An adaptive RKHS regularization for Fredholm integral equations

Fei Lu and Miao-Jung Yvonne Ou [1]

Abstract

Regularization is a long-standing challenge for ill-posed linear inverse problems, and a prototype is the Fredholm integral equation of the first kind with additive Gaussian measurement noise. We introduce a new RKHS regularization adaptive to measurement data and the underlying linear operator. This RKHS arises naturally in a variational approach, and its closure is the function space in which we can identify the true solution. Also, we introduce a small noise analysis to compare regularization norms by sharp convergence rates in the small noise limit. Our analysis shows that the RKHS- and L^2-regularizers yield the same convergence rate when their optimal hyper-parameters are selected using the true solution, and the RKHS-regularizer has a smaller multiplicative constant. However, in computational practice, the RKHS regularizer significantly outperforms the L^2-and l^2-regularizers in producing consistently converging estimators when the noise level decays or the observation mesh refines.

1 Introduction

We consider the inverse problem of recovering the input function in the Fredholm integral equation of the first kind from *discrete noisy* output. Specifically, let $\mathcal{S} \subset \mathbb{R}$ and $\mathcal{T} \subset \mathbb{R}$ be two compact sets. For a given integral kernel K, we aim to recover the function $\phi : \mathcal{S} \to \mathbb{R}$ from a discrete noisy data $\mathbf{y} = \{y(t_i), t_0 < t_1 < \cdots < t_m\}$, which is generated from

$$y(t) = \int_{\mathcal{S}} K(t,s)\phi(s)\nu(ds) + \sigma\dot{W}(t) =: L_K\phi(t) + \sigma\dot{W}(t). \tag{1.1}$$

Here ν is a finite measure on $\mathcal{S}$, and it is suitable for both continuous and discrete models: it is the Lebesgue measure when $\mathcal{S}$ is an interval and an atom measure when $\mathcal{S}$ has finitely many elements. Similarly, we define μ to be an atomic measure with $\mu(t_i) = t_{i+1} - t_i$, which can be viewed as the discrete approximation of the Lebesgue measure. The integral operator L_K in (1.1) maps from $L^2_\nu(\mathcal{S})$ to $L^2_\mu(\mathcal{T})$. The given integral kernel $K : \mathcal{T} \times \mathcal{S} \to \mathbb{R}$ is assumed to be continuous. A typical example is

$$K(t,s) = s^{-2}e^{-st}$$

with $\mathcal{S} \subset [a,b]$ and $\mathcal{T} \subset [0,T]$ for some $b > a > 0$ and $T > 0$, which arises from the magnetic resonance relaxometry (MRR) [3]; see Section 4 for more details.

The measurement noise $\sigma\dot{W}(t)$ is the white noise; that is, the noise at t_i has a Gaussian distribution $\mathcal{N}(0, \sigma^2(t_{i+1} - t_i))$ for each i. We note that such a noise is integrable when the observation mesh refines, i.e., $\max_i(t_{i+1} - t_i)$ vanishes.

Eq.(1.1) is a prototype of ill-posed inverse problems, dating back to Hadamard [8], and it remains a testbed for new regularization methods [9, 12, 20]. It is ill-posed in $L^2_\nu(\mathcal{S})$ in the sense that the least squares solution with mini-norm is sensitive to measurement noise in data [16]. In other words, in the variational formulation of the inverse problem, the *minimal norm minimizer* of a loss functional

$$\mathcal{E}(\phi) := \|L_K\phi - \mathbf{y}\|^2_{L^2_\mu(\mathcal{T})} \tag{1.2}$$

[1]FL: Department of Mathematics, Johns Hopkins University; feilu@math.jhu.edu
MYO: Department of Mathematical Science, University of Delaware; mou@udel.edu

is sensitive to the data $\mathbf{y}$. Such an ill-posedness roots in the fact that the operator $L : L^2_\nu(\mathcal{S}) \to L^2_\mu(\mathcal{T})$ is compact with eigenvalues converging to zero.

Regularization, necessary to avoid amplifying the measurement noise, is a long-standing challenge for ill-posed inverse problems, and there are numerous regularization methods. Roughly speaking, in the variational formulation, a regularization method restricts the subset or subspace of search either by constraints on ϕ or the loss functional (e.g., minimizing $\mathcal{E}(\phi)$ subject to $\|\phi\|_\square \leqslant \alpha_0$ or minimizing $\|\phi\|_\square$ subject to $\mathcal{E}(\phi) \leqslant \alpha_1$); or by adding a penalty to the loss functional $\mathcal{E}$, e.g., the well-known Tikhonov regulation:

$$\phi_\lambda = \arg\min_\phi \mathcal{E}_\lambda(\phi) \text{ with } \mathcal{E}_\lambda(\phi) := \mathcal{E}(\phi) + \lambda\|\phi\|_\square^2. \tag{1.3}$$

Here $\|\phi\|_\square$ is a certain regularization norm and $(\alpha_0, \alpha_1, \lambda)$ are the hyper-parameters. There is tremendous effort in selecting the norm and the hyper-parameter. Once the norm is chosen, the minimization can be solved by either direct or iterative methods. The direct methods solve the linear equations by canonical matrix decompositions, and the selection of λ has been thoroughly studied in [9, 21]. In particular, the L-curve method in [10] is a widely used robust method. The iterative methods are flexible for high-dimensional problems, and we refer to [5, 7, 22] for recent developments.

In this paper, we focus on the selection of the regularization norm. In particular, we introduce a new data-adaptive Reproducing Kernel Hilbert Space (DA-RKHS) norm and compare it with the widely used L^2-norm for regularization.

The new RKHS-regularization is inspired by the data-adaptive RKHS regularization for learning the kernels in operators [4, 11, 13, 14], and arises naturally in the variational approach. The closure of this RKHS is referred to as the *function space of identifiability* (FSOI), where a unique solution can be identified. Specifically, we introduce an exploration measure ρ on $\mathcal{S}$ to quantify how the kernel explores the function $\phi(s)$ at different s. The variational inverse problem is then formulated in $L^2_\rho(\mathcal{S})$.

The RKHS norm will result in a specific regularization norm $\|\phi\|_\square$ that restricts the minimization search inside the RKHS, and hence prevents the errors outside the FSOI from contaminating the regularized solution.

Also, we present *sharp convergence rates* for the RKHS- and L^2-regularized estimators in the small noise limit for the cases where the spectrum of the operator of inversion converges to 0 either exponentially or polynomially. These convergence rates are sharp in the sense that these rates come with explicit multiplicative constants; see Theorem 3.1. These sharp estimates enable a theoretical comparison between the two regularizers: the RKHS- and L^2-regularized estimators both converge linearly, but the RKHS-regularizer outperforms the L^2-regularizer with a smaller multiplicative factor.

Furthermore, we implement the new RKHS regularization in a matrix-decomposition-based algorithm (see Sec.4), which is a variant of DARTR in [14]. This algorithm uses the L-curve method in [10] to select the optimal hyper-parameter in the Tikhonov regularization. Numerical results show that the RKHS-regularizer consistently outperforms the l^2- and L^2-regularizer when either the noise level decays or the observation mesh $\max_i(t_{i+1} - t_i)$ refines. In particular, the RKHS norm results in better L-shaped curves than the other two norms, indicating that the RKHS norm provides a better metric on the function space for the success of the L-curve method. The MATLAB code for our numerical tests is available at https://github.com/LearnDynamics/dartr_Fredholm.git.

There are various regularization norms, including the Euclidean norms (e.g., [9, 19]), the total variation norm $\|\phi'\|_{L^1}$ in Rudin–Osher–Fatemi method in [17], the L^1 norm $\|\phi\|_{L^1}$ in LASSO (e.g., [18]), and the RKHS-norm with a user-specified reproducing kernel R (e.g., [2, 6, 21]) with tuned hyper-parameters for the reproducing kernel. However, these norms are based on pre-assumed properties of the solution and are generic without considering the specific inverse problem. In contrast, our RKHS norm is adaptive to the operator and the observation mesh, our reproducing kernel is intrinsic to the inverse problem and it is implicitly determined without any hyper-parameter to be tuned.

1.1 Main contributions

This study has two main contributions.

- *Introducing a new adaptive RKHS regularization.* We introduce a new adaptive RKHS for Tikhonov regularization. This RKHS has a reproducing kernel determined by the integral kernel and data, and its closure is the linear subspace in which the inverse problem has a unique solution. Thus, the resulting RKHS regularization is adaptive. We propose an implementation algorithm based on matrix decomposition. Also, we numerically demonstrate that the RKHS regularization outperforms the commonly-used L^2 and l^2 regularization by yielding more accurate estimators with more consistent convergence when either the noise vanishes or the observation mesh refines.

- *Introducing a small noise analysis approach for comparing estimators using different regularization norms.* We introduce a small noise analysis to establish a theoretical comparison between two regularization norms: the RKHS and L^2 norms. The analysis shows that when the optimal λ is selected by minimizing the expected error, the convergence rates of the two are the same. However, the adaptive RKHS regularizer outperforms the L^2 regularizer with a smaller multiplicative constant.

To the best of our knowledge, this study is the first to use the adaptive RKHS for regularizing Fredholm-type inverse problems, and it is the first result comparing regularization norms through sharp convergence rates in the small noise limit. The adaptive RKHS (with a reproducing kernel determined by the linear model and data) for regularization applies to general linear machine learning and inverse problems. In particular, it provides an automatic reproducing kernel for the Gaussian process and kernel-based regression. Also, the comparative analysis in this study sets a theoretical framework for comparing regularization norms, which will lead to a better understanding of regularization and the selection of regularization norms for ill-posed linear inverse problems. The limitations of this study are discussed in Section 5.

The rest of the paper is organized as follows. We introduce the adaptive RKHS in Section 2, with a characterization between the RKHS and the function space of identifiability. Section 3 proves the convergence of the RKHS-regularized estimator, and Section 4 presents the algorithm and numerically demonstrates the robust convergence of the estimator when either the noise decays or the observation mesh refines. We discuss the limitations of this study in Section 5 and conclude with future developments of the adaptive RKHS regularization strategy in Section 6.

2 An adaptive RKHS for the inverse problem

In this section, we introduce the RKHS based on the variational formulation of the inverse problem. A core element is an identifiability theory specifying the function space in which the

loss functional has a unique minimizer. Importantly, if the true solution is in this space, the minimizer of the loss functional recovers the solution when there is no measurement noise in the data. Thus, we call this *function space of identifiability* (FSOI). For this FSOI, an RKHS norm is introduced and used as the regularization norm $\|\phi\|_{\square}$ in (1.3). As will be explained later in the section, this choice of penalty term ensures the search of the minimum of (1.3) takes place in the FSOI.

2.1 The function space of identifiability

We first introduce an ambient function space $L^2_\rho(\mathcal{S})$, where the measure ρ is defined as

$$\frac{d\rho}{d\nu}(s) := \frac{1}{Z} \int_{\mathcal{T}} |K(t,s)| \mu(dt), \ \forall s \in \mathcal{S}, \tag{2.1}$$

where $Z = \int_{\mathcal{S}} \int_{\mathcal{T}} |K(t,s)| \mu(dt)\nu(ds)$ is the normalizing constant. From (1.1), we see that ρ quantifies the exploration by the integral kernel K to the unknown input function through the output set $\mathcal{T}$, and hence is referred to as an *exploration measure*.

The major advantage of the space $L^2_\rho(\mathcal{S})$ over the original space $L^2_\nu(\mathcal{S})$ is that it is adaptive to the specific setting of the inverse problem. In particular, this weighted space takes into account the structure of the integral kernel and the data points in $\mathcal{T}$. It can provide better scaling and reduce the ill-conditioning in computation (see [11]). Thus, while the following identifiability theory can be carried out for both $L^2_\rho(\mathcal{S})$ and $L^2_\nu(\mathcal{S})$, we will focus only on $L^2_\rho(\mathcal{S})$.

We define the function space of identifiability by the loss functional.

Definition 2.1 *The function space of identifiability (FSOI) is the largest linear subspace of $L^2_\rho(\mathcal{S})$ where the loss functional $\mathcal{E}$ in (1.2) has a unique minimizer.*

Clearly, the FSOI is the space in which the Fréchet derivative of the quadratic loss functional has a unique zero. This motivates us to study the Fréchet derivative of the loss functional $\mathcal{E}$ in $L^2_\rho(\mathcal{S})$ and identify the operator of inversion when solving the zero of the derivative of $\mathcal{E}$. We start with writing the quadratic term $\langle L_K\phi, L_K\phi \rangle_{L^2_\mu(\mathcal{T})}$ in the loss functional (1.2) into a bilinear form

$$\langle L_K\phi, L_K\psi \rangle_{L^2_\mu(\mathcal{T})} = \int_{\mathcal{S}} \int_{\mathcal{S}} \phi(s)\psi(s')G(s,s')\nu(ds)\nu(ds'), \tag{2.2}$$

where the integral kernel $G : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is defined as

$$G(s,s') := \int_{\mathcal{T}} K(t,s)K(t,s')\mu(dt). \tag{2.3}$$

Note that

$$G(s,s') \leqslant \|K\|_\infty Z \min\{\frac{d\rho}{d\nu}(s), \frac{d\rho}{d\nu}(s')\}, \ \forall (s,s') \in \mathcal{S} \times \mathcal{S}. \tag{2.4}$$

Then, we can weigh it by the exploration measure and define

$$\overline{G}(s,s') := \frac{G(s,s')}{\frac{d\rho}{d\nu}(s)\frac{d\rho}{d\nu}(s')}. \tag{2.5}$$

The continuity of the kernel $K \in C(\mathcal{T} \times \mathcal{S})$ implies the following lemma.

Lemma 2.2 *Assume that $K \in C(\mathcal{T} \times \mathcal{S})$ and recall ρ in (2.1). The following statements hold.*

4

(a) *The function $\overline{G}$ in (2.5) is a Mercer kernel. The operator $\mathcal{L}_{\overline{G}} : L^2_\rho(\mathcal{S}) \to L^2_\rho(\mathcal{S})$ defined by*

$$\mathcal{L}_{\overline{G}}\phi(r) := \int_a^b \phi(s)\overline{G}(r,s)\rho(ds) \tag{2.6}$$

is compact, self-adjoint and positive. Moreover, for $\phi, \psi \in L^2_\rho(\mathcal{S})$, we have

$$\langle L_K\phi, L_K\psi \rangle_{L^2_\mu(\mathcal{T})} = \langle \mathcal{L}_{\overline{G}}\phi, \psi \rangle_{L^2_\rho(\mathcal{S})} = \langle \phi, \mathcal{L}_{\overline{G}}\psi \rangle_{L^2_\rho(\mathcal{S})}. \tag{2.7}$$

(b) *Let $\{\lambda_i\}_{i\geqslant 1}$ be the positive eigenvalues of $\mathcal{L}_{\overline{G}}$ arranged in a descending order. Denote by $\{\psi_i, \psi^0_j\}_{i,j}$ the orthonormal eigenfunctions of $\mathcal{L}_{\overline{G}}$ with ψ_i and ψ^0_j corresponding to eigenvalues λ_i and zero (if any), respectively. Then, $\{\lambda_i\}_{i\geqslant 1}$ is either finite or $\lambda_i \to 0$ as $i \to \infty$, and these eigenfunctions form a complete basis of $L^2_\rho(\mathcal{S})$.*

(c) *The operator $\mathcal{L}_{\overline{G}}$ is a trace-class operator,*

$$\sum_i \lambda_i = \int_\mathcal{S} \overline{G}(s,s)\rho(ds) \leqslant \|K\|_\infty Z\nu(\mathcal{S}) < \infty, \tag{2.8}$$

where $\nu(\mathcal{S})$ is the measure of $\mathcal{S}$ and Z is the normalizing constant in (2.1).

Proof. Since the kernel K is continuous, so is ρ and G in (2.3), thus $\overline{G}$ is also continuous. Note that $\overline{G}$ is symmetric, i.e., $\overline{G}(s,s') = \overline{G}(s',s)$ for any $s, s' \in \mathcal{S}$. Also, it is positive semi-definite, i.e.,

$$\sum_{i=1}^n \sum_{j=1}^n c_i c_j \overline{G}(s_i, s_j) = \int_\mathcal{T} \left(\sum_{i=1}^n c_i \mathbf{1}_\mathcal{S}(s_i) K(t,s_i)/\frac{d\rho}{d\nu}(s_i) \right)^2 \mu(dt) \geqslant 0 \tag{2.9}$$

for any $\{c_j\}_{j=1}^n \subset \mathbb{R}$, $\{s_j\}_{j=1}^k \subset \mathcal{S}$ and $n \in \mathbb{N}$. Moreover, by (2.4), we have $\overline{G} \in L^2(\rho \times \rho)$ because

$$\int_\mathcal{S} \int_\mathcal{S} \overline{G}(s,s')^2 \rho(ds)\rho(ds') \leqslant \|K\|^2_\infty Z^2 < \infty.$$

Thus, $\mathcal{L}_{\overline{G}}$ is compact and positive self-adjoint (e.g., [6, Proposition 4.6]).

Part (b) follows directly from that $\mathcal{L}_{\overline{G}}$ is compact and positive self-adjoint ([6, Theorem 4.7]).

To prove Part (c), recall that from Mercer's theorem, we have

$$\overline{G}(s,s') = \sum_i \lambda_i \psi_i(s)\psi_i(s'),$$

where the convergence is uniform on $\mathcal{S} \times \mathcal{S}$. Then,

$$\sum_i \lambda_i = \int_\mathcal{S} \overline{G}(s,s)\rho(ds) = \int_\mathcal{S} \frac{G(s,s)}{\frac{d\rho}{d\nu}(s)} ds \leqslant \|K\|_\infty Z\nu(\mathcal{S}),$$

where the inequality follows from (2.4). ∎

The next theorem characterizes the FSOI through the Fréchet derivative of the loss functional, highlighting the role of $\mathcal{L}_{\overline{G}}$ as the operator of inversion in the variational approach.

Theorem 2.3 (Function space of identifiability) *Assume that $K \in C(\mathcal{T} \times \mathcal{S})$ and let $\mathbf{y}$ be the data generated with a function $\phi_{true} \in L^2_\rho(\mathcal{S})$ by (1.1). Then, we have the following characterization of the inverse problem.*

5

(a) *There exists a unique $\phi^{\mathbf{y}} \in L^2_\rho(\mathcal{S})$ such that*

$$\langle \phi^{\mathbf{y}}, \psi \rangle_{L^2_\rho(\mathcal{S})} = \langle L_K \psi, \mathbf{y} \rangle_{L^2_\mu(\mathcal{T})}, \quad \forall \psi \in L^2_\rho(\mathcal{S}), \tag{2.10}$$

and it has a decomposition

$$\phi^{\mathbf{y}} = \mathcal{L}_{\overline{G}} \phi_{true} + \phi^\sigma, \tag{2.11}$$

where ϕ^σ has a distribution $\mathcal{N}(0, \sigma^2 \mathcal{L}_{\overline{G}})$, i.e., $\phi^\sigma = \sum_i \sigma \xi_i \lambda_i^{1/2} \psi_i$ with $\{\xi_i\}$ being independent identically distributed standard Gaussian random variables and $\mathbb{E}\left[\|\phi^\sigma\|^2_{L^2_\rho(\mathcal{S})} \right] = \sigma^2 \sum_i \lambda_i$.

(b) *The Fréchet derivative of the loss functional in L^2_ρ is $\nabla \mathcal{E}(\phi) = 2(\mathcal{L}_{\overline{G}}\phi - \phi^{\mathbf{y}})$, where $\phi^{\mathbf{y}}$ is defined in* (2.10).

(c) *The FSOI of $\mathcal{E}$ is $H := \overline{\mathrm{span}\{\psi_i\}_{i:\lambda_i>0}}$ with the closure taken in $L^2_\rho(\mathcal{S})$.*

(d) *Assume that the data is noisy, i.e. $\sigma > 0$. If $\sum_{i:\lambda_i>0} \lambda_i^{-1} < \infty$, the unique minimizer of $\mathcal{E}$ in H is $\widehat{\phi} = \mathcal{L}_{\overline{G}}^{-1} \phi^{\mathbf{y}}$; but if $\sum_{i:\lambda_i>0} \lambda_i^{-1} = \infty$, the solution $\mathcal{L}_{\overline{G}}^{-1} \phi^{\mathbf{y}}$ is ill-defined in $L^2_\rho(\mathcal{S})$ in the sense that $\mathbb{E}\left[\|\mathcal{L}_{\overline{G}}^{-1} \phi^{\mathbf{y}}\|^2_{L^2_\rho(\mathcal{S})} \right] = \infty$.*

(e) *When the observation is noiseless, we have $\widehat{\phi} = \mathcal{L}_{\overline{G}}^{-1} \phi^{\mathbf{y}} = P_H \phi_{true}$, where P_H is the projection operator of H.*

Proof. The existence and uniqueness of $\phi^{\mathbf{y}}$ in Part (a) follow from the Riesz representation theorem. Denote the observation by $\mathbf{y} = L_K \phi_{true} + \sigma \Delta \mathbf{W}$ with $\Delta \mathbf{W} = (W(t_i) - W(t_{i-1}) : 1 \leq i \leq m)$. Then, the decomposition of $\phi^{\mathbf{y}}$ follows from (2.10):

$$\langle \phi^{\mathbf{y}}, \psi \rangle_{L^2_\rho(\mathcal{S})} = \langle L_K \psi, L_K \phi_{true} \rangle_{L^2_\mu(\mathcal{T})} + \langle L_K \psi, \sigma \Delta \mathbf{W} \rangle_{L^2_\mu(\mathcal{T})} = \langle \mathcal{L}_{\overline{G}} \phi_{true}, \psi \rangle_{L^2_\rho(\mathcal{S})} + \langle \psi, \phi^\sigma \rangle_{L^2_\rho(\mathcal{S})},$$

where the first term in the last equation comes from the definitions of the operator $\mathcal{L}_{\overline{G}}$ in (2.6), and the second term comes from the Riesz representation. The distribution of ϕ^σ is $\mathcal{N}(0, \sigma^2 \mathcal{L}_{\overline{G}})$ in the sense that for each $\psi \in L^2_\rho(\mathcal{S})$, the random variable $\langle \psi, \phi^\sigma \rangle_{L^2_\rho(\mathcal{S})} = \langle L_K \psi, \sigma \Delta \mathbf{W} \rangle_{L^2_\mu(\mathcal{T})}$ is Gaussian with mean zero and variance

$$\mathbb{E}[\langle \psi, \phi^\sigma \rangle^2_{L^2_\rho(\mathcal{S})}] = \mathbb{E}[\langle L_K \psi, \sigma \Delta \mathbf{W} \rangle^2_{L^2_\mu(\mathcal{T})}] = \sigma^2 \mathbb{E}[\langle L_K \psi, L_K \psi \rangle_{L^2_\mu(\mathcal{T})}] = \sigma^2 \langle \psi, \mathcal{L}_{\overline{G}} \psi \rangle_{L^2_\rho(\mathcal{S})}.$$

Therefore, we can write $\phi^\sigma = \sum_i \sigma \xi_i \lambda_i^{1/2} \psi_i$ with $\{\xi_i\}$ being i.i.d. standard Gaussian, and $\mathbb{E}\left[\|\phi^\sigma\|^2_{L^2_\rho(\mathcal{S})} \right] = \sigma^2 \sum_i \lambda_i$, where the sum is finite by (2.8).

Part (b) follows directly from the definition of the Fréchet derivative. In fact, by (2.7) and (2.10), we can write the loss functional as

$$\mathcal{E}(\phi) = \langle \mathcal{L}_{\overline{G}} \phi, \phi \rangle_{L^2_\rho(\mathcal{S})} + \langle \phi^{\mathbf{y}}, \phi \rangle_{L^2_\rho(\mathcal{S})} + \|\mathbf{y}\|^2_{L^2_\mu(\mathcal{T})}. \tag{2.12}$$

Then, the Fréchet derivative $\nabla \mathcal{E}(\phi)$ in $L^2_\rho(\mathcal{S})$ is

$$\langle \nabla \mathcal{E}(\phi), \psi \rangle_{L^2_\rho(\mathcal{S})} = \lim_{h \to 0} \frac{\mathcal{E}(\phi + h\psi) - \mathcal{E}(\phi)}{h} = \langle 2(\mathcal{L}_{\overline{G}} \phi - \phi^{\mathbf{y}}), \psi \rangle_{L^2_\rho(\mathcal{S})}, \ \forall \psi \in L^2_\rho(\mathcal{S}).$$

For Part (c), first, note that the quadratic loss functional has a unique minimizer in H because its derivative has a unique zero in H. Meanwhile, note that H is the orthogonal complement of

the null space of $\mathcal{L}_{\overline{G}}$, and $\mathcal{E}(\phi_{true} + \phi^0) = \mathcal{E}(\phi_{true})$ for any ϕ^0 such that $\mathcal{L}_{\overline{G}}\phi^0 = 0$. Thus, H is the largest such linear subspace of $L_\rho^2(\mathcal{S})$, and we conclude that H is the FSOI.

For Part (d), note that

$$\mathcal{L}_{\overline{G}}^{-1}\phi^\sigma = \sum_{i:\lambda_i>0} \lambda_i^{-1/2}\sigma\xi_i\psi_i.$$

Then, if $\sum_{i:\lambda_i>0} \lambda_i^{-1} < \infty$, which happens only when there are finitely many non-zero eigenvalues, we have $\mathcal{L}_{\overline{G}}^{-1}\phi^\sigma \in L_\rho^2(\mathcal{S})$ because $\mathbb{E}\left[\|\mathcal{L}_{\overline{G}}^{-1}\phi^\sigma\|_{L_\rho^2(\mathcal{S})}^2\right] = \sigma^2 \sum_{i:\lambda_i>0} \lambda_i^{-1} < \infty$ by using the fact that $\{\xi_i\}$ are i.i.d. standard Gaussian. Thus, ϕ^σ in $\mathcal{L}_{\overline{G}}(L_\rho^2)$, so is $\phi^{\mathbf{y}}$, and the estimator $\widehat{\phi} = \mathcal{L}_{\overline{G}}^{-1}\phi^{\mathbf{y}}$ is well-defined in L_ρ^2. By Part (b), this estimator is the unique zero of the loss functional's Fréchet derivative in H. Hence it is the unique minimizer of $\mathcal{E}(\phi)$ in H.

In contrary, if $\sum_i \lambda_i^{-1} = \infty$, we have $\mathbb{E}\left[\|\mathcal{L}_{\overline{G}}^{-1}\phi^\sigma\|_{L_\rho^2(\mathcal{S})}^2\right] = \sum_i \lambda_i^{-1}\sigma^2 = \infty$, and hence $\mathcal{L}_{\overline{G}}^{-1}\phi^{\mathbf{y}}$ is ill-defined.

For Part (e), when the data is noiseless, i.e., $\mathbf{y} = L_K\phi_{true}$, we have $\phi^{\mathbf{y}} = \mathcal{L}_{\overline{G}}\phi_{true}$ from Part (a). Hence $\widehat{\phi} = \mathcal{L}_{\overline{G}}^{-1}\phi^{\mathbf{y}} = P_H\phi_{true}$. ∎

Theorem 2.3 reveals the nature of the ill-posedness of this inverse problem, and provides insights on regularization:

- The variational inverse problem is ill-defined outside the FSOI H. Its ill-posedness in H depends on the smallest eigenvalue of the operator $\mathcal{L}_{\overline{G}}$.

- When the data is noiseless, the minimizer of the loss functional is the H-projection of the true input function. In other words, the inverse problem can only recover the H-projection of the true input function.

- When the data is noisy, its minimizer $\mathcal{L}_{\overline{G}}^{-1}\phi^{\mathbf{y}}$ can be ill-defined in L_ρ^2 if $\sum_{i:\lambda_i>0} \lambda_i^{-1} = \infty$, but it is well-defined when $\sum_{i:\lambda_i>0} \lambda_i^{-1} < \infty$.

Based on these observations, the regularization scheme proposed in Section 2.2 is designed to ensure the solution lies in the FSOI and either remove the non-integrable components related to small eigenvalues or reduce the bias caused by the noise.

2.2 An adaptive RKHS regularization

We introduce an adaptive RKHS Tikhonov regularization. This RKHS is defined based on the operator of inversion $\mathcal{L}_{\overline{G}}$; thus, it is adaptive to the original integral operator L and the data settings. The next lemma characterizes this RKHS.

Lemma 2.4 (Characterization of the adaptive RKHS) *Assume $K \in C(\mathcal{T}\times\mathcal{S})$. The RKHS H_G with $\overline{G}$ as its reproducing kernel satisfies the following properties.*

(a) *$H_G := \mathcal{L}_{\overline{G}}^{\frac{1}{2}}(L_\rho^2(\mathcal{S}))$ and its inner product is*

$$\langle\phi,\psi\rangle_{H_G} := \langle\mathcal{L}_{\overline{G}}^{-\frac{1}{2}}\phi, \mathcal{L}_{\overline{G}}^{-\frac{1}{2}}\psi\rangle_{L_\rho^2(\mathcal{S})}.$$

The operator $\mathcal{L}_{\overline{G}}$ is self-adjoint in H_G. Moreover, we have $\langle\phi,\psi\rangle_{L_\rho^2(\mathcal{S})} = \langle\mathcal{L}_{\overline{G}}\phi,\psi\rangle_{H_G}$ for any $\phi \in L_\rho^2(\mathcal{S})$ and $\psi \in H_G$.

(b) *$\{\sqrt{\lambda_i}\psi_i\}_{i=1}^\infty$ is an orthonormal basis of H_G, where $\{\psi_i\}_i$ are the orthonormal eigenfunctions of $\mathcal{L}_{\overline{G}}$ in $L_\rho^2(\mathcal{S})$ corresponding to positive eigenvalues $\{\lambda_i\}_i$.*

(c) *For any $\phi = \sum_{i=1}^{\infty} c_i \psi_i$, with $c_i \in \mathbb{R}$, we have*

$$\langle L_K \phi, L_K \phi \rangle_{L^2_\mu(\mathcal{T})} = \sum_{i=1}^{\infty} \lambda_i c_i^2, \quad \|\phi\|^2_{L^2_\rho} = \sum_{i=1}^{\infty} c_i^2, \quad \|\phi\|^2_{H_G} = \sum_{i=1}^{\infty} \lambda_i^{-1} c_i^2 \text{ if } \phi \in H_G. \quad (2.13)$$

Moreover, the H_G norm is stronger than the L^2_ρ-norm: $\|\phi\|^2_{H_G} \geqslant \lambda_1^{-1} \|\phi\|^2_{L^2_\rho}$.

(d) *$H = \overline{H_G}$ with closure in $L^2_\rho(\mathcal{S})$, where $H = \overline{\mathrm{span}\{\psi_i\}_{i:\lambda_i>0}}$ is the FSOI.*

Proof. The first part is from the standard characterization theorem of RKHS, e.g., [6, Section 4.4] or [14]. When $\phi \in L^2_\rho(\mathcal{S})$, we have $\mathcal{L}_{\overline{G}} \phi \in H_G$. Then, by the definition of the inner product and the symmetry of $\mathcal{L}_{\overline{G}}^{-1/2}$, we have $\langle \mathcal{L}_{\overline{G}} \phi, \psi \rangle_{H_G} = \langle \mathcal{L}_{\overline{G}}^{1/2} \phi, \mathcal{L}_{\overline{G}}^{-1/2} \psi \rangle_{L^2_\rho(\mathcal{S})} = \langle \phi, \psi \rangle_{L^2_\rho(\mathcal{S})}$.

Part (b) follows directly from the characterization of the inner product in Part (a). Part (c) follows directly from (2.7), the orthonormality of the eigenfunctions and the characterization of the inner product. Part (d) follows from that $\overline{H_G} = \overline{\mathrm{span}\{\sqrt{\lambda_i}\psi_i\}_{i:\lambda_i>0}} = \overline{\mathrm{span}\{\psi_i\}_{i:\lambda_i>0}} = H$. ∎

Next, we propose to use the RKHS norm for regularization, that is, use $\|\phi\|_{H_G}$ in the place of $\|\phi\|_\square$ in (1.3). Note that by using the RKHS norm, the loss functional is minimized over H_G, not $H \subset L^2_\rho(\mathcal{S})$, because elements in H may have unbounded H_G norm.

Proposition 2.5 (Regularized estimators) *Consider the L^2_σ-regularized estimator $\widehat{\phi}_\alpha^{L^2_\rho}$, and the H_G-regularized estimator $\widehat{\phi}_\alpha^{H_G}$, which can be expressed in terms of $\mathcal{L}_{\overline{G}}$ as follows:*

$$\widehat{\phi}_\lambda^{L^2_\rho} := \underset{\phi \in L^2_\rho}{\arg\min} \, \mathcal{E}(\phi) + \lambda \|\phi\|^2_{L^2_\rho} = (\mathcal{L}_{\overline{G}} + \lambda I)^{-1} \phi^{\mathbf{y}}, \quad (2.14)$$

$$\widehat{\phi}_\lambda^{H_G} := \underset{\phi \in H_G}{\arg\min} \, \mathcal{E}(\phi) + \lambda \|\phi\|^2_{H_G} = (\mathcal{L}_{\overline{G}}^2 + \lambda I)^{-1} \mathcal{L}_{\overline{G}} \phi^{\mathbf{y}}. \quad (2.15)$$

Express the true function in terms of the eigen-basis functions of $\mathcal{L}_{\overline{G}}$ as $\phi_{true} = \sum_i c_i \psi_i + \sum_j d_j \psi_j^0$. Then, for any $\lambda > 0$, the biases of these two regularized estimators satisfy the following estimates:

$$\left\| \widehat{\phi}_\lambda^{L^2_\rho} - \phi_{true} \right\|^2_{L^2_\rho} = \sum_i (\lambda_i + \lambda)^{-2} (\sigma \lambda_i^{1/2} \xi_i - \lambda c_i)^2 + \sum_j d_j^2, \quad (2.16)$$

$$\left\| \widehat{\phi}_\lambda^{H_G} - \phi_{true} \right\|^2_{L^2_\rho} = \sum_i (\lambda_i^2 + \lambda)^{-2} (\sigma \lambda_i^{3/2} \xi_i - \lambda c_i)^2 + \sum_j d_j^2, \quad (2.17)$$

where ξ_i are i.i.d. standard Gaussian random variables in Theorem 2.3.

Proof. The minimizers' uniqueness and explicit form follow from the Fréchet derivatives of the regularized loss functionals. In fact, by Theorem 2.3(b), the Fréchet derivative of $\mathcal{E}_\lambda(\phi) = \mathcal{E}(\phi) + \lambda \|\phi\|^2_{L^2_\rho(\mathcal{S})}$ in $L^2_\rho(\mathcal{S})$ is

$$\nabla \mathcal{E}_\lambda(\phi) = 2[(\mathcal{L}_{\overline{G}} + \lambda I)\phi - \phi^{\mathbf{y}}].$$

It has a unique zero, $\widehat{\phi}_\lambda^{L^2_\rho} = (\mathcal{L}_{\overline{G}} + \lambda I)^{-1} \phi^{\mathbf{y}}$, which is the unique minimizer of $\mathcal{E}_\lambda$ in L^2_ρ, and thus we have (2.14). Similarly, note that by Lemma 2.4 (a) and (2.12), for any $\phi \in H_G$,

$$\mathcal{E}(\phi) = \langle \mathcal{L}_{\overline{G}}^2 \phi, \phi \rangle_{H_G} + \langle \mathcal{L}_{\overline{G}} \phi^{\mathbf{y}}, \phi \rangle_{H_G} + \|\mathbf{y}\|^2_{L^2_\mu(\mathcal{T})}.$$

8

Then, the Fréchet derivative of $\widetilde{\mathcal{E}}_\lambda = \mathcal{E}(\phi) + \lambda \|\phi\|^2_{H_G}$ in H_G is

$$\nabla\widetilde{\mathcal{E}}_\lambda(\phi) = 2[(\mathcal{L}_{\overline{G}}^2 + \lambda I)\phi - \mathcal{L}_{\overline{G}}\phi^{\mathbf{y}}].$$

Its unique zero, $\widehat{\phi}_\lambda^{H_G} = (\mathcal{L}_{\overline{G}}^2 + \lambda I)^{-1}\mathcal{L}_{\overline{G}}\phi^{\mathbf{y}}$, is the minimizer of $\widetilde{\mathcal{E}}_\lambda$ in H_G as in (2.15).

The eigenvalue characterizations of the biases in (2.16) and (2.17) follow from the decomposition of $\phi^{\mathbf{y}} = \mathcal{L}_{\overline{G}}\phi_{true} + \phi^\sigma$ with $\phi^\sigma = \sum_i \sigma\lambda_i^{1/2}\xi_i\psi_i$ in (2.11) and the facts listed below:

$$\widehat{\phi}_\lambda^{L_\rho^2} = (\mathcal{L}_{\overline{G}} + \lambda I)^{-1}(\mathcal{L}_{\overline{G}}\phi_{true} + \lambda\phi_{true} - \lambda\phi_{true} + \phi^\sigma) = \phi_{true} + (\mathcal{L}_{\overline{G}} + \lambda I)^{-1}(-\lambda\phi_{true} + \phi^\sigma),$$

$$\widehat{\phi}_\lambda^{H_G} = (\mathcal{L}_{\overline{G}}^2 + \lambda I)^{-1}(\mathcal{L}_{\overline{G}}^2\phi_{true} + \mathcal{L}_{\overline{G}}\phi_1^\delta) = \phi_{true} + (\mathcal{L}_{\overline{G}}^2 + \lambda I)^{-1}(-\lambda\phi_{true} + \mathcal{L}_{\overline{G}}\phi^\sigma).$$

Then, we obtain (2.16) and (2.17). ∎

Proposition 2.5 demonstrates the complexity of choosing an optimal hyper-parameter λ. An optimal λ aims to balance the error caused by the noise and the shift from the true solution caused by the regularization, and it depends on the spectrum of the operator, each realization of the noise, and the true solution. Thus, it is important to select an optimal λ adaptive to these factors, and the L-curve method does so.

Also, Proposition 2.5 shows that the true solution's components outside the FSOI remain a bias for both estimators, because there is no information about these components in the data. Thus, in the error analysis presented in Section 3, we focus only on the error inside the FSOI.

3 Convergence of the regularized estimators as the noise decays

A key question in regularization is the choice of the regularization norm $\|\phi\|_\square$ in (1.3) when there is little prior knowledge. Unfortunately, empirical comparisons do not lead to a conclusive answer due to multiple mixed factors, including the random noise, the spectrum of the inversion operator, the smoothness of the solution, the model/discretization error, the selection of the hyper-parameter, and the method of optimization (matrix decomposition for least squares or iterative methods for the minimization). Therefore, to make a meaningful comparison of different regularized estimators, one must extract the essentials from these multiple factors.

We introduce next a comparative analysis framework to compare regularization norms in the small noise limit. The small noise limit approach helps us separate the aforementioned mixed factors and analyze their effects in regularization. Hence, this comparative analysis framework enables us to compare the adaptive RKHS norm and the L_ρ^2 norm, making a small step toward guiding the regularization norm selection in practice.

We show that both the RKHS- and L_ρ^2-regularized estimators converge in mean-squares at order 1 as $\sigma \downarrow 0$ under the assumption that the spectrum of $\mathcal{L}_{\overline{G}}$ converges exponentially or polynomially. The RKHS-regularizer slightly outperforms the L_ρ^2-regularizer by a smaller multiplicative factor. The spectral assumption and the convergence are observed in the numerical tests in Section 4.

In the following comparative analysis, we assume $\phi_{true} = \sum_i c_i\psi_i \in H$. We estimate the mean-square error (MSE) of the L_ρ^2-regularized and the H_G-regularized estimators:

$$e^{L_\rho^2}(\lambda) := \mathbb{E}\left\|\widehat{\phi}_\lambda^{L_\rho^2} - \phi_{true}\right\|^2_{L_\rho^2}, \quad e^{H_G}(\lambda) := \mathbb{E}\left\|\widehat{\phi}_\lambda^{H_G} - \phi_{true}\right\|^2_{L_\rho^2}.$$

Note that by Proposition 2.5, these MSEs are

$$
e^{L_\rho^2}(\lambda) = \sum_i (\lambda_i + \lambda)^{-2}(\sigma^2 \lambda_i + \lambda^2 c_i^2),
$$
$$
e^{H_G}(\lambda) = \sum_i (\lambda_i^2 + \lambda)^{-2}(\sigma^2 \lambda_i^3 + \lambda^2 c_i^2).
$$

(3.1)

The purpose of the following theorem is to give a quantitative comparison of these two regularized estimators.

Theorem 3.1 (Small noise limits of regularized estimators) *Consider the inverse problem* (1.1) *with a true solution* $\phi_{true} = \sum_i c_i \psi_i \in H = \overline{\text{span}\{\psi_i\}}$, *where* $\{\psi_i\}$ *are the eigenfunctions of positive eigenvalues of* $\mathcal{L}_{\overline{G}}$ *in* (2.6). *If the decay rate of the spectrum of* $\mathcal{L}_{\overline{G}}$, *which is determined by the integral kernel* K, *is either exponential or polynomial, the following upper bounds and sharp rates hold for the MSEs in* (3.1) *as the standard deviation of the Gaussian noise* $\sigma \to 0$.

- **Exponential spectrum decay.** *Assume that the spectrum of the operator* $\mathcal{L}_{\overline{G}}$ *decays to 0 exponentially, i.e., there exists* $\theta > 0$ *such that* $\lambda_i = e^{-\theta i}$ *for all* $i \geqslant 1$.

  (a) *If* ϕ_{true} *satisfies the condition* $M_0 := \sup_{i \geqslant 1} \lambda_i^{-1} c_i^2 < \infty$, *which is valid for every function in* H_G *but can also hold for functions not belonging to* H_G, *then we have the following estimates:*

  $$
  \min_{\lambda > 0} e^{H_G}(\lambda) \leqslant e^{H_G}(\sigma^2) \leqslant (1 + M_0)\frac{\pi}{4\theta}\sigma + O(\sigma^2);
  $$
  $$
  \min_{\lambda > 0} e^{L_\rho^2}(\lambda) \leqslant e^{L_\rho^2}(\sigma) \leqslant (1 + M_0)\frac{2}{\theta}\sigma + O(\sigma^2),
  $$

  (3.2)

  where $O(\sigma^2)$ *is the big-O notation.*

  (b) *In particular, for* ϕ_{true} *satisfying* $c_i^2 = \lambda_i$, *we have the following sharp rates of convergence:*

  $$
  e^{H_G}(\lambda_{opt}) = \frac{\pi}{4\theta}\sigma + O(\sigma^2) \quad \text{with } \lambda_{opt} := \arg\min_{\lambda > 0} e^{H_G}(\lambda) = \sigma^2,
  $$
  $$
  e^{L_\rho^2}(\widetilde{\lambda}_{opt}) = \frac{2}{\theta}\sigma + O(\sigma^2) \quad \text{with } \widetilde{\lambda}_{opt} := \arg\min_{\lambda > 0} e^{L_\rho^2}(\lambda) = \sigma + O(\sigma^2).
  $$

  (3.3)

- **Power spectrum decay.** *Assume that the spectrum of the operator* $\mathcal{L}_{\overline{G}}$ *decays polynomially, i.e., there exists* $\theta > 0$ *such that* $\lambda_i = i^{-\theta}$ *for all* $i \geqslant 1$.

  (c) *If* $M_0 := \sup_{i \geqslant 1} \lambda_i^{-1} c_i^2 < \infty$, *then we have the following estimates:*

  $$
  \min_{\lambda > 0} e^{H_G}(\lambda) \leqslant e^{H_G}(\sigma^2) \leqslant (1 + M_0)\sigma \left(C_{H_G} + o(1) \right);
  $$
  $$
  \min_{\lambda > 0} e^{L_\rho^2}(\lambda) \leqslant e^{L_\rho^2}(\sigma) \leqslant (1 + M_0)\sigma (C_{L_\rho^2} + o(1)).
  $$

  (3.4)

  (d) *In particular, for* ϕ_{true} *satisfying* $c_i^2 = \lambda_i$, *we have the following* **sharp rates** *of convergence:*

  $$
  e^{H_G}(\lambda_{opt}) = \sigma \left(C_{H_G} + o(1) \right) \quad \text{with } \lambda_{opt} = \arg\min_{\lambda > 0} e^{H_G}(\lambda) = \sigma^2,
  $$
  $$
  e^{L_\rho^2}(\widetilde{\lambda}_{opt}) = \sigma(C_{L_\rho^2} + o(1)) \quad \text{with } \widetilde{\lambda}_{opt} = \arg\min_{\lambda > 0} e^{L_\rho^2}(\lambda) = \sigma(C_{\widetilde{\lambda}_{opt}} + o(1)),
  $$

  (3.5)

  where the constants are $C_{H_G} = \frac{1}{2}C_{L_\rho^2} = \frac{1}{2\theta}\Gamma(\frac{1}{2} - \frac{1}{2\theta})\Gamma(\frac{1}{2} + \frac{1}{2\theta})$ *and* $C_{\widetilde{\lambda}_{opt}} = \frac{\theta+1}{\theta-1}$.

10

Remark 3.2 (Sharp rates and restrictive assumptions) *The sharp rates in Parts (b) and (d) rely on the restrictive assumptions on the coefficients of ϕ_{true}. These sharp rates show that the RKHS regularizer outperforms the L_ρ^2 regularizer in the sense that it has a smaller multiplicative constant. On the other hand, these restrictive assumptions can be significantly relaxed when a sharp rate is not needed. For example, one only needs a uniform bound for the Picard ratio in Parts (a) and (c) when one only aims for an upper bound; one can also allow perturbations in the form of $\lambda_i = p_i i^{-\theta}$ or $\lambda_i = p_i e^{-\theta i}$ with the perturbations $0 < a \leqslant p_i \leqslant b < \infty$ for all i for some $a, b > 0$ when one is only concerned with the rate but not the multiplicative constant (see [?] for more details).*

To prove the theorem, we first introduce some notations and two technical lemmas. We rewrite the MSEs in (3.1) as

$$e^{HG}(\lambda) = \sigma^2 A(\lambda) + \lambda^2 B(\lambda); \quad e^{L_\rho^2}(\lambda) = \sigma^2 \widetilde{A}(\lambda) + \lambda^2 \widetilde{B}(\lambda). \tag{3.6}$$

where A and B are defined by

$$
\begin{aligned}
A(\lambda) &= \sum_i (\lambda_i^2 + \lambda)^{-2} \lambda_i^3, & B(\lambda) &= \sum_i (\lambda_i^2 + \lambda)^{-2} c_i^2, \\
\widetilde{A}(\lambda) &= \sum_i (\lambda_i + \lambda)^{-2} \lambda_i, & \widetilde{B}(\lambda) &= \sum_i (\lambda_i + \lambda)^{-2} c_i^2.
\end{aligned}
\tag{3.7}
$$

Then, since a minimizer must be a critical point, we have that λ_{opt} must satisfy

$$0 = \frac{d}{d\lambda} e^{HG}(\lambda) = \sigma^2 A'(\lambda) + 2\lambda[B(\lambda) + \frac{\lambda}{2} B'(\lambda)]$$

and similarly for $\widetilde{\lambda}_{opt}$. With $B_1(\lambda) := B(\lambda) + \frac{\lambda}{2} B'(\lambda)$ and $\widetilde{B}_1(\lambda) := \widetilde{B}(\lambda) + \frac{\lambda}{2} \widetilde{B}'(\lambda)$, we obtain the following lemma providing equations for the minimizers of the MSEs.

Lemma 3.3 *The minimizers of e^{HG} and $e^{L_\rho^2}$ in (3.1), denoted by λ_{opt} and $\widetilde{\lambda}_{opt}$, respectively, satisfy*

$$\lambda_{opt} = -\sigma^2 \frac{A'(\lambda_{opt})}{2B_1(\lambda_{opt})}, \quad \widetilde{\lambda}_{opt} = -\sigma^2 \frac{\widetilde{A}'(\widetilde{\lambda}_{opt})}{2\widetilde{B}_1(\widetilde{\lambda}_{opt})}, \tag{3.8}$$

where the functions $A'(\lambda)$ and $A'(\lambda)$ $B_1(\lambda)$, $\widetilde{A}(\lambda)$ and $\widetilde{B}_1(\lambda)$ are given by

$$
\begin{aligned}
A'(\lambda) &= -2 \sum_i (\lambda_i^2 + \lambda)^{-3} \lambda_i^3, & B_1(\lambda) &= \sum_i (\lambda_i^2 + \lambda)^{-3} \lambda_i^2 c_i^2, \\
\widetilde{A}'(\lambda) &= -2 \sum_i (\lambda_i + \lambda)^{-3} \lambda_i, & \widetilde{B}_1(\lambda) &= \sum_i (\lambda_i + \lambda)^{-3} \lambda_i c_i^2.
\end{aligned}
\tag{3.9}
$$

The next lemma estimates these series by Riemann sum when λ is small when the spectrum decays exponentially.

Lemma 3.4 *Assume that $\lambda_i = e^{-\theta i}$ for all $i \geqslant 1$ with $\theta > 0$. Then, for small $\lambda > 0$, we have*

$$A(\lambda) = \sum_i (\lambda_i^2 + \lambda)^{-2}\lambda_i^3 = \frac{1}{2\theta\sqrt{\lambda}}[\arctan\frac{1}{\sqrt{\lambda}} - \frac{\sqrt{\lambda}}{1+\lambda}] + O(1),$$

$$B_c(\lambda) := \sum_i (\lambda_i^2 + \lambda)^{-2}\lambda_i = \frac{1}{2\theta}\lambda^{-3/2}[\arctan\frac{1}{\sqrt{\lambda}} + \frac{\sqrt{\lambda}}{1+\lambda}] + O(1),$$

$$\widetilde{A}(\lambda) = \sum_i (\lambda_i + \lambda)^{-2}\lambda_i = \frac{1}{\theta\lambda(1+\lambda)} + O(1), \tag{3.10}$$

$$\widetilde{A}'(\lambda) = -2\sum_i (\lambda_i + \lambda)^{-3}\lambda_i = \frac{-(1+2\lambda)}{\theta\lambda^2(1+\lambda)^2} + O(1),$$

$$\widetilde{B}_c(\lambda) := \sum_i (\lambda_i + \lambda)^{-2}\lambda_i^2 = \frac{1}{2\theta\lambda(1+\lambda)^2} + O(1).$$

Proof. The proof is based on the Riemann sum approximation of integrals. Note that for $k \in \{1, 3\}$, the function $f(x) = (e^{-2\theta x} + \lambda)^{-2}e^{-\theta x k}$ satisfies $\int_0^\infty f(x)dx = \sum_{i=1}^\infty f(i) + O(1)$. Thus,

$$\sum_i (\lambda_i^2 + \lambda)^{-2}\lambda_i^k = \int_0^\infty (e^{-2\theta x} + \lambda)^{-2}e^{-\theta x k}dx + O(1) = \frac{1}{\theta}\int_0^1 \frac{t^{k-1}dt}{(t^2 + \lambda)^2} + O(1),$$

where we applied a change of variables $t = e^{-\theta x}$ to obtain the second equality. Then, the first two equations in (3.10) follow directly from the facts that $\int_0^1 \frac{t^2 dt}{(t^2+\lambda)^2} = \frac{1}{2\sqrt{\lambda}}[\arctan\frac{1}{\sqrt{\lambda}} - \frac{\sqrt{\lambda}}{1+\lambda}]$ and $\int_0^1 \frac{dt}{(t^2+\lambda)^2} = \frac{1}{2}\lambda^{-3/2}[\arctan\frac{1}{\sqrt{\lambda}} + \frac{\sqrt{\lambda}}{1+\lambda}]$.

Similarly, we obtain the last three equations in (3.10) by using the integrals $\int_0^1 \frac{1}{(t+\lambda)^2}dt$, $\int_0^1 \frac{1}{(t+\lambda)^3}dt$ and $\int_0^1 \frac{t}{(t+\lambda)^3}dt$. ∎

The next lemma estimates these series when the spectrum decays polynomially.

Lemma 3.5 *Assume that $\lambda_i = i^{-\theta}$ for all $i \geqslant 1$ with $\theta > 0$. Denote $C_\theta(s, k, \alpha) = \Gamma(\gamma)\Gamma(k - \gamma)$ with $\gamma = \frac{1}{1+s}(\alpha - \frac{1}{\theta})$, where $s \in \{0, 1\}$, $\alpha \in \{1, 2, 3\}$, and $k \in \{2, 3\}$. Then, for small $\lambda > 0$, we have*

$$A(\lambda) = \sum_i (\lambda_i^2 + \lambda)^{-2}\lambda_i^3 = \frac{1}{2\theta}\lambda^{\gamma-2}[C_\theta(1, 2, 3) + o(1)] \text{ with } \gamma = \frac{1}{2}(3 - \frac{1}{\theta});$$

$$B_c(\lambda) := \sum_i (\lambda_i^2 + \lambda)^{-2}\lambda_i = \frac{1}{2\theta}\lambda^{\gamma-2}[C_\theta(1, 2, 1) + o(1)] \text{ with } \gamma = \frac{1}{2}(1 - \frac{1}{\theta});$$

$$\widetilde{A}(\lambda) = \sum_i (\lambda_i + \lambda)^{-2}\lambda_i = \frac{1}{\theta}\lambda^{\gamma-2}[C_\theta(0, 2, 1) + o(1)] \text{ with } \gamma = (1 - \frac{1}{\theta}); \tag{3.11}$$

$$\widetilde{A}'(\lambda) = -2\sum_i (\lambda_i + \lambda)^{-3}\lambda_i = -\frac{1}{\theta}\lambda^{\gamma-3}[C_\theta(0, 3, 1) + o(1)] \text{ with } \gamma = (1 - \frac{1}{\theta});$$

$$\widetilde{B}_c(\lambda) := \sum_i (\lambda_i + \lambda)^{-3}\lambda_i^2 = \frac{1}{2\theta}\lambda^{\gamma-3}[C_\theta(0, 3, 2) + o(1)] \text{ with } \gamma = (2 - \frac{1}{\theta}).$$

Proof. Again, the proof is based on the Riemann sum approximation of integrals.

To estimate the five terms in unified notations, let

$$F_s(\lambda, \alpha, k) := \sum_i (\lambda_i^{1+s} + \lambda)^{-k} \lambda_i^\alpha, \quad \gamma = \frac{\alpha - \frac{1}{\theta}}{1 + s}.$$

Then, with $s = 1$, we have $A(\lambda) = F_1(\lambda, 3, 2)$ and $B_c(\lambda) = F_1(\lambda, 1, 2)$, which use $k = 2$ and $\alpha \in \{1, 3\}$. Also, with $s = 0$, we have $\widetilde{A}(\lambda) = F_0(\lambda, 1, 2)$, $\widetilde{A}'(\lambda) = F_0(\lambda, 1, 3)$, and $\widetilde{B}_c(\lambda) = F_0(\lambda, 2, 3)$ using $\alpha \in \{1, 2\}$ and $k \in \{2, 3\}$.

Note that the function $f(x) = (x^{-(1+s)\theta} + \lambda)^{-k} x^{-\theta \alpha}$ satisfies $\int_1^\infty f(x) dx = \sum_{i=1} f(i) + O(1)$. Thus, under the assumption $\lambda_i = i^{-\theta}$, we have

$$F_s(\lambda, \alpha, k) = \sum_i (\lambda_i^{1+s} + \lambda)^{-k} \lambda_i^\alpha = \int_1^\infty (x^{-\theta(1+s)} + \lambda)^{-k} x^{-\theta \alpha} dx + O(1)$$

$$= \frac{1}{\theta} \int_0^1 \frac{t^{\alpha - 1 - 1/\theta} dt}{(t^{1+s} + \lambda)^k} + O(1),$$

where we applied a change of variables $t = x^{-\theta}$ (equivalently, $x = t^{-1/\theta}$ and $dx = -\frac{1}{\theta} t^{-1-1/\theta} dt$) to obtain the second equality. We will estimate the last integral with $s = 1$ and $s = 0$ separately.

Consider first the cases with $s = 1$, i.e., the series $A(\lambda) = F_1(\lambda, 3, 2)$ and $B_c(\lambda) = F_1(\lambda, 1, 2)$, which use $k = 2$ and $\alpha \in \{1, 3\}$. By a change of variables $z = t^2/\lambda$ (and hence $t = \lambda^{1/2} z^{1/2}$ and $dt = \frac{1}{2} \lambda^{1/2} z^{-1/2} dz$), and note that $\gamma = \frac{1}{2}(\alpha - \frac{1}{\theta}) < 2$, we have

$$F_1(\lambda, \alpha, 2) = \frac{1}{\theta} \int_0^1 \frac{t^{\alpha - 1 - 1/\theta} dt}{(t^2 + \lambda)^2} = \frac{1}{2\theta} \lambda^{\gamma - 2} \int_0^{1/\lambda} (z + 1)^{-2} z^{\gamma - 1} dz$$

$$= \frac{1}{2\theta} \lambda^{\gamma - 2} \left(\frac{\Gamma(\gamma)\Gamma(2 - \gamma)}{\Gamma(2)} + o(1) \right),$$

where the last equality follows from that fact that $\int_0^\infty (z + 1)^{-k} z^{\gamma - 1} dz = \frac{\Gamma(\gamma)\Gamma(k-\gamma)}{\Gamma(k)}$ and

$$\int_0^{1/\lambda} (z + 1)^{-2} z^{\gamma - 1} dz = \int_0^\infty (z + 1)^{-2} z^{\gamma - 1} dz + o(1).$$

Recall that $\Gamma(2) = 1$. Then, the first two equations in (3.11) follow.

Next, consider the cases with $s = 0$, i.e., the series $\widetilde{A}(\lambda) = F_0(\lambda, 1, 2)$, $\widetilde{A}'(\lambda) = F_0(\lambda, 1, 3)$, and $\widetilde{B}_c(\lambda) = F_0(\lambda, 2, 3)$ using $\alpha \in \{1, 2\}$ and $k \in \{2, 3\}$. By a change of variables $z = t/\lambda$, and note that $\gamma = \alpha - \frac{1}{\theta} < k$, we have

$$F_0(\lambda, \alpha, k) = \frac{1}{\theta} \int_0^1 \frac{t^{\alpha - 1 - 1/\theta} dt}{(t + \lambda)^k} = \frac{1}{\theta} \lambda^{\gamma - k} \int_0^{1/\lambda} (z + 1)^{-k} z^{\gamma - 1} dz$$

$$= \frac{1}{\theta} \lambda^{\gamma - k} \left(\frac{\Gamma(\gamma)\Gamma(k - \gamma)}{\Gamma(k)} + o(1) \right).$$

Then, we obtain the last three equations in (3.11) using the facts that $\Gamma(2) = 1$ and $\Gamma(3) = 2$. ∎

Proof of Theorem 3.1. We prove Parts (a) and (c) by solving algebraic equations to obtain the optimal hyper-parameters using Lemmas 3.4-3.5, and we prove Parts (b) and (d) by direct evaluations.

13

Part (a). The upper bounds follow from direct evaluations of $e^{H_G}(\sigma^2)$ and $e^{L_\rho^2}(\sigma)$. In fact, recall that $e^{H_G}(\sigma^2) = [\sigma^2 A(\lambda) + \lambda^2 B(\lambda)] \,|_{\lambda=\sigma^2}$, and the bound for $\sigma^2 A(\sigma^2)$ comes from (3.10):

$$\sigma^2 A(\sigma^2) = \sigma^2 \left(\frac{1}{2\theta\sqrt{\lambda}}[\arctan\frac{1}{\sqrt{\lambda}} - \frac{\sqrt{\lambda}}{1+\lambda}] + O(1) \right) \,|_{\lambda=\sigma^2} \leqslant \frac{\pi}{4\theta}\sigma + o(\sigma).$$

To bound the term $\lambda^2 B(\lambda)$ with $\lambda = \sigma^2$, note that $M_0 := \sup_i \lambda_i^{-1}c_i^2 < \infty$ since $\sum_i \lambda_i^{-1}c_i^2 < \infty$. Then,

$$\lambda^2 B(\lambda) \,|_{\lambda=\sigma^2} = \lambda^2 \sum_i (\lambda_i^2 + \lambda)^{-2}\lambda_i \lambda_i^{-1}c_i^2 \,|_{\lambda=\sigma^2} \leqslant M_0 \lambda^2 \sum_i (\lambda_i^2 + \lambda)^{-2}\lambda_i \,|_{\lambda=\sigma^2}$$

$$\leqslant \frac{M_0}{2\theta}[\lambda^{1/2}[\arctan\frac{1}{\sqrt{\lambda}} + \frac{\sqrt{\lambda}}{1+\lambda}] + O(1)] \,|_{\lambda=\sigma^2} \leqslant \frac{M_0\pi}{4\theta}\sigma + o(\sigma).$$

where the last inequality follows from the second equation in (3.10). Thus,

$$\min_{\lambda>0} e^{H_G}(\lambda) \leqslant e^{H_G}(\sigma^2) = [\sigma^2 A(\lambda) + \lambda^2 B(\lambda)] \,|_{\lambda=\sigma^2} \leqslant (1 + M_0)\frac{\pi}{4\theta}\sigma + O(\sigma^2).$$

Similarly, recall that $e^{L_\rho^2}(\sigma) = \sigma^2\widetilde{A}(\sigma) + \sigma^2\widetilde{B}(\sigma)$ with $\widetilde{A}$ and $\widetilde{B}$ defined in (3.7). Note that

$$\widetilde{B}(\lambda) = \sum_i (\lambda_i + \lambda)^{-2}c_i^2 = \sum_i (\lambda_i + \lambda)^{-2}\lambda_i\lambda_i^{-1}c_i^2 \leqslant \sum_i (\lambda_i + \lambda)^{-2}\lambda_i M_0 = M_0\widetilde{A}(\lambda).$$

Meanwhile, the upper bound for $\widetilde{A}$ in (3.10) implies that

$$\sigma^2\widetilde{A}(\sigma) \leqslant \sigma^2 \left[\frac{1}{\theta\sigma(1+\sigma)} + O(1) \right] \leqslant \frac{1}{\theta}\sigma + O(\sigma^2).$$

Hence, $e^{L_\rho^2}(\sigma) \leqslant (1 + M_0)\sigma^2\widetilde{A}(\sigma) \leqslant (1 + M_0)\frac{2}{\theta}\sigma + O(\sigma^2)$.

Part (b). Substituting $c_i^2 = \lambda_i$ into the equation for $B_1(\lambda)$ in (3.9), we have $B_1(\lambda) = \sum_i (\lambda_i^2 + \lambda)^{-3}\lambda_i^3 = \frac{-1}{2}A'(\lambda)$. Hence, Eq. (3.8) implies that $\lambda_{opt} = \sigma^2$. Also, substituting $c_i^2 = \lambda_i$ into the equation for $B(\lambda)$ in (3.7), we obtain

$$B(\lambda) = \sum_i (\lambda_i^2 + \lambda)^{-2}\lambda_i = \frac{1}{2\theta}\lambda^{-3/2}[\arctan\frac{1}{\sqrt{\lambda}} + \frac{\sqrt{\lambda}}{1+\lambda}] + O(1).$$

where the second equality follows from the second equation in (3.10) when λ is small (since we will set it to be σ^2). This equation, together with the first equation in (3.10) and (3.6), implies

$$\min_{\lambda>0} e^{H_G}(\lambda) = e^{H_G}(\sigma^2) = [\sigma^2 A(\lambda) + \lambda^2 B(\lambda)] \,|_{\lambda=\sigma^2}$$

$$= \left[\frac{\sigma^2}{2\theta\sqrt{\lambda}}(\arctan\frac{1}{\sqrt{\lambda}} - \frac{\sqrt{\lambda}}{1+\lambda}) + \frac{1}{2\theta}\sqrt{\lambda}(\arctan\frac{1}{\sqrt{\lambda}} - \frac{\sqrt{\lambda}}{1+\lambda}) \right] \Big|_{\lambda=\sigma^2} + O(\sigma^2)$$

$$= \sigma^2\frac{1}{2\theta\sqrt{\lambda}}\arctan\frac{1}{\sqrt{\lambda}} \,|_{\lambda=\sigma^2} + O(\sigma^2) = \frac{\pi}{4\theta}\sigma + O(\sigma^2). \tag{3.12}$$

14

Similarly, to find the minimizer of $e^{L_\rho^2}(\lambda)$, we substitute $c_i^2 = \lambda_i$ into the equation for $\widetilde{B}_1(\lambda)$ in (3.9), we have $\widetilde{B}_1(\lambda) = \sum_i (\lambda_i + \lambda)^{-3}\lambda_i^2 = \frac{1}{2\theta\lambda(1+\lambda)^2} + O(1)$. Consequently, by the equation for $\widetilde{A}'(\lambda)$ in (3.10), the minimizer $\widetilde{\lambda}_{opt}$ satisfies $\lambda = -\sigma^2 \frac{\widetilde{A}'(\lambda)}{2\widetilde{B}_1(\lambda)}$, and it can be solved from

$$2\lambda\widetilde{B}_1(\lambda) = -\sigma^2\widetilde{A}'(\lambda)$$

$$\Leftrightarrow \qquad \frac{1}{\theta(1+\lambda)^2} + 2\lambda O(1) = \sigma^2 \frac{1+2\lambda}{\theta\lambda^2(1+\lambda)^2} + \sigma^2 O(1)$$

$$\Leftrightarrow \qquad \lambda^2 = \sigma^2 + 2\lambda\sigma^2 + O(\lambda^2\sigma^2 + \lambda^3).$$

The positive solution to $\lambda^2 = \sigma^2 + 2\lambda\sigma^2$ is $\sigma^2 + \sigma\sqrt{1+\sigma^2}$. Thus, the minimizer is $\widetilde{\lambda}_{opt} = \sigma + O(\sigma^2)$. Together with the fact that $\widetilde{B}(\lambda) = \sum_i (\lambda_i + \lambda)^{-2}\lambda_i = \widetilde{A}(\lambda)$, we obtain

$$\min_{\lambda > 0} e^{L_\rho^2}(\lambda) = e^{L_\rho^2}(\lambda_{opt}) = \widetilde{A}(\lambda)(\sigma^2 + \lambda^2)\big|_{\lambda = \lambda_{opt}}$$

$$= \frac{1}{\theta\lambda(1+\lambda)}(\sigma^2 + \lambda^2) + O(\sigma^2)\big|_{\lambda = \lambda_{opt}}$$

$$= \frac{1}{\theta}\frac{2\sigma^2 + O(\sigma^3))}{(\sigma + O(\sigma^2))(1 + \sigma + O(\sigma^2))} + O(\sigma^2) = \frac{2}{\theta}\sigma + O(\sigma^2),$$

where we used Eq.(3.10) to obtain the third equality.

Part (c). The upper bounds in Part (d) follow directly from evaluations of the series using the uniform bound.

Part (d). The proof is similar to Part (a) with slightly more computations. First, the same argument in Part (a) leads to $\lambda_{opt} = \sigma^2$. To evaluate $e^{H_G}(\sigma^2) = [\sigma^2 A(\lambda) + \lambda^2 B(\lambda)]\,|_{\lambda = \sigma^2}$, note first that substituting $c_i^2 = \lambda_i$ into the equation for $B(\lambda)$ in (3.7) and using (3.11), we obtain

$$B(\lambda) = \sum_i (\lambda_i^2 + \lambda)^{-2}\lambda_i = B_c(\lambda) = \frac{1}{2\theta}\lambda^{-\frac{1}{2}(3+\frac{1}{\theta})}[C_\theta(1,2,1) + o(1)],$$

$$A(\lambda) = \frac{1}{2\theta}\lambda^{-\frac{1}{2}(1+\frac{1}{\theta})}[C_\theta(1,2,3) + o(1)].$$

Consequently, we obtain

$$\min_{\lambda > 0} e^{H_G}(\lambda) = e^{H_G}(\sigma^2) = \sigma\left(\frac{1}{2\theta}[C_\theta(1,2,1) + C_\theta(1,2,3)] + o(1)\right) := \sigma(C_{H_G} + o(1)). \qquad (3.13)$$

Thus, we have $C_{H_G} = \frac{1}{2\theta}[C_\theta(1,2,1) + C_\theta(1,2,3)]$, which we will compute its value after obtaining a similar expression for C_{L^2}.

To find the minimizer of $e^{L_\rho^2}(\lambda)$, we solve an algebraic equation as in Part (a). Substituting $c_i^2 = \lambda_i$ into the equation for $\widetilde{B}_1(\lambda)$ in (3.9), we have

$$\widetilde{B}_1(\lambda) = \sum_i (\lambda_i + \lambda)^{-3}\lambda_i^2 = \widetilde{B}_c(\lambda) = \frac{1}{2\theta}\lambda^{-\frac{1}{\theta}-1}[C_\theta(0,3,2) + o(1)]$$

using (3.11). Also, we have $\widetilde{A}'(\lambda) = -\frac{1}{\theta}\lambda^{-2-\frac{1}{\theta}}[C_\theta(0,3,1) + o(1)]$. Hence, since the minimizer $\widetilde{\lambda}_{opt}$ satisfies $\lambda = -\sigma^2 \frac{\widetilde{A}'(\lambda)}{2\widetilde{B}_1(\lambda)}$, it can be solved from

$$2\lambda\widetilde{B}_1(\lambda) = -\sigma^2\widetilde{A}'(\lambda)$$

$$\Leftrightarrow \qquad \lambda^{-\frac{1}{\theta}}[C_\theta(0,3,2) + o(1)] = \sigma^2\lambda^{-2-\frac{1}{\theta}}[C_\theta(0,3,1) + o(1)].$$

15

Thus, the minimizer is

$$\widetilde{\lambda}_{opt} = \sigma[C_\theta(0,3,2)^{-1/2}C_\theta(0,3,1)^{1/2} + o(1)] =: \sigma(C_{\widetilde{\lambda}_{opt}} + o(1)). \tag{3.14}$$

Together with the fact that $\widetilde{B}(\lambda) = \sum_i(\lambda_i + \lambda)^{-2}\lambda_i = \widetilde{A}(\lambda) = \frac{1}{\theta}\lambda^{-1-\frac{1}{\theta}}[C_\theta(1,2,3) + o(1)]$ as in Eq.(3.11), we obtain

$$\min_{\lambda>0} e^{L_\rho^2}(\lambda) = e^{L_\rho^2}(\lambda_{opt}) = \widetilde{A}(\lambda)(\sigma^2 + \lambda^2)\big|_{\lambda=\lambda_{opt}}$$

$$= \sigma\frac{1}{\theta}C_\theta(1,2,3)[(1 + C_\theta(0,3,2)^{-1}C_\theta(0,3,1)) + o(1)] =: \sigma(C_{L_\rho^2} + o(1)).$$

At last, we evaluate the constants C_{H_G} in (3.13), $C_{\widetilde{\lambda}_{opt}}$ in (3.14) and $C_{L_\rho^2}$ above. In summary, we have

$$C_{H_G} = \frac{1}{2\theta}[C_\theta(1,2,1) + C_\theta(1,2,3)],$$

$$C_{\widetilde{\lambda}_{opt}} = C_\theta(0,3,2)^{-1/2}C_\theta(0,3,1)^{1/2},$$

$$C_{L_\rho^2} = \frac{1}{\theta}C_\theta(1,2,3)[1 + C_\theta(0,3,2)^{-1}C_\theta(0,3,1)].$$

Here we denote $C_\theta(s,k,\alpha) = \Gamma(\gamma)\Gamma(k-\gamma)$ with $\gamma = \frac{1}{1+s}(\alpha - \frac{1}{\theta})$, where $s \in \{0,1\}$, $\alpha \in \{1+2s, 1+s\}_{s\in\{0,1\}} = \{1,2,3\}$, and $k \in \{2,3\}$.

Noting that $\gamma = \frac{1}{1+s}(\alpha - \frac{1}{\theta}) = \begin{cases} \frac{1}{2}(3 - \frac{1}{\theta}), & \text{if } (s,k,\alpha) = (1,2,3); \\ \frac{1}{2}(1 - \frac{1}{\theta}), & \text{if } (s,k,\alpha) = (1,2,1), \end{cases}$ and using the formula $\Gamma(z+1) = z\Gamma(z)$, we obtain

$$C_\theta(1,2,3) = \Gamma(\frac{3}{2} - \frac{1}{2\theta})\Gamma(\frac{1}{2} + \frac{1}{2\theta}) = \frac{1}{2}(1 - \frac{1}{\theta})\Gamma(\frac{1}{2} - \frac{1}{2\theta})\Gamma(\frac{1}{2} + \frac{1}{2\theta}),$$

$$C_\theta(1,2,1) = \Gamma(\frac{1}{2} - \frac{1}{2\theta})\Gamma(\frac{3}{2} + \frac{1}{2\theta}) = \frac{1}{2}(1 + \frac{1}{\theta})\Gamma(\frac{1}{2} - \frac{1}{2\theta})\Gamma(\frac{1}{2} + \frac{1}{2\theta})$$

Hence,

$$C_{H_G} = \frac{1}{2\theta}C_\theta(1,2,3)[1 + C_\theta(1,2,3)^{-1}C_\theta(1,2,1)] = \frac{1}{2\theta}\Gamma(\frac{1}{2} - \frac{1}{2\theta})\Gamma(\frac{1}{2} + \frac{1}{2\theta}).$$

Similarly, noting that $\gamma = \frac{1}{1+s}(\alpha - \frac{1}{\theta}) = \begin{cases} 2 - \frac{1}{\theta}, & \text{if } (s,k,\alpha) = (0,3,2); \\ 1 - \frac{1}{\theta}, & \text{if } (s,k,\alpha) = (0,3,1), \end{cases}$ we have

$$C_\theta(0,3,2) = \Gamma(2 - \frac{1}{\theta})\Gamma(1 + \frac{1}{\theta}) = (1 - \frac{1}{\theta})\Gamma(1 - \frac{1}{\theta})\Gamma(1 + \frac{1}{\theta}),$$

$$C_\theta(0,3,1) = \Gamma(1 - \frac{1}{\theta})\Gamma(2 + \frac{1}{\theta}) = (1 + \frac{1}{\theta})\Gamma(1 - \frac{1}{\theta})\Gamma(1 + \frac{1}{\theta}).$$

Consequently,

$$C_{\widetilde{\lambda}_{opt}} = C_\theta(0,3,2)^{-1/2}C_\theta(0,3,1)^{1/2} = \frac{\theta + 1}{\theta - 1},$$

$$C_{L_\rho^2} = \frac{1}{\theta}C_\theta(1,2,3)[1 + C_\theta(0,3,2)^{-1}C_\theta(0,3,1)] = \frac{1}{\theta}\Gamma(\frac{1}{2} - \frac{1}{2\theta})\Gamma(\frac{1}{2} + \frac{1}{2\theta}).$$

Clearly, $C_{H_G} = \frac{1}{2}C_{L_\rho^2}$ and we complete the proof. ∎

16

4 Algorithm and numerical examples

We present a direct matrix-decomposition-based algorithm to implement the data-adaptive RKHS regularization. We compare the RKHS regularizer with two commonly used regularizers: the L^2-regularizer that sets $\|\phi\|_\square = \|\phi\|_{L^2_\rho}$ in (1.3) and the l^2-regularizer that sets $\|\phi\|_\square = (\sum_{i=1}^n \phi(s_i)^2)^{1/2}$, which is the l^2 norm of $\boldsymbol{\phi} = (\phi(s_1), \dots, \phi(s_n))$, the discrete representation of ϕ on $\mathcal{S}$.

Numerical settings. We consider (1.1) with $\mathcal{S} := \{s_k\}_{k=1}^n \subset [a, b]$, where $\delta := s_k - s_{k-1} = (b-a)/n$ for $k \geqslant 1$ with the convention $s_0 = a$, that is, $s_k = a + k\delta$ is a uniform mesh. We set $\nu(\{s_k\}) = \delta$ so that (1.1) can be viewed as a discretization of the following Fredholm integral equation

$$y(t) = \int_a^b K(t, s)\phi(s)ds + \sigma\dot{W}(t) =: L_K\phi(t) + \sigma\dot{W}(t), t \in [c, d].$$

We consider two specific kernels:

$$
\begin{aligned}
K_{exp}(t, s) &= s^{-2}e^{-st}; \\
K_{poly}(t, s) &= s^{-1}|\sin(xt + 1)|,
\end{aligned}
\tag{4.1}
$$

where the kernels K_{exp} and K_{poly} lead to operators with exponentially and polynomially decaying spectra, respectively (see Figure 1). Here the kernel K_{exp} arises from the $T2$ problem in magnetic resonance relaxometry [3], where ϕ is the distribution of transverse nuclear relaxation times, and it characterizes the material composition of a sample.

We aim to recover $\boldsymbol{\phi} = (\phi(s_1), \dots, \phi(s_n)) \in \mathbb{R}^n$. This discrete vector provides a piecewise-constant approximation to the function ϕ: $\phi(s) = \sum_{k=1}^n \phi(s_k)\mathbf{1}_{[s_{k-1}, s_k]}(s)$, where $\mathbf{1}_A(s)$ is the indicator function of the set A. Then, Eq.(1.1) can be viewed as a Riemann sum approximation of the above integral:

$$\mathbf{y} = \mathbf{L}\boldsymbol{\phi} + \mathbf{w}, \tag{4.2}$$

where $\mathbf{w} \in \mathbb{R}^m \sim \mathcal{N}(0, \sigma^2\Delta t I_m)$ and

$$
\mathbf{L} = \begin{bmatrix}
K(t_1, s_1)\delta_1 & K(t_1, s_2)\delta_2 & K(t_1, s_3)\delta_3 & \cdots & K(t_1, s_n)\delta_n \\
K(t_2, s_1)\delta_1 & K(t_2, s_2)\delta_2 & K(t_2, s_3)\delta_3 & \cdots & K(t_2, s_n)\delta_n \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
K(t_m, s_1)\delta_1 & K(t_m, s_2)\delta_2 & K(t_m, s_3)\delta_3 & \cdots & K(t_m, s_n)\delta_n
\end{bmatrix}
\tag{4.3}
$$

with $K(t, s)$ being either $K_{exp}(t, s)$ or $K_{poly}(t, s)$.

We set $(a, b, c, d) = (1, 5, 0, 5)$ and $n = 100$. We set the standard deviation of the noise to be $\sigma = \|\mathbf{L}\boldsymbol{\phi}\| \times nsr$, where the noise-to-signal ratio nsr is set to be $nsr = 1$ unless otherwise specified. For the data $\mathbf{y} = (y(t_1), \dots, y(t_m)) \in \mathbb{R}^m$, we consider equally spaced grid $t_i = c + i\Delta t$, $0 \leqslant i \leqslant m$ with $\Delta t = (d - c)/m = 0.01$, i.e., with $m = 500$, unless otherwise specified.

The space $L^2_\rho(\mathcal{S})$. The empirical exploration measure of (4.2) is

$$\rho(s_k) = \frac{1}{z}\sum_{i=1}^m s_k^{-2}e^{-s_k t_i}\delta, \quad k = 1, \dots, n, \tag{4.4}$$

where z is the normalizing constant. In general, when the true kernel K is available, one may compute the exploration measure analytically and then evaluate $\rho(s_k)$ for each s_k in $\mathcal{S}$. For example, the analytical exploration measure for K_{exp} is $\rho(s) \propto \int_c^d s^{-2}e^{-st}dt = s^{-3}(e^{-cs} - e^{-ds})$.

Our goal is to estimate $\boldsymbol{\phi} \in L^2_\rho(\mathcal{S}) \subset \mathbb{R}^n$, which is equipped with the norm $\|f\|^2_{L^2_\rho(\mathcal{S})} = \sum_{k=1}^n f_k^2 \rho(s_k)$ (and we consider only the values on the support of ρ). The support of ρ is where the operator explores the vector $\boldsymbol{\phi}$, for which we have information for an estimation. Thus, $L^2_\rho(\mathcal{S})$ is a proper space with a metric for estimating $\boldsymbol{\phi} \in \mathbb{R}^n$.

The discrete $\boldsymbol{\phi}$ uses the Cartesian basis $\{\mathbf{e}_i\}_{i=1}^n$ of $\mathbb{R}^n$ such that $(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n) = I_n$. In $L^2_\rho(\mathcal{S})$, these basis vectors have a (diagonal) basis matrix

$$\mathbf{B} = (\langle \mathbf{e}_i, \mathbf{e}_j \rangle_{L^2_\rho(\mathcal{S})}) = \text{diag}(\rho(s_k)). \tag{4.5}$$

and its inner product is

$$\langle \boldsymbol{\phi}, \boldsymbol{\psi} \rangle_{L^2_\rho} := \boldsymbol{\phi}^T \mathbf{B} \boldsymbol{\psi}, \quad \boldsymbol{\phi}, \boldsymbol{\psi} \in \mathbb{R}^n.$$

4.1 Algorithm: regularization with adaptive RKHS

We estimate the vector $\boldsymbol{\phi}$ in (4.2) by least squares, which minimizes the loss functional

$$\widehat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi} \in \mathbb{R}^n} \mathcal{E}(\boldsymbol{\phi}), \qquad \mathcal{E}(\boldsymbol{\phi}) = \|\mathbf{y} - \mathbf{L}\boldsymbol{\phi}\|^2 = \sum_{i=1}^m |y_i - (\mathbf{L}\boldsymbol{\phi})_i|^2. \tag{4.6}$$

With $\mathbf{A}^\dagger$ denoting the pseudo-inverse of $\mathbf{A}$, we compute a minimizer as

$$\widehat{\boldsymbol{\phi}} = \mathbf{A}^\dagger \mathbf{b}, \quad \text{where } \mathbf{A} = \mathbf{L}^\top \mathbf{L}, \ \mathbf{b} = \mathbf{L}^\top \mathbf{y}. \tag{4.7}$$

However, this inverse problem is ill-posed. The ill-posedness is rooted in the integral equation (1.1), as its solution involves the inversion of a compact operator with eigenvalues converging to zero [16]. Computationally, the ill-posedness is seen in the ill-conditionedness of the matrix $\mathbf{A}$.

Figure 1 shows the eigenvalues of $\mathbf{A}$ and the operator $\mathcal{L}_{\overline{G}} : L^2_\rho(\mathcal{S}) \to L^2_\rho(\mathcal{S})$. Here the eigenvalues of $\mathbf{A}$ are computed by singular value decomposition (SVD) to achieve better numerical stability than the eigenvalue decomposition. The eigenvalues of $\mathcal{L}_{\overline{G}}$ are computed by using the generalized eigenvalue problem (see [14, Theorem 4.1] or [4, Proposition 5.6])

$$\mathbf{A}V = \mathbf{B}\Lambda V, \quad s.t., V^\top \mathbf{B} V = I_n, \quad \Lambda = \text{Diag}(\lambda_1, \dots, \lambda_n), \tag{4.8}$$

which is computed by (symmetric) generalized eigenvalue decomposition [15]. We observe that the generalized SVD method, while producing stable generalized eigenvalues for these symmetric matrices, does not yield generalized eigenvectors satisfying the above equations. Hence, to take advantage of $\mathbf{A}$ and $\mathbf{B}$ being symmetric positive semi-definite, we use the generalized eigenvalue decomposition algorithm that yields eigenvectors satisfying these equations.
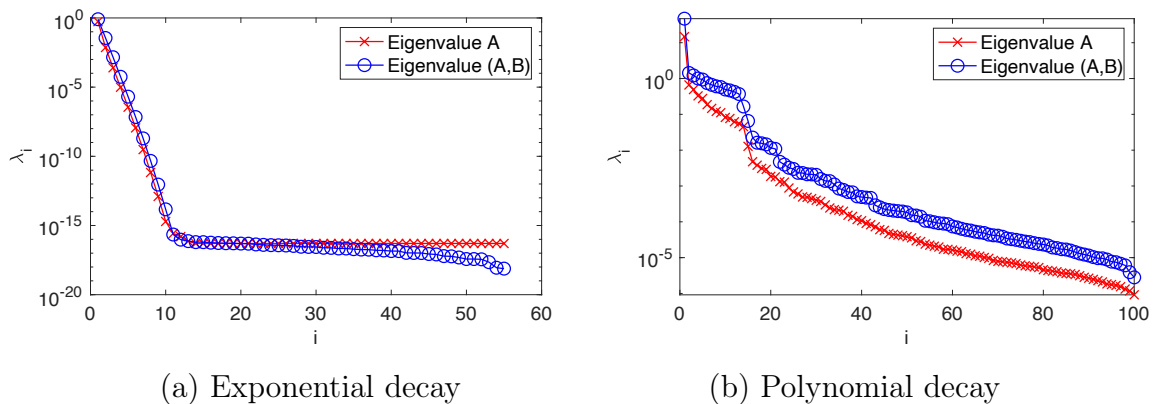

(a) Exponential decay (b) Polynomial decay

Figure 1: Eigenvalues of $\mathbf{A}$ and eigenvalues of $\mathcal{L}_{\overline{G}}$ corresponding to the two kernels in (4.1). Here the eigenvalues of $\mathbf{A}$ are computed by singular value decomposition and the eigenvalues of $\mathcal{L}_{\overline{G}}$ are computed as the generalized eigenvalues of $(\mathbf{A}, \mathbf{B})$.

Figure 1 shows that the eigenvalues of $\mathbf{A}$ and $\mathcal{L}_{\overline{G}}$ decay exponentially in i, providing an example of the assumption on λ_i in Theorem 3.1. In particular, the fast-decaying eigenvalues show that the conditional numbers of $\mathbf{A}$ and $\mathcal{L}_{\overline{G}}$ are extremely large, indicating severe ill-posedness. As a result, regularization is necessary to produce a stable solution from the noisy observations.

We consider Tikhonov regularization that adds a quadratic penalty term $\lambda\|\phi\|_{\mathbf{C}}^2 = \lambda\phi^\top\mathbf{C}\phi$ to the loss function, and solves the minimizer by least squares

$$\widehat{\phi}_\lambda = (\mathbf{A} + \lambda\mathbf{C})^{-1}\mathbf{b} = \arg\min_{\phi\in\mathbb{R}^n} \mathcal{E}(\phi) + \lambda\|\phi\|_{\mathbf{C}}^2. \tag{4.9}$$

The hyper-parameter λ controls the strength of the regularization, and its selection has been thoroughly studied in [9, 10, 21]. We will select it by the L-curve method in [10]. The L-curve is a log-log plot of the curve $l(\lambda) = (y(\lambda), x(\lambda))$ with $y(\lambda)^2 = \|\phi_\lambda\|_{\mathbf{C}}^2$ and $x(\lambda)^2 = \mathcal{E}(\phi_\lambda)$ with ϕ_λ in (4.9). The L-curve method maximizes the curvature to reach a balance between the minimization of the likelihood and the control of the regularization:

$$\lambda_* = \mathrm{argmax}_{\lambda_{\min}\leqslant\lambda\leqslant\lambda_{\max}}\kappa(l(\lambda)), \quad \kappa(l(\lambda)) = \frac{x'y'' - x'y''}{(x'^{\,2} + y'^{\,2})^{3/2}},$$

where λ_{min} and λ_{max} are the smallest and the largest eigenvalues of $\mathbf{A}$.

The main task is to select the norm $\|\cdot\|_{\mathbf{C}}$. We consider the norm of the RKHS H_G (which we explain in detail below), and we compare it with two commonly used norms: the Euclidean norm with $\mathbf{C} = \mathbf{I}$ and the $L_\rho^2(\mathcal{S})$ norm with $\mathbf{C} = \mathbf{B}$ in (4.5). We refer to them as the RKHS, l^2, and $L_\rho^2(\mathcal{S})$ regularization, respectively. They are summarized in Table 1.

Table 1: Three regularizers using the norms of l^2, $L_\rho^2(\mathcal{S})$ and RKHS.

Regularizer name	$\mathbf{C}$	Regularized estimator
l2	$\mathbf{I}$	$\phi_\lambda^{l^2} = (\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{b}$
L2	$\mathbf{B}$	$\phi_\lambda^{L^2} = (\mathbf{A} + \lambda\mathbf{B})^{-1}\mathbf{b}$
RKHS	$\mathbf{C}_{rkhs}$	$\phi_\lambda^{H_G} = (\mathbf{A} + \lambda\mathbf{C}_{rkhs})^{-1}\mathbf{b}$

The RKHS regularization. Let $\{(\lambda_i, \psi_i)\}_{i=1}^n$ be the eigenvalue and eigen-function of $\mathcal{L}_{\overline{G}}$ over $L_\rho^2(\mathcal{S})$, and recall that $\{\psi_i\}$ form an orthonormal basis of $L_\rho^2(\mathcal{S})$. Here they are solved by the generalized eigenvalue problem in (4.8), thus, $\psi_i = \sum_j V_{ji}\mathbf{e}_j$ and $\mathbf{\Psi} = (\psi_1, \ldots, \psi_n) = \mathbf{I}V$. Then, $\phi = \mathbf{I}\phi = \mathbf{\Psi}V^{-1}\phi$. Thus, the H_G norm of ϕ is

$$\begin{aligned}
\|\phi\|_{H_G}^2 &= \langle\phi, \mathcal{L}_{\overline{G}}^{-1}\phi\rangle_{L_\rho^2(\mathcal{S})} = \langle\mathbf{\Psi}V^{-1}\phi, \mathcal{L}_{\overline{G}}^{-1}\mathbf{\Psi}V^{-1}\phi\rangle_{L_\rho^2(\mathcal{S})} \\
&= (V^{-1}\phi)^\top\langle\mathbf{\Psi}, \mathcal{L}_{\overline{G}}^{-1}\mathbf{\Psi}\rangle_{L_\rho^2(\mathcal{S})}V^{-1}\phi = \phi^\top(V^{-1})^\top\Lambda^\dagger V^{-1}\phi \\
&= \|\phi\|_{\mathbf{C}_{rkhs}}^2 \quad \text{with } \mathbf{C}_{rkhs} := (V^{-1})^\top\Lambda^\dagger V^{-1}.
\end{aligned}$$

In particular, if ρ is a uniform measure and the basis matrix is $\mathbf{B} = I_n$, we have $\mathbf{C}_{rkhs} = \mathbf{A}^\dagger$.

In computation, we improve numerical stability by avoiding the pseudo-inverse of a singular matrix. The procedure is as follows. Note that for $\mathbf{C}_* := V\Lambda^{1/2}$, we have $\mathbf{C}_*^\top\mathbf{C}_{rkhs}\mathbf{C}_* = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} := \mathbf{I}_r$, where I_r is the identity matrix with rank r, the number of positive eigenvalues in

Λ. Then, the linear equation $(\mathbf{A} + \lambda \mathbf{C}_{rkhs})c_\lambda = \mathbf{b}$, the equation (4.9) with regularization matrix $\mathbf{C}_{rkhs}$, is equivalent to

$$(\mathbf{C}_* \mathbf{A} \mathbf{C}_* + \lambda \mathbf{I}_r)\widetilde{\phi}_\lambda = \mathbf{C}_* \mathbf{b} \tag{4.10}$$

with $\widetilde{\phi}_\lambda = \mathbf{C}_*^{-1}\phi_\lambda$. We compute $\widetilde{\phi}_\lambda$ in the above equation by least squares with the minimal norm; then we obtain $\phi_\lambda = \mathbf{C}_*\widetilde{\phi}_\lambda$. These treatments avoid the inversions of ill-conditioned or singular matrices and lead to more robust estimators.

The implementation of the RKHS regularization based on matrix decomposition is summarized in Algorithm 1.

Input: The regression triplet $(\mathbf{A}, \mathbf{b}, \mathbf{B})$ consisting of normal matrix $\mathbf{A}$, vector $\mathbf{b}$ in (4.7) and basis matrix $\mathbf{B}$ as in (4.5).
Output: An RKHS regularized estimator $\widehat{\phi}_{\lambda_0}$ and its loss value $\mathcal{E}(\widehat{\phi}_{\lambda_0})$.
 1: Solve the generalized eigenvalue problem $\mathbf{A}V = \mathbf{B}V\Lambda$, $V^\top \mathbf{B}V = I$.
 2: Compute $\widetilde{\mathbf{A}} = \mathbf{C}_*^\top \mathbf{A}\mathbf{C}_*$ and $\widetilde{\mathbf{b}} = \mathbf{C}_*\mathbf{b}$ with $\mathbf{C}_* = V\Lambda^{1/2}$.
 3: Use the L-curve method to find an optimal estimator $\widehat{\phi}_{\lambda_0}$:

- Set the range for λ to be the range of the eigenvalues in Λ.

- For each λ, solve $\widetilde{\phi}_\lambda$ from $(\widetilde{\mathbf{A}} + \lambda I)\widetilde{\phi}_\lambda = \widetilde{\mathbf{b}}$ by least squares with minimal norm and set $\widehat{\phi}_\lambda = \mathbf{C}_*\widetilde{\phi}_\lambda$.

- Select λ_0 maximizing the curvature of the λ-curve $(\log \mathcal{E}(\widehat{\phi}_\lambda), \log(\widehat{\phi}_\lambda^\top \mathbf{C}_{rkhs}\widehat{\phi}_\lambda))$.

- Return the solution $\widetilde{\phi}_{\lambda_0}$ solving $(\widetilde{\mathbf{A}} + \lambda_0 I)\widetilde{\phi}_{\lambda_0} = \widetilde{\mathbf{b}}$.

Algorithm 1: Solving $\mathbf{A}\phi \approx \mathbf{b}$ with basis matrix $\mathbf{B}$ by the adaptive RKHS Regularization.

Function space of identifiability (FSOI). Recall that the FSOI H is the linear subspace of $L_\rho^2(\mathcal{S})$ spanned by the eigenvectors of $\mathcal{L}_{\overline{G}}$ with nonzero eigenvalues. In computation, the FSOI is the eigenspace of $\mathcal{L}_{\overline{G}}$ spanned by the eigenvectors with positive eigenvalues (greater than the numerical precision in practice). Figure 1 suggests that the FSOI is low-dimensional and contains only a few eigenvectors.

Also, we access the performance of an estimator by its accuracy in estimating the projection of the true solution in the FSOI, since the data provides no information for estimating the components outside the FSOI. That is, we compute the error of an estimator ϕ as

$$Err = \|P_H\widehat{\phi} - P_H\phi_*\|_{L_\rho^2(\mathcal{S})}^2, \tag{4.11}$$

where P_H is the projection to H.

4.2 Convergence as the noise level decays

To test the convergence of the regularized estimators as the noise level decays, we set the noise-to-signal ratio to be $nsr \in \{0.125, 0.25, 0.5, 1, 2\}$ and $\Delta t = 0.01$. Recall that $\sigma = \|\mathbf{L}\phi\| \times nsr$.

We consider both the scenario of $\phi_{true} \in$ FSOI and the scenario of $\phi_{true} \notin$ FSOI or nearly not in the FSOI in the sense that it has significant components in the eigenspace of small eigenvalues. We call these scenarios "in FSOI" and "not in FSOI" for simplicity. For the first scenario, we set $\phi_{true} = \psi_2$, the second eigenvector of $\mathcal{L}_{\overline{G}}$. For the second scenario, we consider $\phi_{true} = x^2$. We note that the latter is relatively challenging to recover because of ϕ_{true}'s components in eigenspace with small eigenvalues.
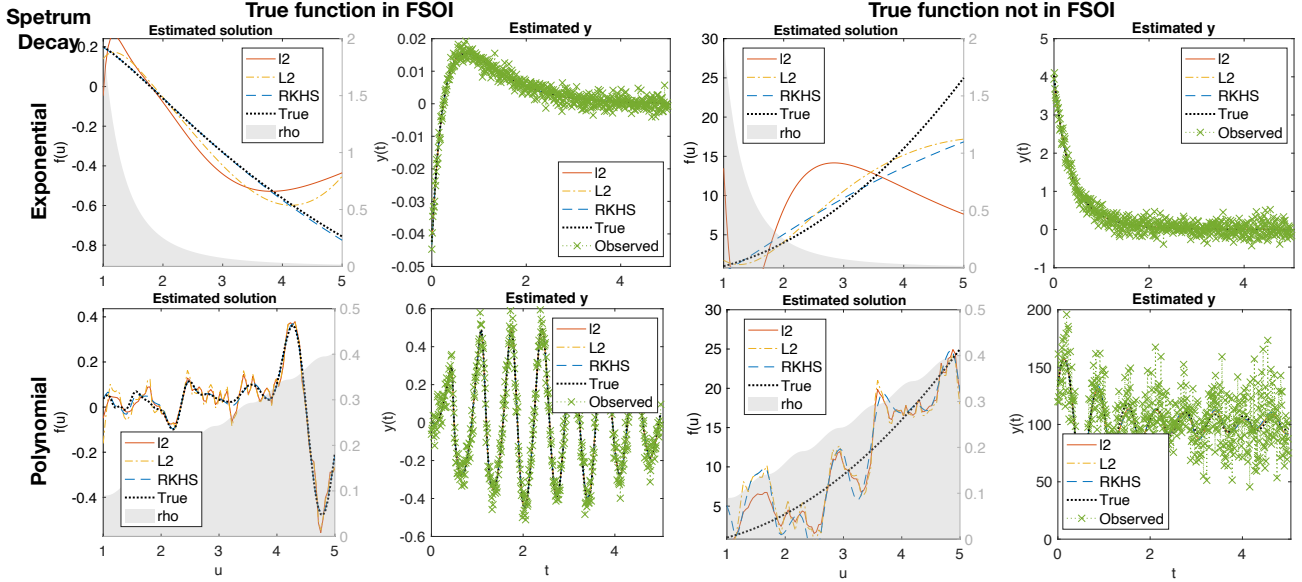
Figure 2: Typical regularized estimators when $nsr = 2$ and their recovery of the signal. All estimators recover the true signal accurately. The RKHS regularizer significantly outperforms the other two regularizers when the true function is in the FSOI (see top-left and bottom-left), but it may slightly underperform the $L_\rho^2(\mathcal{S})$ regularizer when the true function is not in the FSOI (see top- and bottom-right).

Figure 2 shows the typical regularized estimators in the two scenarios and their de-noised output $\widehat{y} = \mathbf{L}\widehat{\phi}$. The left column shows that when the true solution is in the FSOI, the RKHS regularizer significantly outperforms the other two. However, all the regularized estimators can recover the true signal (or de-noise the data) accurately, even though these estimators may deviate largely from the true solution. The right column shows that when the true solution has components outside the FSOI, the RKHS regularizer can be less accurate than the $L_\rho^2(\mathcal{S})$ regularizer. At the same time, both significantly outperform the l^2 regularizer. Yet again, all regularized estimators can de-noise the data accurately even when they have large errors. Thus, this inverse problem is ill-defined, and one must restrict the inverse to be in the FSOI.

Figure 3 demonstrates the convergence of the regularized estimators and their loss values as the noise-to-signal-ratio decreases, with $nsr \in \{0.125, 0.25, 0.5, 1, 2\}$. It presents the mean and standard deviations of the regularized estimators' $L_\rho^2(\mathcal{S})$ errors and the loss values in 100 simulations, each with a different noise realization. The errors are computed for the estimators' projection inside the FSOI as in (4.11). When the true solution is inside the FSOI, the RKHS regularized estimator has errors significantly smaller than those of the other two regularizers, though it has the largest loss values. When the true solution is outside the FSOI, the RKHS regularized estimator has slightly smaller errors than the $L_\rho^2(\mathcal{S})$ regularizer, and both outperform the l^2 regularizer. Importantly, when the true solution is inside FOSI, the RKHS-regularized estimator shows a clear linear error decay in noise for $nsr \in \{0.25, 0.5, 1, 2\}$. In contrast, the L_ρ^2-regularized estimator has an error remaining at a constant level.

Notably in Figure 3, the RKHS-regularizer achieves a rate of convergence about $\sigma^{0.5}$ for the Err defined in (4.11), and this rate agrees with the theoretical rate for the squared error in Theorem 3.1. Such a close agreement between theory and practical computation is amazing since the theoretical rate uses the oracle optimal hyper-parameter based on the true solution, but the practical computation empirically selects the hyper-parameter by the L-curve method.
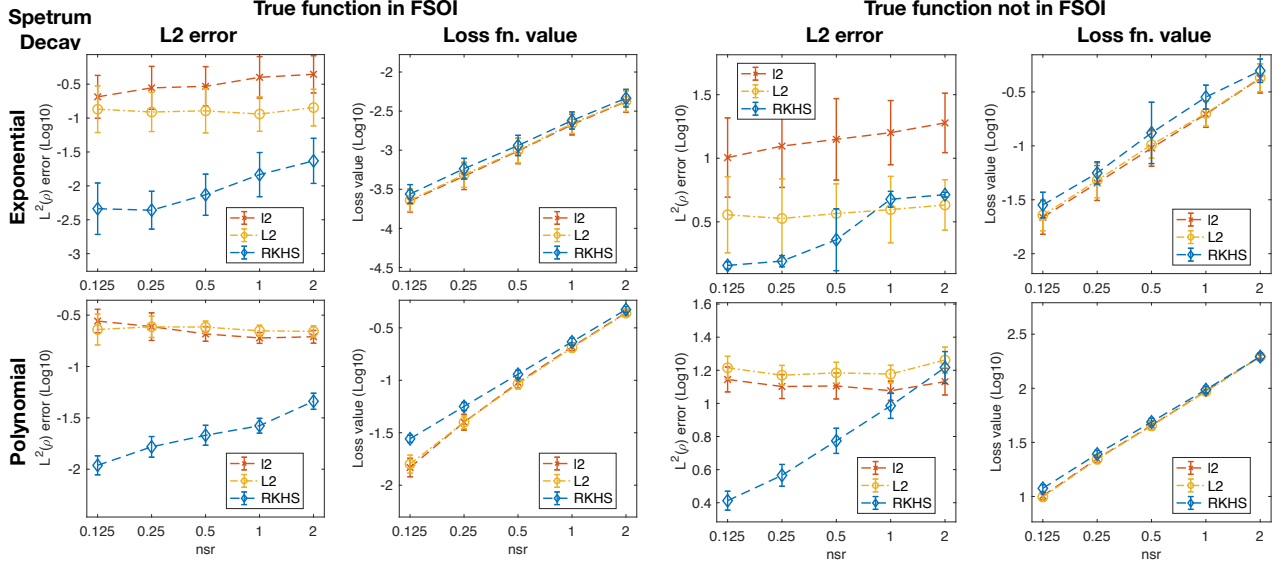
Figure 3: Mean and standard deviations of the projected error Err in (4.11) and the loss functional values in 100 simulations. The RKHS regularizer significantly outperforms the other two regularizers for $\phi_{true} \in$ FSOI (see top- and bottom-left). In all cases, the RKHS regularized estimator has errors consistently decaying as the noise level decreases, and it yields minimal loss functional values that are slightly larger than those of the l^2 and L^2-regularized estimators.

In contrast, the L^2-regularizer does not yield a convergent estimator. Since the two regularizers only differ at the choice of regularization norm $\|\phi\|_\square$, we conclude that the RKHS-norm provides a better metric for the L-curve than the L^2_ρ norm. Figure 4 shows that the L-curve in the RKHS-norm is in a better shape than the other two norms.
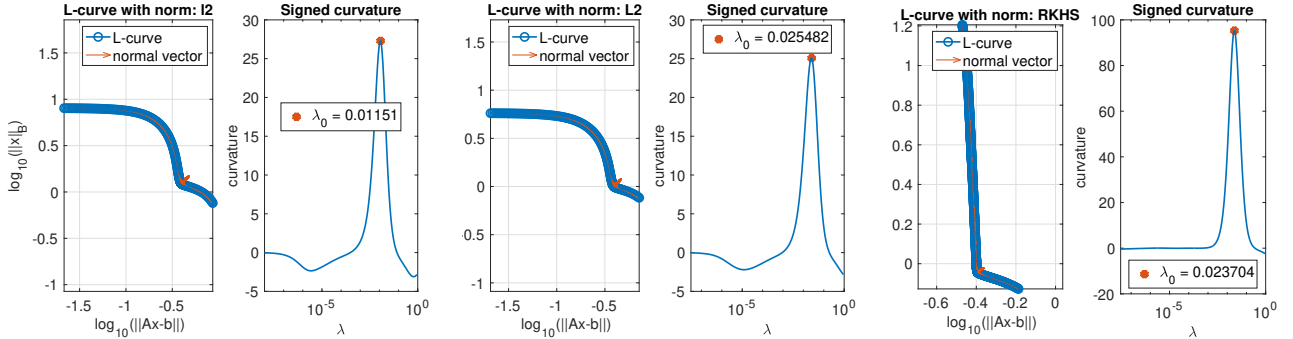

Figure 4: L-curves in a typical simulation for the l^2-, L^2- and RKHS-regularizers. The RKHS-regularizer has a better L-shaped curve than the other two regularizers, indicating that the RKHS-norm is a better metric for regularization than the other two norms.

4.3 Convergence as the observation mesh refines

Next, we test the convergence of the regularized estimator as the observation mesh refines. We set $\Delta t \in 0.005 \times \{1, 2, 4, 8, 16\}$ and $nsr = 1$.

We continue to consider the two scenarios as in Section 4.2 with the same choice of the true functions. However, we compute the FSOI using a finer observation mesh $\Delta t = 0.0005$.

Figure 5 demonstrates the convergence of the regularized estimators and their loss values as the observation mesh increases. It presents the mean and standard deviations of the projected error Err defined in (4.11) and the loss values in 100 simulations, each with an independent noise

realization. The RKHS regularized estimator has errors consistently decaying as the mesh refines, outperforming the other two regularizers when the true solution is inside the FSOI (top- and bottom-left), although its loss functional values are slightly larger. Also, the RKHS-regularizer slightly underperforms the L^2-regularizer when the true solution is outside the FSOI, and both outperforms the l^2 regularizer. Importantly, in all the cases, the RKHS regularized estimator has errors consistently decaying as the time mesh refines.
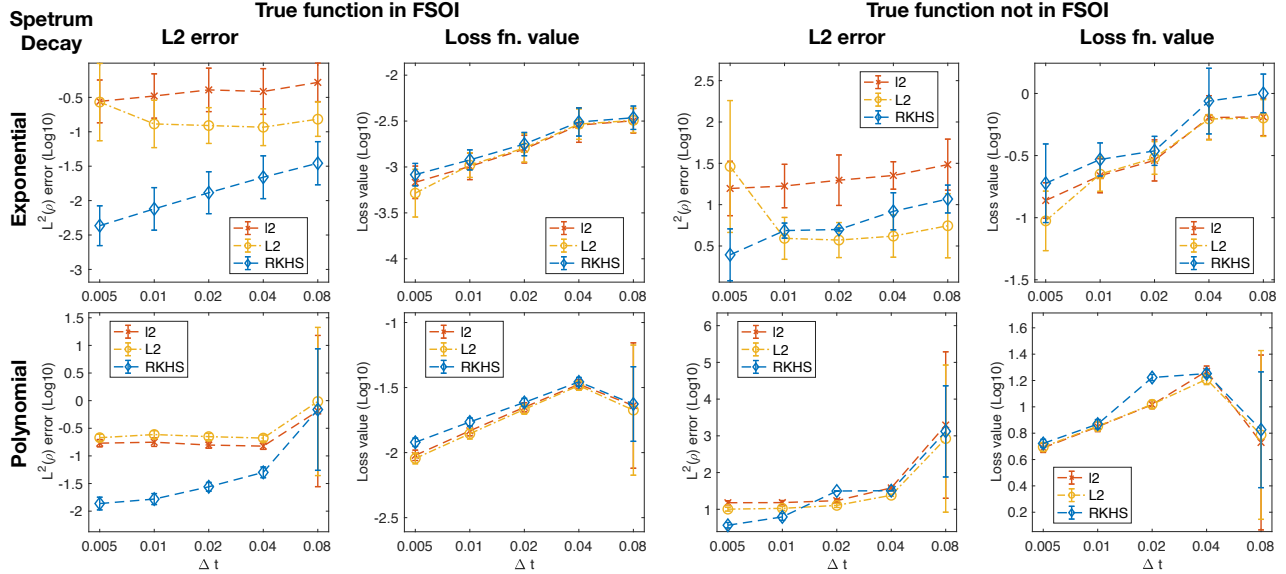
Figure 5: Mean and standard deviations of the projected error Err in (4.11) and the loss functional values in 100 independent simulations. The RKHS regularizer significantly outperforms the other two regularizers when $\phi_{true} \in$ FSOI (see top-and bottom-left). In all cases, only the RKHS regularized estimator has errors decaying consistently with the decrease of Δt.

5 Discussions and limitations

Comparative analysis. A major contribution of this study is the comparative analysis between the L^2_ρ-regularized estimator and the H_G-regularized estimator via the sharp convergence rate in the small noise limit. The sharp convergence rate of the expected L^2_ρ error depends on two factors: spectrum decay of the operator of inversion, and the speed of decay of the true function's cofficients. The key innovation is extracting these factors from multiple mixed factors in regularization. In particular, the expected error under the Gaussian noise assumption helps remove the randomness in measurement error, and the oracle optimal hyper-parameter minimizing the expected L^2_ρ error removes the uncertainty associated with the empirical selection of the hyper-parameter.

Our comparative analysis can be extended in two directions.

- *Comparing other Hilbert space norms for regularization.* In this study, we compared the RKHS-regularizer with the widely-used L^2-regularizer. But our small noise analysis can be applied to comparing other regularization norms as well, such as other RKHS norms with pre-selected reproducing kernels or the H^1 norm. However, it may not be suitable to analyze non-Hilbert space norms, since the analysis presented here relies on the spectral properties of the self-adjoint operator of inversion.

- *Convergence with respect to mesh refinement.* It remains open to studying the convergence

of the regularized estimator when the observation mesh refines. The challenge comes from a complicated balance between three elements: the discretization error, the observation noise-induced error, and the regularization-induced bias.

Practical computations. Another main contribution is the new adaptive RKHS regularization for solving the Fredholm integral equation of the first kind. This study implements the RKHS regularization by a matrix decomposition-based algorithm. Numerical results show that the RKHS-regularizer outperforms the widely used l^2- and L^2-regularizers. However, the matrix decomposition-based method will be computationally costly when the dimension of the matrix is large. For large-scale problems, iterative methods are necessary, and the challenge comes from the computation of the RKHS norm and proper early stopping strategy. These investigations are beyond the scope of this study, and we leave it for future work.

6 Conclusion and future work

We have introduced an adaptive RKHS regularization for ill-posed linear inverse problems. The adaptive RKHS arises in the variational formulation of the inverse problem, and it is determined by the linear operator and the observation mesh. Its closure is the space in which we can identify the true solution.

Importantly, we have introduced a quantitative comparative analysis to compare different regularization norms through the sharp convergence rate of the mean-square error of the regularized estimator in the small noise limit. We have proved that the RKHS-regularized and the widely-used L^2-regularized estimators have the same sharp convergence rate as $\sigma \to 0$, but the RKHS-regularized estimator has a smaller multiplicative factor.

Additionally, we present an implementation of the RKHS-regularization by an algorithm based on matrix decomposition. The algorithm uses the L-curve method to select the optimal hyperparameter. Numerical tests confirm the robust convergence of the RKHS-regularized estimator when either the noise level decays or the observation mesh refines, no matter whether the true function is in the FSOI or not. In the tests, the RKHS regularizer outperforms the widely used l^2- and L^2-regularizers. This implementation relies on the matrix decomposition and L-curve selection of the hyper-parameter, which are sensitive to numerical error when the dimension is high. A future direction is to develop iterative methods that minimize the penalized loss functional via iterations with an early stopping strategy [1, 9], which is suitable for solving high-dimensional problems.

References

[1] Mario Arioli. Generalized golub–kahan bidiagonalization and stopping criteria. *SIAM Journal on Matrix Analysis and Applications*, 34(2):571–592, 2013.

[2] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

[3] Chuan Bi, M. Yvonne Ou, Mustapha Bouhrara, and Richard G. Spencer. Span of regularization for solution of inverse problems with application to magnetic resonance relaxometry of the brain. *Scientific Reports*, 12(1):20194, 2022.

[4] Neil K Chada, Quanjun Lang, Fei Lu, and Xiong Wang. A data-adaptive prior for Bayesian learning of kernels in operators. *arXiv preprint arXiv:2212.14163*, 2022.

[5] Zhiming Chen, Wenlong Zhang, and Jun Zou. Stochastic convergence of regularized solutions and their finite element approximations to inverse source problems. *SIAM Journal on Numerical Analysis*, 60(2):751–780, 2022.

[6] Felipe Cucker and Ding Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, Cambridge, 2007.

[7] Silvia Gazzola, Per Christian Hansen, and James G Nagy. Ir tools: a matlab package of iterative regularization methods and large-scale test problems. *Numerical Algorithms*, 81(3):773–811, 2019.

[8] Jacques Salomon Hadamard. *Lectures on Cauchy's problem in linear partial differential equations*, volume 18. Yale University Press, 1923.

[9] Per Christian Hansen. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. SIAM, 1998.

[10] Per Christian Hansen. The L-curve and its use in the numerical treatment of inverse problems. In *in Computational Inverse Problems in Electrocardiology, ed. P. Johnston, Advances in Computational Bioengineering*, pages 119–142. WIT Press, 2000.

[11] Quanjun Lang and Fei Lu. Identifiability of interaction kernels in mean-field equations of interacting particles. *arXiv preprint arXiv:2106.05565*, 2021.

[12] Gongsheng Li and Zuhair Nashed. A modified Tikhonov regularization for linear operator equations. *Numerical functional analysis and optimization*, 26(4-5):543–563, 2005.

[13] Fei Lu, Qingci An, and Yue Yu. Nonparametric learning of kernels in nonlocal operators. *arXiv preprint arXiv2205.11006*, 2022.

[14] Fei Lu, Quanjun Lang, and Qingci An. Data adaptive RKHS Tikhonov regularization for learning kernels in operators. *Proceedings of Mathematical and Scientific Machine Learning, PMLR 190:158-172*, 2022.

[15] Yuji Nakatsukasa and Nicholas J Higham. Stable and efficient spectral divide and conquer algorithms for the symmetric eigenvalue decomposition and the svd. *SIAM Journal on Scientific Computing*, 35(3):A1325–A1349, 2013.

[16] M Zuhair Nashed and Grace Wahba. Generalized inverses in reproducing kernel spaces: an approach to regularization of linear operator equations. *SIAM Journal on Mathematical Analysis*, 5(6):974–987, 1974.

[17] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[18] Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[19] Andrei Nikolajevits Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math.*, 4:1035–1038, 1963.

[20] Grace Wahba. Convergence rates of certain approximate solutions to fredholm integral equations of the first kind. *Journal of Approximation Theory*, 7(2):167–185, 1973.

[21] Grace Wahba. Practical approximate solutions to linear operator equations when the data are noisy. *SIAM journal on numerical analysis*, 14(4):651–667, 1977.

[22] Ye Zhang and Chuchu Chen. Stochastic asymptotical regularization for linear inverse problems. *Inverse Problems*, 39(1):015007, 2022.