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

Yiping Lu
ICME
Stanford University
Stanford, CA 94305
yplu@stanford.edu

Jose Blanchet
Management Science and Engineering
Stanford University
Stanford, CA 94305
jose.blanchet@stanford.edu

Lexing Ying
Department of Mathematics
Stanford University
Stanford, CA 94305
lexing@stanford.edu

Abstract

In this paper, we study the statistical limits in terms of Sobolev norms of gradient descent for solving inverse problems 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.

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 [1], inverse scattering [2] and 3D reconstruction [3]. In this paper, we study the statistical limit of machine learning methods for 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 [4, 5]. The observations may also come from certain physical laws described by a partial differential equation (PDE) [6, 7]. 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 \( u^* \) through a measurement procedure \( A \). For simplicity, we assume \( A \) 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 \( A \), i.e. \( f^* = Au^* \). To solve this problem, we consider gradient descending over the following general class of objective function

\[
\hat{u} = \arg \min_{u \in \mathcal{H}} \mathbb{E}_{\mathbb{P}_n(x, y)} \frac{1}{2} \langle u(x), A_1 u(x) \rangle - \langle y, A_2 u(x) \rangle,
\]

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 \( A_i, i = 1, 2 \) are two self-adjoint operators that satisfy \( A_1 = A A_2 \). In Section 2, we show that several algorithms, including kernel regression [4, 8] via Sobolev training [9, 10, 11] and solving PDEs via machine learning based algorithm, [12, 13, 2, 14] can be considered as special cases of this formulation.

Recent works [15, 16] 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 [17, 18, 19, 20], 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 [21, 22], detailed definition see Section 2).

We first prove that a properly early stopped gradient descent algorithm over the class of objective functions can achieve statistical optimality. At the same time, although any suitably early stopped gradient flow of the class of loss function can achieve statistical optimality according to our theory, we discover an acceleration effect of using a Sobolev norm as the loss function for kernel based machine learning algorithms. The implicit acceleration of the Sobolev loss function arises because a differential operator can enlarge the small eigenvalue of the kernel integral operator for high frequency functions, leading to better condition numbers and faster convergence in these eigenspaces while maintaining 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 breakthroughs in deep learning and the rapid development of sensors, computational power, and data storage in the past decade have drawn attention to numerically solving PDEs via machine learning methods [23, 24, 12, 25, 13, 2], 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. [25, 26] utilize the Feynman-Kac formulation which turns solving a PDE into a stochastic control problem. The work of [27] 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) [14, 2] and the Physics-Informed neural network (PINN) [12, 13]. DRM [14, 2] 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 [12, 13] 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, [28–29, 30, 31, 52, 33, 34] investigated the regularity of PDEs approximated by a neural network and [28–35, 56, 57, 58] further provided generalization analyses. [15, 16, 39, 40] 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 [41], 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 [42, 43, 44], i.e., we consider the statistical optimality of the implicit regularization effect of optimization algorithm. A concurrent paper [45] considered similar stochastic gradient descent approach for statistical inverse problem.

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 [4, 46, 47, 48]. Statistical optimality of early stopped (stochastic) gradient descent has been widely discussed in [42–49, 50, 18, 17, 51, 52]. The convergence of least square regression in Sobolev norm has been discussed recently in [21, 22]. Recently, training neural networks with stochastic gradient descent in certain regimes has been found to be equivalent to kernel regression [53, 54, 55]. Gradient descent training of neural network in the kernel regime has been found optimal for a wide class of non-parametric functions with both early stopping regularization and ridge regression [56, 57].

1.2 Contribution

- We provide information theoretical lower bounds (Theorem 3.1) for a wide class of inverse problems, including the Sobolev learning rate [21] for the solution of elliptic inverse
problems. We also show that the previous lower bound \([15,16]\) 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 3.2), including PINN \([12,13]\) and Deep Ritz Methods \([14,2]\) for solving PDEs as well as Sobolev training \([11,9,58]\) of kernel methods. We provide \([16]\) a computational feasible estimator and generalize the previous statistical optimality results of gradient descent \([42,18,19]\) 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 \(f^* \in \mathbb{R}^X\) from random observations of \(u^* = A f^*\), where \(A\) is an observation process which is modeled by an operator maps from \(\mathbb{R}^X\) to \(\mathbb{R}^X\). To solve this problem, we write the operator \(A\) in terms of two operators \(A_i\) (\(i = 1, 2\)) with \(A_1 = A_2A_2\) and build our objective function as

\[
\mathbb{E}_\mathcal{P} \left[ \frac{1}{2} \langle u(x), A_1 u(x) \rangle - \langle y, A_2 u(x) \rangle \right],
\]

where \(\mathcal{P}\) 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 \(f(x) = (Au)(x)\). In other words, \(\mathbb{E}(y|x) = f(x)\). The minimizer of objective function \((\cdot)\) is the ground truth function \(u^* = A^{-1}f\) that we are interested in.

Learning with Kernel

Consider the case that \(u\) is parameterized by a Reproducing Kernel Hilbert Space \(u_0(x) = (\theta, K_x)\) (we provide standard notations of RKHS in Appendix A). At the same time, the kernel function has the following representation \(K(s,t) = \sum_{i=1}^n \lambda_i e_i(s)e_i(t)\), where \(e_i\) are orthogonal basis of \(L_2(\rho_X)\) with \(\rho_X\) being the uniform distribution over \(\mathcal{X}\), where \(L_2\) denotes the space of all the square integrable functions. Then \(e_i\) is also the eigenvector of the covariance operator \(\Sigma = \mathbb{E}_{x \sim \mathcal{H}} K_x \otimes K_x\) 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 \(g \otimes h : f \mapsto \langle f, h \rangle\), and \(g\). The covariance matrix \(\Sigma\) is the core of the integral operator technique \([46,8]\) for kernel regression. For any \(f \in \mathcal{H}\), the reproducing property gives \((\Sigma f)(z) = \langle K_z, \Sigma f \rangle_{\mathcal{H}} = \mathbb{E}[f(X)k(X,z)] = \mathbb{E}[f(X)K_z(X)]\). If we consider the mapping \(S : \mathcal{H} \rightarrow L_2(dx)\) defined as a parameterization of a vast class of functions in \(\mathbb{R}^X\) via \(\mathcal{H}\) through the mapping \((Sg)(x) = \langle g, K_x \rangle (\Phi(x) = K_x = K(\cdot,x))\). Its adjoint operator \(S^* : L_2 \rightarrow \mathcal{H}\) then can be defined as \(g \mapsto \int_X g(x)K_x\rho_X(dx)\). \(\Sigma\) is the same as the self-adjoint operator \(S^* S\) and the self-adjoint operator \(\mathcal{L} = SS^* : L_2(dx) \rightarrow L_2(dx)\) can be defined as \((\mathcal{L}f)(x) = \int_X K(x,z)f(z)\rho_X(dz)\). Based on this notation, we present all our assumptions on the underlying kernel.

**Assumption 2.1** (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 \([8]\). We further assume a regularity condition for our kernel \(k(\cdot,\cdot)\) via an \(\ell_\infty\) embedding property follows \([59,60,18,21]\). 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 \(Q = \text{tr}(\Sigma^{1/\alpha}) < \infty\).
- **(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 \(u^*(x) = \langle \theta_x, K_x \rangle_{\mathcal{H}}\), the source condition can also be written as

\[
\|\Sigma^{1/\alpha} \theta_x\|_{\mathcal{H}} < \infty.
\]
\[ \mathcal{A}_1 = \sum_{i=1}^{\infty} p_i e_i \otimes e_i, \mathcal{A}_2 = \sum_{i=1}^{\infty} q_i e_i \otimes e_i \]

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 [61][62], due to the Bochner’s theorem. We further assume \( p < 0, q < 0, \alpha + p > 0 \). We refer the detailed discussion to Remark 4.

- (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

\[ \mathcal{A}_1 = \sum_{i=1}^{\infty} p_i e_i \otimes e_i, \mathcal{A}_2 = \sum_{i=1}^{\infty} q_i e_i \otimes e_i \]

Remark 1. To simplify the technical exposition, we assume that operator \( \mathcal{A}_i (i = 1, 2) \) commute with the kernel covariance operator \( \Sigma \). This assumption is also made in [63]. 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. [63] assumes the operator \( \mathcal{A}_i \) to be bounded operator in operator norm, which can be consider as a special case of ours. At the same time, for numerically solving a PDE/elliptic inverse problem, we take \( \mathcal{A}_1 \) to become the power of the Laplace operator \( \Delta \), which contracts with \( \Sigma \)'s assumption. 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 [64] and the Neural Tangent Kernel [65][66], 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 \( K(x, y) = \sum_{i=1}^{\infty} \psi(w)e^{iw(e-x-y)} \), we know that the eigen-functions are also Fourier modes. There are also works that use Green function as the kernel [67][68], 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 [69][21].

Definition 2.2 (Sobolev Norm). For \( \gamma > 0 \), the \( \gamma \)-power space is

\[ \mathcal{H}^\gamma := \left\{ \sum_{i \geq 1} a_i \lambda_i^{\gamma/2} e_i : \sum_{i \geq 1} a_i^2 \leq \infty \right\} \subset L_2(v), \]

equipped with the \( \gamma \)-power norm via \( \| \sum_{i \geq 1} a_i \lambda_i^{\gamma/2} e_i \|_\gamma := \left( \sum_{i \geq 1} a_i^\gamma \right)^{1/2}. \)

It is obvious that \( \| L^{\gamma/2} f \|_\gamma = \| f \|_{L_2} \) and \( \| f \|_\gamma \leq \| \Sigma^{1/2} f \|_\mathcal{H} \) [21]. The source condition can also be understood as the target function \( u^* \) lies in the \( \beta \)-power Sobolev space. The regularity condition of the kernel function implies a continuously embedding from \( \mathcal{H}^\gamma \rightarrow L_\infty \). Throughout this paper, we consider the convergence rate of \( \hat{u} - u^* \) in \( \gamma \)-power Sobolev norm \( (\gamma > 0) \).

2.1 Examples

Sobolev Training [70][9][11] 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 [71][72][10][73][74][75]. The work of [71] 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 [73][74][75]. 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

\[ -\Delta u + u = f \quad \text{in } \Omega. \]
We mainly focus on analyzing Deep Ritz Method (DRM) \[14\] and Physics Informed Neural Network (PINN) \[12, 13\]. DRM solves the equation (2) via minimizing the following variational form
\[ u^* = \arg \min_{u \in \mathcal{F}} E_{\text{DRM}}(u) := \frac{1}{2} \int_{\Omega} |\nabla u|^2 + u^2 \, dx - \int_{\Omega} f u \, dx, \] (3)
while PINNs solves the equation (4) via minimizing the following strong formula, i.e the residual of the PDE,
\[ u^* = \arg \min_{u \in \mathcal{F}} E_{\text{PINN}}(u) := \frac{1}{2} \int_{\Omega} (\Delta u - u + f)^2 \, dx, \] (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 \[76, 77\]. \[16\] 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, \( A_1 u = \Delta u \) and \( A_2 u = \Delta u \) for all function \( u \in \mathbb{R}^n \). For PINN, \( A_1 u = \Delta^2 u \) and \( A_2 u = \Delta u \) for all function \( u \in \mathbb{R}^n \).

We discover that PINN convergences faster than DRM consistently due to the implicit Sobolev acceleration, matching the observation made in \[78\]. \[61\] considered semi-supervised learning using Laplacian regularization with kernel parameterization. However, this paper does not consider training acceleration, matching the observation made in \[78\].

### 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 \[15, 16\], we formulate the problem in an RKHS. This leads to a different construction of hypothesis and show that \[15, 16\] is a special case of our lower bound using specific kernel and operator \( A_i (i = 1, 2) \) in Section 3.3.

**Theorem 3.1 (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 \( A = (A_2^{-1}, A_1) \) satisfies Assumption [2.1] We have \( n \) i.i.d. random observations \( \{(x_i, y_i) \in X \times Y\}_{i=1}^n \) of \( f^* = Au, u \in H^\gamma \cap L_\infty \), i.e. \( y_i = f^*(x_i) + \eta_i \) where \( \eta_i \) is a mean zero random noise satisfies the momentum assumption \( \mathbb{E}[\eta_i]^m \leq \frac{1}{m!} \sigma^2 L^{m-2} \) for some constants \( \sigma, L > 0 \). Then for all estimators \( H : (X \times Y)^n \rightarrow H^\gamma \) satisfies
\[ \inf_{H} \sup_{u^*} \mathbb{E}[\|H(\{(x_i, y_i)\}_{i=1}^n) - u^*\|_\gamma \geq n^{-\frac{\gamma}{\max\{\gamma, \gamma_n\} + \frac{\gamma_n}{\nu}}}. \]

#### 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 \( \sum_{i=1}^n \frac{1}{2} u(x_i) A_1 u(x_i) - y_i A_2 u(x_i) \), 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 \( A_i u(x) = \langle A_i \theta, K_x \rangle \) \( H = \langle \theta, A_i K_x \rangle \) and express our empirical objective function as
\[ \mathbb{E}_{p_{i}(x,y)} \frac{1}{2} \langle u(x), A_1 u(x) \rangle - \langle y, A_2 u(x) \rangle = \mathbb{E}_{p_{i}(x,y)} \frac{1}{2} \langle u, K_x \rangle \langle A_1 u, K_x \rangle - \langle y, A_2 u, K_x \rangle \]

\[ = \mathbb{E}_{p_{i}(x,y)} \frac{1}{2} \langle u, K_x \otimes A_1 K_x u \rangle - \langle y, A_2 K_x \rangle \]

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

- **Initialization:** Set \( \theta_0 = \theta_0 = 0, \gamma \) 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
\[ \theta_t = \theta_{t-1} + \gamma \sum_{i=1}^n \left( y_i A_2 K_{x_i} - \langle \theta_{t-1}, A_1 K_{x_i} \rangle \right) K_{x_i} \]

with averaging step \( \tilde{\theta}_t = (1 - \frac{1}{t}) \theta_{t-1} + \frac{1}{t} \theta_t \).
Remark. Note that the optimizing dynamics considered here is not the exact gradient descent dynamics over the empirical objective. The gradient of the quadratic term \( \frac{1}{n} \sum_{i=1}^{n} u(x_i) A_1 u(x_i) \) should be \( \frac{1}{n} \sum_{i=1}^{n} (\theta_{t-1} A_1 K_{x_i} + \theta_{t-1} K_{x_i}) K_{x_i} \) but we take instead \( \frac{1}{n} \sum_{i=1}^{n} (\theta_{t-1} A_1 K_{x_i} + \theta_{t-1} K_{x_i}) H K_{x_i} \) 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 \( A_1 \). Without our modification, the statistical rate will become sub-optimal in some cases due to the fact that the variance in the empirical covariance matrix dominates the statistical rate. This observation matches the reason behind the sub-optimality of the Deep Ritz Method discovered in [16].

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

**Theorem 3.2.** Under Assumption [27] we have the following three regimes shown in Figure [7]

- For \( \beta > \frac{\alpha + 2g - p - 1}{2} \alpha \), if we take \( t = n \) and \( \gamma = n^{-\frac{(\alpha + p - 1)}{\alpha + p + 2(p-q)+1}} \), we obtain the following rate
  \[
  \mathbb{E}[||\hat{\theta}_t - u^*||_2^2] = O(n^{-\frac{(\beta - \gamma)\alpha}{\alpha + 2(p-q)+1}}).
  \]

- For \( \alpha \leq \frac{\alpha + 2g - p - 1}{2} \alpha \leq \frac{\mu \alpha}{\alpha + 2(p-q)+1} \), if we take \( t = \eta n^{-\frac{(\alpha + p - 1)}{\alpha + p + 2(p-q)+1}} \) and \( \gamma \) a small enough constant, we obtain the following rate
  \[
  \mathbb{E}[||\hat{\theta}_t - u^*||_2^2] = O(n^{-\frac{(\beta - \gamma)\alpha}{\alpha + 2(p-q)+1}}).
  \]

- For \( \beta > \frac{\mu \alpha}{\alpha + 2(p-q)+1} \), if we take \( t = \eta n^{-\frac{(\alpha + p - 1)}{\alpha + p + 2(p-q)+1}} \) and \( \gamma \) a small enough constant, we obtain the following rate
  \[
  \mathbb{E}[||\hat{\theta}_t - u^*||_2^2] = O(n^{-\frac{(\beta - \gamma)\alpha}{\alpha + 2(p-q)+1}}),
  \]
  which is not an optimal converging rate.

**Sketch of the Proof.** We first rewrite the averaged gradient descent in a more compact formula as \( \eta_0 = 0, \eta_p = \eta_{p-1} + \gamma(A_1 \hat{S}_t g - \hat{\Sigma}_{l,1} \eta_{l-1}) \) where \( \hat{S}_n : \mathcal{H} \rightarrow \mathbb{R}^n \) is defined as \( \hat{S}_n g = \frac{1}{\sqrt{n}} (g(x_1), \cdots, g(x_n)) \), \( \hat{\Sigma}_{l,1} \) is a small enough constant, we obtain

\[
\mathbb{E}[||\hat{\theta}_t - u^*||_2^2] = O(n^{-\frac{(\beta - \gamma)\alpha}{\alpha + 2(p-q)+1}}),
\]

which is not an optimal converging rate.

3.3 Discussion and Implications of Our Theory

**Relationship with [15,16]**. [15,16] provided a lower bound of the form \( n^{-\frac{2d - 2}{2 + d - 2}} \) for a 2t-th order linear PDE \( \Delta^t u = f \) with solution in \( H^2 \), evaluated in \( H^1 \) norm. We shall discuss the relationship between their bound with our \( n^{-\frac{2d - 2}{2 + d - 2}} \) lower bound based on the kernel representation of Sobolev spaces. The numerator \( (\beta - \gamma) \) matches the \( \alpha - s \) term in [15,16]'s lower bound and the \( q - p \) term is the order of the linear PDE which matches the \( t \) term in the denominator in [15,16]'s lower bound. The spectral decay speed of kernel \( \alpha \) 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 \( \mathbb{T}^d = [0,1]^d \). We consider the space of square integrable functions on \( \mathbb{T}^d \) with mean 0 and the Matérn kernel \( K_{\sigma,\ell,v}(x,y) = \sigma^2 2^{\ell-1} (\frac{|x-y|}{\ell})^\ell B_v (\frac{|x-y|}{\ell}) \), where \( B_v \) is the modified Bessel function of section kind. The covariance operator is \( C_\theta = \sigma^2 (\Delta + \tau^2 I)^{-s} \) with orthonormal eigenfunctions \( \phi_m(x) = e^{2\pi i m \langle m, x \rangle} \) and corresponding eigenvalues \( \lambda_m = \sigma^2 (4\pi^2 |m|^2 + \tau^2)^{-s} \) for every \( m \in \mathbb{Z}^d \setminus \{0\} \) [81].
Figure 1: Phase diagram of different regimes for solving inverse problem using stochastic gradient descent. Except for the gray area, GD can achieve the information theoretic rate.

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, \([82, 66, 65, 56]\) proved that the corresponding \( \alpha = \frac{d - 1}{d} \) and the eigenfunctions are spherical harmonics that diagonalize the differential equation.

For the upper bound, \([16]\) established the convergence rate based on the empirical process technique \([48, 59]\), while our paper switches to the integral operator/inverse problem technique \([4, 46, 8]\). 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 \([70]\) \([70]\) 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 \([70]\). However, we introduce a commuting assumption over the differential operator with the kernel integral operator that facilitates our analysis.

Sobolev Implicit Acceleration Below we discuss the implication of the choice of early stopping time \( t = n^{\frac{\alpha + p}{\alpha - 2p - \gamma}} \). First of all, the best early stopping time here does not depend on \( \gamma \), 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 \([83, 84, 85]\). 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 \([86, 87]\) 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 \( \ell_2 \) norm. However the preconditioning matrix in \([86]\) 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 \( \hat{\Sigma}_{Id, A_I} \) and the population one \( \Sigma_{Id, A_I} \) dominates. With the observation that these concentrations have no relationship with the supervision signal, \([16, 80]\) proposed to utilize the semi-supervised learning to reduce the error in this regime. In \([16]\), 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 [16]. However [16] 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 [9] and solving PDEs [11, 58]. 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

$$\int |\nabla u - \nabla f|^2 dx = \int \|\nabla u\|^2 + 2\Delta u \cdot f + \|\nabla f\|^2 dx, \quad (6)$$

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 [9], 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 [9]. 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 $\int \lambda \|u - f\|^2 + \|\nabla u\|^2 + 2\Delta u \cdot f + \|\nabla f\|^2 dx$ 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.

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 [12, 58] in 3 dimensions. The example is a simple Poisson equation (static schrödinger equation) on the torus

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

We first compare the Physics Informed Neural Network [14] and Deep Ritz Method [12, 58] with online random inputs. To enforce the periodic boundary conditions, we add a penalty term $L_p = \int_{(x,y) \in [0,1]^2} (u(x,y,0) - u(x,y,1))^2 + (u(0,x,y) - u(1,x,y))^2 + (u(x,0,y) - u(x,1,y))^2 dxdy$ to match the periodic condition of the function value and another term $L_{b,\text{grad}} = \int_{(x,y) \in [0,1]^2} \|\nabla u(x,y,0) - \nabla u(x,y,1)\|^2 + \|\nabla u(0,x,y) - \nabla u(1,x,y)\|^2 + \|\nabla u(x,0,y) - \nabla u(x,1,y)\|^2 dxdy$ to match the periodic condition of the function value. We tested PINN and Deep Ritz on both $u(x) = \sum_{i=1}^d \sin(2\pi x_i)$ and $u(x) = \sum_{i=1}^d \sin(4\pi x_i)$. We use the same
experiment setting as [78] 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

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

[11] [58] also considered using Sobolev norms as the loss function. [11] showed that the Sobolev norms exhibit an acceleration effect. However, in our setting, we cannot have random samples of $\nabla f$. To avoid information of $\nabla f$ appearing in the objective function, we perform an integration by parts that leads to the following objective function

$$\mathcal{L}_{\text{grad}} = \int \|\nabla u(x) + \nabla f(x)\|_{L^2}^2 dx$$

$$= \int \|\nabla u(x)\|_{L^2}^2 + \|\nabla f(x)\|_{L^2}^2 + 2\nabla \Delta u(x) \nabla u(x) - 2\nabla u(x) \nabla f(x) - 2\nabla \Delta u(x) \cdot \nabla f(x) dx$$

$$= \int \|\nabla u(x)\|_{L^2}^2 + \|\nabla f(x)\|_{L^2}^2 + 2\nabla \Delta u(x) \nabla u(x) + 2\Delta u(x) f(x) + 2\Delta \Delta u(x) \cdot f(x) dx.$$  

We conduct the Sobolev training with the objective function $\mathcal{L}_{\text{pin}} + \lambda \mathcal{L}_{\text{grad}} + \lambda_1 L_b + \lambda \mathcal{L}_{b, \text{grad}}$ and compare it with PINN and DRM. Following mostly the experiment setting in [78], we fix 3000 random samples as the dataset and run stochastic gradient descent with batchsize 50. The result presented in Figure 3 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.

![Figure 4: Solving equation (7) in 3 dimension with 3000 fixed samples using Deep Ritz Method [14], Physics-Informed Neural Network [12] and Sobolev Training.](image)

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

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 [88] [89] to accelerate the training. It is also interesting to generalize our results beyond GD, for example to mirror descent [90] and accelerated gradient descent [91]. In this paper, we did not consider operators with continuous spectrum and it will be interesting to extend our results using the techniques in [92]. 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 [93] [94] considered. It is also interesting to consider inverse problem arising from integral equation where $p > 0$. 


References


Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default [TODO] to [Yes], [No], or [N/A]. You are strongly encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section ??.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [N/A]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [TODO]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]
Appendix.

The appendix is constructed as follows:

- In Appendix [A], 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 [B.1] we consider the statistical optimality of the early stopped gradient descent algorithm. We bound the difference of the gradient descent.

- In Appendix [C] we provide our proof for the lower bound in Section [3.1] using the Fano method.

### A Preliminaries and Notations

This section starts with an overview of reproducing kernel Hilbert space, including Mercer’s decomposition, the integral operator techniques [46, 4, 8, 21] and the relationship between RKHS and the Sobolev space [95]. In order to fit the objective function we considered, we did a slight modification to the original integral operator technique [46, 4, 8].

#### A.1 Reproducing Kernel Hilbert Space

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

$$ g \otimes h : f \mapsto (f, h)_{\mathcal{H}} g. $$

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 [46, 8] consider the covariance operator on the Hilbert space $\mathcal{H}$ defined as $\Sigma = \mathbb{E}_{\rho_X} K_x \otimes K_x$. Then for all $f \in \mathcal{H}$, using the reproducing property, we know that

$$ (\Sigma f)(z) = (K_x, \Sigma f)_{\mathcal{H}} = \mathbb{E}[f(X)K_x(x,z)] = \mathbb{E}[f(X)K_x(x)]. $$

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

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

Next we consider the eigen-decomposition of the integral operator $\mathcal{L}$ via Mercer’s Theorem. There exists an orthonormal basis $\{\psi_i\}$ of $L_2(\mathcal{X})$ consisting of eigenfunctions of kernel integral operator $\mathcal{L}$. At the same time, the kernel function have the following representation $K(s,t) = \sum_{i=1}^{\infty} \lambda_i c_i(s)c_i(t)$ where $c_i$ are orthogonal basis of $L_2(\rho_X)$. Then $c_i$ is also the eigenvector of the covariance operator $\Sigma$ with eigenvalue $\lambda_i > 0$, i.e. $\Sigma c_i = \lambda_i c_i$.

### B Proof of the Upper Bound

In this section, we consider the convergence of the gradient descent algorithm to the target function $f$. In particular, we consider the gradient descent as a special case of a wider class of spectral filter algorithms [49, 79, 18, 19]. In our inverse problem setting, the spectral filter is defined as the estimator of the following form for $\lambda > 0$,

$$ \hat{q}_\lambda = g_\lambda(\hat{\Sigma}_{ld_A})A_2 \hat{S}_n^* g, $$

where $\hat{S}_n g = (g(x_1), \ldots, g(x_n))$ (leads to $\hat{S}_n^*$ maps from $\mathbb{R}^n$ to $\mathcal{H}$) via $\hat{S}_n^*(a_1, a_2, \ldots, a_n) = \frac{1}{n} \sum_{i=1}^{n} a_n K_{x_i} = \sum_{j=1}^{n} \alpha_i \cdot \alpha_j$, $\hat{\Sigma}_{ld_A} = \frac{1}{n} \sum_{i=1}^{n} \alpha_i \otimes \alpha_j$ and $Id$ is the identity operator. The function $q_\lambda : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is a function known as filter, which is an approximation of $x^{-1}$ controlled by
We first introduce several auxiliary lemmas which aims to bound different quantities according to our proof of the upper bound, we consider

$$\lambda > 0, \eta > 0, \alpha \in [0, 1]$$

for some positive $$c_\eta > 0$$. Next we show that the averaged gradient descent can be considered as spectral filter algorithm with filter $$q^\eta(x) = \left(1 - \frac{1 - (1 - \gamma_s)^\ell}{\gamma_s^\ell} \right) \frac{1}{\eta}$$. Let us consider the gradient descent

$$\eta_0 = 0, \eta_u = \eta_{u-1} + \gamma (A_2^T \hat S_n^* \hat y - \hat S_{ld,A_1} \hat y_{k-1})$$

and

$$\bar{\eta}_u = \frac{1}{t} \sum_{i=0}^{t} \eta_i = \frac{1}{t} \sum_{i=0}^{t} \left( I - (\gamma \hat S_{ld,A_1})^i \right) \left( \hat S_{ld,A_1} \right)^{-1} A_2^T \hat S_n^* \hat y.$$

Thus if we take the filter $$q_\ell(x) = \frac{1}{\eta} \left(1 - \frac{1 - (1 - \gamma_s)^\ell}{\gamma_s^\ell} \right)$$, we can have $$\bar{\eta} = \eta (\hat S_{ld,A_1}) A_2^T \hat S_n^* \hat y.$$ As the same time $$x^u r_1(x) = x^u (1 - x^u q_\ell(x)) = x^u \frac{1 - (1 - \gamma_s)^\ell}{\gamma_s^\ell} \leq \frac{\gamma s x^u}{\gamma_s} x^u = \frac{1}{\gamma_s^\ell}$$. 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.

### B.1 Convergence Of the Gradient Descent Algorithm

To conduct our proof of the upper bound, we consider $$g_\lambda = (\Sigma_{ld,A_1} + \lambda I)^{-1} A_2 S^* f_x$$ and decompose the error as $$(\|g_{\lambda} - u^*\| + \|\hat g_{\lambda} - g_{\lambda}\|)$$. We first bound the bias in the general Sobolev norm then come to bound the variance.

#### B.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 $$\mathcal{N}(\lambda) = \mathbb{E}_x \| (\Sigma_{A_1} + \lambda I)^{-1/2} A_2 K_x \|_H^2 = \text{Tr}((\Sigma_{A_1} + \lambda I)^{-1} \Sigma_{A_2,A_1})$$, $$\mathcal{N}_\infty(\lambda) = \sup_{x \in \mathcal{P}(x)} \| (\Sigma_{ld,A_1} + \lambda I)^{-1/2} A_2 K_x \|_H^2$$ and $$\mathcal{N}_\infty^2(\lambda) = \sup_{x \in \mathcal{P}(x)} \| (\Sigma_{ld,A_1} + \lambda I)^{-1/2} K_x \|_H^2$$ which are important important quantities used to bound the variance of our estimator.

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

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

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

$$\mathcal{N}(\lambda) = \text{Tr}((\Sigma_{A_1} + \lambda I)^{-1} \Sigma_{A_2,A_1}) = \sum_{i=1}^{\infty} \frac{\mu_i g_i^2}{\mu_i + \lambda} \leq \sum_{i=1}^{\infty} \frac{1}{1 - \alpha - \frac{2q}{\alpha + p} + \lambda} \leq \int_0^\infty \frac{\tau^{p - 2q}}{1 + \lambda(\tau^{\alpha + p})} d\tau \leq \left(\frac{\lambda^{\frac{2q}{\alpha + p}}}{1 + \lambda^{\alpha + p}}\right)^{\frac{p - 2q}{\alpha + p}}$$

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

$$\text{Trace}((\Sigma_{A_1} + \lambda I)^{-1} \Sigma_{ld,A_1}) \leq D\lambda^{\frac{p - 1}{\alpha + p}}$$

17
Proof. Similarly we use the spectral representation to bound the LHS as

\[
\text{Trace}\left((\Sigma_{\mathcal{A}_1} + \lambda)^{-1}\Sigma_{ld,ld}\right) = \sum_{i=1}^{\infty} \frac{\mu_i}{\mu_i \lambda + \lambda} 
\leq \sum_{i=1}^{\infty} \frac{\lambda^{-\alpha - p} + \lambda}{\lambda^{-\alpha - p}} \leq \int_{0}^{\infty} \frac{\tau^{p}}{1 + \lambda(\tau^{\alpha + p})} d\tau
\leq (\lambda)^{-\frac{1}{p + \alpha}} \int_{0}^{\infty} \frac{(\lambda)^{\frac{p}{p + \alpha}} \tau^{p}}{1 + \tau^{\alpha + p}} = \Omega\left(\lambda^{-\frac{p-1}{p + \alpha}}\right)
\] (11)

Lemma B.3. We denote the following quantity by \(N_{\infty}^{1}\), \(N_{\infty}^{2}\) and \(N_{\infty}^{3}\) can be bounded by

- \(N_{\infty}^{1}(\lambda) = \sup_{x \in p(x)} \| (\Sigma_{ld,\mathcal{A}_1} + \lambda)^{-1/2} K x \|_{H}^{2} \leq \| k_{c}^{\alpha} \|_{\infty}^{2} \lambda^{-\frac{\alpha + p}{\alpha + p}}, \)
- \(N_{\infty}^{2}(\lambda) = \sup_{x \in p(x)} \| (\Sigma_{ld,\mathcal{A}_1} + \lambda)^{-1/2} A_{2} K x \|_{H}^{2} \leq \| k_{c}^{\alpha} \|_{\infty}^{2} \lambda^{-\frac{\alpha + \alpha + 2q}{\alpha + p}}, \)
- \(N_{\infty}^{3}(\lambda) = \sup_{x \in p(x)} \| (\Sigma_{ld,\mathcal{A}_1} + \lambda)^{-1/2} A_{1} K x \|_{H}^{2} \leq \| k_{c}^{\alpha} \|_{\infty}^{2} \lambda^{-\frac{\alpha + \alpha}{\alpha + p}}. \)

Proof. We can also prove the bound from the spectral formulation and the \(l_{\infty}\) embedding property of the kernel function

\[
\| (\Sigma_{ld,\mathcal{A}_1} + \lambda)^{-1/2} K x \|_{H}^{2} = \sum_{i=1}^{\infty} \frac{\mu_i}{\mu_i \lambda + \lambda} e_i^2(x)
\leq \left( \sum_{i=1}^{\infty} \mu_i \frac{e_i^2(x)}{\mu_i \lambda + \lambda} \right) \sup_{i \geq 1} \frac{\mu_i^{-1 - \mu}}{\mu_i \lambda + \lambda} \leq \left( \sum_{i=1}^{\infty} \mu_i \frac{e_i^2(x)}{\mu_i \lambda + \lambda} \right) \sup_{i \geq 1} \frac{i^{-\left(1 - \mu\right) \alpha}}{i^{-\alpha - p} + \lambda}
\leq \lambda^{-\frac{\alpha + \alpha + 2q}{\alpha + p}} \| k_{c}^{\alpha} \|_{\infty}^{2}
\] (12)

and

\[
\| (\Sigma_{ld,\mathcal{A}_1} + \lambda)^{-1/2} A_{2} K x \|_{H}^{2} = \sum_{i=1}^{\infty} \frac{\mu_i q_i^2}{\mu_i \lambda + \lambda} e_i^2(x)
\leq \left( \sum_{i=1}^{\infty} \mu_i \frac{q_i^2 e_i^2(x)}{\mu_i \lambda + \lambda} \right) \sup_{i \geq 1} \frac{\mu_i^{-1 - \mu} q_i^2}{\mu_i \lambda + \lambda} \leq \left( \sum_{i=1}^{\infty} \mu_i \frac{q_i^2 e_i^2(x)}{\mu_i \lambda + \lambda} \right) \sup_{i \geq 1} \frac{i^{-\left(1 - \mu\right) \alpha - 2q}}{i^{-\alpha - p} + \lambda}
\leq \lambda^{-\frac{\alpha + \alpha + 2q}{\alpha + p}} \| k_{c}^{\alpha} \|_{\infty}^{2}
\] (13)

Similarly we have

\[
\| (\Sigma_{ld,\mathcal{A}_1} + \lambda)^{-1/2} A_{1} K x \|_{H}^{2} = \sum_{i=1}^{\infty} \frac{\mu_i p_i^2}{\mu_i \lambda + \lambda} e_i^2(x)
\leq \left( \sum_{i=1}^{\infty} \mu_i \frac{p_i^2 e_i^2(x)}{\mu_i \lambda + \lambda} \right) \sup_{i \geq 1} \frac{\mu_i^{-1 - \mu} p_i^2}{\mu_i \lambda + \lambda} \leq \left( \sum_{i=1}^{\infty} \mu_i \frac{p_i^2 e_i^2(x)}{\mu_i \lambda + \lambda} \right) \sup_{i \geq 1} \frac{i^{-\left(1 - \mu\right) \alpha - 2p}}{i^{-\alpha - p} + \lambda}
\leq \lambda^{-\frac{\alpha + \alpha + 2p}{\alpha + p}} \| k_{c}^{\alpha} \|_{\infty}^{2}
\] (14)

Lemma B.4. For all \(\lambda > 0\), we have

\[
\| \Sigma_{ld,\mathcal{A}_1}^{1/2} (\Sigma_{ld,\mathcal{A}_1} + \lambda)^{-1/2} \|_{\infty}^{2} \leq \lambda^{-\frac{\alpha + p}{\alpha + p}}
\]
Proof. We first bound \( \|\Sigma^{1/2} (\Sigma_{d,A_1} + \lambda)^{-1/2}\|^2 \)

\[
\|\Sigma^{1/2} (\Sigma_{d,A_1} + \lambda)^{-1/2}\|^2 = \sup_{i \geq 1} \frac{1}{\mu_i} \leq \sup_{i \geq 1} \frac{(1-\gamma)\alpha}{i^{-\alpha} + \lambda} \leq \lambda^{-\frac{\alpha+2p}{\alpha+p}}
\]

\( \square \)

Lemma B.5. With probability \( 1 - e^{-t} \), we have

\[
\| (\Sigma_{d,A_1} + \lambda)^{-1/2} (\Sigma_{d,A_1} - \Sigma_{d,A_1}^2 + \lambda)^{-1/2}\|^2 \lesssim \sqrt{n} \frac{\tau N_\infty^1(\lambda)N_\infty^3(\lambda)}{\tau^2}
\]

and as a consequence once \( n \geq \sqrt{\tau} \lambda^{-\frac{\alpha+2p}{\alpha+p}} \), we’ll have

\( \frac{1}{2} \leq \| (\Sigma_{d,A_1} + \lambda)^{-1/2} (\Sigma_{d,A_1} - \Sigma_{d,A_1}^2 + \lambda)^{-1/2}\| \leq 2, \frac{1}{2} \leq \| (\Sigma_{d,A_1} + \lambda)^{-1/2} (\Sigma_{d,A_1} - \Sigma_{d,A_1}^2 + \lambda)^{-1/2}\| \leq 2 \).

Proof. We utilize the concentration result for Hilbert space valued random variable \([96]\) to prove the bound here. Now, we consider the operator \( C_x : \mathcal{H} \to \mathcal{H} \) the operator defined by

\[
C_x f := A_1 f(x) k(x, \cdot) = \langle f, A_1 K_k \rangle K_x,
\]

and consider the random variable \( \xi_x := (\Sigma_{d,A_1} + \lambda)^{-1/2} C_x (\Sigma_{d,A_1} + \lambda)^{-1/2} \). From definition, we know that

\[
\xi_x = \langle f, A_1 (\Sigma_{d,A_1} + \lambda)^{-1/2} K_x \rangle (\Sigma_{d,A_1} + \lambda)^{-1/2} K_x
\]

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

\[
\| (\Sigma_{d,A_1} + \lambda)^{-1/2} (\Sigma_{d,A_1} - \Sigma_{d,A_1}^2 + \lambda)^{-1/2}\|^2 \lesssim \sqrt{\tau} \frac{N_\infty^1(\lambda)N_\infty^3(\lambda)}{n}.
\]

From Lemma B.3, we know that \( N_\infty^1(\lambda) = \sup_{\xi \in \{x \mid \| f \|_H \leq 1 \}} \| (\Sigma_{d,A_1} + \lambda)^{-1/2} K_x \|^2 \leq \| k_0 \|^2 \lambda^{-\frac{\alpha+2p}{\alpha+p}} \)

and \( N_\infty^3(\lambda) = \sup_{\xi \in \{x \mid \| f \|_H \leq 1 \}} \| (\Sigma_{d,A_1} + \lambda)^{-1/2} K_x \|^2 \leq \| k_0 \|^2 \lambda^{-\frac{\alpha+2p}{\alpha+p}} \). Thus once \( n \geq \sqrt{\tau} \lambda^{-\frac{\alpha+2p}{\alpha+p}} \), we’ll have

\( \frac{1}{2} \leq \| (\Sigma_{d,A_1} + \lambda)^{-1/2} (\Sigma_{d,A_1} - \Sigma_{d,A_1}^2 + \lambda)^{-1/2}\| \leq 2, \| (\Sigma_{d,A_1} + \lambda)^{-1/2} (\Sigma_{d,A_1} - \Sigma_{d,A_1}^2 + \lambda)^{-1/2}\| \leq 2 \).

\( \square \)

Theorem B.6 (Bernstein’s Inequality). Let \( (\Omega, \mathcal{B}, P) \) be a probability space, \( H \) be a separable Hilbert space, and \( \xi : \Omega \to H \) with

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

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

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

Lemma B.7 (Lemma 25 in [21]). For \( \lambda > 0 \) and \( 0 \leq \alpha \leq 1 \), the function \( f_{\lambda,\alpha} : [0, \infty) \to \mathbb{R} \) be defined by \( f_{\lambda,\alpha}(t) := \frac{\lambda}{\lambda+p} \). 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 \( t^* = \frac{\lambda}{1-\alpha} \).
Proof. For completeness, we provide the proof here. For function \( f_{\lambda,\alpha}(t) := \frac{\alpha t^\alpha}{\lambda t^\alpha + 1} \) with \( 0 < \alpha < 1 \), we know the derivative of it is \( f'_{\lambda,\alpha}(t) = \frac{\alpha \alpha^{-1}(\alpha + 1) - \alpha^{-1}}{(\lambda + 1) t^\alpha} \). The derivative \( f'_{\lambda,\alpha} \) has a unique root at \( t^* = \alpha \lambda / (1 - \alpha) \), \( f_{\lambda,\alpha} \) attains global maximum at \( t^* \) and
\[
\sup_{t \geq 0} f_{\lambda,\alpha}(t) = f_{\lambda,\alpha}(t^*) = \lambda^{-1} \alpha^{-1}(1 - \alpha)^{1-\alpha} \leq \lambda^{-1}.
\]
At the same time, \( (\alpha^\alpha (1 - \alpha)^{1-\alpha})' = \alpha^\alpha (1 - \alpha)^{1-\alpha} \log \left(t^{-\alpha} \right) \) thus \( \alpha^\alpha (1 - \alpha)^{1-\alpha} \) achieves minimum \( \frac{1}{2} \) when \( \alpha = \frac{1}{2} \). Thus we know \( \lambda^{-1}/2 \leq \sup_{t \geq 0} f_{\lambda,\alpha}(t) \). 

B.1.2 Bias

In this section, we consider the bias introduced by the regularization factor, i.e. the difference between \( g_\lambda = (\Sigma_{Id,A_1} + \lambda I)^{-1} A_2 S^* f_\rho \) and the ground truth solution \( A_1^{-1} A_2 f_\rho \).

Lemma B.8. If \( u^* = A_1^{-1} A_2 f_\rho \in [H]^{\beta} \) holds, then for all \( 0 \leq \gamma \leq \beta \) and \( \lambda > 0 \), the following bounds hold
\[
\|g_\lambda - A_1^{-1} A_2 S^* f_\rho\|_\gamma \leq \frac{\lambda^{-1}}{\lambda^{-1} + \gamma} \|u^*\|_{[H]^{\beta}}.
\]

Here \( g_\lambda = (\Sigma_{Id,A_1} + \lambda I)^{-1} A_2 S^* f_\rho \).

Proof. Since \( u^* = A_1^{-1} A_2 f_\rho \in [H]^{\beta} \), we can use the spectral representation \( u^* = \sum_{i=1}^\infty \alpha_i e_i \) with \( \|u^*\|_{[H]^{\beta}} = \sum_{i=1}^\infty \mu_i^{-\beta} \alpha_i^2 \). At the same time \( A_2 f_\rho = A_1 u^* = \sum_{i=1}^\infty \alpha_i \mu_i e_i \). We also observe that the matrix \((\Sigma_{Id,A_1} + \lambda I)^{-1}\) have the spectral representation \((\Sigma_{Id,A_1} + \lambda I)^{-1} = \sum_{i=1}^{\infty} (\mu_i + \lambda)^{-1} e_i \otimes e_i \) and leads to the spectral representation of the solution
\[
g_\lambda = (\Sigma_{Id,A_1} + \lambda I)^{-1} A_2 S^* f_\rho = \sum_{i=1}^{\infty} \frac{\mu_i \alpha_i^2}{\mu_i + \lambda} \frac{\alpha_i e_i}{\alpha_i e_i} = \sum_{i=1}^{\infty} \frac{\mu_i \alpha_i}{\mu_i + \lambda} \alpha_i e_i
\]

Then we can bound the bias via the spectral representation
\[
\|g_\lambda - A_1^{-1} A_2 f_\rho\|_\gamma^2 = \|g_\lambda - A_1^{-1} A_2 f_\rho\|_\gamma^2 = \left( \sum_{i=1}^{\infty} \frac{\mu_i \alpha_i}{\mu_i + \lambda} e_i \right)^2 = \left( \sum_{i=1}^{\infty} \frac{\lambda}{\mu_i + \lambda} e_i \right)^2 \leq \lambda \left( \frac{\lambda^{1-\alpha}}{\lambda^{-1} + \gamma} \right) \|u^*\|_{[H]^{\beta}} \]
\[
\lambda^2 \left( \sum_{i \geq 1} \frac{\mu_i^{-\beta} \alpha_i^2}{\mu_i + \lambda} \right) \leq \lambda \left( \frac{\lambda^{1-\alpha}}{\lambda^{-1} + \gamma} \right) \|u^*\|_{[H]^{\beta}}
\]

In this section, we also bound a bias over the energy function \( \|A_1 g_\lambda - A_2 f_\rho\|_2 \), which will be used in bounding the variance term.

Lemma B.9. If \( u^* = A_1^{-1} A_2 f_\rho \in [H]^{\beta} \) holds, then for all \( 0 \leq \gamma \leq \beta \) and \( \lambda > 0 \), the following bounds holds
\[
\|A_1 g_\lambda - A_2 f_\rho\|_2 \leq \frac{\lambda^{\frac{1-\alpha}{1-\gamma}}}{\lambda^\frac{1-\alpha}{1-\gamma}} \|u^*\|_{[H]^{\beta}}.
\]

Here \( g_\lambda = (\Sigma_{Id,A_1} + \lambda I)^{-1} A_2 S^* f_\rho \).

Proof. As discussed in the proof of Lemma B.8, we have the spectral representation of \( g_\lambda \) as
\[
g_\lambda = (\Sigma_{Id,A_1} + \lambda I)^{-1} A_2 S^* f_\rho = \sum_{i=1}^{\infty} \frac{\mu_i \alpha_i^2}{\mu_i + \lambda} \alpha_i e_i = \sum_{i=1}^{\infty} \frac{\mu_i \alpha_i}{\mu_i + \lambda} \alpha_i e_i
\]

Thus \( A_1 g_\lambda - A_2 f_\rho = \sum_{i=1}^{\infty} \left( \frac{\mu_i \alpha_i}{\mu_i + \lambda} - \mu_i \alpha_i \right) \alpha_i e_i = -\sum_{i=1}^{\infty} \left( \frac{\mu_i \lambda}{\mu_i + \lambda} \right) \alpha_i e_i \) and we can have the bound of the bias in the energy norm as
\[ \|A_1 g_\lambda - A_2 f_\lambda\|^2 = \left\| \sum_{i=1}^\infty \left( \frac{p_i \lambda}{\mu_i \lambda + \lambda} - p_i \right) a_i e_i \right\|^2_2 = \sum_{i=1}^\infty \left( \frac{p_i \lambda}{\mu_i \lambda + \lambda} - p_i \right)^2 = \lambda^2 \left( \sup_{i \geq 1} \frac{i^{-2p^2} - p}{\lambda + i^{-2p^2}} \right)^2 \sum_{i=1}^\infty p_i \gamma a_i^2 \leq \lambda \frac{a_{\gamma - 2p^2}}{2} \|u^*\|_{\|H\|^p}. \]

\[ \Box \]

**B.1.3 Variance**

In this section, we bound the variance which defined as the difference between \( g_\lambda = (\Sigma_{ld, A_1} + \lambda I)^{-1}A_2 S^* f_\rho \) and \( \hat{g}_\lambda = q_\lambda(\Sigma_{ld, A_1} )A_2 S^* y \) at the scale \( O \left( \left( \sigma^2 + \frac{R_\lambda^2 \lambda^2}{n} \right)^{\lambda(\lambda_0)} \sum_{\lambda} + \frac{(\gamma - \gamma_0) \rho \mu_0 - \rho}{n} + o(\frac{1}{n}) \right) \). We first did the following decomposition

\[ \Sigma_{ld, A_1} \lambda^{\frac{1}{2}} (g_\lambda - \hat{g}_\lambda) = \Sigma_{ld, A_1} \frac{q_\lambda(\Sigma_{ld, A_1} )A_2 S^* y - (\hat{g}_\lambda(\Sigma_{ld, A_1} ))X_{ld, A_1} - I \right) g_\lambda \]

\[ = \Sigma_{ld, A_1} \frac{q_\lambda(\Sigma_{ld, A_1} )A_2 S^* y - (\hat{g}_\lambda(\Sigma_{ld, A_1} ))X_{ld, A_1} - I \right) g_\lambda \]

\[ = \Sigma_{ld, A_1} \frac{q_\lambda(\Sigma_{ld, A_1} )A_2 S^* y - (\hat{g}_\lambda(\Sigma_{ld, A_1} ))X_{ld, A_1} - I \right) g_\lambda \]

\[ + \sum_{ld, A_1} q_\lambda(\Sigma_{ld, A_1} )A_2 S^* y - (\hat{g}_\lambda(\Sigma_{ld, A_1} ))X_{ld, A_1} - I \right) g_\lambda, \]

where we take the random variable \( \xi(x, y) \) as \( \xi(x, y) = (\Sigma_{ld, A_1} + \lambda)^{-1/2}(yA_2 K_x - A_1 g_\lambda(x) K_x) \) which satisfies \( \mathbb{E}Q \xi_2 = (\Sigma_{ld, A_1} + \lambda)^{-1/2}(A_2 fQ - \Sigma_{ld, A_1} g_\lambda) \) where \( fQ = \mathbb{E}Q f(x) K_x \) and \( \Sigma_{Q, A_1} = \mathbb{E}Q K_x \otimes A_1 K_x \) for arbitrary distribution \( Q \) and \( \mathbb{E}Q \xi(x, y) = (\Sigma_{ld, A_1} + \lambda)^{-1/2}(A_2 S^* f_\rho - \Sigma_{ld, A_1} g_\lambda) \). 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 \( \frac{\lambda^{\frac{1}{2}}}{n} \) when the problem is regular. Term (II) and (III) is smaller than the bias.

Our bound of term (III) bounds tighter than \[ \Box \] (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

\[ \left\| \Sigma_{ld, A_1} \frac{q_\lambda(\Sigma_{ld, A_1} )A_2 S^* y - (\hat{g}_\lambda(\Sigma_{ld, A_1} ))X_{ld, A_1} - I \right) g_\lambda \right\|^2_2 \]

\[ \leq \left\| \Sigma_{ld, A_1} \frac{q_\lambda(\Sigma_{ld, A_1} )A_2 S^* y - (\hat{g}_\lambda(\Sigma_{ld, A_1} ))X_{ld, A_1} - I \right) g_\lambda \right\|^2_2 \]

\[ \cdot \left\| \Sigma_{ld, A_1} \frac{q_\lambda(\Sigma_{ld, A_1} )A_2 S^* y - (\hat{g}_\lambda(\Sigma_{ld, A_1} ))X_{ld, A_1} - I \right) g_\lambda \right\|^2_2 \]

where \( \Sigma_{ld, A_1} = \Sigma_{ld, A_1} + \lambda I \) and \( \Sigma_{ld, A_1} = \Sigma_{ld, A_1} + \lambda I \). At the same time, we knows \( \| \Sigma_{ld, A_1} \lambda^{\frac{1}{2}} \left( \xi(x, y) - \mathbb{E}P \xi(x, y) \right) \|_H \]

\[ \leq \lambda \frac{a_{\gamma - 2p^2}}{2} \|u^*\|_{\|H\|^p}. \]

Thus we only need to focus on bounding the concentration error \( \frac{1}{2} \sum_{i=1}^n (\xi(x_i, y_i) - \mathbb{E}P \xi(x_i, y_i)) \). We recall the moment condition to control the noise of the observations. There are constants \( \sigma, L > 0 \) such that

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

21
is satisfied for $\mu$-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 $\xi(x,y) = (\Sigma_{I,d,A_1} + \lambda)^{-1/2}(y_2^2 K_x - A_1 g_1(x) K_x)$ via decomposing the random into three parts $(\Sigma_{I,d,A_1} + \lambda)^{-1/2}(y_2^2 K_x - f^*(x) A_2 K_x)$, $(\Sigma_{I,d,A_1} + \lambda)^{-1/2}(f^*(x) A_2 K_x - A_2 f^*(x) K_x)$ and $(\Sigma_{I,d,A_1} + \lambda)^{-1/2}(A_2 f^*(x) K_x - A_1 g_1(x) K_x)$. Base on the decomposition, we can bound the moments of random variable $\xi(x,y)$ as

\[
\mathbb{E}_P\|\xi(x,y)\|_H^m = \int \left(\|((\Sigma_{I,d,A_1} + \lambda)^{-1/2} A_2 K_x \|_H^m \int \|y - f^*(x)\|^m P(dy|x) + \int \left(\|((\Sigma_{I,d,A_1} + \lambda)^{-1/2} A_2 K_x \|_H^m \int \|A_2 f^*(x) - A_1 g_1\|^m P(dy|x) \right) du(x) \right) \leq \frac{1}{2} m! \sigma^2(L + \|f\|_x)^m \|h_x\|_H^{-m+2} \text{trace}((\Sigma_{I,d,A_1} + \lambda)^{-1}\Sigma_{A_2,A_2}) + \|h_x\|_H^{-m+2} \int |A_2 f^*(x) - A_1 g_1|^2 \mu(x) \leq m! \left(\|h_x\|^m \right)^{m-2} \left(\|h_x\|^2 \|A_2 f^*(x) - A_1 g_1\|^2 \right) \]
\[(\Sigma_{\text{Id},A_1}^{\frac{1}{2}})^2 r(\Sigma_{\text{Id},A_1})\| = \sup_{\sigma \in \sigma(\Sigma_{\text{Id},A_1}^{\frac{1}{2}})} (\sigma + \lambda)^{\frac{1}{2}} r_\lambda(\sigma) \leq \lambda^{1/2}. \tag{19}\]

At the same time, we can bound \[\| (\Sigma_{\text{Id},A_1}^{\lambda})^{-1} A_2 S^* f_\rho \|\] using the spectral representation
\[\| (\Sigma_{\text{Id},A_1}^{\lambda})^{-1} A_2 S^* f_\rho \|^2 = \sum_{i=1}^{\infty} \frac{\mu_i \hat{H}_i^2 |a_i^2|}{(\lambda + \mu_i)^2} \leq \left( \sup_{i \geq 1} \frac{\mu_i \hat{H}_i^2 \mu_i}{(\lambda + \mu_i)^2} \right) \sum_{i \geq 1} \mu_i a_i^2 \leq \left( \lambda \left( \frac{1}{\beta - \gamma} \right) \right)^2 \| u^* \|_{H_\beta}^2 \leq \lambda \frac{\beta}{\sigma} \| u^* \|_{H_\beta}^2, \tag{20}\]

Thus we know that
\[\| (\Sigma_{\text{Id},A_1}^{\lambda})^{-1/2} r(\Sigma_{\text{Id},A_1}) \| \leq \lambda^{-\frac{\sigma + p}{\sigma + \beta}} \lambda^{1/2} \| \Sigma_{\text{Id},A_1}^{\lambda} \|^{-\frac{1}{2}} \leq \lambda^{-\frac{\sigma + p}{\sigma + \beta}}, \]

where the last inequality is because \( p < 0 \) in our assumption.

**Bounded term (II).** In this paragraph, we demonstrate the proof to bound the term \[\Sigma_{\text{Id},A_1}^{\frac{1}{2}} q_{\lambda}(\hat{\Sigma}_{\text{Id},A_1}) (\Sigma_{\text{Id},A_1}^{\lambda})^{-1/2} \mathbb{E}_P \xi(x, y) = \Sigma_{\text{Id},A_1}^{\frac{1}{2}} q_{\lambda}(\hat{\Sigma}_{\text{Id},A_1}) (\Sigma_{\text{Id},A_1}^{\lambda})^{-1} A_2 S^* f_\rho - \Sigma_{\text{Id},A_1} g_\lambda. \] Note that \( \Sigma_{\text{Id},A_1} (\Sigma_{\text{Id},A_1}^{\lambda})^{-1} = I - \lambda (\Sigma_{\text{Id},A_1}^{\lambda})^{-1} \), thus we know that \[\Sigma_{\text{Id},A_1}^{\frac{1}{2}} q_{\lambda}(\hat{\Sigma}_{\text{Id},A_1}) (\Sigma_{\text{Id},A_1}^{\lambda})^{-1/2} (\Sigma_{\text{Id},A_1} + \lambda)^{-1/2} (A_2 S^* f_\rho - \Sigma_{\text{Id},A_1} g_\lambda) = \lambda \Sigma_{\text{Id},A_1}^{\frac{1}{2}} q_{\lambda}(\hat{\Sigma}_{\text{Id},A_1}) (\Sigma_{\text{Id},A_1}^{\lambda})^{-1} A_2 S^* f_\rho. \]

At the same time, according to our assumption on the spectral filter \( q_{\lambda} \), we know that
\[\| (\Sigma_{\text{Id},A_1}^{\lambda})^{-1/2} q_{\lambda}(\hat{\Sigma}_{\text{Id},A_1}) (\Sigma_{\text{Id},A_1}^{\lambda})^{-1/2} \| = \sup_{\sigma \in \sigma(\Sigma_{\text{Id},A_1}^{\lambda})} (\sigma + \lambda) q_{\lambda}(\sigma) \leq 2 q_{\gamma}. \]

Thus we can bound \( \Sigma_{\text{Id},A_1}^{\frac{1}{2}} q_{\lambda}(\hat{\Sigma}_{\text{Id},A_1}) \mathbb{E}_P \xi(x, y) \) via the following decomposition
\[\| \Sigma_{\text{Id},A_1}^{\frac{1}{2}} q_{\lambda}(\hat{\Sigma}_{\text{Id},A_1}) \mathbb{E}_P \xi(x, y) \| = \lambda \| \Sigma_{\text{Id},A_1}^{\frac{1}{2}} q_{\lambda}(\hat{\Sigma}_{\text{Id},A_1}) (\Sigma_{\text{Id},A_1}^{\lambda})^{-1} A_2 S^* f_\rho \| \]
\[\leq \| \Sigma_{\text{Id},A_1}^{\frac{1}{2}} (\Sigma_{\text{Id},A_1}^{\lambda})^{-1/2} \mathbb{E}_P \xi(x, y) \| \leq \| \Sigma_{\text{Id},A_1}^{\lambda} \|^{-\frac{1}{2}} \| \Sigma_{\text{Id},A_1}^{\lambda}^{-1} \| \| (\Sigma_{\text{Id},A_1}^{\lambda})^{-1} \| \| (\Sigma_{\text{Id},A_1}^{\lambda})^{-1} A_2 S^* f_\rho \| \]
\[\leq \lambda^{-\frac{\sigma + p}{\sigma + \beta}} \lambda^{-1/2} \lambda \frac{\beta - \gamma}{\sigma} - \frac{1}{2} \leq \lambda^{-\frac{\sigma + p}{\sigma + \beta}} \]

The last line is because of \[\| \Sigma_{\text{Id},A_1}^{\frac{1}{2}} (\Sigma_{\text{Id},A_1}^{\lambda})^{-1/2} \| \leq \lambda^{-\frac{\sigma + p}{\sigma + \beta}} \] (Lemma B.4), \[\| (\Sigma_{\text{Id},A_1}^{\lambda})^{-1/2} (\Sigma_{\text{Id},A_1}^{\lambda})^{-1/2} \| \leq 2 \] with high probability (Lemma B.5), \[\| (\Sigma_{\text{Id},A_1}^{\lambda})^{-1} \| \leq \lambda^{-1}, \]
\[\| (\Sigma_{\text{Id},A_1}^{\lambda})^{-1} A_2 S^* f_\rho \| \leq \lambda \frac{\beta - \gamma}{\sigma} - \frac{1}{2} \] (proved while bounding term (III)) and \( p < 0 \).

**B.2 Final Bound**

At this time we can have our final bound in Theorem B.2 via combining the bound for bias (Appendix B.1.2) and (Appendix B.1.3)
\[\| \hat{q}_{\lambda} - u^* \|_\gamma^2 \leq \| \hat{q}_{\lambda} - g_{\lambda} \|_\gamma^2 + \| \hat{q}_{\lambda} - u^* \|_\gamma^2 \]
\[\leq \lambda \gamma \frac{\alpha + p}{\alpha + p} + \frac{\gamma^2}{\alpha + p} \left( \sigma^2(\lambda) - \lambda^{-\frac{\sigma + p}{\sigma + \beta}} \right) + \lambda \gamma \frac{\alpha - \mu}{\alpha + p} + \lambda \gamma \frac{\alpha - \mu}{\alpha + p} + \frac{L_\lambda \| h_\beta \| + \| h_\beta \|}{n} \]
\[\leq \lambda \gamma \frac{\alpha + p}{\alpha + p} + \lambda \gamma \frac{2(p - \gamma) + 1}{\alpha + p} + \frac{\gamma \gamma \frac{\alpha + p}{\alpha + p}}{n} + \lambda \gamma \frac{\alpha - \mu}{\alpha + p} + \lambda \gamma \frac{\alpha - \mu}{\alpha + p} + \frac{L_\lambda \| h_\beta \| + \| h_\beta \|}{n^2} \tag{21}\]

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

C Proof of the Lower Bound

C.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 C.1 (Fano’s methods). Assume that $V$ is a uniform random variable over set $\mathcal{V}$, then for any Markov chain $V \rightarrow X \rightarrow \hat{V}$, we always have

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

Lemma C.2 (Varshamov-Gilbert Lemma). Let $D \geq 8$. There exists a subset $\mathcal{V} = \{\tau(0), \ldots, \tau(D/8)\}$ of $D$-dimensional hypercube $\mathcal{H}^D = \{0, 1\}^D$ such that $\tau(0) = (0, 0, \cdots, 0)$ and the $\ell_1$ distance between every two elements is larger than $\frac{D}{8}$

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

C.2 Proof of the Lower Bound

Theorem C.3. Let $(X, B)$ be a measurable space, $H$ be a separable RKHS on $X$ respect to a bounded and measurable kernel $k$ and operator $A = (A_{2}^{-1} \cdot A_{1})$ satisfies Assumption 2.7. We have $n$ random observations $\{(x_i, y_i) \in X \times Y\}_{i=1}^{n}$ of $f = \hat{A}u$, $u \in \mathcal{H} \cap L_\infty$, i.e. $y_i = f(x_i) + \eta_i$ where $\eta_i$ 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 $H : (X \times Y)^{\otimes n} \rightarrow \mathcal{H}$ satisfies

$$\inf_{H \in \mathcal{H}} \sup_{u \in \mathcal{H} \cap L_\infty} \mathbb{E} \|H(\{(x_i, y_i)\}_{i=1}^{n}) - u^*\|^2 \geq n^{-\min\{\alpha(\beta, \gamma, m)\}}$$

Proof. To prove 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

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

If we control $m \leq \epsilon^{-\frac{1}{2(\beta - \gamma)}}$, then we can always keep $u_\omega \in \mathcal{H}^\beta$ for $\|u_\omega\|_{\beta}^2 = \sum_{i=1}^{m} \omega_i^2 \mu_i^{\beta-1} \leq \epsilon \mu_2^{\beta-1} \leq m^{\beta - 1} \epsilon = O(1)$. Similarly, we can select $m \leq \epsilon^{-\frac{1}{2(\gamma - \beta)}}$ to control $\|u_\omega\|_{\gamma}^2 \leq \|u_\omega\|_{\beta}$ to $O(1)$. At the same time, the associated PDE right hand side function $f_\omega = A_{2}^{-1} \cdot A_{1} u_\omega = \left(\frac{\epsilon}{m}\right)^{1/2} \sum_{i=1}^{m} \frac{q \omega_i}{p} \mu_i^{1/2} e_i$.

Using Gilbert-Varshamov Lemma we know that there exists $M \geq 2m/8$ 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 $f_\omega$ and $f_{\omega'}$ can be lower bounded as $\|u_\omega - u_\omega'\|_{\gamma}^2 = \sum_{i=1}^{m} \left(\omega_i - \omega_i'\right)^2 \geq \epsilon/8$. 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 $\eta_i$ is sampled from $\mathcal{N}(0, \min\{\sigma, L\}^2)$ which satisfies the momentum
condition. Then we know that this mutual information can be bounded by the following average of KL divergence via

\[
I(V, X) = \frac{1}{M_e} \sum_{j=1}^{M_e} KL(P_j^n || P_0^n) = \frac{n}{2\sigma^2 M_e} \sum_{j=1}^{M_e} \| f_j - f_0 \|_2^2 \lesssim n\epsilon m^{-\alpha\gamma + 2(p-q)} \tag{22}
\]

Then we apply the Fano’s inequality

\[
\mathbb{P}(\hat{V} \neq V) \geq 1 - \frac{I(V; X) + \log 2}{\log |V|} = 1 - \frac{16C^*}{\log m_e} \frac{n\epsilon m^{-\alpha\gamma - 2(p-q)} + \log 2}{\log 2 \epsilon m_e}
\]

Take \( \epsilon = n^{-\frac{(\max(\beta, \mu) - \gamma)\alpha}{\max(3, p) + 2(p-q) + 1}} \), we know that with constant probability we have

\[
\| H(\{(x_i, y_i)\}_{i=1}^n) - u^* \|_2^2 \gtrsim n^{-\frac{(\max(\beta, \mu) - \gamma)\alpha}{\max(3, p) + 2(p-q) + 2}}
\]