Working Paper. Do not distribute. (2021) RESEARCH NOTE

Sobolev Acceleration and Statistical Optimality for Learning Elliptic Equations via Gradient Descent

Yiping Lu,[⋆,1] Jose Blanchet,[⋆,1,2] and Lexing Ying[⋆,1,3]

[1]ICME, Stanford University, CA, USA
[2]Department of Management Science & Engineering, Stanford University, CA, USA
[3]Department of Mathematics, Stanford University, CA, USA
[⋆]Corresponding author. Email: yplu@stanford.edu; jose.blanchet@stanford.edu; lexing@stanford.edu

Abstract

In this paper, we study the statistical limits in terms of Sobolev norms of gradient descent for solving inverse problem from randomly sampled noisy observations using a general class of objective functions. Our class of objective functions includes Sobolev training for kernel regression, Deep Ritz Methods (DRM), and Physics Informed Neural Networks (PINN) for solving elliptic partial differential equations (PDEs) as special cases. We consider a potentially infinite-dimensional parameterization of our model using a suitable Reproducing Kernel Hilbert Space and a continuous parameterization of problem hardness through the definition of kernel integral operators. We prove that gradient descent over this objective function can also achieve statistical optimality and the optimal number of passes over the data increases with sample size. Based on our theory, we explain an implicit acceleration of using a Sobolev norm as the objective function for training, inferring that the optimal number of epochs of DRM becomes larger than the number of PINN when both the data size and the hardness of tasks increase, although both DRM and PINN can achieve statistical optimality.

Keywords: Kernel Regression, Numerical PDE, Machine Learning, Non-parametric Statistics

1. Introduction

Several learning based methods for solving inverse problems have been proposed recently with state-of-the-art performance across a wide range of tasks, including medical image reconstruction (Ronneberger, Fischer, and Brox 2015), inverse scattering (Khoo, Lu, and Ying 2017) and 3D reconstruction (Sitzmann et al. 2020). In this paper, we study the statistical limit of machine learning methods of solving inverse problems. To be specific, we consider the problem of reconstructing a function from random sampled observations with statistical noise in measurements. We apply gradient descent to a general class of objective functions for the reconstruction. When the observations are the direct observations of the function, the problem is non-parametric function estimation (De Vito et al. 2005; Alexander B Tsybakov 2004). The observations may also come from certain physical laws described by a partial differential equation (PDE) (Stuart 2010; Benning and Burger 2018). Formally, we aim to reconstruct a function f^* based on independently sampled data set $D = \{(x_i, y_i)\}_{i=1}^n$ from an unknown distribution P on $\mathcal{X} \times \mathcal{Y}$, where y_i is the noisy measurement of u^* though a measurement procedure $\mathcal{A}$. For simplicity, we assume $\mathcal{A}$ is self-adjoint in this paper. The conditional mean function $f^*(x) = \mathbb{E}_P(Y|X = x)$ is the ground truth function for observation of u^* through the measurement procedure $\mathcal{A}$, *i.e.* $f^* = \mathcal{A}u^*$. To solve this problem, we consider gradient descending over the following general class of objective function

$$\hat{u} = \arg\min_{u\in\mathcal{H}} \mathbb{E}_{\mathbb{P}_n(x,y)} \frac{1}{2}\left\langle u(x), \mathcal{A}_1 u(x)\right\rangle - \left\langle y, \mathcal{A}_2 u(x)\right\rangle,$$

where $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta(x_i, y_i)$ is the empirical distribution, $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) and $\mathcal{A}_i, i = 1, 2$ are two self-adjoint operators that satisfy $\mathcal{A}_1 = \mathcal{A}\mathcal{A}_2$. In Section 2, we show that several algorithms, including kernel regression (De Vito et al. 2005; Caponnetto and De Vito 2007) via Sobolev training (Czarnecki et al. 2017; Richardson Jr 2008; Son et al. 2021) and solving PDEs via machine learning based algorithm, (Raissi, Perdikaris, and Karniadakis 2019; Sirignano and Spiliopoulos 2018; Khoo, Lu, and Ying 2017; E and Yu 2018) can be considered as special cases of this formulation.

Recent works (Nickl, Geer, and Wang 2020; Lu et al. 2021) have considered the statistical limit of learning of elliptic inverse problem, *i.e.* how many observation of the right hand side function of an elliptic PDE are needed to reach a prescribed performance level. However, none of these papers consider computationally feasible methods for constructing such optimal estimators. In this paper, we consider the statistical optimality of gradient descent (Lin and Rosasco 2017; Pillaud-Vivien, Rudi, and Bach 2018; Lin et al. 2020; Marteau-Ferey et al. 2019), a successful and widely used algorithm in machine learning. We show that proper early stopped gradient descent can achieve information theoretical optimal convergence rate according to a continuous scale of suitable Hilbert norm (*i.e.* Sobolev norms(Fischer and Steinwart 2020; Liu and Li 2020), detailed definition see Section 2).

We first proof that properly early stopped gradient descent over the class of objective functions can achieve statistical optimality. At the same time, although all the gradient flow of the class of loss function can achieve statistical optimality according to our theory, we discover an acceleration effect of using Sobolev norm as loss function for kernel based machine learning algorithms. The implicit acceleration of Sobolev loss function arises because a differential operator can enlarge the small eigenvalue of kernel integral operator for high frequency functions, leading to better condition numbers and faster convergence in these eigenspaces while keeping the statistical optimality. We justify our theoretical finding with several numerical experiments.

1.1 Related Works

Machine Learning Based PDE Solver. Partial differential equations (PDEs) are widely used in many disciplines of science and engineering and play a prominent role in modeling and forecasting the dynamics of multiphysics and multiscale systems. The recent deep learning breakthrough and the rapid development of sensors, computational power, and data storage in the past decade has drawn attention to numerically solving PDEs via machine learning methods (Long et al. 2018; Long, Lu, and Dong 2019; Raissi, Perdikaris, and Karniadakis 2019; Han, Jentzen, and Weinan 2018; Sirignano and Spiliopoulos 2018; Khoo, Lu, and Ying 2017), especially in high dimensions where conventional methods become impractical. Based on the natural idea of representing solutions of PDEs by (deep) neural networks, different loss functions for solving PDEs are proposed. (Han, Jentzen, and Weinan 2018; Han, Lu, and Zhou 2020) utilize the Feynman-Kac formulation which turns solving PDE to a stochastic control problem. The weak adversarial network (Zang et al. 2020) solves the weak formulations of PDEs via an adversarial network. In this paper, we focus on the convergence rate of the Deep Ritz Method (DRM) (E and Yu 2018; Khoo, Lu, and Ying 2017) and the Physics-Informed neural network (PINN) (Raissi, Perdikaris, and Karniadakis 2019; Sirignano and Spiliopoulos 2018). DRM (E and Yu 2018; Khoo, Lu, and Ying 2017) utilizes the variational structure of the PDE, which is similar to the Ritz-Galerkin method in classical numerical analysis of PDEs, and trains a neural network to minimize the variational objective. PINN (Raissi, Perdikaris, and Karniadakis 2019; Sirignano and Spiliopoulos 2018) trains a neural network directly to minimize the residual of the PDE, i.e., using the strong form of the PDE. Theoretical convergence results for deep learning based

PDE solvers has also received considerable attention recently. Specifically, (Lu, Lu, and Wang 2021; Grohs and Herrmann 2020; Marwah, Lipton, and Risteski 2021; Wojtowytsch et al. 2020; Xu 2020; Shin, Zhang, and Karniadakis 2020; Bai et al. 2021) investigated the regularity of PDEs approximated by a neural network and (Lu, Lu, and Wang 2021; Luo and Yang 2020; Duan et al. 2021; Jiao, Lai, Li, et al. 2021; Jiao, Lai, Luo, et al. 2021) further provided generalization analyses. (Nickl, Geer, and Wang 2020; Lu et al. 2021; Hütter and Rigollet 2019; Manole et al. 2021) provided information theoretical optimal lower and upper bounds for solving PDEs from random samples. However, all these papers assume accessibility of the global solution of empirical loss minimization. In contrast, here we consider the gradient descent algorithm for learning the estimator. The most relevant work in connection to is (Nickl and Wang 2020), which considers a polynomial-time Langevin-type algorithms to sample from the posterior measure of the Bayesian inverse methods. Instead of considering the Bayesian setting, here we optimize on the un-regularized objective. However, the estimator is regularized via early stopping (Yao, Rosasco, and Caponnetto 2007; Ali, Kolter, and Tibshirani 2019; Ali, Dobriban, and Tibshirani 2020), *i.e.* we consider the statistical optimality of the implicit regularization effect of optimization algorithm.

Learning with kernel. Supervised least square regression in RKHS has a long history and its generalization ability and mini-max optimality has been thoroughly studied (Caponnetto and De Vito 2007; Smale and Zhou 2007; De Vito et al. 2005; Rosasco, Belkin, and De Vito 2010; Mendelson and Neeman 2010). Statistical optimality of early stopped (stochastic) gradient descent has been widely discussed in (Yao, Rosasco, and Caponnetto 2007; Dieuleveut and Bach 2016; Polyak and Juditsky 1992; Pillaud-Vivien, Rudi, and Bach 2018; Lin and Rosasco 2017; Wei, Yang, and Wainwright 2017; Lei, Hu, and Tang 2021). The convergence of least square regression in Sobolev norm has been discussed recently in (Fischer and Steinwart 2020; Liu and Li 2020). Recently, training neural networks with stochastic gradient descent in certain regimes has been found to be equivalent to kernel regression (Daniely 2017; Lee et al. 2017; Jacot, Gabriel, and Hongler 2018). Gradient descent training of neural network in the kernel regime has been found optimal for non-parametric of a wide class of functions with both early stopping regularization and ridge regression (Nitanda and Suzuki 2020; Hu et al. 2020).

1.2 Contribution

- We provide information theoretical lower bounds (Theorem 1) for a wide class of inverse problems, including the Sobolev learning rate (Fischer and Steinwart 2020) for the solution of elliptic inverse problems. We also show that the previous lower bound (Nickl, Geer, and Wang 2020; Lu et al. 2021) for machine learning solving elliptic equations can be considered as a special case of our lower bound.
- We provide a proof of statistical optimality of the gradient descent algorithm of a general class of objective functions (Theorem 2), including PINN (Raissi, Perdikaris, and Karniadakis 2019; Sirignano and Spiliopoulos 2018) and Deep Ritz Methods (E and Yu 2018; Khoo, Lu, and Ying 2017) for solving PDEs as well as Sobolev training (Son et al. 2021; Czarnecki et al. 2017; J. Yu et al. 2021) of kernel methods. We provide (Lu et al. 2021) a computational feasible estimator and generalize the previous statistical optimality results of gradient descent (Yao, Rosasco, and Caponnetto 2007; Pillaud-Vivien, Rudi, and Bach 2018; Lin et al. 2020) to general Sobolev norm.
- We also characterize the acceleration effect of Sobolev loss function for learning with kernel. The acceleration happens because differential operator can enlarge the small eigenvalues for high frequency functions, leading to better condition number and faster convergence in these eigenspaces while keeping the statistical optimality. Thus when the target function have more high frequency component, the lead of PINN will become larger (Figure 3). We justify our

theoretical finding with several numerical experiments (Figure 2 and Figure 4).

2. Problem Formulation

In this section, we formulate the problem of learning inverse problem using the kernelized gradient descent. As described previously, we aim to reconstruct a function $f^* \in \mathbb{R}^{\mathcal{X}}$ from random observations of $u^* = \mathcal{A}f^*$, where $\mathcal{A}$ is an observation process which is modeled by an operator maps from $\mathbb{R}^{\mathcal{X}}$ to $\mathbb{R}^{\mathcal{X}}$. To solve this problem, we write the operator $\mathcal{A}$ in terms of two operators $\mathcal{A}_i$ $(i = 1, 2)$ with $\mathcal{A}_1 = \mathcal{A}\mathcal{A}_2$ and build our objective function as

$$\mathbb{E}_{\mathbb{P}}\left[\frac{1}{2}\left\langle u(x), \mathcal{A}_1 u(x)\right\rangle - \left\langle y, \mathcal{A}_2 u(x)\right\rangle\right], \tag{1}$$

where $\mathbb{P}$ is the joint distribution of x and y with x sampled from the uniform distribution on $\mathcal{X}$ for simplicity and y as the noisy observation of $f(x) = (\mathcal{A}u)(x)$. In other words, $\mathbb{E}(y|x) = f(x)$. The minimizer of objective function (1) is the ground truth function $u^* = \mathcal{A}^{-1}f$ that we are interested in.

Learning with Kernel Consider the case that u is parameterized by a Reproducing Kernel Hilbert Space $u_\theta(x) = \langle \theta, K_x \rangle$ (we provide standard notations of RKHS in Appendix Appendix 1). At the same time, the kernel function has the following representation $K(s,t) = \sum_{i=1}^{\infty} \lambda_i e_i(s) e_j(t)$, where e_i are orthogonal basis of $\mathcal{L}_2(\rho_{\mathcal{X}})$ with $\rho_{\mathcal{X}}$ being the uniform distribution over $\mathcal{X}$. Then e_i is also the eigenvector of the covariance operator $\Sigma = \mathbb{E}_{x\sim\mathbb{P}} K_x \otimes K_x$ with eigenvalue $\lambda_i > 0$, *i.e.* $\Sigma e_i = \lambda_i e_i$. Here $g \otimes h = gh^\top$ is an operator from $\mathcal{H}$ to $\mathcal{H}$ defined as

$$g \otimes h : f \to \langle f, h \rangle_{\mathcal{H}} g.$$

The covariance matrix Σ is the core of the integral operator technique (Smale and Zhou 2007; Caponnetto and De Vito 2007) for kernel regression. For any $f \in \mathcal{H}$, the reproducing property gives

$$(\Sigma f)(z) = \left\langle K_z, \Sigma f \right\rangle_{\mathcal{H}} = \mathbb{E}[f(X)k(X,z)] = \mathbb{E}[f(X)K_x(X)].$$

If we consider the mapping $S : \mathcal{H} \to L_2(dx)$ defined as a parameterization of a vast class of functions in $\mathbb{R}^{\mathcal{X}}$ via $\mathcal{H}$ through the mapping $(Sg)(x) = \langle g, K_x \rangle$ $(\Phi(x) = K_x = K(\cdot, x))$. Its adjoint operator $S^* : \mathcal{L}_2 \to \mathcal{H}$ then can be defined as $g \to \int_{\mathcal{X}} g(x) K_x \rho_X(dx)$. Σ is the same as the self-adjoint operator S^*S and the self-adjoint operator $\mathcal{L} = SS^* : L_2 \to L_2$ can be defined as

$$(\mathcal{L}f)(x) = (SS^*f)(x) = \int_{\mathcal{X}} K(x,z) f(z) \rho_{\mathcal{X}}(dz).$$

Based on this notation, we present all our assumptions on the underlying kernel.

Assumption 1 (Assumptions on Kernel). *We assume the standard capacity condition on kernel covariance operator with a source condition about the regularity of the target function following (Caponnetto and De Vito 2007). We further assume a regularity condition for our kernel $k(\cdot,\cdot)$ via a ℓ_∞ embedding property follows (Steinwart, Hush, Scovel, et al. 2009; Dicker, Foster, and Hsu 2017; Pillaud-Vivien, Rudi, and Bach 2018; Fischer and Steinwart 2020). These conditions are stated explicitly below.*

- ***(a) Standard assumptions.*** *The kernel feature are bounded almost surely,* i.e. $|k(x,y)| \leq R$ *and the observation y is also bounded by M almost surely.*
- ***(b) Capacity condition.*** *Consider the spectral representation of the kernel covariance operator $\sigma = \sum \lambda_i e_i \otimes e_i$, we assume polynomial decay of eigenvalues of the covariance matrix $\lambda_i \propto i^{-\alpha}$ for some $\alpha > 1$. As a result $Q = tr(\Sigma^{1/\alpha}) < \infty$.*

- ***(c) Source condition.*** *We also impose an assumption on the smoothness of the true function. There exists* $\beta \in (0,1]$ *such that* $u^* = \mathcal{L}^{\beta/2}\phi$ *for some* $\phi \in L^2$. *If* $u^*(x) = \langle \theta_*, K_x \rangle_{\mathcal{H}}$, *the source condition can also be written as*
$$\|\Sigma^{\frac{1-\beta}{2}}\theta_*\|_{\mathcal{H}} < \infty.$$
- ***(d) Capacity conditions on*** $\mathcal{A}_i$. *For theoretical simplicity, we assume that the self-adjoint operators* $\mathcal{A}_i$ *are diagonalizable in the same orthonormal basis* e_i. *Thus we can assume*
$$\mathcal{A}_1 = \sum_{i=1}^{\infty} p_i e_i \otimes e_i, \mathcal{A}_2 = \sum_{i=1}^{\infty} q_i e_i \otimes e_i$$
for positive constants $p_i, q_i > 0$. *We further assume* $p_i \propto i^{-p}$ *and* $q_i \propto i^{-q}$. *This commuting assumptions also made in (Cabannes et al. 2021; Hoop et al. 2021). due to the Bochner's theorem. We further assume* $p < 0, q < 0, \alpha + p > 0$. *We refer the detailed discussion to Remark 1.*
- ***(e) Regularity results on RKHS.*** *For* $\mu \in [0,1]$, *there exists* $\kappa_\mu \geq 0$ *such that* $\Phi(x) \otimes \Phi(x) \leq k_\mu^2 R^{2\mu}\Sigma^{1-\mu}$ *holds almost surely. The regularity assumption here is equivalent to* $\|g\|_{L_\infty}^2 \leq \kappa_\mu^2 R^{2\mu} \|\Sigma^{1/2-\mu/2} g\|_{\mathcal{H}}^2$ *and implies* $\|g\|_{L_\infty} \leq \kappa_\mu R^\mu \|g\|_{\mathcal{H}}^\mu \|g\|_{L_2}^{1-\mu}$ *for every* $g \in \mathcal{H}$. *As a consequence, we know that* $\|\Sigma^{\mu/2-1/2}\Phi(x)\|_{\mathcal{H}} \leq \kappa^\mu R^\mu$ *holds almost surely. (Steinwart, Hush, Scovel, et al. 2009; Fischer and Steinwart 2020; Pillaud-Vivien, Rudi, and Bach 2018)*

Remark 1. *To simplify the technical exposition, we assume that operator* $\mathcal{A}_i (i = 1,2)$ *commute with the kernel covariance operator* Σ. *This assumption is also made in (Hoop et al. 2021; Cabannes et al. 2021). Here we provide several examples that satisfy this assumption. The simplest case is* $\mathcal{A}_1 = \mathcal{A}_2 = id$, *which gives rise to the function regression setting. For numerically solving a PDE, we take* $\mathcal{A}_i$ *to become the power of the Laplace operator* Δ. *If the domain is a sphere, the eigen-functions are spherical harmonics which are also the eigen-functions of a wide class of kernels, examples includes the dot product kernels (Scetbon and Harchaoui 2021) and the Neural Tangent Kernel (Bietti and Bach 2020; Chen and Xu 2020), when the data distribution is uniform distribution. When the domain is the torus, the eigen-functions are Fourier modes. If we consider a shift invariant kernel* $K(x,y) = \psi(x-y)$, *from Bochner's Theorem* $K(x,y) = \sum_{i=1}^n \tilde{\psi}(w) e^{iws} e^{-iwt}$ *we know that the eigen-functions are also Fourier modes. There are also works that use Green function as the kernel (Zhou and Belkin 2011; Fasshauer and Ye 2011), where the three operators will automatically commute with each other.*

In this paper, we consider the convergence of the estimator in Sobolev norm class. We define the different Sobolev spaces via the power space approaches used in (Steinwart and Scovel 2012; Fischer and Steinwart 2020).

Definition 1 (Sobolev Norm). *For* $\gamma > 0$, *the* γ-power space *is*
$$\mathcal{H}^\gamma := \left\{ \sum_{i\geq 1} a_i \lambda_i^{\gamma/2} e_i : \sum_{i \geq 1} a_i^2 \leq \infty \right\} \subset L_2(\nu),$$
equipped with the γ*-power norm via* $\|\sum_{i\geq 1} a_i \lambda_i^{\gamma/2} e_i\|_\gamma := \left(\sum_{i\geq 1} a_i^2\right)^{1/2}$.

It is obvious that $\|\mathcal{L}^{\gamma/2} f\|_\gamma = \|f\|_{L_2}$ and $\|f\|_\gamma \leq \|\Sigma^{\frac{1-\gamma}{2}} f\|_{\mathcal{H}}$ (Fischer and Steinwart 2020). The source condition can also be understood as the target function u^* lies in the β-power Sobolev space. The regularity condition of the kernel function implies a continuously embedding from $\mathcal{H}^\gamma \to L_\infty$. Throughout this paper, we consider the convergence rate of $\hat{u} - u^*$ in γ-power Sobolev norm ($\gamma > 0$).

2.1 *Examples*

Sobolev Training (Shi, Guo, and Zhou 2010; Czarnecki et al. 2017; Son et al. 2021) introduce the idea of training using Sobolev spaces via matching not only the function value but also the derivative of the classifier. Using different Sobolev norms as loss function has also been used widely in image processing, inverse problems, and graphics applications (Yang, Hu, and Lou 2021; Calder, Mansouri, and Yezzi 2010; Richardson Jr 2008; Yu, Schumacher, and Crane 2021; C. Yu et al. 2021; Soliman et al. 2021). The work of (Yang, Hu, and Lou 2021) discovered that different Sobolev loss functions would lead to different implicit bias and that the proper Sobolev preconditioned gradient descent can accelerate the optimization of geometry objectives (Yu, Schumacher, and Crane 2021; C. Yu et al. 2021; Soliman et al. 2021). In this paper, we discover that stochastic gradient descent over Sobolev norm loss class functions can achieve statistical optimal but proper selection of the Sobolev norm loss function can accelerate training. We call this phenomenon **Sobolev Implicit Acceleration** and discuss it in Section 4.

Machine Learning Based PDE Solver. To simplify the exposition, we focus on a prototype elliptic PDE: Poisson's equation on a torus, *i.e.* $\Omega = \mathbb{T}^d = [0,1]^d_{\text{per}}$. Our focus is on the analysis of deep-learning-based numerical methods for the elliptic equations

$$-\Delta u = f \quad \text{in } \Omega. \tag{2}$$

We mainly focus on analyzing Deep Ritz Method (DRM) (E and Yu 2018) and Physics Informed Neural Network (PINN) (Raissi, Perdikaris, and Karniadakis 2019; Sirignano and Spiliopoulos 2018). DRM solves the equation (2) via minimizing the following variational form

$$u^* = \arg\min_{u\in\mathcal{F}} \mathcal{E}^{\text{DRM}}(u) := \frac{1}{2}\int_\Omega |\nabla u|^2\, dx - \int_\Omega f u dx, \tag{3}$$

while PINNs solves the equation (4) via minimizing the following strong formula, *i.e* the residual of the PDE,

$$u^* = \arg\min_{u\in\mathcal{F}} \mathcal{E}^{\text{PINN}}(u) := \frac{1}{2}\int_\Omega (\Delta u - f)^2\, dx, \tag{4}$$

where u is minimized over a parameterized function class $\mathcal{F}$ (for example neural network). Here we consider the function class to be the RKHS space (Chen et al. 2021). (Lu et al. 2021) showed that empirical risk minimization of both objectives can achieve information theoretical optimal bounds. The objective function in 3 and 4 can be considered as special case of objective function (1). For DRM, $\mathcal{A}_1 u = \Delta u$ and $\mathcal{A}_2 u = u$ for all function $u \in \mathbb{R}^{\mathcal{X}}$. For PINN, $\mathcal{A}_1 u = \Delta^2 u$ and $\mathcal{A}_2 u = \Delta u$ for all function $u \in \mathbb{R}^{\mathcal{X}}$.

We discover that PINN convergences faster than DRM consistently due to the implicit Sobolev acceleration, matching the observation made in (Chen, Du, and Wu 2020). (Cabannes et al. 2021) considered semi-supervised learning using Laplacian regularization with kernel parameterization. However, this paper does not consider training with stochastic gradient descent and also does not introduce the source condition assumption that leads to different convergence rate for a hierarchical parameterization of task difficulty.

3. Main Theorem

We present our main results in this section, including an information theoretical lower bound and a matching upper bound with proper selected early stopping time.

3.1 Lower Bounds

This subsection investigates the statistical optimality of the Sobolev convergence rate of solving elliptic problem using stochastic gradient descent. We provide the information theoretical lower bound of learning the elliptic problems. Different from (Nickl, Geer, and Wang 2020; Lu et al. 2021), we formulate the problem in an RKHS. This leads to a different construction of hypothesis and show that (Nickl, Geer, and Wang 2020; Lu et al. 2021) is a special case of our lower bound using specific kernel and operator $\mathcal{A}_i (i = 1, 2)$ in Section 3.3.

Theorem 1 (Lower Bound). *Let (X, B) be a measurable space, H be a separable RKHS on X with respect to a bounded and measurable kernel k and operator $\mathcal{A} = (\mathcal{A}_2^{-1}\mathcal{A}_1)$ satisfies Assumption 1. We have n i.i.d. random observations $\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1}^n$ of $f^* = \mathcal{A}u, u \in \mathcal{H}^\gamma \cap L_\infty$, i.e. $y_i = f^*(x_i) + \eta_i$ where η_i is a mean zero random noise satisfies the momentum assumption $\mathbb{E}|\eta|^m \le \frac{1}{2}m!\sigma^2 L^{m-2}$ for some constants $\sigma, L > 0$. Then for all estimators $H : (\mathcal{X} \times \mathcal{Y})^{\otimes n} \to \mathcal{H}^\gamma$ satisfies*

$$\inf_H \sup_{u^*} \mathbb{E}\|H(\{(x_i, y_i)\}_{i=1}^n) - u^*\|_\gamma^2 \gtrsim n^{-\frac{(\max\{\beta,\mu\}-\gamma)\alpha}{\max\{\beta,\mu\}\alpha+2(q-p)+1}}.$$

3.2 Upper Bounds

This subsection, we consider the (multiple pass) gradient descent over the empirical data of objective function (1). We aim to construct our estimator via optimizing the empirical loss function

$$\sum_{i=1}^n \frac{1}{2} u(x_i)\mathcal{A}_1 u(x_i) - y_i \mathcal{A}_2 u(x_i),$$

where x_i is sampled randomly and y_i is the associated noisy observation introduced in Section 1. We consider a parameterization $u(x) = \langle u, K_x\rangle$ and $\mathcal{A}_i u(x) = \langle \mathcal{A}_i\theta, K_x\rangle_{\mathcal{H}} = \langle \theta, \mathcal{A}_i K_x\rangle_{\mathcal{H}}$ and express our empirical objective function as

$$\begin{aligned} &\mathbb{E}_{\mathbb{P}_n(x,y)} \frac{1}{2}\langle u(x), \mathcal{A}_1 u(x)\rangle - \langle y, \mathcal{A}_2 u(x)\rangle \\ &= \mathbb{E}_{\mathbb{P}_n(x,y)} \frac{1}{2}\langle u, K_x\rangle \langle \mathcal{A}_1 u, K_x\rangle - y\langle \mathcal{A}_2 u, K_x\rangle \\ &= \mathbb{E}_{\mathbb{P}_n(x,y)} \frac{1}{2}\langle u, K_x \otimes \mathcal{A}_1 K_x u\rangle - y\langle u, \mathcal{A}_2 K_x\rangle \end{aligned} \tag{5}$$

Then the gradient descent algorithm can be written as the following procedure:

- **Initialization**: $\theta_0 = \bar{\theta}_0 = 0$, γ is a constant to be determined later which is used as the learning rate in the algorithm.
- **Iteration**: For the t–th iteration, we perform the following gradient descent step

$$\theta_t = \theta_{t-1} + \gamma\frac{1}{n}\sum_{i=1}^n \left(y_i \mathcal{A}_2 K_{x_i} - \langle \theta_{t-1}, \mathcal{A}_1 K_{x_i}\rangle_{\mathcal{H}} K_{x_i}\right)$$

  with an averaging step $\bar{\theta}_t = (1 - \frac{1}{t})\bar{\theta}_{t-1} + \frac{1}{t}\theta_t$.

Remark. Note that the optimizing dynamics considered here is not the exacting gradient descent dynamics over the empirical objective. The gradient of the quadratic term $\frac{1}{n}\sum_{i=1}^n u(x_i)\mathcal{A}_1 u(x_i)$ should be $\frac{1}{n}\sum_{i=1}^n \left(\langle \theta_{t-1}, \mathcal{A}_1 K_{x_i}\rangle_{\mathcal{H}} K_{x_i} + \langle \theta_{t-1}, K_{x_i}\rangle_{\mathcal{H}} \mathcal{A}_1 K_{x_i}\right)$ but we take instead $\frac{1}{n}\sum_{i=1}^n \langle \theta_{t-1}, \mathcal{A}_1 K_{x_i}\rangle_{\mathcal{H}} K_{x_i}$

in our dynamics. In the population expectation, the two dynamics are the same due to the commuting assumption between the kernel integral operator and operator $\mathcal{A}_1$. The implementation of our dynamics can be applied via a stop gradient operator on the $\mathcal{A}_1 u$ during back propagating. The slight variation of gradient descent considered here facilitates the technical analysis.

The following theorem is the main result for upper bounds with the proof details given in the appendix.

Theorem 2. *Under Assumption 1, we have the following three regimes shown in Figure 1.*

- *For* $\beta > \frac{\alpha+2q-p-1}{\alpha}$, *if we take* $t = n$ *and* $\boxed{\gamma = n^{\frac{\alpha+p}{\beta\alpha+2(p-q)+1}-1}}$, *we obtain the following rate*
$$\mathbb{E}[\|\bar{\theta}_t - u^*\|_\gamma^2] = O(n^{-\frac{(\beta-\gamma)\alpha}{\alpha\beta+2(p-q)+1}}).$$
- *For* $\frac{\alpha+2q-p-1}{\alpha} \leq \beta \leq \frac{\mu\alpha+2q-p+1}{\alpha}$, *if we take* $\boxed{t = n^{\frac{\alpha+p}{\beta\alpha+2(p-q)+1}}}$ *and* γ *a small enough constant, we obtain the following rate*
$$\mathbb{E}[\|\bar{\theta}_t - u^*\|_\gamma^2] = O(n^{-\frac{(\beta-\gamma)\alpha}{\alpha\beta+2(p-q)+1}}).$$
- *For* $\beta > \frac{\mu\alpha+2q-p+1}{\alpha}$, *if we take* $\boxed{t = n^{\frac{\alpha+p}{\mu\alpha+p}}}$ *and* γ *a small enough constant, we obtain the following rate*
$$\mathbb{E}[\|\bar{\theta}_t - u^*\|_\gamma^2] = O(n^{-\frac{(\beta-\gamma)\alpha}{\mu\alpha+p}}),$$
which is not an optimal converging rate.

Sketch of the Proof. We first rewrite the averaged gradient descent in a more compact formula as $\eta_0 = 0, \eta_u = \eta_{u-1} + \gamma(\mathcal{A}_2^\top \hat{S}_n^* \hat{y} - \hat{\Sigma}_{Id,\mathcal{A}_1}\eta_{t-1})$ where $\hat{S}_n : \mathcal{H} \to \mathbb{R}^n$ is defined as $\hat{S}_n g = \frac{1}{\sqrt{n}}(g(x_1), \cdots, g(x_n))$, $\hat{\Sigma}_{\mathcal{O}_1,\mathcal{O}_2} = \frac{1}{n}\sum_{i=1}^n \mathcal{O}_1 K_x \otimes \mathcal{O}_2 K_x$ and Id is the identity operator. For the error of GD, we consider early stopping of gradient descent algorithm as a spectral filtering (Gerfo et al. 2008; Pillaud-Vivien, Rudi, and Bach 2018; Blanchard and Mücke 2018; Lin et al. 2020). Our proof is based on standard bias-variance decomposition. For t iteration, GD will behave similarly to ridge regression with γt regularization strength (Yao, Rosasco, and Caponnetto 2007; Pillaud-Vivien, Rudi, and Bach 2018) and this result in bias of $(\frac{1}{\gamma t})^{\frac{(\beta-\gamma)\alpha}{\alpha+p}}$. For the variance, we provide a bound which is related to the effective dimension given by $\mathrm{tr}((\Sigma_{Id,\mathcal{A}_1} + (\frac{1}{\gamma t})I)^{-1}\Sigma_{\mathcal{A}_2^\top \mathcal{A}_2})$ and obtain a final variance of the form $\frac{1}{n}(\gamma t)^{-\frac{\gamma\alpha+p}{\alpha+p}}(\frac{1}{\gamma t})^{-\frac{1}{\alpha+p}}(\frac{1}{\gamma t})^{-\frac{p-2q}{\alpha+p}} + \frac{1}{n}(\frac{1}{\gamma t})^{-\frac{\gamma\alpha+p}{\alpha+p}}(\frac{1}{\gamma t})^{-\frac{\mu\alpha-p}{\alpha+p}}(\frac{1}{\gamma t})^{\frac{\beta\alpha-2q}{\alpha+p}}$. If we only have the first term of variance, we shall achieve information theoretical optimal bound when $\boxed{t = n^{\frac{\alpha+p}{\beta\alpha+2(p-q)+1}}}$. For the section term in the variance is from the convergence of empirical covariance matrix $\hat{\Sigma}_{Id,\mathcal{A}_1}$ to the population one $\Sigma_{Id,\mathcal{A}_1}$. This term can be reduced using semi-supervised learning techniques as in (Murata and Suzuki 2021; Lu et al. 2021).

3.3 Discussion and Implication of Our Theory

Relationship with (Nickl, Geer, and Wang 2020; Lu et al. 2021). (Nickl, Geer, and Wang 2020; Lu et al. 2021) provided a lower bound of the form $n^{-\frac{2\alpha-2s}{2\alpha-4t+d}}$ for a $2t$–th order linear PDE $\Delta^t u = f$ with solution in H^α, evaluated in H^s norm. We shall discuss the relationship between their bound with our $n^{-\frac{(\beta-\gamma)\alpha}{\beta\alpha+2(p-q)+1}}$ lower bound based on the kernel representation of Sobolev spaces. The numerator

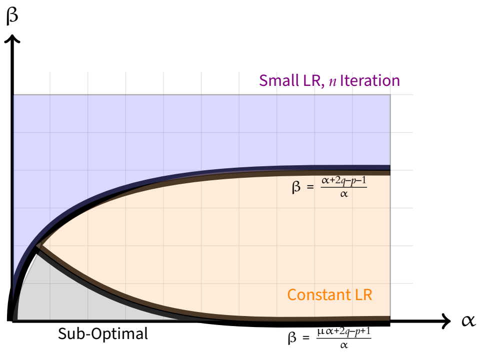


Figure 1. Phase diagram of different regimes for solving inverse problem using stochastic gradient descent. Except the gray regime, GD can achieve information theoretical rate.

$(\beta - \gamma)$ matches the $\alpha - s$ term in (Nickl, Geer, and Wang 2020; Lu et al. 2021)'s lower bound and the $q - p$ term is the order of the linear PDE which matches the t term in the denominator in (Nickl, Geer, and Wang 2020; Lu et al. 2021)'s lower bound. The spectral decay speed of kernel α is always relative to the dimension d. To understand this problem, we consider the following two examples.

For the first example, the kernel is defined on the torus $\mathbb{T}^d = [0, 1]^d_{\text{per}}$. We consider the space of square integrable functions on $\mathbb{T}^d$ with mean 0 and the Matérn kernel $K_{\sigma,l,\nu}(x, y) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{|x-y|}{l}\right)^\nu B_\nu(\frac{|x-y|}{l})$, where B_ν is the modified Bessel function of section kind. The covariance operator is $C_\theta = \sigma^2(-\Delta + \tau^2 I)^{-s}$ with orthonormal eigenfunctions $\phi_m(x) = e^{2\pi i \langle m,x \rangle}$ and corresponding eigenvalues $\lambda_m = \sigma^2(4\pi^2|m|^2 + \tau^2)^{-s}$ for every $m \in \mathbb{Z}^d \backslash \{0\}$ (Stein 1999).

For the second example, we consider the Mercer's decomposition of a translation invariant kernel via Fourier series $K(s-t) = \frac{1}{2\pi} \sum_w \tilde{K}(w) e^{iw(s)} e^{iw(-t)} dw$. The eigenfunctions of the translation invariant kernel is the Fourier modes and the eigenvalues are the Fourier coefficients. As an example, for Neural Tangent Kernel, (Cao and Gu 2019; Chen and Xu 2020; Bietti and Bach 2020; Nitanda and Suzuki 2020) proved that the corresponding $\alpha = \frac{d}{d-1}$ and the eigenfunctions are spherical harmonics that diagonalize the differential equation.

For the upper bound, (Lu et al. 2021) established the convergence rate based on the *empirical process* technique (Mendelson and Neeman 2010; Steinwart, Hush, Scovel, et al. 2009), while our paper switches to the integral operator/inverse problem technique (De Vito et al. 2005; Smale and Zhou 2007; Caponnetto and De Vito 2007). An advantage of the integral operator/inverse problem technique is that it can provide convergence results with respect to a continuous scale of Sobolev norms while the empirical process technique can only be used for the Sobolev norm equivalent to the objective function.

Relationship with (Shi, Guo, and Zhou 2010) (Shi, Guo, and Zhou 2010) also considered learning from data involving function value and gradients under the framework of least-square regularized regression in reproducing kernel Hilbert spaces. In this paper, we only have access to the noisy observation of the function values but still aim to know about the convergence rate with respect to the Sobolev norm. At the same time, we further consider an inverse problem setting with an early stopping regularization, which is not discussed in (Shi, Guo, and Zhou 2010). However, we introduce a commuting assumption over the differential operator with the kernel integral operator that makes the problem easier.

Sobolev Implicit Acceleration Below we discuss the implication of the choice of early stopping time $t = n^{\frac{\alpha+p}{\beta\alpha+2(p-q)+1}}$. First of all, the best early stopping time here does not depend on γ, which means the best model in different Sobolev is the same over the stochastic gradient descent path asymptotically. Secondly, all the components in an iteration step depend on the problem itself except the numerator $\alpha + p$. For differential operators, the p is actually ***negative*** (differential operators have large eigenvalues over high-frequency basis). Thus we can accelerate the training via letting p more negative, *i.e.* using a higher order Sobolev norm as loss can lead to earlier stopping. As an implication, the PINN achieves the statistical optimal solution faster than DRM.

Relationship with implicit bias of frequency Recent work credit the success of deep learning to the fast training in low frequency components (Rahaman et al. 2019; Xu et al. 2019; Kalimeris et al. 2019). However, in our work, with Sobolev preconditioning, the training speed of high frequency part increases, yet achieving statistical optimality in the class of Sobolev norm. This suggests that the implicit bias of frequency is not necessary for good generalization results. We also would like refer to (Amari et al. 2020; Mücke and Reiss 2020) Theorem 8 for the extreme case, where the authors directly invert the population covariance matrix which leads to the same training speed in every eigen-spaces while still maintaining the statistical optimality in ℓ_2 norm. However the preconditioning matrix in (Amari et al. 2020) is the population Fisher information matrix, which requires further sampling of unlabeled data that is not accessible in our setting.

Discussion of the Sub-Optimal Regime In the sub-optimal regime, the concentration error between the empirical covariance matrix $\hat{\Sigma}_{Id,\mathcal{A}_1}$ and the population one $\Sigma_{Id,\mathcal{A}_1}$ dominates. With the observation that these concentrations have no relationship with the supervision signal, (Lu et al. 2021; Murata and Suzuki 2021) proposed to utilize the semi-supervised learning to reduce the error in this regime. In (Lu et al. 2021), Deep Ritz method requires semi-supervised learning while PINN does not for the exact empirical risk minimization solution. In our formulation, if $|p|$ is larger, the sub-optimal regime will become smaller, which contradict with the observation in (Lu et al. 2021). However (Lu et al. 2021) only considers the statistical generalization bound but doesn't take optimization into consideration. We leave designing algorithm with smaller sub-optimal regime as future work.

4. Sobolev implicit acceleration

The Sobolev norm has already been proposed as loss function for training neural network (Czarnecki et al. 2017) and solving PDEs (Son et al. 2021; J. Yu et al. 2021). However, all these papers need a further gradient information of the supervision signal. This does not fit the theoretical framework considered here and hence it is also not fair to compare their algorithms with methods without gradient supervision signal. Thus in this section, we proposed an alternative objective that can perform Sobolev training without gradient supervision loss function. The basic idea is to using an integration by parts

$$\int |\nabla u - \nabla f|^2 dx = \int \|\nabla u\|_2^2 + 2\Delta u \cdot f + \|\nabla f\|_2^2 dx, \tag{6}$$

which leads to an objective function without the gradient of the target function. In this section, we shall show how this idea is applied to different machine learning examples.

4.1 Predicting a Toy Function on Torus

In this section, we conduct experiments to illustrate the Sobolev implicit acceleration for function regression. Different from the Sobolev training (Czarnecki et al. 2017), the objective that we are interested in does not involve the gradient of the target function. As a result, we do not need to train a teacher network to provide the gradient supervision information as done in (Czarnecki et al. 2017). In the toy example, for simplicity we ignore the boundary terms introduced by the integral by part. Here consider estimating a function on the torus, *i.e.* a periodic function. We consider using $\int \lambda \|u - f\|^2 + \|\nabla u\|_2^2 + 2\Delta u \cdot f + \|\nabla f\|_2^2 dx$ as our objective function. The goal is to fit function $y = \sum_{i=1}^{d} \sin(2\pi x_i)$ using Gaussian Kernel and a simple three layer feed-forward network with tanh activation function. We randomly sampled 1000 data in 10 dimension as our dataset and run a gradient descent algorithm. Figure 2 presents our convergence result of the validation error, where the Sobolev norm have shown an acceleration effect for training.

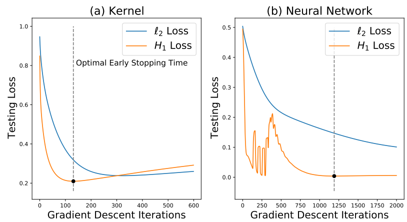


Figure 2. Sobolev Implicit Acceleration of Estimating function using kernel method and Neural Network. We observed that using Sobolev Norm as loss function can accelerate training.

4.2 Solving Partial Differential Equations

In this section, we conduct experiments to illustrate the Sobolev implicit acceleration for solving partial differential equation using PINNs (Raissi, Perdikaris, and Karniadakis 2019; J. Yu et al. 2021) in 3 dimensions. The example is a simple Poisson equation (static schrödinger equation) on the torus

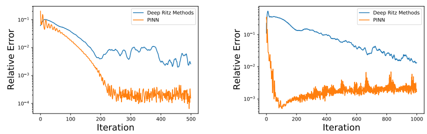
$$\Delta u + u = f \text{ in } \mathcal{T}^d = [0,1]^d_{\text{per}}. \tag{7}$$

We first compare the Physics Informed Neural Network (Raissi, Perdikaris, and Karniadakis 2019) and Deep Ritz Method (E and Yu 2018; Khoo, Lu, and Ying 2017) with online random inputs. To enforce the periodic boundary conditions, we add a penalty term $\mathcal{L}_b = \int_{(x,y)\in[0,1]^2} (u(x,y,0) - u(x,y,1))^2 + (u(0,x,y) - u(1,x,y))^2 + (u(x,0,y) - u(x,1,y))^2 dxdy$ to match the periodic condition of the function value and another term $\mathcal{L}_{b,grad} = \int_{(x,y)\in[0,1]^2} (\nabla u(x,y,0) - \nabla u(x,y,1))^2 + (\nabla u(0,x,y) - \nabla u(1,x,y))^2 + (\nabla u(x,0,y) - \nabla u(x,1,y))^2 dxdy$ to match the periodic condition of the function value. We tested PINN and Deep Ritz on both $u(x) = \sum_{i=1}^{d} \sin(2x_i)$ and $u(x) = \sum_{i=1}^{d} \sin(4x_i)$. We use the same experiment setting as (Chen, Du, and Wu 2020) and keep the learning rate constantly to $1e-3$ to match our theory. 50000 data points are randomly sampled in every batch. The results are shown in Figure 3. PINN converges faster than DRM consistently in terms of iteration number and the lead seems to become significant for more oscillatory problems.

To solve equation (7), we consider minimizing the following Sobolev norm objective function

$$\mathcal{L}(u) := \lambda \|\Delta u + u - f\|^2_{L_2(\Omega)} + \|\nabla \Delta u + \nabla u - \nabla f\|^2_{L_2(\Omega)}.$$

(Son et al. 2021; J. Yu et al. 2021) also considered using Sobolev norms as the loss function. (Son et al. 2021) showed that the Sobolev norms exhibit an acceleration effect. However, in our setting, we cannot have random samples of ∇f. To avoid information of ∇f appearing in the objective



(a) Smooth Problem (b) Harder Problem

Figure 3. We show the convergence result of PINN and Deep Ritz Method for smooth problem $\sum_{i=1}^{d}\sin(2\pi x)$ and harder problem $\sum_{i=1}^{d}\sin(4\pi x)$. PINN convergence faster than DRM for online stream input which also matches our theory and the empirical observation in (Chen, Du, and Wu 2020). The Sobolev Implicit Acceleration will becomes more significant for harder problem as our theory shows.

function, we perform an integration by parts that leads to the following objective function

$$\begin{aligned}\mathcal{L}_{grad} &= \int \|\nabla\Delta u(x) + \nabla u(x) - \nabla f(x)\|_2^2 dx\\ &= \int \|\nabla\Delta u(x)\|_2^2 + \|\nabla u(x)\|_2^2 + \|\nabla f(x)\|_2^2\\ &+ 2\nabla\Delta u(x)\nabla u(x) - 2\nabla u(x)\nabla f(x) - 2\nabla\Delta u(x)\cdot\nabla f(x)dx\\ &= \int \|\nabla\Delta u(x)\|_2^2 + \|\nabla u(x)\|_2^2 + \|\nabla f(x)\|_2^2\\ &+ 2\nabla\Delta u(x)\nabla u(x) + 2\Delta u(x)f(x) + 2\Delta\Delta u(x)\cdot f(x)dx.\end{aligned}$$

We conduct the Sobolev training with the objective function $\mathcal{L}_{pinn} + \lambda\mathcal{L}_{grad} + \lambda_1\mathcal{L}_b + \lambda\mathcal{L}_{b,grad}$ and compare it with PINN and DRM. Following mostly the experiment setting in (Chen, Du, and Wu 2020), we fix 3000 random samples as the dataset and run stochastic gradient descent with batchsize 50. The result presented in Figure 4 show the Sobolev implicit acceleration, i.e., the gradient dynamic of higher order Sobolev norm convergence faster. We do not scale the Sobolev training to online setting as under large batch size the Sobolev training consume too much memory at this point.

(a) Solution by Sobolev Training. (b) Convergence Speed.

Figure 4. Solving equation (7) in 3 dimension with 3000 fixed samples using Deep Ritz Method (E and Yu 2018), Physics-Informed Neural Network (Raissi, Perdikaris, and Karniadakis 2019) and Sobolev Training.

5. Conclusion and Discussion

In this paper, we consider the statistical optimality of gradient descent for solving elliptic inverse problem using a general class of objective functions. Although we can achieve statistical optimality of gradient descent using all the objective functions with proper early stopping time, the early stopping iteration strategy for the optimal solution behaves differently as a function of the sample size. For instance, we observed that PINN convergences faster than the DRM method. Generally speaking, by using a higher order Sobolev norm as loss function, one can accelerate training. The reason is that the differential operator can counteract the kernel integral operator, leading to better condition number for optimization. We call this phenomena *Sobolev implicit acceleration*.

Although we have shown the Sobolev implicit acceleration on several simple examples, the $\Delta^s u$ term is hard to compute in high dimensions, scalable Sobolev training without gradient supervision in higher dimension remains as future work. However, we believe that this direction is promising. For example, we can use MIM method (Lyu et al. 2020; Zhu and Yang 2021) to accelerate the training. It is also interesting to generalize our results beyond GD, for example to mirror descent (Vaškevičius, Kanade, and Rebeschini 2020) and accelerated gradient descent (Pagliana and Rosasco 2019). In this paper, we did not consider operators with continuous spectrum and it will be interesting to extend our results using the techniques in (Colbrook, Horning, and Townsend 2021). Due to technical issue, we have not considered the batch stochastic gradient descent. It will be interesting to characterize the condition under which the stochastic noise in gradient does not degrade the optimal bounds that we obtain. At the same time, we also want to investigate more complex nonlinear inverse problems as (Abraham and Nickl 2019; Monard, Nickl, and Paternain 2021) considered. It is also interesting to consider inverse problem arising from integral equation where $p > 0$.

Acknowledgement

Yiping Lu is supported by the Stanford Interdisciplinary Graduate Fellowship (SIGF). Jose Blanchet is supported in part by the Air Force Office of Scientific Research under award number FA9550-20-1-0397 and NSF grants 1915967, 1820942, 1838576. Lexing Ying is supported by National Science Foundation under award DMS-2011699. Yiping Lu also thanks Yifan Chen, Junbin Huang, Zong Shang, Bin Dong and George Em Karniadakis for helpful comments and feedback.

References

Abraham, Kweku, and Richard Nickl. 2019. On statistical calder\'on problems. *arXiv preprint arXiv:1906.03486.*

Adams, Robert A, and John JF Fournier. 2003. *Sobolev spaces.* Elsevier.

Ali, Alnur, Edgar Dobriban, and Ryan Tibshirani. 2020. The implicit regularization of stochastic gradient flow for least squares. In *International conference on machine learning,* 233–244. PMLR.

Ali, Alnur, J Zico Kolter, and Ryan J Tibshirani. 2019. A continuous-time view of early stopping for least squares regression. In *The 22nd international conference on artificial intelligence and statistics,* 1370–1378. PMLR.

Amari, Shun-ichi, Jimmy Ba, Roger Grosse, Xuechen Li, Atsushi Nitanda, Taiji Suzuki, Denny Wu, and Ji Xu. 2020. When does preconditioning help or hurt generalization? *arXiv preprint arXiv:2006.10732.*

Bai, Genming, Ujjwal Koley, Siddhartha Mishra, and Roberto Molinaro. 2021. Physics informed neural networks (pinns) for approximating nonlinear dispersive pdes. *arXiv preprint arXiv:2104.05584.*

Benning, Martin, and Martin Burger. 2018. Modern regularization methods for inverse problems. *Acta Numerica* 27:1–111.

Bietti, Alberto, and Francis Bach. 2020. Deep equals shallow for relu networks in kernel regimes. *arXiv preprint arXiv:2009.14397.*

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

Cabannes, Vivien, Loucas Pillaud-Vivien, Francis Bach, and Alessandro Rudi. 2021. Overcoming the curse of dimensionality with laplacian regularization in semi-supervised learning. In *Thirty-fifth conference on neural information processing systems.*

Calder, Jeff, A Mansouri, and Anthony Yezzi. 2010. Image sharpening via sobolev gradient flows. *SIAM Journal on Imaging Sciences* 3 (4): 981–1014.

Cao, Yuan, and Quanquan Gu. 2019. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in Neural Information Processing Systems* 32:10836–10846.

Caponnetto, Andrea, and Ernesto De Vito. 2007. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics* 7 (3): 331–368.

Chen, Jingrun, Rui Du, and Keke Wu. 2020. A comprehensive study of boundary conditions when solving pdes by dnns. *arXiv preprint arXiv:2005.04554.*

Chen, Lin, and Sheng Xu. 2020. Deep neural tangent kernel and laplace kernel have the same rkhs. *arXiv preprint arXiv:2009.10683.*

Chen, Yifan, Bamdad Hosseini, Houman Owhadi, and Andrew M Stuart. 2021. Solving and learning nonlinear pdes with gaussian processes. *arXiv preprint arXiv:2103.12959.*

Colbrook, Matthew, Andrew Horning, and Alex Townsend. 2021. Computing spectral measures of self-adjoint operators. *SIAM Review* 63 (3): 489–524.

Czarnecki, Wojciech Marian, Simon Osindero, Max Jaderberg, Grzegorz Świrszcz, and Razvan Pascanu. 2017. Sobolev training for neural networks. *arXiv preprint arXiv:1706.04859.*

Daniely, Amit. 2017. Sgd learns the conjugate kernel class of the network. *arXiv preprint arXiv:1702.08503.*

De Vito, Ernesto, Lorenzo Rosasco, Andrea Caponnetto, Umberto De Giovannini, Francesca Odone, and Peter Bartlett. 2005. Learning from examples as an inverse problem. *Journal of Machine Learning Research* 6 (5).

Dicker, Lee H, Dean P Foster, and Daniel Hsu. 2017. Kernel ridge vs. principal component regression: minimax bounds and the qualification of regularization operators. *Electronic Journal of Statistics* 11 (1): 1022–1047.

Dieuleveut, Aymeric, and Francis Bach. 2016. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics* 44 (4): 1363–1399.

Duan, Chenguang, Yuling Jiao, Yanming Lai, Xiliang Lu, and Zhijian Yang. 2021. Convergence rate analysis for deep ritz method. *arXiv preprint arXiv:2103.13330.*

E, Weinan, and Bing Yu. 2018. The deep ritz method: a deep learning-based numerical algorithm for solving variational problems. *Communications in Mathematics and Statistics* 6 (1): 1–12.

Fasshauer, Gregory E, and Qi Ye. 2011. Reproducing kernels of generalized sobolev spaces via a green function approach with distributional operators. *Numerische Mathematik* 119 (3): 585–611.

Fischer, Simon, and Ingo Steinwart. 2020. Sobolev norm learning rates for regularized least-squares algorithms. *J. Mach. Learn. Res.* 21:205–1.

Gerfo, L Lo, Lorenzo Rosasco, Francesca Odone, E De Vito, and Alessandro Verri. 2008. Spectral algorithms for supervised learning. *Neural Computation* 20 (7): 1873–1897.

Grohs, Philipp, and Lukas Herrmann. 2020. Deep neural network approximation for high-dimensional elliptic pdes with boundary conditions. *arXiv preprint arXiv:2007.05384.*

Han, Jiequn, Arnulf Jentzen, and E Weinan. 2018. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences* 115 (34): 8505–8510.

Han, Jiequn, Jianfeng Lu, and Mo Zhou. 2020. Solving high-dimensional eigenvalue problems using deep neural networks: a diffusion monte carlo like approach. *Journal of Computational Physics* 423:109792.

Hoop, Maarten V de, Nikola B Kovachki, Nicholas H Nelsen, and Andrew M Stuart. 2021. Convergence rates for learning linear operators from noisy data. *arXiv preprint arXiv:2108.12515.*

Hu, Tianyang, Wenjia Wang, Cong Lin, and Guang Cheng. 2020. Regularization matters: a nonparametric perspective on overparametrized neural network. *arXiv preprint arXiv:2007.02486.*

Hütter, Jan-Christian, and Philippe Rigollet. 2019. Minimax rates of estimation for smooth optimal transport maps. *arXiv preprint arXiv:1905.05828.*

Jacot, Arthur, Franck Gabriel, and Clément Hongler. 2018. Neural tangent kernel: convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572.*

Jiao, Yuling, Yanming Lai, Dingwei Li, Xiliang Lu, Yang Wang, and Jerry Zhijian Yang. 2021. *Convergence analysis for the pinns.* arXiv: 2109.01780 `[math.NA]`.

Jiao, Yuling, Yanming Lai, Yisu Luo, Yang Wang, and Yunfei Yang. 2021. Error analysis of deep ritz methods for elliptic equations. *arXiv preprint arXiv:2107.14478.*

Kalimeris, Dimitris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. 2019. Sgd on neural networks learns functions of increasing complexity. *Advances in Neural Information Processing Systems* 32:3496–3506.

Khoo, Yuehaw, Jianfeng Lu, and Lexing Ying. 2017. Solving parametric pde problems with artificial neural networks. *arXiv preprint arXiv:1707.03351.*

Lee, Jaehoon, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. 2017. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165.*

Lei, Yunwen, Ting Hu, and Ke Tang. 2021. Generalization performance of multi-pass stochastic gradient descent with convex loss functions. *J. Mach. Learn. Res.* 22:25–1.

Lin, Junhong, and Lorenzo Rosasco. 2017. Optimal rates for multi-pass stochastic gradient methods. *The Journal of Machine Learning Research* 18 (1): 3375–3421.

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

Liu, Zejian, and Meng Li. 2020. On the estimation of derivatives using plug-in krr estimators. *arXiv preprint arXiv:2006.01350.*

Long, Zichao, Yiping Lu, and Bin Dong. 2019. Pde-net 2.0: learning pdes from data with a numeric-symbolic hybrid deep network. *Journal of Computational Physics* 399:108925.

Long, Zichao, Yiping Lu, Xianzhong Ma, and Bin Dong. 2018. Pde-net: learning pdes from data. In *International conference on machine learning,* 3208–3216. PMLR.

Lu, Jianfeng, Yulong Lu, and Min Wang. 2021. A priori generalization analysis of the deep ritz method for solving high dimensional elliptic equations. *arXiv preprint arXiv:2101.01708.*

Lu, Yiping, Haoxuan Chen, Jianfeng Lu, Lexing Ying, and Jose Blanchet. 2021. Machine learning for elliptic pdes: fast rate generalization bound, neural scaling law and minimax optimality. *arXiv preprint arXiv:2110.06897.*

Luo, Tao, and Haizhao Yang. 2020. Two-layer neural networks for partial differential equations: optimization and generalization theory. *arXiv preprint arXiv:2006.15733.*

Lyu, Liyao, Zhen Zhang, Minxin Chen, and Jingrun Chen. 2020. Mim: a deep mixed residual method for solving high-order partial differential equations. *arXiv preprint arXiv:2006.04146.*

Manole, Tudor, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. 2021. Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364.*

Marteau-Ferey, Ulysse, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. 2019. Beyond least-squares: fast rates for regularized empirical risk minimization through self-concordance. In *Conference on learning theory,* 2294–2340. PMLR.

Marwah, Tanya, Zachary C Lipton, and Andrej Risteski. 2021. Parametric complexity bounds for approximating pdes with neural networks. *arXiv preprint arXiv:2103.02138.*

Mendelson, Shahar, and Joseph Neeman. 2010. Regularization in kernel learning. *The Annals of Statistics* 38 (1): 526–565.

Monard, François, Richard Nickl, and Gabriel P Paternain. 2021. Consistent inversion of noisy non-abelian x-ray transforms. *Communications on Pure and Applied Mathematics* 74 (5): 1045–1099.

Mücke, Nicole, and Enrico Reiss. 2020. Stochastic gradient descent in hilbert scales: smoothness, preconditioning and earlier stopping. *arXiv preprint arXiv:2006.10840.*

Murata, Tomoya, and Taiji Suzuki. 2021. Gradient descent in rkhs with importance labeling. In *International conference on artificial intelligence and statistics,* 1981–1989. PMLR.

Nickl, Richard, Sara van de Geer, and Sven Wang. 2020. Convergence rates for penalized least squares estimators in pde constrained regression problems. *SIAM/ASA Journal on Uncertainty Quantification* 8 (1): 374–413.

Nickl, Richard, and Sven Wang. 2020. On polynomial-time computation of high-dimensional posterior measures by langevin-type algorithms. *arXiv preprint arXiv:2009.05298.*

Nitanda, Atsushi, and Taiji Suzuki. 2020. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. *arXiv preprint arXiv:2006.12297.*

Pagliana, Nicolò, and Lorenzo Rosasco. 2019. Implicit regularization of accelerated methods in hilbert spaces. *arXiv preprint arXiv:1905.13000.*

Pillaud-Vivien, Loucas, Alessandro Rudi, and Francis Bach. 2018. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *arXiv preprint arXiv:1805.10074.*

Pinelis, Iosif F, and Aleksandr Ivanovich Sakhanenko. 1985. Remarks on inequalities for the probabilities of large deviations. *Teoriya Veroyatnostei i ee Primeneniya* 30 (1): 127–131.

Polyak, Boris T, and Anatoli B Juditsky. 1992. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization* 30 (4): 838–855.

Rahaman, Nasim, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. 2019. On the spectral bias of neural networks. In *International conference on machine learning,* 5301–5310. PMLR.

Raissi, Maziar, Paris Perdikaris, and George E Karniadakis. 2019. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics* 378:686–707.

Richardson Jr, WB. 2008. Sobolev gradient preconditioning for image-processing pdes. *Communications in Numerical Methods in Engineering* 24 (6): 493–504.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. U-net: convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention,* 234–241. Springer.

Rosasco, Lorenzo, Mikhail Belkin, and Ernesto De Vito. 2010. On learning with integral operators. *Journal of Machine Learning Research* 11 (2).

Scetbon, Meyer, and Zaid Harchaoui. 2021. A spectral analysis of dot-product kernels. In *International conference on artificial intelligence and statistics,* 3394–3402. PMLR.

Shi, Lei, Xin Guo, and Ding-Xuan Zhou. 2010. Hermite learning with gradient data. *Journal of computational and applied mathematics* 233 (11): 3046–3059.

Shin, Yeonjong, Zhongqiang Zhang, and George Em Karniadakis. 2020. Error estimates of residual minimization using neural networks for linear pdes. *arXiv preprint arXiv:2010.08019.*

Sirignano, Justin, and Konstantinos Spiliopoulos. 2018. Dgm: a deep learning algorithm for solving partial differential equations. *Journal of computational physics* 375:1339–1364.

Sitzmann, Vincent, Julien NP Martel, Alexander W Bergman, David B Lindell, and Gordon Wetzstein. 2020. Implicit neural representations with periodic activation functions. *arXiv preprint arXiv:2006.09661.*

Smale, Steve, and Ding-Xuan Zhou. 2007. Learning theory estimates via integral operators and their approximations. *Constructive approximation* 26 (2): 153–172.

Soliman, Yousuf, Albert Chern, Olga Diamanti, Felix Knöppel, Ulrich Pinkall, and Peter Schröder. 2021. Constrained willmore surfaces. *ACM Transactions on Graphics (TOG)* 40 (4): 1–17.

Son, Hwijae, Jin Woo Jang, Woo Jin Han, and Hyung Ju Hwang. 2021. Sobolev training for the neural network solutions of pdes. *arXiv preprint arXiv:2101.08932.*

Stein, Michael L. 1999. *Interpolation of spatial data: some theory for kriging.* Springer Science & Business Media.

Steinwart, Ingo, Don R Hush, Clint Scovel, et al. 2009. Optimal rates for regularized least squares regression. In *Colt,* 79–93.

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

Stuart, Andrew M. 2010. Inverse problems: a bayesian perspective. *Acta numerica* 19:451–559.

Tsybakov, Alexander B. 2004. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics* 32 (1): 135–166.

Tsybakov, Alexandre B. 2008. *Introduction to nonparametric estimation.* Springer Science & Business Media.

Vaškevičius, Tomas, Varun Kanade, and Patrick Rebeschini. 2020. The statistical complexity of early-stopped mirror descent. *arXiv preprint arXiv:2002.00189.*

Wei, Yuting, Fanny Yang, and Martin J Wainwright. 2017. Early stopping for kernel boosting algorithms: a general analysis with localized complexities. *arXiv preprint arXiv:1707.01543.*

Wojtowytsch, Stephan, et al. 2020. Some observations on partial differential equations in barron and multi-layer spaces. *arXiv preprint arXiv:2012.01484.*

Xu, Jinchao. 2020. The finite neuron method and convergence analysis. *arXiv preprint arXiv:2010.01458.*

Xu, Zhi-Qin John, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. 2019. Frequency principle: fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523.*

Yang, Yunan, Jingwei Hu, and Yifei Lou. 2021. Implicit regularization effects of the sobolev norms in image processing. *arXiv preprint arXiv:2109.06255.*

Yao, Yuan, Lorenzo Rosasco, and Andrea Caponnetto. 2007. On early stopping in gradient descent learning. *Constructive Approximation* 26 (2): 289–315.

Yu, Chris, Caleb Brakensiek, Henrik Schumacher, and Keenan Crane. 2021. Repulsive surfaces. *arXiv preprint arXiv:2107.01664.*

Yu, Chris, Henrik Schumacher, and Keenan Crane. 2021. Repulsive curves. *ACM Transactions on Graphics (TOG)* 40 (2): 1–21.

Yu, Jeremy, Lu Lu, Xuhui Meng, and George Em Karniadakis. 2021. Gradient-enhanced physics-informed neural networks for forward and inverse pde problems. *arXiv preprint arXiv:2111.02801.*

Zang, Yaohua, Gang Bao, Xiaojing Ye, and Haomin Zhou. 2020. Weak adversarial networks for high-dimensional partial differential equations. *Journal of Computational Physics* 411:109409.

Zhou, Xueyuan, and Mikhail Belkin. 2011. Semi-supervised learning by higher order regularization. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics,* 892–900. JMLR Workshop and Conference Proceedings.

Zhu, Quanhui, and Jiang Yang. 2021. A local deep learning method for solving high order partial differential equations. *arXiv preprint arXiv:2103.08915.*

Appendix.

The appendix is constructed as follows:

- In Appendix Appendix 1, we introduce the basic notations of Reproducing Kernel Hilbert space and the associated kernel integral operator. We also put a discussion of how differential operators and Sobolev spaces relates to the kernel setting we considered as a preliminary.
- In Appendix Appendix 2.1, we consider the statistical optimality of the early stopped gradient descent algorithm. We bound the difference of the gradient descent.
- In Appendix Appendix 3, we provide our proof for the lower bound in Section 3.1 using the Fano method.

Appendix 1. Preliminaries and Notations

This section starts with an overview of reproducing kernel Hilbert space, including Mercer's decomposition, the integral operator techniques (Smale and Zhou 2007; De Vito et al. 2005; Caponnetto and De Vito 2007; Fischer and Steinwart 2020) and the relationship between RKHS and the Sobolev space (Adams and Fournier 2003). In order to fit the objective function we considered, we did a slight modification to the original integral operator technique (Smale and Zhou 2007; De Vito et al. 2005; Caponnetto and De Vito 2007).

Appendix 1.1 Reproducing Kernel Hilbert Space

We consider a Hilbert space $\mathcal{H}$ with inner product $\langle , \rangle_{\mathcal{H}}$ is a separable Hilbert space of functions $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$. We call this space a Reproducing Kernel Hilbert space if $f(x) = \langle f, K_x \rangle_{\mathcal{H}}$ for all $K_x \in \mathcal{H} : t \to K(x,t), x \in \mathcal{X}$. Now we consider a distribution ρ on $\mathcal{X} \times \mathcal{Y} (\mathcal{Y} \subset \mathbb{R})$ and denote ρ_X as the margin distribution of ρ on $\mathcal{X}$. We further assume $\mathbb{E}[K(x,x)] < \infty$ and $\mathbb{E}[Y^2] < \infty$. We define $g \otimes h = gh^\top$ is an operator from $\mathcal{H}$ to $\mathcal{H}$ defined as

$$g \otimes h : f \to \langle f, h \rangle_{\mathcal{H}} g.$$

At the same time, we knows that

$$\|f \otimes g\| = \|f\|_{\mathcal{H}} \|g\|_{\mathcal{H}}$$

holds for all $f, g \in \mathcal{H}$.

The integral operator technique(Smale and Zhou 2007; Caponnetto and De Vito 2007) consider the covariance operator on the Hilbert space $\mathcal{H}$ defined as $\Sigma = \mathbb{E}_{\rho_{\mathcal{X}}} K_x \otimes K_x$. Then for all $f \in \mathcal{H}$, using the reproducing property, we know that

$$(\Sigma f)(z) = \langle K_z, \Sigma f \rangle_{\mathcal{H}} = \mathbb{E}[f(X)k(X,z)] = \mathbb{E}[f(X)K_x(X)].$$

If we consider the mapping $S : \mathcal{H} \to L_2(dx)$ defined as a parameterization of a vast class of functions in $\mathbb{R}^{\mathcal{X}}$ via $\mathcal{H}$ through the mapping $(Sg)(x) = \langle g, K_x \rangle$ $(\Phi(x) = K_x = K(\cdot, x))$. Its adjoint operator S^* then can be defined as $S^* : \mathcal{L}_2 \to \mathcal{H} : g \to \int_{\mathcal{X}} g(x) K_x \rho_X(dx)$ and at the same time Σ is the same as the self-adjoint operator S^*S and the self-adjoint operator $\mathcal{L} = SS^* : L_2 \to L_2$ can be defined as

$$(\mathcal{L}f)(x) = (SS^*f)(x) = \int_{\mathcal{X}} K(x,z) f(z) d\rho_{\mathcal{X}}(x), \forall f \in L_2$$

Next we consider the eigen-decomposition of the integral operator $\mathcal{L}$ via Mecer's Theorem. There exists an orthonormal basis $\{\psi_i\}$ of $\mathcal{L}_2(\mathcal{X})$ consisting of eigenfunctions of kernel integral operator $\mathcal{L}$. At the same time, the kernel function have the following representation $K(s,t) = \sum_{i=1}^{\infty} \lambda_i e_i(s) e_j(t)$ where e_i are orthogonal basis of $\mathcal{L}_2(\rho_{\mathcal{X}})$. Then e_i is also the eigenvector of the covariance operator Σ with eigenvalue $\lambda_i > 0$, *i.e.* $\Sigma e_i = \lambda_i e_i$.

Appendix 2. Proof of the Upper Bound

In this section, we consider the convergence of the gradient descent algorithm to the target function 1. In particular, we consider the gradient descent as a special case of a wider class of spectral filter algorithms (Dieuleveut and Bach 2016; Gerfo et al. 2008; Pillaud-Vivien, Rudi, and Bach 2018; Lin et al. 2020). In our inverse problem setting, the spectral filter is defined as the estimator of the following form for $\lambda > 0$,

$$\hat{q}_\lambda = g_\lambda(\hat{\Sigma}_{Id,\mathcal{A}_1})\mathcal{A}_2\hat{S}_n^*\hat{y},$$

where $\hat{S}_n g = (g(x_1), \cdot, g(x_n))$ (leads to $\hat{S}_n^*$ maps from $\mathbb{R}^n$ to $\mathcal{H}$ via $\hat{S}_n^*(a_1, a_2, \cdots, a_n) = \frac{1}{n}\sum_{i=1}^n a_n K_{x_n}$), $\hat{\Sigma}_{\mathcal{O}_1,\mathcal{O}_2} = \frac{1}{n}\sum_{i=1}^n \mathcal{O}_1 K_x \otimes \mathcal{O}_2 K_x$ and Id is the identity operator. The function $q_\lambda : \mathbb{R}^+ \to \mathbb{R}^+$ is a function known as *filter*, which is an approximation of x^{-1} controlled by λ. We further define the error of approximation via $r_\lambda(x) = 1 - xq_\lambda(x)$. **Spectral Filters** need the function q_λ further satisfies

$$\lambda q_\lambda(x) \le c_q, r_\lambda(x)x^u \le c_q\lambda^u, \forall x > 0, \lambda > 0, u \in [0,1],$$

for some positive $c_q > 0$. Next we show that the averaged gradient descent can be considered as spectral filter algorithm with filter $q^\eta(x) = \left(1 - \frac{1-(1-\gamma x)^t}{\gamma t x}\right)\frac{1}{x}$. Let us consider the gradient descent $\eta_0 = 0, \eta_u = \eta_{u-1} + \gamma(\mathcal{A}_2^\top \hat{S}_n^*\hat{y} - \hat{\Sigma}_{Id,\mathcal{A}_1}\eta_{t-1})$, then

$$\begin{aligned}\eta_t &= (I - \gamma\hat{\Sigma}_{Id,\mathcal{A}_1})\eta_{t-1} + \gamma\mathcal{A}_2^\top \hat{S}_n^*\hat{y} = \gamma\sum_{k=0}^{t-1}(I - \gamma\hat{\Sigma}_{Id,\mathcal{A}_1}))^k \mathcal{A}_2^\top \hat{S}_n^*\hat{y} \\ &= \left[I - (I - \gamma\hat{\Sigma}_{Id,\mathcal{A}_1}))^t\right](\hat{\Sigma}_{Id,\mathcal{A}_1}))^{-1}\mathcal{A}_2^\top \hat{S}_n^*\hat{y}\end{aligned} \tag{8}$$

and

$$\bar{\eta}_t = \frac{1}{t}\sum_{i=0}^t \eta_i = \frac{1}{t}\sum_{i=0}^t\left[I - (I - \gamma\hat{\Sigma}_{Id,\mathcal{A}_1}))^i\right](\hat{\Sigma}_{Id,\mathcal{A}_1}))^{-1}\mathcal{A}_2^\top \hat{S}_n^*\hat{y}. \tag{9}$$

Thus if we take the filter $q_t(x) = \frac{1}{x}\left(1 - \frac{1-(1-\gamma x)^t}{\gamma t x}\right)$, we can have $\bar{\eta} = q_t(\Sigma_{Id,\mathcal{A}_1})\mathcal{A}_2^\top \hat{S}_n^*\hat{y}$. At the same time $x^u r_t(x) = x^u(1 - xq_t(x)) = x^u\frac{1-(1-\gamma t)^t}{\gamma t x} \le \frac{(\gamma t x)^{1-u}}{\gamma t x}x^u = \frac{1}{(\gamma t)^u}$. Thus we can consider the gradient descent algorithm for the inverse problem as a spectral filtering algorithm.

Next, we compare the spectral filter of early stopped gradient descent with ridge regression and decompose the risk to bias and variance terms. Via bounding the bias and variance separately, we can achieve information theoretical optimal upper bound for such problems.

Appendix 2.1 Convergence Of the Gradient Descent Algorithm

To conduct our proof of the upper bound, we consider $g_\lambda = (\Sigma_{Id,\mathcal{A}_1} + \lambda I)^{-1}\mathcal{A}_2 S^* f_\rho$ and decompose the error as $\underbrace{(g_\lambda - u^*)}_{\text{Bias}} + \underbrace{(\hat{q}_\lambda - g_\lambda)}_{\text{Variance}}$. We first bound the bias in the general Sobolev norm then come to bound the variance.

Appendix 2.1.1 Auxiliary Lemmas

We first introduce several auxiliary lemmas which aims to bound different quantities according to the effective dimension/capacity of the kernel covariance operator. We define $\mathcal{N}(\lambda) = \mathbb{E}_x\|(\Sigma_{\mathcal{A}_1} + \lambda)^{-1/2}\mathcal{A}_2 K_x\|_H^2 = \mathrm{Tr}((\Sigma_{\mathcal{A}_1} + \lambda)^{-1}\Sigma_{\mathcal{A}_2\mathcal{A}_2})$, $\mathcal{N}_\infty^1(\lambda) = \sup_{x\in\rho(x)}\|(\Sigma_{Id,\mathcal{A}_1} + \lambda)^{-1/2}\mathcal{A}_2 K_x\|_H^2$ and $\mathcal{N}_\infty^2(\lambda) = \sup_{x\in\rho(x)}\|(\Sigma_{Id,\mathcal{A}_1} + \lambda)^{-1/2}K_x\|_H^2$ which are important important quantities used to bound the variance of our estimator.

Lemma 1. *There exists a constant D such that the following inequality is satisfied for* $\lambda > 0$,

$$\mathcal{N}(\lambda) \leq D(\lambda)^{-\frac{1}{p+\alpha}+\frac{p-2q}{p+\alpha}}$$

Proof. We use the spectral representation to bound the effective dimension $\mathcal{N}(\lambda)$ as

$$\begin{aligned}
\mathcal{N}(\lambda) &= \mathrm{Tr}\big((\Sigma_{\mathcal{A}_1}+\lambda)^{-1}\Sigma_{\mathcal{A}_2\mathcal{A}_2}\big) = \sum_{i=1}^{\infty}\frac{\mu_i q_i^2}{\mu_i p_i+\lambda} \\
&\lesssim \sum_{i=1}^{\infty}\frac{i^{-\alpha-2q}}{i^{-\alpha-p}+\lambda} \leq \int_0^{\infty}\frac{\tau^{p-2q}}{1+\lambda(\tau^{\alpha+p})}d\tau \qquad (10)\\
&= (\lambda)^{-\frac{1}{p+\alpha}}\int_0^{\infty}\frac{(\lambda)^{\frac{2q-p}{p+\alpha}}\tau^{p-q}}{1+\tau^{\alpha+p}} = \Omega\left((\lambda)^{-\frac{1}{p+\alpha}-\frac{p-2q}{p+\alpha}}\right)
\end{aligned}$$

□

Lemma 2. *There exists a constant D such that the following inequality is satisfied for* $\lambda > 0$,

$$Trace\big((\Sigma_{\mathcal{A}_1}+\lambda)^{-1}\Sigma_{Id,Id}\big) \leq D\lambda^{\frac{-p-1}{p+\alpha}}$$

Proof. Similarly we use the spectral representation to bound the LHS as

$$\begin{aligned}
\mathrm{Trace}\big((\Sigma_{\mathcal{A}_1}+\lambda)^{-1}\Sigma_{Id,Id}\big) &= \sum_{i=1}^{\infty}\frac{\mu_i}{\mu_i p_i+\lambda} \\
&\lesssim \sum_{i=1}^{\infty}\frac{i^{-\alpha}}{i^{-\alpha-p}+\lambda} \leq \int_0^{\infty}\frac{\tau^{p}}{1+\lambda(\tau^{\alpha+p})}d\tau \qquad (11)\\
&= (\lambda)^{-\frac{1}{p+\alpha}}\int_0^{\infty}\frac{(\lambda)^{\frac{-p}{p+\alpha}}\tau^{p}}{1+\tau^{\alpha+p}} = \Omega\left(\lambda^{\frac{-p-1}{p+\alpha}}\right)
\end{aligned}$$

□

Lemma 3. *We denote the following quantity by* $\mathcal{N}_\infty^1, \mathcal{N}_\infty^2$ *and* $\mathcal{N}_\infty^3$ *can be bounded by*

- $\mathcal{N}_\infty^1(\lambda) = \sup_{x\in\rho(x)}\|(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}K_x\|_H^2 \leq \|k_v^\alpha\|_\infty^2\lambda^{-\frac{\mu\alpha+p}{\alpha+p}}$,
- $\mathcal{N}_\infty^2(\lambda) = \sup_{x\in\rho(x)}\|(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}\mathcal{A}_2K_x\|_H^2 \leq \|k_v^\alpha\|_\infty^2\lambda^{-\frac{\mu\alpha+p+2q}{\alpha+p}}$,
- $\mathcal{N}_\infty^3(\lambda) = \sup_{x\in\rho(x)}\|(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}\mathcal{A}_1K_x\|_H^2 \leq \|k_v^\alpha\|_\infty^2\lambda^{-\frac{\mu\alpha+3p}{\alpha+p}}$.

Proof. We can also prove the bound from the spectral formulation and the l_∞ embedding property of the kernel function

$$\begin{aligned}
\|(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}K_x\|_H^2 &= \sum_{i\geq 1}\frac{\mu_i}{\mu_i p_i+\lambda}e_i^2(x) \\
&\leq \left(\sum_{i\geq 1}\mu_i^{\mu}e_i^2(x)\right)\sup_{i\geq 1}\frac{\mu_i^{1-\mu}}{\mu_i p_i+\lambda} \lesssim \left(\sum_{i\geq 1}\mu_i^{\mu}e_i^2(x)\right)\sup_{i\geq 1}\frac{i^{-(1-\mu)\alpha}}{i^{-\alpha-p}+\lambda} \qquad (12)\\
&\leq \lambda^{-\frac{\mu\alpha+p}{\alpha+p}}\|k_v^{\mu}\|_\infty^2,
\end{aligned}$$

and

$$\begin{aligned}
\|(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}\mathcal{A}_2K_x\|_H^2 &= \sum_{i\geq 1}\frac{\mu_i q_i^2}{\mu_i p_i+\lambda}e_i^2(x)\\
&\leq \left(\sum_{i\geq 1}\mu_i^{\mu}e_i^2(x)\right)\sup_{i\geq 1}\frac{\mu_i^{1-\mu}q_i^2}{\mu_i p_i+\lambda}\lesssim \left(\sum_{i\geq 1}\mu_i^{\mu}e_i^2(x)\right)\sup_{i\geq 1}\frac{i^{-(1-\mu)\alpha-2q}}{i^{-\alpha-p}+\lambda}\\
&\leq \lambda^{-\frac{\mu\alpha+p+2q}{\alpha+p}}\|k_v^{\mu}\|_\infty^2.
\end{aligned} \tag{13}$$

Similarly we have

$$\begin{aligned}
\|(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}\mathcal{A}_1K_x\|_H^2 &= \sum_{i\geq 1}\frac{\mu_i p_i^2}{\mu_i p_i+\lambda}e_i^2(x)\\
&\leq \left(\sum_{i\geq 1}\mu_i^{\mu}e_i^2(x)\right)\sup_{i\geq 1}\frac{\mu_i^{1-\mu}p_i^2}{\mu_i p_i+\lambda}\lesssim \left(\sum_{i\geq 1}\mu_i^{\mu}e_i^2(x)\right)\sup_{i\geq 1}\frac{i^{-(1-\mu)\alpha-2p}}{i^{-\alpha-p}+\lambda}\\
&\leq \lambda^{-\frac{\mu\alpha+2p}{\alpha+p}}\|k_v^{\mu}\|_\infty^2.
\end{aligned} \tag{14}$$

□

Lemma 4. *For all $\lambda > 0$, we have*

$$\|\Sigma^{\frac{1-\gamma}{2}}(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}\|^2 \leq \lambda^{-\frac{\gamma\alpha+p}{\alpha+p}}$$

Proof. We first bound $\|\Sigma^{\frac{1-\gamma}{2}}(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}\|^2$

$$\|\Sigma^{\frac{1-\gamma}{2}}(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}\|^2 = \sup_{i\geq 1}\frac{\mu_i^{1-\gamma}}{\mu_i p_i+\lambda}\lesssim \sup_{i\geq 1}\frac{i^{-(1-\gamma)\alpha}}{i^{-\alpha-p}+\lambda}\leq \lambda^{-\frac{\gamma\alpha+p}{\alpha+p}}$$

□

Lemma 5. *With probability $1-e^{-\tau}$, we have*

$$\|(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}(\hat{\Sigma}_{Id,\mathcal{A}_1}-\Sigma_{Id,\mathcal{A}_1})(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}\|^2 \lesssim \sqrt{\frac{\tau\sqrt{\mathcal{N}_\infty^1(\lambda)\mathcal{N}_\infty^3(\lambda)}}{n}}$$

and as a consequence once $n \gtrsim \tau\lambda^{-\frac{\mu\alpha+2p}{\alpha+p}}$, we'll have

$$\frac{1}{2}\leq \|(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{1/2}(\hat{\Sigma}_{Id,\mathcal{A}_1}+\lambda)^{-1/2}\|\leq 2,\ \frac{1}{2}\leq \|(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}(\hat{\Sigma}_{Id,\mathcal{A}_1}+\lambda)^{1/2}\|\leq 2.$$

Proof. We utilize the concentration result for Hilbert space valued random variable (Pinelis and Sakhanenko 1985) to prove the bound here. Now, we consider the operator $C_x : \mathcal{H}\to\mathcal{H}$ the operator defined by

$$C_x f := \mathcal{A}_1 f(x)k(x,\cdot) = \langle f, \mathcal{A}_1 K_x\rangle K_x,$$

and consider the random variable $\xi_x := (\Sigma_{Id,\mathcal{A}_1} + \lambda)^{1/2} C_x (\Sigma_{Id,\mathcal{A}_1} + \lambda)^{-1/2}$. From definition, we know that

$$\begin{aligned} \xi_x f &= (\Sigma_{Id,\mathcal{A}_1} + \lambda)^{1/2} C_x (\Sigma_{Id,\mathcal{A}_1} + \lambda)^{-1/2} f \\ &= \left\langle f, (\Sigma_{Id,\mathcal{A}_1} + \lambda)^{-1/2} \mathcal{A}_1 K_x \right\rangle (\Sigma_{Id,\mathcal{A}_1} + \lambda)^{1/2} K_x \\ &= ((\Sigma_{Id,\mathcal{A}_1} + \lambda)^{1/2} K_x \otimes (\Sigma_{Id,\mathcal{A}_1} + \lambda)^{-1/2} \mathcal{A}_1 K_x) f. \end{aligned} \tag{15}$$

At the same time, we know that $\|f \otimes g\| = \|f\|_{\mathcal{H}} \|g\|_{\mathcal{H}}$ for all $f, g \in \mathcal{H}$, thus utilizing the concentration results for Hilbert space valued random variable, we have

$$\|(\Sigma_{Id,\mathcal{A}_1} + \lambda)^{-1/2} (\hat{\Sigma}_{Id,\mathcal{A}_1} - \Sigma_{Id,\mathcal{A}_1}) (\Sigma_{Id,\mathcal{A}_1} + \lambda)^{-1/2}\|^2 \lesssim \sqrt{\frac{\tau \sqrt{\mathcal{N}_\infty^1(\lambda) \mathcal{N}_\infty^3(\lambda)}}{n}}.$$

From Lemma 3, we know that $\mathcal{N}_\infty^1(\lambda) = \sup_{x \in \rho(x)} \|(\Sigma_{Id,\mathcal{A}_1} + \lambda)^{-1/2} K_x\|_H^2 \leq \|k_v^\alpha\|_\infty^2 \lambda^{-\frac{\mu\alpha + p}{\alpha + p}}$, and $\mathcal{N}_\infty^3(\lambda) = \sup_{x \in \rho(x)} \|(\Sigma_{Id,\mathcal{A}_1} + \lambda)^{-1/2} \mathcal{A}_1 K_x\|_H^2 \leq \|k_v^\alpha\|_\infty^2 \lambda^{-\frac{\mu\alpha + 3p}{\alpha + p}}$. Thus once $n \gtrsim \tau \lambda^{-\frac{\mu\alpha + 2p}{\alpha + p}}$, we'll have

$$\frac{1}{2} \leq \|(\Sigma_{Id,\mathcal{A}_1} + \lambda)^{1/2} (\hat{\Sigma}_{Id,\mathcal{A}_1} + \lambda)^{-1/2}\| \leq 2, \frac{1}{2} \leq \|(\Sigma_{Id,\mathcal{A}_1} + \lambda)^{-1/2} (\hat{\Sigma}_{Id,\mathcal{A}_1} + \lambda)^{1/2}\| \leq 2.$$

□

Theorem 3 (Bernstein's Inequality). *Let $(\Omega, \mathcal{B}, P)$ be a probability space, H be a separable Hilbert space, and $\xi : \Omega \to H$ with*

$$\mathbb{E}_P \|\xi\|_H^m \leq \frac{1}{2} m! \sigma^2 L^{m-2}$$

for all $m \geq 2$. Then, for $\tau \geq 1$ and $n \geq 1$, the following concentration inequality is satisfied

$$\mathcal{P}^n \left[\left| \|\frac{1}{n} \sum_{i=1}^n \xi(\omega_i) - \mathbb{E}_P \xi \right| \|_H^2 \geq 32 \frac{\tau^2}{n} \left(\sigma^2 + \frac{L^2}{n} \right) \right] \leq 2e^{-\tau}$$

Lemma 6 (Lemma 25 in (Fischer and Steinwart 2020)). *For $\lambda > 0$ and $0 \leq \alpha \leq 1$, the function $f_{\lambda,\alpha} : [0, \infty) \to \mathbb{R}$ be defined by $f_{\lambda,\alpha}(t) := \frac{t^\alpha}{\lambda + t}$. In the case $\alpha = 0$ the function is decreasing and for $\alpha = 1$ the function is increasing. Furthermore*

$$\lambda^{\alpha-1}/2 \leq \sup_{t \geq 0} f_{\lambda,\alpha}(t) \leq \lambda^{\alpha-1}$$

for $0 < \alpha < 1$ the function attain its supremum at $t^* = \frac{\lambda\alpha}{1-\alpha}$

Proof. For completeness, we provide the proof here. For function $f_{\lambda,\alpha}(t) := \frac{t^\alpha}{\lambda + t}$ with $0 < \alpha < 1$, we know the derivative of it is $f'_{\lambda,\alpha}(t) = \frac{\alpha t^{\alpha-1}(\alpha + t) - t^\alpha}{(\lambda + t)^2}$. The derivative $f'_{\lambda,\alpha}$ has a unique root at $t^* = \alpha\lambda/(1 - \alpha)$. $f_{\lambda,\alpha}$ attains global maximum at t^* and

$$\sup_{t \geq 0} f_{\lambda,\alpha}(t) = f_{\lambda,\alpha}(t^*) = \lambda^{\alpha-1} \alpha^\alpha (1 - \alpha)^{1-\alpha} \leq \lambda^{\alpha-1}.$$

At the same time, $(\alpha^\alpha (1 - \alpha)^{1-\alpha})' = \alpha^\alpha (1 - \alpha)^{1-\alpha} \log\left(\frac{\alpha}{1-\alpha}\right)$ thus $\alpha^\alpha (1 - \alpha)^{1-\alpha}$ achieves minimum $\frac{1}{2}$ when $\alpha = \frac{1}{2}$. Thus we know $\lambda^{\alpha-1}/2 \leq \sup_{t \geq 0} f_{\lambda,\alpha}(t)$. □

Appendix 2.1.2 Bias

In this section, we consider the bias introduced by the regularization factor, *i.e.* the difference between $g_\lambda = (\Sigma_{Id,\mathcal{A}_1} + \lambda I)^{-1}\mathcal{A}_2 S^* f_\rho$ and the ground truth solution $\mathcal{A}_1^{-1}\mathcal{A}_2 f_\rho$.

Lemma 7. *If $u^* = \mathcal{A}_1^{-1}\mathcal{A}_2 f_\rho \in [H]^\beta$ holds, then for all $0 \le \gamma \le \beta$ and $\lambda > 0$, the following bounds holds*

$$\|g_\lambda - \mathcal{A}_1^{-1}\mathcal{A}_2 f_\rho\|_\gamma \lesssim \lambda^{\frac{(\frac{\beta-\gamma}{2})\alpha}{\alpha+p}} \|u^*\|_{[H]^\beta}.$$

Here $g_\lambda = (\Sigma_{Id,\mathcal{A}_1} + \lambda I)^{-1}\mathcal{A}_2 S^* f_\rho$.

Proof. Since $u^* = \mathcal{A}_1^{-1}\mathcal{A}_2 f_\rho \in [H]^\beta$, we can use the spectral representation $u^* = \sum_{i=1}^n a_i e_i$ with $\|u^*\|_{[H]^\beta} = \sum_{i=1}^\infty \mu_i^{-\beta} a_i$. At the same time $\mathcal{A}_2 f_\rho = \mathcal{A}_1 u^* = \sum_{i=1}^n a_i p_i e_i$. We also observe that the matrix $(\Sigma_{Id,\mathcal{A}_1} + \lambda I)^{-1}$ have the spectral representation $(\Sigma_{Id,\mathcal{A}_1} + \lambda I)^{-1} = \sum_{i=1}^\infty (\mu_i p_i + \lambda)^{-1} e_i \otimes e_i$ and leads to the spectral representation of the solution

$$g_\lambda = (\Sigma_{Id,\mathcal{A}_1} + \lambda I)^{-1}\mathcal{A}_2 S^* f_\rho = \sum_{i=1}^\infty \frac{\mu_i q_i}{\mu_i p_i + \lambda}\frac{p_i}{q_i} a_i e_i = \sum_{i=1}^\infty \frac{\mu_i p_i}{\mu_i p_i + \lambda} a_i e_i$$

Then we can bound the bias via the spectral representation

$$\begin{aligned}
\|g_\lambda - \mathcal{A}_1^{-1}\mathcal{A}_2 f_\rho\|_\gamma^2 &= \|(\Sigma_{Id,\mathcal{A}_1} + \lambda I)^{-1}\mathcal{A}_2 S^* f_\rho - \mathcal{A}_1^{-1}\mathcal{A}_2 f_\rho\|_\gamma^2 \\
&= \left| \left| \sum_{i=1}^\infty \frac{\mu_i p_i}{\mu_i p_i + \lambda} a_i e_i - a_i e_i \right| \right|^2 = \left| \left| \sum_{i=1}^\infty \frac{\lambda}{\mu_i p_i + \lambda} a_i e_i \right| \right|_\gamma^2 \\
&= \sum_{i=1}^\infty \left(\frac{\lambda}{\mu_i p_i + \lambda} a_i \right)^2 \mu_i^{-\gamma} \qquad (16) \\
&= \lambda^2 \left(\sup_{i\ge 1} \frac{i^{-\alpha(\frac{\beta-\gamma}{2})}}{\lambda + i^{-\alpha-p}} \right)^2 \sum_{i\ge 1} \mu_i^{-\beta} a_i^2 \le \lambda^{\frac{(\beta-\gamma)\alpha}{\alpha+p}} \|u^*\|_{[H]^\beta}^2
\end{aligned}$$

□

In this section, we also bound a bias over the energy function $\|\mathcal{A}_1 g_\lambda - \mathcal{A}_2 f_\rho\|_2^2$, which will be used in bounding the variance term.

Lemma 8. *If $u^* = \mathcal{A}_1^{-1}\mathcal{A}_2 f_\rho \in [H]^\beta$ holds, then for all $0 \le \gamma \le \beta$ and $\lambda > 0$, the following bounds holds*

$$\|\mathcal{A}_1 g_\lambda - \mathcal{A}_2 f_\rho\|_2 \lesssim \lambda^{\frac{\beta\alpha - 2p}{2(\alpha+p)}} \|u^*\|_{[H]^\beta}.$$

Here $g_\lambda = (\Sigma_{Id,\mathcal{A}_1} + \lambda I)^{-1}\mathcal{A}_2 S^* f_\rho$.

Proof. As discussed in the proof of Lemma 7, we have the spectral representation of g_λ as

$$g_\lambda = (\Sigma_{Id,\mathcal{A}_1} + \lambda I)^{-1}\mathcal{A}_2 S^* f_\rho = \sum_{i=1}^\infty \frac{\mu_i q_i}{\mu_i p_i + \lambda}\frac{p_i}{q_i} a_i e_i = \sum_{i=1}^\infty \frac{\mu_i p_i}{\mu_i p_i + \lambda} a_i e_i$$

Thus $\mathcal{A}_1 g_\lambda - \mathcal{A}_2 f_\rho = \sum_{i=1}^\infty \left(\frac{\mu_i p_i^2}{\mu_i p_i + \lambda} - p_i \right) a_i e_i = -\sum_{i=1}^\infty \left(\frac{p_i \lambda}{\mu_i p_i + \lambda} \right) a_i e_i$ and we can have the bound of the bias in the energy norm as

$$
\begin{aligned}
\|\mathcal{A}_1 g_\lambda - \mathcal{A}_2 f_\rho\|_2^2 &= \left| \| \sum_{i=1}^{\infty} \left(\frac{p_i\lambda}{\mu_i p_i + \lambda} - p_i \right) a_i e_i \right| \|_2^2 \\
&= \sum_{i=1}^{\infty} \left(\frac{p_i\lambda}{\mu_i p_i + \lambda} a_i \right)^2 = \lambda^2 \left(\sup_{i\geq 1} \frac{i^{-(\frac{\alpha\beta}{2})-p}}{\lambda + i^{-\alpha-p}} \right)^2 \sum_{i\geq 1} \mu_i^{-\beta} a_i^2 \\
&\lesssim \lambda^{\frac{\beta\alpha-2p}{\alpha+p}} \|u^*\|_{[H]^\beta}.
\end{aligned}
$$

□

Appendix 2.1.3 Variance

In this section, we bound the variance which defined as the difference between between $g_\lambda = (\Sigma_{Id,\mathcal{A}_1} + \lambda I)^{-1}\mathcal{A}_2 S^* f_\rho$ and $\hat{g}_\lambda = q_\lambda(\hat{\Sigma}_{Id,\mathcal{A}_1})\mathcal{A}_2\hat{S}^* y$ at the scale $O\left(\frac{(\sigma^2+R^2\lambda^{2r})\mathcal{N}(\lambda_q)}{n} + \frac{\lambda^{\frac{(\beta-\gamma)\alpha-\mu\alpha-p}{\alpha+p}}}{n} + o(\frac{1}{n})\right)$. We first did the following decomposition

$$
\begin{aligned}
\Sigma^{\frac{1-\gamma}{2}}(g_\lambda - \hat{g}_\lambda) &= \Sigma^{\frac{1-\gamma}{2}} q_\lambda(\hat{\Sigma}_{Id,\mathcal{A}_1})(\mathcal{A}_2\hat{S}^* y - (\hat{\Sigma}_{Id,\mathcal{A}_1}) g_\lambda) + \Sigma^{\frac{1-\gamma}{2}} \left[g_\lambda(\hat{\Sigma}_{Id,\mathcal{A}_1})\hat{\Sigma}_{Id,\mathcal{A}_1} - I \right] g_\lambda \\
&= \Sigma^{\frac{1-\gamma}{2}} q_\lambda(\hat{\Sigma}_{Id,\mathcal{A}_1})(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2} \left[\frac{1}{n}\sum_{i=1}^{n} (\xi(x_i,y_i)) \right] + \Sigma^{\frac{1-\gamma}{2}} r(\hat{\Sigma}_{Id,\mathcal{A}_1}) g_\lambda \\
&= \underbrace{\Sigma^{\frac{1-\gamma}{2}} q_\lambda(\hat{\Sigma}_{Id,\mathcal{A}_1})(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2} \left[\frac{1}{n}\sum_{i=1}^{n} (\xi(x_i,y_i) - \mathbb{E}_P \xi(x,y)) \right]}_{(I)} \\
&\quad + \underbrace{\Sigma^{\frac{1-\gamma}{2}} q_\lambda(\hat{\Sigma}_{Id,\mathcal{A}_1})(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2} \mathbb{E}_P \xi(x,y)}_{(II)} + \underbrace{\Sigma^{\frac{1-\gamma}{2}} r(\hat{\Sigma}_{Id,\mathcal{A}_1}) g_\lambda}_{(III)},
\end{aligned}
\tag{17}
$$

where we take the random variable $\xi(x,y)$ as $\xi(x,y) = (\Sigma_{Id,\mathcal{A}_1} + \lambda)^{-1/2}(y\mathcal{A}_2 K_x - \mathcal{A}_1 g_\lambda(x) K_x)$ which satisfies $\mathbb{E}_Q \xi_2 = (\Sigma_{Id,\mathcal{A}_1} + \lambda)^{-1/2}(\mathcal{A}_2 f_Q - \Sigma^Q_{Id,\mathcal{A}_1} g_\lambda)$ where $f_Q = \mathbb{E}_Q f(x) K_x$ and $\Sigma^Q_{Id,\mathcal{A}_1} = \mathbb{E}_Q K_x \otimes \mathcal{A}_1 K_x$ for arbitrary distribution Q and $\mathbb{E}_{\mathbb{P}}\xi(x,y) = (\Sigma_{Id,\mathcal{A}_1} + \lambda)^{-1/2}(\mathcal{A}_2 S^* f_\rho - \Sigma_{Id,\mathcal{A}_1} g_\lambda)$. We bound different terms (I), (II) and (III) separately and combine them to get the final upper bound. We show that (I) is the mean variance term and is at the scale $\frac{\mathcal{N}(\lambda)}{n} = \frac{\mathrm{Tr}(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1}\Sigma_{\mathcal{A}_2,\mathcal{A}_2}}{n}$ when the problem is regular. Term (II) and (III) is smaller than the bias. Our bound of term (III) bounds tighter than (Pillaud-Vivien, Rudi, and Bach 2018) (the second term, Lemma 10) via the spectral representation.

Bounding term (I). The term (I) is the concentration error of the random variable $\xi(x,y)$ and can be bounded via a Bernstein Inequality. We first bound term (I) via the following decomposition

$$
\begin{aligned}
&\left| \|\Sigma^{\frac{1-\gamma}{2}} q_\lambda(\hat{\Sigma}_{Id,\mathcal{A}_1})(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2} \left[\frac{1}{n}\sum_{i=1}^{n} (\xi(x_i,y_i) - \mathbb{E}_P\xi(x,y)) \right] \right| \|_H^2 \leq \|\Sigma^{\frac{1-\gamma}{2}} (\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1/2}\|^2 \left| \|\frac{1}{n}\sum_{i=1}^{n}(\xi(x_i,y_i) - \mathbb{E}_P\xi) \right| \|_H^2 \\
&\cdot \|(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2}(\hat{\Sigma}^\lambda_{Id,\mathcal{A}_1})^{-1/2}\|^2 \|(\hat{\Sigma}^\lambda_{Id,\mathcal{A}_1})^{1/2} q_\lambda(\hat{\Sigma}_{Id,\mathcal{A}_1})(\hat{\Sigma}^\lambda_{Id,\mathcal{A}_1})^{1/2}\|^2 \|(\hat{\Sigma}^\lambda_{Id,\mathcal{A}_1})^{-1/2}(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2}\|^2,
\end{aligned}
$$

where $\Sigma^\lambda_{Id,\mathcal{A}_1} = \Sigma_{Id,\mathcal{A}_1} + \lambda I$ and $\hat{\Sigma}^\lambda_{Id,\mathcal{A}_1} = \hat{\Sigma}_{Id,\mathcal{A}_1} + \lambda I$. At the same time, we knows $\|\Sigma^{\frac{1-\gamma}{2}}(\Sigma_{Id,\mathcal{A}_1} + \lambda)^{-1/2}\|^2 \leq \lambda^{-\frac{\gamma\alpha+p}{\alpha+p}}$ (From lemma 4) and $\|(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2}(\hat{\Sigma}^\lambda_{Id,\mathcal{A}_1})^{-1/2}\|^2 \leq 2$ (From lemma 5) with high probability. At the same time, we have

$$\|(\hat{\Sigma}^{\lambda}_{Id,\mathcal{A}_1})^{1/2} q_\lambda(\hat{\Sigma}_{Id,\mathcal{A}_1})(\hat{\Sigma}^{\lambda}_{Id,\mathcal{A}_1})^{1/2}\| = \sup_{\sigma\in\sigma(\hat{\Sigma}^{\lambda}_{Id,\mathcal{A}_1})} (\sigma+\lambda)q_\lambda(\sigma) \leq 2c_q.$$

Thus we only need to focus on bounding the concentration error $\frac{1}{n}\sum_{i=1}^n(\xi(x_i,y_i)-\mathbb{E}_P\xi)$. We recall the moment condition to control the noise of the observations. There are constants $\sigma, L>0$ such that

$$\int_{\mathbb{R}} |y-f^*(x)|^m P(dy|x) \leq \frac{1}{2}m!\sigma^2 L^{m-2}$$

is satisfied for μ–almost all $x\in X$ and all $m>2$. Note that the moment condition is satisfied for Gaussian noise with bounded variance or have a bounded observation noise. Then we can bound the second order momentum of the random variable $\xi(x,y) = (\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}(y\mathcal{A}_2K_x - \mathcal{A}_1 g_\lambda(x)K_x)$ via decomposing the random into three parts $(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}(y\mathcal{A}_2K_x - f^*(x)\mathcal{A}_2K_x)$, $(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}(f^*(x)\mathcal{A}_2K_x - \mathcal{A}_2f^*(x)K_x)$ and $(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}(\mathcal{A}_2f^*(x)K_x - \mathcal{A}_1g_\lambda(x)K_x)$. Base on the decomposition, we can bound the moments of random variable $\xi(x,y)$ as

$$\begin{aligned}
\mathbb{E}_P\|\xi(x,y)\|_H^m &= \int\left[\|(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}\mathcal{A}_2K_x\|_H^m\int_{\mathbb{R}}|y-f^*(x)|^m P(dy|x)\right] + \int\left[\|(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}\mathcal{A}_2K_x\|_H^m\|f\|_\infty^m\right]\\
&+\int\left[\|(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}K_x\|_H^m\int_{\mathbb{R}}|\mathcal{A}_2f^*(x)-\mathcal{A}_1g_\lambda|^m P(dy|x)\right]dv(x)\\
&\leq \frac{1}{2}m!\sigma^2(L+\|f\|_{\mathcal{H}})^m\|h_x^1\|_{\mathcal{H}}^{m-2}\mathrm{trace}((\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1}\Sigma_{\mathcal{A}_2,\mathcal{A}_2})\\
&+\|h_x^2\|_{\mathcal{H}}^{m-2}\|\mathcal{A}_2f^*(x)-\mathcal{A}_1g_\lambda\|_{L_\infty}^{m-2}\int|\mathcal{A}_2f^*(x)-\mathcal{A}_1g_\lambda|^2d\mu(x)\\
&\lesssim m!\left(\|h_x^1\|\right)^m\left[\sigma^2\mathrm{trace}((\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1}\Sigma_{\mathcal{A}_2,\mathcal{A}_2})\right] + m!\left(L_\lambda\|h_x^2\|\right)^{m-2}\left[\|h_x^2\|^2\|\mathcal{A}_2f^*(x)-\mathcal{A}_1g_\lambda\|_2^2\right]
\end{aligned}$$

where $L_\lambda = \|\mathcal{A}_2f^*(x)-\mathcal{A}_1g_\lambda\|_{L_\infty}$, $h_x^1 = (\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}\mathcal{A}_2K_x$ and $h_x^2 = (\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}K_x$. The two vectors' norms are bounded in Lemma 3 as $\mathcal{N}^1_\infty(\lambda) = \sup_{x\in\rho(x)}\|(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}K_x\|_H^2 \leq \|k_v^\alpha\|_\infty^2\lambda^{-\frac{\mu\alpha+p}{\alpha+p}}$, and $\mathcal{N}^2_\infty(\lambda) = \sup_{x\in\rho(x)}\|(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}K_x\|_H^2 \leq \|k_v^\alpha\|_\infty^2\lambda^{-\frac{\mu\alpha+p+2q}{\alpha+p}}$. At the same time, we know that $\|\mathcal{A}_1g_\lambda-\mathcal{A}_2f_\rho\|_2 \lesssim \lambda^{\frac{\beta\alpha-2p}{2(\alpha+p)}}\|u^*\|_{[H]^\beta}$ from Lemma 8 and $\mathrm{Trace}((\Sigma_{\mathcal{A}_1}+\lambda)^{-1}\Sigma_{Id,Id}) \leq D\lambda^{\frac{p-1}{p+\alpha}}$ from Lemma 2. Then using Bernstein Inequality (Theorem 3), we knows that with probability $1-2e^{-\tau}$

$$\begin{aligned}
&\|\frac{1}{n}\sum_{i=1}^n(\xi(x_i,y_i)-\mathbb{E}_P\xi(x,y))\|_H^2\\
&\lesssim \frac{32\tau^2}{n}\left(\sigma^2\mathrm{trace}((\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1}\Sigma_{\mathcal{A}_2,\mathcal{A}_2}) + \|h_x^2\|^2\|\mathcal{A}_2f^*(x)-\mathcal{A}_1g_\lambda\|_2^2 + \frac{L_\lambda\|h_x^2\|+\|h_x^1\|}{n}\right) \qquad (18)\\
&\lesssim \frac{\tau^2}{n}\left(\sigma^2(\lambda)^{-\frac{1}{p+\alpha}-\frac{p-2q}{p+\alpha}} + \lambda^{-\frac{\mu\alpha-p}{\alpha+p}}\lambda^{\frac{\alpha\beta-2p}{\alpha+p}} + \frac{L_\lambda\|h_x^2\|+\|h_x^1\|}{n}\right)
\end{aligned}$$

Thus we have the final bound $\left|\|\Sigma^{\frac{1-\gamma}{2}}q_\lambda(\hat{\Sigma}_{Id,\mathcal{A}_1})(\Sigma^{\lambda}_{Id,\mathcal{A}_1})^{1/2}\left[\frac{1}{n}\sum_{i=1}^n(\xi(x_i,y_i)-\mathbb{E}_P\xi(x,y))\right]\right|\|_H^2 \leq \frac{\tau^2}{n}\lambda^{-\frac{\gamma\alpha+p}{\alpha+p}}\left(\sigma^2(\lambda)^{-\frac{1}{p+\alpha}-\frac{p-2q}{p+\alpha}} + \lambda^{-\frac{\mu\alpha-p}{\alpha+p}}\lambda^{\frac{\alpha\beta-2p}{\alpha+p}} + \frac{L_\lambda\|h_x^2\|+\|h_x^1\|}{n}\right)$.

Remark 2. *In this remark, we'll bound the* $L_\lambda = \|\mathcal{A}_2 f^*(x) - \mathcal{A}_1 g_\lambda\|_{L_\infty}$ *here. For the embedding theorem of the* ℓ_∞, $L_\lambda \leq \|\mathcal{A}_2 f^*(x) - \mathcal{A}_1 g_\lambda\|_\mu \lesssim \lambda^{-\frac{(\mu-\beta)+\alpha}{\alpha+p}}$. *From Lemma 3, we know that* $\|h_x^1\|_H^2 \lesssim \lambda^{-\frac{\mu\alpha+p+2q}{\alpha+p}}$ *and* $\|h_x^2\|_H^2 \lesssim \lambda^{-\frac{\mu\alpha+p}{\alpha+p}}$.

Bounding term (III). At last we bound the term $\Sigma^{\frac{1-\gamma}{2}} r(\hat\Sigma_{Id,\mathcal{A}_1}) g_\lambda$ via the following decomposition

$$\begin{aligned}\|\Sigma^{\frac{1-\gamma}{2}} r(\hat\Sigma_{Id,\mathcal{A}_1}) g_\lambda\|_{\mathcal{H}} &= \|\Sigma^{\frac{1-\gamma}{2}} \underbrace{(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1/2}(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2}}_{Id} \underbrace{(\hat\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1/2}(\hat\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2}}_{Id} r(\hat\Sigma^\lambda_{Id,\mathcal{A}_1})(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1}\mathcal{A}_2 S^* f_\rho\|_{\mathcal{H}} \\ &\leq \|\Sigma^{\frac{1-\gamma}{2}}(\Sigma_{Id,\mathcal{A}_1})^{-1/2}\|\|(\Sigma_{Id,\mathcal{A}_1})^{1/2}(\hat\Sigma_{Id,\mathcal{A}_1})^{-1/2}\|\|(\hat\Sigma_{Id,\mathcal{A}_1})^{1/2} r(\hat\Sigma_{Id,\mathcal{A}_1})\|\|(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1}\mathcal{A}_2 S^* f_\rho\|,\end{aligned}$$

where we use $\Sigma^\lambda_{Id,\mathcal{A}_1}$ to denote $\Sigma_{Id,\mathcal{A}_1} + \lambda I$. From Lemma 4 we now that $\|\Sigma^{\frac{1-\gamma}{2}}(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1/2}\|^2 \leq \lambda^{-\frac{\gamma\alpha+p}{\alpha+p}}$. Then we bound the term $\|\Sigma^{\frac{1-\gamma}{2}} r(\hat\Sigma_{Id,\mathcal{A}_1}) g_\lambda\|_{\mathcal{H}}$ using $r_\lambda(x)x^u \lesssim \lambda^u$ and get

$$\|(\hat\Sigma_{Id,\mathcal{A}_1})^{1/2} r(\hat\Sigma_{Id,\mathcal{A}_1})\| = \sup_{\sigma\in\sigma(\hat\Sigma^\lambda_{Id,\mathcal{A}_1})} (\sigma+\lambda)^{1/2} r_\lambda(\sigma) \leq \lambda^{1/2}. \tag{19}$$

At the same time, we can bound $\|(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1}\mathcal{A}_2 S^* f_\rho\|$ using the spectral representation

$$\begin{aligned}\|(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1}\mathcal{A}_2 S^* f_\rho\|^2 &= \sum_{i=1}^\infty \frac{\mu_i p_i^2 a_i^2}{(\lambda+\mu_i p_i)^2} \lesssim \left(\sup_{i\geq 1} \frac{\mu_i p_i^2 \mu_i^\beta}{(\lambda+\mu_i p_i)^2}\right) \sum_{i\geq 1} \mu_i^{-\beta} a_i^2 \\ &\leq \left(\lambda^{\frac{(\frac{1-\beta}{2})\alpha+p}{\alpha+p}-1}\right)^2 \|u^*\|^2_{[H]^\beta} \leq \lambda^{\frac{\beta}{\alpha+p}-1}\|u^*\|^2_{[H]^\beta}\end{aligned} \tag{20}$$

Thus we know that

$$\|(\hat\Sigma_{Id,\mathcal{A}_1})^{1/2} r(\hat\Sigma_{Id,\mathcal{A}_1})\| \lesssim \lambda^{-\frac{\gamma\alpha+p}{2(\alpha+p)}}\lambda^{1/2}\lambda^{\frac{\beta}{2(\alpha+p)}-\frac{1}{2}} \leq \lambda^{-\frac{(\beta-\gamma)\alpha}{2(\alpha+p)}},$$

where the last inequality is because $p < 0$ in our assumption.

Bounding term (II). In this paragraph, we demonstrate the proof to bound the term

$$\Sigma^{\frac{1-\gamma}{2}} q_\lambda(\hat\Sigma_{Id,\mathcal{A}_1})(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2}\mathbb{E}_P\xi(x,y) = \Sigma^{\frac{1-\gamma}{2}} q_\lambda(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2}(\hat\Sigma_{Id,\mathcal{A}_1})(\Sigma_{Id,\mathcal{A}_1}+\lambda)^{-1/2}(\mathcal{A}_2 S^* f_\rho - \Sigma_{Id,\mathcal{A}_1} g_\lambda).$$

Note that $\Sigma_{Id,\mathcal{A}_1}(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1} = I - \lambda(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1}$, thus we knows that $\Sigma^{\frac{1-\gamma}{2}} q_\lambda(\hat\Sigma_{Id,\mathcal{A}_1})(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2}(\Sigma_{Id,\mathcal{A}_1} + \lambda)^{-1/2}(\mathcal{A}_2 S^* f_\rho - \Sigma_{Id,\mathcal{A}_1} g_\lambda) = \lambda\Sigma^{\frac{1-\gamma}{2}} q_\lambda(\hat\Sigma_{Id,\mathcal{A}_1})(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1}\mathcal{A}_2 S^* f_\rho$. At the same time, according to our assumption on the spectral filter q_λ, we know that

$$\|(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2} q(\hat\Sigma^\lambda_{Id,\mathcal{A}_1})(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2}\| = \sup_{\sigma\in\sigma(\hat\Sigma^\lambda_{Id,\mathcal{A}_1})} (\sigma+\lambda) q_\lambda(\sigma) \leq 2c_q.$$

Thus we can bound $\Sigma^{\frac{1-\gamma}{2}} q_\lambda(\hat{\Sigma}_{Id,\mathcal{A}_1})\mathbb{E}_P\xi(x,y)$ via the following decomposition

$$\|\Sigma^{\frac{1-\gamma}{2}} q_\lambda(\hat{\Sigma}_{Id,\mathcal{A}_1})\mathbb{E}_P\xi(x,y)\| = \lambda\|\Sigma^{\frac{1-\gamma}{2}} q_\lambda(\hat{\Sigma}_{Id,\mathcal{A}_1})(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1}\mathcal{A}_2 S^* f_\rho\|$$
$$= \lambda\|\Sigma^{\frac{1-\gamma}{2}} \underbrace{(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1/2}(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2}}_{Id}\underbrace{(\hat{\Sigma}^\lambda_{Id,\mathcal{A}_1})^{-1/2}(\hat{\Sigma}^\lambda_{Id,\mathcal{A}_1})^{1/2}}_{Id} q(\hat{\Sigma}^\lambda_{Id,\mathcal{A}_1})(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2}(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-3/2}\mathcal{A}_2 S^* f_\rho\|_{\mathcal{H}}$$
$$\lesssim \lambda\|\Sigma^{\frac{1-\gamma}{2}}(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1/2}\|\|(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1}\|\|(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2}\|\|(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1}\mathcal{A}_2 S^* f_\rho\|$$
$$\lesssim \lambda\lambda^{-\frac{\gamma\alpha+p}{2(\alpha+p)}}\lambda^{-1}\lambda^{1/2}\lambda^{\frac{\beta}{2(\alpha+p)}-\frac{1}{2}} \lesssim \lambda^{-\frac{(\beta-\gamma)\alpha}{2(\alpha+p)}}$$

The last line is because of $\|\Sigma^{\frac{1-\gamma}{2}}(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1/2}\|^2 \le \lambda^{-\frac{\gamma\alpha+p}{\alpha+p}}$ (Lemma 4), $\|(\Sigma^\lambda_{Id,\mathcal{A}_1})^{1/2}(\hat{\Sigma}^\lambda_{Id,\mathcal{A}_1})^{-1/2}\| \le 2$ with high probability (Lemma 5), $\|(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1}\| \le \lambda^{-1}$, $\|(\Sigma^\lambda_{Id,\mathcal{A}_1})^{-1/2}\mathcal{A}_2 S^* f_\rho\| \le \lambda^{\frac{\beta}{2(\alpha+p)}-\frac{1}{2}}$ (proved while bounding term (III)) and $p < 0$.

Appendix 2.2 Final Bound

At this time we can have our final bound in Theorem 2 via combining the bound for bias (Appendix Appendix 2.1.2) and (Appendix Appendix 2.1.3)

$$\begin{aligned}\|\hat{q}_\lambda - u^*\|_\gamma^2 &\lesssim \|\hat{q}_\lambda - g_\lambda\|_\gamma^2 + \|\hat{q}_\lambda - u^*\|_\gamma^2 \\ &\lesssim \lambda^{\frac{(\beta-\gamma)\alpha}{\alpha+p}} + \frac{\tau^2}{n}\lambda^{-\frac{\gamma\alpha+p}{\alpha+p}}\left(\sigma^2(\lambda)^{-\frac{1}{p+\alpha}-\frac{p-2q}{p+\alpha}} + \lambda^{-\frac{\mu\alpha-p}{\alpha+p}}\lambda^{\frac{\alpha\beta-2p}{\alpha+p}} + \frac{L_\lambda\|h_x^2\|+\|h_x^1\|}{n}\right) \\ &\lesssim \lambda^{\frac{(\beta-\gamma)\alpha}{\alpha+p}} + \frac{\lambda^{-\frac{\gamma\alpha+2(p-q)+1}{p+\alpha}}}{n} + \frac{\lambda^{\frac{(\beta-\gamma)\alpha-\mu\alpha-p}{\alpha+p}}}{n} + \frac{\lambda^{-\frac{\mu\alpha+p+2q}{\alpha+p}}\lambda^{-\frac{\mu\alpha+p+2q}{\alpha+p}}}{n^2}\end{aligned} \tag{21}$$

Case 1. $\beta \le \frac{\mu\alpha+2q-p+1}{\alpha}$ In this situation, $\frac{\lambda^{-\frac{\gamma\alpha+2(p-q)+1}{p+\alpha}}}{n}$ is larger than $\frac{\lambda^{\frac{(\beta-\gamma)\alpha-\mu\alpha-p}{\alpha+p}}}{n}$. Thus $\lambda^{\frac{(\beta-\gamma)\alpha}{\alpha+p}} + \frac{\lambda^{-\frac{\gamma\alpha+2(p-q)+1}{p+\alpha}}}{n}$ is the dominating term of the loss upper bound. Thus we can take $\lambda = n^{-\frac{\alpha+p}{\beta\alpha+2(p-q)+1}}$ and leads to $n^{-\frac{(\beta-\gamma)\alpha}{\beta+2(p-q)+1}}$ upper bound. At the same time, the third term is dominated by the second term.

Case 2. $\beta > \frac{\mu\alpha+2q-p+1}{\alpha}$ In this situation, $\frac{\lambda^{\frac{(\beta-\gamma)\alpha-\mu\alpha-p}{\alpha+p}}}{n}$ is larger than $\frac{\lambda^{-\frac{\gamma\alpha+2(p-q)+1}{p+\alpha}}}{n}$. Thus $\lambda^{\frac{(\beta-\gamma)\alpha}{\alpha+p}} + \frac{\lambda^{\frac{(\beta-\gamma)\alpha-\mu\alpha-p}{\alpha+p}}}{n}$ is the dominating term of the loss upper bound. Thus we can take $\lambda = n^{-\frac{\alpha+p}{\mu\alpha+p}}$ and leads to $n^{-\frac{(\beta-\gamma)\alpha}{\mu\alpha+p}}$ upper bound. At the same time, the third term is also dominated by the second term.

Appendix 3. Proof of the Lower Bound

Appendix 3.1 Preliminaries on Tools for Lower Bounds

In this section, we repeat the standard tools we use to establish the lower bound. The main tool we use is the Fano's inequality and the Varshamov-Gilber Lemma.

Lemma 9 (Fano's methods)**.** *Assume that V is a uniform random variable over set $\mathcal{V}$, then for any Markov chain $V \to X \to \hat{V}$, we always have*

$$\mathcal{P}(\hat{V} \neq V) \ge 1 - \frac{I(V;X)+\log 2}{\log(|\mathcal{V}|)}$$

Lemma 10 (Varshamov-Gillbert Lemma,(Alexandre B Tsybakov 2008) Theorem 2.9)**.** *Let $D \geq 8$. There exists a subset $\mathcal{V} = \{\tau^{(0)}, \cdots, \tau^{(2^{D/8})}\}$ of D–dimensional hypercube $\mathcal{H}^D = \{0,1\}^D$ such that $\tau^{(0)} = (0,0,\cdots,0)$ and the ℓ_1 distance between every two elements is larger than $\frac{D}{8}$*

$$\sum_{l=1}^{D} \|\tau^{(j)} - \tau^{(k)}\|_{\ell_1} \geq \frac{D}{8}, \text{for all } 0 \leq j,k \leq 2^{D/8}$$

Appendix 3.2 Proof of the Lower Bound

Theorem 4. *Let (X, B) be a measurable space, H be a separable RKHS on X respect to a bounded and measurable kernel k and operator $\mathcal{A} = (\mathcal{A}_2^{-1}\mathcal{A}_1)$ satisfies Assumption 1. We have n random observations $\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1}^n$ of $f^* = \mathcal{A}u, u \in \mathcal{H}^\gamma \cap L_\infty$,* i.e. *$y_i = f^*(x_i) + \eta_i$ where η_i is a random noise satisfies the momentum assumption $\mathbb{E}|\eta|^m \leq \frac{1}{2}m!\sigma^2 L^{m-2}$ for some constant $\sigma, L > 0$. Then for all estimators $H : (\mathcal{X} \times \mathcal{Y})^{\otimes n} \to \mathcal{H}^\gamma$ satisfies*

$$\inf_{H} \sup_{u^* \in \mathcal{H}^\beta \cap L_\infty} \mathbb{E}\|H(\{(x_i, y_i)\}_{i=1}^n) - u^*\|_\gamma^2 \gtrsim n^{-\frac{(\max\{\beta,\mu\}-\gamma)\alpha}{\max\{\beta,\mu\}\alpha+2(q-p)+1}}$$

Proof. To proof the lower bound, we use the standard Fano methods via reducing the lower bound to multiple hypothesis testing. We construct our hypothesis using binary strings $\omega = (\omega_1, \cdots, \omega_m) \in \{0,1\}^m$ (m to be determined later) by defining

$$u_\omega = \left(\frac{\epsilon}{m}\right)^{1/2} \sum_{i=1}^{m} \omega_i \mu_{i+m}^{\gamma/2} e_{i+m}.$$

If we control $m \lesssim \epsilon^{-\frac{1}{\alpha\beta-\alpha\gamma}}$, then we can always keep $u_\omega \in \mathcal{H}^\beta$ for $\|u_\omega\|_\beta^2 = \frac{\epsilon}{m}\sum_{i=1}^m \omega_i^2 \mu_{i+m}^{-(\beta-\gamma)} \lesssim \epsilon\mu_{2m}^{-(\beta-\gamma)} \lesssim m^{\alpha(\beta-\gamma)}\epsilon = O(1)$. Similarly, we can select $m \lesssim \epsilon^{-\frac{1}{\alpha\mu-\alpha\gamma}}$ to control $\|u_\omega\|_{L^\infty} \leq \|u_\omega\|_\mu \leq O(1)$. At the same time, the associated PDE right hand side function$f_\omega = \mathcal{A}_2^{-1}\mathcal{A}_1 u_\omega = \left(\frac{\epsilon}{m}\right)^{1/2} \sum_{i=1}^m \frac{q_i \omega_i}{p_i} \mu_{i+m}^{\gamma/2} e_{i+m}$.

Using Gilbert-Varshamov Lemma we know that there exists $M \geq 2^{m/8}$ binary strings $\omega^{(1)}, \cdots, \omega^{(k)} \in \{0,1\}^m$ with $\omega^{(0)} = (0, \cdots, 0)$ subject to

$$\sum_{i=1}^{m} \left(\omega_i^{(j)} - \omega_i^{(k)}\right)^2 \geq m/8$$

holds for all $j \neq k$. As consequence, the distance between f_ω and $f_{\omega'}$ can be lower bounded as $\|u_\omega - u_{\omega'}\|_\gamma^2 = \frac{\epsilon}{m}\sum_{i=1}^m (\omega_i - \omega_i')^2 \geq \epsilon/8$. To apply the Fano method, we still need to bound the mutual information between the uniform distribution over all the hypothesis and the distribution of the observed data. We take η_i is sampled form $\mathcal{N}(0, \min\{\sigma, L\}^2)$ which satisfies the momentum condition. Then we know that this mutual information can be bounded by the following average of KL divergence(Alexandre B Tsybakov 2008) via

$$I(V, X) = \frac{1}{M_\epsilon}\sum_{j=1}^{M_\epsilon} KL(P_j^n \| P_0^n) = \frac{n}{2\bar{\sigma}^2 M_\epsilon}\sum_{j=1}^{M_\epsilon} \|f_j - f_0\|_{L_2}^2 \lesssim n\epsilon m_\epsilon^{-\alpha\gamma+2(p-q)} \tag{22}$$

Then we apply the Fano's inequality

$$\mathbb{P}(\hat{V} \neq V) \geq 1 - \frac{I(V;X) + \log 2}{\log |V|} = 1 - \frac{\frac{16C^{\gamma}}{\min\{\sigma,L\}^2} n\epsilon m_{\epsilon}^{-\alpha\gamma-2(p-q)} + \log 2}{\frac{\log 2}{8} m_{\epsilon}}$$
$$= 1 - O\left(n\epsilon\epsilon^{\frac{1+\alpha\gamma+2(p-q)}{\alpha(\max\{\beta,\mu\}-\gamma)}}\right)$$

Take $\epsilon = n^{-\frac{(\max\{\beta,\mu\}-\gamma)\alpha}{\max\{\beta,\mu\}\alpha+2(p-q)+1}}$, we know that with constant probability we have

$$\|H(\{(x_i, y_i)\}_{i=1}^n) - u^*\|_{\gamma}^2 \gtrsim n^{-\frac{(\max\{\beta,\mu\}-\gamma)\alpha}{\max\{\beta,\mu\}\alpha+2(p-q)+1}}$$

$\square$