Stochastic Gradient Descent Meets Distribution Regression

Nicole Mücke
muecke@tu-berlin.de
Berlin Mathematics Research Center MATH+
TU Berlin

Abstract

Stochastic gradient descent (SGD) provides a simple and efficient way to solve a broad range of machine learning problems. Here, we focus on distribution regression (DR), involving two stages of sampling: Firstly, we regress from probability measures to real-valued responses. Secondly, we sample bags from these distributions for utilizing them to solve the overall regression problem. Recently, DR has been tackled by applying kernel ridge regression and the learning properties of this approach are well understood. However, nothing is known about the learning properties of SGD for two stage sampling problems. We fill this gap and provide theoretical guarantees for the performance of SGD for DR. Our bounds are optimal in a mini-max sense under standard assumptions.

1 Introduction

In a standard non-parametric least squares regression model, the aim is to predict a response $Y \in \mathcal{Y}$ from a covariate X on some domain $\tilde{\mathcal{X}}$. Popular approaches are kernel methods (Hofmann et al., 2008), where one defines on $\tilde{\mathcal{X}}$ a reproducing kernel K associated to a *reproducing kernel Hilbert space* $\mathcal{H}_K$ (RKHS) (Aronszajn, 1950; Steinwart and Christmann, 2008). The overall aim is to minimize the least squares error over $\mathcal{H}_K$ by applying a suitable regularization method, involving the kernel and based on an i.i.d. sample, drawn according to some unknown distribution on $\tilde{\mathcal{X}} \times \mathcal{Y}$. We will later refer to such data as a "first-stage" sample.

In this paper, we study *distribution regression* (DR) (Póczos et al., 2013), where the covariate is a probability distribution. Typically, we do not observe this distribution directly, but rather, we observe a "second-stage sample" drawn from that, amounting to a regression model with measurement error.

Distribution regression has been analyzed in various settings, e.g. multiple instance learning (Dooly et al., 2002; Maron and Lozano-Pérez, 1998; Dietterich et al., 1997; Chevaleyre and Zucker, 2001; Wagstaff et al., 2008), in an online setting (Zhi-Gang et al., 2013), in semi-supervised-learning (Zhou and Xu, 2007) or active learning (Settles et al., 2008).

A popular approach for regression on the domain of distributions is to embed the distributions into a Hilbert space. This can be achieved by e.g. *kernel mean embeddings* (Smola et al., 2007; Muandet et al., 2017), utilizing another appropriate reproducing kernel mapping these distributions into an RKHS. The idea is then to introduce the kernel K as a similarity measure between the embedded distributions and to use a traditional kernel machine to solve the overall learning problem.

The learning properties of kernel regularized least squares algorithms based on mean embeddings and with two stages of sampling are rarely analyzed. The first work establishing the learning properties of kernel ridge regression (KRR) is Szabó et al. (2016), where optimal bounds are derived under suitable assumptions on the learning problem and the second-stage sample size. Recently, Fang et al. (2020) also considered KRR, with a slight improvement of results. However, to the best of our knowledge, an analysis of other kernel based regularization methods is missing.

While KRR performs an explicit regularization to avoid overfitting, stochastic gradient descent (SGD) performs an implicit regularization as an iterative algorithm. Many variants of SGD are known for one-stage least squares regression, ranging from considering one pass over the data (Smale and Yao, 2006; Tarres and Yao, 2014; Ying and Pontil, 2008) to multiple passes

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

(Bertsekas, 1997; Rosasco and Villa, 2015; Hardt et al., 2016; Lin et al., 2016; Pillaud-Vivien et al., 2018), with mini-batching (Lin and Rosasco, 2017) or (tail-)averaging (Dieuleveut and Bach, 2016; Mücke et al., 2019; Mücke and Reiss, 2020).

While SGD is a workhorse in machine learning, the learning properties of this algorithm in a two stage-sampling setting based on mean embeddings are not yet analyzed. We aim at providing an algorithm with reduced computational complexity for two-stage sampling problems, compared to KRR, which is known to scale poorly with large sample sizes.

Contributions. We analyze the distribution regression problem in the RKHS framework and extend the previous approaches in Szabó et al. (2016) and Fang et al. (2020) from two-stage kernel ridge regression to two-stage tail-averaging stochastic gradient descent with mini-batching. Our main result is a computational-statistical efficiency trade-off analysis, resulting in finite sample bounds on the excess risk. In particular, we overcome the saturation effect of KRR.

We give a minimum number of the second-stage sample size which is required to obtain the same best possible learning rates as for the classical one-stage SGD algorithm. For well-specified models, i.e. the regression function belongs to the RKHS where SGD is performed, we achieve minimax optimal rates with a single pass over the data. These bounds match those for classical kernel regularization methods. In the misspecified case, i.e. the regression function does not belong to the RKHS, our bounds also match those for the one-stage sample methods with multiple passes over the data.

Moreover, we investigate the interplay of all parameters determining the SGD algorithm, i.e. mini-batch size, step size and stopping time and show that the same error bounds can be achieved under various choices of these parameters.

On our way we additionally establish the learning properties of tail-averaging two-stage gradient descent which is necessary for deriving our error bounds for SGD. Due to space restrictions, this is fully worked out in the Appendix, Section 2.

Our results are the first for distribution regression using SGD and a two-stage sampling strategy.

Outline. In Section 2 we introduce the distribution regression problem in detail. We introduce our main tool, kernel mean embeddings, and explain the classical non-parametric regression setting in reproducing kernel Hilbert spaces. In addition, we define our second-stage SGD estimator. Section 3 collects our main results for different settings, followed by a detailed discussion in Section 4. All proofs are deferred to the Appendix.

2 The Distribution Regression Problem

In this section we introduce the distribution regression problem in detail. Let us begin with with some notation. We let $(\mathcal{X}, \tau)$ be a compact topological space and denote by $\mathcal{B}(\mathcal{X})$ the Borel σ-algebra induced by the topology τ. The set $\mathcal{M}^+(\mathcal{X})$ denotes the set of Borel probability measures on the measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, endowed with the weak topology. We furthermore assume that there exists a constant $M > 0$ such that $\mathcal{Y} \subseteq [-M, M]$.

Our approach is based on two stages of sampling:

1. We are given data $\{(x_j, y_j)\}_{j=1}^n \subset \mathcal{M}^+(\mathcal{X}) \times \mathcal{Y}$, i.e., each input x_j is a probability distribution with corresponding label y_j. Each pair (x_j, y_j) is i.i.d. sampled from a meta distribution $\mathcal{M}$ on $\mathcal{M}^+(\mathcal{X}) \times \mathcal{Y}$. However, we do not observe x_j directly.
2. Instead, for each distribution x_j we obtain samples $\{x_{j,i}\}_{i=1}^N \subset \mathcal{X}$, drawn i.i.d. according to x_j. The observed data are $\hat{\mathbf{z}} = \{(\{x_{j,i}\}_{i=1}^N, y_j)\}_{j=1}^n$.

2.1 Our Tool: Kernel Mean Embeddings

Following the previous approaches in Szabó et al. (2016), Fang et al. (2020) we employ *kernel mean embeddings* to map the distributions $\{x_j\}_{j=1}^n$ into a Hilbert space. To be more specific, we let $\mathcal{H}_G$ be a *reproducing kernel Hilbert space* (RKHS) with a Mercer kernel $G : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, i.e., G is symmetric, continuous and positive semidefinite (Aronszajn, 1950; Steinwart and Christmann, 2008). Moreover, we make the following

Assumption 2.1 (Boundedness I)**.** *The kernel G is bounded, i.e.*

$$\sup_{s \in \mathcal{X}} G(s, s) =: \gamma^2 < \infty \,, \quad a.s. \,,$$

w.r.t. any probability measure on $\mathcal{X}$.

The associated mean embedding is a map $\mu : \mathcal{M}^+(\mathcal{X}) \to \mathcal{H}_G$, defined as

$$\mu_x := \mu(x) := \int_{\mathcal{X}} G(s, \cdot) \, dx(s) \,. \tag{1}$$

Kernel mean embeddings were introduced in e.g. Smola et al. (2007) as a technique for comparing distributions without the need for density estimation as an

intermediate step and thus have a broad applicability, see also Muandet et al. (2017) and references therein.

Of particular interest are *characteristic* kernels, i.e. the map $\mu : x \mapsto \mu_x$ is injective (Fukumizu et al., 2004). Those type of kernels are essential since $||\mu_x - \mu_{x'}||_{\mathcal{H}_G} = 0$ is equivalent to $x = x'$ and there is no loss of information when mapping a distribution into a characteristic RKHS[1].

It is well known that *universal* kernels[2] are characteristic, see e.g. Theorem 1 in Smola et al. (2007). Examples include the *exponential kernel, binomial kernel* or the *Gaussian RBF kernel*. Thus, a kernel mean embedding serves as a suitable tool for measuring the similarity between two distributions.

For controlling the two stage sampling process we shall employ this property and compare each distribution x_j from the first stage sample with its empirical distribution $\hat{x}_j := \frac{1}{N}\sum_{i=1}^{N} \delta_{x_{j,i}}$ obtained from the second stage of sampling by mapping them into the RKHS $\mathcal{H}_G$ by means of the kernel mean embedding (1).

Thus, the first stage data for DR are

$$D := \{(\mu_{x_j}, y_j)\}_{j=1}^n \subset \mu(\mathcal{M}^+(\mathcal{X})) \times \mathcal{Y} \,,$$

while the second stage data are

$$\hat{D} := \{(\mu_{\hat{x}_j}, y_j)\}_{j=1}^n \subset \mu(\mathcal{M}^+(\mathcal{X})) \times \mathcal{Y} \,,$$

with the associated mean embeddings

$$\mu_{x_j} = \int_{\mathcal{X}} G(s,\cdot)\, dx_j(s) \,, \quad \mu_{\hat{x}_j} = \frac{1}{N}\sum_{i=1}^{N} G(x_{j,i},.) \,.$$

Both datasets now belong to the same space, making the two stage sampling problem accessible for further investigations applying classical kernel methods, as amplified below.

2.2 One-Stage Least Squares Regression

We let ρ be a probability measure on $\mathcal{Z} := \mu(\mathcal{M}^+(\mathcal{X})) \times \mathcal{Y}$ with marginal distribution ρ_μ on the image $\mu(\mathcal{M}^+(\mathcal{X})) \subset \mathcal{H}_G$. In least squares regression we aim to minimize the risk with respect to the least squares loss, i.e.

$$\min_{\mathcal{H}} \mathcal{E}(f) \,, \quad \mathcal{E}(f) := \int_{\mathcal{Z}} (f(\mu_x) - y)^2 \, d\rho \qquad (2)$$

over a suitable hypotheses space $\mathcal{H}$. Here, we assume that $\mathcal{H} = \mathcal{H}_K$ is a RKHS associated with a kernel K on $\mu(\mathcal{M}^+(\mathcal{X}))$, satisfying:

[1] A RKHS is called *characteristic* if it's associated kernel is characteristic

[2] A continuous kernel is called *universal*, if it's associated RKHS is dense in the space of continuous bounded functions on the compact domain $\mathcal{X}$ (Steinwart and Christmann, 2008).

Assumption 2.2 (Boundedness II)**.** *The kernel K is bounded, i.e.*

$$\sup_{\tilde{\mu}\in\mu(\mathcal{M}^+(\mathcal{X}))} K(\tilde{\mu},\tilde{\mu}) =: \kappa^2 < \infty \,, \quad \rho_\mu - a.s. \,.$$

Note that under this assumption, the RKHS $\mathcal{H}_K$ can be continuously embedded into $L^2(\mu(\mathcal{M}^+(\mathcal{X})), \rho_\mu)$ and we henceforth denote this inclusion by $S_K : \mathcal{H}_K \hookrightarrow L^2(\mu(\mathcal{M}^+(\mathcal{X})), \rho_\mu)$.

The minimizer of (2) over $L^2(\mu(\mathcal{M}^+(\mathcal{X})), \rho_\mu)$ is known to be the *regression function*

$$f_\rho(\mu_x) = \int_{\mathcal{Y}} y \, d\rho(y|\mu_x) \,, \quad \mu_x \in \mu(\mathcal{M}^+(\mathcal{X})) \,,$$

where $\rho(\cdot|\mu_x)$ denotes the conditional distribution on $\mathcal{Y}$ given $\mu_x \in \mu(\mathcal{M}^+(\mathcal{X}))$. Note that our assumption $\mathcal{Y} \subseteq [-M, M]$ implies that $||f_\rho||_\infty \leq M$.

Classical kernel based approaches for least-squares regression to (approximately) solve (2) employ some kind of explicit or implicit regularization. Among them, and well understood, are Kernel Ridge Regression (Caponnetto and De Vito, 2006; Fischer and Steinwart, 2017), Kernel PCA, Gradient Descent (Blanchard and Mücke, 2018; Lin et al., 2020) or Stochastic Gradient Descent (Dieuleveut and Bach, 2016; Lin et al., 2016; Lin and Rosasco, 2017; Mücke et al., 2019; Mücke and Reiss, 2020).

All these methods use the first stage data $D = \{(\mu_{x_j}, y_j)\}_{j=1}^n$ to build an estimator f_D with an appropriate amount of regularization and the overall aim is to achieve a small *excess risk*

$$\mathcal{E}(f_D) - \inf_{f\in\mathcal{H}} \mathcal{E}(f) \,,$$

with high probability with respect to the data D.

2.3 Solving DR With Two-Stage Sampling SGD

Remember we do not directly have access to the first stage data D but by means of the tool of kernel mean embeddings we are able to use the second stage data $\hat{D}$ for our algorithm. Our aim is to perform a variant of stochastic gradient descent. To this end, let $i_\cdot = i(\cdot)$ denote a map defining the strategy with which the data are selected at each iteration $t = 0, ..., T$. The most common approach, which we follow here, is sampling each point uniformly at random with replacement. We additionally consider *mini-batching*, where a batch of size $b \in \{1, ..., n\}$ of data points at each iteration is selected. Formally, the $j_1, ..., j_{bT}$ are iid random variables, distributed according to the uniform distribution on $\{1, ..., n\}$.

Starting with $\hat{h}_0 \in \mathcal{H}_K$, our SGD recursion is given by

$$\hat{h}_{t+1} = \hat{h}_t - \eta \frac{1}{b} \sum_{i=b(t-1)+1}^{bt} (\hat{h}_t(\mu_{\hat{x}_{j_i}}) - y_{j_i}) K_{\mu_{\hat{x}_{j_i}}} \ ,$$

where we write $K_\mu := K(\mu, \cdot)$ and where $\eta > 0$ is the stepsize. The number of passes after T iterations is $\lfloor bT/n \rfloor$.

We are particularly interested in tail-averaging the iterates, that is

$$\bar{\hat{h}}_T := \frac{2}{T} \sum_{t=\lfloor T/2 \rfloor + 1}^{T} \hat{h}_t \ . \tag{3}$$

The idea of averaging the iterates goes back to Robbins and Monro (1951), Polyak and Juditsky (1992), see also Shamir and Zhang (2013). More recently, in Dieuleveut and Bach (2016) full averaging, i.e. summing up the iterates from $t = 1$ to $t = T$, was shown to lead to the possibility of choosing larger/ constant stepsizes. However, it is also known to lead to *saturation*, i.e. the rates of convergence do not improve anymore in certain well-specified cases and thus leads to suboptimal bounds in the high smoothness regime. This has been alleviated in Mücke et al. (2019) by considering tail-averaging, see also Mücke and Reiss (2020).

Note that our SGD algorithm only has access to the observed input samples $\{x_{j,i}\}_{i=1}^N$, $j = 1, ..., n$ through their mean embeddings $\{\mu_{\hat{x}_j}\}_{j=1}^n$.

Main goals: We analyze the excess risk[3]

$$\mathbb{E}_{\hat{D}|D}[\mathcal{E}(\bar{\hat{h}}_T) - \mathcal{E}(f_\rho)] = \mathbb{E}_{\hat{D}|D}[||S_K \bar{\hat{h}}_T - f_\rho||^2_{L^2}]$$

and study the interplay of all parameter b, η, T determining the SGD algorithm. We derive finite-sample high probability bounds, presenting computational-statistical efficiency trade-offs in our main Theorem 3.4. In addition, we give fast rates of convergence as the sample sizes n grows large and give an answer to the question

How many second-stage samples N do we need to obtain best possible learning rates, comparable to one-stage learning ?

Our bounds depend on the difficulty of the problem. More precisely, we shall investigate the learning properties of (3) in two different basic settings:

1. Well-specified Model: Here, we assume that the regression function f_ρ belongs to the RKHS $\mathcal{H}_K$. We analyze this setting in Section 3.1 and give high probability bounds, matching the known optimal bounds for one stage regularization methods and two stage kernel ridge regression.

2. Mis-specified Model: In this case the regression function is assumed to not to belong to the RKHS $\mathcal{H}_K$. These bounds are presented in Section 3.2 and still match the known optimal bounds in the so called *easy learning* regime, to be refined below. For so called *hard learning* problems, our bounds still match the best known ones for classical one-stage kernel methods.

3 Main Results

This section is devoted to presenting our main results. Before we go into more detail, we formulate our assumptions on the learning setting. The first one considers the reproducing kernel that we define on the set $\mu(\mathcal{M}^+(\mathcal{X}))$.

Assumption 3.1 (Hölder Property)**.** *Let $\alpha \in (0, 1]$ and $L > 0$. We assume that the mapping $K_{(\cdot)} : \mu(\mathcal{M}^+(\mathcal{X})) \to \mathcal{H}_K$ defined as $\tilde{\mu} \mapsto K(\tilde{\mu}, \cdot)$ is (α, L)-Hölder continuous, i.e.*

$$||K_{\mu_1} - K_{\mu_2}||_{\mathcal{H}_K} \leq L ||\mu_1 - \mu_2||^\alpha_{\mathcal{H}_G} \ ,$$

for all $\mu_1, \mu_2 \in \mu(\mathcal{M}^+(\mathcal{X}))$.

The next assumption refers to the *regularity* of the regression function f_ρ. It is a well established fact in learning theory that the regularity of f_ρ describes the hardness of the learning problem and has an influence of the rate of convergence of any algorithm. To smoothly measure the regularity of f_ρ we introduce the *kernel integral operator* $L_K = S_K S_K^* : L^2(\mu(\mathcal{M}^+(\mathcal{X})), \rho_\mu) \to L^2(\mu(\mathcal{M}^+(\mathcal{X})), \rho_\mu)$, defined by

$$L_K f(\tilde{\mu}) := \int_{\mu(\mathcal{M}^+(\mathcal{X}))} K(\mu', \tilde{\mu}) f(\mu') \, d\rho_\mu(\mu') \ .$$

Note that under Assumption 2.2, L_K is positive, self-adjoint, trace-class and hence compact, with $||L_K|| \leq trace(L_K) \leq \kappa^2$, see e.g. Steinwart and Scovel (2012).

Assumption 3.2 (Regularity Condition)**.** *We assume that for some $r > 0$ the regression function f_ρ satisfies*

$$f_\rho = L_K^r h_\rho \ , \quad h_\rho \in L^2(\mu(\mathcal{M}^+(\mathcal{X})), \rho_\mu) \ ,$$

with $||h_\rho||_{L^2} \leq R$, for some $R < \infty$.

This assumption is also known as a *source condition*. We recall here that powers of L_K are defined by spectral calculus, see for instance Reed (2012). The larger the parameter r, the smoother is f_ρ. We have the range inclusions $\mathcal{R}ange(L_K^r) \subseteq \mathcal{R}ange(L_K^{r'})$ if $r \geq r'$ with $\mathcal{R}ange(L_K^r) \subseteq \mathcal{H}_K$ for any $r \geq \frac{1}{2}$. Thus, if

[3] $\mathbb{E}_{\hat{D}|D}$ denotes the conditional expectation with respect to the sample $\hat{D}$ given D.

$r \geq \frac{1}{2}$, then f_ρ belongs to $\mathcal{H}_K$ under Assumption 3.2 and we are in the well-specified case. For more general smoothness assumptions we refer to Mücke and Reiss (2020).

Our last condition refers to the capacity of the RKHS $\mathcal{H}_K$. Given $\lambda > 0$, we define the *effective dimension*

$$\mathcal{N}(\lambda) := trace\big(L_K(L_K + \lambda)^{-1}\big) \ .$$

This key quantity can be used to describe the complexity of $\mathcal{H}_K$.

Assumption 3.3 (Effective Dimension)**.** *We assume that for some $\nu \in (0, 1]$, $c_\nu < \infty$, the effective dimension obeys*

$$\mathcal{N}(\lambda) \leq c_\nu \lambda^{-\nu} \ . \tag{4}$$

This assumption is common in the nonparametric regression setting, see e.g. Zhang (2003) or Caponnetto and De Vito (2006); Lin et al. (2020). Roughly speaking, it quantifies how far L_K is from being finite rank. This assumption is satisfied if the eigenvalues $(\sigma_j)_{j\in\mathbb{N}}$ of L_K have a polynomial decay $\sigma_j \leq c' j^{-\frac{1}{\nu}}$, $c' \in \mathbb{R}_+$. Since L_K is trace class, the above assumption is always satisfied with $\nu = 1$ and $c_\nu = \kappa^2$. Smaller values of ν lead to faster rates of convergence.

Being now well prepared, we state our main result.

Theorem 3.4. *Suppose Assumptions 2.1, 2.2, 3.1, 3.2 and 3.3 are satisfied. Let further $\delta \in (0, 1]$, $\eta < \frac{1}{4\kappa^2}$, $\lambda = (\eta T)^{-1}$ and assume*

$$n \geq \frac{32\kappa^2 \log(4/\delta)}{\lambda} \log\left(e\mathcal{N}(\lambda)\left(1 + \frac{\lambda}{||L_K||}\right)\right) .$$

Then with probability not less than $1 - \delta$, the excess risk for the second stage tail-averaging SGD algorithm (3) *satisfies*

$$\begin{aligned}
&\mathbb{E}_{\hat{D}|D}[||S_K \bar{\tilde{h}}_T - f_\rho||_{L^2}] \leq C \ \log(6/\delta)\big(\ (\eta T)^{-r} + \\
&+ \ \sqrt{\frac{\mathcal{N}(\lambda)}{n}} \ + \ \frac{(\eta T)^{\frac{1}{2}-r}}{\sqrt{n}} \ + \\
&+ \log(T)\frac{\eta T}{N^{\frac{\alpha}{2}}}\Big(1 + 1_{(0,\frac{1}{2}]}(r)(\eta T)^{\max\{\nu, 1-2r\}}\Big) + \\
&+ \sqrt{\frac{\eta}{b}(\eta T)^{\nu-1}}\left(\frac{\eta T}{\sqrt{n}} + \frac{\sqrt{\eta T}}{N^{\frac{\alpha}{2}}}\right)^{1/2}\Bigg) ,
\end{aligned}$$

for some constant $C < \infty$, depending on the parameters $\gamma, \kappa, \alpha, L, M$, but not on n or N.

Note that for the sake of clarity and due to space restrictions we only report the leading error terms. A full statement of this Theorem including all lower order terms with it's proof is given in the Appendix, Section 3.

From Theorem 3.4 we can now draw some conclusions. Below, we will give rates of convergence, depending on different a priori assumptions on the hardness of the learning problem.

3.1 Well-specified Case

Here, we give rates of convergence for the most easiest learning problem where our model is well-specified and the regression function lies in the same space as our second-stage SGD iterates, namely in $\mathcal{H}_K$.

Corollary 3.5 (Learning Rates Well-Specified Model)**.** *Suppose all assumptions of Theorem 3.4 are satisfied. Let $r \geq \frac{1}{2}$, $\eta_0 < \frac{1}{4\kappa^2}$ and choose*

$$N_n \geq \log^{2/\alpha}(n)\left(\frac{R^2 n}{M^2}\right)^{\frac{2r+1}{\alpha(2r+\nu)}} .$$

Then, for any n sufficiently large, the excess risk satisfies with probability at least $1 - \delta$ w.r.t. the data D

$$\mathbb{E}_{\hat{D}|D}[||S_K \bar{\tilde{h}}_{T_n} - f_\rho||_{L^2}] \leq C \log(6/\delta) R\left(\frac{M^2}{R^2 n}\right)^{\frac{r}{2r+\nu}} ,$$

for each of the following choices:

1. *One-pass SGD: $b_n = 1$, $\eta_n = \eta_0 \frac{R^2}{M^2}\left(\frac{M^2}{R^2 n}\right)^{\frac{2r+\nu-1}{2r+\nu}}$ and $T_n = \frac{R^2}{M^2} n$,*
2. *Early stopping and one-pass SGD: $b_n = n^{\frac{2r+\nu-1}{2r+\nu}}$, $\eta_n = \eta_0$ and $T_n = \left(\frac{R^2 n}{M^2}\right)^{\frac{1}{2r+\nu}}$,*
3. *Batch-GD: $b_n = n$, $\eta_n = \eta_0$ and $T_n = \left(\frac{R^2 n}{M^2}\right)^{\frac{1}{2r+\nu}}$.*

We comment on these results in Section 4.

3.2 Mis-specified Case

In this subsection we investigate the mis-specified case and further distinguish between two cases:

1. $r \leq \frac{1}{2}$ **but** $2r + \nu > 1$**:** This setting is sometimes called *easy problems*.

2. $r \leq \frac{1}{2}$ **but** $2r + \nu \leq 1$**:** This setting is dubbed *hard problem*, see Pillaud-Vivien et al. (2018).

Corollary 3.6 (Learning Rates Mis-Specified Model; $2r + \nu > 1$)**.** *Suppose all assumptions of Theorem 3.4 are satisfied. Let $r \leq \frac{1}{2}$, $2r + \nu > 1$, $\eta_0 < \frac{1}{4\kappa^2}$ and choose*

$$N_n \geq \log^{2/\alpha}(n)\left(\frac{R^2}{M^2} n\right)^{\frac{2+\nu}{\alpha(2r+\nu)}} . \tag{5}$$

Then, for any n sufficiently large, the excess risk satisfies with probability at least $1-\delta$ *w.r.t. the data D*

$$\mathbb{E}_{\hat{D}|D}[||S_K\bar{\hat{h}}_{T_n} - f_\rho||_{L^2}] \leq C\log(6/\delta)R\left(\frac{M^2}{R^2n}\right)^{\frac{r}{2r+\nu}},$$

for each of the following choices:

1. *Multi-pass SGD:* $b_n = \sqrt{n}$, $\eta_n = \eta_0$ *and* $T_n = \left(\frac{R^2n}{M^2}\right)^{\frac{1}{2r+\nu}}$,
2. *Batch GD:* $b_n = n$, $\eta_n = \eta_0$ *and* $T_n = \left(\frac{R^2n}{M^2}\right)^{\frac{1}{2r+\nu}}$.

Corollary 3.7 (Learning Rates Mis-Specified Model; $2r+\nu \leq 1$)**.** *Suppose all assumptions of Theorem 3.4 are satisfied. Let* $K > 1$, $r \leq \frac{1}{2}$, $2r+\nu \leq 1$, $\eta_0 < \frac{1}{4\kappa^2}$ *and choose*

$$N_n \geq \left(\frac{R^2n}{M^2\log^K(n)}\right)^{\frac{3-2r}{\alpha}}.$$

Then, for any n sufficiently large, the excess risk satisfies with probability at least $1-\delta$ *w.r.t. the data D*

$$\mathbb{E}_{\hat{D}|D}[||S_K\bar{\hat{h}}_{T_n} - f_\rho||_{L^2}] \leq C\log(6/\delta)R\left(\frac{M^2\log^K(n)}{R^2n}\right)^{r},$$

for each of the following choices:

1. $b_n = 1$, $\eta_n = \left(\frac{M^2\log^K(n)}{R^2n}\right)^{2r+\nu}$ *and* $T_n = \left(\frac{R^2n}{M^2\log^K(n)}\right)^{2r+\nu+1}$,
2. $b_n = \left(\frac{R^2n}{M^2\log^K(n)}\right)^{2r+\nu}$, $\eta_n = \eta_0$ *and* $T_n = \frac{R^2n}{M^2\log^K(n)}$,
3. $b_n = \frac{R^2n}{M^2\log^K(n)}$, $\eta_n = \eta_0$ *and* $T_n = \frac{R^2n}{M^2\log^K(n)}$.

Again, we comment on these results in Section 4 in detail.

4 Discussion of Results

We now comment on our results in more detail and also compare, in possible cases, to previous results.

High level comments. Let us briefly describe the nature of our results. In all our bounds above, we are able to establish optimal/ best known rates of convergence if the sample size of the second-stage sample is sufficiently large. In Corollary 3.5, we need

$$N_n \geq \log(n)n^{\frac{2r+1}{\alpha(2r+\nu)}}.$$

While choosing a smaller size comes with computational savings, it would reduce the statistical efficiency. In addition, increasing this number beyond this value would not lead to any gain in statistical accuracy, but would worsen computational requirements. The same phenomenon is observed in Corollary 3.6 and Corollary 3.7.

We also observe an influence of the degree of smoothness of the kernel applied. Choosing a smoother kernel, i.e. a large Hölder index $\alpha \in (0,1]$ reduces the number of samples required, the lowest is achieved for $\alpha = 1$.

Finally, smoother regression functions (corresponding to large r) are easier to reconstruct, i.e. N_n gets smaller for increasing r.

Comparison to one-stage kernel methods. Optimal learning bounds for traditional one-stage regularization (kernel) methods are known under various assumptions. For "easy learning" problems, i.e. if $2r+\nu > 1$, the optimal learning rate is of order $\mathcal{O}(n^{-\frac{r}{2r+\nu}})$ if the amount of regularization is chosen appropriately, see Caponnetto and De Vito (2006), Lin et al. (2020), Blanchard and Mücke (2018). Our results in Corollary 3.5 and Corollary 3.6 match these optimal bounds, provided the number N_n of second-stage samples is chosen sufficiently large, depending on the number of first-stage samples.

For "hard learning" problems, i.e. if $2r+\nu \leq 1$, the best known learning rates for one-stage regularization methods are of order $\mathcal{O}\left(\left(\frac{\log^K(n)}{n}\right)^r\right)$, $K > 1$, see Fischer and Steinwart (2017), Lin et al. (2020), Pillaud-Vivien et al. (2018). Our bounds from Corollary 3.7 also match this bound if N_n is sufficiently large.

Comparison to two-stage KRR. The first paper establishing learning theory for distribution regression using a two stage sampling strategy is Szabó et al. (2016). In this paper, the authors consider a two-stage kernel ridge regression estimator (KRR) and derive optimal rates in the well-specified case $\frac{1}{2} \leq r \leq 1$ if the number of second-stage samples is sufficiently large. More precisely, if

$$N_n \geq \log(n)n^{\frac{2r+1}{\alpha(2r+\nu)}},$$

the rate $\mathcal{O}(n^{-\frac{r}{2r+\nu}})$ given in that paper matches our optimal rate from Corollary 3.5, under the same number N_n. However, for mis-specified models, the results in this paper take not the capacity condition (4) into account[4] and differ from our bounds. If $0 < r \leq \frac{1}{2}$, the rate obtained is $\mathcal{O}(n^{-\frac{r}{r+2}})$ if

$$N_n \geq \log(n)n^{\frac{2(r+1)}{\alpha(r+2)}}.$$

[4]This amounts to considering the worst case with $\nu = 1$.

Compared to our result in Corollary 3.6 with $\nu = 1$, this number is smaller that ours in (5), but it only gives suboptimal bounds. Our result shows that increasing the number of second-stage samples N_n leads to optimal rates also this setting. We also emphasize that KRR suffers from saturation. Using tail-ave SGD instead, we can overcome this issue and establish optimality also for $r \geq 1$.

We also refer to Fang et al. (2020) where for KRR in the well specified case $\frac{1}{2} \leq r \leq 1$, the logarithmic pre-factor for N_n could be removed.

However, for the "hard learning" regime, to the best of our knowledge, no learning rates taking Assumption (4) into account are known for two-stage sampling, except our Corollary 3.7. Thus, we cannot compare our results in this case.

Some additional remarks specific for SGD. Finally, we give some comments specifically related to the SGD algorithm we are applying and compare our results with those known for SGD in the one-stage sampling setting. In all our results we precisely describe the interplay of all parameters guiding the algorithm: batch-size b, stepsize $\eta > 0$ and stopping time T.

All our results show that different parameter choices allow to achieve the same error bound. As noted above, the bound in Corollary 3.5 are mini-max optimal, i.e. there exists a corresponding lower bound (provided the eigenvalues of L_K satisfy a polynomial lower bound $\sigma_j \geq cj^{-1/\nu}$). In addition, these bounds and the parameter choices coincide with those in Mücke et al. (2019). In particular we achieve statistically optimal bounds with a single pass over the data also in the two-stage sampling setting if $f_\rho \in \mathcal{H}_K$, see Corollary 3.5, 1. and 2. . We also recover the known bound for a stochastic version of gradient descent in Corollary 3.5, 3, see Blanchard and Mücke (2018), Lin et al. (2020).

Moreover, as pointed out in Mücke et al. (2019), combining mini-batching with tail-averaging brings some benefits. Indeed, in Lin and Rosasco (2017) it is shown that a large stepsize of order $\log(n)^{-1}$ can be chosen if the mini-batch size is of order $b_n = \mathcal{O}(n^{\frac{2r}{2r+\nu}})$ with a number $\mathcal{O}(n^{\frac{1}{2r+\nu}})$ of passes. Mücke et al. (2019) show that with a comparable number of passes it is allowed to use a larger constant step-size with a much smaller mini-batch size. We observe the same phenomenons in the two-stage sampling setting, provided N_n is sufficiently large. Finally, Corollary 3.5 also shows that increasing the mini-batch size beyond a critical value does not yield any benefit.

However, if $f_\rho \notin \mathcal{H}_K$ we do not achieve the best known bounds with a single pass and multiple passes are necessary. As in the well-specified case, we can achieve these bounds with a large constant stepsize and increasing the mini-batch size beyond a certain value does not yield any benefit. Here, we want to stress once more that our results are the first for distribution regression using SGD and a two-stage sampling strategy.

Acknowledgements

This work is funded by the Deutsche Forschungsgemeinschaft (DFG) under Excellence Strategy *The Berlin Mathematics Research Center MATH+* (EXC-2046/1, project ID:390685689).

The author is also thankful to three anonymous reviewers who gave useful and kind comments.

References

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Dimitri P Bertsekas. A new class of incremental gradient methods for least squares problems. *SIAM Journal on Optimization*, 7(4):913–926, 1997.

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

Andrea Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2006.

Yann Chevaleyre and Jean-Daniel Zucker. Solving multiple-instance and multiple-part learning problems with decision trees and rule sets. application to the mutagenesis problem. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 204–214. Springer, 2001.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *Ann. Statist.*, 44(4):1363–1399, 08 2016. doi: 10.1214/15-AOS1391.

Daniel R Dooly, Qi Zhang, Sally A Goldman, and Robert A Amar. Multiple-instance learning of real-valued data. *Journal of Machine Learning Research*, 3(Dec):651–678, 2002.

Zhiying Fang, Zheng-Chu Guo, and Ding-Xuan Zhou. Optimal learning rates for distribution regression. *Journal of Complexity*, 56:101426, 2020.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. *arXiv preprint arXiv:1702.07254*, 2017.

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008.

Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research 18*, 2017.

Junhong Lin, Raffaello Camoriano, and Lorenzo Rosasco. Generalization properties and implicit regularization for multiple passes sgm. *International Conference on Machine Learning*, 2016.

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *Advances in neural information processing systems*, pages 570–576, 1998.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2), 2017. ISSN 1935-8245. doi: 10.1561/2200000060. URL `http://dx.doi.org/10.1561/2200000060`.

Nicole Mücke and Enrico Reiss. Stochastic gradient descent in hilbert scales: Smoothness, preconditioning and earlier stopping. *stat*, 1050:18, 2020.

Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating sgd saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems*, pages 12568–12577, 2019.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pages 8114–8124, 2018.

Barnabás Póczos, Aarti Singh, Alessandro Rinaldo, and Larry Wasserman. Distribution-free distribution regression. In *Artificial Intelligence and Statistics*, pages 507–515. PMLR, 2013.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, jul 1992. ISSN 0363-0129. doi: 10.1137/0330046. URL `http://dx.doi.org/10.1137/0330046`.

Michael Reed. *Methods of modern mathematical physics: Functional analysis*. Elsevier, 2012.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in neural information processing systems*, pages 1289–1296, 2008.

O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization:convergence results and optimal averaging schemes. In *Proceedings of the 30thInternational Conference on Machine Learning*, 2013.

Steve Smale and Yuan Yao. Online learning algorithms. *Foundations of computational mathematics*, 6(2):145–170, 2006.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

Ingo Steinwart and Clint Scovel. Mercers theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35(3):363–417, 2012.

Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

Pierre Tarres and Yuan Yao. Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9):5716–5735, 2014.

Kiri L Wagstaff, Terran Lane, and Alex Roper. Multiple-instance regression with structured data. In *2008 IEEE International Conference on Data Mining Workshops*, pages 291–300. IEEE, 2008.

Yiming Ying and Massimiliano Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

T. Zhang. Effective dimension and generalization of kernel learning. *Advances in Neural Information Processing Systems 2003*, 2003.

Wang Zhi-Gang, Zhao Zeng-Shun, and Zhang Chang-Shui. Online multiple instance regression. *Chinese Physics B*, 22(9):098702, 2013.

Zhi-Hua Zhou and Jun-Ming Xu. On the relation between multi-instance learning and semi-supervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1167–1174, 2007.