MINIMAX OPTIMAL KERNEL OPERATOR LEARNING VIA MULTILEVEL TRAINING

Jikai Jin
School of Mathematical Sciences
Peking University
Beijing, China
jkjin@pku.edu.cn

Yiping Lu
Institute for Computational & Mathematical Engineering
Stanford University
Stanford, CA, US
yplu@stanford.edu

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

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

ABSTRACT

Learning mappings between infinite-dimensional function spaces have achieved empirical success in many disciplines of machine learning, including generative modeling, functional data analysis, causal inference, and multi-agent reinforcement learning. In this paper, we study the statistical limit of learning a Hilbert-Schmidt operator between two infinite-dimensional Sobolev reproducing kernel Hilbert spaces (RKHSs). We establish the information-theoretic lower bound in terms of the Sobolev Hilbert-Schmidt norm and show that a regularization that learns the spectral components below the bias contour and ignores the ones above the variance contour can achieve the optimal learning rate. At the same time, the spectral components between the bias and variance contours give us flexibility in designing computationally feasible machine learning algorithms. Based on this observation, we develop a multilevel kernel operator learning algorithm that is optimal when learning linear operators between infinite-dimensional function spaces.

1 INTRODUCTION

Supervised learning of operators between two infinite-dimensional spaces has attracted attention in several areas of application of machine learning, including, scientific computing (Lu et al., 2019; Li et al., 2020; de Hoop et al., 2021; Li et al., 2018; 2021b), functional data analysis (Crambes & Mas, 2013; Hörmann & Kidziński, 2015; Wang et al., 2020a), mean-field games (Guo et al., 2019; Wang et al., 2020b), conditional kernel mean embedding (Song et al., 2009; 2013; Muandet et al., 2017) and econometrics (Singh et al., 2019; Muandet et al., 2020; Dikkala et al., 2020; Singh et al., 2020). Despite the empirical success of operator learning, the statistical limit of learning an infinite-dimensional operator has not been investigated studied. In this paper, we study the problem of learning Hilbert Schmidt operators between infinite-dimensional Sobolev RKHSs $\mathcal{H}_K^{\beta}$ and $\mathcal{H}_L^{\gamma}$ with given kernels $k$ and $l$, respectively with $\beta, \gamma \in [0, 1)$ (Adams & Fournier, 2003; Christmann & Steinwart, 2008; Fischer & Steinwart, 2020). Our goal is to derive the optimal sample complexity for linear operator learning, i.e. how much data is required to achieve a certain performance level.

We first establish an information-theoretic lower bound for learning a Hilbert-Schmidt operator between Sobolev spaces with respect to a general Sobolev norm. Our information-theoretic lower bound indicates that the optimal learning rate is determined by the minimum of two polynomial rates; one is purely decided by the input Sobolev reproducing kernel Hilbert space and its evaluating norm, while the other one is purely determined by the output space along with its evaluating norm.

The rate is novel in that all existing results (Fischer & Steinwart, 2020; Li et al., 2022; de Hoop et al., 2021) only establish rates that depend on the parameter of input space. The reason is all previous
works (Talwai et al., 2022; Li et al., 2022; de Hoop et al., 2021) only consider the case of the output space as a subspace of a trace bounded reproducing kernel Hilbert space but not a general Sobolev space. We refer to Remark 2.1 for detailed comparisons.

To design a learning algorithm for approximating an infinite-dimensional operator, we need to learn a finite-dimensional restriction instead of the whole operator, as the latter would result in infinite variance. The finite-dimensional selection leads to bias error but decreases the variance. A natural task is then to study the shape of regularization that can lead to the optimal bias-variance trade-off and achieve the optimal learning rate. In this paper, we consider the bias and variance contour at the scale of optimal learning. Once the regularization enables one to learn all the spectral parts above the bias contour and below the variance contour, the learning is optimal. Finally, utilizing the region between the bias contour and variance contour, we developed a multilevel training algorithm (Lye et al., 2021; Li et al., 2021a) which first learns the mapping on low frequency and then successively fine-tunes the machine learning models to fit the high-frequency output. The intuition of our algorithm aligns with the original motivation of multilevel Monte Carlo (Giles, 2008; 2015): we use the next level to reduce bias while keeping the variance at the same scale. We demonstrate that such a multilevel algorithm can achieve an optimal non-parametric rate for linear operator learning.

1.1 Related Work

**Machine Learning Based PDE Solver** Solving partial differential equations (PDEs) plays a prominent role in many scientific and engineering disciplines, such as physics, chemistry, operation management, macro-economy, etc. The recent deep learning breakthrough has drawn attention to solving PDEs via machine learning methods (Raissi et al., 2019; Han et al., 2018; Sirignano & Spiliopoulos, 2018; Yu et al., 2018; Khoo et al., 2019; Chen et al., 2021). The statistical power and computational cost of these problem is well-studied by recent papers (Lu et al., 2021; 2022; Nickl et al., 2020; Nickl & Wang, 2020). This paper focuses on operator learning (Chen & Chen, 1995; Long et al., 2018; Long et al., 2019; Feliu-Faba et al., 2020; Khoo et al., 2021; Lu et al., 2019; Li et al., 2020; Kovachki et al., 2021; Stepaniants, 2021), i.e. learning a map between two infinite-dimensional function spaces. For example, one can learn a PDE solver that maps from the boundary condition to the solution or an inverse problem that maps from the boundary measurement to the coefficient field. Regarding the mathematical foundation of operator learning, (Liu et al., 2022) considers the learning rate of non-parametric operator learning. However, non-parametric functional data analysis often suffers from slower-than-polynomial convergence rates (Mas, 2012) due to the small ball probability problem for the probability distributions in infinite dimensional spaces (Delaigle & Hall, 2010). The most relevant works are (Lin et al., 2011; Reimherr, 2015; de Hoop et al., 2021), which consider the rates for learning a linear operator. For the comparison between our work and (de Hoop et al., 2021), see Remark 2.1.

**Learning with Kernel.** Supervised least square regression in RKHS and its generalization capability have been thoroughly studied (Caponnetto & De Vito, 2007; Smale & Zhou, 2007; De Vito et al., 2005; Rosasco et al., 2010; Mendelson & Neeman, 2010). The minimax optimality with respect to the Sobolev norm has been discussed recently in (Fischer & Steinwart, 2020; Liu & Li, 2020; Lu et al., 2022). Our paper is highly related to recent works (Schuster et al., 2020; Mollenhauer & Kolta, 2020; Talwai et al., 2022; Li et al., 2022; Park & Muandet, 2020; Singh et al., 2019; 2020) on identifying the Sobolev norm learning rate for the kernel mean embedding (Song et al., 2009; 2013; Muandet et al., 2017), which can also be formulated as learning an operator. The difference between our work and (Talwai et al., 2022; Li et al., 2022) sees Remark 2.1. A concurrent paper (Balasubramanian et al., 2022) considers a unified RKHS methodology for functional data analysis. Our paper provided a refined analysis and provided information theoretical optimal rates for this problem.

**Multilevel Monte Carlo** By combining biased estimators with multiple stepizes, multilevel Monte Carlo (MLMC) (Giles, 2008; 2015) dramatically improves the rate of convergence and achieves in many settings the canonical square root convergence rate associated with unbiased Monte Carlo (Rhee & Glynn, 2015; Blanchet & Glynn, 2015). Multilevel Monte Carlo can also be used for a random variable with infinite variance (Blanchet & Liu, 2016; Chen et al., 2020). To the best of our knowledge, this is the first paper that provides optimal sample complexity for multilevel Monte Carlo type algorithm for infinite variance problems in the non-parametric regime. Very recently,
(Lye et al., 2021; Li et al., 2021a) developed a multilevel machine learning Monte Carlo algorithm (ML2MC) / multilevel fine-tuning algorithm for learning solution maps by first learning the map on the coarsest grid and then successively fine-tuning the network on samples generated at finer grids. The authors also showed that, following the telescoping in MLMC, the multilevel training procedure could reduce the generalization error without spending more time on generating training samples. (Schäfer & Owhadi, 2021; Boullé et al., 2022) consider such multi-scale algorithm for learning Green’s function. However, the statistical power of such an algorithm is still under investigation. Another difference with (Boullé et al., 2022) is that we consider the Green function in $H^{-1}$ norm rather than the $\ell_1$ norm used in (Boullé et al., 2022). In this paper, we qualify a specific setting where this multilevel procedure can and is necessary to achieve the minimax optimal learning rate.

1.2 Contribution

- We derive a novel information-theoretic lower bound of learning a linear operator between two infinite-dimensional Sobolev reproducing kernel Hilbert spaces. The optimal learning rate is a minimum of two polynomial rates, one only dependent on the parameters of the input space while the other only on the parameters of the output space. The first rate aligns with the previous works (Li et al., 2022), while the second is novel to the literature.

- We study the shape of regularization that can lead to the optimal learning rate. One should learn all the spectral parts under the bias contour at the level of the optimal learning rate but not the spectral components above the variance contour at the level of the learning rate. This enables the estimator to enjoy an optimal balance of bias-variance.

- We qualify a specific setting where a multilevel training procedure (Lye et al., 2021; Li et al., 2021a) is necessary and capable of achieving a minimax optimal learning rate for learning a linear operator. We perform the optimal learning rate via $O((\ln \ln n)$ ensemble of ridge regression models. This differs from finite-dimensional operator learning, where a single-level estimator can be optimal.

2 Problem Formulation

2.1 Preliminary

Let $P_K$ be a distribution over the input space $H_K$ and define covariance operator $C_{KK} = \mathbb{E}_{u \sim P_K} u \otimes u$. Consider its spectral decomposition $C_{KK} = \sum_{i=1}^{\infty} \mu_i^2 e_i \otimes e_i$, where $\{\mu_i\}_{i=1}^{\infty}$ is an orthogonal eigenbasis and $\mu_i$ is the corresponding eigenvalues of $C_{KK}$ (here the $g \otimes h$ is an operator defined as $g \otimes h = gh^* : f \rightarrow \langle f, h \rangle g$). In the typical machine learning applications, the test distribution is the same as the training distribution, so we can assume that $H_K = \left\{ \sum_i a_i \mu_i^2 e_i : \{a_i\}_{i=1}^{\infty} \in \ell_2 \right\}$ without loss of generality. Note that this automatically holds in the context of learning the conditional mean embedding (CME) (Fischer & Steinwart, 2020; Talwai et al., 2022; Li et al., 2022).

Following Christmann & Steinwart (2008); Fischer & Steinwart (2020), we define the interpolation Sobolev space $H_K^\beta = \left\{ f = \sum_i a_i (\mu_i^{\beta/2} e_i) : \{a_i\}_{i=1}^{\infty} \in l^2 \right\}$ for any $\beta > 0$, equipped with Sobolev norm defined by the inner product $\left( \sum_i a_i (\mu_i^{\beta/2} e_i), \sum_i b_i (\mu_i^{\beta/2} e_i) \right)_{H_K^\beta} = \sum_i a_i b_i$. For the output space, we fix a user-specified distribution $Q_L$ and a reproducing Kernel Hilbert Space. We can similarly define the covariance operator $C_{QL}$ and the Sobolev space $H_L^{\gamma}$. Natural choices of $Q_L$ include some distribution on kernel functions $\{\ell(y, \cdot) : y \in Y\}$ of $H_L$ induced by some distribution $Q_L$ on $Y$, so that $C_{QL}$ is a kernel integral operator with respect to $Q_L$ and $H_L^{\gamma}$ is an interpolation space between $H_L$ and $L^2(Q_L)$; see Example 2.1 for a specific example.

Following Li et al. (2022), in this paper we consider the Hilbert-Schmidt norm between two Sobolev Spaces for all the operators, which is defined as following.

**Definition 2.1** $((\beta, \gamma))-\text{norm}$ Let $T : H_K \rightarrow H_L$ be a possibly unbounded linear operator. $I_{1,\beta,P_K} : H_K \rightarrow H_K^\beta, \beta \in (0, 1)$ is the canonical embedding mapping that takes $u \in H_K$ to the same element $u$ in the larger space $H_K^\beta$, and $I_{1,\gamma,Q_L} : H_L \rightarrow H_L^{\gamma}, \gamma < 1$ is similarly defined.
Then the \((\beta, \gamma)\)-norm of \(T\) is defined as
\[
\|T\|_{\beta, \gamma} = \left\| \left( I_{1_\gamma \gamma} \right|^*_{Q_L} \right\|_{\text{HS}(\mathcal{H}_K, \mathcal{H}_L)} \circ T \circ I_{1_\beta \beta} P_K \right\|_{\text{HS}(\mathcal{H}_K, \mathcal{H}_L)} = \left\| C_{Q_L}^{-((1-\gamma)/2)} \circ T \circ C_{K}^{((1-\beta)/2)} \right\|_{\text{HS}(\mathcal{H}_K, \mathcal{H}_L)},
\]
where we omit the dependence of \(\|\cdot\|_{\beta, \gamma}\) on \(P_K\) and \(Q_L\) since it will always be clear from context.

### 2.2 Problem Formulation

We consider the problem of learning an unknown linear operator \(A_0 : \mathcal{H}_K \mapsto \mathcal{H}_L\) between two reproducing kernel Hilbert spaces corresponding to kernels \(k\) and \(l\) respectively. We are given \(N\) noisy data pairs \((u_i, v_i)\), \(1 \leq i \leq N\) related by
\[
v_i = A_0 u_i + \varepsilon_i \tag{1}
\]
where \(u_i \overset{i.i.d.}{\sim} P_K\) for some unknown distribution \(P_K\) and \(\varepsilon_i\) is the noise drawn from some distribution with zero mean that may depend on \(u_i\). We use \(P_{KL}\) for the joint distribution of \((u_i, v_i)\). Denote \(C_{KK} = \mathbb{E}_{u \sim P_K} u \otimes u\), \(C_{KL} = \mathbb{E}_{(u, v) \sim P_{KL}} u \otimes v\) and its adjoint \(C_{LK} = \mathcal{C}_{KL}^*\) be uncentered cross-covariance operators associated with \(P_{KL}\). Then we can reformulate the ground truth operator as \(A_0 = C_{KK} C_{KL}^\dagger\), where \(\dagger\) is the pseudo-inverse (Talwai et al., 2022; Li et al., 2022).

With the goal of understanding the relative difficulty of learning different types of linear operators, we investigate the sample efficiency of learning \(A_0\) under certain source assumptions imposed on the data model (1). Source condition (Caponnetto & De Vito, 2007; Mendelson & Neeman, 2010; Steinwart et al., 2009; Rosasco et al., 2010; Fischer & Steinwart, 2020) assumes that the learning target lies in a parameterized function class and study the learning rate for different problems with different hardness. Specifically, the source condition assume that the learning target is bounded in certain Sobolev norm. In this paper, we consider learning an operator with bounded \((\beta, \gamma)\)-norm, which is the Hilbert-Schmidt norm that maps from \(\mathcal{H}_K^\beta\) to \(\mathcal{H}_L^\gamma\). We consider the generalization error/convergence rate under another \((\beta', \gamma')\)-norm as in Fischer & Steinwart (2020); Lu et al. (2022); Talwai et al. (2022); Li et al. (2022).

**Remark 2.1** Although recent works have considered similar problems in the context of conditional mean embedding (Talwai et al., 2022; Li et al., 2022) and functional data analysis (de Hoop et al., 2021), in all these papers, the output space is a trace bounded RKHS (de Hoop et al., 2021). Assumption 2.14 (vi) rather than the general parameterized Sobolev space in our paper.

We then list all the assumptions imposed on the underlying kernel for our theoretical results. We follow the standard capacity assumptions and embedding properties used in kernel regression (Fischer & Steinwart, 2020; Talwai et al., 2022; Li et al., 2022).

**Assumption 2.1 (Capacity Condition of the Covariance)** The eigenvalues \(\{\mu_i\}_{i \geq 1}\) of the covariance operator \(C_{KK} = \mathbb{E}_{u \sim P_K} u \otimes u\) satisfies \(\mu_i \propto i^{-\frac{1}{2}}\) for some \(p \in (0, 1)\). Similarly, the eigenvalues \(\{\rho_i\}_{i \geq 1}\) of the covariance operator \(C_{KL} = \mathbb{E}_{(u, v) \sim P_{KL}} u \otimes v\) satisfies \(\rho_i \propto i^{-\frac{1}{2}}\) for some \(q \in (0, 1)\).

**Assumption 2.2 \((\ell_\infty, \infty)\) Embedding Property of the Input RKHS** There exists a smallest \(\alpha \in (0, 1)\) such that \(\left\| (I_{1_\alpha \alpha} P_K)^\dagger f \right\|_{\mathcal{H}_K^\alpha} \leq A_1\) a.s. under \(P_K\) for some \(A_1 < +\infty\).

**Assumption 2.3 \((\ell_\infty, \infty)\) Embedding Property of the Output RKHS** There exists \(A_2 < +\infty\) such that \(\|g\|_{\mathcal{H}_L} \leq A_2\) holds for all \(g\) in the range of \(A_0\), except from a \(Q_L\)-null set.

**Assumption 2.4 (Moment Condition)** There exists an operator \(V : \mathcal{H}_L \mapsto \mathcal{H}_L\) with \(\text{tr}(V) \leq \sigma^2\) such that for every \(u \in \mathcal{H}_K\), we have \(\mathbb{E}_{v \sim P_{KL}(u)} \left[ (v - A_0 u)^2 \otimes (v - A_0 u) \right] \leq \frac{1}{2} (2k)! R^{2k-2} V\), holds for all \(k \geq 2\).

**Assumption 2.5 (Source Condition)** \(A_0\) is bounded under \((\beta, \gamma)\)-norm i.e. \(\|A_0\|_{\beta, \gamma} \leq B\).
2.3 Examples

In this section, we will introduce two examples of our theory. The first one is about learning a differential operator, for example inferring an advection-diffusion model (Portone & Moser, 2022) from observations or predicting the future (Long et al., 2018; Lu et al., 2019; Li et al., 2020; Feliu-Faba et al., 2020; Huang et al., 2021). The second example is about learning conditional mean embedding (Song et al., 2009; 2013; Muandet et al., 2017), which represents a conditional distribution as an RKHS element. Thus conditional distribution regression can be reduced to kernel operator learning. Our theory can also be used for linear inverse problems such as radial electrical impedance tomography (EIT) (Mueller & Siltanen, 2012), and the severely ill-posed inverse boundary problem for the Helmholtz equation with unknown wave-number parameter (Agapiou et al., 2014). For detailed discussion, we refer to (de Hoop et al., 2021; Section 1.3).

Example 2.1 (Learning differential operators) Suppose that the ground-truth operator \( A_0 = \Delta \) where \( \Delta \) is the Laplacian and \( t \in \mathbb{Z} \). Let \( \mathcal{H}_K = \mathcal{H}^{m+2l}([0,1]) \) be the Sobolev space with smoothness \( m + 2l \) on \([0,1]\) and \( \mathcal{H}_L = \mathcal{H}^m([0,1]) \), then \( A_0 \) is a bounded operator from \( \mathcal{H}_K \) to \( \mathcal{H}_L \) which corresponds to the \( \beta = \gamma = 1 \) case. However, we will see below that we can obtain a better characterization of the learning error using our theory.

Consider for example that the input has mean zero and the Matérn-type covariance operator \( C_{KK} = \sigma^2 (\Delta + \tau^2 I)^{-s} \). Its eigenvalues satisfy \( \mu_n \propto n^{-2s} \). On the other hand, we choose \( Q_Y \) to be a distribution supported on \( \{ \ell(y, \cdot) \mid y \in [0,1] \} \) induced by a uniform distribution on \([0,1]\), where \( \ell \) is the kernel function of \( H_L \). Then \( C_{Q_Y} \) is essentially the kernel integral operator on \( H_L \) w.r.t. the uniform distribution, and its eigenvalues are \( \rho_n \propto n^{-2m} \). The assumption \( \|A_0\|_{\beta, \gamma} < +\infty \) is satisfied if and only if \( (1-\gamma)m < (1-\beta)s - \frac{1}{2} \) (i.e. \( \gamma > 1 - \frac{2(1-\beta)s-1}{2m} \))

Example 2.2 (Conditional mean embedding) Suppose that we would like to learn the conditional distribution \( Y \mid X \) from a data set \( \{(x_i, y_i) \mid 1 \leq i \leq N\} \subset X \times Y \) where \( x_i \overset{\text{i.i.d.}}{\sim} P_K \). Let \( \mathcal{H}_K \) and \( \mathcal{H}_L \) be two RKHSs on \( X \) and \( Y \) respectively, with measurable kernel \( k(\cdot, \cdot) \) and \( \ell(\cdot, \cdot) \). Then we can define a conditional mean embedding (CME) operator \( C_{Y \mid X} \) that satisfies
\[
C_{Y \mid X} k(x, \cdot) = \mathbb{E}_{Y \mid x} \ell(Y, \cdot) =: \mu_{Y \mid x}, \quad \text{and} \quad \mathbb{E}_{Y \mid x} g(Y) = \langle g, \mu_{Y \mid x} \rangle \quad \forall x \in X.
\]
We choose \( A_0 = C_{Y \mid x} \). In this case, \( C_{KK} = \mathbb{E}_{P_K} k(X, \cdot) \otimes k(X, \cdot) \). Assumption 2.2 states that \( \sup_{x \in X} k^0(x, x) = \Lambda_1 \), while Assumption 2.3 is equivalent to \( \sup_{x \in X} \|\mu_{Y \mid x}\| \leq \Lambda_2 \) (for simplicity we only focus on the case \( \zeta = 1 \)). According to Assumption 2.5, we assume that \( \|C_{Y \mid X}\|_{\beta, \gamma} \leq B \).

The mis-specified setting where \( \beta < 1 \) has been studied in previous work (Fischer & Steinwart, 2020; Taiwai et al., 2022; Li et al., 2022). However, they only consider the case \( \gamma = 1 \). Our results also cover the case \( \gamma < 1 \), which allows us to obtain theoretical guarantees for computing conditional expectation of the larger function class \( \mathcal{H}_L \).

3 Information Theoretic Lower Bound

In this section, we provide an information-theoretic lower bound via the Fano method for the convergence rate of the operator learning problem formulated in Section 2.

Theorem 3.1 Suppose that \( P_K \) and \( Q_L \) are probability distributions on Hilbert spaces \( \mathcal{H}_K \) and \( \mathcal{H}_L \) respectively such that Assumptions 2.1 and 2.2 hold. Then for any estimator \( \mathcal{L} : (\mathcal{H}_K \times \mathcal{H}_L)^{\otimes N} \mapsto \mathcal{H}_K \otimes \mathcal{H}_L \), there exists a linear operator \( A_0 \) and a joint data distribution \( P_{KL} \) with marginal distribution \( P_K \) on \( \mathcal{H}_K \) satisfying Assumptions 2.3 to 2.5, such that with probability \( \geq 0.99 \) over \((u_i, v_i) \overset{\text{i.i.d.}}{\sim} P_{KL} \), we have \( \|\mathcal{L} ((u_i, v_i))_{i=1}^N \|_{\beta, \gamma} \geq N^{-\min \left\{ \frac{\max \{\alpha, \beta\} - \beta'}{\max \{\alpha, \beta\} + \beta'} , \frac{\zeta}{\zeta + 1} \right\} } \).

Remark 3.1 Our lower bound is composed of a minimum of two parts. The first rate \( N^{-\max \{\alpha, \beta\}} \) is the minimax optimal Sobolev learning rate for kernel regression (Fischer & Steinwart, 2020; Taiwai et al., 2022; Li et al., 2022; Lu et al., 2022) and is fully determined by the parameter of the input Sobolev reproducing kernel Hilbert space. Our second rate \( N^{-\frac{\zeta}{\zeta + 1}} \) is novel to the literature.
This bound shows how the infinite-dimensional problem is different from the finite-dimensional regression problem and is fully determined by the output Sobolev reproducing kernel Hilbert space parameter. Our lower bound shows that the hardness of learning a linear operator is determined by the harder part between the input and output spaces. We will explain why the lower bound has such a structure in Remark 4.2 and Figure 2.

4 ON THE SHAPE OF REGULARIZATION

In this section, we aim to understand the shape of regularization so that the constructed estimator \( \hat{A} \) based on \( N \) i.i.d. data \( \{(u_i, v_i)\}_{i=1}^N \sim P_{KL}^\mathcal{H} \) for \( 1 \leq i \leq N \) enjoys an optimal learning rate. Compared with existing approaches where a regularized least-squares estimator can achieve statistical optimality (Fischer & Steinwart, 2020; Talwai et al., 2022; Li et al., 2022; de Hoop et al., 2021) under \( (\beta, 1) \)-norm, we study the learning rate under the \( (\beta', \gamma') \)-norm \( (\beta' \in (0, \beta), \gamma' \in (\gamma, 1)) \) which is defined in Definition 2.1 as \( \| \hat{A} - A_0 \|_{\beta', \gamma'} = \left\| C_{Q_L}^{-\frac{1}{\beta'}} \left( \hat{A} - A_0 \right) C_{K_K}^{-\frac{1}{\gamma'}} \right\|_{\text{HS}(\mathcal{H}_K, \mathcal{H}_L)} \). The norm of the additional \( C_{Q_L}^{-\frac{1}{\beta'}} \) term is unbounded which make our setting harder than the convergence in \( (\beta, 1) \)-norm in existing works. Since \( C_{Q_L}^{-\frac{1}{\beta'}} \) is bounded when restricted to the finite-dimensional space span \( \{\rho_i^2 f_i : 1 \leq i \leq n\} \), we should also include another bias-variance trade-off via regularizing in the output shape. As a result, we are interested in answering the following question

"What is the optimal way to combine the regularization in the input space and regularization in the output space? i.e. What is the optimal shape of regularization?"

To answer this question, we investigate the problem in the spectral space, i.e. considering the spectral representation of operator \( A_0 = \sum_{i,j=1}^{+\infty} a_{ij} \mu_i^{\beta} e_i \otimes \rho_j^{1-\beta} f_j \). The problem of estimating \( A \) then reduces to learning the coefficients “matrix” \( (a_{ij})_{i,j=1}^{+\infty} \). The source condition Assumption 2.5 enforces \( \sum_{i,j=1}^{+\infty} a_{ij}^2 \leq B \). We show in Appendix B.1.1 that regularizing the basis \( e_i \otimes f_j \) will introduce a bias of order \( \left\| a_{ij} \mu_i^{\beta} e_i \otimes \rho_j^{1-\beta} f_j \right\|^2 = a_{ij}^2 \mu_i^{\beta-\beta'} \rho_j^{\gamma'-\gamma} \propto i^{\frac{\beta-\beta'}{2}} j^{\frac{1}{2} - \frac{\gamma'-\gamma}{2}} \) under the \( (\beta', \gamma') \)-norm. On the other hand, when \( \alpha \leq \beta + p \), we show in Appendix B.1.2 that the variance of learning \( (i, j) \) from noisy data scales as \( \frac{1}{N} \mu_i^{\beta-\beta'} \rho_j^{\gamma'-\gamma} \propto i^{\frac{\beta-\beta'}{2}} j^{\frac{1}{2} - \frac{\gamma'-\gamma}{2}} \). Since the variance would accumulate for a fixed \( j \), learning \( (i, j) \) for \( i \leq i_{\text{max}} \) results in a variance of \( \frac{1}{N} i_{\text{max}}^{\beta-\beta'} j^{\frac{1}{2} - \frac{\gamma'-\gamma}{2}} \). (Similar analysis can be carried out for the \( \alpha > \beta + p \) case as well, but the variance now scales as \( \frac{1}{N} i_{\text{max}}^{\beta-\beta'} j^{\frac{1}{2} - \frac{\gamma'-\gamma}{2}} \); see Appendix B for detailed derivations.) In summary, we need to make bias-variance trade off in the \((i, j)\)-plane, i.e. decide whether we should learn or regularize over the basis \( e_i \otimes f_j \).

4.1 REGULARIZATION VIA VARIANCE CONTOUR

The underlying idea of regularization is that some components are intrinsically hard to learn due to large variance; these components are then neglected by adding regularization and are counted as bias. Thus, the remaining components are easy to learn due to controllable variance. This intuition works well when the estimation error results from the noise of the data and is well-studied in a line of works (Fischer & Steinwart, 2020; de Hoop et al., 2021; Talwai et al., 2022; Li et al., 2022). This idea still works in our setting, but we need to re-evaluate the bias and variance of each component. Since we work with the Hilbert-Schmidt norm, this can be done in a coordinate-wise manner, meaning that we can look at each \( a_{ij} \) separately and decide whether to neglect it (contribute to bias) or to learn it from data (contribute to variance).

Since the variance term measures the hardness of learning, we naturally introduce the notion of variance contour, which is a curve on the \( \mathbb{R}^2_+ \) plane on which all points induce the same order of variance (here we work with real coordinates for convenience, although we only care about integer
even without the additional problem-dependent parameters $\gamma, \gamma'$, where $\hat{\lambda}$ produce a certain level of bias i.e. the specific level of bias contour.

Specifically, we consider deciding the regularization strength according to the spectral elements in-duce a certain level of bias i.e. the specific level of bias contour. In this section, we introduce another scheme to design the optimal estimator via learning all the spectral component under a certain bias contour. We have showed that if we learn all the spectral components under certain variance contour and reg-

Theorem 4.1

Consider the estimator $\hat{A}$ defined by (3) and (4). A reasonable regularization scheme is then to learn all coordinates $(i, j) \in \mathbb{Z}_+^2$ below the curve $\ell_{C, \text{var}}$ and ‘regularize out’ the remaining coordinates that are difficult to learn due to large variances. This can gives us the estimator with smallest estimator at give variance level. This observation motivates us to construct our estimator as

$$\hat{A} = \sum_{j=1}^{N^N} \left( \rho_j^{1/2} \otimes \rho_j^{1/2} \right) \hat{C}_{LK} \left( \hat{C}_{KK} + \lambda_j I \right)^{-1},$$

where $\hat{C}_{LK} = \frac{1}{N} \sum_{i=1}^N v_i \otimes u_i$, $\lambda_j (1 \leq j \leq y_N = C^{1/\gamma'})$ are the regularization coefficients imposed on different dimensions of the output space. According to (2) and noting that $\mu_i \propto i^{-\frac{1}{2}}$, we define

$$\lambda_j = \max \left\{ \left( j^{\frac{1}{2} - \frac{1}{\gamma'}} N \max \left\{ 1 - \frac{\beta - \beta'}{\max(x, \beta + p)}, \frac{1 - \gamma'}{1 - \gamma} \right\} \right)^{-\frac{1}{\beta + p}}, c_0 \left( \frac{N}{\log N} \right)^{-\frac{1}{\alpha}} \right\},$$

with $C = N^{\max \left\{ 1 - \frac{\beta - \beta'}{\max(x, \beta + p)}, \frac{1 - \gamma'}{1 - \gamma} \right\}}$ in (2). The additional $N^{-\frac{1}{\alpha}}$ term in (4) is needed for controlling the error of approximating $C_{KK}$ via $\hat{C}_{KK}$ (cf. Theorem D.3) which is standard in the Sobolev learning literature Fischer & Steinwart (2020); Talwai et al. (2022); Lu et al. (2022). The following theorem describes the convergence rate of our estimator defined by (3) and (4).

Theorem 4.1

Consider the estimator $\hat{A}$ defined by (3) and (4). Suppose that Assumptions 2.1 to 2.5 hold, then there exists a universal constant $C$ such that with probability $\geq 1 - e^{-\tau}$, we have

$$\| \hat{A} - A_0 \|_{\beta', \gamma'}^2 \leq C \tau^2 \left( \frac{N}{\log N} \right)^{-\frac{1}{\max (\alpha, \beta + p) \cdot \frac{\gamma' - \gamma}{1 - \gamma}}} \log^2 N.$$

Remark 4.1

Compared with Theorem 3.1, our upper bound is optimal up to logarithmic factors when $\alpha \leq \beta$. The optimal learning rate in the $\alpha > \beta$ regime is an outstanding problem for decades, even without the additional problem-dependent parameters $\gamma, \gamma'$ (see e.g. the discussions following Fischer & Steinwart, 2020, Theorem 2)). In this paper, we do not address this problem either.

4.2 Regularization via Bias Contour

We have showed that if we learn all the spectral components under certain variance contour and regular-
The sequences are chosen in a staircase manner as plotted in Figure 3 (for formal definitions see [108x304]Consider the estimator

\[ \hat{A} \]

Theorem 4.2

regularization on the output space leading to a convergence rate of

\[ O \left( \frac{1}{N^{\gamma}} \right) \]

the two yields \( O \left( N^{-\frac{\beta - \beta'}{\beta + p + \gamma}} \right) \) convergence rate. It is the same learning rate as the two kernel regression curves meet when \( y = 1 \). Right: When \( \frac{\beta + p}{\beta + p + \gamma} \geq \frac{1}{4} \), the two contours yield the same regularization on the output space leading to a convergence rate of \( O \left( N^{-\frac{\beta - \beta'}{4}} \right) \).

\( (x^*, y^*) \) on the variance contour with maximal contribution to bias. Naturally, we can also construct our estimator using a bias contour that passes through \( (x^*, y^*) \). In this case, we may define

\[ \lambda_j = \max \left\{ \frac{1}{N^{\min \{ \frac{\beta - \beta'}{\beta + p + \gamma}, \frac{\gamma}{4} \}}}, e^0 \left( \frac{N_{\log N}}{1} \right)^{-\frac{1}{2}} \right\} \]

for similar reasons as Section 4.1, which also yields optimal rate as stated in Theorem 4.2 below.

Remark 4.2 (On the optimal shape of regularization) The discussion in Sections 4.1 and 4.2 reveals another understanding of our information theoretic lower bound. Firstly, we should learn all the spectral components under the bias contour otherwise the bias will exceed the lower bound. Secondly, we should not learn any spectral component over the variance contour since otherwise the variance will exceed the lower bound. Thus the bias contour should always be under the variance contour, otherwise no estimator can be designed. The bias and variance contours at the level of optimal learning rate are plotted in Figure 2. They only meet at \( (x^*, y^*) \) with \( x^* = 1 \) or \( y^* = 1 \), which has the largest contribution to the bias (resp. variance) among all points on the variance (resp. bias) contour, thus dominating the estimation error. When the two curves meet at \( y^* = 1 \), it reduces to the original kernel regression case. When the two curves meet at \( x^* = 1 \), it leads to our new rate that depends on the output space.

Theorem 4.2 Consider the estimator \( \hat{A} \) defined by (3) with \( \lambda_j \) defined above. Suppose that Assumptions 2.1 to 2.5 hold, then there exists a universal constant \( C \), such that

\[
\| \hat{A} - A_0 \|_{\beta, \gamma} \leq C\tau^2 \left( \frac{N_{\log N}}{1} \right)^{-\min \{ \frac{\beta - \beta'}{\beta + p + \gamma}, \frac{\gamma}{4} \}} \log^2 N \mbox{ holds with probability } \geq 1 - e^{-\tau}.
\]

5 MultiLevel Kernel Operator Learning

In this section, we study a multilevel machine learning algorithm [Lye et al., 2021]; [Li et al., 2021a]; [Boullé et al., 2022] but at each level we consider a cost-accuracy trade-off [De Hoop et al., 2022] to control the variance at a proper scale. We show that the multilevel level algorithm can cover all the spectral component below the bias contour and achieve the optimal learning rate. Our idea is similar to the multilevel Monte Carlo [Giles, 2008; 2015], which reduces bias from multilevel algorithm. Our multilevel estimator differs from the DeepONet [Lu et al., 2019] and the PCA-Net [Bhattacharya et al., 2020] since we add different regularizations for each level. Our theory indicates that the multilevel approach outperforms previous ones and achieves the optimal learning rate.

The basic idea is to design a minimum number of machine learning estimators that cover all the spectral elements under the bias contour but do not exceed the variance contour at the same time. To achieve this, we choose sequences \( \{x_i\} \) and \( \{y_i\} \) for \( 1 \leq i \leq L_N \) where \( y_i \) denotes the \( i \)-th level and \( x_i \) controls the corresponding regularization via the regularization coefficient \( \lambda_i^{(K)} = x_i^{-\frac{1}{2}} \). The sequences are chosen in a staircase manner as plotted in Figure 3 (for formal definitions see
The following theorem shows that the estimator (5) can achieve the optimal convergence rate with

\[ \frac{\beta' - \beta}{\max\{\alpha, \beta' + p\}} \neq \frac{\gamma - \gamma'}{1 - \gamma}, \]

where the bias and variance contours overlap and we set \( x_{n+1} = \frac{1}{2} x_n \).

Each rectangular represents a certain level of regularization.

Appendix C). The eigenbasis \( \{\rho_j f_j\} \) of the output space is divided into different levels by \( \{y_i\} \).

The main idea behind our multilevel method is that different levels of the output need to be learned with different regularization. Formally, we define our multilevel estimator as

\[
\hat{A}_{ml} = \sum_{i=0}^{L_N} \left( \sum_{y_{i-1} < y_i} \rho_j^\frac{1}{2} f_j \otimes \rho_j^\frac{1}{2} f_j \right) \hat{C}_{LK} \left( \hat{C}_{KK} + \lambda^{(K)} I \right)^{-1}. \tag{5}
\]

The following theorem shows that the estimator (5) can achieve the optimal convergence rate with

\[ L_N = \mathcal{O}(\ln \ln N) \]

when \( \frac{\beta - \beta'}{\max\{\alpha, \beta' + p\}} \neq \frac{\gamma - \gamma'}{1 - \gamma} \). We also show that \( \mathcal{O}(\ln N) \) estimator is needed for the case when \( \frac{\beta - \beta'}{\max\{\alpha, \beta' + p\}} = \frac{\gamma - \gamma'}{1 - \gamma} \) (Figure 3 Right) in Appendix C.

**Theorem 5.1** Suppose that Assumptions 2.1 to 2.5 hold, then there exists a sequence \( \{\hat{y}_i\} \) with \( L_N = \mathcal{O}(\ln N) \) when \( \frac{\beta - \beta'}{\max\{\alpha, \beta' + p\}} \neq \frac{\gamma - \gamma'}{1 - \gamma} \) and \( \mathcal{O}(\ln \ln N) \) otherwise, such that the estimator \( \hat{A}_{ml} \) satisfies

\[
\|\hat{A}_{ml} - A_0\|_{\beta', \gamma'}^2 \leq C_T^2 \left( \frac{N}{\log N} \right) - \min\{ \frac{\beta' - \beta}{\max\{\alpha, \beta' + p\}}, \frac{\gamma - \gamma'}{1 - \gamma} \} \log^2 N \text{ with probability } \geq 1 - e^{-\tau}, \]

where \( C \) is a universal constant.

**Remark 5.1** Our multilevel algorithm first apply the regression algorithm on low-frequency projections of the output samples with small regularization and then successively fine-tune the regression model on high-frequency projections of the output samples with stronger regularization, which matches the empirical use (Li et al., 2021a; Lye et al., 2021).

6 CONCLUSION AND DISCUSSION

We considered the sample complexity of learning an operator between two infinite-dimensional Sobolev kernel Hilbert spaces. We provided an information theoretical lower bound for this problem along with a multi-level machine learning algorithm. Our lower bound is determined by the harder rate of two polynomial rates: one is fully determined by the hardness of the input space, while the other is fully controlled by the hardness of the output space. The second rate is new to the literature. We explained our bound from the viewpoint of variance and bias counters in Remark 4.2 and Figure 2. The optimal estimator should learn all the spectral elements under the bias contour but learn no information above the variance contour. To meet this requirement, we combined the idea of multi-level Monte Carlo with kernel operator learning, using successive levels to fit higher frequency information while keeping the variance at the same scale to reduce the bias. Our paper is the first on the non-parametric statistical optimality for multi-level algorithms. We leave estimation from discretely observed functional covariates with noise as future work (Zhou et al.; 2022).

REFERENCES


Hang Zhou, Fang Yao, and Huiming Zhang. Functional linear regression for discretely observed data: From ideal to reality.

A Proof of the Lower Bound

In this section, we follow the lower bound proof in Fischer & Steinwart (2020) to give a lower bound of the convergence rate in our operator learning setting.

A.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-Gilbert Lemma.

Lemma A.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

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

In our proof we will use a version from Fischer & Steinwart (2020).

Lemma A.2 (Fischer & Steinwart, 2020, Theorem 20) Let $M \geq 2$, $(\Omega, \mathcal{A})$ be a measurable space, $P_0, P_1, \ldots, P_M$ be probability measures on $(\Omega, \mathcal{A})$ with $P_j \ll P_0$ for all $j = 1, \ldots, M$, and $0 < \alpha_* < \infty$ with

$$\frac{1}{M} \sum_{j=1}^{M} KL(P_j||P_0) \leq \alpha_*.$$ 

Then, for all measurable functions $\Psi : \Omega \rightarrow \{0, 1, \ldots, M\}$, the following bound is satisfied

$$\max_{j=0,1,\ldots,M} P_j(\omega \in \Omega : \Psi(\omega) \neq j) \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - \frac{3\alpha_*}{\log(M)} - \frac{1}{2\log(M)}\right).$$

Lemma A.3 (Varshamov-Gilbert Lemma, Tsybakov (2008) Theorem 2.9) Let $D \geq 8$. There exists a subset $\mathcal{V} = \{\tau^{(0)}, \ldots, \tau^{(2D/8)}\}$ of $D$-dimensional hypercube $\mathcal{H}^D = \{0, 1\}^D$ such that $\tau^{(0)} = (0, 0, \ldots, 0)$ and the $\ell_1$ distance between every two elements is larger than $\frac{D}{8}$.

A.2 Proof of the Lower Bound

To prove our lower bound, we construct a sequence of linear operators as follows:

$$A_\omega = \sqrt{\frac{32\varepsilon}{m_1K}} \sum_{i=1}^{m_1} \sum_{j=1}^{K} \omega_{ij} \mu_{i+m_1j+1}^{1/2} \mu_{j+m_2}^{1/2} f_{j+m_2} \otimes e_{i+m_1}, \quad \omega_{ij} \in \{0, 1\}$$

where $m_1$ and $m_2$ are hyper-parameters (scale as $\text{poly}(N)$ and will be selected later) and $K$ is a constant that will be specified afterwards. It’s easy to check that

$$\|A_\omega - A_\omega'\|_{\beta', \gamma'} \leq \frac{32\varepsilon}{m_1K} \sum_{i=1}^{m_1} \sum_{j=1}^{K} (\omega_{ij} - \omega'_{ij})^2$$

By Gilbert-Varshamov Lemma it is possible to select $M_\varepsilon \geq 2^{m_1K/8}$ binary strings

$$\omega^{(1)}, \omega^{(2)}, \ldots, \omega^{(M_\varepsilon)} \in \{0, 1\}^{m_1K}$$

such that $\|\omega^{(i)} - \omega^{(j)}\|_2^2 \geq 4\varepsilon$. Let $\Omega$ be the collection of this strings.

We now select the hyper-parameters to satisfies the assumptions made in Section 2. First we have

$$\|A_\omega\|_{\beta, \gamma}^2 \leq \frac{32\varepsilon}{m_1K} \sum_{i=1}^{m_1} \sum_{j=1}^{K} (\beta - \beta') (\gamma' - \gamma) \lesssim \varepsilon (2m_1)^{\frac{\beta - \beta'}{\beta}} (2m_2)^{\frac{\gamma' - \gamma}{\gamma}}$$

14
where the last step follows from Assumption 2.1. Similarly, we have $\|A_\var{\omega}\|_{\alpha, 1}^2 \lesssim \varepsilon (2m_1)^{\frac{1-\beta'}{2}} (2m_2)^{\frac{\gamma'-1}{2}}$. To the assumptions made in Section 2, we should make

$$
(2m_1)^{\frac{\max\{a, \beta\}-\frac{\beta'}{2}}{r}} (2m_2)^{\frac{\gamma'-2}{2}} \lesssim \varepsilon^{-1}
$$

(6)

be satisfied. To be specific, with the previous selection of hyper-parameters, we can have $\|A_\var{\omega}\|_{\beta, \gamma} = O(1)$ and

$$
\sup_{g \in \text{range}(A_\var{\omega})} \|g\| \leq \sup_f \|A_\var{\omega}\|_\alpha \cdot \left\| \left( I_{H_\var{\omega}, P_\var{\omega}} \right)^\dagger f \right\|_{H_\var{\omega}} < +\infty
$$

where the last step follows from our assumption on the input distribution Assumption 2.2. This verifies that Assumptions 2.3 and 2.5 hold for $A_\var{\omega}, \forall \omega \in \Omega$.

We now construct the hypothesis (probability distributions) as follows: for $\forall \omega \in \{0, 1\}^{m_1}$, define

$$
P_\var{\omega}(df, dg) = dN(C_\var{\omega} f, \Sigma) \cdot dP_\var{\omega}(f)
$$

where the covariance operator $\Sigma = \frac{\sigma^2}{K} \sum_{j=1}^K \rho_j f_j + m_2 \otimes f_j + m_2$ for some constant $\sigma > 0$.

It’s then easy to see that $\text{tr}(\Sigma) = \sigma^2$, which satisfies Assumption 2.4. Note that the range of $A_\var{\omega}$ is $\text{span}(f_{m_2})$ and $\Sigma$ is non-degenerate on this subspace. As a result, we can view $P_\var{\omega}, \omega \in \Omega$ as distributions on $H_\var{\omega} \times \text{span}(f_{j+m_2} : 1 \leq j \leq K)$, and we have for $\var{\omega}, \var{\omega}' \in \Omega$ that

$$
KL(P_\var{\omega} || P_\var{\omega}') = E_{f \sim P_\var{\omega}} \left( KL(P_\var{\omega}(df) || P_\var{\omega}'(dg)) \right)
$$

$$
= E_{f \sim P_\var{\omega}} \left( KL(N(C_\var{\omega} f, \Sigma)||N(C_\var{\omega}' f, \Sigma)) \right)
$$

$$
\leq \sigma^{-2} K E_{f \sim P_\var{\omega}} \langle (A_\var{\omega} - A_\var{\omega}') f, \Sigma (A_\var{\omega} - A_\var{\omega}') f \rangle
$$

$$
= \frac{32\varepsilon}{m_1 \sigma^2} E_{f \sim P_\var{\omega}} \left( \sum_{i=1}^{m_1} \sum_{j=1}^{m_1} (\omega_{ij} - \omega'_{ij})^2 \mu_{j+m_1}^{1-\gamma'/2} \rho_j f_i f_{i+m_1} \right)^2_{H_\var{\omega}}
$$

$$
= \frac{32\varepsilon}{m_1 \sigma^2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_1} (\omega_{ij} - \omega'_{ij})^2 \mu_{j+m_1}^{1-\gamma'/2} \rho_j f_i f_{i+m_1} \lesssim \varepsilon^{-2} m_1^{-\frac{1-\beta'}{2}} m_2^{-\frac{1-\gamma'}{2}}
$$

where the last step follows from $E_{P_\var{\omega} u \otimes u} = C_{KK} = \sum_{i=1}^{\infty} \mu_i^2 e_i \otimes e_i$ and recall that $K$ is a constant. Hence we deduce that

$$
\frac{1}{M_\varepsilon} \sum_{\omega \in \Omega} KL(P_\var{\omega} || P_\var{\omega}') \lesssim \sigma^{-2} m_1^{-\frac{\beta'}{2}} m_2^{-\frac{1-\gamma'}{2}} =: \alpha^*
$$

Applying Lemma A.2, we find that when

$$
\alpha^* \lesssim \log M_\varepsilon \leftrightarrow \varepsilon \lesssim n^{-1} m_1^{\frac{\beta'}{2}} m_2^{-\frac{1-\gamma'}{2}}
$$

there exists a hypothesis $P_\var{\omega_0}$ such that for any estimator $\hat{C}_\var{\omega_0}$,

$$
\left\{ \left\| \hat{A}_\var{\omega_0} - A_\var{\omega_0} \right\|_{\beta', \gamma'} \gtrsim \varepsilon \right\} \supset \left\{ \omega_0 \neq \arg \min_{\omega \in \Omega} \| A_\omega - A_{\omega_0} \|_{\beta', \gamma'} \right\}
$$

holds with high probability.

Finally, we need to choose optimal $m_1$ and $m_2$ under the constraint (6). It turns out that either $m_1 = 1$ or $m_2 = 1$, and the resulting lower bound is

$$
\left\| \hat{A} - A \right\|_{\beta', \gamma'} \gtrsim \min\left\{ n^{-\frac{\max\{a, \beta\}-\frac{\beta'}{2}}{\max\{a, \beta\}+\frac{\beta'}{2}}}, n^{-\frac{\gamma'-\gamma}{\gamma'-1}} \right\}
$$
B  PROOF OF THE UPPER BOUND

In this section, we upper-bound the learning error of estimator (3) which defined as

$$\hat{A} = \sum_{j=1}^{y_N} \left( \rho_j^\frac{1}{2} f_j \otimes \rho_j^\frac{1}{2} f_j \right) \hat{C}_{LK} \left( \hat{C}_{KK} + \lambda_j I \right)^{-1},$$  \hspace{1cm} (7)$$

where $\lambda_j, 1 \leq j \leq y_N = N \frac{a}{\bar{a}} \max \left\{ 1 - \frac{\beta - \beta'}{\max \left( a \delta_{ij} + \rho_j \right) \gamma}, \frac{y_j'}{y_j' + \gamma} \right\}$ are regularization coefficients that we impose on different dimensions of the output space. In this section, we consider the following two ways to select regularization coefficients in Section 4:

- We regularize all spectral component below certain variance contour, i.e. we set regularization strength $\lambda_j = \max \left\{ \left( \frac{\beta - \beta'}{\max \left( a \delta_{ij} + \rho_j \right) \gamma}, \frac{y_j'}{y_j' + \gamma} \right) \right\}$ (4).

- We regularize all spectral component below certain bias contour, i.e. we set regularization strength $\lambda_j = \max \left\{ \left( \frac{\beta - \beta'}{\max \left( a \delta_{ij} + \rho_j \right) \gamma}, \frac{y_j'}{y_j' + \gamma} \right) \right\}$ (19).

To obtain the upper bound for our estimator, we decompose the learning error $E(\hat{A}) = \| \hat{A} - A_0 \|_{\beta', \gamma'}$ in to bias and variance via

$$E(A) \leq \left\| \hat{A} - A_0 \right\|_{\beta', \gamma'} + \left\| A_\lambda - A_0 \right\|_{\beta', \gamma'},$$

where

$$A_\lambda = \sum_{j=1}^{y_N} \left( \rho_j^\frac{1}{2} f_j \otimes \rho_j^\frac{1}{2} f_j \right) C_{KL} \left( C_{KK} + \lambda_j I \right)^{-1}.$$  \hspace{1cm} (8)$$

B.1  REGULARIZATION VIA VARIANCE COUNTER

In the following, we separately bound the bias term and the variance term. We first assume $\alpha \leq \beta + p$ in Appendix B.1.1 and Appendix B.1.2, then the case $\alpha > \beta + p$ is treated in Appendix B.1.3. Finally in Appendix B.2, we establish the same convergence rate for regularization via bias contour.

B.1.1  BIAS

Lemma B.1 $\| A_0 - A_\lambda \|_{\beta', \gamma'} \leq N^{-\min\left\{ \frac{\beta - \beta'}{\beta + p}, \frac{1}{\gamma'} \right\}}$.

Proof sketch: Since $\| A_0 \|_{\beta, \gamma} \leq B$, we can write $A_0 := \sum_{i=1}^{+\infty} \sum_{j=1}^{+\infty} a_{ij} \mu_i^\frac{\beta}{2} \rho_j^1 f_j \otimes e_i$ where the coefficient matrix $A_0 = (a_{ij})_{1 \leq i, j \leq +\infty}$ satisfies $\| A_0 \|_F \leq B^2$. The definition (8) implies that for $1 \leq j \leq y_N$ and $i \geq 1$ we have

$$\langle \rho_j^\frac{1}{2} f_j, A_\lambda \mu_i^\frac{\beta}{2} e_i \rangle = \langle \rho_j^\frac{1}{2} f_j, C_{KL} \left( C_{KK} + \lambda_j I \right)^{-1} \mu_i^\frac{1}{2} e_i \rangle$$

$$= \langle \rho_j^\frac{1}{2} f_j, A_0 \left( C_{KL} \left( C_{KK} + \lambda_j I \right)^{-1} \right)^{-1} \mu_i^\frac{1}{2} e_i \rangle = \frac{\mu_i^\frac{1}{2} a_{ij}}{\mu_i + \lambda_j} \rho_j^\frac{1}{2} a_{ij}.$$
The bias term can be bounded as follows:

\[
\|A_0 - A_\lambda\|_{\beta', \gamma'}^2 \leq \sum_{i,j=1}^{+\infty} \left< \rho_j^\frac{1}{\beta'} f_j, C_{Q_K}^{-\frac{1-\gamma'}{\beta'}} (A_0 - A_\lambda) C_{K^K}^{\frac{1-\gamma'}{\beta'}} \mu_1^\frac{1}{\mu} e_i \right>^2
\]

\[
= \sum_{j=1}^{y_N} \frac{1}{\mu_j} (\mu_j + \lambda_j)^2 a_{ij}^2
\]

\[
\leq \sum_{j=1}^{y_N} \rho_j' \gamma' \max_{i \neq 1} \left( \frac{\mu_i - \beta'}{\mu_i + \lambda_j} \right)^2 \cdot \sum_{i=1}^{+\infty} a_{ij}^2
\]

\[
\leq \sum_{j=1}^{y_N} j^{-\frac{\gamma'}{\beta'}} \lambda_j^{-\beta'} \sum_{i=1}^{+\infty} a_{ij}^2 \leq B^2 \max_{1 \leq j \leq y_N} j^{-\frac{\gamma'}{\beta'}} \lambda_j^{-\beta'}.
\]

We now prove that

\[
j^{-\frac{\gamma'}{\beta'}} \lambda_j^{-\beta'} \geq N^{-\min\left\{ \frac{\beta' + p}{\beta + p}, \frac{1-\gamma'}{1-\gamma} \right\}}, \quad \forall 1 \leq j \leq y_N.
\]

**Case 1.** If \( \lambda_j = c_0 \left( \frac{N}{\log N} \right)^{\frac{1}{\beta'}} \), then

\[
j^{-\frac{\gamma'}{\beta'}} \lambda_j^{-\beta'} \geq N^{-\min\left\{ \frac{\beta' + p}{\beta + p}, \frac{1-\gamma'}{1-\gamma} \right\}}, \quad \forall 1 \leq j \leq y_N.
\]

where we use \( \alpha \leq \beta + p \) in the final step.

**Case 2.** If \( \lambda_j = N^{\max\left\{ \frac{\beta' + p}{\beta + p}, \frac{1-\gamma'}{1-\gamma} \right\}} \), we need to consider two sub-cases:

- If \( \frac{\beta' + p}{\beta + p} > \frac{1-\gamma'}{1-\gamma} \), then we have \( \lambda_j = \left( N^{\frac{\beta' + p}{\beta + p}} j^{-\frac{1-\gamma'}{\beta'}} \right)^{\frac{1}{\beta + p}} \) and thus

\[
j^{-\frac{\gamma'}{\beta'}} \lambda_j^{-\beta'} = j^{-\frac{\gamma'}{\beta'}} \left( N^{\frac{\beta' + p}{\beta + p}} j^{-\frac{1-\gamma'}{\beta'}} \right)^{\frac{\beta' - \beta}{\beta + p}} = N^{\frac{\beta' - \beta}{\beta + p}} j^{-\frac{1-\gamma'}{\beta'}} \left( \frac{\beta' - \beta}{\beta + p} \right) \geq N^{\frac{\beta' - \beta}{\beta + p}}.
\]

- If \( \frac{\beta' + p}{\beta + p} < \frac{1-\gamma'}{1-\gamma} \), then similarly we have \( \lambda_j = \left( N^{\frac{1-\gamma'}{1-\gamma}} j^{-\frac{1-\gamma'}{\beta'}} \right)^{\frac{1}{\beta + p}} \) and

\[
j^{-\frac{\gamma'}{\beta'}} \lambda_j^{-\beta'} = j^{-\frac{\gamma'}{\beta'}} \left( N^{\frac{1-\gamma'}{1-\gamma}} j^{-\frac{1-\gamma'}{\beta'}} \right)^{\frac{\beta' - \beta}{\beta + p}} \geq j^{-\frac{\gamma'}{\beta'}} \left( N^{\frac{1-\gamma'}{1-\gamma}} j^{-\frac{1-\gamma'}{\beta'}} \right)^{\frac{\beta' - \beta}{\beta + p}} = N^{\frac{\beta' - \beta}{\beta + p}}.
\]

Hence, in all cases (10) holds and we have that

\[
\|A_0 - A_\lambda\|_{\beta', \gamma'}^2 \lesssim N^{-\min\left\{ \frac{\beta' + p}{\beta + p}, \frac{1-\gamma'}{1-\gamma} \right\}}.
\]
B.1.2 Variance

The variance term can be rewritten in the following way:

\[
\mathcal{V} = \left\| \hat{A} - A_{\lambda} \right\|^2_{\beta', \gamma'} = \left\| C_{KK}^{-1/2} \left( \hat{A} - A_{\lambda} \right) C_{KK}^{1/2} \right\|^2_{HS}
\]

\[
\begin{aligned}
&= \sum_{i,j=1}^{\infty} \left\langle \rho_j^2 f_j, C_{KK}^{1/2} \left( \hat{A} - A_{\lambda} \right) C_{KK}^{-1/2} \mu_i e_i \right\rangle^2 \\
&= \sum_{j=1}^{\infty} \rho_j^{-(1-\gamma)} \sum_{i=1}^{\infty} \left\langle \rho_j^2 f_j, \left[ \hat{C}_{LK} \left( \hat{C}_{KK} + \lambda_j I \right)^{-1} - C_{LK} \left( C_{KK} + \lambda_j I \right)^{-1} \right] \mu_i e_i \right\rangle^2 \\
&= \sum_{j=1}^{\infty} \rho_j^{-(1-\gamma)} \sum_{i=1}^{\infty} \left\langle (C_{KK} + \lambda_j I)^{-1/2} \left[ \hat{C}_{KL} - \left( \hat{C}_{KK} + \lambda_j I \right) (C_{KK} + \lambda_j I)^{-1} C_{KL} \right] \rho_j^2 f_j, \right\rangle^2 \\
&= \sum_{j=1}^{\infty} \rho_j^{-(1-\gamma)} U_j \left( \rho_j^2 f_j, G_j \left( \sum_{i=1}^{\infty} \mu_i^{2-\beta'} \mu_i + \lambda_j e_i \otimes e_i \right) G_j \rho_j^2 f_j \right) \\
&\lesssim \sum_{j=1}^{\infty} \rho_j^{-(1-\gamma)} \left\| G_j \right\|^2 \lambda_j^{-\beta'} \left\| U_j \rho_j^2 f_j \right\|^2
\end{aligned}
\]

In (12), (12a) uses the definition of the Hilbert-Schmidt norm; (12b) follows from the definition of \( \hat{A} \) (cf.(3)) and the fact that for any \( j \geq y_N \), we have \( \left\langle \rho_j^2 f_j, \left( \hat{A} - A_{\lambda} \right) \mu_i^2 e_i \right\rangle = 0 \); (12c) is obtained from re-arranging and (12d) follows from

\[
\sum_{i=1}^{\infty} \frac{\mu_i^{2-\beta'}}{\mu_i + \lambda_j} \lesssim \lambda_j^{-\beta'}
\]

Note that

\[
U_j = (C_{KK} + \lambda_j I)^{-1/2} \left[ \hat{C}_{KL} - C_{KL} - \left( \hat{C}_{KK} - C_{KK} \right) (C_{KK} + \lambda_j I)^{-1} C_{KL} \right]
\]

\[
= \frac{1}{N} \sum_{k=1}^{N} (C_{KK} + \lambda_j I)^{-1/2} u_k \otimes v_k - \mathbb{E}_{P_{KL}} u_k \otimes A_{0} u_k - (u_k \otimes u_k - \mathbb{E}_{P_{KL}} u_k \otimes u_k) (C_{KK} + \lambda_j I)^{-1} C_{KK} A_{0}
\]

\[
= \frac{1}{N} \sum_{k=1}^{N} (C_{KK} + \lambda_j I)^{-1/2} (u_k \otimes (v_k - A_{0} u_k))
\]

\[
:= U_j^1
\]

\[
+ \frac{1}{N} \sum_{k=1}^{N} (C_{KK} + \lambda_j I)^{-1/2} \left[ u_k \otimes A_{0} u_k - \mathbb{E}_{P_{KL}} u_k \otimes A_{0} u_k - (u_k \otimes u_k - \mathbb{E}_{P_{KL}} u_k \otimes u_k) (C_{KK} + \lambda_j I)^{-1} C_{KK} A_{0} \right]
\]

\[
:= U_j^2 = \lambda_j \frac{1}{N} \sum_{k=1}^{N} (C_{KK} + \lambda_j I)^{-1/2} \left[ u_k \otimes A_{0} (C_{KK} + \lambda_j I)^{-1} u_k - \mathbb{E}_{P_{KL}} u_k \otimes A_{0} (C_{KK} + \lambda_j I)^{-1} u_k \right]
\]

The \( U_j^1 \) term is the variance of observational noise and \( U_j^2 \) term is the variance of regularized bias. Thus the \( U_j^1 \) term is the dominating term. Plugging the above decomposition into (12), we deduce...
Recall that $V = 2 (V_1 + V_2)$ where

$$V_1 \lesssim \max_{1 \leq j \leq n_N} \| G_j \|^2 \sum_{j=1}^{n_N} \| \frac{1}{N} \sum_{k=1}^N \left( v_k - A_0 u_k, \rho_j^2 f_j \right) (C_{KK} + \lambda_j I)^{-\frac{1}{2}} u_k \|^2$$

$$V_2 \lesssim \max_{1 \leq j \leq n_N} \| G_j \|^2 \sum_{j=1}^{n_N} \| \frac{1}{N} \sum_{k=1}^N \left( \hat{E} - E \right) \left( \left( A_0 (C_{KK} + \lambda_j I)^{-1} u_k, \rho_j^2 f_j \right) (C_{KK} + \lambda_j I)^{-\frac{1}{2}} u_k \right) \|^2$$

where $\hat{E}[X] = \frac{1}{N} \sum_{k=1}^N X_k$ denotes the empirical mean. Define the event

$$E_{1,j} = \left\{ G_j = \left( [P_{ij} (C_{KK})]^{\frac{1}{2}} [P_{ij} (\hat{C}_{KK})]^{\frac{1}{2}} \right] \leq 2 \sqrt{\alpha_1} \right\}.$$

Recall that $m_N \leq c_0 \left( \frac{N}{\log N} \right)^{\frac{p}{q}}$, by Theorem D.3, we know that $E_{1,j}$ holds with probability $\geq 1 - 2e^{-\alpha_1}$. As a result $E_1 = \cap_{j=1}^{n_N} E_{1,j}$ holds with probability $\geq 1 - 2m_N e^{-\alpha_1}$. We assume event $E_1$ holds in all the following proof.

**Bounding** $V_1$. Let

$$X_{j,k} = j^{\frac{1 - \gamma'}{2} - \frac{\alpha'}{2}} \left( v_k - A_0 u_k, \rho_j^2 f_j \right) (C_{KK} + \lambda_j I)^{-\frac{1}{2}} u_k \in H_K$$

and $X_k = (X_{j,k} : 1 \leq j \leq n_N) \in H_K^{\otimes N}$. Then we have $V_1 \lesssim \left\| \frac{1}{N} \sum_{k=1}^N X_k \right\|^2$ where the norm here defined for $H_K^{\otimes N}$ is induced by $(a, b) = \sum_{i=1}^{n_N} (a_i, b_i)_{H_K}$. Note that $X_k, k = 1, 2, \ldots, N$ are i.i.d. random variables with mean zero, and

$$E \| X_1 \|^{2t} = E_{P_{KL}} \left[ \left( \sum_{j=1}^{n_N} \| X_{j,k} \|^2 \right)^t \right]$$

$$= E_{P_{KL}} \left[ \left( \sum_{j=1}^{n_N} j^{\frac{1 - \gamma'}{2} - \frac{\alpha'}{2}} \lambda_j^{-\beta'} \left( v_k - A_0 u_k, \rho_j^2 f_j \right)^2 \left\| (C_{KK} + \lambda_j I)^{-\frac{1}{2}} u \right\|^2 \right) \right]$$

$$\leq \max_{1 \leq j \leq n_N} \sup_{u \in \operatorname{sup}(P_{KL})} \left( j^{\frac{1 - \gamma'}{2} - \frac{\alpha'}{2}} \lambda_j^{-\beta'} \left\| (C_{KK} + \lambda_j I)^{-\frac{1}{2}} u \right\|^2 \right)^{t-1}$$

$$\lesssim \left( \sum_{j=1}^{n_N} \lambda_j^{-\beta'} \left\| (C_{KK} + \lambda_j I)^{-\frac{1}{2}} u \right\|^2 \right)^{t-1} \leq G_1$$

$$E_{(u,v) \sim P_{KL}} \left[ \| v - A_0 u \|^{2t-2} \left( \sum_{j=1}^{n_N} j^{\frac{1 - \gamma'}{2} - \frac{\alpha'}{2}} \lambda_j^{-\beta'} \left( v - A_0 u, \rho_j^2 f_j \right)^2 \left\| (C_{KK} + \lambda_j I)^{-\frac{1}{2}} u \right\|^2 \right) \right]$$

$$= G_2$$

By Lemma D.2 we have

$$G_1 \lesssim j^{\frac{1 - \gamma'}{2} - \frac{\alpha'}{2}} \lambda_j^{-(\beta' + \alpha)}.$$
where \( \sigma^2 = \langle \rho_{ij}^2 f_j, V \rho_{ij}^2 f_j \rangle \). As a result, we have

\[
G_2 \leq E_{P_K} \left[ \frac{1}{2}(2t)!R^{2t-2} \sum_{j=1}^{nN} \sigma^2_j \left\| (C_{KK} + \lambda_j I)^{-\frac{1}{2}} u \right\|^2 \right] \leq \frac{1}{2}(2t)!R^{2t-2} \sigma^2 \max_{1 \leq j \leq nN} \frac{1}{\lambda_j^{(p+p')}} \lambda_j^{-(p+p')},
\]

where in the second step we use \( \sum_{j=1}^{\infty} \sigma^2_j = tr(V) = \sigma^2 \) and

\[
E_{P_K} \left[ \left\| (C_{KK} + \lambda_j I)^{-\frac{1}{2}} u \right\|^2 \right] \leq tr \left( E_{P_K} \left[ (C_{KK} + \lambda_j I)^{-\frac{1}{2}} u \otimes (C_{KK} + \lambda_j I)^{-\frac{1}{2}} u \right] \right)
= tr \left( \sum_{i=1}^{\infty} \frac{\mu_i}{\mu_i + \lambda_j} e_i \otimes e_i \right)
= \sum_{i=1}^{\infty} \frac{\mu_i}{\mu_i + \lambda_j}
\leq \lambda_j^{-p}.
\]

We have shown that for some constant \( c_1 > 0 \),

\[
E\|X_1\|^{2t} \leq \frac{1}{2}(2t)!\sigma^2 \max_{1 \leq j \leq nN} \frac{1}{\lambda_j^{(p+p')}} \left( c_1 R^2 \max_{1 \leq j \leq nN} \frac{1}{\lambda_j^{(p+p)}} \right)^{t-1}.
\]

By Bernstein’s inequality, the event

\[
E_2 := \left\{ \frac{1}{N} \sum_{k=1}^{N} X_k \right\}^2 \leq 6a_2 \left( \frac{\sigma^2 \max_{1 \leq j \leq nN} \frac{1}{\lambda_j^{(p+p')}}}{N} + \frac{c_1 R^2 \max_{1 \leq j \leq nN} \frac{1}{\lambda_j^{(p+p)}}}{N^2} \right)
\]

holds with probability \( \geq 1 - 2e^{-a_2} \). By our definition of \( \lambda_j \), we have

\[
\max_{1 \leq j \leq nN} \frac{1}{\lambda_j^{(p+p')}} \lesssim N^{\min \left\{ \frac{p}{p+p'}, \frac{1}{2} \right\}}
\]

and \( \lambda_j \gtrsim N^{-\frac{1}{2}} \) (which implies that the \( \frac{1}{N^2} \) term is dominated by the \( \frac{1}{N} \) term). Hence, under \( \mathcal{E}_1 \cap \mathcal{E}_2 \) we have

\[
\mathcal{V}_1 \lesssim a_1 a_2 \sigma^2 N^{-\min \left\{ \frac{p}{p+p'}, \frac{1}{2} \right\}}
\]

with probability \( \geq 1 - 2a_2 e^{-a_2} \).

**Bounding \( \mathcal{V}_2 \).** For any \( j \in \mathbb{Z}_+ \) we have

\[
E_{u \sim P_K} \left[ \left\langle A_0 (C_{KK} + \lambda_j I)^{-1} u, \rho_j^\frac{1}{2} f_j \right\rangle \right]^2
= E_{u \sim P_K} \left\langle \rho_j^\frac{1}{2} f_j, E_{P_K} \left[ A_0 (C_{KK} + \lambda_j I)^{-1} u \otimes A_0 (C_{KK} + \lambda_j I)^{-1} u \right] \rho_j^\frac{1}{2} f_j \right\rangle
= \left\langle \rho_j^\frac{1}{2} f_j, A_0 (C_{KK} + \lambda_j I)^{-1} C_{Q_{KK}} (C_{KK} + \lambda_j I)^{-1} \rho_j^\frac{1}{2} f_j \right\rangle
= \rho_j^{1-\gamma} \left\langle C_{Q_{KK}} \rho_j^{\frac{1}{2}} f_j, (C_{KK} + \lambda_j I)^{-1} C_{Q_{KK}} (C_{KK} + \lambda_j I)^{-1} \right\langle \rho_j^\frac{1}{2} f_j \right\rangle
\leq j^{\frac{1}{2} - \gamma} \lambda_j^{-(2-\beta)} \left\langle C_{Q_{KK}} \rho_j^{\frac{1}{2}} f_j \right\rangle^2 \equiv \mathcal{D}_{1,2}
\]

where (15a) follows from \( E_{P_K} u \otimes u = C_{KK} \), (15b) uses the fact that \( C_{KK} \) and \( C_{KK} + \lambda_j I \) commute, and lastly (15c) follows from

\[
\left\| (C_{KK} + \lambda_j I)^{-1} C_{Q_{KK}} (C_{KK} + \lambda_j I)^{-1} \right\|_{\mathcal{H}_K} \propto \lambda_j^{-(2-\beta)}.
\]
Let \( Y_{j,k} = \langle A_0 (C_{KK} + \lambda_j I)^{-1} u_k, \rho_j^{1/2} f_j \rangle (C_{KK} + \lambda_j I)^{-\frac{1}{2}} u_k \in \mathcal{H}_K \)

and

\[ Y_k = (Y_{j,k} : 1 \leq j \leq y_N) \in \mathcal{H}_K^{n_N}. \]

Then we have

\[ \mathcal{V}_2 \lesssim \left\| \frac{1}{N} \sum_{k=1}^N Y_k \right\|_{\mathcal{H}_K^{n_N}}^2. \]

Note that \( Y_k, k = 1, 2, \cdots, N \) are i.i.d. random variables, and

\[ \mathbb{E}\|Y_1\|^{2t} = \mathbb{E}\left[ \left( \sum_{j=1}^{n_N} |Y_{j,k}|^2 \right)^t \right] = \mathbb{E}_{P_K} \left[ \left( \sum_{j=1}^{n_N} \sup_{u \in \text{supp}(P_K)} \left( \sum_{j=1}^{n_N} \frac{1}{\lambda_j^{2-\beta'}} \langle A_0 (C_{KK} + \lambda_j I)^{-1} u, \rho_j^{1/2} f_j \rangle^2 \right) \right)^t \right] \]

\[ \leq \sup_{u \in \text{supp}(P_K)} \left( \sum_{j=1}^{n_N} \frac{1}{\lambda_j^{2-\beta'}} \langle A_0 (C_{KK} + \lambda_j I)^{-1} u, \rho_j^{1/2} f_j \rangle^2 \right)^{t-1} \sum_{j=1}^{n_N} \frac{1}{\lambda_j^{2-\alpha}} \lambda_j^{-(\beta' + \alpha - \beta)} D_{j,2}. \]

For any \( j \in \mathbb{Z}_+ \) and \( u \in \text{supp}(P_K) \) we have

\[ \sum_{j=1}^{n_N} \frac{1}{\lambda_j^{1-\beta'}} \lambda_j^{-(\beta' + \alpha)} \langle A_0 (C_{KK} + \lambda_j I)^{-1} \lambda_j u, \rho_j^{1/2} f_j \rangle^2 \]

\[ \leq \sum_{j=1}^{n_N} \frac{1}{\lambda_j^{1-\beta'}} \lambda_j^{-(\beta' + \alpha)} \left( 2 \langle A_0 u, \rho_j^{1/2} f_j \rangle^2 + 2 \rho_j^{1-\gamma} \langle C_{Q_K}^{-\frac{1}{2}} A_0 (C_{KK} + \lambda_j I)^{-1} C_{KK} u, \rho_j^{1/2} f_j \rangle^2 \right) \]

\[ \leq \max_{1 \leq j \leq y_N} \frac{1}{\lambda_j^{1-\beta'}} \lambda_j^{-(\beta' + \alpha)} + \sum_{1 \leq j \leq y_N} \lambda_j^{-(\beta' + \alpha)} \lambda_j^{\max \{ \alpha - \beta, 0 \} \left( \langle C_{Q_K}^{-\frac{1}{2}} A_0 C_{KK}^{-\frac{1}{2}} \rangle \right)^* \rho_j^{1/2} f_j \|}^2 \]

\[ \leq \max_{1 \leq j \leq y_N} \frac{1}{\lambda_j^{1-\beta'}} \lambda_j^{-(\beta' + \alpha)} + \lambda_j^{-(\beta' + \alpha) - \max \{ \alpha - \beta, 0 \}} \]

where (17a) uses the AM-GM inequality, (17b) follows from the assumption that \( \|A_0 u\| \leq A_2 \) is uniformly bounded, and that

\[ \left\| C_{KK}^{-\frac{1}{2}} (C_{KK} + \lambda_j I)^{-1} C_{KK} u \right\| = \left\| C_{KK}^{-\frac{1}{2}} (C_{KK} + \lambda_j I)^{-1} (C_{KK}^{-1})^* u \right\| \lesssim \lambda_j^{\max \{ \alpha - \beta, 0 \}}. \]

by Assumption 2.2, and lastly (17c) follows from \( \|A_0\|_{\beta, \gamma} \leq B \).

Plugging into (16), we deduce that

\[ \mathbb{E}\|Y_1\|^{2t} \lesssim \sup_{u \in \text{supp}(P_K)} \left( \max_{1 \leq j \leq y_N} \frac{1}{\lambda_j^{1-\beta'}} \lambda_j^{-(\beta' + \alpha)} + \lambda_j^{-(\beta' + \alpha) - \max \{ \alpha - \beta, 0 \}} \right)^{t-1} \sum_{j=1}^{n_N} \frac{1}{\lambda_j^{1-\beta'}} \lambda_j^{-(\beta' + \alpha - \beta)} D_{j,2} \]

\[ \lesssim \sup_{u \in \text{supp}(P_K)} \left( \max_{1 \leq j \leq y_N} \frac{1}{\lambda_j^{1-\beta'}} \lambda_j^{-(\beta' + \alpha)} + \lambda_j^{-(\beta' + \alpha) - \max \{ \alpha - \beta, 0 \}} \right)^{t-1} \max_{1 \leq j \leq y_N} \frac{1}{\lambda_j^{1-\beta'}} \lambda_j^{-(\beta' + \alpha - \beta)} \]
where the last step follows from $\sum_{j=1}^{+\infty} D_{j;2} = \|A_0\|_{\beta,\gamma}^2$.

By Bernstein’s inequality, there exists a constant $C_3$ such that the event
\[
E_3 = \left\{ \mathcal{V}_2 \leq 6a_3C_3 \left( \frac{j^{1-\frac{1}{2} - (\beta' + \alpha - \beta)}}{N} + \max_{1 \leq j \leq n_N} \lambda_j^{-(\beta' + \alpha)} \left( \frac{1}{N^2} + \lambda_j^{-\max(\alpha, \beta, 0)} \right) \right) \right\}
\]
holds with probability $\geq 1 - 2e^{-a_3}$.

The definition of $\lambda_N$ implies that the $\frac{1}{N^2}$ term is dominated by the $\frac{1}{N}$ term, so
\[
\mathcal{V}_2 \leq a_1a_3 \frac{1}{N} \max_{1 \leq j \leq n_N} j^{1-\frac{1}{2} - (\beta' + \alpha - \beta)} \lesssim N^{-\min\left\{ \frac{\beta - \beta'}{2(\beta' + \alpha - \beta)}, \frac{\gamma'}{\gamma - \gamma'} \right\}}
\]
holds under $E_1 \cap E_3$. To summarize, under $E_1 \cap E_2 \cap E_3$ which holds with probability $\geq 1 - 2n_Ne^{-a_3} - 2e^{-a_2} - 2e^{-a_3}$, we have
\[
\mathcal{V} \leq 2a_1 \max\{a_2, a_3\} (\mathcal{V}_1 + \mathcal{V}_2) \lesssim N^{-\min\left\{ \frac{\beta - \beta'}{2(\beta' + \alpha - \beta)}, \frac{\gamma'}{\gamma - \gamma'} \right\}}.
\]
Recall that the bias term is upper bounded in (11). This gives the final upper bound
\[
\left\| \hat{A} - A_0 \right\|_{\beta', \gamma'} \lesssim N^{-\min\left\{ \frac{\beta - \beta'}{2(\beta' + \alpha - \beta)}, \frac{\gamma'}{\gamma - \gamma'} \right\}}.
\]

### B.1.3 The hard-learning regime

In the previous sections, we focus on the case where $\alpha \leq \beta + \gamma$ and establish an upper bound for the convergence rate via an optimal bias-variance trade-off. The opposite case, $\alpha > \beta + \gamma$ is referred to as the hard-learning regime, for which the optimal rate is not known for several decades even in the case of $\gamma = 1$ (cf. the discussion following (Fischer & Steinwart, 2020, Theorem 2)). In the hard learning regime the $V_2$ term becomes the leading terms.

In this section, we use the technique developed in previous sections to obtain an upper bound in the hard-learning regime. To do this, we need to re-define the truncation set $S_N$ as follows:
\[
S_N = \left\{ (x, y) \in \mathbb{Z}^2 \mid x^{\beta + \gamma - \beta} y^{1 - \gamma'} \lesssim N^{1-\min\left\{ \frac{\beta - \beta'}{2(\beta' + \alpha - \beta)}, \frac{\gamma'}{\gamma - \gamma'} \right\}} \text{ and } x \leq c_0 \left( \frac{N}{\log N} \right)^{\frac{2}{\beta'}} \right\}.
\]

The definition implies that the variance can be controlled by $N^{-\min\left\{ \frac{\beta - \beta'}{2(\beta' + \alpha - \beta)}, \frac{\gamma'}{\gamma - \gamma'} \right\}}$ and it remains to focus on the bias term.

Similar to the derivations in Appendix B.1.1, we have
\[
\|A_0 - T_N(A_0)\|_{\beta', \gamma'} \lesssim \max_{(i,j) \notin S_N} i^{\beta - \beta'} j^{-\gamma' + \gamma}.
\]

The maximum value of the right hand side can be achieved in either of the following two cases:

- $i = O(1)$. Then we have $j \gtrsim N^{1-\frac{\beta - \beta'}{2(\beta' + \alpha - \beta)}} \left( 1-\min\left\{ \frac{\beta - \beta'}{2(\beta' + \alpha - \beta)}, \frac{\gamma'}{\gamma - \gamma'} \right\} \right)$ so that
  \[
i^{\beta - \beta'} j^{-\frac{\gamma' - \gamma}{\beta'}} \lesssim N^{-\frac{\gamma' - \gamma}{\beta'}} \left( 1-\min\left\{ \frac{\beta - \beta'}{2(\beta' + \alpha - \beta)}, \frac{\gamma'}{\gamma - \gamma'} \right\} \right) \lesssim N^{-\frac{\gamma' - \gamma}{\beta'}}.
  \]

- $j = O(1)$. In this case we must have $i \lesssim N^{\min\left\{ \frac{\beta - \beta'}{2(\beta' + \alpha - \beta)}, \frac{\gamma'}{\gamma - \gamma'} \right\}}$, otherwise it falls into $S_N$ by definition. Hence we have
  \[
i^{\beta - \beta'} j^{-\frac{\gamma' - \gamma}{\beta'}} \lesssim N^{-\min\left\{ \frac{\beta - \beta'}{2(\beta' + \alpha - \beta)}, \frac{\gamma'}{\gamma - \gamma'} \right\}}.
  \]

On the other hand, for the variance term we still have
\[
\mathcal{V}_1 \lesssim \frac{1}{N} \max_{1 \leq j \leq n_N} J_{1-\frac{1}{2} - (\beta' + \alpha - \beta)} \left( \frac{1}{N^2} + \lambda_j^{-\max(\alpha, \beta, 0)} \right) \text{ and } \mathcal{V}_2 \lesssim \frac{1}{N} \max_{1 \leq j \leq n_N} J_{1-\frac{1}{2} - (\beta' + \alpha - \beta)} \left( \frac{1}{N^2} + \lambda_j^{-\max(\alpha, \beta, 0)} \right),
\]

where
\[
E_3 = \left\{ \mathcal{V}_2 \leq 6a_3C_3 \left( \frac{j^{1-\frac{1}{2} - (\beta' + \alpha - \beta)}}{N} + \max_{1 \leq j \leq n_N} \lambda_j^{-(\beta' + \alpha)} \left( \frac{1}{N^2} + \lambda_j^{-\max(\alpha, \beta, 0)} \right) \right) \right\}
\]
As a result, we can obtain the following convergence rate:

\[ \left\| \hat{A} - A_0 \right\|_{\beta', \gamma'} \lesssim N^{-\min\left\{ \frac{\beta - \beta'}{2(1 - \gamma' \gamma)}, \frac{\gamma' - \gamma}{2(1 - \gamma' \gamma)} \right\}}. \]

## B.2 Regularization via Bias Contour

In this subsection, we analyze the convergence rate of regularization via bias contour (cf. Figure 2). Specifically, we consider the estimator (3) with the choice

\[ \lambda_j = \max \left\{ \left( j - \frac{\gamma' - \gamma}{q} N \min\left\{ \frac{\beta' - \beta}{\max(\alpha, \beta + p)}, \frac{\gamma' - \gamma}{1 - \gamma' \gamma} \right\} \right)^{-\frac{1}{2(1 - \gamma' \gamma)}}, c_0 \left( \frac{N}{\log N} \right)^{-\frac{1}{6}} \right\}. \]  

(19)

It now remains to plug the above \( \lambda_j \) into our bounds for bias and variance derived in the previous subsections.

### Bounding the bias term

It follows from (9) that

\[ \left\| A_0 - A_\lambda \right\|^2_{\beta', \gamma'} \lesssim \max_{1 \leq j \leq y_N} j^{-\frac{\gamma' - \gamma}{q}} \lambda_j^{\beta - \beta'} \]

\[ \lesssim \max \left\{ N^{-\min\left\{ \frac{\beta - \beta'}{\max(\alpha, \beta + p)}, \frac{\gamma' - \gamma}{1 - \gamma' \gamma} \right\}}, c_0 \left( \frac{N}{\log N} \right)^{-\frac{\beta - \beta'}{6}} \right\} \]

\[ \lesssim N^{-\min\left\{ \frac{\beta - \beta'}{\max(\alpha, \beta + p)}, \frac{\gamma' - \gamma}{1 - \gamma' \gamma} \right\}}. \]

### Bounding the variance term

It follows from (14) and (18) that the variance is bounded by

\[ \left\| \hat{A} - A_\lambda \right\|^2_{\beta', \gamma'} \lesssim \frac{1}{N} \max_{1 \leq j \leq y_N} j^{-\frac{\gamma' - \gamma}{q}} \left( \beta' + \max(\alpha - \beta, p) \right). \]

As before, we consider the cases \( \alpha \leq \beta + p \) and \( \alpha > \beta + p \) separately.

- **If \( \alpha \leq \beta + p \), then it follows that**

\[ \left\| \hat{A} - A_\lambda \right\|^2_{\beta', \gamma'} \lesssim \frac{1}{N} \max_{1 \leq j \leq y_N} j^{-\frac{\gamma' - \gamma}{q}} \lambda_j^{-\beta' + \beta + p} \]

\[ \lesssim \frac{1}{N} \max_{1 \leq j \leq y_N} j^{-\frac{\gamma' - \gamma}{q}} \left( j^{-\frac{\gamma' - \gamma}{q}} N \min\left\{ \frac{\beta - \beta'}{\max(\alpha + p, \beta + p)}, \frac{\gamma' - \gamma}{1 - \gamma' \gamma} \right\} \right)^{-\frac{\beta' + \beta + p}{2(1 - \gamma' \gamma)}} \]

\[ \lesssim \frac{1}{N} \max_{1 \leq j \leq y_N} j^{-\frac{\gamma' - \gamma}{q}} \left( j^{-\frac{\gamma' - \gamma}{q}} \frac{\beta' + \beta + p}{\max(\alpha + p, \beta + p)} \min\left\{ \frac{\beta - \beta'}{\max(\alpha + p, \beta + p)}, \frac{\gamma' - \gamma}{1 - \gamma' \gamma} \right\} \right) \]

\[ = \frac{1}{N} \max_{j \in \{1, y_N\}} \left( j^{-\frac{\gamma' - \gamma}{q}} \frac{\beta' + \beta + p}{\max(\alpha + p, \beta + p)} \min\left\{ \frac{\beta - \beta'}{\max(\alpha + p, \beta + p)}, \frac{\gamma' - \gamma}{1 - \gamma' \gamma} \right\} \right) \]

\[ = N^{-\min\left\{ \frac{\beta - \beta'}{\max(\alpha + p, \beta + p)}, \frac{\gamma' - \gamma}{1 - \gamma' \gamma} \right\}} \]

where we use \( y_N^{-\frac{\gamma' - \gamma}{q}} = N^{\min\left\{ \frac{\beta - \beta'}{\max(\alpha + p, \beta + p)}, \frac{\gamma' - \gamma}{1 - \gamma' \gamma} \right\}} \) by definition.

- **If \( \alpha > \beta + p \), then similarly we have**

\[ \left\| \hat{A} - A_\lambda \right\|^2_{\beta', \gamma'} \lesssim \frac{1}{N} \max_{1 \leq j \leq y_N} j^{-\frac{\gamma' - \gamma}{q}} \lambda_j^{-\beta' + \alpha - \beta} \]

\[ \lesssim \frac{1}{N} \max_{1 \leq j \leq y_N} j^{-\frac{\gamma' - \gamma}{q}} \left( j^{-\frac{\gamma' - \gamma}{q}} N \min\left\{ \frac{\beta - \beta'}{\alpha}, \frac{\gamma' - \gamma}{1 - \gamma' \gamma} \right\} \right)^{-\frac{\beta' + \alpha - \beta}{2(1 - \gamma' \gamma)}} \]

\[ = \frac{1}{N} \max_{j \in \{1, y_N\}} \left( j^{-\frac{\gamma' - \gamma}{q}} \frac{\beta' + \alpha - \beta}{\alpha} \min\left\{ \frac{\beta - \beta'}{\alpha}, \frac{\gamma' - \gamma}{1 - \gamma' \gamma} \right\} \right) \]

\[ \leq N^{-\min\left\{ \frac{\beta - \beta'}{\alpha}, \frac{\gamma' - \gamma}{1 - \gamma' \gamma} \right\}}. \]
Hence we deduce that
\[ \| \hat{A} - A_0 \|_{H_{L}^{2-\gamma'}}^2 \lesssim N^{-\min\left\{ \frac{\beta - \beta'}{1 - \gamma'}, \frac{\gamma' - \gamma}{1 - \gamma} \right\}}, \]
as desired.

### B.3 Implication of the Upper Bound

In this section, we discuss the implications of our upper bounds under the \((\beta', \gamma')\)-norm.

Note that \( \| C_{Q_K}^{-1, \beta'} v \|_{H_{L}^{2-\gamma'}} = \| v \|_{H_{L}^{2-\gamma'}} \) for all \( v \in L_2(Q_K) \) (if one side of the equation is +\( \infty \) then so is the other), we have that
\[
\mathbb{E}_{u \sim P_K} \left( \| \hat{A} - A_0 \|_{H_{L}^{2-\gamma'}}^2 \right) = \mathbb{E}_{u \sim P_K} \left( \| C_{Q_K}^{-1, \beta'} (\hat{A} - A_0) u \|_{H_{L}^{2-\gamma'}}^2 \right)
= \text{tr} \left( C_{Q_K}^{-1, \beta'} (\hat{A} - A_0) \mathbb{E}_{u \sim P_K} u \otimes u \left( C_{Q_K}^{-1, \beta'} (\hat{A} - A_0) \right) \right)^2
\lesssim \| \hat{A} - A_0 \|_{\beta', \gamma'}^2,
\]
where the last step follows from \( \mathbb{E}_{u \sim P_K} u \otimes u = C_{P_K} \). Note that the above derivations hold for any \( 0 \leq \beta' < \beta \), so choosing \( \beta' = 0 \) yields the best upper bound. We can see from (20) that our analysis implies an upper bound of the expected error of the learned solution evaluated under the \( H_{L}^{2-\gamma'} \)-norm. On the other hand, it is also possible to obtain a uniform convergence rate when \( \beta' \geq \alpha \):
\[
\| (\hat{A} - A_0) u \|_{H_{L}^{2-\gamma'}} = \| C_{Q_K}^{-1, \beta'} (\hat{A} - A_0) u \|_{H_{L}^{2-\gamma'}} \lesssim \| \hat{A} - A_0 \|_{\beta', \gamma'} \cdot \| C_{P_K}^{-1, \beta'} u \|_{H_{L}^{2-\gamma'}} \lesssim \| \hat{A} - A_0 \|_{\beta', \gamma'}.
\]

### C Proofs for the multi-level operator learning algorithm

In this section, we analyze the convergence rate of our multi-level algorithm described in Section 5. We define \( \eta_1 = \min \left\{ \frac{\beta - \beta'}{\max\{\alpha, \beta + p\} \gamma' - \gamma}, \frac{\gamma' - \gamma}{1 - \gamma} \right\} \) and \( \eta_2 = \max \left\{ 1 - \frac{\beta - \beta'}{\max\{\alpha, \beta + p\} \gamma' - \gamma}, 1 - \gamma, 1 - \gamma \right\} = 1 - \eta_1 \). We first restrict ourselves to the case when \( \frac{\beta - \beta'}{\max \{\alpha, \beta + p\} \gamma' - \gamma} \neq \frac{\gamma' - \gamma}{1 - \gamma}; \) the special case when the two terms are equal will be separately treated in Appendix C.1. For the optimal bias and variance contours \( \ell_{C_{1, \text{bias}}} \) and \( \ell_{C_{2, \text{var}}} \) with \( C_1 = N^{\eta_1} \) and \( C_2 = N^{\eta_2} \), we define a sequence \( \{x_n\} \) as follows:
\[
\begin{align*}
x_0 &= \max \left\{ \frac{1}{2} N^{\frac{\beta'}{\beta - \beta'}} n, c_0 \left( \frac{N}{\log N} \right)^{-\frac{\beta - \beta'}{2}} \right\}, \\
y_n &= \text{the solution of } x_n \frac{\beta' + \max\{\alpha, \beta + p\} \gamma' - \gamma}{\beta - \beta'} y_n^{-\gamma'} = N^{\eta_2}, \quad n \geq 0 \\
x_{n+1} &= \text{the solution of } x_n \frac{\beta - \beta'}{p} y_n^{\gamma' - \gamma} = N^{\eta_1}, \quad n \geq 0.
\end{align*}
\]
We first derive an explicit recursive formula for \( \{x_n\} \).

**Lemma C.1** Let \( u = \frac{\beta' + \max\{\alpha, \beta + p\} \gamma' - \gamma}{\beta - \beta'} \frac{\beta - \beta'}{\beta - \beta'} > 0 \), then
\[
\begin{align*}
(1). & \text{ if } u > 1, \text{ then } N^{-\frac{\beta'}{\beta - \beta'}} x_{n+1} = \left( N^{-\frac{\beta'}{\beta - \beta'}} x_n \right)^u, \\
(2). & \text{ if } u < 1, \text{ then } x_{n+1} = x_n^u.
\end{align*}
\]
Proof:

(1). Suppose that \( u > 1 \), then we have \( \eta_1 = \frac{\beta - \beta'}{\max\{\alpha, \beta + p\}} \) and \( \eta_2 = 1 - \eta_1 \). It follows from (21b) and (21c) that

\[
x_{n+1} = N \max\{\frac{p}{\alpha, (\beta + p)}\} y_n - \gamma' \frac{\gamma}{\alpha, (\beta + p)}
\]

\[
= N \max\{\frac{p}{\alpha, (\beta + p)}\} \left( N^\frac{\eta_2}{\gamma} x_n - \frac{\beta' - \max\{\alpha, \beta, p\}}{\eta_2 \gamma} \right)
\]

\[
= N \max\{\frac{p}{\alpha, (\beta + p)}\} \left( N^\frac{\eta_2}{\gamma} x_n \right)^u.
\]

(2). Suppose that \( u < 1 \), then we have \( \eta_1 = \frac{\gamma' - \gamma}{\gamma} \) and \( \eta_2 = 1 - \eta_1 \), so that \( \frac{\eta_2}{\gamma} = \frac{\gamma' - \gamma}{\gamma} \), and it follows from (21b) and (21c) that \( x_n \to x_{n+1} \), thus \( x_{n+1} = x_n^u \).}

Lemma C.1 implies that when \( u \neq 1 \), the sequence \( \{x_n\} \) decreases super-exponentially. Thus, there exists \( L_N = O(\log \log N) \) such that \( x_n \leq 2 \) for all \( n \geq L_N \).

Let \( \lambda_i^{(K)} = x_i^{-\frac{1}{2}} \) and \( \lambda_i^{(L)} = y_i^{-\frac{1}{2}} \), then we construct the following estimator:

\[
\hat{A}_{m1} = \sum_{i=0}^{L_N} \left( \sum_{y_{i-1} \leq j < y_i} \rho_j^2 f_j \otimes \rho_j^2 f_j \right) \hat{C}_{XY} \left( \hat{C}_{KK} + \lambda_i^{(K)} \right)^{-1}
\]  

where \( y_{-1} := 0 \). Note that each summand in the above equation is essentially a regularized least-squares estimator and learns a rectangular region. The following theorem states that the estimator \( \hat{A}_{m1} \) can achieve minimax optimal convergence rate.

Theorem C.1 Consider the estimator \( \hat{A}_{m1} \) defined by (5). Suppose that Assumptions 2.1 to 2.5 hold, then there exists a universal constant \( C \), such that

\[
\left\| \hat{A}_{m1} - A_0 \right\|_{\beta', \gamma'}^2 \leq C \left( \frac{N}{\log N} \right)^{-\min\left\{ \frac{\beta - \beta'}{\max\{\alpha, (\beta + p)\}}, \frac{\gamma' - \gamma}{1 - \gamma} \right\}} \log^2 \frac{1}{N}
\]

holds with probability \( \geq 1 - e^{-\tau} \).

Proof: The proof of Theorem 5.1 is similar to that of Theorems 4.1 and 4.2. We consider the bias-variance decomposition of the estimation error

\[
\left\| \hat{A}_{m1} - A_0 \right\|_{\beta', \gamma'} \leq \left\| \hat{A}_{m1} - \hat{A}_{m1}^\lambda \right\|_{\beta', \gamma'} + \left\| \hat{A}_{m1}^\lambda - A_0 \right\|_{\beta', \gamma'}
\]

where

\[
\hat{A}_{m1}^\lambda = \sum_{i=0}^{L_N} \left( \sum_{y_{i-1} \leq j < y_i} \rho_j^2 f_j \otimes \rho_j^2 f_j \right) \hat{C}_{XY} \left( \hat{C}_{KK} + \lambda_i^{(K)} \right)^{-1}.
\]

Bounding the bias term. Since \( \|A_0\|_{\beta, \gamma} \leq B \), we can write

\[
A_0 := \sum_{i=1}^{+\infty} \sum_{j=1}^{+\infty} a_{ij} \mu_j^\beta \rho_j^1 2 f_j \otimes e_i
\]

where the coefficient matrix \( A_0 = (a_{ij})_{i,j} \subseteq +\infty \) satisfies \( \|A_0\|_F^2 \leq B^2 \). We fix \( (i,j) \in \mathbb{Z}_+^2 \) and assume WLOG that \( y_{m-1} \leq j < y_{m+1} \) for some \( m \geq 0 \), where \( y_{L_N+1} = +\infty \). It follows from (23) that

\[
\left\langle \rho_j^2 f_j, \hat{A}_{m1}^\lambda \right\rangle_{\mu_i^\beta} = \sum_{k=0}^{L_N} \left( \sum_{y_{k-1} \leq j < y_k} \rho_j^2 f_j \otimes \rho_j^2 f_j \right) \rho_j^2 f_j, \hat{C}_{XY} \left( \hat{C}_{KK} + \lambda_i^{(K)} \right)^{-1} \mu_i^\beta e_i
\]

\[
= \frac{\mu_i^\beta}{\mu_i^\beta + \lambda_m^\beta \mu_i^\beta} \frac{1 - \beta}{1 - \beta} a_{ij}.
\]
Thus

$$
\| A_0 - \hat{A}_m \|_{\beta', \gamma'}^2 = \left\| C_{Q_L}^{-1 - \gamma'} \left( \hat{A}_m - A_0 \right) C_{K K}^{1 - \beta'} \right\|_{HS}^2 
$$

$$
= \sum_{i, j=1}^{+\infty} \left\langle \rho_j f_j, C_{Q_L}^{-1 - \gamma'} \left( \hat{A}_m - A_0 \right) C_{K K}^{1 - \beta'} \mu_i e_i \right\rangle^2 
$$

$$
= \sum_{i, j=1}^{+\infty} \left( \frac{\lambda_{m_j}^{(K)}}{\mu_i + \lambda_{m_j}^{(K)}} \right)^2 \mu_j^{\beta' - \beta} \rho_j^{-\gamma} a_{ij}^2 
$$

$$
\leq \sum_{j=1}^{+\infty} \rho_j^{-\gamma} \left( \sum_{i=1}^{+\infty} a_{ij}^2 \right) \max_{i, j=1} \mu_i^{\beta' - \beta} \left( \frac{\lambda_{m_j}^{(K)}}{\mu_i + \lambda_{m_j}^{(K)}} \right)^2 
$$

$$
\leq B^2 \max_{j \geq 1} \rho_j^{-\gamma} \left( \lambda_{m_j}^{(K)} \right)^{\beta' - \beta'} \lesssim B^2 \max_{j \geq 1} j^{-\gamma' - \gamma} x_{m_j}^{-\beta' - \beta'} 
$$

$$
\leq B^2 \gamma_{m_j}^{-\gamma} x_{m_j}^{-\beta' - \beta'} \lesssim N^{-\eta_1} 
$$

where we recall that $\eta_1 = \min \left\{ \frac{\beta' - \beta}{\max\{n, \beta + p\}}, \frac{\gamma' - \gamma}{\gamma - \gamma} \right\}$ and the last step follows from (21c).

**Bounding the variance term.** The variance term can be rewritten in the following way:

$$
V = \left\| \hat{A}_m - A_\lambda \right\|_{\beta', \gamma'}^2 
$$

$$
= \left\| C_{Q_L}^{-1 - \gamma'} \left( \hat{A}_m - A_\lambda \right) C_{K K}^{1 - \beta'} \right\|_{HS}^2 
$$

$$
= \sum_{i, j=1}^{+\infty} \left\langle \rho_j f_j, C_{Q_L}^{-1 - \gamma'} \left( \hat{A}_m - A_\lambda \right) C_{K K}^{1 - \beta'} \mu_i e_i \right\rangle^2 
$$

$$
= \sum_{j=1}^{+\infty} \rho_j^{-\gamma} \left( \sum_{i=1}^{+\infty} a_{ij}^2 \right) \max_{i, j=1} \mu_i^{\beta' - \beta} \left( \frac{\lambda_{m_j}^{(K)}}{\mu_i + \lambda_{m_j}^{(K)}} \right)^2 
$$

$$
\leq \sum_{j=1}^{+\infty} \rho_j^{-\gamma} \left( \sum_{i=1}^{+\infty} a_{ij}^2 \right) \max_{i, j=1} \mu_i^{\beta' - \beta} \left( \frac{\lambda_{m_j}^{(K)}}{\mu_i + \lambda_{m_j}^{(K)}} \right)^2 
$$

$$
\leq B^2 \max_{j \geq 1} \rho_j^{-\gamma} \left( \lambda_{m_j}^{(K)} \right)^{\beta' - \beta'} \lesssim B^2 \max_{j \geq 1} j^{-\gamma' - \gamma} x_{m_j}^{-\beta' - \beta'} 
$$

$$
\leq B^2 \gamma_{m_j}^{-\gamma} x_{m_j}^{-\beta' - \beta'} \lesssim N^{-\eta_1} 
$$

for reasons similar to (12). It now remains to bound $\| G_{m_j} \|$ and $\| U_{m_j} \lambda_j^{-\gamma} \|$ for $1 \leq j \leq L_N$. Note that these quantities have already been bounded in [Appendix B.1.2](#) with $\lambda_{m_j}$ replaced with $\lambda_j$ (there we use a different regularization for each $j$). Hence, those bounds can be directly applied here, so there exists a constant $C > 0$ such that

$$
V \lesssim C a^2 \frac{1}{N} \max_{1 \leq j \leq L_N} j^{-\gamma} \lambda_{m_j}^{-\beta + \max\{n, \beta, p\}} 
$$

$$
\lesssim C a^2 \frac{1}{N} \max_{1 \leq j \leq L_N} j^{-\gamma} \lambda_{m_j}^{-\beta + \max\{n, \beta, p\}} 
$$

$$
\lesssim C a^2 \frac{1}{N} \max_{1 \leq j \leq L_N} j^{-\gamma} \lambda_{m_j}^{-\beta + \max\{n, \beta, p\}} 
$$

26
with probability \( \geq 1 - N e^{-\alpha} \). Since \( j \leq y_{m_j} \), by (21b) we have
\[
\frac{1 - \gamma'}{\gamma} \lambda_{m_j} \leq \frac{1 - \gamma'}{\gamma} y_{m_j} = N^\eta_2.
\]
Hence
\[
\mathcal{V} \leq \frac{1}{N} \max_{1 \leq j \leq L_N} \frac{1 - \gamma'}{\gamma} \lambda_{m_j} \leq N^\eta_2 = N^\eta.
\]
Combining the bias and variance bounds, the conclusion directly follows. \( \square \)

### C.1 Special Case:

\[
\frac{\beta - \beta'}{\max(\alpha, \beta + p)} = \frac{\gamma' - \gamma}{1 - \gamma}
\]

Note that Lemma C.1 does not cover the case \( u = 1 \), or equivalently \( \frac{\beta - \beta'}{\max(\alpha, \beta + p)} \neq \frac{\gamma' - \gamma}{1 - \gamma} \). This case is special since the bias contour coincides with the variance contour, and we need to modify our construction of the multilevel estimator.

We define two sequences \( \{x_n\}, \{y_n\} \) as follows:
\[
x_0 = \max \left\{ \frac{1}{2} \left( \frac{n}{m^2} + \eta_2 \right), c_0 \left( \frac{1}{\log N} \right)^{\frac{1}{\beta_2}} \right\} ;
\]
\[
x_n = \frac{1}{2} x_{n-1}, \quad y_n = \frac{1}{2} y_{n-1} = \frac{1}{2} y_{n-1}
\]
where we recall that \( \eta_1 = \frac{\beta - \beta'}{\max(\alpha, \beta + p)} = \frac{\gamma' - \gamma}{1 - \gamma} \). In this case, there exists \( L_N = \mathcal{O}(\log N) \) such that \( x_n < 1 \) for all \( n \geq L_N \). Let \( \lambda^{(K)}_i = x_i^{\frac{1}{\beta_2}} \), then we construct the following estimator:
\[
\hat{A}_m = \sum_{i=0}^{L_N} \left( \sum_{y_{i-1} < y_i} \rho_{ij} \hat{f}_j \otimes \hat{f}_j \right) \hat{C}_{IK} \left( \hat{C}_{IK} + \lambda^{(K)}_i I \right)^{-1}.
\]

Similar to Theorem C.1, we can establish the following result:

**Theorem C.2** Consider the estimator \( \hat{A}_m \) defined by (26). Suppose that Assumptions 2.1 to 2.5 hold, then there exists a universal constant \( C \), such that
\[
\| \hat{A}_m - A_0 \|_{\beta', \gamma'}^2 \leq C \tau^2 \left( \frac{N}{\log N} \right)^{-\min \left\{ \frac{\beta - \beta'}{\max(\alpha, \beta + p)} : \frac{\gamma' - \gamma}{1 - \gamma} \right\}} \log^2 N
\]
holds with probability \( \geq 1 - e^{-\tau} \).

**Proof**: The proof of Theorem C.2 is similar to that of Theorems 4.1 and 4.2. We consider the bias-variance decomposition
\[
\| \hat{A}_m - A_0 \|_{\beta', \gamma'} \leq \| \hat{A}_{m_1} - \hat{A}_{m_1} \|_{\beta', \gamma'} + \| \hat{A}_{m_1} - A_0 \|_{\beta', \gamma'}
\]
where
\[
\hat{A}_{m_1} = \sum_{i=0}^{L_N} \left( \sum_{y_{i-1} < y_i} \rho_{ij} \hat{f}_j \otimes \hat{f}_j \right) \hat{C}_{IK} \left( \hat{C}_{IK} + \lambda^{(K)}_i I \right)^{-1}.
\]
as defined in (26).

**Bounding the bias term.** Let \( A_0 := \sum_{i,j} a_{ij} \mu_i^{\frac{1}{\beta_2}} \rho_{ij}^{1 - \frac{1}{\beta_2}} f_j \otimes e_i \) with coefficient matrix \( A_0 = (a_{ij})_{i,j=1}^{+\infty} \) such that \( \|A_0\|_F \leq B^2 \). We fix \( (i, j) \in \mathbb{Z}_+^2 \) and assume WLOG that \( y_{m_j-1} \leq j < y_{m_j} \) for some \( m_j \geq 0 \), where \( y_{L_N+1} = +\infty \). It follows from (27) that
\[
\left( \rho_{ij}^{\frac{1}{\beta_2}} f_j, \lambda^{(K)}_i \mu_i^{\frac{1}{\beta_2}} e_i \right) = \frac{\mu_i}{\mu_i + \lambda^{(K)}_j} \rho_{ij}^{\frac{1}{\beta_2}} f_j \otimes e_i.
\]
Thus we can proceed as in (24) to deduce that
\[ \|A_0 - A_\lambda\|_{p, \gamma}^2 \leq \max_{j \geq 1} \rho_{j}^{\gamma - \gamma}(\lambda(K))^{\beta - \beta'} \]
\[ \leq \max \left\{ \max_{1 \leq j \leq N} \frac{y_{m_j}^{\gamma - \gamma} - 2 - \beta'}{x_{m_j}^{\gamma - \gamma}}, \frac{y_{L_N}^{\gamma - \gamma}}{x_{m_j}^{\gamma - \gamma}} \right\} \]
\[ \leq \max \left\{ \max_{1 \leq j \leq N} \frac{y_{m_j}^{\gamma - \gamma} - 2 - \beta'}{x_{m_j}^{\gamma - \gamma}}, \frac{y_{L_N}^{\gamma - \gamma}}{x_{m_j}^{\gamma - \gamma}} \right\} \]

The definition (25) implies that
\[ y_{m_j}^{\gamma - \gamma} - 2 - \beta' \leq 2 \frac{\beta - \beta'}{x_{m_j}^{\gamma - \gamma}} \]
\[ \leq 2 \frac{\beta - \beta'}{x_{m_j}^{\gamma - \gamma}} N^{-\eta}. \]

On the other hand, since \( x_{L_N} < 1 \), by (25) implies that \( y_{L_N}^{\gamma - \gamma} \leq N^{-\eta} \). Therefore, for the bias term \( \|A_0 - A_\lambda\|_{p, \gamma}^2 \leq N^{-\eta} \).

**Bounding the variance term.** Repeating the arguments in (25), we can deduce that there exists a constant \( C > 0 \) such that
\[ \mathcal{V} \leq Ca \left\{ \max_{1 \leq j \leq N} \frac{1 - \gamma}{\lambda m_j} \right\} \]
\[ \leq Ca \left\{ \max_{1 \leq j \leq N} \frac{1 - \gamma}{\lambda m_j^{\beta + \max(\alpha - \beta, \rho)}} \right\} \]
\[ \leq C a \left\{ \max_{1 \leq j \leq N} \frac{1 - \gamma}{\lambda m_j^{\beta + \max(\alpha - \beta, \rho)}} \right\} \]
with probability \( \geq 1 - N e^{-a} \).

Combining the bias and variance bounds, we arrive at the desired conclusion. \( \square \)

The conclusion of Theorem 5.1 then follows from Theorems C.1 and C.2.

## D Auxiliary Results

**Lemma D.1** We have \( \|T\|_{p, \gamma} = \left\| C_{Q_L}^{-(1-\gamma)/2} \circ T \circ C_{KK}^{(1-\beta)/2} \right\|_{HS(H_\gamma, H_\gamma)} \).

**Proof:** We recall from the definition that \( \|T\|_{p, \gamma} = \left\| (I_{1, \gamma, Q_L})^\dagger \circ T \circ I_{1, \gamma, Q_L} \right\|_{HS(H_\gamma, H_\gamma)} \), so that
\[ \|T\|_{p, \gamma}^2 = \left\| (I_{1, \gamma, Q_L})^\dagger \circ T \circ I_{1, \gamma, Q_L} \right\|_{HS(H_\gamma, H_\gamma)}^2 \]
\[ = \sum_{i,j=1}^{+\infty} \left\langle \rho_j^2 f_j, (I_{1, \gamma, Q_L})^\dagger \circ T \circ I_{1, \gamma, Q_L} \mu_i^{1-\gamma} e_i \right\rangle_{H_\gamma}^2 \]
\[ = \sum_{i,j=1}^{+\infty} \left\langle \rho_j^2 f_j, (I_{1, \gamma, Q_L})^\dagger \circ T \mu_i^{1-\gamma} e_i \right\rangle_{H_\gamma}^2 \]
\[ = \sum_{i,j=1}^{+\infty} \left\langle \rho_j^2 f_j, T \mu_i^{1-\gamma} e_i \right\rangle_{H_\gamma}^2 \]
\[ = \sum_{i,j=1}^{+\infty} \left\langle \rho_j^2 f_j, C_{Q_L}^{-(1-\gamma)/2} \circ T \circ C_{KK}^{(1-\beta)/2} \mu_i^{1-\gamma} e_i \right\rangle_{H_\gamma}^2 \]
\[ = \left\| C_{Q_L}^{-(1-\gamma)/2} \circ T \circ C_{KK}^{(1-\beta)/2} \right\|_{HS}^2 \]
as desired. \( \square \)

**Lemma D.2** Under Assumption 2.2, we have
\[ \left\| (C_{KK} + \lambda I)^{-{\gamma/2}} u \right\| \leq \lambda^{-{\gamma/2}} A_1 \quad P_K \text{-a.s.} \]
Proof: By Assumption 2.2 we have $\|C_{KK}^{\frac{1}{2}n} u\|_{\mathcal{H}_K} \leq A_1$, so that
\[ \left\| (C_{KK} + \lambda I)^{-\frac{1}{2}} u \right\| \leq \left\| (C_{KK} + \lambda I)^{-\frac{1}{2}} \right\| \cdot \left\| C_{KK}^{\frac{1}{2}n} u \right\| \leq \lambda^{-\frac{n}{2}} \cdot A_1 \]
as desired.

\[ \square \]

D.1 CONCENTRATION INEQUALITIES

Theorem D.1 (Fischer & Steinwart, 2020, Theorem 27) Let $(\Omega, \mathcal{B}, P)$ be a probability space, $\mathcal{H}$ be a separable Hilbert space and $X : \Omega \to \text{HS}(H; H)$ be a random variable with self-adjoint values. Furthermore, assume that $\|X\|_F \leq B$, $P$-a.s. and $V$ be a positive semi-definite matrix with $E_P(X^2) \preceq V$, i.e. $V - E_P(X^2)$ is positive semi-definite. Then, for $g(V) := \log (2e \text{tr}(V)\|V\|^{-1})$, $\tau \geq 1$, and $n \geq 1$, the following concentration inequality is satisfied
\[ P^n \left( (\omega_1, \ldots, \omega_n) \in \Omega^n : \left\| \frac{1}{n} \sum_{i=1}^n X(\omega_i) - E_P X(\omega) \right\| \geq \frac{A \tau B g(V)}{n} + \sqrt{\frac{2\tau \|V\| g(V)}{n}} \right) \leq 2 e^{-\tau}. \]

Theorem D.2 (Fischer & Steinwart, 2020, Theorem 26) Let $(\Omega, \mathcal{B}, P)$ be a probability space, $H$ be a separable Hilbert space, and $\xi : \Omega \to H$ be a random variable with $E_P \|\xi\|_H^n \leq \frac{1}{2} m! \sigma^2 L^{m-2}$ for all $m \geq 2$. Then, for $\tau \geq 1$ and $n \geq 1$, the following concentration inequality is satisfied
\[ P^n \left( (\omega_1, \ldots, \omega_n) \in \Omega^n : \left\| \frac{1}{n} \sum_{i=1}^n \xi(\omega_i) - E_P \xi \right\|_H^2 \geq 32 \frac{\tau^2}{n} \left( \sigma^2 + \frac{L^2}{n} \right) \right) \leq 2 e^{-\tau}. \]

The following theorem shows that the regularized covariance $C_{KK} + \lambda I$ can be estimated with small error when $\lambda$ is above a certain threshold. Although it is well-known (Fischer & Steinwart, 2020; Talwai et al., 2022), we still recall it below for completeness.

Theorem D.3 Recall that $\hat{C}_{KK} = E_{P_K} u \otimes u$ and $\tilde{C}_{KK} = \frac{1}{N} \sum_{i=1}^N u_i \otimes u_i$ where $u_i \sim \text{i.i.d.} P_K$. Suppose that Assumption 2.2 holds and $N \geq A^2 g_\lambda \lambda^{-\alpha}$, where $g_\lambda = \log (2e N_{P_K} (\lambda) \|C_{KK}^{1/2} + \lambda^2\|_{\mathcal{H}_K}^{1/2})$ and $N_{P_K} (\lambda) = \text{tr} \left( (C_{KK} + \lambda I)^{-1} C_{KK} \right)$ is the effective dimension, then with probability at least $1 - e^{-\tau}$, we have
\[ \left\| (C_{KK} + \lambda I)^{-\frac{1}{2}} (C_{KK} - \hat{C}_{KK}) (C_{KK} + \lambda I)^{-\frac{1}{2}} \right\| \leq \sqrt{\frac{A^2 g_\lambda}{N \lambda^\alpha}} \leq 0.1. \] (28)

Proof: Let $X(u) = (C_{KK} + \lambda I)^{-\frac{1}{2}} u \otimes u (C_{KK} + \lambda I)^{-\frac{1}{2}}$ where $u \in \mathcal{H}_K$, then the LHS of (28) can be expressed as $\left\| \frac{1}{n} \sum_{i=1}^N X(u_i) - E_{u \sim P_K} X(u) \right\|_F$. We hope to apply Theorem D.1 and start with verifying the assumptions.

Since $E_{P_K} X = (C_{KK} + \lambda I)^{-\frac{1}{2}} C_{KK} (C_{KK} + \lambda I)^{-\frac{1}{2}}$ and $\|X(u)\| = \|X(u)\|_F = \left\| (C_{KK} + \lambda I)^{-\frac{1}{2}} u \right\|^2 \leq A^2 \left\| (C_{KK} + \lambda I)^{-\frac{1}{2}} \right\|^2 \leq A^2 \lambda^{-\alpha},$ so that there exists $V = O \left( \lambda^{-\alpha} (C_{KK} + \lambda I)^{-\frac{1}{2}} C_{KK} (C_{KK} + \lambda I)^{-\frac{1}{2}} \right)$ such that $E_{P_K} X^2 \preceq V$. It’s easy to see that $\|V\| \preceq \lambda^{-\alpha}$ and $\text{tr}(V) \preceq N_{P_K} (\lambda)$. The conclusion then follows from Theorem D.1 with $B = O(\lambda^{-\alpha})$ and $g(V) = g_\lambda$. \[ \square \]

Corollary D.1 Under the notations and assumptions of Theorem D.3, there exists a constant $C_1 > 0$ with probability $\geq 1 - e^{-\tau}$ we have
\[ \left\| (C_{KK} + \lambda I)^{\frac{1}{2}} (\hat{C}_{KK} + \lambda I)^{-1} (C_{KK} + \lambda I)^{\frac{1}{2}} \right\| \leq C_1. \] (29)
**Proof:** By Theorem D.3 we have

\[
\left\| (C_{KK} + \lambda I)^{\frac{1}{2}} \left( \hat{C}_{KK} - C_{KK} \right) (C_{KK} + \lambda I)^{\frac{1}{2}} \right\| \\
= \left\| \left( I - (C_{KK} + \lambda I)^{-\frac{1}{2}} \left( C_{KK} + \lambda I \right)^{-\frac{1}{2}} \right)^{-1} \right\| \\
\leq 2
\]

with probability \( \geq 1 - e^{-\lambda} \), as desired. \( \square \)