of a class of functions mapping to $[-1, 1]$. Since $\mathcal{F}_{\widetilde{M}}$ maps to $[-\widetilde{M}, \widetilde{M}]$, we rederive the analogous result here. The most that one sample affects $\mathcal{R}_n(\widetilde{L} \circ \mathcal{F}_{\widetilde{M}})$ is $2L_{max}/n$. We have

$$\mathbb{P}\left[\left| \mathcal{R}_n(\widetilde{L} \circ \mathcal{F}_{\widetilde{M}}) - \mathbb{E}\mathcal{R}_n(\widetilde{L} \circ \mathcal{F}_{\widetilde{M}}) \right| \geqslant t \right] \leqslant 2e^{-2nt^2/(4L_{max}^2)}.$$

Thus, with probability $1 - \delta/2$,

$$\begin{aligned}
2\mathbb{E}\mathcal{R}_n(\widetilde{L} \circ \mathcal{F}_{\widetilde{M}}) &\leqslant 2\mathcal{R}_n(\widetilde{L} \circ \mathcal{F}_{\widetilde{M}}) + 4L_{max} \sqrt{\frac{\log(4/\delta)}{2n}} \\
&\leqslant 4L_L \mathcal{R}_n(\mathcal{F}_{\widetilde{M}}) + 4L_{max} \sqrt{\frac{\log(8/\delta)}{2n}}.
\end{aligned}$$

The second line follows from part 4 of Theorem 12, which states that, for $\widetilde{L} : \mathbb{R} \to \mathbb{R}$ with Lipschitz constant $L_{\widetilde{L}}$ and satisfying $\widetilde{L}(0) = 0$, $\mathbb{E}\mathcal{R}_n(\widetilde{L} \circ \mathcal{F}_{\widetilde{M}}) \leqslant 2L_{\widetilde{L}} \mathbb{E}\mathcal{R}_n(\mathcal{F}_{\widetilde{M}})$. The reasoning from the proof also applies to the empirical Rademacher average, giving $\mathcal{R}_n(\widetilde{L} \circ \mathcal{F}_{\widetilde{M}}) \leqslant 2L_{\widetilde{L}} \mathcal{R}_n(\mathcal{F}_{\widetilde{M}})$. Since $\widetilde{L}$ has the same Lipschitz constant as L, we use the notation L_L.

Finally, with probability at least $1 - \delta$,

$$\left| R(f) - \hat{R}(f) \right| \leqslant 4L_L \mathcal{R}_n(\mathcal{F}_{\widetilde{M}}) + 7L_{max} \sqrt{\frac{\log(8/\delta)}{2n}}.$$

■

Next, the following lemma gives an upper bound on the covering number $N(\eta, \mathcal{F}_{\widetilde{M}}, \|\cdot\|_\infty)$ of $\mathcal{F}_{\widetilde{M}}$ with respect to the supremum norm.

Lemma 6 (adapted from Theorem 2.7.1 of Van Der Vaart and Wellner, 1996) *There exists a constant K depending only on d such that, for every $\eta > 0$,*

$$\log N(\eta, \mathcal{F}_{\widetilde{M}}, \|\cdot\|_\infty) \leqslant K\left(\frac{\widetilde{M}}{\eta}\right)^{d/2}.$$

Proof Following Van Der Vaart and Wellner (1996), every $f \in \mathcal{F}_{\widetilde{M}}$ is continuous on the open unit ball B by assumption, Taylor's theorem applies everywhere. Fix $\delta = \varepsilon^{1/2} \leqslant 1$, and take the δ-net of points $x_1, \ldots, x_m$ in B, where the number of points m is less than or equal to the volume of B times a constant that only depends on d. Then for all vectors k whose sum of entries $\overline{k} := \sum_i k_i$ do not exceed 1 (this includes the zero vector and standard basis vectors), define for each f the vector

$$A_k f = \left(\left\lfloor \frac{D^k f(x_1)}{\delta^{2-\overline{k}}} \right\rfloor, \ldots, \left\lfloor \frac{D^k f(x_k)}{\delta^{2-\overline{k}}} \right\rfloor, \ldots, \left\lfloor \frac{D^k f(x_m)}{\delta^{2-\overline{k}}} \right\rfloor\right).$$

The vector $\delta^{2-\overline{k}} A_k f$ thus consists of the $D^k f(x_i)$ values discretized on a mesh with grid width $\delta^{2-\overline{k}}$.

Now, if $f, g \in \mathcal{F}_{\widetilde{M}}$ are such that $A_k f = A_k g$ for all k, then $\|f - g\|_\infty \leqslant C\varepsilon$ for a constant C, implying that for each x there exists an x_i such that $\|x - x_i\| \leqslant \delta$. The remainder term in the Taylor expansion

$$(f-g)(x) = \sum_{k \leqslant 1} D^k(f-g)(x_i)\frac{(x-x_i)^k}{k!} + R$$

is bounded by the mesh width δ: indeed, we may consider an integral form via the Fundamental Theorem of Calculus:

$$\begin{aligned}
(f-g)(x) &= (f-g)(x_i) + \int_{x_i}^{x} D(f-g)(s)\,\mathrm{d}s \\
&= (f-g)(x_i) + \int_{x_i}^{x} D(f-g)(x_i) + D(f-g)(s) - D(f-g)(x_i)\,\mathrm{d}s \\
&= (f-g)(x_i) + D(f-g)(x-x_i) + \int_{x_i}^{x} D(f-g)(s) - D(f-g)(x_i)\,\mathrm{d}s \\
&\leqslant (f-g)(x_i) + D(f-g)(x-x_i) + C\|x-x_i\|^2
\end{aligned}$$

where the last line follows from $f - g \in \mathcal{F}_{\widetilde{M}}$. Therefore we see that the remainder term $|R| \propto \|x - x_i\|^2$, and we may next substitute the mesh width bounding this quantity:

$$|f-g|(x) \propto \sum_{\overline{k} \leqslant 1} \delta^{2-\overline{k}} \prod_{i=1}^{d} \frac{\delta_i^{\overline{k}}}{k_i!} + \delta^2 \leqslant \delta^2(e^d + 1).$$

Thus, there exists $C = C(d)$ such that the covering number $N(C_\varepsilon, \mathcal{F}_{\widetilde{M}}, \|\cdot\|_\infty)$ is bounded by the number of different matrices $\{Af\}$ whose rows are the vectors $A_k f$ for k such that $\overline{k} \leqslant 1$ and f ranges over $\mathcal{F}_{\widetilde{M}}$. There are $d+1$ such vectors. Now, by definition of $A_k f$ and using $|D^k f(x_i)| \leqslant \widetilde{M}$ for all i the number of values of each element in each row is at most $2\widetilde{M}/\delta^{2-\overline{k}} + 1 \leqslant 2\widetilde{M}\delta^{-2} + 1$. Thus, each column of $\{Af\}$ has at most $(2\widetilde{M}\delta^{-2}+1)^{d+1}$ values. Note that this already suffices to produce a finite bound. Following Van Der Vaart and Wellner (1996) and applying Taylor's theorem again yields a less crude bound $\#\{Af\} \leqslant (2\delta^{-2}+1)^{d+1} C^{m-1}$ where C is a constant depending only on d. We may replace δ in this expression by $\varepsilon^{1/2}$ and m by its upper bound $\mathrm{Vol}(B)\varepsilon^{-d/2}$. Now, the lemma follows by taking logarithms, bounding $\log(1/\varepsilon)$ by $K(1/\varepsilon)^{d/2}$, and combining all constant terms into K. ∎

We are now ready to provide risk bounds in the following theorem.

Theorem 7 *Suppose we set $\widetilde{M} := n^{1/(2\widetilde{d})}$, where $\widetilde{d} := \max\{d, 5\}$, and let $\mathcal{F}_{\widetilde{M}}$ be the class of functions with $C^{1,1}$ norm bounded above by $\widetilde{M}$.*

(i) For $0 < \delta < 1$,

$$\mathbb{P}\left[\sup_{f \in \mathcal{F}_{\widetilde{M}}} \left|R(f) - \widehat{R}(f)\right| < \varepsilon\right] > (1-\delta)(1 - e^{-n^{2/\max\{d,5\}}/2\sigma^2}),$$

where ε is a monotonically-decreasing function of n for large enough n and $\lim_{n\to\infty} \varepsilon = 0$.

(ii)

$$\sup_{f \in \mathcal{F}_{\widetilde{M}}} \left|R(f) - \widehat{R}(f)\right| \xrightarrow{a.s.} 0.$$

Proof $\mathcal{F}_{\widetilde{M}}$ is a sequence of function classes with increasing $C^{1,1}$ norm. We set the rate $\widetilde{M} := n^{1/(2\widetilde{d})}$, where $\widetilde{d} := \max\{d, 5\}$, so that f^* is a candidate for large enough n. We aim to use Lemma 5 to prove the desired probability statement, but our loss function is unbounded since Y can be arbitrarily large. To circumvent this, we also let the maximum value of $\{y_i\}$ increase with n; samples violating this condition are part of the error probability. Write $y = f^*(x) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2)$. We condition on the event $\mathcal{E} := \left\{\max_{1\leqslant i\leqslant n} |\xi_i| \leqslant \sigma\sqrt{2\log 2n} + n^{1/\widetilde{d}}\right\}$. Theorem 7.1 of Ledoux (2005) gives the following bound for suprema of Gaussian processes:

$$\mathbb{P}\left[\max_{1\leqslant i\leqslant n} |\xi_i| < \mathbb{E}\max_{1\leqslant i\leqslant n} |\xi_i| + r\right] > 1 - e^{-r^2/2\sigma^2}.$$

The following is well-known (Boucheron, Lugosi, and Massart, 2013):

$$\mathbb{E}\max_{1\leqslant i\leqslant n} |\xi_i| \leqslant \sigma\sqrt{2\log 2n}.$$

Thus, $\mathbb{P}(\mathcal{E}) > 1 - e^{-n^{2/\tilde{d}}/2\sigma^2}$.

Since the loss function is bounded after conditioning on $\mathcal{E}$, we can compute:

$$\begin{aligned} L_{max} &< \sup_{x,y,f} (f(x) - y)^2 \\ &< \sup_{x,y,f} (|f(x)| + |y|)^2 \\ &< \left(\widetilde{M} + M^* + \sigma\sqrt{2\log 2n} + n^{1/\tilde{d}}\right)^2 \\ &= \left(n^{1/(2\tilde{d})} + M^* + \sigma\sqrt{2\log 2n} + n^{1/\tilde{d}}\right)^2 \\ &:= \widetilde{L}_{max}. \end{aligned}$$

We also find the Lipschitz constant as follows, where $f_1, f_2 \in \mathcal{F}_{\widetilde{M}}$:

$$\begin{aligned} \sup_{x,y,f_1,f_2} \left|(f_1(x) - y)^2 - (f_2(x) - y)^2\right| &= \sup_{x,y,f_1,f_2} |(-2y + f_1(x) + f_2(x))(f_1(x) - f_2(x))| \\ &\leqslant \sup_{x,y,f_1,f_2} |-2y + f_1(x) + f_2(x)| \, \|f_1 - f_2\|_\infty. \end{aligned}$$

This implies:

$$\begin{aligned} L_L &\leqslant \sup_{x,f_1,f_2} |f_1(x) + f_2(x)| + 2\sup_y |y| \\ &< 2\left(\widetilde{M} + M^* + \sigma\sqrt{2\log 2n} + n^{1/\tilde{d}}\right) \\ &= 2\left(n^{1/(2\tilde{d})} + M^* + \sigma\sqrt{2\log 2n} + n^{1/\tilde{d}}\right) \\ &:= \widetilde{L}_L. \end{aligned}$$

Next, we bound the Rademacher complexity using the entropy integral:

$$\begin{aligned} \mathcal{R}_n(\mathcal{F}_{\widetilde{M}}) &\leqslant \inf_{\gamma \geqslant 0} \left\{ 4\gamma + 12\int_\gamma^{\widetilde{M}} \sqrt{\frac{\log N(\eta, \mathcal{F}_{\widetilde{M}}, \|\cdot\|_{\mathcal{L}_2(S_S)})}{n}} d\eta \right\} \\ &\leqslant \inf_{\gamma \geqslant 0} \left\{ 4\gamma + 12\int_\gamma^{\widetilde{M}} \sqrt{\frac{\log N(\eta, \mathcal{F}_{\widetilde{M}}, \|\cdot\|_{\infty})}{n}} d\eta \right\} \\ &\leqslant \inf_{\gamma \geqslant 0} \left\{ 4\gamma + 12\int_\gamma^{\widetilde{M}} \sqrt{\frac{K\widetilde{M}^{d/2}}{n\eta^{d/2}}} d\eta \right\}. \end{aligned}$$

The second inequality is standard, and the third is from substituting the covering number bound from Lemma 6. The integral is different for $d \neq 4$ and $d = 4$. In the first case,

$$\mathcal{R}_n(\mathcal{F}_{\widetilde{M}}) \leqslant \inf_{\gamma \geqslant 0} \left\{ 4\gamma + 12\frac{\sqrt{K}\widetilde{M}^{d/4}\left(4\widetilde{M}^{1-d/4} - 4\gamma^{1-d/4}\right)}{(4-d)\sqrt{n}} \right\},$$

and the infimum is achieved at $\gamma = 81^{1/d} K^{2/d} \widetilde{M} n^{-2/d}$. When $d = 4$,

$$\mathcal{R}_n(\mathcal{F}_{\widetilde{M}}) \leqslant \inf_{\gamma \geqslant 0} \left\{ 4\gamma + 12 \frac{\sqrt{K}\widetilde{M}}{\sqrt{n}} (\log \widetilde{M} - \log \gamma) \right\},$$

which is minimized at $\gamma = 3\sqrt{K}\widetilde{M} n^{-1/2}$. Substituting in γ and $\widetilde{M}$ gives us

$$\widetilde{R} := \begin{cases} \dfrac{4n^{1/(2\tilde{d})} \left(-\frac{12\sqrt{K}}{\sqrt{n}} + 81^{1/d} d K^{2/d} n^{-2/d} \right)}{d-4} & : d \neq 4 \\ \dfrac{6\sqrt{K} n^{1/(2\tilde{d})} (2 + \log n - \log 9 - \log K)}{\sqrt{n}} & : d = 4, \end{cases}$$

so that $\mathcal{R}_n(\mathcal{F}_{\widetilde{M}}) \leqslant \widetilde{R}$.

Set

$$\varepsilon := 4\widetilde{L}_L \widetilde{R} + 7\widetilde{L}_{max} \sqrt{\frac{\log(8/\delta)}{2n}}.$$

Each term goes to zero, so $\lim_{n\to\infty} \varepsilon = 0$. Additionally, $\partial\varepsilon/\partial n = O\left(-n^{-\left(2\tilde{d}+1\right)/\left(2\tilde{d}\right)}\right)$. If n is sufficiently large, $\partial\varepsilon/\partial n < 0$, and ε is decreasing in n. Finally, applying Lemma 5 yields the first part of the theorem.

To strengthen the result to almost-sure convergence, we appeal to the Borel-Cantelli lemma. It is enough to show that

$$\sum_n \mathbb{P}\left[\sup_{f \in \mathcal{F}_{\widetilde{M}}} \left| R(f) - \widehat{R}(f) \right| > \varepsilon' \,\middle|\, \mathcal{E} \right] + \sum_n e^{-n^{2/\tilde{d}}/2\sigma^2} < \infty,$$

where $\varepsilon' > 0$ is an arbitrary, fixed value. The second series converges by comparison with the integral

$$\int_0^\infty e^{-n^{2/\tilde{d}}/2\sigma^2} dn = \left(2\sigma^2\right)^{\tilde{d}/2} \Gamma\left(1 + \tilde{d}/2\right).$$

Each term in the first series is bounded above by $\min\{1, \delta\}$, with δ satisfying $\varepsilon' = 4L_L \mathcal{R}_n(\mathcal{F}_{\widetilde{M}}) + 7L_{max}\sqrt{\log(8/\delta)/2n}$. For a given ε', a solution does not exist if n is too small. When n is large enough, we have the following:

$$\begin{aligned} \frac{\delta}{8} &= \exp\left\{ -2n \left(\frac{\varepsilon' - 4L_L \mathcal{R}_n(\mathcal{F}_{\widetilde{M}})}{7L_{max}} \right)^2 \right\} \\ &\leqslant \exp\left\{ -2n \left(\frac{\varepsilon' - 4\widetilde{L}_L \widetilde{R}}{7\widetilde{L}_{max}} \right)^2 \right\} \\ &:= \widetilde{\delta} \end{aligned}$$

The second line follows because $\widetilde{L}_L\widetilde{R} \to 0$ and $\partial\left(\widetilde{L}_L\widetilde{R}\right)/\partial n < 0$. Eventually, $0 < \widetilde{L}_L\widetilde{R} < \varepsilon'$ and $\left(\varepsilon' - 4\widetilde{L}_L\widetilde{R}\right)^2 \leqslant \left(\varepsilon' - 4L_L\mathcal{R}_n(\mathcal{F}_{\widetilde{M}})\right)^2$.

Asymptotically, $\log\widetilde{\delta} = O\left(-n^{1-4/\tilde{d}}\right)$. Furthermore, its derivative is $O\left(-n^{-4/\tilde{d}}\right)$, so $\widetilde{\delta}$ is decreasing for large n. Since

$$\int_0^\infty e^{-n^{1-4.1/\tilde{d}}} dn = \left(1 - 4.1/\widetilde{d}\right)^{-1} \Gamma\left\{\left(1 - 4.1/\widetilde{d}\right)^{-1}\right\},$$

the integral test shows the tail of $\sum_n \widetilde{\delta}$ is finite, proving

$$\sup_{f \in \mathcal{F}_{\widetilde{M}}} \left|R(f) - \widehat{R}(f)\right| \xrightarrow{a.s.} 0.$$

■

Finally, the following theorem establishes almost sure convergence of the empirical risk minimizer.

Theorem 8 *Let $X \sim P_X$, where P_X has density p on B such that $0 < c \leqslant \inf_x p$ for some constant c. Let f^* be the true regression function in that observations follow $S_{Y|X} \sim \mathcal{N}(f^*(X), \sigma^2)$. Suppose we set $\widetilde{M} := n^{1/(2\tilde{d})}$, where $\widetilde{d} := \max\{d, 5\}$, and let $\widehat{f} \in \mathcal{F}_{\widetilde{M}}$ be the empirical risk minimizer.*

(i) For $0 < \delta < 1$,

$$\mathbb{P}\left[\sup_{x \in B}\left|\widehat{f} - f^*\right| < \beta\right] > (1-\delta)(1 - e^{-n^{2/\max\{d,5\}}/2\sigma^2}),$$

where β is a monotonically-decreasing function of n for large enough n and $\lim_{n\to\infty}\beta = 0$.

(ii)

$$\sup_{x \in B}\left|\widehat{f} - f^*\right| \xrightarrow{a.s.} 0.$$

Proof We again condition on the event $\mathcal{E} := \left\{\max_{1\leqslant i\leqslant n}|\xi_i| \leqslant \sigma\sqrt{2\log 2n} + n^{1/\tilde{d}}\right\}$. However, we now set $\widetilde{M} := n^{1/(16\tilde{d}^2)}$. The conclusions of Theorem 7 still hold with the appropriate modifications made to any constants depending on $\widetilde{M}$. To relate the uniform risk bound to the difference between $\widehat{f}$ and f^*, we start by decomposing the risk. With probability at least $(1-\delta)(1 - e^{-n^{2/\tilde{d}}/2\sigma^2})$ over the sample,

$$\begin{aligned}
\mathbb{E}_X\left[\left(\widehat{f} - f^*\right)^2\right] &= R(\widehat{f}) - R(f^*) \\
&\leqslant \left|R(\widehat{f}) - \widehat{R}(\widehat{f})\right| + \left|\widehat{R}(f^*) - R(f^*)\right| \\
&\leqslant 2\sup_{f \in \mathcal{F}_{\widetilde{M}}}\left|R(f) - \widehat{R}(f)\right| \\
&< 2\varepsilon.
\end{aligned}$$

Combining this with Chebyshev's inequality, we have

$$\begin{aligned}\mathbb{P}_X\left[\left|\widehat{f}-f^*\right|>\alpha\right] &< \mathbb{E}_X\left[\left(\widehat{f}-f^*\right)^2\right]\Big/\alpha^2 \\ &< 2\varepsilon\alpha^{-2}\end{aligned}$$

for $\alpha > 0$. In other words, $\widehat{f}$ lies within a tube of radius α, except on a set $A \subset B$ such that $P_X(A) < 2\varepsilon\alpha^{-2}$.

Let $h := \sup_{x\in A}\left|\widehat{f}-f^*\right| - \alpha$. Because $\widehat{f}$ and f^* are Lipschitz, h is constrained by the inequality

$$\frac{(f^*(x)+M^*r+\alpha+h)-(f^*(x)+\alpha)}{r} \leqslant \widetilde{M},$$

where x is on the boundary of A, and r is the inradius of A. This implies that $h \leqslant \widetilde{M}r$. We can maximize this by taking A to be the $d-$dimensional ball of radius

$$\widetilde{r} := \left(\frac{2\varepsilon\alpha^{-2}\Gamma\left(1+\frac{d}{2}\right)}{c\pi^{d/2}}\right)^{1/d},$$

where c is a constant bounding the density p away from zero. This shows

$$\sup_{x\in B}\left|\widehat{f}-f^*\right| < \widetilde{M}\widetilde{r}+\alpha.$$

Set $\alpha := n^{-1/(10\tilde{d})}$. Then $\widetilde{M}\widetilde{r} = O(n^{\rho})$ and $\partial\left(\widetilde{M}\widetilde{r}\right)/\partial n = O(-n^{\rho-1})$, where

$$\rho := \begin{cases} -3/\left(50d\right)+1/400 & : d \leqslant 5 \\ \left(5-59d\right)/\left(80d^3\right) & : d > 5. \end{cases}$$

Now, defining $\beta := \widetilde{M}\widetilde{r}+\alpha$ gives the first part of the theorem.

Almost-sure convergence follows from a similar argument as part (ii) of Theorem 7. It suffices to show that, for arbitrary $\beta' > 0$,

$$\sum_n \mathbb{P}\left[\sup_{x\in B}\left|\widehat{f}-f^*\right| > \beta' \,\middle|\, \mathcal{E}\right] + \sum_n e^{-n^{2/\tilde{d}}/2\sigma^2} < \infty,$$

where we have already shown convergence of the second series. Let

$$\varepsilon' := \left(\frac{\beta'-\alpha}{\widetilde{M}}\right)^d \left(\frac{c\pi^{d/2}}{\Gamma\left(1+d/2\right)}\right)\left(\frac{\alpha^2}{2}\right)$$

be the solution when setting $\beta' = \widetilde{M}\widetilde{r}+\alpha$ and solving for ε. For fixed β' and large n, there is a corresponding $\varepsilon' > 0$. Note that ε' is not fixed, but decreasing in n. In fact,

$\varepsilon' = O\left(n^{-\left(5d+16\tilde{d}\right)/\left(80\tilde{d}^2\right)}\right)$. Since $\widetilde{L}_L\widetilde{R} = O\left(n^{-119/400}\right)$ for $d < 5$ and $O\left(n^{-(16d-1)/\left(16d^2\right)}\right)$ for $d \geqslant 5$, eventually $0 < \widetilde{L}_L\widetilde{R} < \varepsilon'$. Therefore,

$$\widetilde{\delta} := 8\exp\left\{-2n\left(\frac{\varepsilon' - 4\widetilde{L}_L\widetilde{R}}{7\widetilde{L}_{max}}\right)^2\right\}$$

is an upper bound for the tail of the first series. Observe that $\log\left(\widetilde{\delta}/8\right) = O\left(-n^{1-d/\left(8\tilde{d}^2\right)-22/\left(5\tilde{d}\right)}\right)$ and $\partial\left(\log\left(\widetilde{\delta}/8\right)\right)/\partial n = O\left(-n^{-d/\left(8\tilde{d}^2\right)-22/\left(5\tilde{d}\right)}\right)$. Comparison with the integral

$$\int_0^\infty e^{-n^{1-d/\left(8\tilde{d}^2\right)-22.1/\left(5\tilde{d}\right)}}dn = 1\Big/\left(1 - d/\left(8\tilde{d}^2\right) - 22.1/\left(5\tilde{d}\right)\right) \times \Gamma\left\{\left(1 - d/\left(8\tilde{d}^2\right) - 22.1/\left(5\tilde{d}\right)\right)^{-1}\right\}$$

is enough to give almost-sure uniform convergence of the empirical risk minimizer. ■

2.2 Vaidya's Algorithm

Given the value of M, it remains to solve the optimization problem (8). This is a convex program over a set that is not a polytope, and can be solved using interior point methods that respect a barrier constraint for the $C^{1,1}$ norm. Here we detail a solution by way of Vaidya's algorithm (Vaidya, 1996), making use of a slightly modification of an efficient implementation provided by Anstreicher (1997). Vaidya's algorithm is a cutting-plane method which seeks a feasible point in an arbitrary convex set $K \subset S_0 := \left\{x \in \mathbb{R}^k \;\middle|\; \|x\|_\infty \leqslant \rho\right\}$. Note that Anstreicher (1997) assumes $\rho = 1$. The set K is specified by a separation oracle: given a point $y \in \mathbb{R}^k$, the oracle either certifies that $y \in K$, or returns a separating hyperplane between y and K (i.e., a vector w such that $K \subset \left\{x \;\middle|\; w \cdot (x - y) \leqslant 0\right\}$). The algorithm initializes with an interior point $x_0 = 0$ and polytope S_0, and maintains a polytope $S_t \supset K$ and an interior point x_t of S_t at each iteration t, where S_t is defined via the separation oracle. At each iteration t, a constraint is either added or deleted, and the polytope S_t is specified by no more than $201k$ constraints throughout the algorithm. One of the strengths of Vaidya's algorithm is that it comes with complexity guarantees: that after $T = O(k(L + \log\rho))$ calls to the separation oracle, we have

$$\lambda(S_T) < \lambda(2^{-L}B^k),$$

where $\lambda(\cdot)$ denotes the k-dimensional volume and $L = \Omega(\log k)$ is a user-specified constant. Thus, the algorithm certifies that if no feasible point is found within T iterations, the volume of K is less than that of a k-dimensional ball of radius 2^{-L}. We remark that the value of T in our case (of $\rho \neq 1$ in general) is easily determined via an argument along the lines of Lemma 3.1 of Anstreicher (1997), and is given as

$$T \geqslant \frac{k\left[1.4L + 2\log k + 2\log(1 + 1/\epsilon) + 0.5\log\left(\frac{1+\tau}{1-\epsilon}\right) + 2\log(\rho) - \log(2)\right]}{\Delta V}, \tag{9}$$

where $\epsilon = 0.005$ and $\tau = .007$ are parameters of the algorithm, and $\Delta V = 0.00037$. The algorithm uses a total of $O(k(L + \log \rho)\xi + k^4(L + \log \rho))$ operations using standard linear algebra, where ξ is the cost of evaluating the separation oracle.

The feasibility algorithm may be applied to minimize an arbitrary convex function $g(\cdot)$ as follows. The minimization problem is essentially a feasibility problem in which we seek a point $\hat{x}$ in the set $K \cap \left\{x \;\middle|\; g(x) - g(x^\star) \leqslant \gamma\right\}$, where $\gamma > 0$ is an error tolerance and $x^\star$ is any minimizer of g on K. If we find a point $y \in K$, we instead use the oracle specified by any subgradient $w \in \partial g(y)$ to localize an optimal solution. If $0 \in \partial g(y)$, then y is an optimal point, and we are done. Otherwise, we use the hyperplane $\left\{x \;\middle|\; w \cdot (x - y) \leqslant 0\right\}$ within which the set $\left\{x \;\middle|\; g(x) \leqslant g(y)\right\}$ is contained, and proceed as in the feasibility case. If an optimal $x^\star$ was not found in T iterations, we find an approximate solution as follows. Let $\mathcal{T} \subset \{1, 2, \ldots, T\}$ denote the steps for which an $x_t \in K$ was found, after T iterations we return

$$\hat{x}_T \in \operatorname*{argmin}_{x_s,\, s \in \mathcal{T}} g(x_s). \tag{10}$$

Note that $g(x^\star)$ is not known, so we cannot directly evaluate whether any estimate $\hat{x}_T \in K$ satisfies the error tolerance. However, given information on the geometry of K and on the objective function, we may choose T to guarantee that this is the case. Fix $x^\star$ to be any optimal solution, and define $K_\epsilon(x^\star) := x^\star + \epsilon(K - x^\star)$ which contains the points in K in a small neighborhood around $x^\star$. Now let $x^\star_\epsilon$ denote the worst possible $x \in K_\epsilon(x^\star)$ in terms of having the largest value of g over all possible optimal solutions $x^\star$. Nemirovski (1995) defines an ϵ-solution to be any $x \in K$ such that

$$g(x) \leqslant g(x^\star_\epsilon),$$

and provides the following theorem.

Theorem 9 *Assume that after T steps the method has not terminated with an optimal solution. Then given that $\mathcal{T} \neq \varnothing$, any solution $\hat{x}_T$ of equation (10) is an ϵ-solution for any ϵ such that*

$$\epsilon^k > \frac{\lambda(S_T)}{\lambda(K)}.$$

If the function g is convex and continuous on K, then any ϵ-solution x satisfies

$$g(x) - g(x^\star) \leqslant \epsilon \left(\sup_{x \in K} g(x) - g(x^\star) \right).$$

Before finding the requisite number of iterations T, let us first derive the separation oracles, starting with the oracle for $K_1(M) := \left\{P \;\middle|\; \|P\|_{C^{1,1}} \leqslant M\right\}$. Presume at the current step of the algorithm, we have a set of function values and gradients $\{f(a), D_a f\}_{a \in E}$. which generate a candidate 1-field P. By Theorem 2 and equation (6), we may find the $a, b \in E, a \neq b$ such that $\Gamma^1(P; E) = \|P\|_{C^{1,1}(E)}$ in $n(n-1)/2$ operations. Thus, to determine whether the 1-field at the current step is contained in the constraint set, we simply check if $\|P\|_{C^{1,1}(E)} \leqslant M$. Otherwise, we must return a separating hyperplane in the space of 1-fields. Let $a^\star, b^\star \in E$, $a^\star \neq b^\star$ be any elements of E that solve (7), with $a^\star$ denoting the

first element of E in the numerator of (7). Specifying the separating hyperplane requires finding the $x \in \mathbb{R}^d$ that solves (7), that is, $x \in \mathbb{R}^d$ such that

$$\Gamma^1(P;E) = 2 \sup_{x \in B^d\left(\frac{a^\star+b^\star}{2}, \frac{|a^\star-b^\star|}{2}\right)} \frac{P_{a^\star}(x) - P_{b^\star}(x)}{|a^\star - x|^2 + |b^\star - x|^2}. \tag{11}$$

Equation (11) is a nonlinear fractional program, and is equivalent to minimizing the ratio

$$R(x) := \frac{N(x)}{D(x)}, \tag{12}$$

where $N(x) = |a^\star - x|^2 + |b^\star - x|^2$ and $D(x) = 2(P_{a^\star}(x) - P_{b^\star}(x)) > 0$. Here, we additionally know that the minimizer of (12) attains the optimal value $1/\Gamma^1(P;E)$ due to equation (6). Jagannathan (1966) and Dinkelbach (1967) showed that for $N(x)$ continuous, $D(x) > 0$ continuous, the solution to $\min_{x \in \mathcal{X}} R(x)$ over a compact subset $\mathcal{X} \subset \mathbb{R}^d$ is $z \in \mathcal{X}$ if and only if $z \in \mathcal{X}$ is also an optimal solution for

$$\min_{x \in \mathcal{X}} N(x) - R(z)D(x).$$

Plugging in the optimal value $R(y) = 1/\Gamma^1(P;E)$ yields the minimization

$$\min_x |a^\star - x|^2 + |b^\star - x|^2 - \left(\frac{2(P_{a^\star}(x) - P_{b^\star}(x))}{\Gamma^1(P;E)}\right).$$

Thus finding the x which solves (11) amounts to minimizing a convex quadratic in x. The solution is

$$z = \left(\frac{a^\star + b^\star}{2}\right) + \left(\frac{D_{a^\star} f - D_{b^\star} f}{2\Gamma^1(P;E)}\right).$$

The separation oracle for feasibility is thus specified as follows. For a candidate 1-field P, if $\|P\|_{C^{1,1}(E)} \leqslant M$, then certify that P is a feasible 1-field. Otherwise, separate all other 1-fields $\tilde{P} = \left\{\tilde{f}(a), \widetilde{D_a f}\right\}_{a \in E}$ from P via

$$\left\{\tilde{P} \;\middle|\; \frac{2(\tilde{P}_{a^\star}(z) - \tilde{P}_{b^\star}(z))}{|a^\star - z|^2 + |b^\star - z|^2} \leqslant \Gamma^1(P;E)\right\},$$

or equivalently

$$(\tilde{f}(a^\star) + \widetilde{D_a f} \cdot (z - a^\star)) - (\tilde{f}(b^\star) + \widetilde{D_{b^\star} f} \cdot (z - b^\star)) \leqslant \Gamma^1(P;E)\left(|a^\star - z|^2 + |b^\star - z|^2\right).$$

The separation oracle for $K_1(M)$ is thus equivalent to using the vector

$$w_{a^\star, b^\star} = \begin{pmatrix} 1 \\ z - a^\star \\ -1 \\ -(z - b^\star) \end{pmatrix} \in \mathbb{R}^{2(d+1)}$$

and the scalar $u_{a^\star,b^\star} = \Gamma^1(P;E)\left(|a^\star - z|^2 + |b^\star - z|^2\right)$ to define the hyperplane

$$\left\{v \in \mathbb{R}^{2(d+1)} \;\middle|\; w_{a^\star,b^\star} \cdot v \leqslant u_{a^\star,b^\star}\right\}. \tag{13}$$

Appropriately concatenating $w_{a^\star,b^\star}$ and v with zeros over the remaining possible choices of a, b thus defines a k-dimensional separating hyperplane, and this separating hyperplane is constructed in $O(n^2 + d)$ operations.

Note that the objective function in (8) is equivalent to

$$g(P) = \frac{1}{n}|y - f|^2.$$

To construct the separation oracle for $K_2(\gamma) := \left\{P \;\middle|\; g(P) - g(P^\star) \leqslant \gamma\right\}$, taking the gradient with respect to f, we have $w = \frac{2}{n}(f - y)$. Thus, given a current feasible field P with function values f, the separating hyperplane is specified as

$$\left\{\tilde{f} \in \mathbb{R}^n \;\middle|\; w \cdot \tilde{f} \leqslant u\right\}, \tag{14}$$

where $u = w \cdot f$. Suitably concatenating w and $\tilde{f}$ with zeros to form vectors in $\mathbb{R}^k$ thus specifies the separation oracle for $K_2(\gamma)$, and requires $O(n)$ operations to evaluate.

Now, to find the requisite number of iterations T to find a feasible point in $K_1(M) \cap K_2(\gamma)$, we sandwich the set $K_1(M)$ with Euclidean balls. The next result characterizes the Euclidean ball inside $K_1(M)$.

Lemma 10 *Assume $|a - b| \geqslant r > 0$ for all $a, b \in E$, $a \neq b$. Then*

$$\rho_1 B^k \subset K_1(M),$$

where

$$\rho_1 = \left(\frac{r^2 M}{8(1+r)}\right)\sqrt{n}.$$

Proof Let $P = \{f(a), D_a f\}_{a \in E}$ be any 1-field, and assume that

$$|P_a| = \sqrt{|f(a)|^2 + |D_a f|^2} \leqslant \rho$$

for all $a \in E$, so P is represented by a vector in $(\rho\sqrt{n})B^k$. Note that the numerator of (7) may be written as

$$P_a(x) - P_b(x) = (f(a) - f(b)) + \frac{1}{2}(D_a f + D_b f) \cdot (b - a) + (D_a f - D_b f) \cdot \left(x - \frac{a+b}{2}\right).$$

Thus,

$$\begin{aligned} |P_a(x) - P_b(x)| &\leqslant 2\rho + \rho|b - a| + 2\rho\left|x - \frac{a+b}{2}\right| \\ &\leqslant 2\rho(1 + |b - a|), \end{aligned}$$

where we have used the fact that $x \in B^d\left(\frac{a+b}{2}, \frac{|a-b|}{2}\right)$ in (7). The denominator is minimized at $x = \frac{a+b}{2}$ with minimal value $|a-b|^2/2$. Thus

$$\begin{aligned}\Gamma^1(P; E) &\leqslant \frac{8\rho(1+|b-a|)}{|b-a|^2} \\ &\leqslant 8\rho(r^{-1} + r^{-2}).\end{aligned}$$

It follows that $\|P\|_{C^{1,1}} \leqslant M$ if $\rho \leqslant \frac{Mr^2}{8(1+r)}$. ■

We now derive a bounding ball for $K_1(M)$ which indicates the value of ρ to use in Vaidya's algorithm.

Lemma 11 *Assume E is an $\epsilon-$net of B_d, the unit ball in $\mathbb{R}^d$, where $\epsilon < 1/10$. Suppose for all distinct $a, b \in E$, $|a-b| > r$. Then*

$$K_1(M) \subset \rho_2 B^k,$$

where

$$\rho_2 = \sqrt{n}\left[\frac{|y|^2}{n} + 4\left(\frac{10|y|}{r} + \frac{5M}{2}\right)^2\right]^{1/2}.$$

Proof Let $\prod_f$ denote the projection of any subspace of $\mathbb{R}^k$ onto the n-dimensional subspace corresponding to $D_a f = 0$ for all $a \in E$. If $y \in \prod_f K_1(M)$, then an optimal solution is given by $\{y(a), 0\}_{a\in E}$. Thus we may presume that $y \notin \prod_f K_1(M)$, whence

$$|f| \leqslant |y|.$$

To bound $|D_a f|$ given $a \in E$, note that by equation (7) we have

$$\frac{2|P_a(x) - P_b(x)|}{|a-x|^2 + |b-x|^2} \leqslant M$$

for any $b \in E\backslash\{a\}$ and $x \in B^d\left(\frac{a+b}{2}, \frac{|a-b|}{2}\right)$. Choosing $x = b$ implies that

$$\begin{aligned}|D_a f \cdot (b-a)| &\leqslant |f(a) - f(b)| + \frac{M}{2}|a-b|^2 \\ &\leqslant 2|y| + \frac{M}{2}|a-b|^2.\end{aligned}$$

By the conditions of the lemma, given a, there exist $b \in E\backslash\{a\}$ such that $|D_a f \cdot (b-a)| \geqslant \frac{|D_a f||(b_1-a)|}{10}$. It follows that

$$|D_a f| \leqslant \frac{20|y|}{r} + 5M.$$

Thus

$$\begin{aligned}|P|^2 &= |f|^2 + \sum_{a\in E} |D_a f|^2 \\ &\leqslant |y|^2 + 4n\left(\frac{10|y|}{r} + \frac{5M}{2}\right)^2.\end{aligned}$$

■

We arrive at the number of iterations required such that $g(P) - g(P^\star) \leqslant \gamma$.

Theorem 12 *Let $\gamma > 0$ be an error tolerance parameter, let $P^\star$ be any optimal solution to (8), and assume $0 < r \leqslant |a-b| \leqslant R$ for all $a, b \in E$. Applying Vaidya's algorithm for minimization as in (10) using the separation oracles specified in (13) and (14) yields an approximate solution $\hat{P}_T$ to (8) such that*

$$g(\hat{P}_T) - g(P^\star) \leqslant \gamma$$

where we choose

$$L \geqslant \log_2 \left(\frac{4|y|^2}{n\gamma\rho_1} \right),$$

with ρ_1 as stated in lemma 10 and T is given in equation (9) using $\rho = \rho_2$ from lemma 11.

Proof Recall that if $y \in \prod_f K_1(M)$, then an optimal 1-field is returned without calling Vaidya's algorithm via $\{y(a), 0\}_{a \in E}$. Thus we may presume that $|f| \leqslant |y|$ on $\prod_f K_1(M)$. Thus

$$g(P) = \frac{1}{n}|y - f|^2 \leqslant \frac{4|y|^2}{n}$$

for any $P \in K_1(M)$. Thus we set $\epsilon = \frac{n\gamma}{4|y|^2}$, and apply theorem 9. From lemma 10, we have that $\lambda(K_1(M)) \geqslant \rho_1^k \lambda(B^k)$. Since $\lambda(S_T) < 2^{-kL}\lambda(B^k)$, it suffices to choose L such that

$$2^{-kL}\rho_1^{-k} \leqslant \left(\frac{n\gamma}{4|y|^2} \right)^k,$$

from which the statement results. ■

2.3 Wells' Construction for $\hat{f}$ Given Lip($\nabla \hat{f}$)

Given $M = \mathrm{Lip}(\nabla \hat{f})$ and the gradients $\{D_a f\}_{a \in E}$, it remains to construct the interpolant $\hat{f} \in C^{1,1}(\mathbb{R}^d)$. We may now apply solution methods from the noiseless function interpolation problem. We summarize the solution provided by Wells et al. (1973) here.

Wells' construction takes $E \subset \mathbb{R}^d$, the 1-field $P : E \to \mathcal{P}$ consisting of function values $\{f(a)\}_{a \in E}$ and gradients $\{D_a f\}_{a \in E}$, and a value $M = \mathrm{Lip}(\nabla \hat{f})$ as inputs. A necessary condition for Wells' construction to hold is that

$$f(b) \leqslant f(a) + \frac{1}{2}(D_a f + D_b f) \cdot (b - a) + \frac{M}{4}|b-a|^2 - \frac{1}{4M}|D_a f - D_b f|^2, \quad \forall a, b \in E, \quad (15)$$

for which the optimal objective function value and gradients returned by the methods in Sections 1.2 and 2.2 satisfy.

For all $a \in E$, Wells defines the shifted points

$$\tilde{a} = a - \frac{D_a f}{M},$$

and associates a type of distance function for any $x \in \mathbb{R}^d$ to that point,

$$d_a(x) = f(a) - \frac{1}{2M}|D_a f|^2 + \frac{M}{4}|x - \tilde{a}|^2.$$

Using the shifted points, every subset $S \subset E$ is associated with several new sets:

$$\begin{aligned}
\widetilde{S} &= \left\{\tilde{a} \mid a \in S\right\}, \\
S_H &= \text{the smallest affine space containing } \widetilde{S}, \\
\widehat{S} &= \text{the convex hull of } \widetilde{S}, \\
S_E &= \left\{x \in \mathbb{R}^d \mid d_a(x) = d_b(x) \text{ for all } a, b \in S\right\}, \\
S_* &= \left\{x \in \mathbb{R}^d \mid d_a(x) = d_b(x) \leqslant d_c(x) \text{ for all } a, b \in S,\ c \in E\right\}, \\
S_C &= S_H \cap S_E.
\end{aligned}$$

Note that $S_H \perp S_E$, so S_C is a singleton. Wells next defines the collection of subsets

$$\mathcal{K} = \left\{S \subset E \mid \exists x \in S_* \text{ such that } d_S(x) < d_{E\setminus S}(x)\right\},$$

and a new collection of sets $\{T_S\}_{S\in\mathcal{K}}$, where

$$T_S = \frac{1}{2}(\widehat{S} + S_*) = \left\{\frac{1}{2}(y+z) \mid y \in \widehat{S},\ z \in S_*\right\}, \quad S \in \mathcal{K}. \tag{16}$$

The collection $\{T_S\}_{S\in\mathcal{K}}$ form a partition of $\mathbb{R}^d$ in the sense that overlapping sets have Lebesgue measure 0. On each set T_S, Wells defines a function $\hat{f}_S : T_S \to \mathbb{R}$ which is a local piece of the interpolating function $\hat{f}$:

$$\hat{f}_S(x) = d_S(S_C) + \frac{M}{2}\mathrm{dist}(x, S_H)^2 - \frac{M}{2}\mathrm{dist}(x, S_E)^2, \quad x \in T_S,\ S \in \mathcal{K}, \tag{17}$$

where as usual for sets $A, B \subset \mathbb{R}^d$, we have $\mathrm{dist}(A, B) = \inf_{x\in A,\, y\in B}|x - y|$. The final function $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ is then defined using (16) and (17):

$$\hat{f}(x) = \hat{f}_S(x), \quad \text{if } x \in T_S. \tag{18}$$

The gradient of $\hat{f}_S$ is

$$\nabla\hat{f}_S(x) = \frac{M}{2}(z - y), \quad \text{where } x = \frac{1}{2}(y+z),\ y \in \widehat{S}, z \in S_*.$$

The function $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ of (18) satisfies the following:

Theorem 13 (Wells' Construction) *Given a finite set $E \subset \mathbb{R}^d$, a 1-field $P : E \to \mathcal{P}$, and a constant M satisfying (15), the function $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ defined by (17) is in $C^{1,1}(\mathbb{R}^d)$ and satisfies*

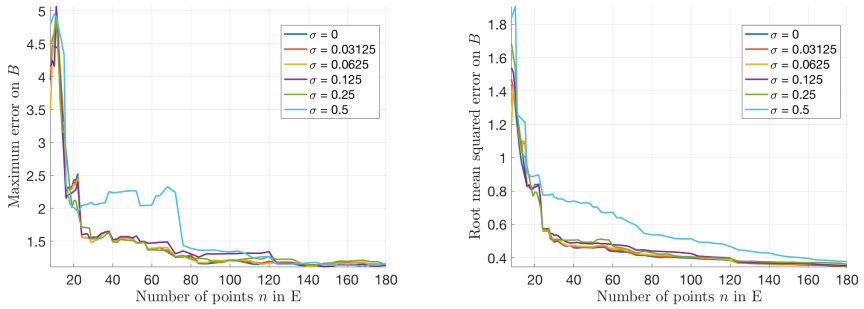
1. *$J_a\hat{f} = P_a$ for all $a \in E$.*

2. $Lip(\nabla \hat{f}) = M$.

Herbert-Voss et al. (2017) provide efficient algorithms to implement Wells' construction. We refer the reader to Herbert-Voss et al. (2017) for the details, but state briefly the computational cost of their methods. Let $m = |\mathcal{K}|$, and note that in the worst case $m = O(n^{\lceil d/2 \rceil})$. As stated by Herbert-Voss et al. (2017), a pessimistic bound on the storage as well as the number of computations for the one time work is $O(m^2)$. The query work is then also bounded by $O(m^2)$, however using more efficient querying algorithms to find the set $T_S \in \mathcal{K}$ to which a given $x \in \mathbb{R}^d$ belongs can lessen the work significantly. Using a tree structure, for example, will require $O(\log m) = O(\log n)$ work per query if the tree is balanced, but this need not be the case in general.

3. Simulation

We numerically compute the empirical risk minimizer $\widehat{f}$ over the functional class $\mathcal{F}_M = \{f \mid \|f\|_{C^{1,1}} = \mathrm{Lip}(\nabla f) \leqslant M\}$. To do so, we solve the optimization problem (8) for a 1-field P over the finite set E. This can be done efficiently with the algorithm described in Section 2.2, or using any constrained, convex optimization algorithm (such as interior point methods). The 1-field P is extended to a function in $\mathcal{F}_M$, which is $\widehat{f}$, using the algorithm described by Herbert-Voss et al. (2017).



(a) Maximum error

(b) Root mean square error

Figure 1: Generalization error as a function of $|E| = n$

The underlying function $f : \mathbb{R}^2 \to \mathbb{R}$ is taken as:

$$\forall\, x = (x_1, x_2) \in \mathbb{R}^2, \quad f(x_1, x_2) = \begin{cases} \cos(\pi x_1)\sin(\pi x_2)\exp\left(-\frac{1}{1-|x|^2}\right), & |x| < 1, \\ 0, & |x| \geqslant 1, \end{cases}$$

which is supported in the unit ball B. The training points E are sampled uniformly from B, and noisy function values $y(a) = f(a) + e_a$ for $a \in E$ are recorded, where e_a is i.i.d. Gaussian white noise with standard deviation σ. To approximate the generalization error between $\widehat{f}$ and f on B, the unit cube in which B is inscribed is sampled on a grid containing 2^{14} points (2^7 along each axis). As such, all errors over B described below are numerically approximated on the intersection of this grid with B.

The error $\sup_{x\in B} |f(x) - \widehat{f}(x)|$ between f and the empirical risk minimizer is plotted in Figure 1(a) as a function of n, for various values of the noise standard deviation σ. The value of M grows with n according to $M = O(1/n^{10})$, as in the proof of Theorem 7. Figure 1(b) plots the generalization error for the quadratic loss, i.e., $\left(\int_B |f(x) - \widehat{f}(x)|^2 \, dx\right)^{1/2}$. Both figures show that the generalization error generally decreases as n (and correspondingly M) increase.

Figure 2 gives a qualitative assessment of the empirical risk minimizer $\widehat{f}$ in the noiseless setting ($\sigma = 0$), by plotting the original function f and several versions of $\widehat{f}$ for selected values of n. As expected, the empirical risk minimizer $\widehat{f}$ visually appears to better match the underlying function f as n increases. Figure 3 fixes $n = 84$ and plots $\widehat{f}$ for increasing values of the noise standard deviation. While for large noise ($\sigma = 0.5$) the empirical risk minimizer $\widehat{f}$ deviates noticeably from f, for lower noise values ($\sigma \leqslant 0.25$), $\widehat{f}$ is a stable approximation of f.

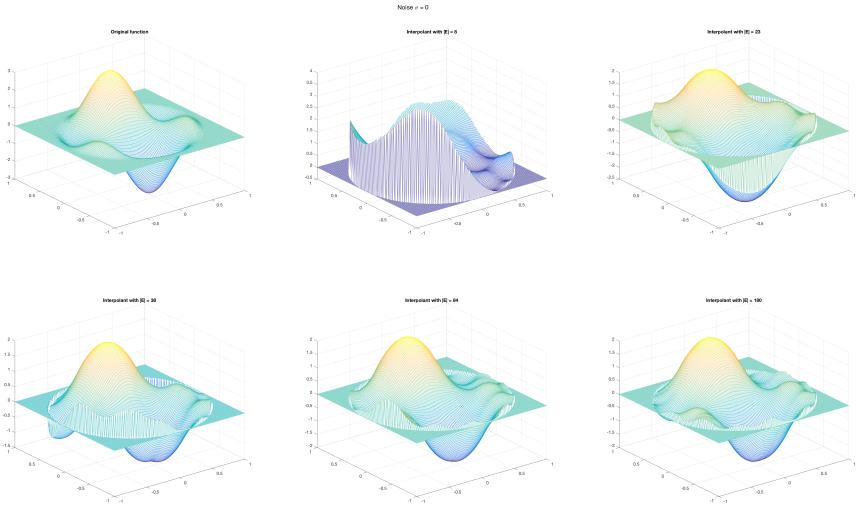


Figure 2: Upper left: The original function f. Subsequent plots, moving left to right and then to the second row: the empirical risk minimizer $\widehat{f}$ computed with $n = 8, 23, 38, 84, 180$ samples, respectively, and for $\sigma = 0$.

4. Discussion

In this paper, we extend the function interpolation problem considered by Herbert-Voss et al. (2017) to the regression setting, where function values $f(a)$ are observed with uncertainty over finite $a \in E$. We impose smoothness on the approximating function by considering regression solutions in the class of $C^{1,1}(\mathbb{R}^d)$ functions. Minimizing the risk over this function class is computationally tractable optimization problem, requiring $O((d+1)^2 n^2)$ calls to a separation oracle using Vaidya's algorithm. We present a separating hyperplane that

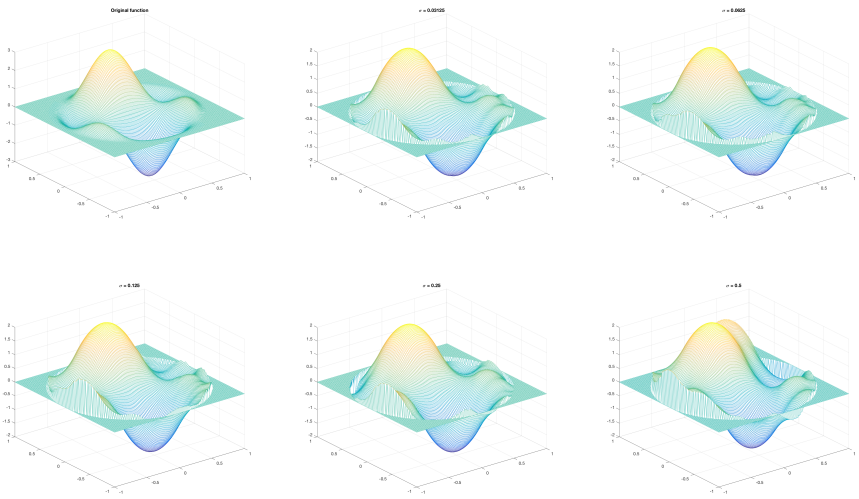


Figure 3: Upper left: The original function f. Subsequent plots, moving left to right and then to the second row: the empirical risk minimizer $\widehat{f}$ computed from noisy data with $\sigma = 2^{-j}$ for $j = 5, 4, 3, 2, 1$, respectively, and for $n = 84$ samples.

requires $O(n^2)$ operations, and given the output of Vaidya's algorithm, reconstruct the interpolant using efficient implementations of Wells' construction proposed by Herbert-Voss et al. (2017).

We derive uniform bounds relating the empirical risk of the regression solution to the true risk using empirical processes methods. The covering number of the class of $C^{1,1}(\mathbb{R}^d)$ functions is known and can be used to derive the covering number of Lipschitz loss classes. Our loss class is unbounded, but by conditioning on a suitable bound that increases with n, we obtain high probability bounds establishing that the algorithm is Probably-Approximately-Correct. As a consequence of the uniform risk bounds, almost sure convergence of the empirical risk minimizer is also guaranteed. These theoretical contributions are supported by numerical results via simulation.

References

Kurt M Anstreicher. On Vaidya's volumetric cutting plane method for convex programming. *Mathematics of Operations Research*, 22(1):63–89, 1997.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

Gleb Beliakov. Interpolation of Lipschitz functions. *Journal of Computational and Applied Mathematics*, 196:20–44, 2006.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pages 169–207. Springer, 2004.

Werner Dinkelbach. On nonlinear fractional programming. *Management Science*, 13(7): 492–498, 1967.

Charles Fefferman and Bo'az Klartag. Fitting a C^m-smooth function to data I. *Annals of Mathematics*, pages 315–346, 2009.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient regression in metric spaces via approximate Lipschitz extension. In Edwin Hancock and Marcello Pelillo, editors, *Similarity-Based Pattern Recognition, SIMBAD 2013*, volume 7953 of *Lecture Notes in Computer Science*, pages 43–58. Springer, Berlin, Heidelberg, 2013.

Ariel Herbert-Voss, Matthew J Hirn, and Frederick McCollum. Computing minimal interpolants in $C^{1,1}(\mathbb{R}^d)$. *Revista Matemática Iberoamericana*, 33(1):29–66, 2017.

R Jagannathan. On some properties of programming problems in parametric form pertaining to fractional programming. *Management Science*, 12(7):609–615, 1966.

Rasmus Kyng, Anup Rao, Sushant Sachdeva, and Daniel A. Spielman. Algorithms for Lipschitz learning on graphs. *JMLR: Workshop and Conference Proceedings*, 40:1–34, 2015.

Erwan Le Gruyer. Minimal Lipschitz extensions to differentiable functions defined on a Hilbert space. *Geometric and Functional Analysis*, 19(4):1101–1118, 2009.

Michel Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2005.

Colin McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141 (1):148–188, 1989.

Arkadi Nemirovski. *Information-Based Complexity of Convex Programming*. http://www2.isye.gatech.edu/~nemirovs/Lect_EMCO.pdf, 1995.

Karthik Sridharan and Nathan Srebro. Note on refined Dudley integral covering number bound, 2010.

Pravin M Vaidya. A new algorithm for minimizing convex functions over convex sets. *Mathematical Programming*, 73(3):291–341, 1996.

Aad W Van Der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, 2004.

John C Wells et al. Differentiable functions on Banach spaces with Lipschitz derivatives. *Journal of Differential Geometry*, 8(1):135–152, 1973.

Hassler Whitney. Analytic extensions of differentiable functions defined in closed sets. *Transactions of the American Mathematical Society*, 36(1):63–89, 1934.