A SIEVE STOCHASTIC GRADIENT DESCENT ESTIMATOR FOR ONLINE NONPARAMETRIC REGRESSION IN SOBOLEV ELLIPSOIDS

BY TIANYU ZHANG*, AND NOAH SIMON†

Department of Biostatistics, University of Washington. *zyy@uw.edu; †nrsimon@uw.edu

The goal of regression is to recover an unknown underlying function that best links a set of predictors to an outcome from noisy observations. In nonparametric regression, one assumes that the regression function belongs to a pre-specified infinite-dimensional function space (the hypothesis space). In the online setting, when the observations come in a stream, it is computationally-preferable to iteratively update an estimate rather than refitting an entire model repeatedly. Inspired by nonparametric sieve estimation and stochastic approximation methods, we propose a sieve stochastic gradient descent estimator (Sieve-SGD) when the hypothesis space is a Sobolev ellipsoid. We show that Sieve-SGD has rate-optimal mean squared error (MSE) under a set of simple and direct conditions. The proposed estimator can be constructed with a low computational (time and space) expense: We also formally show that Sieve-SGD requires almost minimal memory usage among all statistically rate-optimal estimators.

1. Introduction. It is commonly of interest to understand the association between a number of features (or predictors) and a quantitative outcome. To this end, one often estimates an underlying regression function that best links these two quantities from noisy observations. More formally, suppose we obtain \( n \) samples, \((X_i, Y_i)\), where \( X_i \in \mathcal{X} \subset \mathbb{R}^p \) denotes a \( p \)-vector of features from the \( i \)-th sample we observe, and \( Y_i \in \mathbb{R} \) denotes the \( i \)-th outcome. Further suppose that each pair \((X_i, Y_i)\) is independently and identically distributed (i.i.d.) from a fixed but unknown distribution \( \rho \) over \( \mathcal{X} \times \mathbb{R} \). A common target of estimation is the conditional mean \( f_\rho(X) := \mathbb{E}_\rho[Y|X] \). Under extremely mild conditions, this conditional mean is the optimal function for predicting \( Y \) from \( X \) with regard to mean squared error. More formally,

\[
 f_\rho = \arg\min_{f \in L^2_{\rho_X}} \mathbb{E}_\rho \left[ (Y - f(X))^2 \right],
\]

where \( L^2_{\rho_X} \) is the collection of all \( \rho_X \)-mean square integrable functions and \( \rho_X \) is the marginal distribution of \( X \). Our goal is to estimate \( f_\rho \) from our collection of observed data.

In order to make a tractable estimation of \( f_\rho \) from data, we need to make additional assumptions on its smoothness/structure: The entire \( L^2_{\rho_X} \) space is too big to search within \([4, 31]\). We often formally assume that \( f_\rho \) belongs to a pre-specified function space \( \mathcal{F} \subsetneq L^2_{\rho_X} \). This \( \mathcal{F} \) is known as the hypothesis space of the regression problem.

If \( \mathcal{F} \) can be indexed by a finite-dimensional parameter set \( \Theta \subset \mathbb{R}^d \), \( d \in \mathbb{N}^+ \), we refer to \( \mathcal{F} \) as a parametric function space or a parametric class. One common parametric class is \( \mathcal{F} = \{ X^\top \beta \mid \beta \in \mathbb{R}^d \} \), the set of all linear functions of \( X \). Parametric classes can impose overly restrictive assumptions on the form of the regression function that may not be realistic in practice. As such, it has become popular to assume less restrictive structure: It is common to define the hypothesis space based on constraints on derivatives, monotonicity, or other shape-related properties. Such an \( \mathcal{F} \) is most naturally written as an infinite-dimensional subset of \( L^2_{\rho_X} \). Commonly used examples of \( \mathcal{F} \) in the statistics community include Hölder balls,

---

MSC 2010 subject classifications: 62G08.

Keywords and phrases: Stochastic Gradient Descent, Nonparametric Regression, Online Learning.
Sobolev spaces [22, 37, 51], reproducing kernel Hilbert spaces (RKHS) [5, 11] and Besov spaces [24]. These are known as nonparametric function spaces, as they cannot naturally be parametrized using a finite length vector. The Sobolev ellipsoid, in particular, is a simple and useful abstraction of many important function spaces [51]. Therefore, we focus on them exclusively as the hypothesis spaces in this paper.

In this paper, we propose an estimator for online nonparametric regression. In online estimation, the data are seen sequentially, one sample at a time. After each sample is observed, our estimate of $f_\rho$ must be updated, as a prediction may be required at any point in time before all the available samples are processed. In an online problem with $n$ observations, we must sequentially construct $n$ estimates. This is in contrast to the classical batch learning setting where we collect all the data initially and perform estimation only once. In the online setting, it is generally computationally infeasible to repeatedly refit the whole model from scratch for each new observation. Thus, online algorithms are generally carefully developed to permit more tractable updates after each new observation [14, 30].

An ideal estimator in online settings should be: i) statistically rate-optimal, i.e. achieve the minimax-rate for estimating $f_\rho$ over $\mathcal{F}$; and ii) computationally inexpensive to construct/update. In this paper, we present such an online nonparametric estimator for use when the hypothesis space is a Sobolev ellipsoid, which we term the Sieve Stochastic Gradient Descent estimator (Sieve-SGD). This method can be thought of as an online version of the classical projection estimator [49], where the latter is a specific example of sieve estimators [21, 42]. We use the more general term “sieve” in naming our method to emphasize its non-parametric nature and avoid confusion with the term “stochastic projection” [50]. We will show that Sieve-SGD can achieve rate-optimal estimation error for $\mathcal{F}$ a Sobolev ellipsoid and asymptotically uses minimal memory (up to a log factor) among all rate-optimal estimators. In addition, our estimator has the same computational cost (up to a constant) as merely examining each allocated memory location every time a new sample $X_i$ is collected. This intimates that in scenarios when our estimator has near optimal space complexity, it may also have near optimal time complexity (though formal investigation of lower bounds for time complexity in this problem is beyond the scope of the current manuscript).

The structure of our paper continues as follows. In Section 2 we briefly cover classical results for batch, nonparametric estimation in Sobolev ellipsoids, focusing on projection estimators (which motivate our method). In Section 3 we return to the online setting and explore intuition for how one might combine projection estimation and stochastic gradient descent (SGD) [7]. The latter is a well-studied method that has been applied fruitfully to online parametric regression problems. This will help motivate our proposed method, which, as we will see, can be thought of as an SGD estimator with a parameter space of increasing dimension. In Section 4 we discuss existing nonparametric SGD estimators, and identify some notable drawbacks of current methods. In Section 5, we introduce the formal construction of Sieve-SGD and analyze its computational expense. From there, we show that our estimator has a dramatically smaller “dimension” than existing methods and discuss how this helps to reduce the computational expense. In Section 6, we give a theoretical analysis of the statistical properties of Sieve-SGD. In constructing our estimator, we need to decide how quickly to grow the dimension it projects onto. Under minimal assumptions, we characterize the required growth rate and learning rate for our estimator to be statistically and computationally (near) optimal. We will also investigate under what conditions such an optimality result is adaptive/insensitive to our choice of the “dimension-specific learning rate”. Section 7 provides simulation studies to illustrate our theoretical results. Finally, in Section 8, we have some further discussion of Sieve-SGD and possible future research directions.

**Notation:** In this paper, we use $C$ to denote a generic constant that does not depend on sample size $n$ (The value of $C$ may be different in different parts of the manuscript). Additionally the notation $a_n = \Theta(b_n)$ means $a_n = O(b_n)$ and $b_n = O(a_n)$. The function $\lfloor x \rfloor$ maps
$x$ to the largest integer smaller than $x$. For a vector $x \in \mathbb{R}^p$, $x^{(i)}$ is the $i$-th component of $x$. The notation $x \lor y$ (resp. $x \land y$) is shorthand for $\max\{x, y\}$ (resp. $\min\{x, y\}$). The $\| \cdot \|_\infty$ norm of a continuous function $f$ is defined as $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$, where $\mathcal{X}$ is the domain of $f$.

2. Batch Learning and the Projection Estimator. In this section we consider estimation in the classical batch setting where our estimate is constructed once after all $n$ samples are observed. We will begin by formally introducing a Sobolev ellipsoid: This is the hypothesis space we will use throughout this manuscript. This will be followed by presenting the classical projection estimator [49].

Consider a user-specified measure $\nu$ whose support contains $\mathcal{X}$, and the corresponding square-integrable function space $L^2_\nu$. In many interesting cases $\nu$ can be simply taken as Lebesgue measure over $\mathcal{X}$ but it is not necessary in the general form of our theory. To define a Sobolev ellipsoid in $L^2_\nu$, suppose we have a complete orthonormal basis $\{\psi_j, j = 1, 2, \ldots \} \subset L^2_\nu$. This means

i) For any $f \in L^2_\nu$, there exists a unique sequence $(\theta_j)_{j=1}^\infty \in \ell^2$ such that

$$\lim_{N \to \infty} \int \left| f(z) - \sum_{j=1}^N \theta_j \psi_j(z) \right|^2 d\nu(z) = 0 \quad \text{(completeness)}$$

where $\ell^2$ is the space of square convergent series.

ii) $\{\psi_j\}$ is an orthonormal system:

$$\int \psi_i(z) \psi_j(z) d\nu(z) = \delta_{ij} \quad \text{(orthonormality)}$$

where $\delta_{ij}$ is the Kronecker delta.

We define the Sobolev ellipsoid $W(s, Q, \{\psi_j\})$ as:

$$W(s, Q, \{\psi_j\}) := \left\{ f = \sum_{j=1}^\infty \theta_j \psi_j \left| \sum_{j=1}^\infty (\theta_j j^s)^2 \leq Q^2 \right\}$$

We refer to $(\theta_j)_{j=1}^\infty$ as the (general) Fourier coefficients of a function $f$. Throughout this manuscript, we assume the measure $\nu$, basis functions $\psi_j$ and the regularity parameter $s$ are all known. When it is clear which $\psi_j$ we are using, we will denote a Sobolev ellipsoid simply by $W(s, Q)$. We may also use the further simplified notation $W(s)$ because the diameter $Q$ usually plays a secondary role in our theoretical analysis and the proposed method is adaptive to it. Intuitively, by saying a function $f$ belongs to a Sobolev ellipsoid, we are requiring its coefficients $\{\theta_j\}$ to converge to zero faster than $j^{-(s+1/2)}$ (if not, the sum $\sum_{j=1}^\infty (\theta_j j^s)^2$ would diverge to infinity). The larger $s$ is, the faster the decay of $\theta_j$ will be, and thus the stronger our assumption is.

Sobolev ellipsoids are popular spaces to study for two reasons: 1) They impose a useful structure for theory and computations, especially as a basic example of hypothesis spaces with finite metric entropy; and 2) Many natural spaces of regular functions are Sobolev ellipsoids. For example, if $\mathcal{X} = [0, 1]$ with $\nu$ as Lebesgue measure, then for any $s > 0$, the periodic Sobolev space

$$\mathcal{F} = \left\{ f \in L^2_\nu \mid \int \left( f^{(s)}(x) \right)^2 dx < Q^2, f^{(k)}(0) = f^{(k)}(1), k = 0, 1, \ldots, s - 1 \right\}$$
can be written as a Sobolev ellipsoid, using an orthogonal basis of trigonometric functions [51, Chapter 2]. More generally, for many important RKHSs $\langle \mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}} \rangle$, it is possible to find a set of $\psi_j$ such that $\mathcal{W}(s, Q, \{\psi_j\}) = \{ f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq Q \}$, i.e. a ball in an RKHS is a Sobolev ellipsoid (see [12, 47]): Under mild conditions [44], a Mercer kernel $K(s, t) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ has the following Mercer representation:

$$
K(s, t) = \sum_{j \in J} \lambda_j \psi_j(s) \psi_j(t),
$$

where $\lambda_j > 0$, $J$ is at most countably infinite. And $\{\psi_j\}$ is an orthonormal system (in $L^2_\nu$) w.r.t. some measure $\nu$ on $\mathcal{X}$, and any function $f \in \mathcal{H}$ can be written as $f = \sum_{j \in J} \theta_j \psi_j$. It is also known that the RKHS-norm can be identified as $\|f\|^2_{\mathcal{H}} = \sum_{j \in J} \theta_j^2 \lambda_j^{-1}$. So a ball in the RKHS, i.e. $\{ f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq Q \}$, is the same as a Sobolev ellipsoid spanned by $\{\psi_j\}$ when $J = \mathbb{N}^+$ and $\lambda_j = j^{-2s}$. This is the case for many Sobolev-type kernels (for example, p.454 in [16]). When $J$ is finite dimensional (polynomial kernels) or $\lambda_j$ decays exponentially fast in $j$ (Gaussian kernel, p.455 in [16]), a ball in the RKHS can be characterized as some “generalized” Sobolev ellipsoid.

In everything that follows we will assume that $f_\rho$, our target of estimation, lives in a known Sobolev ellipsoid $\mathcal{W}(s, Q, \{\psi_j\})$; with $\{\psi_j\}$ specified, and orthonormal w.r.t. a specified measure $\nu$ (not necessarily equal to $\rho_X$); and $s$ known (we allow $Q$ to be unknown).

The Projection Estimator is a classical estimator naturally associated with a Sobolev ellipsoid. We can treat it as a special case of general sieve estimation [21, Chapter 10]: The estimates can be characterized by a sequence of finite dimensional linear spaces of increasing dimension (the dimension increases with sample size). For any given $f \in \mathcal{W}(s, Q)$, the magnitude of its Fourier coefficients must asymptotically decrease with $j$ fast enough. Thus, it might be sensible to consider an estimator that discards the basis functions far into the tail. This is precisely what the projection estimator does. More formally, for a user-specified truncation level $J_n$, the projection estimator is given by

$$
\hat{f}_{n, J_n} = \sum_{j=1}^{J_n} \hat{\theta}_j \psi_j
$$

where $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_{J_n})^\top$ is the solution of the least square problem:

$$
\min_{\theta \in \mathbb{R}^{J_n}} \sum_{i=1}^{n} \left( Y_i - \sum_{j=1}^{J_n} \theta_j \psi_j(X_i) \right)^2
$$

It has been shown (e.g. [49], Theorem 1.9) that when we choose $J_n = \Theta(n^{\frac{1}{2s+1}})$, the projection estimator is a rate-optimal estimator over $\mathcal{W}(s, Q)$, i.e.

$$
\limsup_{n \to \infty} \sup_{f_\rho \in \mathcal{W}(s, Q)} E \left[ \left\| \hat{f}_{n, J_n} - f_\rho \right\|_2^2 \right] = O(n^{-\frac{2s}{2s+1}})
$$

This result is usually shown in the literature for $X_i$ equally spaced, or drawn from a uniform distribution. But in our theoretical analysis (Section 6), we allow $\rho_X$ to be a much more general distribution.

Sieve-SGD is inspired by this (batch) projection estimator. The key here is that the number of basis functions we need to use can be dramatically smaller than the sample size, and their analytical forms do not depend on the data (usually reproducing kernel methods use basis functions “centered” at the feature vectors $X_i$). This possibility has been rarely explored [58] by existing nonparametric online estimation research.
3. Online Learning and Stochastic Approximation. We now move to the online learning setting where observations are collected sequentially from a data stream, and an estimate of our function is required after each sample. Such an infinite data stream may really exist, for example, with simulated samples as in reinforcement learning. Or the stream may serve as an abstraction used with large-scale data sets where it is not favorable to handle all the samples at once. It is generally computationally prohibitive to use a method developed for the “batch” setting and completely refit it after each observation. Instead methods that iteratively update at once. It is generally computationally prohibitive to use a method developed for the “batch” abstraction used with large-scale data sets where it is not favorable to handle all the samples example, with simulated samples as in reinforcement learning. Or the stream may serve as an of our function is required after each sample. Such an infinite data stream may really exist, for

\[ \sum_{i=1}^{n} \frac{1}{i^{2+x}} \]

requires computation of \( \Theta(n^{2+\frac{2}{x+1}}) \). Refitting a projection estimator (from scratch) after each observation \( i = 1, \ldots, n \) with \( J_i = \lfloor \frac{i}{2+x} \rfloor \) would require an accumulated computation of \( \sum_{i=1}^{n} i^{1+\frac{2}{x+1}} = \Theta(n^{2+\frac{2}{x+1}}) \). This scales worse than quadratically in \( n \). Our goal in the online nonparametric setting is to find a statistically rate-optimal estimator whose computation scales only slightly worse than linearly in \( n \).

Online learning has been thoroughly studied for parametric \( \mathcal{F} \). Many proposed methods are based on the concept of stochastic approximation [30]. One of the most popular methods in stochastic approximation is Stochastic Gradient Descent (SGD) [7]. In the parametric setting, SGD gives a statistically rate-optimal estimator \( \hat{f}_n \) whose population mean-square-error \( E\|\hat{f}_n - f_\rho\|^2_{L^2_X} \) is of order \( O(\frac{1}{n}) \) [2, 3, 17]. Both vanilla SGD and its variants have been applied to general convex loss functions and are shown to be statistically rate-optimal under mild conditions [14, 38].

3.1. Parametric SGD. To motivate stochastic optimization in the nonparametric setting, we first give more details on SGD for parametric classes. Here we consider a specific class of functions \( \mathcal{F} = \{ f = \sum_{j=1}^{d} \beta(j) \psi_j, \beta \in \mathbb{R}^d \} \) for a set of pre-specified basis functions \( \psi_j : \mathbb{R}^p \rightarrow \mathbb{R}, j = 1, \ldots, d \). We use this example to illustrate the principle of (parametric) SGD. Solving \( \arg\min_{f \in \mathcal{F}} E\left[ (Y - f(X))^2 \right] \) reduces to solving

\[ \min_{\beta \in \mathbb{R}^d} \ell(\beta) := \min_{\beta \in \mathbb{R}^d} E\left[ \left( Y - \sum_{j=1}^{d} \beta(j) \psi_j(X) \right)^2 \right] \]

We assume the minimizer of \( \ell(\beta) \) exists and denote it as \( \beta^* \).

If we knew the true joint distribution \( \rho \) of \( (X, Y) \) (which never happens in practice), then equation (10) is just a numerical optimization problem which does not involve data. We could use gradient descent to solve it. The gradient of \( \ell \) at any point \( \beta \) is

\[ \nabla \ell(\beta) = -2E\left[ \left( Y - \sum_{j=1}^{d} \beta(j) \psi_j(X) \right) (\psi_1(X), ..., \psi_d(X))^\top \right] \]

Thus, the gradient descent updating rule one could use is:

\[ \hat{\beta}_0 = 0 \]

\[ \hat{\beta}_n = \hat{\beta}_{n-1} - \gamma_n \nabla \ell(\hat{\beta}_{n-1}) \]

where \( \{ \gamma_n \} \) is a pre-specified sequence of step-sizes (or learning rate) and \( \hat{\beta}_n \in \mathbb{R}^d \) is the sequence of approximations of \( \beta^* \).

In practice, we do not know the joint distribution \( \rho \): we must use data to estimate \( \beta^* \). In the framework of SGD, this is done by using the data to get unbiased estimates of the gradients and substituting the estimates into our updating rule (12). In particular we note
that $\nabla \ell(\beta) := -2 \left( Y_i - \sum_{j=1}^{d} \beta^{(j)} \psi_j(X_i) \right) (\psi_1(X_i), ..., \psi_d(X_i))^\top$ is an unbiased estimator of the gradient $\nabla \ell(\beta)$ based on one sample. This results in the SGD updating rule.

$$\hat{\beta}_0 = 0$$

$$\hat{\beta}_n = \hat{\beta}_{n-1} - \gamma_n \nabla \ell(\hat{\beta}_{n-1})$$

$$= \hat{\beta}_{n-1} + 2\gamma_n \left( Y_n - \sum_{j=1}^{d} \hat{\beta}^{(j)}_{n-1} \psi_j(X_n) \right) (\psi_1(X_n), ..., \psi_d(X_n))^\top$$

So our estimator $\hat{f}_n$ of $f_\rho$ has the following functional update rule, derived from (13):

$$\hat{f}_n = \hat{f}_{n-1} + 2\gamma_n \left( Y_n - \hat{f}_{n-1}(X_n) \right) \sum_{j=1}^{d} \psi_j(X_n) \psi_j.$$  

Here we have shifted to considering our estimator $\hat{f}_n$ as a function, rather than thinking about $\hat{\beta}_n$ as a vector of coefficients. This will be important in the nonparametric setting.

3.2. From parametric SGD to nonparametric SGD. In this subsection we discuss the intuition in moving from SGD in a finite dimensional parametric space to an infinite dimensional space.

We assume $f_\rho \in W(s,Q,\{\psi_j\}) \subset L_\nu^2$. Since $\psi_j$ is a complete basis of $L_\nu^2$, we can always find an expansion of $f_\rho$ w.r.t. $\{\psi_j\}$:

$$f = \sum_{j=1}^{\infty} \theta_j \psi_j.$$  

In Subsection 3.1, we already discussed the SGD updating rule for a $d$-dimensional model $f(X) = \sum_{j=1}^{d} \beta^{(j)} \psi_j(X)$. In the nonparametric scenario, the number of basis function is increased from $d$ to infinity: This causes problems if care is not taken.

One might naturally consider applying a direct analog to the finite-dimensional SGD rule (14) here (we omit the constant 2):

$$\hat{f}_n = \hat{f}_{n-1} + \gamma_n \left( Y_n - \hat{f}_{n-1}(X_n) \right) \sum_{j=1}^{\infty} \psi_j(X_n) \psi_j.$$  

Unfortunately we run into a severe problem: The series $\sum_{j=1}^{\infty} \psi_j(X_n) \psi_j$ does not converge even if all $\psi_j$ are bounded (it is direct to check when $X_n = 0$ and $\psi_j$ are trigonometric functions). However, as we assume $f_\rho \in W(s)$, we know that those higher order components, $\psi_j$, $j \gg 1$ should have very small coefficients. Thus, one natural solution is to use a different step size per component, that decreases as $j$ increases. By doing “less fitting” for larger $j$, we can stabilize our update (smaller variance), and yet might still appropriately fit the overall regression function. In particular one might modify (16) to

$$\hat{f}_n = \hat{f}_{n-1} + \gamma_n \left( Y_n - \hat{f}_{n-1}(X_n) \right) \sum_{j=1}^{\infty} t_j \psi_j(X_n) \psi_j,$$

where the component-specific (or dimension-specific) learning rates $t_j > 0$ are monotonically decreasing with $j$. For $t_j$ decreasing fast enough and uniformly bounded $\psi_j$, the function series $\sum_{j=1}^{\infty} t_j \psi_j(X_n) \psi_j$ is absolutely convergent. Now (17) becomes a sensible nonparametric

SGD updating rule when the hypothesis space is a Sobolev ellipsoid. In addition, sometimes \( \sum_{j=1}^{\infty} t_j \psi_j(X_n) \psi_j \) actually has a simply characterized closed form (in particular, for many RKHS). In such cases, (17) results in a relatively straightforward algorithm. More specifically, one can show that when \( t_j = j^{-2s} \) and \( \gamma_n = \Theta(n^{-\frac{1}{2s+1}}) \), the average

\[
\hat{f}_n := \frac{1}{n} \sum_{i=1}^{n} \hat{f}_i
\]

is a rate-optimal estimator of \( f_\rho \in W(s) \). This was recently proposed (though motivated quite differently) in the context of RKHS hypothesis spaces [13]. The authors there engage directly with the kernel function for the RKHS (though their updating rule is equivalent to eq (17)). This will be discussed in more detail in Section 4. Our work engages and extends these ideas (in combination with sieve estimation) to form a statistically rate-optimal online estimator with greatly reduced computational and memory complexity.

4. Related work. Nonparametric online learning is a relatively new area. A few remarkable functional stochastic approximation algorithms have been proposed in the last two decades [9, 13, 34, 48, 55]. The key ideas in that body of work are intimately related to those mentioned in Section 3.2, however, they engage those ideas from a different direction: They assume that the hypothesis function space \( \mathcal{F} \) is an RKHS, and then leverage the kernel in that space. In particular, the RKHS structure makes it possible to take the gradient of the evaluation functional \( L_x(f) := f(x) \), with respect to the RKHS inner product \( \langle \cdot, \cdot \rangle_K \), i.e.

\[
L_x(f + \epsilon g) = f(x) + \epsilon g(x) = L_x(f) + \epsilon \langle g, K_x \rangle_K.
\]

Thus, \( K_x(\cdot) := K(x, \cdot) \in \mathcal{F} \) is the gradient of functional \( L_x \) at \( f \). However, one cannot do this in the general \( L^2_{\rhoX} \) space where the evaluation functional is no longer a bounded operator.

Thus when \( \mathcal{F} \) is an RKHS associated with kernel \( K \), there is a simple nonparametric SGD updating rule for minimizing \( E[(Y - f(X))^2] \) over \( \mathcal{F} \):

\[
\hat{f}_0 = 0
\]

\[
\hat{f}_n = \hat{f}_{n-1} + \gamma_n \left( Y_n - \hat{f}_{n-1}(X_n) \right) K(X_n, \cdot)
\]

Here, because the gradient is taken with respect to the RKHS inner product, we do not have the issue encountered in (16) where our series representation of the “gradient” actually did not converge. In fact, by working with the RKHS inner-product, we implicitly carry out the proposal of Section 3.2 and decrease the component-specific learning rate of higher order terms. More specifically, we usually have the Mercer expansion of the kernel function:

\[
K(x, z) = \sum_{j=1}^{\infty} t_j \psi_j(x) \psi_j(z),
\]

with respect to an orthonormal basis \( \{\psi_j\} \) of \( L^2_\nu \). For many common RKHS, we have \( t_j = \Theta(j^{-u}) \) for some \( u > 1 \) [16, Appendix A]. Thus, (20) corresponds precisely to the previously discussed update (17). Most popular RKHS have a kernel \( K(x, z) \) with a closed form representation, and thus, rather than having to store an infinite number of coefficients, after \( n \) steps the estimate from (20) would take the form of a weighted linear combination of \( n \) kernel functions [13]:

\[
\hat{f}_n = \sum_{i=1}^{n} b_i K(X_i, \cdot).
\]
Although such estimators (with one more Polyak averaging step (18)) have been shown to give rate-optimal MSE [13], updating them with a new observation \((X_{n+1}, Y_{n+1})\) usually involves evaluating \(n\) kernel functions at \(X_{n+1}\), with computational expense of order \(\Theta(n)\). This is in contrast with the constant update cost of \(\Theta(d)\) in parametric SGD, where \(d\) is the dimension of the parameter space. Thus, the time expense of nonparametric kernel SGD will accumulate at order \(\Theta(n^2)\). Also, one is required to store the \(n\) feature-values \(\{X_i\}_{i=1}^n\) to evaluate the estimator which results in \(\Theta(n)\) space expense. This relatively large time and space complexity indicates that those kernel-based SGD estimators are not ideal as methods that are nominally designed to deal with large data sets.

There has been some work in the literature aimed at improving the computational aspects of kernel SGD methods [43, 32, 29]. These methods select a subset of the \(n\) kernel functions centered at the feature vectors and use them as basis functions to construct estimators (which is also related to Nyström projection). Neither the statistical performance nor the computational expense of the aforementioned work is guaranteed to be optimal. Also, the theoretical analysis in that work typically requires the noise variable to have extremely light tails.

There has also been recent work [9, 34] aimed at improving kernel SGD algorithms by leveraging approximate second order information (SGD only uses the first order information). The estimator in [34] is shown to give rate-optimal MSE and have better (theoretical) computational efficiency than the vanilla kernel SGD mentioned above. However, these algorithms are usually dramatically more complicated and have a couple of hyper parameters that need to be tuned.

There is another branch of research also called “online nonparametric regression” that engages with a different but related setting [18, 40]. They do not aim to directly minimize the (population) generalization error. Their definition of “regret” is based on comparing a running average of prediction error and the empirical risk minimizer’s training error. Formally, it is defined as \(\sum_{i=1}^n l(\hat{Y}_i, Y_i) - \inf_{f \in \mathcal{F}} \sum_{i=1}^n l(f(X_i), Y_i)\), where \(\hat{Y}_i\) is the prediction of the algorithm based on the first \(i - 1\) observations, \(l\) is a convex loss and \(\mathcal{F}\) is the hypothesis function space. While this is an interesting area of research, and might be used to engage with population generalization error (using online-to-batch techniques), it is a less direct treatment than what we are considering in this work.

5. Online Learning and the Projection Estimator: Sieve-SGD. In this section, we combine ideas from the projection estimator (in the batch learning setting), and stochastic gradient descent to develop an estimator that is suitable for online nonparametric regression. The estimator we will propose achieves the minimax rate for MSE over a Sobolev ellipsoid, and is much more computationally efficient than standard kernel SGD methods.

As a reminder, the kernel SGD estimator based on (20) has minimax rate optimal MSE. When \(\sum_{j=1}^\infty t_j \psi_i(s) \psi_j(t)\) has an available closed form, it requires \(\Theta(n)\) memory and has \(\Theta(n^2)\) time expense for sequentially processing \(n\) observations. We aim to improve this and furthermore to propose an effective estimator appropriate for cases where \(\sum_{j=1}^\infty t_j \psi_i(s) \psi_j(t)\) has no closed form.

Motivated by the projection estimator, we opt to use truncated series in the updating rule, modifying (17) (or equivalently (20)) to get

\[
\hat{f}_n = \hat{f}_{n-1} + \gamma_n \left( Y_n - \hat{f}_{n-1}(X_n) \right) \sum_{j=1}^{J_n} t_j \psi_j(X_n) \psi_j
\]

Here \(J_n\) is an increasing sequence of integers that grows as we collect more observations. When \(J_n\) is larger, the updating rule (23) is closer to our original form (17); however, a
smaller \( J_n \) is desirable because it results in a lower computational expense. Part of our task is identifying a “minimal” \( J_n \) that still maintains favorable statistical properties.

It turns out there are 2 ways to control the bias-variance tradeoff. One can use the truncation level \( J_n \), or the component specific step sizes \( t_j \). If the truncation level is used, then the methodology is more analogous to a projection estimator. In this case, so long as \( t_j \) is not too large (controlling the variance in the dynamics of SGD) or too small (controlling the bias term), we would get (near) optimal statistical performance for a relatively wide range of choices for \( t_j \). We give formal results related to this in Section 6.3. If, instead, we control the bias-variance tradeoff using \( t_j \), then our estimator is more analogous to kernel-SGD. In this case, the first order terms for bias and variance are determined by the sequence \( \{ t_j \} \) and \( J_n \) just needs to be sufficiently large (such that we do not induce excess bias). We give formal results for this in Section 6.2. This second way to control the tradeoff is similar to using a truncated basis for penalized regression in the batch learning setting. For example, in [23] and [54, Section 5.2], the authors propose to estimate \( f_\rho \) by solving a penalized regression spline problem, where they use a reduced spline basis for improved computation (rather than including a knot at every point). The bias/variance trade-off there is controlled via the penalty: They are careful to include enough basis elements so that the use of a reduced basis only contributes a second order term to the bias.

We will next give details of our proposal. For this proposal we are assuming that \( f_\rho \in W(s,Q,\{\psi_j\}) \subset L^2_\nu \), and that \( s \) is known. Based on this, we choose our component-specific step-sizes as \( t_j = j^{-2\omega} \) (for some \( 1/2 < \omega \leq s \)). We also define

\[
K^\omega_{x,J_n}(\cdot) = \sum_{j=1}^{J_n} j^{-2\omega} \psi_j(x) \psi_j(\cdot).
\]

In addition to simplifying exposition, this notation relates our method to (21). The function \( K^\omega_{x,J_n}(\cdot) \) can be seen as a truncated approximation of the kernel function

\[
K^\omega_{x,\infty}(\cdot) = \sum_{j=1}^{\infty} j^{-2\omega} \psi_j(x) \psi_j(\cdot)
\]

that drops all the \( \psi_j \) with index \( j > J_n \).

5.1. Sieve Stochastic Gradient Descent. We now explicitly give our Sieve Stochastic Gradient Descent algorithm (Sieve-SGD) for estimation of \( f_\rho \) in a Sobolev ellipsoid \( W(s,Q,\{\psi_j\}) \).

Let \( J_n = [n^\alpha] \) for some specified \( \alpha > 0 \) and \( \omega \in (\frac{1}{2}, s] \). The parameter \( \alpha \) is usually taken between \( \frac{1}{2\nu+1} \) and 1. We use \( \gamma_i \) to denote the step size (learning rate) of the \( i \)-th update and typically choose \( \gamma_i = \Theta(i^{-\frac{1}{2\nu+1}}) \).

Proposed Algorithm: Sieve Stochastic Gradient Descent (Sieve-SGD)

Set \( \alpha, \omega > 0 \), step size \( \{\gamma_i\} \) and basis functions \( \{\psi_j\} \). Initialize \( \tilde{f}_0 = \hat{f}_0 = 0 \).

For \( i = 1, 2, \ldots \) :

1. Calculate \( J_i = [i^\alpha] \), collect data pair \( (X_i,Y_i) \).
2. Update \( \tilde{f}_i \):

\[
\tilde{f}_i = \tilde{f}_{i-1} + \gamma_i \left( Y_i - \tilde{f}_{i-1}(X_i) \right) \sum_{j=1}^{J_i} j^{-2\omega} \psi_j(X_i) \psi_j
\]

\[
= \tilde{f}_{i-1} + \gamma_i \left( Y_i - \tilde{f}_{i-1}(X_i) \right) K^\omega_{X_i,J_i}
\]
3. Polyak averaging: Update $\tilde{f}_i$ by

$$
\tilde{f}_i = \frac{1}{i+1} \sum_{k=0}^{i} \hat{f}_k
$$

(27)

$$
= \frac{i}{i+1} \tilde{f}_{i-1} + \frac{1}{i+1} \hat{f}_i
$$

We refer to the function $\tilde{f}_i$ as the Sieve-SGD estimate of $f_\rho$. We will later show that $\tilde{f}_i$ has rate-optimal MSE for estimating any $f_\rho \in W(s)$. Here we use the language of “updating a function”, but in practice one would update the coefficient vector corresponding to the functions $\{\psi_j\}_{j=1}^{J_n}$. In Appendix A we attach a presentation of the algorithm that works directly with the coefficients. This estimator is quite simple, though it does require apriori selection/knowledge of $\{\psi_j\}$ and $s$ (which can be done using a held-out validation set in practice). Unfortunately showing its favorable statistical properties (in Section 6) is somewhat more complex!

5.2. Computational expense. After examining the updating rule above, one can see that $\hat{f}_i$ has the form:

$$
\hat{f}_i(x) = \sum_{j=1}^{J_n} b_j \psi_j(x)
$$

(28)

This requires storing the coefficients $\{b_j\}_{j=1}^{J_n}$ in memory. The main computational burden of each update step is calculating $\hat{f}_{i-1}(X_i)$ and $K_{\hat{f}_{i-1},J_i}^\psi$. Both require evaluating $J_i$ basis functions at $X_i$. Thus, the computational expense of the “Update $\hat{f}_i$” step above is of order $J_i = \Theta(i^\alpha)$, when we take evaluating one basis function at one point as $O(1)$. And the total expense of processing $n$ samples is of order $\Theta(n^{1+\alpha})$. The space expense is of the same order $\Theta(i^\alpha)$. We need only store coefficients of $J_i$ basis functions. In Section 6.4 we will show that, under mild conditions, this memory complexity is near optimal among all estimators with rate-optimal MSE.

This compares favorably with standard kernel SGD (22) which uses $i$ basis functions at step $i$: Our estimator uses fewer when $\alpha < 1$; as we will show later, $\alpha$ can be taken as small as $1/2s+1$ which is a substantial improvement. In practice, the parameter $\alpha$ can either be selected based on our assumptions about $s$ (belief on the smoothness of $f_\rho$) or heuristically tuned for empirical performance.

5.3. General Convex loss. Although the main focus of this paper is regression with squared-error loss, our algorithm has a straightforward extension to general convex loss. Suppose we want to minimize the population loss

$$
E[\ell(Y, f(X))]
$$

(29)

over all functions $f \in W(s, Q, \{\psi_j\})$ and the loss function $\ell(Y, \cdot)$ is convex for each $Y$. In this case, we need only modify step 2 of the Sieve-SGD estimator in Section 5.1. Given loss $\ell(\cdot, \cdot)$, the updating rule for $\hat{f}_i$ takes the general form:

2’ Update $\hat{f}_i$:

$$
\hat{f}_i = \hat{f}_{i-1} + \gamma_i \frac{\partial}{\partial v} \ell(u, v) \bigg|_{(Y_i, \hat{f}_{i-1}(X_i))} K_{\hat{f}_{i-1},J_i}^\psi
$$

(30)
For example, with $Y = \{1, -1\}$ considering nonparametric logistic regression, the loss function one would use is
\[ \ell(Y, f(X)) = \log(1 + \exp(-Yf(X))). \]
In this case, we have
\[
\left(1 + \exp(Y_if_i(X_i))\right)^{-1}Y_i \in \mathbb{R}
\]

Theoretical guarantees for Sieve-SGD using general convex loss are beyond the scope of this paper. However, in Section 7 we provide simulated experiments that show the empirical performance of Sieve-SGD for nonparametric logistic regression. These empirical results intimate that perhaps similar theoretical guarantees to those shown for squared-error-loss hold in a more general setting.

5.4. Choice of Basis Functions & Multivariate Problems. In practice, there are many available choices of univariate $\psi_j$ that in general lead to interesting (Sobolev-type) hypothesis spaces. For example,
\[
\psi_1(x) = 1, \quad \psi_j = \sqrt{2} \cos((j-1)\pi x), \quad \text{for} \ j \geq 2.
\]
This set of basis functions are the “eigenfunctions” of Sobolev spaces over $[0, 1]$ (Appendix A.2 in [39]), which means they are orthogonal w.r.t to the Lebesgue inner product and the Sobolev inner product simultaneously. The corresponding Sobolev ellipsoid does not impose periodicity assumptions of $f_\rho$, and is very convenient to use in practice. Among many other choices, we can also use algebraic polynomials, or a combination of algebraic polynomial and (periodic) Fourier basis [15].

In most applications, the covariate $X_i$’s take value in $\mathbb{R}^p$ where $p > 1$. In some situations, there are some “canonical” choices of basis function $\psi(x) : \mathbb{R}^p \rightarrow \mathbb{R}$ that people might use for identifying their (multivariate) Sobolev ellipsoid. For example, when considering estimating a function on a sphere $S^2$, $\psi_j$ could be taken as the orthonormal spherical harmonics ([27], [36]).

In many situations, the basis functions $\psi_j$ can conveniently be taken as a tensor product of a one-dimensional complete basis, and Sieve-SGD can be directly applied in this multivariate setting. If we are using a univariate Sobolev ellipsoid to represent a ball in an RKHS, then the ellipsoid defined by the tensor product basis will correspond to a ball in the RKHS spanned by the tensor product kernel (though care needs to be taken with the ordering of the basis vectors). Some technical details and numerical examples on this can be found in Appendix B and the reference therein. In all of these cases, our theoretical results will hold (so long as the function $f_\rho$ belongs to the specified space).

A common alternative approach in multivariate problems is to impose some additional structure on the hypothesis space to make estimation more tractable. This is particularly true when the feature dimension $p$ is large. One popular model is the nonparametric additive model [46, 25, 57], which is thought to effectively balance model flexibility and interpretability. For $x \in \mathbb{R}^p$, we might consider assuming/imposing an additive structure on the regression function:
\[
f_\rho(x) = \sum_{k=1}^{p} f_{\rho,k}(x^{(k)})
\]
where each of the component functions $f_{\rho,k}$ belong to a Sobolev ellipsoid $W_k(s_k, Q_k, \{\psi_{jk}\})$.
For ease of exposition, in (33), we assume $E[Y] = 0$ to avoid the need for a common intercept term. For a more complete version with common intercept, see Appendix B. For a fixed dimension $p$, when all $W_k = W^*$ (for some Sobolev ellipsoid $W^*$), the minimax rate for estimating such an additive model is identical (up to a multiplicative constant $p$) to the
minimax rate in the analogous one-dimension nonparametric regression problem with the same hypothesis space $W^*$ [41, 46]. For the additive model (33), the updating rule (26) of Sieve-SGD could be replaced by:

$$
\hat{f}_i = \hat{f}_{i-1} + \gamma_i \left( Y_i - \sum_{k=1}^{p} \hat{f}_{i-1,k} \left( X_i^{(k)} \right) \right) \sum_{k=1}^{p} \sum_{j=1}^{J_{ik}} j^{-2\omega_k} \psi_{jk} \left( X_i^{(k)} \right) \psi_{jk}
$$



here $J_{ik}$ is the truncation level of $k$-th dimension when the sample size $= i$ and $\hat{f}_{i-1,k}$ is the estimate of $f_{\rho,k}$. Most of the theory that we develop in Section 6 could apply here.

6. Generalization Guarantees of Sieve-SGD. In this section, we show Sieve-SGD achieves the minimax rate for nonparametric estimation in Sobolev ellipsoids under mild assumptions. We also show that Sieve-SGD has near minimal memory complexity among all estimators that are minimax rate-optimal for estimation in a Sobolev ellipsoid. The conditions on the hyperparameters can be used as theoretical guidance when applying Sieve-SGD to real data problems.

6.1. Model Assumptions. We begin by listing the conditions we will require in our proof. They reflect different aspects of the problem: independent observations (A1), distribution of $X$ (A2), the hypothesis space assumed for $f_\rho$ (A3) and tail behaviour of the noise (A4). These conditions ensure the MSE rate-optimality of Sieve-SGD.

A1 (i.i.d. data) The data points $(X_n, Y_n)_{n \in \mathbb{N}} \in \mathcal{X} \times \mathbb{R}$ are independently, identically sampled from a distribution $\rho(X, Y)$.

A2 (feature distribution) Let $\nu$ be a user-specified measure that is strictly positive on $\mathcal{X}$. Assume the distribution of feature $X$, $\rho_X$, is absolutely continuous w.r.t. $\nu$. Let $p_X = \frac{d\rho_X}{d\nu}$ denote its Radon–Nikodym derivative. We assume for some $u, \ell$ such that $0 < \ell < u < \infty$:

$$\ell \leq p_X(x) \leq u \quad \text{for all } x \in \mathcal{X}$$

A3 (Sobolev ellipsoid) Let $\{\psi_j\}_{j=1}^{\infty}$ be a set of uniformly bounded ($\|\psi_j\|_\infty \leq M$), continuous, orthonormal basis of $L^2_\nu$. We assume the regression function $f_\rho$ falls in a Sobolev ellipsoid, with basis functions given by $\{\psi_j\}$, i.e. for some $s > \frac{1}{2}, Q < \infty$,

$$f_\rho \in W(s, Q, \{\psi_j\})$$

A4 (noise) One of the following two assumptions is satisfied by the noise variable $\epsilon = Y - f_\rho(X)$:

- $\epsilon$ is bounded by some $C_\epsilon$ almost surely.
- $\epsilon$ is independent of the features, $X$, and has a finite second moment $E[\epsilon^2] = C_\epsilon^2$.

Note 1: The lower bound requirement of $p_X$ in A2 may be due to artifacts in our proof. In reality, especially when the dimension of our feature-space $X$ is large, such an requirement may be hard to satisfy. According to our simulation results, Sieve-SGD still achieves the minimax rate even when $p_X$ has a strictly smaller support than $\nu$. As compared with other work in nonparametric online learning [13, 48, 55], our assumptions are more direct. We discuss this in detail later in this section.

Note 2: In assumption A3, we do not require $\psi_j$ to be orthonormal w.r.t. $\rho_X$ (and it is in general not true), but only require them to be orthonormal w.r.t. the known measure $\nu$. In many cases $\nu$ might be taken to be Lesbesgue (or uniform) measure over a domain containing $\mathcal{X}$, as this is the canonical measure under which function spaces such as Sobolev spaces and Besov spaces are defined. As long as the density function $p_X$ satisfies A2, using the
Also set the overall learning rate to be $\gamma$ (or $\omega$) favorable statistical and computational properties. Here is the main result in this setting: the “incorrect” version very soon in Section 6.3, and show that it can actually still have a-

6.2. Rate Optimality when $t_j = j^{-2s}$. In this section, we present the rate-optimality results of Sieve-SGD when we choose the component-specific learning rate to be $t_j = j^{-2s}$ (or $\omega = s$ in (26)). In this setting, our theoretical analysis treats Sieve-SGD as a truncated-version (in the basis expansion domain) of a “correct” kernel SGD procedure (we will discuss the “incorrect” version very soon in Section 6.3, and show that it can actually still have favorable statistical and computational properties). Here is the main result in this setting:

**Theorem 6.1.** Assume A1-A4. Set the component-specific learning rate as $t_j = j^{-2s}$. Also set the overall learning rate to be $\gamma_n = \gamma_0 n^{-\frac{1}{2s+1}}$ with $\gamma_0 \leq (2M^2\zeta(2s))^{-1}$, where $\zeta(k) = \sum_{i=1}^{\infty} i^{-k}$. Choose the number of basis functions to be $J_n \geq n^s \log^2 n \vee 1$ for an arbitrary $\alpha \geq \frac{1}{2s+1}$.

Then the MSE of Sieve-SGD (27) converges at the following rate

$$E\|\hat{f}_n - f_\rho\|_{L^2_X}^2 = O\left(n^{-\frac{2s}{2s+\alpha}}\right).$$

This implies that Sieve-SGD is a minimax rate-optimal estimator of $f_\rho$ over $W(s, Q, \{\psi_j\})$.

We now discuss our assumptions and results in more detail, and relate them to what is currently in the literature.

**Note 1:** In the analysis of many reproducing kernel methods for nonparametric estimation [13, 48, 56], the spectrum of the covariance operator plays an important role in controlling the statistical behavior of estimators. It is conventional in the community to make assumptions associated with this spectrum, which we find less natural than our related assumptions A2 and A3. The covariance operator is an analog of the covariance matrix in infinite dimensional spaces. For our problem setting, one natural covariance operator $T_X$ is defined as:

$$T_X : L^2_{p_X} \rightarrow L^2_{p_X},$$

$$g \mapsto \int_X g(\tau) \left( \sum_{j=1}^{\infty} j^{-2s} \psi_j(\tau) \psi_j \right) d\rho_X (\tau).$$

(37)

A direct analysis of the spectrum of $T_X$ is hard. However, there is a simpler operator that we have in hand which we can relate $T_X$ to:

$$T_\nu : L^2_{p_\nu} \rightarrow L^2_{p_\nu},$$

$$g \mapsto \int_X g(\tau) \left( \sum_{j=1}^{\infty} j^{-2s} \psi_j(\tau) \psi_j \right) d\nu (\tau).$$

(38)

We know the eigensystem of $T_\nu$: It is exactly $(j^{-2s}, \psi_j)$ (eigenvalue, eigenfunction). It is direct to check because $\{\psi_j\}$’s are orthonormal w.r.t. $\nu$, so $\int \psi_j(\tau) \sum_{j=1}^{\infty} j^{-2s} \psi_j(\tau) \psi_j d\nu (\tau) =$
14

\[ j^{-2s} \psi_j. \] As an additional contribution, our work shows that with the simple assumptions A2 & A3, we can get knowledge about \( T_X \)'s eigenvalues from those of \( T_\nu \).

**Lemma 6.2.** Given assumptions A2, A3, the \( j \)-th eigenvalue, \( \lambda_j \), (sorted in a decreasing order) of \( T_X \) satisfies \( \lambda_j = \Theta(j^{-2s}) \).

Moreover, the upper bound of the density in A2 ensures the upper bound in Lemma 6.2 \( (\lambda_j = O(j^{-2s})) \), and the lower bound of the density ensures the other half of the result. The proof of the above Lemma uses the underlying connection between the eigenvalues of an operator and its metric entropy. For rigorous definitions and proof of Lemma 6.2, see Appendix C.

Although the exact result of Lemma 6.2 is not used in the proof of Theorem 6.1 (or Theorem 6.3). We still present it here since it may be of interest itself and the stated results is less technical and easier to comprehend. The proof of the more technical version (Lemma C.14) follows very closely to that of Lemma 6.2. In that more general version, we investigate the spectrum of covariance operators of form:

\[
T_X^\omega : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}
\]

where

\[
K_{J_n}^s : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}
\]

\[
(s,t) \mapsto \sum_{j=1}^{J_n} j^{-2s} \psi_j(s) \psi_j(t)
\]

While we discuss our work in the context of Sobolev ellipsoids, there is an equivalent formulation directly in RKHS. See Appendix C for more discussion. Although an explicit form for \( K_{J_n}^s \) is not in general necessary or accessible for Sieve-SGD, the existence (i.e. the absolute convergence of the infinite sum) of \( K_{J_n}^s \) is a direct consequence of A3. This is enough for theoretical analysis. For kernel SGD methods, a fixed kernel (with \( J_n = \infty \)) is used and there is only one relevant RKHS. This means, on average, kernel SGD is applying the same procedure each iteration; but for Sieve-SGD, we need to engage with a series of increasing RKHSs (on average, Sieve-SGD may not be doing the same thing between iterations). As a side contribution, we present how to handle such a more technically involved case.

**Note 2:** In contrast to our assumption A3, the hypothesis spaces in [13, 48, 55, 34] are described in terms of “\( T_X \)” and its eigen-decomposition. This unfortunately obfuscates difficulties related to verifying those conditions when analyzing their statistical performance (though applying the learning algorithm in practice does not need the knowledge of the eigensystem). In particular because \( \rho_X \) is involved in the definition of \( T_X \) (37), we need knowledge of (generally unknown) \( \rho_X \) to characterize \( T_X \), and understand its eigenvalues and eigenfunctions.

More specifically, in the literature we mentioned above, it is often assumed that for some \( r \in [1/2, 1] \) (Definition C.6):

\[
\|T_X^{-r}(f_\rho)\|_{L^2_{\rho_X}} < \infty
\]

This can be related to a Sobolev ellipsoid-type condition

\[
\|T_X^{-r}(f_\rho)\|_{L^2_{\rho_X}} = \sum_{j=1}^{\infty} \lambda_j^{-2r} \theta_j^2 < \infty \quad \text{where} \quad f_\rho = \sum_{j=1}^{\infty} \theta_j \phi_j
\]

where \((\lambda_j, \phi_j)_{j=1}^{\infty}\) are the eigenvalue and eigenfunctions of operator \( T_X \), and \( \phi_j \)'s are orthonormal w.r.t. \( L^2_{\rho_X} \). Unfortunately, we cannot directly engage with \((\lambda_j, \phi_j)_{j=1}^{\infty}\), since calculating them requires knowledge of \( \rho_X \). Thus, assumptions formulated in the language of
$T_{\rho_X}$ are difficult to directly understand. In contrast, our assumptions translate to analyzing the spectrum of $T_{\rho}$, which has no dependence on $\rho_X$, and its spectrum can be directly calculated (as noted above).

**Note 3:** For parametric SGD methods, usually a bound on the second moment of the gradient vector is required to guarantee rate-optimal performance (both theoretically and in practice). Formally, for optimization problem (10), it is usually required that $E[\|\nabla \ell(\beta)\|^2] \leq R^2 < \infty$ for all $\beta \in \mathbb{R}^d$ [6, 14].

For nonparametric stochastic approximation, there is a similar regularity requirement for the “gradient”. The assumptions A2-A3 are enough to ensure this for Sieve-SGD. In our proof, we show that there exists a number $R < \infty$ such that for all $x \in X$ and any $J_n$, we have $\|K_{x,J_n}\|_K^2 \leq R^2$. This result is listed in Lemma D.1 where $R^2 = M^2\zeta^2(2s)$ and $\zeta(k) = \sum_{i=1}^{\infty} i^{-k}$. In Theorem 6.1, we required $\gamma_0$ to be smaller than $(2M^2\zeta^2(2s))^{-1}$ to ensure our theoretical guarantees.

**Note 4:** For completeness, here we state the minimax-rate of our nonparametric regression problem over a Sobolev ellipsoid:

$$\liminf_{n \to \infty} \inf_{\hat{f}} \sup_{f \in W(s,Q,\psi)} E\left[ n^{\frac{2s}{2s+1}} \|\hat{f} - f\|_{L_{X^n}}^2 \right] \geq C$$

where the infimum ranges over all possible functions $\hat{f}$ that are sufficiently measurable. For a derivation of this lower bound, see [52, Chapter 15]. Many other online methods we mentioned in Section 4 can achieve this lower bound, however, their computational expense is in general unfavorable compared with the proposed method.

Also, the bound (36) should not be understood as a dimension-free result. When the feature $X \in \mathbb{R}^p$ is a multivariate vector, the parameter $s$ should be treated as $s = s^*/p$, where $s^*$ is, for example, the order of derivative that we assume the regression function $f_{\rho}$ has. Plugging this into the result presented in Theorem 6.1 gives a dimension-dependent bound of order $n^{-\frac{s^*}{2s+1}}$ in which both the smoothness assumption and dimension show up in the exponent. Such a bound is minimax optimal when learning in a (large) homogeneous multivariate space [45]. In practice, one can usually achieve better performance by leveraging other low dimensional structure (See Section 5.4 and Appendix B).

6.3. **Robustness of $t_j$ for Properly Chosen $J_n$.** In Section 6.2 we presented the optimality guarantees of Sieve-SGD in the case when the component-specific learning rate is chosen as the most “natural” value, i.e. $t_j = j^{-2s}$. In that case, Sieve-SGD is statistically optimal so long as the number of basis functions does not increase too slowly, that is, $J_n \geq n^{\frac{1}{2s+1}} \log^2(n)$. Specifically, when $J_n = \infty$, the Sieve-SGD updating rule reduces to the kernel SGD updating rule (20) with kernel $K^{s\psi}_n(X_n, \cdot) = \sum_{j=1}^{\infty} j^{-2s} \psi_j(X_n) \psi_j(\cdot)$. So long as we have access to the closed-form of $K^{s\psi}_n(X_n, \cdot)$, the corresponding kernel SGD estimator is also statistically optimal under the same conditions. In such a scenario, Sieve-SGD can be seen as a truncated-version of a “correct” kernel SGD method with much better computational properties.

Although truncating the kernel in the spectral domain may be seen as an extension of kernel SGD, it can alternatively be seen as related to projection estimators: In that case however, two pieces of Theorem 6.1 may seem unnatural: 1) the strict requirement on $t_j$ ($= j^{-2s}$); and 2) The fact that we only lower bound the truncation rate, rather than requiring a precise value for the growth of $J_n$. In the case of the original Theorem 6.1, the bias-variance tradeoff is actually not balanced via truncation. Instead, it is balanced directly using the $t_j$. The required lower bound on the truncation rate is just given to ensure that we do not accrue excess bias. Alternatively, to better parallel projection estimators, it might seem more natural to directly
use the number of basis functions \( J_n \) to control the bias-variance tradeoff (there is nothing akin to \( t_j \) in (8)). In this section we will explore this idea: That if we are more precise in specifying \( J_n \), perhaps we can be more flexible with \( t_j \).

More specifically, we are interested in milder conditions on the sequence \((t_j)\) such that if we properly select the rate at which our “dimension” increases (i.e. the rate at which \( J_n \) grows), Sieve-SGD would still attain its favorable statistical and computational properties. Since we will be using \( J_n \) as the tuning parameter to balance bias and variance, one might expect kernel SGD, which sets \( J_n = \infty \), would not always have optimal statistical performance for all sequences \((t_j)\) satisfying the “milder” conditions. This is confirmed via the following theorem: For Sieve-SGD one can actually use large component-wise step-sizes that need only satisfy \( t_j < j^{-1} \) for any smoothness class \( W(s) \), so long as the truncation level is appropriately set; while the largest \( t_j \) that can be used for kernel-SGD (without truncation) needs to ensure \( t_j < j^{-(s+1/2)} \), which depends on the smoothness of \( f_\rho \).

**Theorem 6.3.** Assume A1-A4. Set the component-specific learning rate to be \( t_j = j^{-2\omega} \) with \( \frac{1}{2} < \omega \leq s \). Choose the learning rate to be \( \gamma_n = \gamma_0 n^{-2t_j^{-1}} \), with \( \gamma_0 \leq M^2 \zeta(2\omega)/2 \). Choose the number of basis functions to be \( J_n = n^{2t_j^{-1}} \log^2 n \lor 1 \).

Then the MSE of Sieve-SGD (27) converges at the following near optimal rate

\[
E\|\hat{f}_n - f_\rho\|_{L^2_{\cdot X}}^2 = O\left(n^{-2t_j^{-1}} \log^2 n\right)
\]  

**Note 1:** The requirement of \( t_j < j^{-1} \) is to guarantee a finite “second moment” of the gradient (as in Note 3 under Theorem 6.1). In this theorem, once this minimal requirement is satisfied, the decay rate of \( t_j \) does not influence either the statistical guarantees, nor the computational expense of the estimators — both of these are determined entirely by the truncation level. As we will discuss very soon in Section 6.4, the choice of \( J_n = n^{2t_j^{-1}} \log^2 n \) in Theorem 6.3 and Theorem 6.1 would result in algorithms that are both statistically and computationally near-optimal up to a polylog term.

**Note 2:** The most direct form of the projection estimator determines the basis functions’ coefficients by solving a (unpenalized) least square problem (8) in which there are no learning rates involved. It is the truncation level \( J_n \) that determines the bias-variance trade-off and statistical performance. In Theorem 6.3 we present a stochastic approximation analog to that result. From a reproducing-kernel methodology perspective, Theorem 6.1 investigates the cases when the capacity of the kernel (\( \omega \)) matches the source smoothness (\( s \)); in Theorem 6.3 we show under what conditions the mismatch between these two quantities does not affect the statistical (and computational) properties of Sieve-SGD. It is very common to discuss the generalization ability of a reproducing kernel method in the literature when the kernel capacity does not match the source smoothness. For example, in [13] the authors use a pair of parameters \( (r, \alpha) \) to state the hypothesis space and the capacity of the kernels. The smoothness of the hypothesis space is determined by the product of the two parameters \( r\alpha \). When \( r \neq 1/2 \), they are considering using a kernel whose capacity does not match the smoothness of \( f_\rho \). Their proposed method must modify the learning rate properly to recover rate-optimality (or it is impossible due to saturation).

Comparing their results with Theorem 6.3, there are \( \omega \) such that the kernel SGD estimator, using kernel \( K^{\omega}_{\cdot\cdot}(X_n, \cdot) = \sum_{j=1}^{\infty} j^{-2\omega} \psi_j(X_n) \psi_j(\cdot) \), may not be optimal no matter how we modify the learning rate \( \gamma_n \) (described as “saturation” in [13]). Whereas for Sieve-SGD using the truncated “kernels” \( K^{\cdot\cdot}_{j}(X_n, \cdot) = \sum_{j=1}^{J_n} j^{-2\omega} \psi_j(X_n) \psi_j(\cdot) \), the statistical and computational performance can still be jointly near optimal. Theorem 6.3 is formally similar to such a “source-capacity” discussion, but the results are quite different in nature — in particular
it is the truncation level that “saves” us, and allows a much larger mismatch between kernel capacity and source smoothness. 

**Note 3:** The overall proof structures of Theorem 6.1 and Theorem 6.3 are similar; the difference is, in the proof of Theorem 6.1 we need Lemma D.4 and related technical results, but for Theorem 6.3 we use Lemma E.1 instead.

**Note 4:** We also provide some intuition for using a decreasing learning rate $\gamma_n$: For rate-optimal parametric SGD methods with averaging, the learning rate $\gamma_n$ can be taken as a constant $\gamma_0$. However, the employed constant $\gamma_0$ is inversely proportional to the dimension of parameter (assuming each dimension of the feature has a bounded support) [3], which is, in some sense, consistent with our results (though we have seen no other results in the literature that engage with a dimension that increases as the learning process proceeds). We require the learning rate to be a decreasing sequence so that it can “cancel out” the effect of increasing estimator-dimension: The increasing dimension would have resulted in a noise variance that is increasing if care was not taken.

6.4. Near optimal space expense. In this section we will show that Sieve-SGD is asymptotically (nearly) space-optimal for estimating $f_\rho$ in a Sobolev ellipsoid under the conditions listed in Section 6.1. We will show that, even with computer round-off error, Sieve-SGD only needs $\Theta(n^{\frac{1}{2r+1}} \log^3 n)$ bits to achieve the minimax rate for MSE (or off by a $\log^2(n)$ term when $\omega \neq s$ as stated in Theorem 6.3), and further, that there is no estimator with $o(n^{\frac{1}{2r+1}})$ bits of space expense that can achieve the minimax-rate for estimating $f_\rho \in W(s, Q)$. In our analysis we note that computers cannot store decimals in infinite precision, and formally deal with a modified version of our algorithm that stores coefficients in fixed precision (that grows in $n$): This necessitates the extra $\log(n)$ term (compared with the number of basis function needed in Theorem 6.1 and 6.3). The modified algorithm with fixed, but growing precision still results in the same MSE when round-off error is not considered.

We first give a more formal definition of the space expense of an estimator in our analysis. A regression estimator can be seen as a mapping $M_n$ from the data $Z^n_1 = \{(X_i, Y_i) \mid i = 1, 2, \ldots, n\}$ to a function $\hat{f}_n \in \mathcal{F}$. For any such $M_n$ that can be engaged by a computer, must be decomposable into an “encoder-decoder” pair $(E_n, D_n)$. Here $E_n$ represents the “encoder” that compresses the information into computer memory. Formally, we define $E_n$ to be a mapping from $Z^n_1$ to a binary sequence of length $b_n$. And the corresponding $D_n$ is the “decoder” of the binary sequence that translates the information saved in memory back to a mathematical object $\hat{f}_n$. Generally, the binary sequence length $b_n$ will increase with $n$: As more information is contained in the data, we need more memory to store an increasingly accurate estimate of our regression function.

Given an estimator that can decomposed into a pair $(E_n, D_n)$, one can see that the decomposition is not unique. There are, in fact, infinitely many pairs that are trivially different from each other for any such estimator. Moreover, $E_n, D_n$’s can be random mappings if we allow random algorithms: For example, random forests include additional randomness due to bootstrapping/variable selection. In order to be more precise regarding memory complexity constraints, we introduce the following formalization.

**Definition 6.4 ($b_n$-sized estimator).** Given a sequence of integers $(b_i)_{i \in \mathbb{N}}$, we say an estimator $M_n : (\mathcal{X} \times \mathbb{R})^n \rightarrow \mathcal{F}$ is a $b_n$-sized estimator if it satisfies the following conditions:

1. For every $n$, there exists an encoder mapping $E_n : (\mathcal{X} \times \mathbb{R})^n \rightarrow \{0, 1\}^{b_n}$, and a decoder mapping $D_n : \{0, 1\}^{b_n} \rightarrow \mathcal{F}$ such that

$$M_n = D_n \circ E_n$$  (44)

2. The decoder $D_n$ is a known, fixed mapping. $E_n$ can be either a random or fixed mapping.

We use the sample mean as a toy example to illustrate the above definition. In practice, the sample mean is usually a 64-sized estimator of the population mean. Here 64 stands for the number of bits needed to represent a double-precision floating point number. In this case the size $b_n = 64$ does not increase with sample size $n$. However not every real number can be arbitrarily precisely specified by a fixed-length floating-point number, so a careful asymptotic analysis of estimation of the mean suggests that perhaps we should store a sample mean with growing levels of precision, i.e. $b_n$ would need to grow with $n$. A binary sequence of length $s$ can specify $2^s$ real numbers, so to achieve the $O(n^{-1})$ statistically optimal bound for mean estimation, a $\log(n)$-sized version of sample mean is formally required. In practice, 64-bit precision is generally more than enough for mean estimation. Nevertheless, in this manuscript we aim to give a more formal and precise asymptotic analysis of our Sieve-SGD estimator.

Readers who are more familiar with computational complexity theory may find our definition similar to a (probabilistic) Turing machine. However, in our framework the machine does not use binary sequences on tapes as input and output; nor do we need to identify the basic operations on the "machine". We aimed to remove unnecessary complexity for readers with a more statistical background. Discussion of Turing machines using finite length working tape can be found in [1, Chapter 4].

To construct Sieve-SGD estimators that achieve (near) optimal MSE, we only need to store the coefficients of the $J_n = \Theta(n^{\frac{1}{2} s + 1})$ basis functions. However, as in our example with the sample mean, we need to be careful about the precision with which we store those coefficients. We need to determine: i) how small we require the round-off error to be in order to maintain the statistical optimality of Sieve-SGD, and ii) how much space expense is required to achieve such precision. In Appendix F.1 we identify how round-off error is introduced into the system and how it decreases as more bits are used to store each coefficient. In Corollary F.2 we show that by using $\Theta(\log n)$ bits per coefficient, a $O(n^{\frac{1}{2} s + 1})$-sized version of Sieve-SGD can achieve the same optimal convergence rate as in the infinite precision setting (or equivalently round-off-error-free setting).

Combining the above result with the following theorem, we can conclude that no MSE rate-optimal estimator can require less memory by a polynomial factor than Sieve-SGD.

**Theorem 6.5.** Let $b_n$ be a sequence of integers, and $b_n = o\left(n^{\frac{1}{2} s + 1}\right)$. Let $M(b_n)$ be the collection of all $b_n$-sized estimators, then we have

\[
\lim_{n \to \infty} \inf_{M_n \in M(b_n)} \sup_{f \in W(s,Q,\{\psi_j\})} \mathbb{E}\left[ n^{\frac{1}{2} s + 1} \left\| M_n(Z_1^n) - f \right\|_{L^2_{X}}^2 \right] = \infty
\]

i.e. no such $b_n$-sized estimators can be rate-optimal.

This theorem tells us we cannot find any minimax rate-optimal $o(n^{\frac{1}{2} s + 1})$-sized estimator. Thus the best rate-optimal estimator one can expect to find is a $\Theta(n^{\frac{1}{2} s + 1})$-sized estimator: Sieve-SGD’s space expense only misses this lower bound by a poly-logarithmic factor.

We give the proof of the above theorem in Appendix F.2. Although here we focus on regression in Sobolev spaces, the technique used can be applied to other hypothesis spaces. The proof is based on the concept that metric-entropy is the minimal number of bits needed to represent an arbitrary function from a function space up to $\epsilon$-error, which can be traced back to [28]. Also, following a very similar argument, one can prove that no constant-sized estimator can be rate-optimal (or even consistent) for parametric regression problems. We discuss this further in the Appendix F.2. We also include some discussion of the time expense in Section 8.
7. Simulation study.

7.1. Sieve-SGD for Online Regression. In this section, we illustrate both the statistical and computational properties of Sieve-SGD with simulated examples. The two examples we use have different $f_\rho$, $W(s, Q, \{\psi_j\})$ and $\rho_X$. The user-specified measure $\nu$ is taken as the uniform distribution over $[0, 1]$ in both. We provide the details of our simulation settings in Table 1. These two examples are designed for verifying our theoretical guarantees: The $f_\rho$ we use have known explicit series expansion or is constructed explicitly using the basis function $\psi_j$ (to ensure the truth is hard enough to learn in the assumed Sobolev ellipsoid). In Appendix B we provide more numerical examples to better mimic the practical application: we engage with multivariate features and compare Sieve-SGD with many popular machine learning methods.

Example 1 In this example, we examine the empirical performance of Sieve-SGD and compare it with two other methods in batch or online nonparametric regression: kernel ridge regression (KRR) [52] and kernel SGD [13]. We will see that the relationship between generalization error $\|\hat{f}_n - f_\rho\|_2^2$ and sample size corresponds well with our theoretical expectations presented in Theorem 6.1 (Fig 1).

The true regression function we chose for Example 1 is also used in the analysis of kernel SGD [13]. In that paper, kernel SGD with Polyak averaging is compared with other (kernel-based) nonparametric online estimators [48, 55], and has been shown to have similar or better performance, so we include only kernel SGD with averaging as the reference (kernel-based) nonparametric online estimators $\Theta(n^3)$ per update). We also note that although KRR performs slightly better than online methods, its time expense (which is of order $\Theta(n^3)$ per update) is dramatically more than online-estimators (kernel SGD $\Theta(n)$, Sieve-SGD $\Theta(J_n)$, per update).

We compare the empirical performance of Sieve-SGD under two different distributions of $X$. In Fig 1 panel (A), $X$ has an uniform distribution over $[0, 1]$ and in panel (B) it has a distribution with a strictly smaller support (uniform over $[0.25, 0.75]$). The trigonometric basis functions we use are orthonormal w.r.t. the Lebesgue measure over $[0, 1]$ (panel (A)) but not w.r.t. the one in panel (B). Although only the feature distribution in panel (A) satisfies the distribution assumption in A2, in both cases Sieve-SGD achieves the optimal-rate. This is a heuristic evidence indicating the lower bound requirement in A2 may be due to some artifacts in the proof.

| Example 1 | Example 2 |
|-----------|-----------|
| True $f_\rho$ | $B_4(x)$ | $4\sqrt{2} \sum_{j=1}^{50} (-1)^{j+1} j^{-4} \sin((2j-1)\pi x/2)$ |
| Ellipsoid para. $s$ | $2$ | $n^{0.21}$ |
| $J_n$ | $n^{0.10}$ & $n^{0.15}$ & $n^{0.43}$ | $3$ |
| $t_j$ | $j^{-1.02}$ & $j^{-4}$ | $j^{-6}$ |
| $\psi_j(x)$ | $\sin(2\pi [j/2]x)$, $j$ is even | $\sqrt{2} \sin(\frac{(2j-1)\pi x}{2})$ |
| Kernel $K(s, t)$ | $-\frac{1}{2\pi} B_4(\{s - t\})$ | $\min(s, t)$ |
| Noise | $\text{Unif}[-0.2, 0.2]$ or $\text{Unif}[-0.2, 0.2]$ | $\text{Normal}(0,1)$ |
| $\gamma_0$ | $3$ | $1$ |

Example 2 In this example, we consider the performance of Sieve-SGD under different $J_n = [n^\alpha]$ (number of basis functions). The $f_\rho$ we use is explicitly constructed with basis functions $\psi_j(x) = \sqrt{2} \sin((2j-1)\pi x/2)$ and we tune the proposed method based on the...
Example 1, $\log_{10}\left(\|f_n - f_\rho\|^2\right)$ against $\log_{10} n$. The Black line has slope $= -4/5$, which represents the optimal-rate. Each curve is calculated as the average of 100 repetitions. (A) $X$ is uniformly distributed over $[0, 1]$. In this setting, SNR $\sim 3$. (B) $X$ has a distribution in which $\psi_j$ are not orthonormal. We present the results with very large noise, SNR $\sim 0.03$. Due to different computational costs, we chose different maximum $n$ for different methods.

As we can see from Fig 2 (A), when $\alpha = 0.15 & 0.43$, Sieve-SGD is rate-optimal as expected. When $\alpha = 0.10$, we are using too few basis functions, which results in the sub-optimal statistical performance. Such a suboptimality is because of the bias term: there are too few basis functions used. In fact, the parameter setting $\alpha = 0.1$ is so small that there are only 3 basis functions used when $n = 10^5$. To verify the above statement, we can briefly calculate when the second and the third basis functions are added in: $(10^{3})^{0.1} \sim 2$, this corresponds to the first acceleration of the learning rate around $\log_{10}(n) = 3$; similarly, $(10^{4.8})^{0.1} \sim 3$, which explains the second one.

In Fig 2 (B), we show the CPU time for reference. For Sieve-SGD, the accumulated CPU time should be on the order of $\Theta(n^{1+\alpha})$: The larger $\alpha$, the more basis functions required, the slower the algorithm. We also include the CPU time of kernel SGD with averaging as a benchmark, which has a cumulative computational expense of order $\Theta(n^2)$. The code is written in R (4.0.4), and runs on (the CPU of) a machine with 1 Intel Core m3 processor, 1.2 GHz, with 8 GB of RAM.

7.2. Sieve-SGD for Alternative Convex Losses. In this section, we provide the results of an experiment applying Sieve-SGD to online nonparametric logistic regression. Although this manuscript gives no theoretical guarantees in this setting, it is still of interest to see the empirical performance of Sieve-SGD for general convex loss. Here, the distribution of class labels $Y$ was generated by $Y \sim 2 \text{Ber}(g(X)) - 1$, where $(g(x))^{-1} = 1 + \exp(-5(1 - 2|x - \theta|))$.
When we apply the Sieve-SGD estimator \((30)\) to this problem, we assume the minimizer \(f^*\) of loss \(E[\ell(Y, f(X)) = E[\log(1 + \exp(-Y f(X)))]\) is \(f^* = 5(1 - 2|x - 0.5|)\).

We try several \(\alpha = 0.10, 0.33, 0.50\), all with \(\gamma_0 = 6\). As we can see from Fig 3, the regret \(E[\ell(f_n) - \ell(f^*)]\) converges to zero at an apparent rate of \(n^{-2/3}\) when \(\alpha = 0.33, 0.50\) (which would agree with our result for squared error loss). When the number of basis functions increases too slowly (here is \(\alpha = 0.10\)), the regret decreases slowly after \(\sim 10\) observations (for similar reason of overflowing bias term as we noted in section 7.1).

8. Discussion. In this paper, we considered online nonparametric regression in a Sobolev ellipsoid. We proposed the Sieve Stochastic Gradient Descent estimator (Sieve-SGD), an online estimator inspired by both a) the nonparametric projection estimator, which is a special realization of general sieve estimators; and b) estimators constructed using stochastic gradient descent algorithms. By using an increasing number of basis functions, Sieve-SGD has rate-optimal estimation error and is computationally very efficient.

For online learning problems with general convex losses, the optimal estimation rate depends on both the hypothesis space and loss function (e.g. whether it is Lipschitz or strongly convex). In this paper we did not establish theoretical guarantees for Sieve-SGD when applied to general convex loss, however, we gave some empirical evidence that it can perform well there. We believe our proof techniques might be extended beyond squared-error loss, perhaps using ideas in \([3, 10, 34, 35]\).
any computational restrictions. By relating their statistical problem with a known NP problem \cite{1}, they can usually show the sub-optimality of polynomial-time algorithms under the famous conjecture $P \neq NP$. However, for the nonparametric regression problem in this paper, there is a polynomial-time estimator that can achieve the global minimax-rate. It is of theoretical interest to know if there are any statistically rate-optimal online estimators that require less than $\Theta(n^{1+\frac{1}{2s+1}})$ time-expense: We hypothesize that there are not.

\textbf{Acknowledgments.} N.S and T.Z. were both supported by NIH grant R01HL137808.

\section*{REFERENCES}
\begin{enumerate}
\item \textsc{Arora, S. and Barak, B.} (2009). \textit{Computational complexity: a modern approach}. Cambridge University Press.
\item \textsc{Babichev, D. and Bach, F.} (2018). \textit{Constant Step Size Stochastic Gradient Descent for Probabilistic Modeling}. \textit{stat} \textbf{1050} 21.
\item \textsc{Bach, F. and Moulines, E.} (2013). Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In \textit{Advances in neural information processing systems} 773–781.
\item \textsc{Belkin, M., Hsu, D., Ma, S. and Mandal, S.} (2018). Reconciling modern machine learning and the bias-variance trade-off. \textit{arXiv preprint arXiv:1812.11118}.
\item \textsc{Berlinet, A. and Thomas-Agnan, C.} (2011). \textit{Reproducing kernel Hilbert spaces in probability and statistics}. Springer Science & Business Media.
\item \textsc{Borkar, V. S.} (2009). \textit{Stochastic approximation: a dynamical systems viewpoint} \textbf{48}. Springer.
\item \textsc{Bottou, L.} (2010). Large-scale machine learning with stochastic gradient descent. In \textit{Proceedings of COMPSTAT’2010} 177–186. Springer.
\item \textsc{Cai, T. T., Liang, T., Rakhlin, A. et al.} (2017). Computational and statistical boundaries for submatrix localization in a large noisy matrix. \textit{The Annals of Statistics} \textbf{45} 1403–1430.
\item \textsc{Calandriello, D., Lazaric, A. and Valko, M.} (2017). Efficient second-order online kernel learning with adaptive embedding. In \textit{Advances in Neural Information Processing Systems} 6140–6150.
\item \textsc{Caponnetto, A. and De Vito, E.} (2007). Optimal rates for the regularized least-squares algorithm. \textit{Foundations of Computational Mathematics} \textbf{7} 331–368.
\item \textsc{Christmann, A. and Steinwart, I.} (2008). Support vector machines.
\item \textsc{Cucker, F. and Smale, S.} (2002). On the mathematical foundations of learning. \textit{Bulletin of the American mathematical society} \textbf{39} 1–49.
\end{enumerate}
[13] Dieuleveut, A. and Bach, F. (2016). Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics* 44: 1363–1399.

[14] Duchi, J. C. (2014). Multiple Optimality Guarantees in Statistical Learning. PhD thesis, UC Berkeley.

[15] Eubank, R. and Speckman, P. (1990). Curve fitting by polynomial-trigonometric regression. *Biometrika* 77: 1–9.

[16] Fasshauer, G. E. and McCourt, M. J. (2015). *Kernel-based approximation methods using Matlab* 19. World Scientific Publishing Company.

[17] Frostig, R., Ge, R., Kakade, S. M. and Sidford, A. (2015). Competing with the empirical risk minimizer in a single pass. In *Conference on learning theory* 728–763.

[18] Gaillard, P. and Gerchinovitz, S. (2015). A chaining algorithm for online nonparametric regression. In *Conference on Learning Theory* 764–796.

[19] Gao, C., Ma, Z., Ren, Z., Zhou, H. H. et al. (2015). Minimax estimation in sparse canonical correlation analysis. *The Annals of Statistics* 43: 2168–2197.

[20] Gao, C., Ma, Z., Zhou, H. H. et al. (2017). Sparse CCA: Adaptive estimation and computational barriers. *The Annals of Statistics* 45: 2074–2101.

[21] Geer, S. A. and van de Geer, S. (2000). *Empirical Processes in M-estimation* 6. Cambridge university press.

[22] Györfi, L., Kohler, M., Krzyzak, A. and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

[23] Hall, P. and Opsomer, J. D. (2005). Theory for penalised spline regression. *Biometrika* 92: 105–118.

[24] Hardle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (2012). Wavelets, approximation, and statistical applications 129. Springer Science & Business Media.

[25] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

[26] Hernández, E. and Weiss, G. (1996). *A first course on wavelets*. CRC press.

[27] Kennedy, R. A., Sadeghi, P., Khalid, Z. and McEwen, J. D. (2013). Classification and construction of closed-form kernels for signal representation on the 2-sphere. In *Wavelets and Sparsity XV* 8858: 88580M. International Society for Optics and Photonics.

[28] Kolmogorov, A. N. and Tikhomirov, V. M. (1959). $\varepsilon$-entropy and $\varepsilon$-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk* 14: 3–86.

[29] Koppel, A., Warnell, G., Stump, E. and Ribeiro, A. (2019). Parsimonious online learning with kernels via sparse projections in function space. *The Journal of Machine Learning Research* 20: 83–126.

[30] Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications* 35. Springer Science & Business Media.

[31] Liang, T. and Rakhlin, A. (2018). Just interpolate: Kernel" ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387*.

[32] Lu, J., HoI, S. C., Wang, J., Zhao, P. and Liu, Z.-Y. (2016). Large scale online kernel learning. *The Journal of Machine Learning Research* 17: 1613–1655.

[33] Ma, Z., Wu, Y. et al. (2015). Computational barriers in minimax submatrix detection. *The Annals of Statistics* 43: 1089–1116.

[34] Marteau-Ferey, U., Bach, F. and Rudí, A. (2019). Globally Convergent Newton Methods for Ill-conditioned Generalized Self-concordant Losses. In *Advances in Neural Information Processing Systems* 7634–7644.

[35] Marteau-Ferey, U., Ostrovskii, D., Bach, F. and Rudí, A. (2019). Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. *arXiv preprint arXiv:1902.03046*.

[36] Michel, V. (2012). *Lectures on Constructive Approximation: Fourier, Spline, and Wavelet Methods on the Real Line, the Sphere, and the Ball*. Springer Science & Business Media.

[37] Nemirovski, A. (2000). Topics in non-parametric. *Ecole d’Été de Probabilités de Saint-Flour* 28 85.

[38] Nemirovski, A. S. and Yudin, D. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience series in Discrete Mathematics. John Wiley.

[39] Novak, E. and Wozniakowski, H. (2008). *Tractability of multivariate problems. Vol. 1: Linear information*.

[40] Rakhlin, A. and Sridharan, K. (2015). Online nonparametric regression with general loss functions. *arXiv preprint arXiv:1501.06598*.

[41] Raskutti, G., Yu, B. and Wainwright, M. J. (2009). Lower bounds on minimax rates for nonparametric regression with additive sparsity and smoothness. *Advances in Neural Information Processing Systems* 22: 1563–1570.

[42] Shen, X. (1997). On methods of sieves and penalization. *The Annals of Statistics* 2555–2591.
[43] SI, S., KUMAR, S. and LI, Y. (2018). Nonlinear Online Learning with Adaptive Nyström Approximation. arXiv preprint arXiv:1802.07887.

[44] STEINWART, I. and SCOVEI, C. (2012). Mercer’s theorem on general domains: On the interaction between measures, kernels, and RKHSs. Constructive Approximation 35 363–417.

[45] STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. The annals of Statistics 1348–1360.

[46] STONE, C. J. (1985). Additive regression and other nonparametric models. The annals of Statistics 689–705.

[47] SUN, H. (2005). Mercer theorem for RKHS on noncompact sets. Journal of Complexity 21 337–349.

[48] TARRES, P. and YAO, Y. (2014). Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. IEEE Transactions on Information Theory 60 5716–5735.

[49] TSYBAKOV, A. (2008). Introduction to Nonparametric Estimation. Springer Science & Business Media.

[50] VEMPALA, S. S. (2005). The random projection method 65. American Mathematical Soc.

[51] WAHBA, G. (1990). Spline models for observational data 59. Siam.

[52] WAINWRIGHT, M. J. (2019). High-dimensional statistics: A non-asymptotic viewpoint 48. Cambridge University Press.

[53] WANG, T., BERTHET, Q., SAMWORTH, R. J. et al. (2016). Statistical and computational trade-offs in estimation of sparse principal components. The Annals of Statistics 44 1896–1930.

[54] WOOD, S. N. (2017). Generalized additive models: an introduction with R. CRC press.

[55] YING, Y. and PONTILI, M. (2008). Online gradient descent learning algorithms. Foundations of Computational Mathematics 8 561–596.

[56] YUAN, M., CAI, T. T. et al. (2010). A reproducing kernel Hilbert space approach to functional linear regression. The Annals of Statistics 38 3412–3444.

[57] YUAN, M. and ZHOU, D.-X. (2016). Minimax optimal rates of estimation in high dimensional additive models. The Annals of Statistics 44 2564–2593.

[58] ZHANG, T. and SIMON, N. (2021). An Online Projection Estimator for Nonparametric Regression in Reproducing Kernel Hilbert Spaces. arXiv preprint arXiv:2104.00780.

[59] ZHANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In Conference on Learning Theory 921–948.