Minimax Optimal Kernel Operator Learning via Multilevel Training

Jikai Jin
School of Mathematical Sciences
Peking University
Beijing, China, 100871
jkjin@pku.edu.cn

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

Learning mappings between infinite dimensional function spaces has 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. 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 that above the variance contour can achieve optimal learning rate. At the same time, the spectral components between the bias and variance contours give us the 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

The supervised learning of operators between two infinite-dimensional spaces has attracted attention in many machine learning applications, such as scientific computation [1, 2, 3, 4, 5], functional data analysis [6, 7, 8], learning mean-field games [9, 10], conditional probability regression [11, 12, 13] and econometrics [14, 15, 16]. Despite the empirical success of operator learning, the statistical limit of learning an infinite-dimensional operator is poorly studied. In this paper, we study the problem of learning Hilbert Schmidt operators between infinite-dimensional Sobolev reproducing kernel Hilbert spaces $\mathcal{H}_K^\beta$ and $\mathcal{H}_L^\gamma$ with given kernels k and l respectively and $\beta, \gamma \in [0, 1)$ [17, 18, 19]. Our goal is to derive the optimal sample complexity to learn the linear operator, *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 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 the sense that all existing results [19, 20, 3] only establish rates that depend on the parameter

NeurIPS 2021 AI for Science Workshop.

of input space. The reason is all previous works [21, 20, 3] 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 detail 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 contour and the variance contour at the scale of optimal learning. Once the regularization enables one to learn all the spectral part above the bias contour and below the variance contour, the learning is optimal. Finally, utilizing the region in between the bias contour and variance contour , we developed a multilevel training algorithm [22, 23] 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 [24, 25]: we use the next level to reduce bias while keeping the variance at the same scale.We demonstrate that such multilevel algorithm can achieve optimal non-parametric rate for linear operator learning.

1.1 Related Work

Machine Learning Based PDE Solver Solving PDEs plays a prominent role in many scientific and engineering discipline, such as physics, chemistry, operation management, macro-economy, etc. The recent deep learning breakthrough has drawn attention to solving PDEs via machine learning methods [26, 27, 28, 29, 30, 31]. The statistical power and computational cost of these problem is well-studied by recent papers [32, 33, 34, 35]. This paper focuses on operator learning [36, 37, 38, 39, 40, 1, 2, 41, 42], *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. In terms of the mathematical foundation, [43] considers the learning rate of non-parametric operator learning. However, non-parametric functional data analysis often suffers from slower-than-polynomial convergence rates [44], due to the small ball probability problem for the probability distributions in infinite dimensional spaces [45]. The most relevant works are [46, 47, 3], which consider the rates for learning a linear operator. For the comparison between our work and [3], see Remark 2.1.

multilevel Monte Carlo By combining biased estimators with multiple stepsizes, multilevel Monte Carlo (MLMC) [24, 25] dramatically improves the rate of convergence and achieves in many settings the canonical square root convergence rate associated with unbiased Monte Carlo [48, 49]. Multilevel Monte Carlo can also be used for random variable with infinite variance [50, 51]. 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, [22, 23] developed a multilevel machine learning Monte Carlo algorithm (ML2MC) / multilevel fine-tuning algorithm for learning solution maps, by first learning the map on 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 can reduce the generalization error without spending more time on generating training samples. [52, 53] consider such multi-scale algorithm for learning Green's function. However, the statistical power of such algorithm is still under investigation. Another difference with [53] is that we consider the Green function in H^{-1} norm rather than the ℓ_1 norm used in [53]. 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 [20], while the second lower bound 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 parts above the variance contour at the level of learning rate. This enables the estimator to enjoy an optimal balance of bias-variance.

- We qualify a specific setting where a multilevel training procedure [22, 23] is necessary and capable of achieving a minimax optimal learning rate for learning a linear operator. We achieve the optimal learning rate via $O(\ln \ln n)$ ensemble of ridge regression models. This is different 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 $\mathcal{H}_K$ and define covariance operator $\mathcal{C}_{KK} = \mathbb{E}_{u \sim P_K} u \otimes u$. Consider its spectral decomposition $\mathcal{C}_{KK} = \sum_{i=1}^{+\infty} \mu_i^2 e_i \otimes e_i$, where $\{\mu_i^{\frac{1}{2}} e_i\}_{i=1}^{+\infty}$ is an orthogonal eigenbasis and $\{\mu_i\}$ is the corresponding eigenvalues of $\mathcal{C}_{KK}$ (here the $g \otimes h$ is an operator defined as $g \otimes h = gh^* : f \to \langle f, h \rangle g$). In the typical machine learning applications, the test distribution is the same as the training distribution, so we can assume that $\mathcal{H}_K = \left\{ \sum_i a_i \mu_i^{\frac{1}{2}} e_i : \{a_i\}_{i=1}^{\infty} \in \ell_2 \right\}$ without loss of generality. Note that this automatically holds in the context of learning the conditional mean embedding (CME) [19, 21, 20].

Following [18, 19], we define the interpolation Sobolev space $\mathcal{H}_K^{\beta} = \left\{ f = \sum_i a_i (\mu_i^{\frac{\beta}{2}} e_i) : \{a_i\}_{i=1}^{\infty} \in l^2 \right\}$ for any $\beta > 0$, equipped with Sobolev norm defined by the inner product $\left\langle \sum_i a_i (\mu_i^{\beta/2} e_i), \sum_i b_i (\mu_i^{\beta/2} e_i) \right\rangle_{\mathcal{H}_K^{\beta}} = \sum_i a_i b_i$. 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 $\mathcal{C}_{Q_L}$ and the Sobolev space $\mathcal{H}_L^{\gamma}$. Natural choices of Q_L include some distribution on kernel functions $\{\ell(y, \cdot) : y \in Y\}$ of $\mathcal{H}_L$ induced by some distribution Q_L on Y, so that $\mathcal{C}_{Q_L}$ is a kernel integral operator with respect to Q_L and $\mathcal{H}_L^{\gamma}$ is an interpolation space between $\mathcal{H}_L$ and $\mathcal{L}^2(Q_L)$; see Example 2.1 for a specific choice of Q_L and its practical implications.

Following [20], 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 **((β, γ)-norm)** *Let $T : \mathcal{H}_K \mapsto \mathcal{H}_L$ be a possibly unbounded linear operator. $I_{1,\beta,P_K} : \mathcal{H}_K \mapsto \mathcal{H}_K^{\beta}, \beta \in (0,1)$ is the canonical embedding mapping that takes $u \in \mathcal{H}_K$ to the same element u in the larger space $\mathcal{H}_K^{\beta}$, and $I_{1,\gamma,Q_L} : \mathcal{H}_L \mapsto \mathcal{H}_L^{\gamma}, \gamma \in (0,1)$ is similarly defined. Then the (β, γ)-norm of T is defined as*

$$\|T\|_{\beta,\gamma} = \left\| (I_{1,\gamma,Q_L}^*)^{\dagger} \circ T \circ I_{1,\beta,P_K}^* \right\|_{\mathrm{HS}(\mathcal{H}_K^{\beta}, \mathcal{H}_L^{\gamma})} = \left\| \mathcal{C}_{Q_L}^{-(1-\gamma)/2} \circ T \circ \mathcal{C}_{KK}^{(1-\beta)/2} \right\|_{\mathrm{HS}(\mathcal{H}_K, \mathcal{H}_L)},$$

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 $\mathcal{A}_0 : \mathcal{H}_K \mapsto \mathcal{H}_L$ between two reproducing kernel Hilbert spaces corresponding to kernl k and l respectively. We are given N noisy data pairs $(u_i, v_i), 1 \leqslant i \leqslant N$ related by $v_i = \mathcal{A}_0 u_i + \varepsilon_i$, where $u_i \overset{\text{i.i.d.}}{\sim} P_K$ for some unknown distribution P_K and ε_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 $\mathcal{C}_{KK} = \mathbb{E}_{u \sim P_K} u \otimes u$, $\mathcal{C}_{KL} = \mathbb{E}_{(u,v) \sim P_{KL}} u \otimes v$ and its adjoint $\mathcal{C}_{LK} = \mathcal{C}_{KL}^*$ be uncentered cross-covariance operators associated with P_{KL}. Then we can reformulate the ground turth operator as $\mathcal{A}_0 = \mathcal{C}_{LK} \mathcal{C}_{KK}^{\dagger}$, where $\dagger$ is the pseudo-inverse [21, 20]. With the goal of understanding the relative difficulty of learning different types of linear operators, we investigate the sample efficiency of learning $\mathcal{A}_0$ under certain source assumptions imposed on the data model. Source condition [54, 55, 56, 57, 19] 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 (β, γ)-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 (β', γ')-norm as in [19, 33, 21, 20].

Remark 2.1 *Although recent works have considered similar problems in the context of conditional mean embedding [21, 20] and functional data analysis [3]. In all these papers, the output space is a trace bounded reproducing kernel Hilbert space [3, 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 [19, 21, 20].

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

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

Assumption 2.3 (ℓ_∞ Embedding Property of the Output RKHS) *There exists $A_2 < +\infty$ such that $\|g\|_{\mathcal{H}_L} \leqslant A_2$ holds for all g in the range of $\mathcal{A}_0$, 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 $\operatorname{tr}(V)\leqslant\sigma^2$ such that for every $u\in\mathcal{H}_K$, We have $\mathbb{E}_{v\sim P_{KL}(\cdot|u)}\left[((v-\mathcal{A}_0u)\otimes(v-\mathcal{A}_0u))^k\right]\preceq\frac{1}{2}(2k)!R^{2k-2}V.$ holds for all $k\geq 2$.*

Assumption 2.5 (Source Condition) *$\mathcal{A}_0$ is bounded under (β,γ)-norm i.e. $\|\mathcal{A}_0\|_{\beta,\gamma}\leqslant B$ for some $B\in(0,+\infty)$.*

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 [58] from observations or predicting the future [37, 1, 2, 39, 59]. The second example is about learning conditional mean embedding [11, 12, 13], which represents a conditional distribution as an RKHS element. Thus conditional distribution regression can be reduced to a kernel operator learning. Our theory can also be used for linear inverse problem such as radial electrical impedance tomography (EIT) [60] and the severely ill-posed inverse boundary problem for the Helmholtz equation with unknown wave-number parameter [61]. For detailed discussion, we refer to [3, Section 1.3]

Example 2.1 (Learning differential operators) *Suppose that the ground-truth operator $\mathcal{A}_0=\Delta^t$ where Δ is the Laplacian and $t\in\mathbb{Z}$. Let $\mathcal{H}_K=\mathcal{H}^{m+2t}([0,1])$ be the Sobolev space with smoothness $m+2t$ on $[0,1]$ and $\mathcal{H}_L=\mathcal{H}^m([0,1])$, then $\mathcal{A}_0$ 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 $C_{KK}=\sigma^2\left(-\Delta+\tau^2 I\right)^{-s}$. Its eigenvalues satisfy $\mu_n\propto n^{-2s}$. On the other hand, we choose Q_Y to be a distribution supported on $\{\ell(y,\cdot):y\in[0,1]\}$ induced by a uniform distribution on $[0,1]$, where ℓ is the kernel function of $\mathcal{H}_L$. Then $\mathcal{C}_{Q_Y}$ is essentially the kernel integral operator on $\mathcal{H}_L$ w.r.t. the uniform distribution, and its eigenvalues are $\rho_n\propto n^{-2m}$. The assumption $\|\mathcal{A}_0\|_{\beta,\gamma}<+\infty$ is satisfied if and only if $(1-\gamma)m<(1-\beta)s-\frac{1}{2}\Rightarrow\gamma>1-\frac{2(1-\beta)s-1}{2m}$.

Example 2.2 (Conditional mean embedding) *Suppose that we would like to learn the conditional distribution $P(y\mid x)$ from a data set $\{(x_i,y_i):1\leqslant i\leqslant N\}\subset X\times Y$ 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 $C_{Y|X}$ that satisfies*

$$C_{Y|X}k(x,\cdot)=\mathbb{E}_{Y|x}\ell(Y,\cdot)=:\mu_{Y|x},\ \text{and}\ \mathbb{E}_{Y|x}g(Y)=\langle g,\mu_{Y|x}\rangle\,\forall x\in X.$$

We choose $\mathcal{A}_0=C_{Y|x}$. In this case, $\mathcal{C}_{KK}=\mathbb{E}_{P_K}k(X,\cdot)\otimes k(X,\cdot)$. Assumption 2.2 states that $\sup_{x\in X}k^\alpha(x,x)=A_1$, while Assumption 2.3 is equivalent to $\sup_{x\in X}\left\|\mu_{Y|x}\right\|\leqslant A_2$ (for simplicity we only focus on the case $\zeta=1$). According to Assumption 2.5, we assume that $\left\|C_{Y|X}\right\|_{\beta,\gamma}\leqslant B$.

The mis-specified setting where $\beta < 1$ has been studied in previous work [19, 21, 20]. However, they only consider the case $\gamma = 1$. Our results also cover the case $\gamma < 1$, which allows us to obtain theoretical guarantee for computing conditional expectation of the larger function class $\mathcal{H}_L^{\gamma}$.

3 Information Theoretic Lower Bound

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

Theorem 3.1 *Suppose that $\mathcal{H}_K$ and $\mathcal{H}_L$ are two Hilbert spaces, P_K and Q_L are probability distributions on $\mathcal{H}_K$ and $\mathcal{H}_L$ respectively such that Assumptions 2.1 and 2.2 hold. Then for any estimator $\mathcal{L} : (\mathcal{H}_K \times \mathcal{H}_L)^{\otimes N} \mapsto \mathrm{HS}\left(\mathcal{H}_K^{\beta}, \mathcal{H}_L^{\gamma}\right)$, there exists a linear operator $\mathcal{A}_0$ and a joint data distribution P_{XY} with marginal distribution P_K on $\mathcal{H}_K$ satisfying Assumptions 2.3 to 2.5, such that with probability $\geqslant 0.99$ over $(u_i, v_i) \overset{\text{i.i.d.}}{\sim} P_{XY}$ we have*

$$\left\|\mathcal{L}\left(\{(u_i, v_i)\}_{i=1}^N\right) - \mathcal{A}_0\right\|_{\beta',\gamma'}^2 \gtrsim N^{-\min\left\{\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{\gamma'-\gamma}{(1-\gamma)}\right\}}.$$

Remark 3.1 *Our lower bound is composed of a minimum of two parts. The first rate $N^{-\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}}$ is the minimax optimal Sobolev learning rate for kernel regression [19, 21, 20, 33] and is fully determined by the parameter of the input Sobolev reproducing kernel Hilbert space. Our second rate $N^{-\frac{\gamma'-\gamma}{1-\gamma}}$ is novel to the literature. This bound shows that how the infinite dimensional problem is different from finite dimensional regression problem and is fully determined by the parameter of the output Sobolev reproducing kernel Hilbert space. 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 structure in Remark 4.1 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 $\hat{\mathcal{A}}$ based on N i.i.d. data $\{(u_i, v_i)\}_{i=1}^n \sim P_{KL}^{\otimes n}$ for $1 \leqslant i \leqslant N$ enjoys an optimal learning rate.

Compared with existing approaches where a regularized least-squares estimator can achieve statistical optimality [19, 21, 20, 3] under $(\beta, 1)$-norm, we study the learning rate under the (β, γ)-norm ($\beta' \in (0, \beta), \gamma' \in (\gamma, 1)$) which is defined in Definition 2.1 as $\left\|\hat{\mathcal{A}} - \mathcal{A}_0\right\|_{\beta',\gamma'} = \left\|\mathcal{C}_{Q_L}^{-\frac{1-\gamma'}{2}}\left(\hat{\mathcal{A}} - \mathcal{A}_0\right)\mathcal{C}_{KK}^{\frac{1-\beta'}{2}}\right\|_{\mathrm{HS}}$. The norm of the additional $\mathcal{C}_{Q_L}^{-\frac{1-\gamma'}{2}}$ term is unbounded which make our setting harder than the convergence in $(\beta, 1)$-norm in existing works. Since $\mathcal{C}_{Q_L}^{-\frac{1-\gamma'}{2}}$ is bounded when restricted to the finite-dimensional space $\mathrm{span}\left(\rho_i^{\frac{1}{2}} f_i : 1 \leqslant i \leqslant n\right)$, 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 $\mathcal{A}_0 = \sum_{i,j=1}^{+\infty} a_{ij}\mu_i^{\frac{\beta}{2}} e_i \otimes \rho_j^{1-\frac{\gamma}{2}} f_j$. The problem of estimating $\mathcal{A}$ 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 $\left\|a_{ij}\mu_i^{\frac{\beta}{2}} e_i \otimes \rho_j^{1-\frac{\gamma}{2}} f_j\right\|_{\beta',\gamma'}^2 = a_{ij}^2\mu_i^{\beta-\beta'}\rho_j^{\gamma'-\gamma} \propto i^{-\frac{\beta-\beta'}{p}} j^{-\frac{\gamma'-\gamma}{q}}$ under the (β', γ')-norm. On the other hand, when $\alpha \leqslant \beta + p$, we show in Appendix B.1.2 that the variance of learning (i, j) from noisy data scales as $\frac{1}{N}\mu_i^{-\beta'}\rho_j^{-(1-\gamma')} \propto \frac{1}{N} i^{\frac{\beta'}{p}} j^{\frac{1-\gamma'}{q}}$. Since the variance would accumulate for a

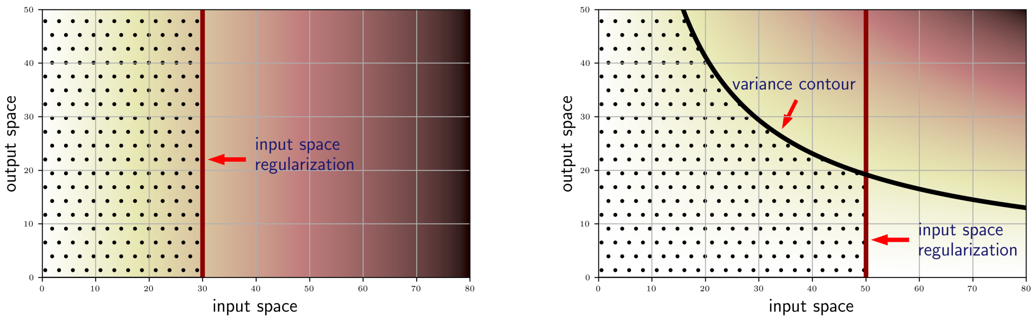

Figure 1: An illustration of our proposed regularization scheme. Left: the regularized least-squares estimator studied in previous works [19, 3, 21] which only regularizes on the input space. Right: our double regularization scheme via variance contour can achieve the optimal convergence rate in our setting.

fixed j, learning (i,j) for $i \leqslant i_{\max}$ results in a variance of $\propto \frac{1}{N} i_{\max}^{\frac{\beta'+p}{p}} j^{\frac{1-\gamma'}{q}}$. (Similar analysis can be carried out for the $\alpha > \beta + p$ case as well, but the variance now scales as $\frac{1}{N} i_{\max}^{\frac{\beta'+\alpha-\beta}{p}} j^{\frac{1-\gamma'}{q}}$; 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. 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 [19, 3, 21, 20]. 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 points). Formally, we fix an arbitrary constant $C > 0$ and define

$$\ell_{C,\texttt{var}} = \left\{ (x,y) \in \mathbb{R}_+^2 : x^{\frac{\beta'+\max\{\alpha-\beta,p\}}{p}} y^{\frac{1-\gamma'}{q}} = C \right\}. \tag{1}$$

A reasonable regularization scheme is then to learn all coordinates $(i,j) \in \mathbb{Z}_+^2$ *below* the curve $\ell_{C,\texttt{var}}$ 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

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

where $\lambda_j (1 \leqslant j \leqslant y_N = C^{\frac{q}{1-\gamma'}})$ are the regularization coefficients imposed on different dimensions of the output space. According to (1) and noting that $\mu_i \propto i^{-\frac{1}{p}}$, we define

$$\lambda_j = \max \left\{ \left(j^{-\frac{1-\gamma'}{q}} N^{\max\left\{1 - \frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{1-\gamma'}{1-\gamma}\right\}} \right)^{-\frac{1}{\beta'+p}}, c_0 \left(\frac{N}{\log N} \right)^{-\frac{1}{\alpha}} \right\}, \tag{3}$$

with $C = N^{\max\left\{1 - \frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{1-\gamma'}{1-\gamma}\right\}}$ in (1). The additional $N^{-\frac{1}{\alpha}}$ term in (3) is needed for controlling the error of approximating $\mathcal{C}_{KK}$ via $\hat{\mathcal{C}}_{KK}$ (cf. Theorem D.3) which is standard in the Sobolev learning

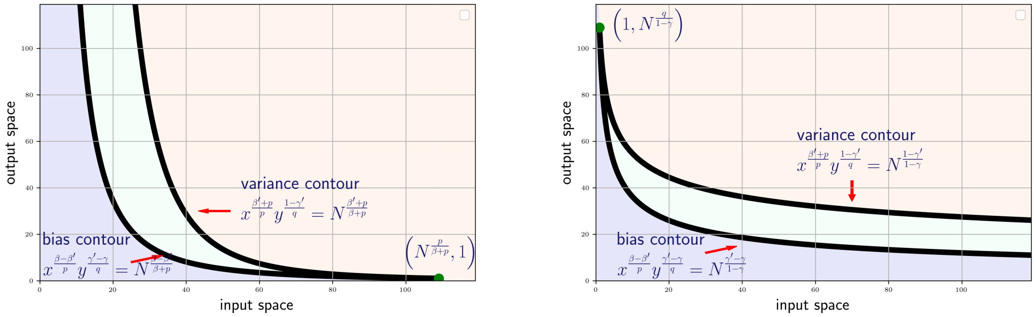

Figure 2: The plot of the bias contour and the variance contour. For simplicity, we only plot the case $\alpha \leqslant \beta + p$ here. The variance contour is always above the bias contour. Left: When $\frac{\beta'+p}{\beta+p} \geqslant \frac{1-\gamma'}{1-\gamma}$, the two yields $\mathcal{O}\left(N^{-\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}}\right)$ convergence rate. It is the same learning rate as the two kernel regression curves meet when $y = 1$. Right: When $\frac{\beta'+p}{\beta+p} \geqslant \frac{1-\gamma'}{1-\gamma}$, the two contours yield the same regularization on the output space leading to a convergence rate of $\mathcal{O}\left(N^{-\frac{\gamma'-\gamma}{1-\gamma}}\right)$

literature [19, 21, 33]. The following theorem describes the convergence rate of our estimator defined by (2) and (3).

Theorem 4.1 *Consider the estimator $\hat{\mathcal{A}}$ defined by (2) and (3). Suppose that Assumptions 2.1 to 2.5 hold, then there exists a universal constant C such that with probability $\geqslant 1 - e^{-\tau}$, we have*

$$\left\|\hat{\mathcal{A}} - \mathcal{A}_0\right\|_{\beta',\gamma'}^2 \leqslant C\tau^2 \left(\frac{N}{\log N}\right)^{-\min\left\{\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{\gamma'-\gamma}{1-\gamma}\right\}} \log^2 N.$$

4.2 Regularization via Bias Contour

We have showed that if we learn all the spectral components under certain variance contour and regularize all other component can achieve optimal rate. In this section, we introduce another scheme to design the optimal estimator via learning all the spectral component under a certain bias contour. Specifically, we consider deciding the regularization strength according to the spectral elements induce a certain level of bias i.e. the *bias contour* $\ell_{C',\texttt{bias}} = \left\{(x,y) \in \mathbb{R}_+^2 : x^{\frac{\beta-\beta'}{p}} y^{\frac{\gamma'-\gamma}{q}} = C'\right\}$. does not coincide with $\ell_{C,\texttt{var}}$ for any C' up to constant scaling. Thus, there exists a point (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 $\lambda_j = \max\left\{\left(j^{-\frac{\gamma'-\gamma}{q}} N^{\min\left\{\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{\gamma'-\gamma}{1-\gamma}\right\}}\right)^{-\frac{1}{\beta-\beta'}}, c_0\left(\frac{N}{\log N}\right)^{-\frac{1}{\alpha}}\right\}$ for similar reasons as Section 4.1, which also yields optimal rate as stated in Theorem 4.2 below.

Remark 4.1 (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.*

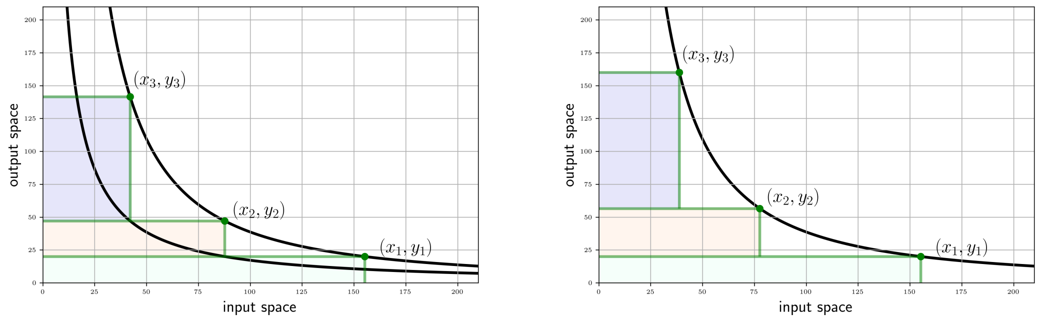

Figure 3: Construction of the sequence $\{(x_i, y_i)\}$. Left: the case $\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}} \neq \frac{\gamma'-\gamma}{1-\gamma}$. Right: the case $\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}} = \frac{\gamma'-\gamma}{1-\gamma}$, where the bias and variance contours overlap and we set $x_{n+1} = \frac{1}{2}x_n$. Each rectangular represents a certain level of regularization.

Theorem 4.2 *Consider the estimator $\hat{A}$ defined by (2) with λ_j defined above. Suppose that Assumptions 2.1 to 2.5 hold, then there exists a universal constant C, such that $\left\|\hat{A} - \mathcal{A}_0\right\|^2_{\beta',\gamma'} \leqslant C\tau^2 \left(\frac{N}{\log N}\right)^{-\min\left\{\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{\gamma'-\gamma}{1-\gamma}\right\}} \log^2 N$ holds with probability $\geqslant 1 - e^{-\tau}$.*

5 MultiLevel Kernel Operator Learning

In this section, we study a multilevel machine learning algorithm [22, 23, 53] but at each level we consider a cost-accuracy trade-off [62] 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 [24, 25], which reduces bias from multilevel algorithm. Our multilevel estimator differs from the DeepONet [1] and the PCA-Net [63] 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 $1 \leqslant i \leqslant L_N$ where y_i denotes the i-th level and x_i controls the corresponding regularization via the regularization coefficient $\lambda_i^{(K)} = x_i^{-\frac{1}{p}}$. The sequences are chosen in a staircase manner as plotted in Figure 3 (for formal definitions see Appendix C). The eigenbasis $\left\{\rho_j^{\frac{1}{2}} f_j\right\}$ of the output space is divided into defferent 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

$$\hat{\mathcal{A}}_{\mathtt{ml}} = \sum_{i=0}^{L_N} \left(\sum_{y_{i-1} \leqslant j < y_i} \rho_j^{\frac{1}{2}} f_j \otimes \rho_j^{\frac{1}{2}} f_j \right) \hat{C}_{LK} \left(\hat{C}_{KK} + \lambda_i^{(K)} I \right)^{-1}. \tag{4}$$

The following theorem shows that the estimator (4) 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 $O(\ln N)$ 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 $\{y_i\}_{1 \leqslant i \leqslant L_N}$ with $L_N = \mathcal{O}(\ln N)$ when $\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}} = \frac{\gamma'-\gamma}{1-\gamma}$ and $\mathcal{O}(\ln \ln N)$ otherwise, such that the estimator $\hat{\mathcal{A}}_{\mathtt{ml}}$ satisfies $\left\|\hat{\mathcal{A}}_{\mathtt{ml}} - \mathcal{A}_0\right\|^2_{\beta',\gamma'} \leqslant C\tau^2 \left(\frac{N}{\log N}\right)^{-\min\left\{\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{\gamma'-\gamma}{1-\gamma}\right\}} \log^2 N$ with probability $\geqslant 1 - e^{-\tau}$, where C is a universal constant.*

References

[1] Lu Lu, Pengzhan Jin, and George Em Karniadakis. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv preprint arXiv:1910.03193*, 2019.

[2] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

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

[4] Zhi-Zhou Li, Yi-Chen Tao, Xue-Dong Wang, and Liang-Sheng Liao. Organic nanophotonics: self-assembled single-crystalline homo-/heterostructures for optical waveguides. *ACS Photonics*, 5(9):3763–3771, 2018.

[5] Zhizhou Li, Chun Yan Gao, Dingkai Su, Chuancheng Jia, and Xuefeng Guo. Principles of molecular machines at the single-molecule scale. *ACS Materials Letters*, 3(10):1484–1502, 2021.

[6] Christophe Crambes and André Mas. Asymptotics of prediction in functional linear regression with functional outputs. *Bernoulli*, pages 2627–2651, 2013.

[7] Siegfried Hörmann and Lukasz Kidziński. A note on estimation in hilbertian linear models. *Scandinavian journal of statistics*, 42(1):43–62, 2015.

[8] Daren Wang, Zifeng Zhao, Yi Yu, and Rebecca Willett. Functional linear regression with mixed predictors. *arXiv preprint arXiv:2012.00460*, 2020.

[9] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. *Advances in Neural Information Processing Systems*, 32, 2019.

[10] Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Breaking the curse of many agents: Provable mean embedding q-iteration for mean-field reinforcement learning. In *International Conference on Machine Learning*, pages 10092–10103. PMLR, 2020.

[11] Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009.

[12] Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

[13] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[14] Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

[15] Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, 33:2710–2721, 2020.

[16] Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33:12248–12262, 2020.

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

[18] Andreas Christmann and Ingo Steinwart. Support vector machines. 2008.

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

[20] Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Optimal rates for regularized conditional mean embedding learning. *arXiv preprint arXiv:2208.01711*, 2022.

[21] Prem Talwai, Ali Shameli, and David Simchi-Levi. Sobolev norm learning rates for conditional mean embeddings. In *International Conference on Artificial Intelligence and Statistics*, pages 10422–10447. PMLR, 2022.

[22] Kjetil O Lye, Siddhartha Mishra, and Roberto Molinaro. A multi-level procedure for enhancing accuracy of machine learning algorithms. *European Journal of Applied Mathematics*, 32(3):436–469, 2021.

[23] Zhihan Li, Yuwei Fan, and Lexing Ying. Multilevel fine-tuning: Closing generalization gaps in approximation of solution maps under a limited budget for training data. *Multiscale Modeling & Simulation*, 19(1):344–373, 2021.

[24] Michael B Giles. Multilevel monte carlo path simulation. *Operations research*, 56(3):607–617, 2008.

[25] Michael B Giles. Multilevel monte carlo methods. *Acta numerica*, 24:259–328, 2015.

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

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

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

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

[30] Yuehaw Khoo, Jianfeng Lu, and Lexing Ying. Solving for high-dimensional committor functions using artificial neural networks. *Research in the Mathematical Sciences*, 6(1):1–13, 2019.

[31] Yifan Chen, Bamdad Hosseini, Houman Owhadi, and Andrew M Stuart. Solving and learning nonlinear pdes with gaussian processes. *Journal of Computational Physics*, 447:110668, 2021.

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

[33] Yiping Lu, Jose Blanchet, and Lexing Ying. Sobolev acceleration and statistical optimality for learning elliptic equations via gradient descent. *arXiv preprint arXiv:2205.07331*, 2022.

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

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

[36] Tianping Chen and Hong Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, 6(4):911–917, 1995.

[37] Zichao Long, Yiping Lu, Xianzhong Ma, and Bin Dong. Pde-net: Learning pdes from data. In *International Conference on Machine Learning*, pages 3208–3216. PMLR, 2018.

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

[39] Jordi Feliu-Faba, Yuwei Fan, and Lexing Ying. Meta-learning pseudo-differential operators with deep neural networks. *Journal of Computational Physics*, 408:109309, 2020.

[40] Yuehaw Khoo, Jianfeng Lu, and Lexing Ying. Solving parametric pde problems with artificial neural networks. *European Journal of Applied Mathematics*, 32(3):421–435, 2021.

[41] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces. *arXiv preprint arXiv:2108.08481*, 2021.

[42] George Stepaniants. Learning partial differential equations in reproducing kernel hilbert spaces. *arXiv preprint arXiv:2108.11580*, 2021.

[43] Hao Liu, Haizhao Yang, Minshuo Chen, Tuo Zhao, and Wenjing Liao. Deep nonparametric estimation of operators between infinite dimensional spaces. *arXiv preprint arXiv:2201.00217*, 2022.

[44] André Mas. Lower bound in regression for functional data by representation of small ball probabilities. *Electronic Journal of Statistics*, 6:1745–1778, 2012.

[45] Aurore Delaigle and Peter Hall. Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38(2):1171–1193, 2010.

[46] Lin Lin, Jianfeng Lu, and Lexing Ying. Fast construction of hierarchical matrix representation from matrix–vector multiplication. *Journal of Computational Physics*, 230(10):4071–4087, 2011.

[47] Matthew Reimherr. Functional regression with repeated eigenvalues. *Statistics & Probability Letters*, 107:62–70, 2015.

[48] Chang-han Rhee and Peter W Glynn. Unbiased estimation with square root convergence for sde models. *Operations Research*, 63(5):1026–1043, 2015.

[49] Jose H Blanchet and Peter W Glynn. Unbiased monte carlo for optimization and functions of expectations via multi-level randomization. In *2015 Winter Simulation Conference (WSC)*, pages 3656–3667. IEEE, 2015.

[50] Jose Blanchet and Zhipeng Liu. Malliavin-based multilevel monte carlo estimators for densities of max-stable processes. In *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 75–97. Springer, 2016.

[51] Guanting Chen, Alex Shkolnik, and Kay Giesecke. Unbiased simulation estimators for path integrals of diffusions. In *2020 Winter Simulation Conference (WSC)*, pages 277–288. IEEE, 2020.

[52] Florian Schäfer and Houman Owhadi. Sparse recovery of elliptic solvers from matrix-vector products. *arXiv preprint arXiv:2110.05351*, 2021.

[53] Nicolas Boullé, Seick Kim, Tianyi Shi, and Alex Townsend. Learning green's functions associated with time-dependent partial differential equations. *Journal of Machine Learning Research*, 23(218):1–34, 2022.

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

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

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

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

[58] Teresa Portone and Robert D Moser. Bayesian inference of an uncertain generalized diffusion operator. *SIAM/ASA Journal on Uncertainty Quantification*, 10(1):151–178, 2022.

[59] Xiang Huang, Zhanhong Ye, Hongsheng Liu, Beiji Shi, Zidong Wang, Kang Yang, Yang Li, Bingya Weng, Min Wang, Haotian Chu, et al. Meta-auto-decoder for solving parametric partial differential equations. *arXiv preprint arXiv:2111.08823*, 2021.

[60] Jennifer L Mueller and Samuli Siltanen. *Linear and nonlinear inverse problems with practical applications*. SIAM, 2012.

[61] Sergios Agapiou, Andrew M Stuart, and Yuan-Xiang Zhang. Bayesian posterior contraction rates for linear severely ill-posed inverse problems. *Journal of Inverse and Ill-posed Problems*, 22(3):297–321, 2014.

[62] Maarten De Hoop, Daniel Zhengyu Huang, Elizabeth Qian, and Andrew M Stuart. The cost-accuracy trade-off in operator learning with neural networks. *arXiv preprint arXiv:2203.13181*, 2022.

[63] Kaushik Bhattacharya, Bamdad Hosseini, Nikola B Kovachki, and Andrew M Stuart. Model reduction and neural networks for parametric pdes. *arXiv preprint arXiv:2005.03180*, 2020.

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

A Proof of the lower bound

In this section, we follow the lower bound proof in **(author?)** [19] 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-Gilber Lemma.

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

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

In our proof we will use a version from **(author?)** [19].

Lemma A.2 *[19, Theorem 20] Let $M \geqslant 2, (\Omega, \mathcal{A})$ 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*

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

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

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

Lemma A.3 (Varshamov-Gillbert Lemma,[64] Theorem 2.9) *Let $D \geq 8$. There exists a subset $\mathcal{V} = \{\tau^{(0)}, \cdots, \tau^{(2^{D/8})}\}$ of $D-$dimensional hypercube $\mathcal{H}^D = \{0,1\}^D$ such that $\tau^{(0)} = (0, 0, \cdots, 0)$ and the ℓ_1 distance between every two elements is larger than $\frac{D}{8}$*

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

A.2 Proof of the Lower Bound

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

$$C_\omega = \sqrt{\frac{32\varepsilon}{m_1 K}} \sum_{i=1}^{m_1}\sum_{j=1}^{K} \omega_{ij} \mu_{i+m_1}^{\beta'/2} \rho_{j+m_2}^{1-\gamma'/2} f_{j+m_2} \otimes e_{i+m_1}, \quad \omega_{ij} \in \{0,1\}$$

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

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

By Gilbert-Varshamov Lemma it is possible to select $M_\varepsilon \geqslant 2^{m_1 K/8}$ binary strings

$$\omega^{(1)}, \omega^{(2)}, \cdots, \omega^{(M_\varepsilon)} \in \{0,1\}^{m_1 K}$$

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

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

$$\|C_\omega\|^2_{\beta,\gamma} \leqslant \frac{32\varepsilon}{m_1 K}\sum_{i=1}^{m_1}\sum_{j=1}^{K} \mu_{i+m_1}^{-(\beta-\beta')} \rho_{j+m_2}^{-(\gamma'-\gamma)} \lesssim \varepsilon\left(2m_1\right)^{\frac{\beta-\beta'}{p}}\left(2m_2\right)^{\frac{\gamma'-\gamma}{q}}$$

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

$$(2m_1)^{\frac{\max\{\alpha,\beta\}-\beta'}{p}}(2m_2)^{\frac{\gamma'-\gamma}{q}} \lesssim \varepsilon^{-1} \tag{5}$$

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

$$\sup_{g\in\mathrm{range}(C_\omega)} \|g\|_{\mathcal{H}_L} \leqslant \sup_f \|C_\omega\|_{\alpha 1} \cdot \left\|\left(I_{1,\alpha,P_K}^*\right)^\dagger f\right\|_{\mathcal{H}_K^\alpha} < +\infty$$

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 $C_\omega, \forall\omega\in\Omega$.

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

$$P_\omega(\mathrm{d}f,\mathrm{d}g) = \mathrm{d}\mathcal{N}\left(C_\omega f,\Sigma\right)(g)\cdot \mathrm{d}P_K(f)$$

where the covariance operator $\Sigma = \frac{\sigma^2}{K}\sum_{j=1}^K \rho_{j+m_2} f_{j+m_2}\otimes f_{j+m_2}$ for some constant $\sigma > 0$. It's then easy to see that $\mathrm{tr}\left(\Sigma\right) = \sigma^2$, which satisfies Assumption 2.4 .Note that the range of C_ω is $\mathrm{span}(f_{m_2})$ and Σ is non-degenerate on this subspace. As a result, we can view $P_\omega, \omega\in\Omega$ as distributions on $\mathcal{H}_K\times\mathrm{span}\left(f_{j+m_2} : 1\leqslant j\leqslant K\right)$, and we have for $\forall\omega,\omega'\in\Omega$ that

$$\begin{aligned}
KL\left(P_\omega\|P_{\omega'}\right) &= \mathbb{E}_{f\sim P_K}\left[KL\left(P_\omega(\mathrm{d}g\mid f)\|P_{\omega'}(\mathrm{d}g\mid f)\right)\right]\\
&= \mathbb{E}_{f\sim P_K}\left[KL\left(\mathcal{N}(C_\omega f,\Sigma)\|\mathcal{N}(C_{\omega'} f,\Sigma)\right)\right]\\
&= \mathbb{E}_{f\sim P_K}\left\langle (C_\omega - C_{\omega'})f, \Sigma^\dagger(C_\omega - C_{\omega'})f\right\rangle\\
&\leqslant \sigma^{-2}K\mathbb{E}_{f\sim P_K}\left\langle (C_\omega - C_{\omega'})f, (C_\omega - C_{\omega'})f\right\rangle\\
&= \frac{32\varepsilon}{m_1\sigma^2}\mathbb{E}_{f\sim P_K}\left\|\sum_{i=1}^{m_1}\sum_{j=1}^{K}(\omega_{ij}-\omega'_{ij})\mu_{i+m_1}^{\beta'/2}\rho_{j+m_2}^{1-\gamma'/2}\left\langle f, e_{i+m_1}\right\rangle f_{j+m_2}\right\|_{\mathcal{H}_L}^2\\
&= \frac{32\varepsilon}{m_1\sigma^2}\mathbb{E}_{f\sim P_K}\sum_{j=1}^{K}\rho_{j+m_2}^{1-\gamma'}\left(\sum_{i=1}^{m_1}(\omega_{ij}-\omega'_{ij})\mu_{i+m_1}^{\beta'/2}\left\langle f, e_{i+m_1}\right\rangle\right)^2\\
&= \frac{32\varepsilon}{m_1\sigma^2}\sum_{i=1}^{m_1}\sum_{j=1}^{K}(\omega_{ij}-\omega'_{ij})^2\mu_{i+m_1}^{\beta'}\rho_{j+m_2}^{1-\gamma'} \lesssim \varepsilon\sigma^{-2}m_1^{-\frac{\beta'}{p}}m_2^{-\frac{1-\gamma'}{q}}
\end{aligned}$$

where the last step follows from $\mathbb{E}_{P_K} f\otimes f = C_{KK} = \sum_{i=1}^\infty \mu_i e_i\otimes e_i$ and recall that K is a constant. Hence we deduce that

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

Applying Lemma A.2, we find that when

$$\alpha^* \lesssim \log M_\varepsilon \Leftrightarrow \varepsilon \lesssim n^{-1}m_1^{\frac{\beta'}{p}}m_2^{\frac{1-\gamma'}{q}},$$

there exists a hypothesis P_{ω_0} such that for any estimator $\hat{C}_{\omega_0}$,

$$\left\{\left\|\hat{C}_{\omega_0} - C_{\omega_0}\right\|_{\beta',\gamma'}^2 \gtrsim \varepsilon\right\} \supset \left\{\omega_0 \neq \arg\min_{\omega\in\Omega}\|C_\omega - C_{\omega_0}\|_{\beta',\gamma'}\right\}$$

holds with high probability.

Finally, we need to choose optimal m_1 and m_2 under the constraint (5). It turns out that either $m_1 = 1$ or $m_2 = 1$, and the resulting lower bound is

$$\left\|\hat{C}_{X\to Y} - C_{X\to Y}\right\|_{\beta',\gamma'} \gtrsim n^{-\min\left\{\frac{\max\{\alpha,\beta\}-\beta'}{2(\max\{\alpha,\beta\}+p)},\frac{\gamma'-\gamma}{2(1-\gamma)}\right\}}.$$

B Proof of the upper bound

In this section, we upper-bound the learning error of estimator (2) which defined as

$$\hat{\mathcal{A}} = \sum_{j=1}^{y_N} \left(\rho_j^{\frac{1}{2}} f_j \otimes \rho_j^{\frac{1}{2}} f_j\right) \hat{\mathcal{C}}_{LK} \left(\hat{\mathcal{C}}_{KK} + \lambda_j I\right)^{-1}, \tag{6}$$

where $\lambda_j, 1 \leqslant j \leqslant y_N = N^{\frac{q}{1-\gamma'}\max\{1-\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{1-\gamma'}{1-\gamma}\}}$ 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 $\lambda_j = \max\left\{\left(j^{-\frac{1-\gamma'}{q}} N^{\max\left\{1-\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{1-\gamma'}{1-\gamma}\right\}}\right)^{-\frac{1}{\beta'+p}}, c_0\left(\frac{N}{\log N}\right)^{-\frac{1}{\alpha}}\right\}$ (3).
- We regularize all spectral component below certain bias contour, *i.e.* we set regularization strength $\lambda_j = \max\left\{\left(j^{-\frac{\gamma'-\gamma}{q}} N^{\min\left\{\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{\gamma'-\gamma}{1-\gamma}\right\}}\right)^{-\frac{1}{\beta-\beta'}}, c_0\left(\frac{N}{\log N}\right)^{-\frac{1}{\alpha}}\right\}$ (18).

To obtain the upper bound for our estimator, we decompose the learning error $\mathcal{E}(\hat{\mathcal{A}}) = \left\|\hat{\mathcal{A}} - \mathcal{A}_0\right\|_{\beta',\gamma'}$ in to bias and variance via

$$\mathcal{E}(\mathcal{A}) \leqslant \underbrace{\left\|\hat{\mathcal{A}} - \mathcal{A}_\lambda\right\|_{\beta',\gamma'}}_{\text{variance term}} + \underbrace{\|\mathcal{A}_\lambda - \mathcal{A}_0\|_{\beta',\gamma'}}_{\text{bias term}},$$

where

$$\mathcal{A}_\lambda = \sum_{j=1}^{y_N} \left(\rho_j^{\frac{1}{2}} f_j \otimes \rho_j^{\frac{1}{2}} f_j\right) \mathcal{C}_{KL} \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1}. \tag{7}$$

B.1 Regularization via Variance Counter

In the following, we separately bound the bias term and the variance term. We first assume $\alpha \leqslant \beta + p$ 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 $\|\mathcal{A}_0 - \mathcal{A}_\lambda\|_{\beta',\gamma'}^2 \lesssim N^{-\min\left\{\frac{\beta-\beta'}{\beta+p}, \frac{\gamma'-\gamma}{1-\gamma}\right\}}$.

***Proof sketch*:** Since $\|\mathcal{A}_0\|_{\beta,\gamma} \leqslant B$, we can write $\mathcal{A}_0 := \sum_{i=1}^{+\infty}\sum_{j=1}^{+\infty} a_{ij}\mu_i^{\frac{\beta}{2}}\rho_j^{1-\frac{\gamma}{2}} f_j \otimes e_i$ where the coefficient matrix $A_0 = (a_{ij})_{1\leqslant i,j \leqslant +\infty}$ satisfies $\|A_0\|_F^2 \leqslant B^2$. The definition (7) implies that for $1 \leqslant j \leqslant y_N$ and $i \geqslant 1$ we have

$$\begin{aligned}\left\langle \rho_j^{\frac{1}{2}} f_j, \mathcal{A}_\lambda \mu_i^{\frac{1}{2}} e_i\right\rangle &= \left\langle \rho_j^{\frac{1}{2}} f_j, \mathcal{C}_{KL}\left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1} \mu_i^{\frac{1}{2}} e_i\right\rangle \\ &= \left\langle \rho_j^{\frac{1}{2}} f_j, \mathcal{A}_0\mathcal{C}_{KK}\left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1} \mu_i^{\frac{1}{2}} e_i\right\rangle = \frac{\mu_i^{\frac{1+\beta}{2}}}{\mu_i + \lambda_j}\rho_j^{\frac{1-\gamma}{2}} a_{ij}.\end{aligned}$$

The bias term can be bounded as follows:

$$\begin{aligned}\|\mathcal{A}_0-\mathcal{A}_\lambda\|^2_{\beta',\gamma'} &= \sum_{i,j=1}^{+\infty}\left\langle \rho_j^{\frac{1}{2}} f_j, \mathcal{C}_{QK}^{-\frac{1-\gamma'}{2}}\left(\mathcal{A}_0-\mathcal{A}_\lambda\right)\mathcal{C}_{KK}^{\frac{1-\beta'}{2}}\mu_i^{\frac{1}{2}}e_i\right\rangle^2 \\ &= \sum_{j=1}^{y_N}\sum_{i=1}^{+\infty}\mu_i^{\beta-\beta'}\rho_j^{\gamma'-\gamma}\frac{\lambda_j^2}{(\mu_i+\lambda_j)^2}a_{ij}^2 \\ &\leqslant \sum_{j=1}^{y_N}\rho_j^{\gamma'-\gamma}\max_{i\geqslant 1}\left(\mu_i^{\beta-\beta'}\frac{\lambda_j^2}{(\mu_i+\lambda_j)^2}\right)\cdot\sum_{i=1}^{+\infty}a_{ij}^2 \\ &\lesssim \sum_{j=1}^{y_N} j^{-\frac{\gamma'-\gamma}{q}}\lambda_j^{-(\beta-\beta')}\sum_{i=1}^{+\infty}a_{ij}^2 \lesssim B^2\max_{1\leqslant j\leqslant y_N} j^{-\frac{\gamma'-\gamma}{q}}\lambda_j^{-(\beta-\beta')}.\end{aligned} \tag{8}$$

We now prove that

$$j^{\frac{\gamma'-\gamma}{q}}\lambda_j^{\beta-\beta'} \gtrsim N^{\min\left\{\frac{\beta-\beta'}{\beta+p},\frac{\gamma'-\gamma}{1-\gamma}\right\}}, \quad \forall 1\leqslant j\leqslant y_N. \tag{9}$$

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

$$j^{\frac{\gamma'-\gamma}{q}}\lambda_j^{\beta-\beta'} \geqslant \lambda_j^{\beta-\beta'} \gtrsim N^{\frac{\beta-\beta'}{\alpha}} \geqslant N^{\frac{\beta-\beta'}{\beta+p}}$$

where we use $\alpha\leqslant\beta+p$ in the final step.

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

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

$$j^{\frac{\gamma'-\gamma}{q}}\lambda_j^{\beta-\beta'} = j^{\frac{\gamma'-\gamma}{q}}\left(N^{\frac{\beta'+p}{\beta+p}}j^{-\frac{1-\gamma'}{q}}\right)^{\frac{\beta-\beta'}{\beta'+p}} = N^{\frac{\beta-\beta'}{\beta+p}}j^{\frac{1-\gamma'}{q}\left(\frac{\gamma'-\gamma}{1-\gamma'}-\frac{\beta-\beta'}{\beta'+p}\right)} \geqslant N^{\frac{\beta-\beta'}{\beta+p}}.$$

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

$$j^{\frac{\gamma'-\gamma}{q}}\lambda_j^{\beta-\beta'} = j^{\frac{\gamma'-\gamma}{q}}\left(N^{\frac{1-\gamma'}{1-\gamma}}j^{-\frac{1-\gamma'}{q}}\right)^{\frac{\beta-\beta'}{\beta'+p}} \geqslant y_N^{\frac{\gamma'-\gamma}{q}}\left(N^{\frac{1-\gamma'}{1-\gamma}}y_N^{-\frac{1-\gamma'}{q}}\right)^{\frac{\beta-\beta'}{\beta'+p}} = N^{\frac{\gamma'-\gamma}{1-\gamma}}.$$

Hence, in all cases (9) holds and we have that

$$\|\mathcal{A}_0-\mathcal{A}_\lambda\|^2_{\beta',\gamma'} \lesssim N^{-\min\left\{\frac{\beta-\beta'}{\beta+p},\frac{\gamma'-\gamma}{1-\gamma}\right\}}. \tag{10}$$

□

B.1.2 Variance

The variance term can be rewritten in the following way:

$$
\begin{aligned}
\mathcal{V} = \left\|\hat{\mathcal{A}} - \mathcal{A}_\lambda\right\|_{\beta',\gamma'}^2 &= \left\|C_{Q_K}^{-\frac{1-\gamma'}{2}}\left(\hat{\mathcal{A}} - \mathcal{A}_\lambda\right)C_{KK}^{\frac{1-\beta'}{2}}\right\|_{\text{HS}}^2 \\
&= \sum_{i,j=1}^{+\infty}\left\langle \rho_j^{\frac{1}{2}} f_j, C_{Q_K}^{-\frac{1-\gamma'}{2}}\left(\hat{\mathcal{A}} - \mathcal{A}_\lambda\right)C_{KK}^{\frac{1-\beta'}{2}}\mu_i^{\frac{1}{2}} e_i\right\rangle^2 && \text{(11a)} \\
&= \sum_{j=1}^{n_N}\rho_j^{-(1-\gamma')}\sum_{i=1}^{+\infty}\left\langle \rho_j^{\frac{1}{2}} f_j, \left[\hat{\mathcal{C}}_{LK}\left(\hat{\mathcal{C}}_{KK} + \lambda_j I\right)^{-1} - \mathcal{C}_{LK}\left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1}\right]\mu_i^{1-\frac{\beta'}{2}} e_i\right\rangle^2 && \text{(11b)} \\
&= \sum_{j=1}^{n_N}\rho_j^{-(1-\gamma')}\sum_{i=1}^{+\infty}\Bigg\langle \underbrace{\left(\mathcal{C}_{KK} + \lambda_j I\right)^{-\frac{1}{2}}\left[\hat{\mathcal{C}}_{KL} - \left(\hat{\mathcal{C}}_{KK} + \lambda_j I\right)\left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1}\mathcal{C}_{KL}\right]}_{=:U_j}\rho_j^{\frac{1}{2}} f_j, \\
&\qquad \underbrace{\left(\mathcal{C}_{KK} + \lambda_j I\right)^{\frac{1}{2}}\left(\hat{\mathcal{C}}_{KK} + \lambda_j I\right)^{-1}\left(\mathcal{C}_{KK} + \lambda_j I\right)^{\frac{1}{2}}}_{=:G_j}\frac{\mu_i^{1-\frac{\beta'}{2}}}{\sqrt{\mu_i + \lambda_j}}e_i\Bigg\rangle^2 && \text{(11c)} \\
&= \sum_{j=1}^{n_N}\rho_j^{-(1-\gamma')}\left\langle U_j\rho_j^{\frac{1}{2}} f_j, G_j\left(\sum_{i=1}^{+\infty}\frac{\mu_i^{2-\beta'}}{\mu_i + \lambda_j}e_i \otimes e_i\right)G_j U_j\rho_j^{\frac{1}{2}} f_j\right\rangle \\
&\lesssim \sum_{j=1}^{n_N} j^{\frac{1-\gamma'}{q}}\|G_j\|^2\lambda_j^{-\beta'}\left\|U_j\rho_j^{\frac{1}{2}} f_j\right\|^2 && \text{(11d)}
\end{aligned}
$$

In (11), (11a) uses the definition of the Hilbert-Schmidt norm; (11b) follows from the definition of $\hat{\mathcal{A}}$ (cf.(2)) and the fact that for any $j \geqslant y_N$, we have $\left\langle \rho_j^{\frac{1}{2}} f_j, \left(\hat{\mathcal{A}} - \mathcal{A}_\lambda\right)\mu_i^{\frac{1}{2}} e_i\right\rangle = 0$; (11c) is obtained from re-arranging and (11d) follows from $\left\|\sum_{i=1}^{+\infty}\frac{\mu_i^{2-\beta'}}{\mu_i+\lambda_j}e_i \otimes e_i\right\| = \max_{i\geqslant 1}\frac{\mu_i^{1-\beta'}}{\mu_i+\lambda_j} \lesssim \lambda_j^{-\beta'}$ and $\rho_j \lesssim j^{-\frac{1}{q}}$.

Note that

$$
\begin{aligned}
U_j &= \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-\frac{1}{2}}\left[\hat{\mathcal{C}}_{KL} - \mathcal{C}_{KL} - \left(\hat{\mathcal{C}}_{KK} - \mathcal{C}_{KK}\right)\left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1}\mathcal{C}_{KL}\right] \\
&= \frac{1}{N}\sum_{k=1}^{N}\left(\mathcal{C}_{KK} + \lambda_j I\right)^{-\frac{1}{2}}\left[u_k \otimes v_k - \mathbb{E}_{P_{KL}} u_k \otimes \mathcal{A}_0 u_k - \left(u_k \otimes u_k - \mathbb{E}_{P_{KL}} u_k \otimes u_k\right)\left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1}\mathcal{C}_{KK}\mathcal{A}_0^*\right] \\
&= \underbrace{\frac{1}{N}\sum_{k=1}^{N}\left(\mathcal{C}_{KK} + \lambda_j I\right)^{-\frac{1}{2}}\left(u_k \otimes \left(v_k - \mathcal{A}_0 u_k\right)\right)}_{:=U_j^1} \\
&\quad + \underbrace{\frac{1}{N}\sum_{k=1}^{N}\left(\mathcal{C}_{KK} + \lambda_j I\right)^{-\frac{1}{2}}\left[u_k \otimes \mathcal{A}_0 u_k - \mathbb{E}_{P_{KL}} u_k \otimes \mathcal{A}_0 u_k - \left(u_k \otimes u_k - \mathbb{E}_{P_{KL}} u_k \otimes u_k\right)\left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1}\mathcal{C}_{KK}\mathcal{A}_0^*\right]}_{:=U_j^2=\lambda_j\frac{1}{N}\sum_{k=1}^{N}\left(\mathcal{C}_{KK}+\lambda_j I\right)^{-\frac{1}{2}}\left(u_k\otimes\mathcal{A}_0\left(\mathcal{C}_{KK}+\lambda_j I\right)^{-1}u_k-\mathbb{E}_{P_{KL}}u_k\otimes\mathcal{A}_0\left(\mathcal{C}_{KK}+\lambda_j I\right)^{-1}u_k\right)}.
\end{aligned}
$$

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 (11), we deduce that $\mathcal{V} \leqslant 2\left(\mathcal{V}_1 + \mathcal{V}_2\right)$ where

$$\begin{aligned}
\mathcal{V}_1 &\lesssim \max_{1\leqslant j\leqslant n_N} \|G_j\|^2 \sum_{j=1}^{n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-\beta'} \underbrace{\left\| \frac{1}{N} \sum_{k=1}^{N} \left[\left\langle v_k - \mathcal{A}_0 u_k, \rho_j^{\frac{1}{2}} f_j \right\rangle \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-\frac{1}{2}} u_k \right] \right\|^2}_{:=\mathcal{V}_{1,j}^2} \\
\mathcal{V}_2 &\lesssim \max_{1\leqslant j\leqslant n_N} \|G_j\|^2 \sum_{j=1}^{n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{2-\beta'} \underbrace{\left\| \left(\hat{\mathbb{E}} - \mathbb{E}\right) \left[\left\langle \mathcal{A}_0 \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1} u_k, \rho_j^{\frac{1}{2}} f_j \right\rangle \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-\frac{1}{2}} u_k \right] \right\|^2}_{:=\mathcal{V}_{2,j}^2}
\end{aligned} \tag{12}$$

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

$$E_{1,j} = \left\{ G_j = \left\| \left[\mathcal{P}_{i_j}\left(\mathcal{C}_{KK}\right)\right]^{\frac{1}{2}} \left[\mathcal{P}_{i_j}\left(\hat{\mathcal{C}}_{KK}\right)\right]^{\dagger} \left[\mathcal{P}_{i_j}\left(\mathcal{C}_{KK}\right)\right]^{\frac{1}{2}} \right\| \leqslant 2\sqrt{a_1}. \right\}.$$

Recall that $m_N \leqslant c_0 \left(\frac{N}{\log N}\right)^{\frac{p}{\alpha}}$, by Theorem D.3, we know that $E_{1,j}$ holds with probability $\geqslant 1 - 2e^{-a_1}$. As a result $E_1 = \cap_{j=1}^{n_N} E_{1,j}$ holds with probability $\geqslant 1 - 2n_N e^{-a_1}$. We assume event E_1 holds in all the following proof.

Bounding $\mathcal{V}_1$. Let

$$X_{j,k} = j^{\frac{1-\gamma'}{2q}} \lambda_j^{-\frac{\beta'}{2}} \left\langle v_k - \mathcal{A}_0 u_k, \rho_j^{\frac{1}{2}} f_j \right\rangle \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-\frac{1}{2}} u_k \in \mathcal{H}_K$$

and $X_k = (X_{j,k} : 1 \leqslant j \leqslant n_N) \in \mathcal{H}_K^{y_N}$. Then we have $\mathcal{V}_1 \lesssim \left\| \frac{1}{N}\sum_{k=1}^{N} X_k \right\|^2$ where the norm here defined for $\mathcal{H}_K^{\otimes y_N}$ is induced by $\langle a, b\rangle = \sum_{i=1}^{n_N} \langle a_i, b_i\rangle_{\mathcal{H}_K}$. Note that $X_k, k = 1, 2, \cdots, N$ are i.i.d. random variables with mean zero, and

$$\begin{aligned}
&\mathbb{E}\left\|X_1\right\|^{2t} = \mathbb{E}_{P_{KL}} \left[\left(\sum_{j=1}^{n_N} \|X_{j,k}\|^2 \right)^t \right] \\
&= \mathbb{E}_{P_{KL}} \left[\left(\sum_{j=1}^{n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-\beta'} \left\langle v_1 - \mathcal{A}_0 u_1, \rho_j^{\frac{1}{2}} f_j \right\rangle^2 \left\| \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-\frac{1}{2}} u \right\|^2 \right)^t \right] \\
&\leqslant \max_{1\leqslant j\leqslant y_N} \sup_{u\in\mathrm{supp}(P_K)} \left(\underbrace{j^{\frac{1-\gamma'}{q}} i_j^{\frac{\beta'}{p}} \left\| \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-\frac{1}{2}} u \right\|^2}_{=:G_1} \right)^{t-1} \cdot \\
&\quad \underbrace{\mathbb{E}_{(u,v)\sim P_{KL}} \left[\left\|v - \mathcal{A}_0 u\right\|^{2t-2} \left(\sum_{j=1}^{n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-\beta'} \left\langle v - \mathcal{A}_0 u, \rho_j^{\frac{1}{2}} f_j \right\rangle^2 \left\| \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-\frac{1}{2}} u \right\|^2 \right) \right]}_{=:G_2}
\end{aligned}$$

By Lemma D.2 we have

$$G_1 \lesssim j^{\frac{1-\gamma'}{q}} \lambda_j^{-(\beta'+\alpha)}.$$

For G_2, note that for fixed u, Assumption 2.4 implies that

$$\begin{aligned}
&\mathbb{E}_{v|u} \left[\left\|v - \mathcal{A}_0 u\right\|^{2t-2} \left(\sum_{j=1}^{n_N} j^{\frac{1-\gamma'}{q}} i_j^{\frac{\beta'}{p}} \left\langle v - \mathcal{A}_0 u, \rho_j^{\frac{1}{2}} f_j \right\rangle^2 \left\| \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-\frac{1}{2}} u \right\|^2 \right) \right] \\
&\leqslant \frac{1}{2}(2t)! R^{2t-2} \sum_{j=1}^{n_N} \sigma_j^2 j^{\frac{1-\gamma'}{q}} \lambda_j^{-\beta'} \left\| \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-\frac{1}{2}} u \right\|^2.
\end{aligned}$$

where $\sigma_j^2 = \left\langle \rho_j^{\frac{1}{2}} f_j, V \rho^{\frac{1}{2}} f_j \right\rangle$. As a result, we have

$$G_2 \leqslant \mathbb{E}_{P_K} \left[\frac{1}{2}(2t)! R^{2t-2} \sum_{j=1}^{n_N} \sigma_j^2 j^{\frac{1-\gamma'}{q}} i_j^{\frac{\beta'}{p}} \left\| (\mathcal{C}_{KK} + \lambda_j I)^{-\frac{1}{2}} u \right\|^2 \right] \leqslant \frac{1}{2}(2t)! R^{2t-2} \sigma^2 \max_{1 \leqslant j \leqslant n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-(p+\beta')},$$

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

$$\begin{aligned} \mathbb{E}_{P_K} \left[\left\| (\mathcal{C}_{KK} + \lambda_j I)^{-\frac{1}{2}} u \right\|^2 \right] &\leqslant \operatorname{tr}\left(\mathbb{E}_{P_K} \left[(\mathcal{C}_{KK} + \lambda_j I)^{-\frac{1}{2}} u \otimes (\mathcal{C}_{KK} + \lambda_j I)^{-\frac{1}{2}} u \right] \right) \\ &= \operatorname{tr}\left(\sum_{i=1}^{+\infty} \frac{\mu_i^2}{\mu_i + \lambda_j} e_i \otimes e_i \right) \\ &= \sum_{i=1}^{+\infty} \frac{\mu_i}{\mu_i + \lambda_j} \\ &\lesssim \lambda_j^{-p}. \end{aligned}$$

We have shown that for some constant $c_1 > 0$,

$$\mathbb{E}\|X_1\|^{2t} \leqslant \frac{1}{2}(2t)! \sigma^2 \max_{1 \leqslant j \leqslant n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-(p+\beta')} \cdot \left(c_1 R^2 \max_{1 \leqslant j \leqslant n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-(\beta'+p)} \right)^{t-1}.$$

By Bernstein's inequality, the event

$$E_2 := \left\{ \left\| \frac{1}{N} \sum_{k=1}^{N} X_k \right\|^2 \leqslant 6a_2 \left(\frac{\sigma^2 \max_{j \in [y_N]} j^{\frac{1-\gamma'}{q}} \lambda_j^{-(\beta'+p)}}{N} + \frac{c_1 R^2 \max_{1 \leqslant j \leqslant n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-(\beta'+\alpha)}}{N^2} \right) \right\} \tag{13}$$

holds with probability $\geqslant 1 - 2e^{-a_2}$. By our definition of λ_j, we have

$$\max_{1 \leqslant j \leqslant n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-(\beta'+p)} \lesssim N^{\max\left\{ \frac{\beta'+p}{\beta+p}, \frac{1-\gamma'}{1-\gamma} \right\}}$$

and $\lambda_j \gtrsim N^{-\frac{1}{\alpha}}$ (which implies that the $\frac{1}{N^2}$ term is dominated by the $\frac{1}{N}$ term). Hence, under $E_1 \cap E_2$ we have

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

with probability $\geqslant 1 - 2n_N e^{-a_2}$.

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

$$\begin{aligned} &\mathbb{E}_{u \sim P_K} \left[\left\langle \mathcal{A}_0 (\mathcal{C}_{KK} + \lambda_j I)^{-1} u, \rho_j^{\frac{1}{2}} f_j \right\rangle^2 \right] \\ &= \mathbb{E}_{u \sim P_K} \left\langle \rho_j^{\frac{1}{2}} f_j, \mathbb{E}_{P_K} \left[\mathcal{A}_0 (\mathcal{C}_{KK} + \lambda_j I)^{-1} u \otimes \mathcal{A}_0 (\mathcal{C}_{KK} + \lambda_j I)^{-1} u \right] \rho_j^{\frac{1}{2}} f_j \right\rangle \\ &= \left\langle \rho_j^{\frac{1}{2}} f_j, \mathcal{A}_0 (\mathcal{C}_{KK} + \lambda_j I)^{-1} \mathcal{C}_{KK} (\mathcal{C}_{KK} + \lambda_j I)^{-1} \mathcal{A}_0^* \rho_j^{\frac{1}{2}} f_j \right\rangle & \text{(14a)} \\ &= \rho_j^{1-\gamma} \left\langle \left(\mathcal{C}_{Q_K}^{-\frac{1-\gamma}{2}} \mathcal{A}_0 \mathcal{C}_{KK}^{\frac{1-\beta}{2}} \right)^* \rho_j^{\frac{1}{2}} f_j, (\mathcal{C}_{KK} + \lambda_j I)^{-1} \mathcal{C}_{KK}^{\beta} (\mathcal{C}_{KK} + \lambda_j I)^{-1} \left(\mathcal{C}_{Q_K}^{-\frac{1-\gamma}{2}} \mathcal{A}_0 \mathcal{C}_{KK}^{\frac{1-\beta}{2}} \right)^* \rho_j^{\frac{1}{2}} f_j \right\rangle & \text{(14b)} \\ &\lesssim j^{-\frac{1-\gamma}{q}} \lambda_j^{-(2-\beta)} \underbrace{\left\| \left(\mathcal{C}_{Q_K}^{-\frac{1-\gamma}{2}} \mathcal{A}_0 \mathcal{C}_{KK}^{\frac{1-\beta}{2}} \right)^* \rho_j^{\frac{1}{2}} f_j \right\|^2}_{=:D_{j,2}} & \text{(14c)} \end{aligned}$$

where (14a) follows from $\mathbb{E}_{P_K} u \otimes u = \mathcal{C}_{KK}$, (14b) uses the fact that $\mathcal{C}_{KK}$ and $\mathcal{C}_{KK} + \lambda_j I$ commute, and lastly (14c) follows from $\left\| (\mathcal{C}_{KK} + \lambda_j I)^{-1} \mathcal{C}_{KK}^{\beta} (\mathcal{C}_{KK} + \lambda_j I)^{-1} \right\|_{\mathcal{H}_K} \propto \lambda_j^{-(2-\beta)}$.

Let

$$Y_{j,k} = \left\langle \mathcal{A}_0 \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1} u_k, \rho_j^{\frac{1}{2}} f_j \right\rangle \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-\frac{1}{2}} u_k \in \mathcal{H}_K$$

and

$$Y_k = \left(Y_{j,k} : 1 \leqslant j \leqslant y_N\right) \in \mathcal{H}_K^{n_N}.$$

Then we have

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

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

$$\begin{aligned}
&\mathbb{E}\|Y_1\|^{2t} = \mathbb{E}\left[\left(\sum_{j=1}^{n_N} \|Y_{j,k}\|^2\right)^t\right] \\
&= \mathbb{E}_{P_K}\left[\left(\sum_{j=1}^{n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{2-\beta'} \left\langle \mathcal{A}_0 \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1} u_1, \rho_j^{\frac{1}{2}} f_j \right\rangle^2 \left\| \mathcal{C}_{KK}^{-\frac{1}{2}} \mathcal{I}_{i_j}\left(u_1\right) \right\|^2 \right)^t\right] \\
&\leqslant \sup_{u \in \operatorname{supp}(P_K)} \left(\sum_{j=1}^{n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{2-\beta'} \left\langle \mathcal{A}_0 \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1} u, \rho_j^{\frac{1}{2}} f_j \right\rangle^2 \left\| \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-\frac{1}{2}} u \right\|^2 \right)^{t-1} \cdot \\
&\quad \sum_{j=1}^{n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{2-\beta'} \mathbb{E}\left[\left\langle \mathcal{A}_0 \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1} u, \rho_j^{\frac{1}{2}} f_j \right\rangle^2\right] \sup_{u \in \operatorname{supp}(P_K)} \left\| \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-\frac{1}{2}} u \right\|^2 \\
&\lesssim \sup_{u \in \operatorname{supp}(P_K)} \left(\sum_{j=1}^{n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{2-\beta'-\alpha} \left\langle \mathcal{A}_0 \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1} u, \rho_j^{\frac{1}{2}} f_j \right\rangle^2 \right)^{t-1} \cdot \sum_{j=1}^{n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-(\beta'+\alpha-\beta)} D_{j,2}.
\end{aligned} \tag{15}$$

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

$$\sum_{j=1}^{n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-(\beta'+\alpha)} \left\langle \mathcal{A}_0 \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1} \lambda_j u, \rho_j^{\frac{1}{2}} f_j \right\rangle^2$$

$$\leqslant \sum_{j=1}^{n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-(\beta'+\alpha)} \left(2\left\langle \mathcal{A}_0 u, \rho_j^{\frac{1}{2}} f_j \right\rangle^2 + 2\rho_j^{1-\gamma} \left\langle \mathcal{C}_{Q_K}^{-\frac{1-\gamma}{2}} \mathcal{A}_0 \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1} \mathcal{C}_{KK} u, \rho_j^{\frac{1}{2}} f_j \right\rangle^2\right) \tag{16a}$$

$$\lesssim \max_{1 \leqslant j \leqslant y_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-(\beta'+\alpha)} + \sum_{j=1}^{n_N} \lambda_j^{-(\beta'+\alpha)} \lambda_j^{-\max\{\alpha-\beta,0\}} \left\| \left(\mathcal{C}_{Q_K}^{-\frac{1-\gamma}{2}} \mathcal{A}_0 \mathcal{C}_{KK}^{\frac{1-\beta}{2}}\right)^* \rho_j^{\frac{1}{2}} f_j \right\|^2 \tag{16b}$$

$$\lesssim \max_{1 \leqslant j \leqslant y_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-(\beta'+\alpha)} + \max_{1 \leqslant j \leqslant y_N} \lambda_j^{-(\beta'+\alpha)-\max\{\alpha-\beta,0\}}. \tag{16c}$$

where (16a) uses the AM-GM inequality, (16b) follows from the assumption that $\|\mathcal{A}_0 u\| \leqslant A_2$ is uniformly bounded, and that

$$\left\| \mathcal{C}_{KK}^{-\frac{1-\beta}{2}} \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1} \mathcal{C}_{KK} u \right\| = \left\| \mathcal{C}_{KK}^{1-\frac{\alpha-\beta}{2}} \left(\mathcal{C}_{KK} + \lambda_j I\right)^{-1} \left(\mathcal{C}_{KK}^{-\frac{1-\alpha}{2}} u\right) \right\| \lesssim \lambda_j^{-\max\{\alpha-\beta,0\}}.$$

by Assumption 2.2, and lastly (16c) follows from $\|\mathcal{A}_0\|_{\beta,\gamma} \leqslant B$.

Plugging into (15), we deduce that

$$\begin{aligned}
&\mathbb{E}\|Y_1\|^{2t} \\
&\lesssim \sup_{u \in \operatorname{supp}(P_K)} \left(\max_{1 \leqslant j \leqslant n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-(\beta'+\alpha)} + \max_{1 \leqslant j \leqslant n_N} \lambda_j^{-(\beta'+\alpha)-\max\{\alpha-\beta,0\}}\right)^{t-1} \cdot \sum_{j=1}^{n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-(\beta'+\alpha-\beta)} D_{j,2} \\
&\lesssim \sup_{u \in \operatorname{supp}(P_K)} \left(\max_{1 \leqslant j \leqslant n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-(\beta'+\alpha)} + \max_{1 \leqslant j \leqslant n_N} \lambda_j^{-(\beta'+\alpha)-\max\{\alpha-\beta,0\}}\right)^{t-1} \max_{1 \leqslant j \leqslant n_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-(\beta'+\alpha-\beta)}
\end{aligned}$$

where the last step follows from $\sum_{j=1}^{+\infty} D_{j,2} = \|\mathcal{A}_0\|_{\beta,\gamma}^2$.

By Bernstein's inequality, there exists a constant C_3 such that the event

$$E_3 = \left\{ \mathcal{V}_2 \leqslant 6a_3C_3 \left(\frac{j^{\frac{1-\gamma'}{q}} \lambda_j^{-(\beta'+\alpha-\beta)}}{N} + \frac{\max_{1\leqslant j\leqslant n_N} \lambda_j^{-(\beta'+\alpha)} \left(j^{\frac{1-\gamma'}{q}} + \lambda_j^{-\max\{\alpha-\beta.0\}} \right)}{N^2} \right) \right\} \tag{17}$$

holds with probability $\geqslant 1 - 2e^{-a_3}$.

The definition of λ_N implies that the $\frac{1}{N^2}$ term is dominated by the $\frac{1}{N}$ term, so

$$\mathcal{V}_2 \lesssim a_1a_3 \frac{1}{N} \max_{1\leqslant j\leqslant y_N} j^{\frac{1-\gamma}{q}} \lambda_j^{-(\beta'+\alpha-\beta)} \lesssim N^{-\min\left\{\frac{\beta-\beta'}{\beta+p}, \frac{\gamma'-\gamma}{1-\gamma}\right\}}$$

holds under $E_1 \cap E_3$. To summarize, under $E_1 \cap E_2 \cap E_3$ which holds with probability $\geqslant 1 - 2n_Ne^{-a_1} - 2e^{-a_2} - 2e^{-a_3}$, we have

$$\mathcal{V} \leqslant 2a_1 \max\{a_2, a_3\} \left(\mathcal{V}_1 + \mathcal{V}_2\right) \lesssim N^{-\min\left\{\frac{\beta-\beta'}{\beta+p}, \frac{\gamma'-\gamma}{1-\gamma}\right\}}.$$

Recall that the bias term is upper bounded in (10). This gives the final upper bound

$$\left\|\hat{\mathcal{A}} - \mathcal{A}_0\right\|_{\beta',\gamma'} \lesssim N^{-\min\left\{\frac{\beta-\beta'}{2(\beta+p)}, \frac{\gamma'-\gamma}{2(1-\gamma)}\right\}}.$$

B.1.3 The hard-learning regime

In the previous sections, we focus on the case where $\alpha \leqslant \beta + p$ and establish an upper bound for the convergence rate via an optimal bias-variance trade-off. The opposite case, $\alpha > \beta + p$ 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 [19, Theorem 2]). In the hard learning regime the $\mathcal{V}_2$ 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:

$$S_N = \left\{ (x,y) \in \mathbb{Z}^2 \,\middle|\, x^{\frac{\beta'+\alpha-\beta}{p}} y^{\frac{1-\gamma'}{q}} \leqslant N^{1-\min\left\{\frac{\beta-\beta'}{\alpha}, \frac{\gamma'-\gamma}{1-\gamma}\right\}} \text{ and } x \leqslant c_0 \left(\frac{N}{\log N}\right)^{\frac{p}{\alpha}} \right\}.$$

The definition implies that the variance can be controlled by $N^{-\min\left\{\frac{\beta-\beta'}{2\alpha}, \frac{\gamma'-\gamma}{2(1-\gamma)}\right\}}$ and it remains to focus on the bias term.

Similar to the derivations in Appendix B.1.1, we have

$$\|\mathcal{A}_0 - \mathcal{T}_N\left(\mathcal{A}_0\right)\|_{\beta',\gamma'}^2 \lesssim \max_{(i,j)\notin S_N} i^{-\frac{\beta-\beta'}{p}} j^{-\frac{\gamma'-\gamma}{q}}.$$

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

- $i = \mathcal{O}(1)$. Then we have $j \gtrsim N^{\frac{q}{1-\gamma'}\left(1-\min\left\{\frac{\beta-\beta'}{\alpha}, \frac{\gamma'-\gamma}{1-\gamma}\right\}\right)}$ so that
$$i^{-\frac{\beta-\beta'}{p}} j^{-\frac{\gamma'-\gamma}{q}} \lesssim N^{-\frac{\gamma'-\gamma}{1-\gamma'}\left(1-\min\left\{\frac{\beta-\beta'}{\alpha}, \frac{\gamma'-\gamma}{1-\gamma}\right\}\right)} \leqslant N^{-\frac{\gamma'-\gamma}{1-\gamma}}.$$
- $j = \mathcal{O}(1)$. In this case we must have $i \lesssim N^{\min\left\{\frac{p}{\alpha}, \frac{p}{\beta-\beta'} \frac{\gamma'-\gamma}{1-\gamma}\right\}}$, otherwise it falls into S_N by definition. Hence we have
$$i^{-\frac{\beta-\beta'}{p}} j^{-\frac{\gamma'-\gamma}{q}} \leqslant i^{-\frac{\beta-\beta'}{p}} \lesssim N^{-\min\left\{\frac{\beta-\beta'}{\alpha}, \frac{\gamma'-\gamma}{1-\gamma}\right\}}.$$

On the other hand, for the variance term we still have $\mathcal{V}_1 \lesssim \frac{1}{N} \max_{1\leqslant j\leqslant n_N} j^{\frac{1-\gamma'}{q}} i_j^{\frac{\beta'+p}{p}}$ and $\mathcal{V}_2 \leqslant \frac{1}{N} \max_{1\leqslant j\leqslant n_N} j^{\frac{1-\gamma'}{q}} i_j^{\frac{\beta'+\alpha-\beta}{p}}$, so that

$$\mathcal{V} \lesssim \frac{1}{N} \max_{1\leqslant j\leqslant n_N} j^{\frac{1-\gamma'}{q}} i_j^{\frac{\beta'+\alpha-\beta}{p}} \leqslant N^{-\min\left\{\frac{\beta-\beta'}{\beta+p}, \frac{\gamma'-\gamma}{1-\gamma}\right\}}.$$

As a result, we can obtain the following convergence rate:

$$\left\|\hat{\mathcal{A}} - \mathcal{A}_0\right\|_{\beta',\gamma'} \lesssim N^{-\min\left\{\frac{\beta-\beta'}{2\alpha}, \frac{\gamma'-\gamma}{2(1-\gamma)}\right\}}.$$

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 (2) with the choice

$$\lambda_j = \max\left\{\left(j^{-\frac{\gamma'-\gamma}{q}} N^{\min\left\{\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{\gamma'-\gamma}{1-\gamma}\right\}}\right)^{-\frac{1}{\beta-\beta'}}, c_0\left(\frac{N}{\log N}\right)^{-\frac{1}{\alpha}}\right\}. \tag{18}$$

It now remains to plug the above λ_j into our bounds for bias and variance derived in the previous subsections.

Bounding the bias term. It follows from (8) that

$$\begin{aligned}
\|\mathcal{A}_0 - \mathcal{A}_\lambda\|_{\beta,\gamma'}^2 &\lesssim \max_{1\leqslant j\leqslant y_N} j^{-\frac{\gamma'-\gamma}{q}} \lambda_j^{\beta-\beta'} \\
&\lesssim \max\left\{N^{-\min\left\{\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{\gamma'-\gamma}{1-\gamma}\right\}}, c_0\left(\frac{N}{\log N}\right)^{-\frac{\beta-\beta'}{\alpha}}\right\} \\
&\lesssim N^{-\min\left\{\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{\gamma'-\gamma}{1-\gamma}\right\}}.
\end{aligned}$$

Bounding the variance term It follows from (13) and (17) that the variance is bounded by

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

As before, we consider the cases $\alpha \leqslant \beta + p$ and $\alpha > \beta + p$ separately.

- If $\alpha \leqslant \beta + p$, then it follows that

$$\begin{aligned}
\left\|\hat{\mathcal{A}} - \mathcal{A}_\lambda\right\|_{\beta',\gamma'}^2 &\lesssim \frac{1}{N} \max_{1\leqslant j\leqslant y_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-\left(\beta'+p\right)} \\
&\lesssim \frac{1}{N} \max_{1\leqslant j\leqslant y_N} j^{\frac{1-\gamma'}{q}} \left(j^{-\frac{\gamma'-\gamma}{q}} N^{\min\left\{\frac{\beta-\beta'}{\beta+p}, \frac{\gamma'-\gamma}{1-\gamma}\right\}}\right)^{\frac{\beta'+p}{\beta-\beta'}} \\
&\lesssim \frac{1}{N} \max_{1\leqslant j\leqslant y_N} j^{\frac{\gamma'-\gamma}{q}\left(\frac{1-\gamma'}{\gamma'-\gamma} - \frac{\beta'+p}{\beta-\beta'}\right)} N^{\frac{\beta'+p}{\beta-\beta'}\min\left\{\frac{\beta-\beta'}{\beta+p}, \frac{\gamma'-\gamma}{1-\gamma}\right\}} \\
&= \frac{1}{N} \max_{j\in\{1,y_N\}} j^{\frac{\gamma'-\gamma}{q}\left(\frac{1-\gamma'}{\gamma'-\gamma} - \frac{\beta'+p}{\beta-\beta'}\right)} N^{\frac{\beta'+p}{\beta-\beta'}\min\left\{\frac{\beta-\beta'}{\beta+p}, \frac{\gamma'-\gamma}{1-\gamma}\right\}} \\
&= N^{\min\left\{\frac{\beta-\beta'}{\beta+p}, \frac{\gamma'-\gamma}{1-\gamma}\right\}\max\left\{\frac{\beta'+p}{\beta-\beta'}, \frac{1-\gamma'}{\gamma'-\gamma}\right\}-1} \\
&= N^{-\min\left\{\frac{\beta-\beta'}{\beta+p}, \frac{\gamma'-\gamma}{1-\gamma}\right\}},
\end{aligned}$$

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

- If $\alpha > \beta + p$, then similarly we have

$$\begin{aligned}
\left\|\hat{\mathcal{A}} - \mathcal{A}_\lambda\right\|_{\beta',\gamma'}^2 &\lesssim \frac{1}{N} \max_{1\leqslant j\leqslant y_N} j^{\frac{1-\gamma'}{q}} \lambda_j^{-\beta'+\alpha-\beta} \\
&\lesssim \frac{1}{N} \max_{1\leqslant j\leqslant y_N} j^{\frac{1-\gamma'}{q}} \left(j^{-\frac{\gamma'-\gamma}{q}} N^{\min\left\{\frac{\beta-\beta'}{\alpha}, \frac{\gamma'-\gamma}{1-\gamma}\right\}}\right)^{\frac{\beta'+\alpha-\beta}{\beta-\beta'}} \\
&= \frac{1}{N} \max_{j\in\{1,y_N\}} j^{\frac{1-\gamma'}{q}} \left(j^{-\frac{\gamma'-\gamma}{q}} N^{\min\left\{\frac{\beta-\beta'}{\alpha}, \frac{\gamma'-\gamma}{1-\gamma}\right\}}\right)^{\frac{\beta'+\alpha-\beta}{\beta-\beta'}} \\
&\leqslant N^{-\min\left\{\frac{\beta-\beta'}{\alpha}, \frac{\gamma'-\gamma}{1-\gamma}\right\}}.
\end{aligned}$$

Hence we deduce that

$$\left\|\hat{\mathcal{A}} - \mathcal{A}_\lambda\right\|_{\beta',\gamma'}^2 \lesssim N^{-\min\left\{\frac{\beta-\beta'}{\alpha}, \frac{\gamma'-\gamma}{1-\gamma}\right\}},$$

as desired.

B.3 Implication of the upper bound

In this section, we discuss the implications of our upper bounds under the (β', γ')-norm.

Note that $\left\|\mathcal{C}_{Q_K}^{-\frac{1-\gamma'}{2}} v\right\|_{\mathcal{H}_L} = \|v\|_{\mathcal{H}_L^{2-\gamma'}}$ for all $v \in L_2(Q_K)$ (if one side of the equation is $+\infty$ then so is the other), we have that

$$\begin{aligned}
\mathbb{E}_{u\sim P_K} \left\|\left(\hat{\mathcal{A}} - \mathcal{A}_0\right) u\right\|_{\mathcal{H}_L^{2-\gamma'}}^2 &= \mathbb{E}_{u\sim P_K} \left\|\mathcal{C}_{Q_K}^{-\frac{1-\gamma'}{2}} \left(\hat{\mathcal{A}} - \mathcal{A}_0\right) u\right\|_{\mathcal{H}_L}^2 \\
&= \operatorname{tr}\left(\mathcal{C}_{Q_K}^{-\frac{1-\gamma'}{2}} \left(\hat{\mathcal{A}} - \mathcal{A}_0\right) \mathbb{E}_{u\sim P_K} u \otimes u \left(\mathcal{C}_{Q_K}^{-\frac{1-\gamma'}{2}} \left(\hat{\mathcal{A}} - \mathcal{A}_0\right)\right)^*\right) \\
&\lesssim \left\|\hat{\mathcal{A}} - \mathcal{A}_0\right\|_{\beta',\gamma'}^2,
\end{aligned} \tag{19}$$

where the last step follows from $\mathbb{E}_{u\sim P_K} u \otimes u = \mathcal{C}_{P_K}$. Note that the above derivations hold for any $0 \leqslant \beta' < \beta$, so choosing $\beta' = 0$ yields the best upper bound. We can see from (19) that our analysis implies an upper bound of the expected error of the learned solution evaluated under the $\mathcal{H}_L^{2-\gamma'}$ norm. On the other hand, it is also possible to obtain a *uniform* convergence rate when $\beta' \geqslant \alpha$:

$$\begin{aligned}
\left\|\left(\hat{\mathcal{A}} - \mathcal{A}_0\right) u\right\|_{\mathcal{H}_L^{2-\gamma'}} &= \left\|\mathcal{C}_{Q_K}^{-\frac{1-\gamma'}{2}} \left(\hat{\mathcal{A}} - \mathcal{A}_0\right) u\right\|_{\mathcal{H}_L} \\
&\leqslant \left\|\hat{\mathcal{A}} - \mathcal{A}_0\right\|_{\beta',\gamma'} \cdot \left\|\mathcal{C}_{P_K}^{-\frac{1-\beta'}{2}} u\right\|_{\mathcal{H}_K} \lesssim \left\|\hat{\mathcal{A}} - \mathcal{A}_0\right\|_{\beta',\gamma'}.
\end{aligned}$$

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 $\eta_1 = \min\left\{\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{\gamma'-\gamma}{1-\gamma}\right\}$ and $\eta_2 = \max\left\{1 - \frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{1-\gamma'}{1-\gamma}\right\} = 1 - \eta_1$. We first restrict ourselves to the case when $\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}} \neq \frac{\gamma'-\gamma}{1-\gamma}$; the special case when the two terms are equal will be separately treated in Appendix C.1. For the optimal bias and variance contours $\ell_{C_1,\texttt{bias}}$ and $\ell_{C_2,\texttt{var}}$ with $C_1 = N^{\eta_1}$ and $C_2 = N^{\eta_2}$, we define a sequence $\{x_n\}$ as follows:

$$x_0 = \max\left\{\frac{1}{2} N^{\frac{p}{\beta'+p}\eta_2}, c_0 \left(\frac{N}{\log N}\right)^{-\frac{1}{\alpha}}\right\} \tag{20a}$$

$$y_n = \text{the solution of } x_n^{\frac{\beta'+\max\{\alpha-\beta,p\}}{p}} y^{\frac{1-\gamma'}{q}} = N^{\eta_2}, \quad n \geqslant 0 \tag{20b}$$

$$x_{n+1} = \text{the solution of } x^{\frac{\beta-\beta'}{p}} y_n^{\frac{\gamma'-\gamma}{q}} = N^{\eta_1}, \quad n \geqslant 0. \tag{20c}$$

We first derive an explicit recursive formula for $\{x_n\}$.

Lemma C.1 *Let $u = \frac{\beta'+\max\{\alpha-\beta,p\}}{\beta-\beta'} \frac{\gamma'-\gamma}{1-\gamma'} > 0$, then*

(1). if $u > 1$, then

$$N^{-\frac{p}{\beta+p}} x_{n+1} = \left(N^{-\frac{p}{\beta+p}} x_n\right)^u.$$

(2). if $u < 1$, then

$$x_{n+1} = x_n^u.$$

***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 (20b) and (20c) that

$$\begin{aligned} x_{n+1} &= N^{\frac{p}{\max\{\alpha,\beta+p\}}} y_n^{-\frac{\gamma'-\gamma}{q}\frac{p}{\beta-\beta'}} \\ &= N^{\frac{p}{\max\{\alpha,\beta+p\}}} \left(N^{\eta_2} x_n^{-\frac{\beta'+\max\{\alpha-\beta,p\}}{p}} \right)^{-\frac{\gamma'-\gamma}{1-\gamma'}\frac{p}{\beta-\beta'}} \\ &= N^{\frac{p}{\max\{\alpha,\beta+p\}}} \left(N^{-\frac{p}{\max\{\alpha,\beta+p\}}} x_n \right)^u . \end{aligned}$$

(2). Suppose that $u < 1$, then we have $\eta_1 = \frac{\gamma'-\gamma}{1-\gamma}$ and $\eta_2 = \frac{1-\gamma'}{1-\gamma}$, so that $\frac{\eta_1}{\eta_2} = \frac{\gamma'-\gamma}{1-\gamma'}$, and it follows from (20b) and (20c) that $x_n^{\frac{\beta'+\max\{\alpha-\beta,p\}}{p}\frac{\gamma'-\gamma}{1-\gamma'}} = x_{n+1}^{\frac{\beta-\beta'}{p}}$, 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 $L_N = \mathcal{O}(\log\log N)$ such that $x_n \leqslant 2$ for all $n \geqslant L_N$.

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

$$\hat{\mathcal{A}}_{\mathtt{ml}} = \sum_{i=0}^{L_N} \left(\sum_{y_{i-1} \leqslant j < y_i} \rho_j^{\frac{1}{2}} f_j \otimes \rho_j^{\frac{1}{2}} f_j \right) \hat{\mathcal{C}}_{YX} \left(\hat{\mathcal{C}}_{KK} + \lambda_i^{(K)} I \right)^{-1} \tag{21}$$

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 $\hat{\mathcal{A}}_{\mathtt{ml}}$ can achieve minimax optimal convergence rate.

Theorem C.1 *Consider the estimator $\hat{\mathcal{A}}_{\mathtt{ml}}$ defined by (4). Suppose that Assumptions 2.1 to 2.5 hold, then there exists a universal constant C, such that*

$$\left\| \hat{\mathcal{A}}_{\mathtt{ml}} - \mathcal{A}_0 \right\|_{\beta',\gamma'}^2 \leqslant C\tau^2 \left(\frac{N}{\log N} \right)^{-\min\left\{ \frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{\gamma'-\gamma}{1-\gamma} \right\}} \log^2 N$$

holds with probability $\geqslant 1 - e^{-\tau}$.

***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

$$\left\| \hat{\mathcal{A}}_{\mathtt{ml}} - \mathcal{A}_0 \right\|_{\beta',\gamma'} \leqslant \left\| \hat{\mathcal{A}}_{\mathtt{ml}} - \hat{\mathcal{A}}_{\mathtt{ml}}^{\lambda} \right\|_{\beta',\gamma'} + \left\| \hat{\mathcal{A}}_{\mathtt{ml}}^{\lambda} - \mathcal{A}_0 \right\|_{\beta',\gamma'}$$

where

$$\hat{\mathcal{A}}_{\mathtt{ml}}^{\lambda} = \sum_{i=0}^{L_N} \left(\sum_{y_i \leqslant j < y_{i+1}} \rho_j^{\frac{1}{2}} f_j \otimes \rho_j^{\frac{1}{2}} f_j \right) \mathcal{C}_{YX} \left(\mathcal{C}_{KK} + \lambda_i^{(K)} I \right)^{-1} . \tag{22}$$

Bounding the bias term. Since $\|\mathcal{A}_0\|_{\beta,\gamma} \leqslant B$, we can write

$$\mathcal{A}_0 := \sum_{i=1}^{+\infty} \sum_{j=1}^{+\infty} a_{ij} \mu_i^{\frac{\beta}{2}} \rho_j^{1-\frac{\gamma}{2}} f_j \otimes e_i$$

where the coefficient matrix $A_0 = (a_{ij})_{1 \leqslant i,j \leqslant +\infty}$ satisfies $\|A_0\|_F^2 \leqslant B^2$. We fix $(i,j) \in \mathbb{Z}_+^2$ and assume WLOG that $y_{m_j - 1} \leqslant j < y_{m_j}$ for some $m \geqslant 0$, where $y_{L_N+1} = +\infty$. It follows from (22) that

$$\begin{aligned} \left\langle \rho_j^{\frac{1}{2}} f_j, \hat{\mathcal{A}}_{\mathtt{ml}}^{\lambda} \mu_i^{\frac{1}{2}} e_i \right\rangle &= \sum_{k=0}^{L_N} \left\langle \left(\sum_{y_{k-1} \leqslant j < y_k} \rho_j^{\frac{1}{2}} f_j \otimes \rho_j^{\frac{1}{2}} f_j \right) \rho_j^{\frac{1}{2}} f_j, \mathcal{C}_{YX} \left(\mathcal{C}_{KK} + \lambda_k^{(K)} I \right)^{-1} \mu_i^{\frac{1}{2}} e_i \right\rangle \\ &= \frac{\mu_i}{\mu_i + \lambda_m^{(K)}} \rho_j^{\frac{1-\gamma}{2}} \mu_i^{-\frac{1-\beta}{2}} a_{ij} . \end{aligned}$$

Thus

$$\begin{aligned}\left\|\mathcal{A}_0-\hat{\mathcal{A}}^{\lambda}_{\mathtt{ml}}\right\|^2_{\beta',\gamma'} &= \left\|\mathcal{C}_{Q_L}^{-\frac{1-\gamma}{2}}\left(\hat{\mathcal{A}}^{\lambda}_{\mathtt{ml}}-\mathcal{A}_0\right)\mathcal{C}_{KK}^{\frac{1-\beta'}{2}}\right\|^2_{\mathrm{HS}}\\
&= \sum_{i,j=1}^{+\infty}\left\langle \rho_j^{\frac{1}{2}} f_j, \mathcal{C}_{Q_L}^{-\frac{1-\gamma'}{2}}\left(\hat{\mathcal{A}}^{\lambda}_{\mathtt{ml}}-\mathcal{A}_0\right)\mathcal{C}_{KK}^{\frac{1-\beta'}{2}}\mu_i^{\frac{1}{2}} e_i\right\rangle^2\\
&= \sum_{i,j=1}^{+\infty}\left(\frac{\lambda^{(K)}_{m_j}}{\mu_i+\lambda^{(K)}_{m_j}}\right)^2 \mu_i^{\beta-\beta'}\rho_j^{\gamma'-\gamma}a_{ij}^2\\
&= \sum_{j=1}^{+\infty}\rho_j^{\gamma'-\gamma}\left(\sum_{i=1}^{+\infty}a_{ij}^2\right)\max_{i\geqslant 1}\mu_i^{\beta-\beta'}\left(\frac{\lambda^{(K)}_{m_j}}{\mu_i+\lambda^{(K)}_{m_j}}\right)^2\\
&\lesssim \sum_{j=1}^{+\infty}\rho_j^{\gamma'-\gamma}\left(\lambda^{(K)}_{m_j}\right)^{\beta-\beta'}\left(\sum_{i=1}^{+\infty}a_{ij}^2\right)\lesssim B^2\max_{j\geqslant 1}\rho_j^{\gamma'-\gamma}\left(\lambda^{(K)}_{m_j}\right)^{\beta-\beta'}\\
&\leqslant B^2\max_{j\geqslant 1}\rho_j^{\gamma'-\gamma}x_{m_j}^{-\frac{\beta-\beta'}{p}}\lesssim B^2\max_{j\geqslant 1}j^{-\frac{\gamma'-\gamma}{q}}x_{m_j}^{-\frac{\beta-\beta'}{p}}\\
&\leqslant B^2 y_{m_j-1}^{-\frac{\gamma'-\gamma}{q}}x_{m_j}^{-\frac{\beta-\beta'}{p}}\lesssim N^{-\eta_1}\end{aligned} \tag{23}$$

where we recall that $\eta_1=\min\left\{\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}},\frac{\gamma'-\gamma}{1-\gamma}\right\}$ and the last step follows from (20c).

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

$$\begin{aligned}\mathcal{V} &= \left\|\hat{\mathcal{A}}_{\mathtt{ml}}-\mathcal{A}^{\lambda}_{\mathtt{ml}}\right\|^2_{\beta',\gamma'}\\
&= \left\|C_{Q_L}^{-\frac{1-\gamma'}{2}}\left(\hat{\mathcal{A}}_{\mathtt{ml}}-\mathcal{A}^{\lambda}_{\mathtt{ml}}\right)C_{KK}^{\frac{1-\beta'}{2}}\right\|^2_{\mathrm{HS}}\\
&= \sum_{i,j=1}^{+\infty}\left\langle \rho_j^{\frac{1}{2}} f_j, C_{Q_L}^{-\frac{1-\gamma'}{2}}\left(\hat{\mathcal{A}}_{\mathtt{ml}}-\mathcal{A}^{\lambda}_{\mathtt{ml}}\right)C_{KK}^{\frac{1-\beta'}{2}}\mu_i^{\frac{1}{2}} e_i\right\rangle^2\\
&= \sum_{j=1}^{z_N}\rho_j^{-(1-\gamma')}\sum_{i=1}^{+\infty}\left\langle \rho_j^{\frac{1}{2}} f_j, \left[\hat{\mathcal{C}}_{YX}\left(\hat{\mathcal{C}}_{KK}+\lambda_{m_j}I\right)^{-1}-\mathcal{C}_{YX}\left(\mathcal{C}_{KK}+\lambda_{m_j}I\right)^{-1}\right]\mu_i^{1-\frac{\beta'}{2}}e_i\right\rangle^2\\
&= \sum_{j=1}^{z_N}\rho_j^{-(1-\gamma')}\sum_{i=1}^{+\infty}\Bigg\langle\underbrace{\left(\mathcal{C}_{KK}+\lambda_{m_j}I\right)^{-\frac{1}{2}}\left[\hat{\mathcal{C}}_{KL}-\left(\hat{\mathcal{C}}_{KK}+\lambda_{m_j}I\right)\left(\mathcal{C}_{KK}+\lambda_{m_j}I\right)^{-1}\mathcal{C}_{KL}\right]}_{=:U_{m_j}}\rho_j^{\frac{1}{2}} f_j,\\
&\qquad\underbrace{\left(\mathcal{C}_{KK}+\lambda_{m_j}I\right)^{\frac{1}{2}}\left(\hat{\mathcal{C}}_{KK}+\lambda_{m_j}I\right)^{-1}\left(\mathcal{C}_{KK}+\lambda_{m_j}I\right)^{\frac{1}{2}}}_{=:G_{m_j}}\frac{\mu_i^{1-\frac{\beta'}{2}}}{\sqrt{\mu_i+\lambda_j}}e_i\Bigg\rangle^2\\
&= \sum_{j=1}^{z_N}\rho_j^{-(1-\gamma')}\left\langle U_{m_j}\rho_j^{\frac{1}{2}} f_j, G_{m_j}\left(\sum_{i=1}^{+\infty}\frac{\mu_i^{2-\beta'}}{\mu_i+\lambda_{m_j}}e_i\otimes e_i\right)G_{m_j}U_{m_j}\rho_j^{\frac{1}{2}} f_j\right\rangle\\
&\lesssim \sum_{j=1}^{z_N}j^{\frac{1-\gamma'}{q}}\|G_{m_j}\|^2\lambda_{m_j}^{-\beta'}\left\|U_{m_j}\rho_j^{\frac{1}{2}} f_j\right\|^2\end{aligned}$$

for reasons similar to (11). It now remains to bound $\left\|G_{m_j}\right\|$ and $\left\|U_{m_j}\rho_j^{\frac{1}{2}} f_j\right\|$ for $1\leqslant j\leqslant L_N$. Note that these quantities have already been bounded in Appendix B.1.2 with λ_{m_j} replaced with λ_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

$$\mathcal{V}\leqslant Ca^2\frac{1}{N}\max_{1\leqslant j\leqslant L_N} j^{\frac{1-\gamma'}{q}}\lambda_{m_j}^{-\left(\beta'+\max\{\alpha-\beta,p\}\right)}$$

with probability $\geqslant 1 - Ne^{-a}$. Since $j \leqslant y_{m_j}$, by (20b) we have

$$j^{\frac{1-\gamma'}{q}} \lambda_{m_j}^{-\left(\beta'+\max\{\alpha-\beta,p\}\right)} \lesssim y_{m_j}^{\frac{1-\gamma'}{q}} x_{m_j}^{\frac{\beta'+\max\{\alpha-\beta,p\}}{p}} = N^{\eta_2}.$$

Hence

$$\mathcal{V} \lesssim \frac{1}{N} \max_{1 \leqslant j \leqslant L_N} j^{\frac{1-\gamma'}{q}} \lambda_{m_j}^{-\left(\beta'+\max\{\alpha-\beta,p\}\right)} \leqslant N^{\eta_2 - 1} = N^{\eta_1}.$$

Combining the bias and variance bounds, the conclusion directly follows. □

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 $\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}} = \frac{\gamma'-\gamma}{1-\gamma}$. 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:

$$\begin{aligned} x_0 &= \max\left\{\frac{1}{2} N^{\frac{p}{\beta'+p}\eta_2}, c_0 \left(\frac{N}{\log N}\right)^{-\frac{1}{\alpha}}\right\} \\ x_n &= \frac{1}{2} x_{n-1} \\ y_n &= \text{the solution of } x_n^{\frac{\beta-\beta'}{p}} y_n^{\frac{\gamma'-\gamma}{q}} = N^{\eta_1}, \end{aligned} \tag{24}$$

where we recall that $\eta_1 = \frac{\beta-\beta'}{\max\{\alpha,\beta+p\}} = \frac{\gamma'-\gamma}{1-\gamma}$. In this case, there exists $L_N = \mathcal{O}(\ln N)$ such that $x_n < 1$ for all $n \geqslant L_N$. Let $\lambda_i^{(K)} = x_i^{-\frac{1}{p}}$, then we construct the following estimator:

$$\hat{\mathcal{A}}_{\mathtt{ml}}^{\lambda} = \sum_{i=0}^{L_N} \left(\sum_{y_{i-1} \leqslant j < y_i} \rho_j^{\frac{1}{2}} f_j \otimes \rho_j^{\frac{1}{2}} f_j \right) \hat{\mathcal{C}}_{LK} \left(\hat{\mathcal{C}}_{KK} + \lambda_i^{(K)} I\right)^{-1}. \tag{25}$$

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

Theorem C.2 *Consider the estimator $\hat{\mathcal{A}}_{\mathtt{ml}}$ defined by (25). Suppose that Assumptions 2.1 to 2.5 hold, then there exists a universal constant C, such that*

$$\left\|\hat{\mathcal{A}}_{\mathtt{ml}} - \mathcal{A}_0\right\|_{\beta',\gamma'}^2 \leqslant C\tau^2 \left(\frac{N}{\log N}\right)^{-\min\left\{\frac{\beta-\beta'}{\max\{\alpha,\beta+p\}}, \frac{\gamma'-\gamma}{1-\gamma}\right\}} \log^2 N$$

holds with probability $\geqslant 1 - e^{-\tau}$.

***Proof*:** The proof of Theorem C.2 is similar to that of Theorems 4.1 and 4.2. We consider the bias-variance decomposition

$$\left\|\hat{\mathcal{A}}_{\mathtt{ml}} - \mathcal{A}_0\right\|_{\beta',\gamma'} \leqslant \left\|\hat{\mathcal{A}}_{m1} - \hat{\mathcal{A}}_{m1}^{\lambda}\right\|_{\beta',\gamma'} + \left\|\hat{\mathcal{A}}_{m1}^{\lambda} - \mathcal{A}_0\right\|_{\beta',\gamma'}$$

where

$$\mathcal{A}_{\mathtt{ml}}^{\lambda} = \sum_{i=0}^{L_N} \left(\sum_{y_{i-1} \leqslant j < y_i} \rho_j^{\frac{1}{2}} f_j \otimes \rho_j^{\frac{1}{2}} f_j \right) \mathcal{C}_{LK} \left(\mathcal{C}_{KK} + \lambda_i^{(K)} I\right)^{-1}. \tag{26}$$

as defined in (25).

Bounding the bias term. Let $\mathcal{A}_0 := \sum_{i=1}^{+\infty} \sum_{j=1}^{+\infty} a_{ij} \mu_i^{\frac{\beta}{2}} \rho_j^{1-\frac{\gamma}{2}} f_j \otimes e_i$ with coefficient matrix $A_0 = (a_{ij})_{i,j=1}^{+\infty}$ such that $\|A_0\|_F^2 \leqslant B^2$. We fix $(i,j) \in \mathbb{Z}_+^2$ and assume WLOG that $y_{m_j-1} \leqslant j < y_{m_j}$ for some $m_j \geqslant 0$, where $y_{L_N+1} = +\infty$. It follows from (26) that

$$\left\langle \rho_j^{\frac{1}{2}} f_j, \mathcal{A}_{\mathtt{ml}}^{\lambda} \mu_i^{\frac{1}{2}} e_i \right\rangle = \frac{\mu_i}{\mu_i + \lambda_{m_j}^{(K)}} \rho_j^{\frac{1-\gamma}{2}} \mu_i^{-\frac{1-\beta}{2}} a_{ij}.$$

Thus we can proceed as in (23) to deduce that

$$\begin{aligned}\left\|\mathcal{A}_0 - \mathcal{A}^{\lambda}_{\mathtt{ml}}\right\|^2_{\beta',\gamma'} &\leqslant \max_{j\geqslant 1} \rho_j^{\gamma'-\gamma}\left(\lambda^{(K)}_{m_j}\right)^{\beta-\beta'}\\ &\lesssim \max\left\{\max_{1\leqslant j\leqslant L_N} j^{-\frac{\gamma'-\gamma}{q}} x_{m_j}^{-\frac{\beta-\beta'}{p}}, y_{L_N}^{-\frac{\gamma'-\gamma}{q}}\right\}\\ &\leqslant \max\left\{\max_{1\leqslant j\leqslant L_N} y_{m_j-1}^{-\frac{\gamma'-\gamma}{q}} x_{m_j}^{-\frac{\beta-\beta'}{p}}, y_{L_N}^{-\frac{\gamma'-\gamma}{q}}\right\}\end{aligned}$$

The definition (24) implies that

$$y_{m_j-1}^{-\frac{\gamma'-\gamma}{q}} x_{m_j}^{-\frac{\beta-\beta'}{p}} \leqslant 2^{\frac{\beta-\beta'}{p}} y_{m_j-1}^{-\frac{\gamma'-\gamma}{q}} x_{m_j-1}^{-\frac{\beta-\beta'}{p}} \leqslant 2^{\frac{\beta-\beta'}{p}} N^{-\eta_1}.$$

On the other hand, since $x_{L_N} < 1$, by (24) implies that $y_{L_N}^{-\frac{\gamma'-\gamma}{q}} \lesssim N^{-\eta_1}$. Therefore, for the bias term $\left\|\mathcal{A}_0 - \mathcal{A}^{\lambda}_{\mathtt{ml}}\right\|^2_{\beta',\gamma'} \lesssim N^{-\eta_1}$.

Bounding the variance term. Repeating the arguments in (24), we can deduce that there exists a constant $C > 0$ such that

$$\mathcal{V} \leqslant Ca^2\frac{1}{N}\max_{1\leqslant j\leqslant L_N} j^{\frac{1-\gamma'}{q}}\lambda_{m_j}^{-\left(\beta'+\max\{\alpha-\beta,p\}\right)} \leqslant Ca^2\frac{1}{N}\max_{1\leqslant j\leqslant L_N} y_{m_j}^{\frac{1-\gamma'}{q}} x_{m_j}^{\frac{\left(\beta'+\max\{\alpha-\beta,p\}\right)}{p}} \lesssim N^{-\eta_1}$$

with probability $\geqslant 1 - Ne^{-a}$.

Combining the bias and variance bounds, we arrive at the desired conclusion. □

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

D Auxiliary results

Lemma D.1 *We have* $\|T\|_{\beta,\gamma} = \left\|\mathcal{C}_{Q_L}^{-(1-\gamma)/2}\circ T\circ \mathcal{C}_{KK}^{(1-\beta)/2}\right\|_{\mathrm{HS}(\mathcal{H}_K,\mathcal{H}_L)}$.

***Proof*:** We recall from the definition that $\|T\|_{\beta,\gamma} = \left\|(I_{1,\gamma,Q_L})^{\dagger}\circ T\circ I^*_{1\beta,P_K}\right\|_{\mathrm{HS}(\mathcal{H}_K^\beta,\mathcal{H}_L^\gamma)}$, so that

$$\begin{aligned}\|T\|^2_{\beta,\gamma} &= \left\|(I_{1,\gamma,Q_L})^{\dagger}\circ T\circ I^*_{1\beta,P_K}\right\|^2_{\mathrm{HS}(\mathcal{H}_K^\beta,\mathcal{H}_L^\gamma)}\\ &= \sum_{i,j=1}^{+\infty}\left\langle \rho_j^{\frac{\gamma}{2}} f_j, \left(I^*_{1,\gamma,Q_L}\right)^{\dagger}\circ T\circ I^*_{1,\beta,P_K}\mu_i^{\frac{\beta}{2}} e_i\right\rangle^2_{\mathcal{H}_L^\gamma}\\ &= \sum_{i,j=1}^{+\infty}\left\langle \rho_j^{\frac{\gamma}{2}} f_j, \left(I^*_{1,\gamma,Q_L}\right)^{\dagger}\circ T\mu_i^{1-\frac{\beta}{2}} e_i\right\rangle^2_{\mathcal{H}_L^\gamma}\\ &= \sum_{i,j=1}^{+\infty}\left\langle \rho_j^{\frac{\gamma}{2}} f_j, T\mu_i^{1-\frac{\beta}{2}} e_i\right\rangle^2_{\mathcal{H}_L}\\ &= \sum_{i,j=1}^{+\infty}\left\langle \rho_j^{\frac{1}{2}} f_j, \mathcal{C}_{Q_L}^{-(1-\gamma)/2}\circ T\circ \mathcal{C}_{KK}^{(1-\beta)/2}\mu_i^{\frac{1}{2}} e_i\right\rangle^2_{\mathcal{H}_L}\\ &= \left\|\mathcal{C}_{Q_L}^{-(1-\gamma)/2}\circ T\circ \mathcal{C}_{KK}^{(1-\beta)/2}\right\|^2_{\mathrm{HS}}\end{aligned}$$

as desired. □

Lemma D.2 *Under Assumption 2.2, we have*

$$\left\|\left(C_{P_K}+\lambda_X I\right)^{-\frac{1}{2}} u\right\| \leqslant \lambda_X^{-\frac{\alpha}{2}}\cdot A_1 \quad P_K\text{-}a.s.$$

***Proof*:** By Assumption 2.2 we have $\left\|C_{P_K}^{-\frac{1-\alpha}{2}} u\right\|_{\mathcal{H}_K} \leqslant A_1$, so that

$$\left\|(C_{P_K} + \lambda_X I)^{-\frac{1}{2}} u\right\| \leqslant \left\|(C_{P_K} + \lambda_X I)^{-\frac{\alpha}{2}}\right\| \cdot \left\|C_{P_K}^{-\frac{1-\alpha}{2}} u\right\| \leqslant \lambda_X^{-\frac{\alpha}{2}} \cdot A_1$$

as desired. □

D.1 Concentration inequalities

Theorem D.1 *[19, Theorem 27] Let $(\Omega, \mathcal{B}, P)$ be a probability space, $\mathcal{H}$ be a separable Hilbert space and $X : \Omega \to \mathrm{HS}(H; H)$ be a random variable with self-adjoint values. Furthermore, assume that $\|X\|_F \leqslant B, P - a.s.$ and V be a positive semi-definite matrix with $\mathbb{E}_P\left(X^2\right) \preccurlyeq V$, i.e. $V - \mathbb{E}_P\left(X^2\right)$ is positive semi-definite. Then, for $g(V) := \log\left(2e\,\mathrm{tr}(V)\|V\|^{-1}\right), \tau \geqslant 1$, and $n \geqslant 1$, the following concentration inequality is satisfied*

$$P^n\left((\omega_1, \ldots, \omega_n) \in \Omega^n : \left\|\frac{1}{n}\sum_{i=1}^n X(\omega_i) - \mathbb{E}_P X(\omega)\right\| \geqslant \frac{4\tau B g(V)}{3n} + \sqrt{\frac{2\tau\|V\|g(V)}{n}}\right) \leqslant 2e^{-\tau}.$$

Theorem D.2 *[19, 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*

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

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

$$P^n\left((\omega_1, \ldots, \omega_n) \in \Omega^n : \left\|\frac{1}{n}\sum_{i=1}^n \xi(\omega_i) - \mathbb{E}_P\xi\right\|_H^2 \geqslant 32\frac{\tau^2}{n}\left(\sigma^2 + \frac{L^2}{n}\right)\right) \leqslant 2e^{-\tau}$$

The following theorem shows that the regularized covariance $\mathcal{C}_{KK} + \lambda I$ can be estimated with small error when λ is above a certain threshold. Although it is well-known [19, 21], we still recall it below for completeness.

Theorem D.3 *Recall that $\mathcal{C}_{KK} = \mathbb{E}_{P_K} u \otimes u$ and $\hat{\mathcal{C}}_{KK} = \frac{1}{N}\sum_{i=1}^N u_i \otimes u_i$ where $u_i \overset{\text{i.i.d.}}{\sim} P_K$. Suppose that Assumption 2.2 holds and $N \gtrsim A_1^2 \tau g_\lambda \lambda^{-\alpha}$, where $g_\lambda = \log\left(2e\mathcal{N}_{P_K}(\lambda)\frac{\|\mathcal{C}_{KK}\|+\lambda}{\|\mathcal{C}_{KK}\|}\right)$ and $\mathcal{N}_{P_K}(\lambda) = \mathrm{tr}\left((\mathcal{C}_{KK} + \lambda I)^{-1}\mathcal{C}_{KK}\right)$ is the effective dimension, then with probability at least $1 - e^{-\tau}$, we have*

$$\left\|(\mathcal{C}_{KK} + \lambda I)^{-\frac{1}{2}}\left(\mathcal{C}_{KK} - \hat{\mathcal{C}}_{KK}\right)(\mathcal{C}_{KK} + \lambda i)^{-\frac{1}{2}}\right\| \lesssim \sqrt{\frac{A_1^2 \tau g_\lambda}{N\lambda^\alpha}} \leqslant 0.1. \tag{27}$$

***Proof*:** Let $X(u) = (\mathcal{C}_{KK} + \lambda I)^{-\frac{1}{2}} u \otimes u (\mathcal{C}_{KK} + \lambda I)^{-\frac{1}{2}}$ where $u \in \mathcal{H}_K$, then the LHS of (27) can be expressed as $\left\|\frac{1}{N}\sum_{i=1}^N X(u_i) - \mathbb{E}_{u\sim P_K} X(u)\right\|$. We hope to apply Theorem D.1 and start with verifying the assumptions.

Since $\mathbb{E}_{P_K} X = (\mathcal{C}_{KK} + \lambda I)^{-\frac{1}{2}} \mathcal{C}_{KK} (\mathcal{C}_{KK} + \lambda I)^{-\frac{1}{2}}$ and $\|X(u)\| = \|X(u)\|_F = \left\|(\mathcal{C}_{KK} + \lambda I)^{-\frac{1}{2}} u\right\|^2 \leqslant A_1^2\left\|(\mathcal{C}_{KK} + \lambda I)^{-\frac{1-\alpha}{2}}\right\|^2 \lesssim A_1^2\lambda^{-\alpha}$, so that there exists $V = \mathcal{O}\left(\lambda^{-\alpha}(\mathcal{C}_{KK} + \lambda I)^{-\frac{1}{2}}\mathcal{C}_{KK}(\mathcal{C}_{KK} + \lambda I)^{-\frac{1}{2}}\right)$ such that $\mathbb{E}_{P_K} X^2 \preccurlyeq V$. It's easy to see that $\|V\| \lesssim \lambda^{-\alpha}$ and $\mathrm{tr}(V) \lesssim \mathcal{N}_{P_K}(\lambda)$. The conclusion then follows from Theorem D.1 with $B = \tilde{\mathcal{O}}(\lambda^{-\alpha})$ and $g(V) = g_\lambda$. □

Corollary D.1 *Under the notations and assumptions of Theorem D.3, there exists a constant $C_1 > 0$ with probability $\geqslant 1 - e^{-\tau}$ we have*

$$\left\|(\mathcal{C}_{KK} + \lambda I)^{\frac{1}{2}}\left(\hat{\mathcal{C}}_{KK} + \lambda I\right)^{-1}(\mathcal{C}_{KK} + \lambda I)^{\frac{1}{2}}\right\| \leqslant C_1. \tag{28}$$

***Proof*:** By Theorem D.3 we have

$$\begin{aligned}
&\left\|(\mathcal{C}_{KK}+\lambda I)^{\frac{1}{2}}\left(\hat{\mathcal{C}}_{KK}-\mathcal{C}_{KK}\right)^{-1}(\mathcal{C}_{KK}+\lambda I)^{\frac{1}{2}}\right\| \\
=&\left\|\left(I-(\mathcal{C}_{KK}+\lambda I)^{-\frac{1}{2}}(\mathcal{C}_{KK}-\hat{\mathcal{C}}_{KK})(\mathcal{C}_{KK}+\lambda I)^{-\frac{1}{2}}\right)^{-1}\right\| \\
\leqslant& \, 2
\end{aligned}$$

with probability $\geqslant 1-e^{-\lambda}$, as desired. □