A Universal Sampling Method for Reconstructing Signals with Simple Fourier Transforms

Haim Avron
Tel Aviv University
haimav@post.tau.ac.il

Michael Kapralov
EPFL
michael.kapralov@epfl.ch

Cameron Musco
Microsoft Research
camusco@microsoft.com

Christopher Musco
Princeton University
cmusco@cs.princeton.edu

Ameya Velingker
Google Research
ameyav@google.com

Amir Zandieh
EPFL
amir.zandieh@epfl.ch

December 20, 2018

Abstract

Reconstructing continuous signals based on a small number of discrete samples is a fundamental problem across science and engineering. In practice, we are often interested in signals with "simple" Fourier structure – e.g., those involving frequencies within a bounded range, a small number of frequencies, or a few blocks of frequencies.[1] More broadly, any prior knowledge about a signal's Fourier power spectrum can constrain its complexity. Intuitively, signals with more highly constrained Fourier structure require fewer samples to reconstruct.

We formalize this intuition by showing that, roughly, a continuous signal from a given class can be approximately reconstructed using a number of samples proportional to the *statistical dimension* of the allowed power spectrum of that class. We prove that, in nearly all settings, this natural measure tightly characterizes the sample complexity of signal reconstruction.

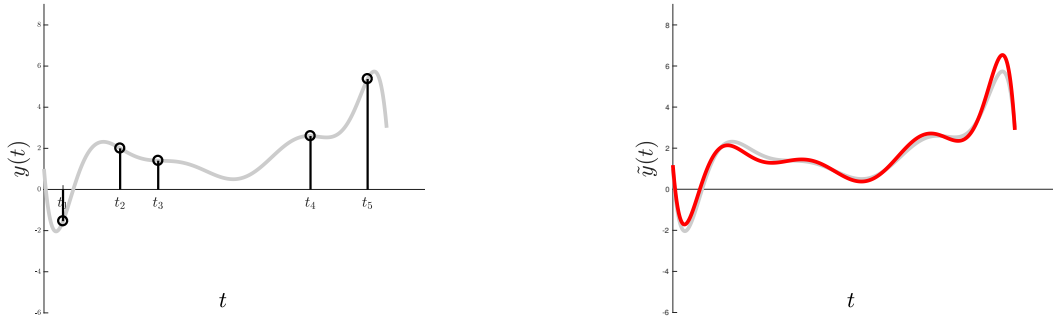
Surprisingly, we also show that, up to logarithmic factors, a universal non-uniform sampling strategy can achieve this optimal complexity for *any class of signals*. We present a simple, efficient, and general algorithm for recovering a signal from the samples taken. For bandlimited and sparse signals, our method matches the state-of-the-art. At the same time, it gives the first computationally and sample efficient solution to a broad range of problems, including multiband signal reconstruction and kriging and Gaussian process regression tasks in one dimension.

Our work is based on a novel connection between randomized linear algebra and the problem of reconstructing signals with constrained Fourier structure. We extend tools based on statistical leverage score sampling and column-based matrix reconstruction to the approximation of continuous linear operators that arise in the signal reconstruction problem. We believe that these extensions are of independent interest and serve as a foundation for tackling a broad range of continuous time problems using randomized methods.

[1] I.e. bandlimited, sparse, and multiband signals, respectively.

1 Introduction

Consider the following fundamental function fitting problem, pictured in Figure 1. We can access a continuous signal $y(t)$ at any time $t \in [0, T]$. We wish to select a finite set of sample times $t_1, \ldots, t_q$ such that, by observing the signal values $y(t_1), \ldots, y(t_q)$ at those samples, we are able to find a good approximation $\tilde{y}$ to y over the entire range $[0, T]$. We also study the problem in a noisy setting, where for each sample t_i, we only observe $y(t_i) + n(t_i)$ for some fixed noise function n.



(a) Observed signal y sampled at times $t_1, \ldots, t_q$. (b) Reconstructed signal $\tilde{y}$ based on samples.

Figure 1: Our basic function fitting problem requires reconstructing a continuous signal based on a small number of (possibly noisy) discrete samples.

We seek to understand:

1. How many samples q are required to approximately reconstruct y and how should we select these samples?
2. After sampling, how can we find and represent $\tilde{y}$ in a computationally efficient way?

Answering these questions requires assumptions about the underlying signal y. In particular, for the information at our samples $t_1, \ldots, t_q$ to be useful in reconstructing y on the entirety of $[0, T]$, the signal must be smooth, structured, or otherwise "simple" in some way.

Across science and engineering, by far one of the most common ways in which structure arises is through various assumptions about $\hat{y}$, the *Fourier transform* of y:

$$\hat{y}(\xi) = \int_{-\infty}^{\infty} y(t) e^{-2\pi i t \xi} \, dt.$$

Our goal is to understand signal reconstruction under natural constraints on the complexity of $\hat{y}$.

1.1 Classical sampling theory and bandlimited signals

Classically, the most standard example of such a constraint is requiring y to be *bandlimited*, meaning that $\hat{y}$ is only non-zero for frequencies ξ with $|\xi| \leq F$ for some bandlimit F. In this case, we recall the famous sampling theory of Nyquist, Shannon, and others [Whi15, Kot33, Nyq28, Sha49]. This theory shows that y can be reconstructed exactly using sinc interpolation (i.e, Whittaker-Shannon interpolation) if $1/2F$ uniformly spaced samples of y are taken per unit of time (the 'Nyquist rate').

Unfortunately, this theory is asymptotic: it requires infinite samples over the entire real line to interpolate y, even at a single point. When a finite number of samples are taken over an interval $[0, T]$, sinc interpolation is not a good reconstruction method, either in theory or in practice [Xia01].[2]

[2] Approximation bounds can be obtained by truncating the Whittaker-Shannon method; however, they are weak, depending *polynomially*, rather than *logarithmically*, on the desired error ϵ (see Appendix A, Example 25).

This well-known issue was resolved through a seminal line of work by Slepian, Landau, and Pollak [SP61, LP61, LP62], who presented a set of explicit basis functions for interpolating bandlimited functions when a finite number of samples are taken from a finite interval. Their so-called "prolate spheroidal wave functions" can be combined with numerical quadrature methods [XRY01, STR06, KZW+17] to obtain sample efficient (and computationally efficient) methods for bandlimited reconstruction. Overall, this work shows that roughly $O(FT + \log(1/\epsilon))$ samples from $[0, T]$ are required to interpolate a signal with bandlimit F to accuracy ϵ on that same interval.[3]

1.2 More general Fourier structure

While the aforementioned line of work is beautiful and powerful, in today's world, we are interested in far more general constraints than bandlimits. For example, there is wide-spread interest in *Fourier-sparse* signals [Don06], where $\hat{y}$ is only non-zero for a small number of frequencies, and *multiband* signals, where the Fourier transform is confined to a small number of intervals. Methods for recovering signals in these classes have countless applications in communication, imaging, statistics, and a wide variety of other disciplines [Eld15].

More generally, in statistical signal processing, a *prior distribution*, specified by some probability measure μ, is often assumed on the frequency content of y [EU06, RvdVU05]. For signals with bandlimit F, μ would be the uniform probability measure on $[-F, F]$. Alternatively, instead of assuming a hard bandlimit, a zero-centered Gaussian prior on $\hat{y}$ can encode knowledge that higher frequencies are less likely to contribute significantly to y, although they may still be present. Such a prior naturally suits a Bayesian approach to signal reconstruction [HS93] and, in fact, is essentially equivalent to assuming y is a stationary stochastic process with a certain covariance function (see Section 3 and Appendix G). Under various names, including "Gaussian process regression" and "kriging," likelihood estimation under a covariance prior is the dominant statistical approach to fitting continuous signals in many scientific disciplines, from geostatistics to economics to medical imaging [Rip05, RW06].

1.3 Our contributions

Despite their clear importance, accurate methods for fitting continuous signals under most common Fourier transform priors are not well understood, even 50 years after the groundbreaking work of Slepian, Landau, and Pollak on the bandlimited problem. The only exception is Fourier sparse signals: the *noiseless* interpolation problem can be solved using classical methods [dP95, Pis73, BM86], and recent work has resolved the much more difficult noisy case [CKPS16, CP18].

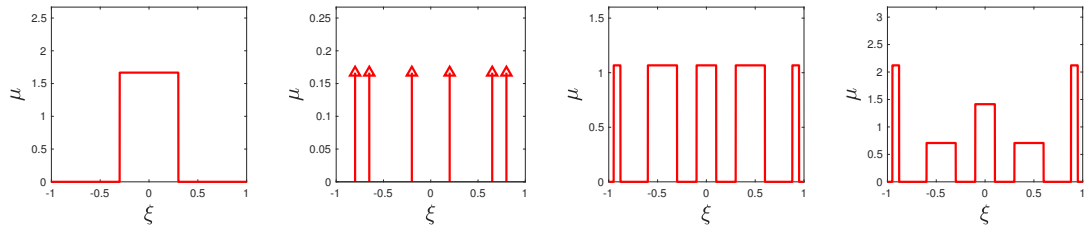
In this paper, we address the problem far more generally. Our contributions are as follows:

1. We tightly characterize the information theoretic sample complexity of reconstructing y under any Fourier transform prior, specified by probability measure μ. In essentially all settings, we can prove that this complexity scales nearly linearly with a natural *statistical dimension* parameter associated with μ. See Theorem 1.

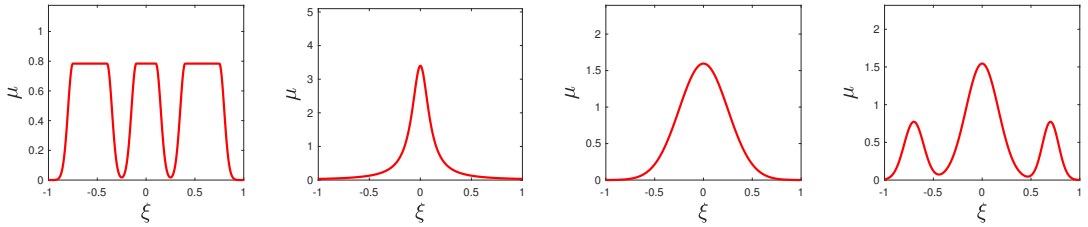
2. We present a method for sampling from y that achieves the aforementioned statistical dimension bound to within a polylogarithmic factor. Our approach is randomized and *universal*: we prove that it is possible to draw $t_1, \ldots, t_q$ from a fixed non-uniform distribution over $[0, T]$ that is *independent of* μ, i.e., "spectrum-blind." In other words, the same sampling scheme works for bandlimited, sparse, or more general priors. See Theorem 2.

[3] We formalize our notion of accuracy in Section 2.

3. We show that y can be recovered from $t_1, \ldots, t_q$ using a simple, efficient, and completely general interpolation method. In particular, we just need to solve a kernel ridge regression problem using $y(t_1), \ldots, y(t_q)$, with an appropriately chosen kernel function for μ. This method runs in $O(q^3)$ time and is already widely used for signal reconstruction in practice, albeit with suboptimal strategies for choosing $t_1, \ldots, t_q$. See Theorem 3.



(a) Bandlimited. (b) Sparse. (c) Multiband. (d) Non-uniform multiband.



(e) Smooth multiband. (f) Cauchy-Lorentz. (g) Gaussian. (h) Gaussian mixture.

Figure 2: Examples of Fourier transform "priors" induced by various measures μ (we plot the corresponding density). Our algorithm can reconstruct signals under any of these priors.

Overall, this approach gives the first finite sample, provable approximation bounds for all common Fourier-constrained signal reconstruction problems beyond bandlimited and sparse functions.

Our results are obtained by drawing on a rich set of tools from randomized numerical linear algebra, including sampling methods for approximate regression and deterministic column-based low-rank approximation methods [BSS14, CNW16]. Many of these methods view matrices as sums of rank-1 outer products and approximate them by sampling or deterministically selecting a subset of these outer products. We adapt these tools to the approximation of continuous operators, which can be written as the (weak) integral of rank-1 operators. For example, our universal time domain sampling distribution is obtained using the notion of *statistical leverage* [SS11, AM15, DM16], extended to a continuous Fourier transform operator that arises in the signal reconstruction problem. We hope that, by extending many of the fundamental contributions of randomized numerical linear algebra to build a toolkit for 'randomized operator theory', our work will offer a starting point for progress on many signal processing problems using randomized methods.

2 Formal statement of results

As suggested, we formally capture Fourier structure through any probability measure μ over the reals.[4] We often refer to μ as a "prior", although we will see that it can be understood beyond the

[4]Formally, we consider the measure space $(\mathbb{R}, \mathcal{B}, \mu)$ where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$.

context of Bayesian inference. The simplicity of a set of constraints will be quantified by a natural *statistical dimension* parameter for μ, defined in Section 2.1.

For signals with bandlimit F, μ is the uniform probability measure on $[-F, F]$. For multiband signals, it is uniform on the union of k intervals, while for Fourier-sparse functions, μ is uniform on a union of k Dirac measures. More general priors are visualized in Figure 2. Those based on Gaussian or Cauchy-Lorentz distributions are especially common in scientific applications, and we will discuss examples shortly. For now, we begin with our main problem formulation.

Problem 1. *Given a known probability measure μ on $\mathbb{R}$, for any $t \in [0, T]$, define the inverse Fourier transform of a function $g(\xi)$ with respect to μ as*

$$\left[\mathcal{F}_\mu^* g\right](t) \stackrel{\text{def}}{=} \int_{\mathbb{R}} g(\xi) e^{2\pi i \xi t} \, d\mu(\xi). \tag{1}$$

Suppose our input y can be written as $y = \mathcal{F}_\mu^* x$ for some frequency domain function $x(\xi)$ and, for any chosen t, we can observe $y(t) + n(t)$ for some fixed noise function $n(t)$. Then, for error parameter ϵ, our goal is to recover an approximation $\tilde{y}$ satisfying

$$\|y - \tilde{y}\|_T^2 \leq \epsilon \|x\|_\mu^2 + C\|n\|_T^2, \tag{2}$$

where $\|x\|_\mu^2 \stackrel{\text{def}}{=} \int_{\mathbb{R}} |x(\xi)|^2 \, d\mu(\xi)$ is the energy of the function x with respect to μ, while $\|z\|_T^2 \stackrel{\text{def}}{=} \frac{1}{T}\int_0^T |z(t)|^2 dt$, so that $\|y - \tilde{y}\|_T^2$ is our mean squared error and $\|n\|_T^2$ is the mean squared noise level. $C \geq 1$ is a fixed positive constant.

Unlike the $\|x\|_\mu^2$ term in (2), which we can control by adjusting ϵ, we can never hope to recover y to accuracy better than $\|n\|_T^2$. Accordingly, we consider $\|n\|_T^2$ to be small and are happy with any solution of Problem 1 that is within a constant factor of optimal – i.e., where $C = O(1)$.

Problem 1 captures signal reconstruction under all standard Fourier transform constraints, including bandlimited, multiband, and sparse signals.[5] The error in (2) naturally scales with the average energy of the signal over the allowed frequencies. For more general priors, $\|x\|_\mu^2$ will be larger when y contains a significant component of frequencies with low density in μ.[6] For a given number of samples, we would thus incur larger error in (2) in comparison to a signal that uses more "likely" frequencies.

As an alternative to Problem 1, we can formulate signal fitting from a Bayesian perspective. We assume that n is independent random noise, and y is a stationary stochastic process with expected power spectral density μ. This assumption on y's power spectral density is equivalent to assuming that y has covariance function (a.k.a. autocorrelation) $\hat{\mu}(t)$, which is the type of prior used in kriging and Gaussian process regression. While we focus on the formulation of Problem 1 in this work, we give an informal discussion of the Bayesian setup in Appendix G.

Examples and applications

As discussed in Section 1.2, "hard constraint" versions of Problem 1, such as bandlimited, sparse, and multiband signal reconstruction, have many applications in communications, imaging, audio,

[5] For sparse or multiband signals, Problem 1 assumes frequency or band locations are known *a priori*. There has been significant work on algorithms that can recover y when these locations are not known [ME09, Moi15, PS15, CKPS16]. Understanding this more complicated problem in the multiband case is an important future direction.

[6] Informally, decreasing $d\mu(\xi)$ by a factor of $c > 1$ requires increasing $x(\xi)$ by a factor of c to give the same time domain signal. This increases $x(\xi)^2$ by a factor of c^2 and so increases its contribution to $\|x\|_\mu^2$ by a factor of $c^2/c = c$.

and other areas of engineering. Generalizations of the multiband problem to non-uniform measures (see Figure 2d) are also useful in various communication problems [ME10].

On the other hand, "soft constraint" versions of the problem are widely applied in scientific applications. In medical imaging, images are often denoised by setting μ to a heavy-tailed Cauchy-Lorentz measure on frequencies [Fud89, LH95, BWvOGD01]. This corresponds to assuming an exponential covariance function for spatial correlation. Exponential covariance and its generalization, Matérn covariance, are also common in the earth and geosciences [Rip89, Rip05], as well as in general image processing [PPV02, RVU06].

A Gaussian prior μ, which corresponds to Gaussian covariance, is also used to model both spatial and temporal correlation in medical imaging [FJT94, WMN+96] and is very common in machine learning. Other choices for μ are practically unlimited. For example, the popular ArcGIS kriging library also supports the following covariance functions: circular, spherical, tetraspherical, pentaspherical, rational quadratic, hole effect, k-bessel, and j-bessel, and stable [ESR18].

2.1 Sample complexity

With Problem 1 defined, our first goal is to characterize the number of samples required to reconstruct y, as a function of the *accuracy parameter* ϵ, the *range* T, and the *measure* μ. We do so using what we refer to as the *Fourier statistical dimension* of μ, which corresponds to the standard notion of statistical or 'effective dimension' for regularized function fitting problems [HTF02, Zha05].

Definition 2 (Fourier statistical dimension)**.** *For a probability measure μ on $\mathbb{R}$ and time length T, define the kernel operator $\mathcal{K}_\mu : L_2(T) \to L_2(T)$[7] as:*

$$[\mathcal{K}_\mu z](t) \stackrel{\text{def}}{=} \int_{\xi \in \mathbb{R}} e^{2\pi i \xi t} \left[\frac{1}{T} \int_{s \in [0,T]} z(s) e^{-2\pi i \xi s} \, ds \right] d\mu(\xi). \tag{3}$$

Note that $\mathcal{K}_\mu$ is self-adjoint, positive semidefinite and trace-class.[8] *The Fourier statistical dimension for μ, T, and error ϵ is denoted by $s_{\mu,\epsilon}$ and defined as:*

$$s_{\mu,\epsilon} \stackrel{\text{def}}{=} \operatorname{tr}(\mathcal{K}_\mu(\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1}), \tag{4}$$

where $\mathcal{I}_T$ is the identity operator on $L_2(T)$. Letting $\lambda_i(\mathcal{K}_\mu)$ denote the i^{th} largest eigenvalue of $\mathcal{K}_\mu$, we may also write

$$s_{\mu,\epsilon} = \sum_{i=1}^{\infty} \frac{\lambda_i(\mathcal{K}_\mu)}{\lambda_i(\mathcal{K}_\mu) + \epsilon}. \tag{5}$$

Note that $\mathcal{K}_\mu$ and $s_{\mu,\epsilon}$ as defined above, and $\mathcal{F}_\mu$ as defined in Problem 1 all depend on T and thus could naturally be denoted $\mathcal{F}_{\mu,T}$, $\mathcal{K}_{\mu,T}$, and $s_{\mu,\epsilon,T}$. However, since T is fixed throughout our results, for conciseness we do not use T in our notation for these and related notions.

It is not hard to see that $s_{\mu,\epsilon}$ increases as ϵ decreases, meaning that we will require more samples to obtain a more accurate solution to Problem 1. The operator $\mathcal{K}_\mu$ corresponds to taking the Fourier transform of a time domain input $z(t)$, scaling that transform by μ, and then taking the inverse Fourier transform. Readers familiar with the literature on bandlimited signal reconstruction will recognize $\mathcal{K}_\mu$ as the natural generalization of the frequency limiting operator studied in the work

[7] $L_2(T)$ denotes the complex-valued square integrable functions with respect to the uniform measure on $[0, T]$.
[8] See Section 3 for a formal explanation of these facts.

of Landau, Slepian, and Pollak on prolate spheroidal wave functions [SP61, LP61, LP62]. In that work, it is established that a quantity nearly identical to $s_{\mu,\epsilon}$ bounds the sample complexity of solving Problem 1 for bandlimited functions.

Our first technical result is that this is actually true *for any prior* μ.

Theorem 1 (Main result, sample complexity)**.** *For any probability measure* μ*, Problem 1 can be solved using* $q = O\left(s_{\mu,\epsilon} \cdot \log s_{\mu,\epsilon}\right)$ *noisy signal samples* $y(t_1) + n(t_1), \ldots, y(t_q) + n(t_q)$*.*

What does Theorem 1 imply for common classes of functions with constrained Fourier transforms? Table 1 includes a list of upper bounds on $s_{\mu,\epsilon}$ for many standard priors.

Fourier prior, μ	Statistical dimension, $s_{\mu,\epsilon}$	Proof
k-sparse	k	Since $\mathcal{K}_\mu$ has rank k.
bandlimited to $[-F, F]$	$O\left(FT + \log(1/\epsilon)\right)$	Theorem 48.
multiband, widths $F_1, \ldots, F_s$	$O\left(\sum_i F_i T + s\log(1/\epsilon)\right)$	Theorem 53.[9]
Gaussian, variance F	$O\left(FT\sqrt{\log(1/\epsilon)} + \log(1/\epsilon)\right)$	Theorem 54.
Cauchy-Lorentz, scale F	$O\left(FT\sqrt{1/\epsilon} + \sqrt{1/\epsilon}\right)$	Theorem 55.

Table 1: Statistical dimension upper bounds for common Fourier interpolation problems. Our result (Theorem 1) requires $O(s_{\mu,\epsilon} \cdot \log s_{\mu,\epsilon})$ samples.

A complexity of $O(s_{\mu,\epsilon} \cdot \log s_{\mu,\epsilon})$ equates to $\tilde{O}(k)$ samples for k-sparse functions and $\tilde{O}(FT + \log 1/\epsilon)$ for bandlimited functions. Up to log factors, these bounds are tight for these well studied problems. In Section 6, we show that Theorem 1 is actually tight for all common Fourier transform priors: $\Omega(s_{\mu,\epsilon})$ time points are required for solving Problem 1 as long as $s_{\mu,\epsilon}$ grows slower than $1/\epsilon^p$ for some $p < 1$. This property holds for all μ in Table 1. We conjecture that our lower bound can be extended to hold even without this weak assumption.

To compliment the sample complexity bound of Theorem 1, we introduce a *universal method* for selecting samples $t_1, \ldots, t_q$ that nearly matches this complexity. Our method selects samples at random, in a way that *does not depend* on the specific prior μ.

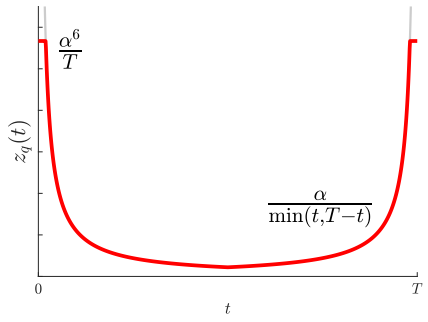
Theorem 2 (Main result, sampling distribution)**.** *For any sample size* q*, there is a fixed probability density* p_q *over* $[0, T]$ *such that, if* q *time points* $t_1, \ldots, t_q$ *are selected independently at random according to* p_q*, and* $q \geq c \cdot s_{\mu,\epsilon} \cdot \log^2 s_{\mu,\epsilon}$ *for some fixed constant* c*, then it is possible to solve Problem 1 with probability 99/100 using the noisy signal samples* $y(t_1) + n(t_1), \ldots, y(t_q) + n(t_q)$*.*[10]

Theorem 2 is our main technical contribution. By achieving near optimal sample complexity with a universal distribution, it shows that wide range of Fourier constrained interpolation problems considered in the literature are more closely related than previously understood.

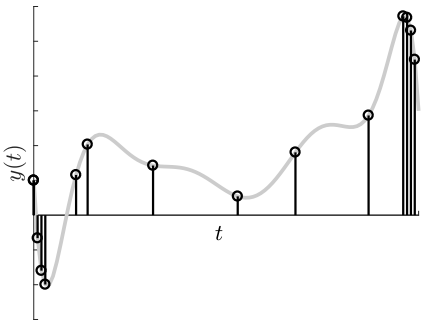
Moreover, p_q (which is formally defined in Theorem 17) is very simple to describe and sample from. As may be intuitive from results on polynomial interpolation, bandlimited approximation, and other function fitting problems, it is more concentrated towards the endpoints of $[0, T]$, so our sampling scheme selects more time points in these regions. The density is shown in Figure 3.

[9] Just as Theorem 48 intuitively matches the Nyquist sampling rate, Theorem 53 intuitively matches the Landau rate for asymptotic recovery of multiband functions [Lan67a].

[10] In Section 5.4, we formally quantify the tradeoff between success probability and sample complexity.



(a) Density for selecting time points.



(b) Example set of nodes sampled according to p_q.

Figure 3: A visualization of the universal sampling distribution, p_q, which can be used for reconstructing a signal under any Fourier transform prior μ. To obtain p_q for a given number of samples q, choose α so that $q = \Theta(\alpha \log^2 \alpha)$. Set $z_q(t)$ equal to $\alpha / \min(t, T-t)$, except near 0 and T, where the function is capped at $z_q(t) = \alpha^6$. Construct p_q by normalizing z_q to integrate to 1.

2.2 Algorithmic complexity

While Theorem 2 immediately yields an approach for selecting samples $t_1, \ldots, t_q$, it is only useful if we can *efficiently* solve Problem 1 given the noisy measurements $y(t_1) + n(t_1), \ldots, y(t_q) + n(t_q)$. We show that this is possible for a broad class of constraint measures. Specifically, we need only assume that we can efficiently compute the positive-definite kernel function[11]:

$$k_\mu(t_1, t_2) = \int_{\xi \in \mathbb{R}} e^{-2\pi i (t_1 - t_2)\xi} d\mu(\xi). \tag{6}$$

The above integral can be approximated via numerical quadrature, but for many of the aforementioned applications, it has a closed-form. For example, when μ is supported on just k frequencies, it is a sum of these frequencies. When μ is uniform on $[-F, F]$, $k_\mu(t_1, t_2) = \text{sinc}(2\pi F(t_1 - t_2))$. For multiband signals with s bands, $k_\mu(t_1, t_2)$ is a sum of s modulated sinc functions. In fact, $k_\mu(t_1, t_2)$ has a closed-form for all μ illustrated in Figure 2. Further details are discussed in Appendix F. Assuming a subroutine for computing $k_\mu(t_1, t_2)$, our main algorithmic result is as follows:

Theorem 3. *(Main result, algorithmic complexity) There is an algorithm that solves Problem 1 with probability* 99/100 *which uses* $O\left(s_{\mu,\epsilon} \cdot \log^2(s_{\mu,\epsilon})\right)$ *time domain samples (sampled according to the distribution given by Theorem 2) and runs in* $\tilde{O}(s_{\mu,\epsilon}^\omega + s_{\mu,\epsilon}^2 \cdot Z)$ *time, assuming the ability to compute* $k_\mu(t_1, t_2)$ *for any* $t_1, t_2 \in [0, T]$ *in* Z *time.*[12] *The algorithm returns a representation of* $\tilde{y}(t)$ *that can be evaluated in* $\tilde{O}(s_{\mu,\epsilon} \cdot Z)$ *time for any* t.

For bandlimited, Gaussian, or Cauchy-Lorentz priors μ, $Z = O(1)$. For s sparse signals or multiband signals with s blocks, $Z = O(s)$.

We note that, while Theorem 3 holds when $\tilde{O}\left(s_{\mu,\epsilon}\right)$ samples are taken, $s_{\mu,\epsilon}$ may be not be known and thus it may be unclear how to set the sample size. In our full statement of the Theorem in Section 5.4 we make it clear that any upper bound on $s_{\mu,\epsilon}$ suffices to set the sample size. The sample complexity will depend on how tight this upper bound is. In Appendix E we give upper bounds on $s_{\mu,\epsilon}$ for a number of common μ, which can be plugged into Theorem 3.

[11] When y is real valued, it makes sense to consider symmetric μ. In this case, k_μ is also real valued. However, in general it may be complex valued.

[12] For conciseness, we use $\tilde{O}(z)$ to denote $\tilde{O}(z \log^c z)$, where c is some fixed constant (usually ≤ 2). In formal theorem statements we give c explicitly. $\omega < 2.373$ is the current exponent of fast matrix multiplication [Wil12].

2.3 Our approach

Theorems 1, 2, and 3 are achieved through a simple and practical algorithmic framework. In Section 4, we show that Problem 1 can be modeled as a least squares regression problem with ℓ_2 regularization. As long as we can compute $k_\mu(t_1, t_2)$, we can solve this problem using *kernel ridge regression*, a popular function fitting technique in nonparametric statistics [STC04].

Naively, the kernel regression problem is infinite dimensional: it needs to be solved over the *continuous* time domain $[0, T]$ to solve our signal reconstruction problem. This is where sampling comes in. We need to discretize the problem and establish that our solution over a fixed set of time samples nearly matches the solution over the continuous interval. To bound the error of discretization, we turn to a tool from randomized numerical linear algebra: *statistical leverage score sampling* [SS11, DM16]. We show how to *randomly* discretize Problem 1 by sampling time points with probability proportional to an appropriately defined non-uniform leverage score distribution on $[0, T]$. The required number of samples is $O(s_{\mu,\epsilon} \log s_{\mu,\epsilon})$, which proves Theorem 1.

Unfortunately, the leverage score distribution does not have a closed-form, varies depending on ϵ, T, and μ, and likely cannot be sampled from exactly. To prove Theorem 2, we show that for any μ, for large enough q, the closed form distribution p_q *upper bounds* the leverage score distribution. This upper bound closely approximates the true leverage score distribution and, therefore, can be used in its place during sampling, losing only a $\log s_{\mu,\epsilon}$ factor in the sample complexity.

The leverage score distribution roughly measures, for each time point t, how large $|y(t)|^2$ can be compared to $\|y\|_T^2$ when y's Fourier transform is constrained by μ (i.e., when $\|x\|_\mu^2$ as defined in Problem 1 is bounded). To upper bound this measure we turn to another powerful result from the randomized numerical linear algebra literature: every matrix contains a small subset of columns that span a near-optimal low-rank approximation to that matrix [Sar06, BMD09, DR10]. In other words, every matrix admits a near-optimal low-rank approximation with *sparse column support*. By extending this result to continuous linear operators, we prove that the smoothness of a signal whose Fourier transform has $\|x\|_\mu^2$ bounded can be bounded by the smoothness of an $O(s_{\mu,\epsilon})$ sparse Fourier function. This lets us apply recent results of [CKPS16, CP18] that bound $|y(t)|^2$ in terms of $\|y\|_T^2$ for any sparse Fourier function y . Intuitively, our result shows that the simplicity of sparse Fourier functions governs the simplicity of *any class* of Fourier constrained functions.

The above argument yields Theorem 2. Since we can sample from p_q in $O(1)$ time, we can efficiently sample the time domain to $O(s_{\mu,\epsilon} \cdot \log^2 s_{\mu,\epsilon})$ points and then solve Problem 1 by applying kernel ridge regression to these points, which takes $\tilde{O}(s_{\mu,\epsilon}^\omega + s_{\mu,\epsilon}^2 \cdot Z)$ time, assuming the ability to compute $k_\mu(\cdot, \cdot)$ in Z time. This yields the algorithmic result of Theorem 3.

2.4 Roadmap

The rest of this paper is devoted to proving Theorems 1, 2, and 3, and is structured as follows:

Section 3 We lay out basic notation that is used throughout the paper.

Section 4 We reduce Problem 1 to a kernel ridge regression problem and explain how to randomly discretize and solve this problem via leverage score sampling, proving Theorem 1.

Section 5 We give an upper bound on the leverage score distribution for general priors, proving Theorems 2 and 3.

Section 6 We prove that, under a mild assumption, the statistical dimension tightly characterizes the sample complexity of solving Problem 1, and thus that our results are nearly optimal.

Section 7 We conclude with a discussion of open questions.

We defer an in depth overview of related work to Appendix A. In Appendix B we give operator theory preliminaries. In Appendix C we prove our extensions of a number of randomized linear algebra primitives to continuous operators. In Appendix D, we bound the statistical dimension for the important case of bandlimited functions. We use this result in Appendix E to prove statistical dimension bounds for multiband, Gaussian, and Cauchy-Lorentz priors (shown in Table 1). In Appendix F, we show how to compute the kernel function k_μ for these common priors. In Appendix G, we discuss a Bayesian approach to signal reconstruction under a Fourier transform prior.

3 Notation

Let μ be a probability measure on $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$. Let $L_2(\mu)$ denote the space of complex-valued square integrable functions with respect to μ. For $a, b \in L_2(\mu)$, let $\langle a, b\rangle_\mu$ denote $\int_{\xi\in\mathbb{R}} a(\xi)^* b(\xi)\, d\mu(\xi)$ where for any $x \in \mathbb{C}$, x^* is its complex conjugate. Let $\|a\|_\mu^2$ denote $\langle a, a\rangle_\mu$. Let $\mathcal{I}_\mu$ denote the identity operator on $L_2(\mu)$. Note that for any μ, $L_2(\mu)$ is a separable Hilbert space and thus has a countably infinite orthonormal basis [HN01].

We overload notation and use $L_2(T)$ to denote the space of complex-valued square integrable functions with respect to the uniform probability measure on $[0, T]$. It will be clear from context that T is not a measure. For $a, b \in L_2(T)$, let $\langle a, b\rangle_T$ denote $\frac{1}{T}\int_0^T a(t)^* b(t)\, dt$ and let $\|a\|_T^2$ denote $\langle a, a\rangle_T$. Let $\mathcal{I}_T$ denote the identity operator on $L_2(T)$.

Define the Fourier transform operator $\mathcal{F}_\mu : L_2(T) \to L_2(\mu)$ as:

$$\left[\mathcal{F}_\mu f\right](\xi) = \frac{1}{T}\int_0^T f(t) e^{-2\pi i t\xi} dt. \tag{7}$$

The adjoint of $\mathcal{F}_\mu$ is the unique operator $\mathcal{F}_\mu^* : L_2(\mu) \to L_2(T)$ such that for all $f \in L_2(T), g \in L_2(\mu)$ we have $\langle g, \mathcal{F}_\mu f\rangle_\mu = \langle \mathcal{F}_\mu^* g, f\rangle_T$. It is not hard to see that $\mathcal{F}_\mu^*$ is the inverse Fourier transform operator with respect to μ as defined in Section 2, equation (1):

$$\left[\mathcal{F}_\mu^* g\right](t) \stackrel{\text{def}}{=} \int_{\mathbb{R}} g(\xi) e^{2\pi i \xi t}\, d\mu(\xi). \tag{8}$$

Note that the kernel operator $\mathcal{K}_\mu : L_2(T) \to L_2(T)$ originally defined in (3) is equal to

$$\mathcal{K}_\mu = \mathcal{F}_\mu^* \mathcal{F}_\mu.$$

$\mathcal{K}_\mu$ is self-adjoint, positive semidefinite and trace-class and an integral operator with kernel k_μ:

$$[\mathcal{K}_\mu z](t) = \frac{1}{T}\int_0^T k_\mu(s, t) z(s) ds,$$

where k_μ is as defined in (6). The trace of $\mathcal{K}_\mu$ is equal to 1.[13] We will also make use of the Gram operator: $\mathcal{G}_\mu \stackrel{\text{def}}{=} \mathcal{F}_\mu \mathcal{F}_\mu^*$. $\mathcal{G}_\mu$ is also self-adjoint, positive semidefinite, and trace-class.

Remark: It may be useful for the reader to informally regard $\mathcal{F}_\mu$ as an infinite matrix with rows indexed by $\xi \in \mathbb{R}$ and columns indexed by $t \in [0, T]$. Following the definition of $\mathcal{F}_\mu$ above, and assuming that μ has a density p, this infinite matrix has entries given by:

$$\mathcal{F}_\mu(\xi, t) = \sqrt{\frac{p(\xi)}{T}} \cdot e^{-2\pi i t\xi}. \tag{9}$$

The results we apply on leverage score sampling can all be seen as extending results for finite matrices from the randomized numerical linear algebra literature to this infinite matrix.

[13] Since the kernel is a Fourier transform of a probability measure, it is Hermitian positive definite (Bochner's Theorem). Then we can conclude that $\mathcal{K}_\mu$ is trace-class from Mercer's theorem, and calculate $\operatorname{tr}(\mathcal{K}_\mu) = \frac{1}{T}\int_0^T k_\mu(t, t) dt = 1$.

4 Function fitting with least squares regression

Least squares regression provides a natural approach to solving the interpolation task of Problem 1. In particular, consider the following regularized minimization problem over functions $g \in L_2(\mu)$[14]:

$$\min_{g \in L_2(\mu)} \|\mathcal{F}_\mu^* g - (y+n)\|_T^2 + \epsilon \|g\|_\mu^2. \tag{10}$$

The first term encourages us to find a function g whose inverse Fourier transform is close to our measured signal $y+n$. The second term encourages us to find a low energy solution – ultimately, we solve (10) based on only a small number of samples $y(t_1), \ldots, y(t_k)$, and smoother, lower energy solutions will better generalize to the entire interval $[0, T]$. We remark that it is well known that least squares approximations benefit from regularization even in the noiseless case [CDL13].

We first state a straightforward fact: if we minimize (10), even to a coarse approximation, then we are able to solve Problem 1.

Claim 4. *Let $y = \mathcal{F}_\mu^* x$, $n \in L_2(T)$ be an arbitrary noise function, and for any $C \geq 1$, let $\tilde{g} \in L_2(\mu)$ be a function satisfying:*

$$\|\mathcal{F}_\mu^* \tilde{g} - (y+n)\|_T^2 + \epsilon \|\tilde{g}\|_\mu^2 \leq C \cdot \min_{g \in L_2(\mu)} \left[\|\mathcal{F}_\mu^* g - (y+n)\|_T^2 + \epsilon \|g\|_\mu^2 \right].$$

Then

$$\|\mathcal{F}_\mu^* \tilde{g} - y\|_T^2 \leq 2C\epsilon \|x\|_\mu^2 + 2(C+1)\|n\|_T^2.$$

Proof. Since $y = \mathcal{F}_\mu^* x$, $\min_{g \in L_2(\mu)} \left[\|\mathcal{F}_\mu^* g - (y+n)\|_T^2 + \epsilon \|g\|_\mu^2 \right] \leq \|n\|_T^2 + \epsilon \|x\|_\mu^2$. Thus, $\|\mathcal{F}_\mu^* \tilde{g} - (y+n)\|_T^2 \leq C\epsilon \|x\|_\mu^2 + C\|n\|_T^2$. The claim then follows via triangle inequality:

$$\begin{aligned} \|\mathcal{F}_\mu^* \tilde{g} - y\|_T - \|n\|_T &\leq \|\mathcal{F}_\mu^* \tilde{g} - (y+n)\|_T \\ \|\mathcal{F}_\mu^* \tilde{g} - y\|_T &\leq \sqrt{C\epsilon \|x\|_\mu^2 + C\|n\|_T^2} + \|n\|_T \\ \|\mathcal{F}_\mu^* \tilde{g} - y\|_T^2 &\leq 2C\epsilon \|x\|_\mu^2 + 2(C+1)\|n\|_T^2. \end{aligned}$$

□

Claim 4 shows that approximately solving the regression problem in (10), with regularization parameter ϵ gives a solution to Problem 1 with parameter $2C\epsilon$ (decreasing the regularization parameter to $\frac{\epsilon}{2C}$ will let us solve with parameter ϵ). But how can we solve the regression problem efficiently? Not only does the problem involve a possibly infinite dimensional parameter vector g, but the objective function also involves the continuous time interval $[0, T]$.

4.1 Random discretization via leverage function sampling

The first step is to deal with the latter challenge, i.e., that of a continuous time domain. We show that it is possible to *randomly discretize* the time domain of (10), thereby reducing our problem to a regression problem on a finite set of times $t_1, \ldots, t_q$. In particular, we can sample time points with probability proportional to the so-called *ridge leverage function*, a specific non-uniform distribution that has been applied widely in randomized algorithms for regression and other linear algebra problems on discrete matrices [AM15, CLV16, CMM17, MM17, MW17].

[14] The fact that the minimum is attainable is a simple consequence of the extreme value theorem, since the search space can be restricted to $\|g\|_\mu^2 \leq \|(y+n)\|_T^2/\epsilon$.

While we cannot compute the leverage function explicitly for our problem, an issue highlighted in [Bac17], our main result (Theorem 2) uses a simple, but very accurate, closed form approximation in its place. We start with the definition of the ridge leverage function:

Definition 3 (Ridge leverage function)**.** *For a probability measure μ on $\mathbb{R}$, time length $T > 0$, and $\epsilon \geq 0$, we define the ϵ-ridge leverage function for $t \in [0, T]$ as*[15]*:*

$$\tau_{\mu,\epsilon}(t) = \frac{1}{T} \cdot \max_{\{\alpha \in L_2(\mu):\, \|\alpha\|_\mu > 0\}} \frac{\left|[\mathcal{F}_\mu^* \alpha](t)\right|^2}{\|\mathcal{F}_\mu^* \alpha\|_T^2 + \epsilon \|\alpha\|_\mu^2}. \tag{11}$$

Intuitively, the ridge leverage function at time t is an upper bound of how much a function can "blow up" at t when its Fourier transform is constrained by μ. The denominator term $\|\mathcal{F}_\mu^* \alpha\|_T^2$ is the average squared magnitude of the function $F_\mu^* \alpha$, while the numerator term, $|[\mathcal{F}_\mu^* \alpha](t)|^2$, is the squared magnitude at t. The regularization term $\epsilon \|\alpha\|_\mu^2$ reflects the fact that, to solve (10), we only need to bound the smoothness for functions with bounded Fourier energy under μ. As observed in [PBV18], the ridge leverage function can be viewed as a type of *Christoffel function*, studied in the literature on orthogonal polynomials and approximation theory [PBV18, Nev86, Tot00, BE12].

The larger the leverage "score" $\tau_{\mu,\epsilon}(t)$, the higher the probability we will sample time t, to ensure that our sample points well reflect any possibly significant components or 'spikes' of the function y. Ultimately, the integral of the ridge leverage function $\int_0^T \tau_{\mu,\epsilon}(t)dt$ determines how many samples we require to solve (10) to a given accuracy. Theorem 5 below states the already known fact that the ridge leverage function integrates to the statistical dimension [AKM+17], which will ultimately allow us to achieve the $\tilde{O}(s_{\mu,\epsilon})$ sample complexity bound of Theorems 1 and 2. Theorem 5 also gives two alternative characterizations of the leverage function that will prove useful. The theorem is proven in Appendix C, using techniques for finite matrices, adapted to the operator setting.

Theorem 5 (Leverage function properties)**.** *Let $\tau_{\mu,\epsilon}(t)$ be the ridge leverage function (Definition 3) and define $\varphi_t \in L_2(\mu)$ by $\varphi_t(\xi) \stackrel{\text{def}}{=} e^{-2\pi i t \xi}$. We have:*

- *The ridge leverage function integrates to the statistical dimension:*

$$\int_0^T \tau_{\mu,\epsilon}(t)dt = s_{\mu,\epsilon} \stackrel{\text{def}}{=} \operatorname{tr}(\mathcal{K}_\mu(\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1}). \tag{12}$$

- *Inner Product characterization:*

$$\tau_{\mu,\epsilon}(t) = \frac{1}{T} \cdot \langle \varphi_t, (\mathcal{G}_\mu + \epsilon \mathcal{I}_\mu)^{-1} \varphi_t \rangle_\mu. \tag{13}$$

- *Minimization Characterization:*

$$\tau_{\mu,\epsilon}(t) = \frac{1}{T} \cdot \min_{\beta \in L_2(T)} \frac{\|\mathcal{F}_\mu \beta - \varphi_t\|_\mu^2}{\epsilon} + \|\beta\|_T^2. \tag{14}$$

In Theorem 6, we give our formal statement that the ridge leverage function can be used to randomly sample time domain points to discretize the regression problem in (10) and solve it

[15] Formally $L_2(T)$ is a space of equivalence classes of functions that differ at a set of points with measure 0. For notational simplicity, here and throughout we use $\mathcal{F}_\mu^* \alpha$ to denote the specific representative of the equivalence class $\mathcal{F}_\mu^* \alpha \in L_2(T)$ given by (8). In this way, we can consider the pointwise value $[\mathcal{F}_\mu^* \alpha](t)$, which we could alternatively express as $\langle \varphi_t, \alpha \rangle_\mu$, for $\varphi_t(\xi) \stackrel{\text{def}}{=} e^{-2\pi i t \xi}$.

approximately. While complex in appearance, readers familiar with randomized linear algebra will recognize Theorem 6 as closely analogous to standard approximate regression results for leverage score sampling from finite matrices [CW13]. As discussed, since we are typically unable to sample according to the true ridge leverage function, we give a general result, showing that sampling with any upper bound function with a finite integral suffices.

Theorem 6 (Approximate regression via leverage function sampling)**.** *Assume that* $\epsilon \leq \|\mathcal{K}_\mu\|_{\text{op}}$.[16] *Consider a measurable function* $\tilde{\tau}_{\mu,\epsilon}(t)$ *with* $\tilde{\tau}_{\mu,\epsilon}(t) \geq \tau_{\mu,\epsilon}(t)$ *for all* t *and let* $\tilde{s}_{\mu,\epsilon} = \int_0^T \tilde{\tau}_{\mu,\epsilon}(t)dt$. *Let* $s = c \cdot \tilde{s}_{\mu,\epsilon} \cdot (\log \tilde{s}_{\mu,\epsilon} + 1/\delta)$ *for sufficiently large fixed constant* c *and let* $t_1, \ldots, t_s$ *be time points selected by drawing each randomly from* $[0, T]$ *with probability proportional to* $\tilde{\tau}_{\mu,\epsilon}(t)$. *For* $j \in 1, \ldots, s$, *let* $w_j = \sqrt{\frac{1}{sT} \cdot \frac{\tilde{s}_{\mu,\epsilon}}{\tilde{\tau}_{\mu,\epsilon}(t_j)}}$. *Let* $\mathbf{F} : \mathbb{C}^s \to L_2(\mu)$ *be the operator defined by:*

$$[\mathbf{F}\, g](\xi) = \sum_{j=1}^{s} w_j \cdot g(j) \cdot e^{-2\pi i \xi t_j}$$

and $\mathbf{y}, \mathbf{n} \in \mathbb{R}^s$ *be the vectors with* $\mathbf{y}(j) = w_j \cdot y(t_j)$ *and* $\mathbf{n}(j) = w_j \cdot n(t_j)$. *Let:*

$$\tilde{g} = \arg\min_{g \in L_2(\mu)} \left[\|\mathbf{F}^* g - (\mathbf{y} + \mathbf{n})\|_2^2 + \epsilon \|g\|_\mu^2 \right] \tag{15}$$

With probability $\geq 1 - \delta$*:*

$$\|\mathcal{F}_\mu^* \tilde{g} - (y + n)\|_T^2 + \epsilon \|\tilde{g}\|_\mu^2 \leq 3 \min_{g \in L_2(\mu)} \left[\|\mathcal{F}_\mu^* g - (y + n)\|_T^2 + \epsilon \|g\|_\mu^2 \right]. \tag{16}$$

A generalized version of this result is proven in Appendix C, which holds even when $\tilde{g}$ is only an approximate minimizer of (15).

Theorem 6 shows that $\tilde{g}$ obtained from solving the discretized regression problem provides an approximate solution to (10) and by Claim 4, $\tilde{y} = \mathcal{F}_\mu^* \tilde{g}$ solves Problem 1 with parameter $\Theta(\epsilon)$. If we have $\tilde{\tau}_{\mu,\epsilon}(t) = \tau_{\mu,\epsilon}(t)$, Theorem 6 combined with Claim 4 shows that Problem 1 with parameter $\Theta(\epsilon)$ can be solved with sample complexity $O\left(s_{\mu,\epsilon} \cdot \log s_{\mu,\epsilon}\right)$, since by (12), $\int_0^T \tau_{\mu,\epsilon}(t)dt = s_{\mu,\epsilon}$. Note that, by simply decreasing the regularization parameter in (10) by a constant factor, we can solve Problem 1 with parameter ϵ. The asymptotic complexity is identical since, by (14), for any $c \leq 1$ and any $t \in [0, T]$, $\tau_{\mu,c\epsilon}(t) \leq \frac{1}{c}\tau_{\mu,\epsilon}(t)$ and so:

$$s_{\mu,c\epsilon} \leq \frac{1}{c} s_{\mu,\epsilon}. \tag{17}$$

This proves the sample complexity result of Theorem 1. However, since it is not clear that sampling according to $\tau_{\mu,\epsilon}(t)$ can be done efficiently (or at all), it does not yet give an algorithm yielding this complexity.[17] This issue will be addressed in Section 5, where we prove Theorem 2.

We prove Theorem 6 in Appendix C. We show that leverage function sampling satisfies, with good probability, an affine embedding guarantee: that $\|\mathbf{F}^* g - (\mathbf{y}+\mathbf{n})\|_2^2 + \epsilon\|g\|_\mu^2$ closely approximates $\|\mathcal{F}_\mu^* g - (y+n)\|_T^2 + \epsilon\|g\|_\mu^2$ *for all* $g \in L_2(\mu)$. Thus, a (near) optimal solution to the discretized problem, $\min_{g \in L_2(\mu)} \left[\|\mathbf{F}^* g - (\mathbf{y} + \mathbf{n})\|_2^2 + \epsilon\|g\|_\mu^2\right]$, gives a near optimal solution to the original problem, $\min_{g \in L_2(\mu)} \left[\|\mathcal{F}_\mu^* g - (y + n)\|_T^2 + \epsilon\|g\|_\mu^2\right]$. Our proof of the affine embedding property is analogous to existing proofs for finite dimensional matrices [CW13, ACW17].

[16] If $\epsilon > \|\mathcal{K}_\mu\|_{\text{op}}$ then (10) is solved to a constant approximation factor by the trivial solution $g = 0$.

[17] We conjecture that the existential sample complexity can in fact be upper bounded by $O(s_{\mu,\epsilon})$ by adapting deterministic sampling methods for finite matrices to the operator setting [CNW16], like we do in Lemma 46.

4.2 Efficient solution of the discretized problem

Given an upper bound on the ridge leverage function $\tilde{\tau}_{\mu,\epsilon}(t) \geq \tau_{\mu,\epsilon}(t)$, we can apply Theorem 6 to approximately solve the ridge regression problem of (10) and therefore Problem 1 by Claim 4. In Section 5 we show how to obtain such an upper bound for any μ using a universal distribution.

First, however, we demonstrate how to apply Theorem 6 algorithmically. Specifically, we show how to solve the randomly discretized problem of (15) efficiently. Combined with Theorem 6 and our bound on $\tau_{\mu,\epsilon}(t)$ given in Section 5, this yields a randomized algorithm (Algorithm 1) for Problem 1. The formal analysis of Algorithm 1 is given in Theorem 7.

Algorithm 1 Time Point Sampling and Signal Reconstruction

input: Probability measure $\mu(\xi)$, $\epsilon, \delta > 0$, time bound T, and function $y : [0,T] \to \mathbb{R}$. Ridge leverage function upper bound $\tilde{\tau}_{\mu,\epsilon}(t) \geq \tau_{\mu,\epsilon}(t)$ with $\tilde{s}_{\mu,\epsilon} = \int_0^T \tilde{\tau}_{\mu,\epsilon}(t)dt$.
output: $t_1, \ldots, t_s \in [0,T]$ and $\mathbf{z} \in \mathbb{C}^s$.

1: Let $s = c \cdot \tilde{s}_{\mu,\epsilon} \cdot \left(\log \tilde{s}_{\mu,\epsilon} + \frac{1}{\delta}\right)$ for a sufficiently large constant c.
2: Independently sample $t_1, \ldots, t_s \in [0,T]$ with probability proportional to $\tilde{\tau}_{\mu,\epsilon}(t)$ and set the weight $w_i := \sqrt{\frac{1}{sT} \cdot \frac{\tilde{s}_{\mu,\epsilon}}{\tilde{\tau}_{\mu,\epsilon}(t_i)}}$.
3: Let $\mathbf{K} \in \mathbb{C}^{s \times s}$ be the matrix with $\mathbf{K}(i,j) = w_i w_j \cdot k_\mu(t_i, t_j)$.
4: Let $\bar{\mathbf{y}} \in \mathbb{C}^s$ be the vector with $\bar{\mathbf{y}}(i) = w_i \cdot [y(t_i) + n(t_i)]$.
5: Compute $\bar{\mathbf{z}} := (\mathbf{K} + \epsilon \mathbf{I})^{-1} \bar{\mathbf{y}}$.
6: **return** $t_1, \ldots, t_s \in [0,T]$ and $\mathbf{z} \in \mathbb{C}^s$ with $\mathbf{z}(i) = \bar{\mathbf{z}}(i) \cdot w_i$.

Algorithm 2 Evaluation of Reconstructed Signal

input: Probability measure $\mu(\xi)$, $t_1, \ldots, t_s \in [0,T]$, $\mathbf{z} \in \mathbb{C}^s$, and evaluation point $t \in [0,T]$.
output: Reconstructed function value $\tilde{y}(t)$.

1: For $i \in \{1, \ldots, s\}$, compute $k_\mu(t_i, t) = \int_{\xi \in \mathbb{R}} e^{-2\pi i (t_i - t)} d\mu(\xi)$.
2: **return** $\tilde{y}(t) = \sum_{i=1}^s \mathbf{z}(i) \cdot k_\mu(t_i, t)$.

Theorem 7 (Efficient signal reconstruction given leverage function upper bounds)**.** *Assume that $\epsilon \leq \|\mathcal{K}_\mu\|_{\text{op}}$.*[18] *Algorithm 1 returns $t_1, \ldots, t_s \in [0,T]$ and $\mathbf{z} \in \mathbb{C}^s$ such that $\tilde{y}(t) = \sum_{i=1}^s \mathbf{z}(i) \cdot k_\mu(t_i, t)$ (as computed in Algorithm 2) satisfies with probability $\geq 1 - \delta$:*

$$\|\tilde{y} - y\|_T^2 \leq 6\epsilon \|x\|_\mu^2 + 8\|n\|_T^2.$$

Suppose we can sample $t \in [0,T]$ with probability proportional to $\tilde{\tau}_{\mu,\epsilon}(t)$ in time W and compute the kernel function $k_\mu(t_1, t_2) = \int_{\xi \in \mathbb{R}} e^{-2\pi i (t_1 - t_2)} d\mu(\xi)$ in time Z. Algorithm 1 queries $y + n$ at s points and runs in $O\left(s \cdot W + s^2 \cdot Z + s^\omega\right)$ time[19] *where $s = O\left(\tilde{s}_{\mu,\epsilon} \cdot (\log \tilde{s}_{\mu,\epsilon} + 1/\delta)\right)$. Algorithm 2 evaluates $\tilde{y}(t)$ in $O(s \cdot Z)$ time for any t.*

Proof. In Step 2 of Algorithm 1, $t_1, \ldots, t_s$ are sampled according to $\tilde{\tau}_{\mu,\epsilon}(t)$, which upper bounds $\tau_{\mu,\epsilon}(t)$. We can thus apply Theorem 6. If the constant c in Step 1 is set large enough, with probability $\geq 1 - \delta$, letting $\mathbf{F}, \mathbf{y}$, and $\mathbf{n}$ be as defined in that theorem, (16) holds for

$$\tilde{g} = \arg\min_{g \in L_2(\mu)} \left[\|\mathbf{F}^* g - (\mathbf{y} + \mathbf{n})\|_2^2 + \epsilon \|g\|_\mu^2\right].$$

[18] As discussed for Theorem 6, if $\epsilon > \|\mathcal{K}_\mu\|_{\text{op}}$, Problem 1 is trivially solved by $\tilde{y} = 0$.

[19] Here $\omega < 2.373$ is the exponent of fast matrix multiplication. s^ω is the theoretically fastest runtime required to invert a dense $s \times s$ matrix. We note that the s^ω term may be thought of as s^3 in practice, and potentially could be accelerated using a variety of techniques for fast (regularized) linear system solvers.

Therefore, letting $\tilde{y} \stackrel{\text{def}}{=} \mathcal{F}_\mu^* \tilde{g}$ and applying Claim 4, with probability $\geq 1 - \delta$,

$$\|\tilde{y} - y\|_T^2 \leq 6\epsilon \|x\|_\mu^2 + 8\|n\|_T^2. \tag{18}$$

Further, the minimizer $\tilde{g}$ is indeed unique and can be written as (see Lemma 38 in Appendix C):

$$\tilde{g} = \mathbf{F}(\mathbf{K} + \epsilon \mathbf{I})^{-1}(\mathbf{y} + \mathbf{n}) = \mathbf{F}(\mathbf{K} + \epsilon \mathbf{I})^{-1}\bar{\mathbf{y}}$$

where $\mathbf{K} = \mathbf{F}^*\mathbf{F}$ is as defined in Step 3 of Algorithm 1 and $\bar{\mathbf{y}} = \mathbf{y} + \mathbf{n}$ is formed in Step 4. If we let $\bar{\mathbf{z}} = (\mathbf{K} + \epsilon \mathbf{I})^{-1}\bar{\mathbf{y}}$ and let $\mathbf{z}$ have $\mathbf{z}(i) = \bar{\mathbf{z}}(i) \cdot w_i$ as in Steps 5 and 6, we can see that:

$$\begin{aligned}\tilde{y} = \mathcal{F}_\mu^* \tilde{g} &= \sum_{i=1}^{s} \bar{\mathbf{z}}(i) \cdot w_i \cdot k_\mu(t_i, t) \\ &= \sum_{i=1}^{s} \mathbf{z}(i) \cdot k_\mu(t_i, t),\end{aligned}$$

giving the expression returned in Algorithm 2. Combined with (18), this completes the accuracy bound of the theorem. The runtime and sample complexity bounds follow from observing that:

- $s \cdot W$ time is required to sample $t_1, \ldots, t_s$ in Step 2.
- $s^2 \cdot Z$ time is required to form $\mathbf{K}$ in Step 3.
- s queries to $y + n$ are required to form $\bar{\mathbf{y}}$ in Step 4.
- $O(s^\omega)$ time is required to compute $\bar{\mathbf{z}} := (\mathbf{K} + \epsilon \mathbf{I})^{-1}\bar{\mathbf{y}}$ in Step 5. This runtime could potentially be improved with a variety of fast system solvers. We take s^ω as a simple upper bound.
- $O(s \cdot Z)$ time is required to compute $k(t_1, t), \ldots, k(t_s, t)$ to evaluate $\tilde{y}(t)$ in Algorithm 2.

This completes the proof of Theorem 7. □

Remark: As discussed, in Section 5 we will give a ridge leverage function upper bound that can be sampled from in $W = O(1)$ time and closely bounds the true leverage function for any μ, giving $\tilde{s}_{\mu,\epsilon} = O(s_{\mu,\epsilon} \cdot \log s_{\mu,\epsilon})$. Using this upper bound to sample time domain points, our sample complexity s is thus within a $O(\log s_{\mu,\epsilon})$ factor of the best possible using Theorem 6, which we would achieve if sampling using the true ridge leverage function.

In Appendix D we prove a tighter leverage function bound than the one in Section 5 for bandlimited signals, removing the logarithmic factor in this case. It is not hard to see that for general μ we can also achieve optimal sample complexity by further subsampling $t_1, \ldots, t_s$ using the ridge leverage scores of $\mathbf{K}^{1/2}$. These scores can be computed in $\tilde{O}(s \cdot s_{\mu,\epsilon}^2)$ time using known techniques for finite kernel matrices [MM17]. Subsampling $O\left(\frac{s_{\mu,\epsilon} \log s_{\mu,\epsilon}}{\delta^2}\right)$ time domain points according to these scores lets us approximately solve the discretized problem of (15) to error $(1 + \delta)$.

Applying the more general version of Theorem 6 stated in Appendix C, this yields an approximate solution to (10) and thus to Problem 1. For constant δ, we need just $O(s_{\mu,\epsilon} \cdot \log s_{\mu,\epsilon})$ time samples to to solve the subsampled regression problem, matching the best possible sample complexity of Theorem 6. By the lower bound given in Section 6, Theorem 24, this complexity is within a $O(\log s_{\mu,\epsilon})$ factor of optimal in nearly all settings. We conjecture that one can in fact achieve within an $O(1)$ factor of the optimal sample complexity by applying deterministic selection methods to $\mathbf{F}$ [CNW16], similar to the techniques used to prove Lemma 46.

5 A near-optimal spectrum blind sampling distribution

In the previous section, we showed how to solve Problem 1 given the ability to sample time points according to the ridge leverage function $\tau_{\mu,\epsilon}$. In general, this function depends strongly on T, μ, and ϵ, and it is not clear if it can be computed or sampled from directly.

Nevertheless, in this section we show that it is possible to efficiently obtain samples from a function that *very closely* approximates the true leverage function for *any* constraint measure μ. In particular we describe a set of closed form functions $\tilde{\tau}_\alpha(t)$, each parameterized by $\alpha > 0$. $\tilde{\tau}_\alpha$ upper bounds the leverage function $\tau_{\mu,\epsilon}$ for any μ and ϵ, as long as the statistical dimension $s_{\mu,\epsilon} \leq O(\alpha)$. Our upper bound satisfies

$$\int_0^T \tilde{\tau}_\alpha(t)dt = O(s_{\mu,\epsilon} \cdot \log s_{\mu,\epsilon}),$$

which means it can be used in place of the true ridge leverage function to give near optimal sample complexity via Theorem 6 and 7. This result is proven formally in Theorem 17, which as a consequence immediately yields our main technical result, Theorem 2. The majority of this section is devoted towards building tools necessary for proving Theorem 17.

5.1 Uniform leverage bound via Fourier sparsification

We seek a simple closed form function that upper bounds the leverage function $\tau_{\mu,\epsilon}$. Ultimately, we want this upper bound to be very tight, but a natural first question is whether it should exists at all. Is it possible to prove any finite upper bound on $\tau_{\mu,\epsilon}$ without using specific knowledge of μ?

We answer this first question by showing that $\tau_{\mu,\epsilon}$ can be upper bounded by a constant function. Specifically, we show that for $t \in [0, T]$, $\tau_{\mu,\epsilon}(t) \leq C$ for $C = \text{poly}(s_{\mu,\epsilon})$. This upper bound depends on the statistical dimension, but importantly, it does not depend on μ. Formally we show:

Theorem 8 (Uniform leverage function bound)**.** *For all $t \in [0, T]$ and $\epsilon \leq 1$*[20]

$$\tau_{\mu,\epsilon}(t) \leq \frac{2^{41}(s_{\mu,\epsilon})^5 \log^3(40 s_{\mu,\epsilon})}{T}.$$

While Theorem 8 appears to give a relatively weak bound, proving this statement is a key technical challenge. Ultimately, it is used in Section 5.3 as one of two main ingredients in proving the much tighter leverage function bound that yields Theorem 17 and Theorem 2.

Towards a proof of Theorem 8, we consider the operator $\mathcal{F}_\mu$ defined in Section 3. Since $\mathcal{F}_\mu$ has statistical dimension $s_{\mu,\epsilon}$, $\mathcal{K}_\mu = \mathcal{F}_\mu^* \mathcal{F}_\mu$ can have at most $2s_{\mu,\epsilon}$ eigenvalues $\geq \epsilon$:

$$s_{\mu,\epsilon} = \sum_{i=1}^{\infty} \frac{\lambda_i(\mathcal{K}_\mu)}{\lambda_i(\mathcal{K}_\mu) + \epsilon} \geq \sum_{i:\lambda_i(\mathcal{K}_\mu) \geq \epsilon} \frac{\lambda_i(\mathcal{K}_\mu)}{\lambda_i(\mathcal{K}_\mu) + \epsilon} \geq \frac{|i : \lambda_i(\mathcal{K}_\mu) \geq \epsilon|}{2}. \tag{19}$$

Thus, if we project $\mathcal{F}_\mu$ onto the span of $\mathcal{K}_\mu$'s top $2s_{\mu,\epsilon}$ eigenfunctions (when μ is uniform on an interval these are the prolate spherical wave functions of Slepian and Pollak [SP61]) we will approximate $\mathcal{K}_\mu$ up to its small eigenvalues. The total mass of these eigenvalues is bounded by:

$$\sum_{i:\lambda_i(\mathcal{K}_\mu) \leq \epsilon} \lambda_i(\mathcal{K}_\mu) \leq 2\epsilon \cdot \sum_{i:\lambda_i(\mathcal{K}_\mu) \leq \epsilon} \frac{\lambda_i(\mathcal{K}_\mu)}{\lambda_i(\mathcal{K}_\mu) + \epsilon} \leq 2\epsilon \cdot s_{\mu,\epsilon}.$$

[20] If $\epsilon > 1 = \text{tr}(\mathcal{K}_\mu)$, Problem 1 is trivially solved by returning $\tilde{y} = 0$.

Alternatively, instead of projecting onto the span of the eigenfunctions, we can approximate $\mathcal{K}_\mu$ nearly optimally by projecting $\mathcal{F}_\mu$ onto the span of a subset of $O(s_{\mu,\epsilon})$ of its "rows" – i.e. frequencies in the support of μ. For finite linear operators, is well known that such a subset exists: the problem of finding these subsets has been studied extensively in the literature on randomized low-rank matrix approximation under the name *column subset selection* [Sar06, BMD09, DR10]. In Appendix C we show that an analogous result extends to the continuous operator $\mathcal{F}_\mu$:

Theorem 9 (Frequency subset selection)**.** *For some $s \leq \lceil 36 \cdot s_{\mu,\epsilon} \rceil$ there exists a set of distinct frequencies $\xi_1, \ldots, \xi_s \in \mathbb{C}$ such that, if $\mathbf{C}_s : L_2(T) \to \mathbb{C}^s$ and $\mathbf{Z} : L_2(\mu) \to \mathbb{C}^s$ are defined by:*

$$[\mathbf{C}_s g](j) = \frac{1}{T}\int_0^T g(t) e^{-2\pi i \xi_j t} dt \qquad\qquad \mathbf{Z} = (\mathbf{C}_s \mathbf{C}_s^*)^{-1}\mathbf{C}_s \mathcal{F}_\mu^*, \text{[21]} \tag{20}$$

then

$$\operatorname{tr}(\mathcal{K}_\mu - \mathbf{C}_s^* \mathbf{Z}\mathbf{Z}^* \mathbf{C}_s) \leq 4\epsilon \cdot s_{\mu,\epsilon}. \tag{21}$$

Note that, if $\varphi_t \in L_2(\mu)$ is defined $\varphi_t(\xi) = e^{-2\pi i t \xi}$ and $\boldsymbol{\phi}_t \in \mathbb{C}^s$ is defined $\boldsymbol{\phi}_t(j) = \varphi_t(\xi_j)$, we have:

$$\operatorname{tr}(\mathcal{K}_\mu - \mathbf{C}_s^* \mathbf{Z}\mathbf{Z}^* \mathbf{C}_s) = \frac{1}{T}\int_{t\in[0,T]} \|\varphi_t - \mathbf{Z}^* \boldsymbol{\phi}_t\|_\mu^2 \, dt.$$

Leverage function bound proof sketch. With Theorem 9 in place, we explain how to use this result to prove Theorem 8, i.e., to establish a universal bound on the leverage function of $\mathcal{F}_\mu$. For the sake of exposition, we use the term "row" of an operator $\mathcal{A} : L_2(\mu) \to L_2(T)$ to refer to the corresponding operator restricted to some time t. We use the term "column" of an operator as the adjoint of a row of $\mathcal{A}^* : L_2(T) \to L_2(\mu)$, i.e., the adjoint operator restricted to some frequency ξ.

By Theorem 9, $\mathbf{C}_s^*\mathbf{Z} : L_2(\mu) \to L_2(T)$ (the projection of $\mathcal{F}_\mu^*$ onto the range of $\mathbf{C}_s$) closely approximates the operator $\mathcal{F}_\mu^*$ yet has columns spanned by just $O(s_{\mu,\epsilon})$ frequencies: $\xi_1, \ldots, \xi_s$. Thus, for any $\alpha \in L_2(\mu)$, $\mathbf{C}_s^*\mathbf{Z}\alpha \in L_2(T)$ is just an $O(s_{\mu,\epsilon})$ sparse Fourier function. Using the maximization characterization of Definition 3, we can thus bound the time domain ridge leverage function of $\mathbf{C}_s^*\mathbf{Z}$ by appealing to known smoothness bounds for Fourier sparse functions [CP18], even for $\epsilon = 0$. When $\epsilon = 0$, the ridge leverage function is known as the *standard leverage function* in the randomized numerical linear algebra literature, and we will refer to them as such.

We can use a similar argument to bound the row norms of the residual operator $[\mathcal{F}_\mu^* - \mathbf{C}_s^*\mathbf{Z}]$. The columns of this residual operator are each spanned by $O(s_{\mu,\epsilon})$ frequencies, and so are again sparse Fourier functions whose smoothness we can bound. This smoothness ensures that no row can have norm significantly higher than average.

Finally, we note that the time domain ridge leverage function of $\mathcal{F}_\mu$ is approximated to within a constant factor by the sum of the standard row leverage function of $\mathbf{C}_s^*\mathbf{Z}$ along with row norms of $\mathcal{F}_\mu - \mathbf{C}_s^*\mathbf{Z}$. This gives us a bound on $\mathcal{F}_\mu$'s ridge leverage function. We prove this formally below:

Theorem 10 (Ridge leverage function approximation)**.** *Let $\mathbf{C}_s$ and $\mathbf{Z}$ be the operators guaranteed to exist by Theorem 9. Let $\ell(t)$ be the standard leverage function of t in $\mathbf{C}_s^*\mathbf{Z}$:*[22]

$$\ell(t) \stackrel{\text{def}}{=} \max_{\{\alpha \in L_2(\mu):\, \|\alpha\|_\mu > 0\}} \frac{1}{T} \cdot \frac{|[\mathbf{C}_s^*\mathbf{Z}\alpha](t)|^2}{\|\mathbf{C}_s^*\mathbf{Z}\alpha\|_T^2}.$$

[21] The fact that $\xi_1, \ldots, \xi_s$ are distinct ensures that $(\mathbf{C}_s\mathbf{C}_s^*)^{-1}$ exists.

[22] Analogously to how $[\mathcal{F}_\mu^*\alpha](t)$ is used in Definition 3, while $L_2(T)$ is formally a space of equivalence classes of functions, here we use $\mathbf{C}_s^*\mathbf{Z}\alpha$ to denote the specific representative of the equivalence class $\mathbf{C}_s^*\mathbf{Z}\alpha \in L_2(T)$ given by $[\mathbf{C}_s^*\mathbf{Z}\alpha](t) = \sum_{j=1}^s [\mathbf{Z}\alpha](j) \cdot e^{2\pi i \xi_j t} = \langle \boldsymbol{\phi}_t, \mathbf{Z}\alpha \rangle_{\mathbb{C}^s}$. In this way, we can consider the pointwise value $[\mathbf{C}_s^*\mathbf{Z}\alpha](t)$.

Let $r(t)$ be the residual:

$$\frac{1}{T} \cdot \|\varphi_t - \mathbf{Z}^* \boldsymbol{\phi}_t\|_\mu^2$$

where φ_t and $\boldsymbol{\phi}_t$ are as defined in Theorem 9. Then for all t:

$$\tau_{\mu,\epsilon}(t) \leq 2 \cdot \left(\ell(t) + \frac{r(t)}{\epsilon} \right)$$

Proof. For any $\alpha \in L_2(\mu)$ we can write $[\mathcal{F}_\mu^* \alpha](t) = \langle \varphi_t, \alpha \rangle_\mu$ and $[\mathbf{C}_s^* \mathbf{Z} \alpha](t) = \langle \boldsymbol{\phi}_t, \mathbf{Z}\alpha \rangle_{\mathbb{C}^s} = \langle \mathbf{Z}^* \boldsymbol{\phi}_t, \alpha \rangle_\mu$. By the maximization characterization of the ridge leverage function in Definition 3,

$$\begin{aligned}
\tau_{\mu,\epsilon}(t) &= \frac{1}{T} \cdot \max_{\{\alpha \in L_2(\mu) : \|\alpha\|_\mu > 0\}} \frac{\langle \varphi_t, \alpha \rangle_\mu^2}{\|\mathcal{F}_\mu^* \alpha\|_T^2 + \epsilon \|\alpha\|_\mu^2} \\
&\leq \frac{2}{T} \cdot \max_{\{\alpha \in L_2(\mu) : \|\alpha\|_\mu > 0\}} \left(\frac{\langle \mathbf{Z}^* \boldsymbol{\phi}_t, \alpha \rangle_\mu^2}{\|\mathcal{F}_\mu^* \alpha\|_T^2} + \frac{\langle \varphi_t - \mathbf{Z}^* \boldsymbol{\phi}_t, \alpha \rangle_\mu^2}{\epsilon \|\alpha\|_\mu^2} \right) \\
&\leq \frac{2}{T} \cdot \max_{\{\alpha \in L_2(\mu) : \|\alpha\|_\mu > 0\}} \left(\frac{\langle \mathbf{Z}^* \boldsymbol{\phi}_t, \alpha \rangle_\mu^2}{\|\mathbf{C}_s^* \mathbf{Z} \alpha\|_T^2} + \frac{\|\varphi_t - \mathbf{Z}^* \boldsymbol{\phi}_t\|_\mu^2}{\epsilon} \right) \\
&= 2 \cdot \left(\ell(t) + \frac{r(t)}{\epsilon} \right)
\end{aligned}$$

where the second to last line follows from observing that due to Cauchy-Schwarz,

$$\langle \varphi_t - \mathbf{Z}^* \boldsymbol{\phi}_t, \alpha \rangle_\mu^2 \leq \|\alpha\|_\mu^2 \cdot \|\varphi_t - \mathbf{Z}^* \boldsymbol{\phi}_t\|_\mu^2,$$

and that, letting $\mathcal{P}_s = \mathbf{C}_s^* (\mathbf{C}_s \mathbf{C}_s^*)^{-1} \mathbf{C}_s$:

$$\begin{aligned}
\|\mathcal{F}_\mu^* \alpha\|_T^2 &= \langle \alpha, \mathcal{F}_\mu \mathcal{F}_\mu^* \alpha \rangle_\mu \\
&\geq \langle \alpha, \mathcal{F}_\mu \mathcal{P}_s \mathcal{F}_\mu^* \alpha \rangle_\mu \\
&= \langle \alpha, \mathbf{Z}^* \mathbf{C}_s \mathbf{C}_s^* \mathbf{Z} \alpha \rangle_\mu = \|\mathbf{C}_s^* \mathbf{Z} \alpha\|_T^2.
\end{aligned}$$

In the above, the inequality is due to the fact that $\mathcal{P}_s$ is an orthogonal projection, so $\mathcal{P}_s \preceq \mathcal{I}_\mu$. This completes the proof. □

With Theorem 10 in place, we now bound $\bar{\tau}_{\mu,\epsilon}(t) = 2\left(\ell(t) + \frac{r(t)}{\epsilon}\right)$, which yields a uniform bound on the true ridge leverage scores.

Lemma 11. *Let $\ell(t), r(t)$ be as defined in Theorem 10 and $\bar{\tau}_{\mu,\epsilon}(t) \stackrel{\text{def}}{=} 2 \cdot \left(\ell(t) + \frac{r(t)}{\epsilon}\right)$. For all $t \in [0, T]$:*

$$\bar{\tau}_{\mu,\epsilon}(t) \leq \frac{15400 (36 s_{\mu,\epsilon} + 2)^5 \log^3(36 s_{\mu,\epsilon} + 2)}{T}.$$

Combining Lemma 11 with Theorem 10 yields Theorem 8. We just simplify the constants by noting that for $\epsilon \leq 1$, $s_{\mu,\epsilon} \geq \frac{\operatorname{tr}(\mathcal{K}_\mu)}{2} = \frac{1}{2}$ and so $36 s_{\mu,\epsilon} + 2 \leq 40 s_{\mu,\epsilon}$.

Proof of Lemma 11. We separately bound the leverage score $\ell(t)$ and residual $r(t)$ components of $\bar{\tau}_{\mu,\epsilon}(t)$ using a similar argument based on the smoothness of sparse Fourier functions for both. Specifically, for both bounds we employ the following smoothness bound of Chen et al.:

Lemma 12 (Follows from Lemma 5.1 of [CKPS16]). *For any* $f(t) = \sum_{j=1}^{k} v_j e^{2\pi i \xi_j t}$,

$$\max_{x \in [0,T]} \frac{|f(x)|^2}{\|f\|_T^2} = 1540 \cdot k^4 \log^3 k.$$

Proof. This follows from Lemma 5.1 of [CKPS16], which gives the bound without an explicit constant. It is not hard to check that their proof gives the constant of 1540 stated above. □

Bounding the leverage scores $\ell(t)$ of $\mathbf{C}_s^*\mathbf{Z}$.

For every $\alpha \in L_2(\mu)$, $\mathbf{C}_s^*\mathbf{Z}\alpha$ is an $s = O(s_{\mu,\epsilon})$ sparse Fourier function. Specifically, we have:

$$[\mathbf{C}_s^*\mathbf{Z}\alpha](t) = \sum_{j=1}^{s} [\mathbf{Z}\alpha](j) \cdot e^{2\pi i \xi_j t},$$

for frequencies $\xi_1, \ldots, \xi_s \in \mathbb{C}$ given by Theorem 9. We can thus directly apply Lemma 12 giving for any $t \in [0, T]$:

$$\begin{aligned}
\ell(t) &\stackrel{\text{def}}{=} \max_{\{\alpha \in L_2(\mu) : \|\alpha\|_\mu > 0\}} \frac{1}{T} \cdot \frac{|[\mathbf{C}_s^*\mathbf{Z}\alpha](t)|^2}{\|\mathbf{C}_s^*\mathbf{Z}\alpha\|_T^2} \\
&\leq \max_{\{\alpha \in L_2(\mu) : \|\alpha\|_\mu > 0\}} \left[\frac{1}{T} \cdot \max_{t' \in [0,T]} \frac{|[\mathbf{C}_s^*\mathbf{Z}\alpha](t')|^2}{\|\mathbf{C}_s^*\mathbf{Z}\alpha\|_T^2} \right] \\
&\leq \frac{1540}{T} \cdot s^4 \log^3 s
\end{aligned} \tag{22}$$

Bounding the residuals $r(t)$.

We first give some intuition. To bound the squared row norms of the residual $\mathcal{F}_\mu^* - \mathbf{C}_s^*\mathbf{Z}$ we show that each "column" of this residual is an $s + 1 = O(s_{\mu,\epsilon})$ sparse Fourier function. Thus, applying Lemma 12, no entry's squared value can significantly exceed the average squared value in the column. This lets us show that no squared row norm $r(t)$ can significantly exceed the average squared row norm, which is bounded by Theorem 9.

Concretely, define $\vartheta_\xi \in L_2(T)$ by $\vartheta_\xi(t) \stackrel{\text{def}}{=} e^{2\pi i t \xi}$, and notice that given $g \in L_2(T)$ the function $\xi \mapsto \langle \vartheta_\xi, g \rangle_T$ is equal to $\mathcal{F}_\mu g$ in the $L_2(T)$ sense (i.e., is a member of the equivalence class $\mathcal{F}_\mu g$). For $\xi \in \mathbb{R}$, let $\mathbf{z}_\xi \in \mathbb{C}^s$ be given by $\mathbf{z}_\xi(j) = \langle \vartheta_\xi, \mathbf{C}_s^*(\mathbf{C}_s\mathbf{C}_s^*)^{-1}\mathbf{e}_j \rangle_T^*$ where $\mathbf{e}_j$ is the j^{th} standard basis vector in $\mathbb{C}^s$. The function $\xi \mapsto \langle z_\xi, \boldsymbol{\phi}_t \rangle = \sum_{j=1}^{s} \mathbf{z}_\xi^*(j) e^{-2\pi i \xi_j t}$ is equal in the $L_2(\mu)$ sense to $\mathbf{Z}^*\boldsymbol{\phi}_t$. Let us define:

$$r_\xi(t) = e^{-2\pi i \xi t} - \sum_{j=1}^{s} \mathbf{z}_\xi^*(j) e^{-2\pi i \xi_j t}.$$

For a fixed t, consider the function $\xi \mapsto r_\xi(t)$, which we denote by $r_\cdot(t)$. We have $r_\cdot(t) = \varphi_t - \mathbf{Z}^*\boldsymbol{\phi}_t$, again in the $L_2(\mu)$ sense. Thus, we can write

$$\begin{aligned}
r(t) &= \frac{1}{T}\|\varphi_t - \mathbf{Z}^*\boldsymbol{\phi}_t\|_\mu^2 \\
&= \frac{1}{T}\int_{\xi \in \mathbb{R}} |r_\xi(t)|^2 \, d\mu(\xi).
\end{aligned} \tag{23}$$

Further, for a fixed ξ, if we consider the function $t \mapsto r_\xi(t)$, which we denote by $r_\xi(\cdot)$, we notice that it is a $s+1 = O(s_{\mu,\epsilon})$ sparse Fourier function, so applying Lemma 12 we have for any $\xi \in \mathbb{R}$ and $t \in [0,T]$:

$$\frac{|r_\xi(t)|^2}{\|r_\xi(\cdot)\|_T^2} \leq 1540(s+1)^4 \log^3(s+1). \tag{24}$$

Combining (24) with (23) we can thus bound for any $t \in [0,T]$:

$$\begin{aligned} r(t) &\leq 1540(s+1)^4 \log^3(s+1) \cdot \frac{1}{T} \int_{\xi \in \mathbb{R}} \|r_\xi(\cdot)\|_T^2 \, d\mu(\xi) \\ &= 1540(s+1)^4 \log^3(s+1) \cdot \frac{1}{T^2} \int_{w \in [0,T]} \int_{\xi \in \mathbb{R}} |r_\xi(w)|^2 \, d\mu(\xi) \, dw \\ &= 1540(s+1)^4 \log^3(s+1) \cdot \frac{1}{T^2} \int_{w \in [0,T]} \|\varphi_w - \mathbf{Z}^* \boldsymbol{\phi}_w\|_\mu^2 \, dw \end{aligned} \tag{25}$$

where the last bound again follows from (23). By Theorem 9 we have $\frac{1}{T}\int_{w\in[0,T]} \|\varphi_w - \mathbf{Z}^*\boldsymbol{\phi}_w\|_\mu^2 \, dw \leq 4\epsilon \cdot s_{\mu,\epsilon}$. Plugging into (25) and using that we can choose $s \leq 36 \cdot s_{\mu,\epsilon} + 1$, for all $t \in [0,T]$:

$$r(t) \leq \frac{\epsilon \cdot 6160(36 s_{\mu,\epsilon} + 2)^5 \log^3(36 s_{\mu,\epsilon} + 2)}{T}. \tag{26}$$

Combining (22) and (26) completes the proof of Lemma 11 since $\bar{\tau}_{\mu,\epsilon}(t) \stackrel{\text{def}}{=} 2 \cdot \left(\ell(t) + \frac{r(t)}{\epsilon}\right)$ and thus

$$\bar{\tau}_{\mu,\epsilon}(t) \leq \frac{15400(36 s_{\mu,\epsilon} + 2)^5 \log^3(36 s_{\mu,\epsilon} + 2)}{T}.$$

□

Theorem 8 gives a universal uniform bound on the ridge leverage scores corresponding to measure μ in terms of $s_{\mu,\epsilon}$. If we directly sample time points according to the uniform distribution over $[0,T]$, this theorem shows that $\text{poly}(s_{\mu,\epsilon})$ samples and $\text{poly}(s_{\mu,\epsilon})$ runtime suffice to apply Theorem 7 and solve Problem 1 with good probability. This is already a surprising result, showing that the simplest sampling scheme, uniform random sampling, can give bounds in terms of the optimal complexity $s_{\mu,\epsilon}$ for *any* μ. Existing methods with similar complexity, such as those that interpolate bandlimited signals using prolate spheroidal wave functions [XRY01, STR06] require nonuniform sampling. Methods that use uniform sampling, such as truncated Whittaker-Shannon, have sample complexity depending polynomially rather than logarithmically on the desired error ϵ.

5.2 Gap-based leverage score bound

Our final result gives a much tighter bound on the ridge leverage scores than the uniform bound of Theorem 8. The key idea is to show that the bound is loose for t bounded away from the edges of $[0,T]$. Specifically we have:

Theorem 13 (Gap-Based Leverage Score Bound)**.** *For all t,*

$$\tau_{\mu,\epsilon}(t) \leq \frac{s_{\mu,\epsilon}}{\min(t, T-t)}.$$

Proof. Consider $t \in [0, T/2]$. We will show that $\tau_{\mu,\epsilon}(t) \leq \frac{s_{\mu,\epsilon}}{t}$. A symmetric proof will hold for $t \in [T/2, T]$, giving the theorem. We define an auxiliary operator: $\mathcal{F}_{\mu,t} : L_2(T) \to L_2(\mu)$ which is given by restricting the integration in $\mathcal{F}_\mu$ to $[0, t]$. Specifically, for $f \in L_2(T)$ we have:

$$[\mathcal{F}_{\mu,t} f](\xi) = \frac{1}{T} \int_0^t f(s) e^{-2\pi i s\xi} \, ds. \tag{27}$$

We can see that $[\mathcal{F}^*_{\mu,t} g](s) = \int_{\mathbb{R}} g(\xi) e^{2\pi i s\xi} \, d\mu(\xi)$ for $s \in [0, t]$ and $[\mathcal{F}^*_{\mu,t} g](s) = 0$ for $s \in (t, T]$. We will use the leverage score of some $s \in [0, t]$ in the restricted operator $\mathcal{F}_{\mu,t}$ to upper bound those of t in $\mathcal{F}_\mu$. We start by defining these scores analogously to Definition 3 for $\mathcal{F}_\mu$.

Definition 4 (Restricted ridge leverage scores)**.** *For probability measure μ on $\mathbb{R}$, time length T, $t \in [0, T]$ and $\epsilon \geq 0$, define the ϵ-ridge leverage score of $s \in [0, t]$ in $\mathcal{F}_{\mu,t}$ as:*

$$\tau_{\mu,\epsilon,t}(s) = \frac{1}{T} \cdot \max_{\{\alpha \in L_2(\mu) : \|\alpha\|_\mu > 0\}} \frac{|[\mathcal{F}_{\mu,t}\alpha](s)|^2}{\|\mathcal{F}^*_{\mu,t}\alpha\|_T^2 + \epsilon\|\alpha\|_\mu^2}.$$

We have the following leverage score properties, analogous to those given for $\mathcal{F}_\mu$ in Theorem 5:

Theorem 14 (Restricted leverage score properties)**.** *Let $\tau_{\mu,\epsilon,t}(s)$ be as defined in Definition 4.*

- *The leverage scores integrate to the statistical dimension:*

$$\int_0^t \tau_{\mu,\epsilon,t}(s) \, ds = s_{\mu,\epsilon,t} \stackrel{\text{def}}{=} \operatorname{tr}(\mathcal{F}^*_{\mu,t}\mathcal{F}_{\mu,t}(\mathcal{F}^*_{\mu,t}\mathcal{F}_{\mu,t} + \epsilon\mathcal{I}_T)^{-1}). \tag{28}$$

- *Inner Product Characterization: Letting $\varphi_s \in L_2(\mu)$ have $\varphi_s(\xi) = e^{-2\pi i s\xi}$ for $s \in [0, t]$,*

$$\tau_{\mu,\epsilon,t}(s) = \frac{1}{T} \cdot \langle \varphi_s, (\mathcal{F}_{\mu,t}\mathcal{F}^*_{\mu,t} + \epsilon\mathcal{I}_\mu)^{-1}\varphi_s \rangle_\mu. \tag{29}$$

- *Minimization Characterization:*

$$\tau_{\mu,\epsilon,t}(s) = \frac{1}{T} \cdot \min_{\beta \in L_2(T)} \frac{\|\mathcal{F}_{\mu,t}\beta - \varphi_s\|_\mu^2}{\epsilon} + \|\beta\|_T^2. \tag{30}$$

We first show that the restricted leverage scores of Definition 4 are not too large on average.

Claim 15 (Restricted statistical dimension bound)**.**

$$\int_0^T \tau_{\mu,\epsilon,t}(s) \, ds \leq s_{\mu,\epsilon}. \tag{31}$$

Proof. Via (28) we have $\int_0^t \tau_{\mu,\epsilon,t}(s) \, ds = s_{\mu,\epsilon,t}$ which we can write as:

$$s_{\mu,\epsilon,t} = \operatorname{tr}(\mathcal{F}^*_{\mu,t}\mathcal{F}_{\mu,t}(\mathcal{F}^*_{\mu,t}\mathcal{F}_{\mu,t} + \epsilon\mathcal{I}_T)^{-1}) = \sum_{i=1}^{\infty} \frac{\lambda_i(\mathcal{F}^*_{\mu,t}\mathcal{F}_{\mu,t})}{\lambda_i(\mathcal{F}^*_{\mu,t}\mathcal{F}_{\mu,t}) + \epsilon}.$$

From Claim 35 we have $\mathcal{F}_{\mu,t}\mathcal{F}^*_{\mu,t} \preceq \mathcal{F}_\mu\mathcal{F}^*_\mu = \mathcal{G}_\mu$. Since $(\mathcal{F}_{\mu,t}\mathcal{F}^*_{\mu,t} + \epsilon\mathcal{I}_\mu)^{-1/2}(\mathcal{F}_{\mu,t}\mathcal{F}^*_{\mu,t})(\mathcal{F}_{\mu,t}\mathcal{F}^*_{\mu,t} + \epsilon\mathcal{I}_\mu)^{-1/2} = \mathcal{I}_\mu - \epsilon(\mathcal{F}_{\mu,t}\mathcal{F}^*_{\mu,t} + \epsilon\mathcal{I}_\mu)^{-1}$ and $(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}\mathcal{G}_\mu(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2} = \mathcal{I}_\mu - \epsilon(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}$ we have from Claim 27 $(\mathcal{F}_{\mu,t}\mathcal{F}^*_{\mu,t} + \epsilon\mathcal{I}_\mu)^{-1/2}(\mathcal{F}_{\mu,t}\mathcal{F}^*_{\mu,t})(\mathcal{F}_{\mu,t}\mathcal{F}^*_{\mu,t} + \epsilon\mathcal{I}_\mu)^{-1/2} \preceq (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}\mathcal{G}_\mu(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}$, and since the trace is monotone for trace-class operators, we have

$$\begin{aligned} s_{\mu,\epsilon,t} &= \operatorname{tr}((\mathcal{F}_{\mu,t}\mathcal{F}^*_{\mu,t} + \epsilon\mathcal{I}_\mu)^{-1/2}(\mathcal{F}_{\mu,t}\mathcal{F}^*_{\mu,t})(\mathcal{F}_{\mu,t}\mathcal{F}^*_{\mu,t} + \epsilon\mathcal{I}_\mu)^{-1/2}) \\ &\leq \operatorname{tr}((\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}\mathcal{G}_\mu(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}) = s_{\mu,\epsilon} \end{aligned}$$

which gives the claim. □

From Claim 15 we immediately have:

Claim 16. *There exists* $s^\star \in [0,t]$ *with* $\tau_{\mu,\epsilon,t}(s^\star) \leq \frac{s_{\mu,\epsilon}}{t}$.

Proof. Assume for the sake of contradiction that $\tau_{\mu,\epsilon,t}(s) > \frac{s_{\mu,\epsilon}}{t}$ for all $s \in [0,t]$. The by (28),

$$\int_0^t \tau_{\mu,\epsilon,t}(s)\, ds > t \cdot \frac{s_{\mu,\epsilon}}{t} = s_{\mu,\epsilon}.$$

This contradicts Claim 15, giving the claim. □

We now show that the leverage score of $s^\star$ in $\mathcal{F}_{\mu,t}$ upper bounds the leverage score of t in $\mathcal{F}_\mu$, completing the proof of Theorem 13. We apply the minimization characterization of Theorem 14, equation (30), showing that by simply shifting an optimal solution for $s^\star$ we can show the existence of a good solution for t, upper bounding its leverage score by that of $s^\star$ and giving $\tau_{\mu,\epsilon}(t) \leq \tau_{\mu,\epsilon,t}(s^\star) \leq \frac{s_{\mu,\epsilon}}{t}$ by Claim 16.

Formally, by Claim 16 and (30), there is some $\beta^\star \in L_2(T)$ achieving:

$$\frac{1}{T} \cdot \frac{\|\mathcal{F}_{\mu,t}\beta^\star - \varphi_{s^\star}\|_\mu^2}{\epsilon} + \|\beta^\star\|_T^2 = \tau_{\mu,\epsilon,t}(s^\star) \leq \frac{s_{\mu,\epsilon}}{t}. \tag{32}$$

We can assume without loss of generality that $\beta^\star(s) = 0$ for $s \notin [0,t]$, since $\mathcal{F}_{\mu,t}\beta^\star$ is unchanged if we set $\beta^\star(s) = 0$ on this range and since doing this cannot increase $\|\beta\|_T^2$. Now, let $\bar{\beta} \in L_2(T)$ be given by $\bar{\beta}(s) = \beta^\star(s - (t - s^\star))$. That is, $\bar{\beta}$ is just $\beta^\star$ shifted from the range $[0,t]$ to the range $[t - s^\star, 2t - s^\star]$. Note that since we are assuming $t \leq T/2$, $[t - s^\star, 2t - s^\star] \subset [0,T]$. For any ξ:

$$\begin{aligned}
[\mathcal{F}_\mu \bar{\beta}](\xi) &= \frac{1}{T}\int_0^T \bar{\beta}(s) e^{-2\pi i s\xi} ds \\
&= \frac{1}{T}\int_{t-s^\star}^{2t-s^\star} \beta^\star(s - (t - s^\star)) e^{-2\pi i s\xi} ds \\
&= \frac{1}{T}\int_0^t \beta^\star(s) e^{-2\pi i (s + (t-s^\star))\xi} ds \\
&= [\mathcal{F}_{\mu,t}\beta^\star](\xi) \cdot e^{-2\pi i (t - s^\star)\xi}.
\end{aligned} \tag{33}$$

Now,

$$\varphi_t(\xi) = e^{-2\pi i t\xi} = e^{-2\pi i (t-s^\star)\xi} \cdot \varphi_{s^\star}(\xi).$$

Combined with (33) this gives:

$$\begin{aligned}
\|\mathcal{F}_\mu \bar{\beta} - \varphi_t\|_\mu^2 = \int_\xi \left|[\mathcal{F}_\mu \bar{\beta}](\xi) - \varphi_t\right|^2 d\mu(\xi) &= \int_\xi \left|([\mathcal{F}_{\mu,t}\beta^\star](\xi) - \varphi_{s^\star}) \cdot e^{-2\pi i (t-s^\star)\xi}\right|^2 d\mu(\xi) \\
&= \int_\xi \left|([\mathcal{F}_{\mu,t}\beta^\star](\xi) - \varphi_{s^\star})\right|^2 d\mu(\xi) \\
&= \|\mathcal{F}_{\mu,t}\beta^\star - \varphi_{s^\star}\|_\mu^2.
\end{aligned} \tag{34}$$

Finally, noting that $\|\bar{\beta}\|_T = \|\beta^\star\|_T$ and applying the minimization characterization of Theorem 5, the bound in (34) along with (32) gives:

$$\tau_{\mu,\epsilon}(t) \leq \frac{1}{T} \cdot \frac{\|\mathcal{F}_\mu \bar{\beta} - \varphi_t\|_\mu^2}{\epsilon} + \|\bar{\beta}\|_T^2 = \frac{\|\mathcal{F}_{\mu,t}\beta^\star - \varphi_{s^\star}\|_\mu^2}{\epsilon} + \|\beta^\star\|_T^2 \leq \frac{s_{\mu,\epsilon}}{t},$$

which completes the theorem.

□

5.3 Nearly tight leverage score bound

Combining Theorems 8 and 13 gives our tight, spectrum blind leverage score bound:

Theorem 17 (Spectrum Blind Leverage Score Bound)**.** *For any $\alpha, T \geq 0$ let $\tilde{\tau}_\alpha(t)$ be given by:*

$$\tilde{\tau}_\alpha(t) = \begin{cases} \frac{\alpha}{256 \cdot \min(t, T-t)} & \text{for } t \in [T/\alpha^6, T(1-1/\alpha^6)] \\ \frac{\alpha^6}{T} & \text{for } t \in [0, T/\alpha^6] \cup [T(1-1/\alpha^6), T]. \end{cases}$$

For any probability measure μ, $T \geq 0$, $0 \leq \epsilon \leq 1$ and $t \in [0, T]$, if $\alpha \geq 256 \cdot s_{\mu,\epsilon}$:

$$\tau_{\mu,\epsilon}(t) \leq \tilde{\tau}_\alpha(t) \text{ and } \tilde{s}_\alpha \stackrel{\text{def}}{=} \int_0^T \tilde{\tau}_\alpha(t)\, dt \leq \frac{\alpha \cdot \log \alpha}{19}.$$

A visualization of $\tilde{\tau}_\alpha$ is given in Figure 3.

Proof. The fact that $\tau_{\mu,\epsilon}(t) \leq \tilde{\tau}_\alpha(t)$ follows from Theorems 8 and 13:

- For $t \in [T/\alpha^6, T(1-1/\alpha^6)]$, by Theorem 13 if $\alpha \geq 256 \cdot s_{\mu,\epsilon}$ we have

$$\tilde{\tau}_\alpha(t) = \frac{\alpha}{256 \cdot \min(t, T-t)} \geq \tau_{\mu,\epsilon}(t).$$

- For $t \in [0, T/\alpha^6] \cup [T(1-1/\alpha^6), T]$, by Theorem 8 we can bound,

$$\tau_{\mu,\epsilon}(t) \leq \frac{2^{41} s_{\mu,\epsilon}^5 \log^3(40 s_{\mu,\epsilon})}{T} \leq \frac{2^{47} s_{\mu,\epsilon}^6}{T} \leq \frac{\alpha^6}{T}$$

  for $\alpha \geq 256 \cdot s_{\mu,\epsilon}$. Note that the second inequality uses that $\log^3(40x) \leq 64x$ for any x.

The integral of the approximate scores $\tilde{s}_\alpha$ is bounded as:

$$\begin{aligned} \int_0^T \tilde{\tau}_\alpha(t)\, dt &= \int_{T/\alpha^6}^{T(1-1/\alpha^6)} \frac{\alpha}{256 \cdot \min(t, T-t)}\, dt + 2\int_0^{T/\alpha^6} \frac{\alpha^6}{T}\, dt \\ &= \frac{2}{256} \int_{T/\alpha^6}^{T/2} \frac{\alpha}{t}\, dt + 2 \\ &= \frac{\alpha}{128} \cdot [\log(T/2) - \log(T/\alpha^6)] + 2 \\ &\leq \frac{6\alpha \log \alpha}{128} + 2 \leq \frac{\alpha \log \alpha}{19}. \end{aligned} \tag{35}$$

where the last inequality follows since for $\epsilon \leq 1$, $s_{\mu,\epsilon} \geq 1/2$ and so $\alpha \geq 128$. □

5.4 Putting it all together: generic signal reconstruction

Finally, we combine the leverage score bound of Theorem 17 with Theorem 7 to give our main algorithmic result, Theorem 3 (and as a corollary, Theorem 2). We state the full theorem below:

Theorem 3 (Main result, algorithmic complexity)**.** *Consider any measure μ, for which we can compute the kernel function $k_\mu(t_1, t_2) = \int_{\xi \in \mathbb{R}} e^{-2\pi i (t_1 - t_2)} d\mu(\xi)$ for any $t_1, t_2 \in [0, T]$ in time Z.*

Let $\tilde{\tau}_\alpha(t)$ be as defined in Theorem 17. For any $\epsilon \leq \|\mathcal{K}_\mu\|_{\text{op}}$ and $T > 0$, let $\tilde{\tau}_{\mu,\epsilon}(t) = \tilde{\tau}_\alpha(t)$ for $\alpha = \beta \cdot s_{\mu,\epsilon}$ with $\beta \geq 256$. Algorithm 1 applied with $\tilde{\tau}_{\mu,\epsilon}(t)$ and failure probability δ returns

$t_1, \ldots, t_s \in [0, T]$ and $\mathbf{z} \in \mathbb{C}^s$ such that $\tilde{y}(t) = \sum_{i=1}^{s} \mathbf{z}(i) \cdot k_\mu(t_i, t)$ solves Problem 1 with parameter 6ϵ and probability $\geq 1 - \delta$. That is, with probability of at least $1 - \delta$:

$$\|\tilde{y} - y\|_T^2 \leq 6\epsilon \|x\|_\mu^2 + 8\|n\|_T^2.$$

The algorithm queries $y + n$ at s points and runs in $O\left(s^2 \cdot Z + s^\omega\right)$ time where

$$s = O\left(\beta \cdot s_{\mu,\epsilon} \log(\beta \cdot s_{\mu,\epsilon}) \cdot [\log(\beta \cdot s_{\mu,\epsilon}) + 1/\delta]\right) = \tilde{O}\left(\frac{\beta \cdot s_{\mu,\epsilon}}{\delta}\right).$$

The output $\tilde{y}(t)$ can be evaluated in $O(s \cdot Z)$ time for any t using Algorithm 2.

Note that if we want to solve Problem 1 with parameter ϵ, it suffices to apply Theorem 3 with parameter $\epsilon' = \epsilon/6$. The asymptotic complexity will be identical since, by (17), $s_{\mu,\epsilon/6} \leq 6 s_{\mu,\epsilon}$.

Proof. The theorem follows directly from Theorem 7, along with Theorem 17 which shows that, for $\alpha = \beta \cdot s_{\mu,\epsilon}$ with $\beta \geq c_1$ and $\tilde{\tau}_{\mu,\epsilon}(t) = \tilde{\tau}_\alpha(t)$ we have:

1. $\tilde{\tau}_{\mu,\epsilon}(t) \geq \tau_{\mu,\epsilon}(t)$ for all $t \in [0, T]$.
2. $\tilde{s}_{\mu,\epsilon} = \int_0^T \tilde{\tau}_{\mu,\epsilon}(t)\, dt = O\left(\beta \cdot s_{\mu,\epsilon} \log(\beta \cdot s_{\mu,\epsilon})\right)$.

The runtime bound follows after noting that we can sample according to τ_α in $W = O(1)$ time using inverse transform sampling since it is straightforward to derive an explicit expression for the CDF and compute the inverse (see (35)).

□

6 Lower bound

We conclude by showing that the statistical dimension $s_{\mu,\epsilon}$ tightly characterizes the sample complexity of solving Problem 1, under a very mild assumption on μ that holds for all natural constraints we discuss in this paper. Thus, Theorem 1 is tight up to logarithmic factors.

We first define a quantity, $n_{\mu,\epsilon}$ that gives a natural lower bound on $s_{\mu,\epsilon}$. For any μ, ϵ, let

$$n_{\mu,\epsilon} \stackrel{\text{def}}{=} \sum_{i=1}^{\infty} \mathbb{I}[\lambda_i(\mathcal{K}_\mu) \geq \epsilon]. \tag{36}$$

That is, $n_{\mu,\epsilon}$ is the number of eigenvalues of $\mathcal{K}_\mu$ that are larger than ϵ. As shown in (19), we always have $n_{\mu,\epsilon} \leq 2 s_{\mu,\epsilon}$. We first prove that solving Problem 1 requires $\Omega(n_{\mu,\epsilon})$ samples. We then show that, under a very mild constraint on μ (which holds for all μ we consider including sparse, bandlimited, multiband, Gaussian, and Cauchy-Lorentz), $n_{\mu,\epsilon} = \Omega(s_{\mu,\epsilon})$. Thus, $s_{\mu,\epsilon}$ gives a tight bound on the query complexity of solving Problem 1.

Theorem 18 (Lower bound in terms of eigenvalue count)**.** *Consider a measure μ, an error parameter $\epsilon > 0$, and any (possibly randomized) algorithm that solves Problem 1 with probability $\geq 2/3$ for any function y and makes at most r (possibly adaptive) queries on any input. Then $r \geq n_{\mu,72\epsilon}/20$.*

Proof. We describe a distribution on inputs y on which any deterministic algorithm that takes $r = o(n_{\mu,72\epsilon})$ samples on any input fails with probability $> 1/3$. The theorem then follows by Yao's principle.

Notation: Let $v_1, \ldots, v_{n_{\mu,72\epsilon}} \in L_2(\mu)$ be the eigenfunctions of $\mathcal{G}_\mu$ corresponding to its top $n_{\mu,72\epsilon}$ eigenvalues. Let $\mathbf{Z} : L_2(\mu) \to \mathbb{C}^{n_{\mu,72\epsilon}}$ be the operator with v_i as its i^{th} row – i.e., $[\mathbf{Z}g](i) = \langle v_i, g\rangle_\mu$. Note that $\mathbf{Z}$ has orthonormal rows. Let $\mathbf{D} \in \mathbb{R}^{n_{\mu,72\epsilon} \times n_{\mu,72\epsilon}}$ be a diagonal matrix with $\mathbf{D}_{ii} = \sqrt{\lambda_i(\mathcal{K}_\mu)}$. Let $\mathbf{U} = \mathcal{F}_\mu^* \mathbf{Z}^* \mathbf{D}^{-1}$. We can see that $\mathbf{Z}\mathcal{F}_\mu \mathcal{F}_\mu^* \mathbf{Z}^* = \mathbf{Z}\mathcal{G}_\mu \mathbf{Z}^* = \mathbf{D}^2$ and hence, $\mathbf{U}^*\mathbf{U} = \mathbf{D}^{-1}\mathbf{Z}\mathcal{F}_\mu \mathcal{F}_\mu^* \mathbf{Z}^* \mathbf{D}^{-1} = \mathbf{I}$. While not needed for our proof, we can check that $\mathbf{U} : \mathbb{C}^{n_{\mu,72\epsilon}} \to L_2(T)$ is an operator with columns corresponding to all eigenfunctions of $\mathcal{K}_\mu$ with eigenvalue $\geq 72\epsilon$.

Hard Input Distribution: Let $\mathbf{c} \in \mathbb{R}^{n_{\mu,72\epsilon}}$ be a random vector with each entry distributed independently as a Gaussian: $\mathbf{c}(i) \sim \mathcal{N}(0, \frac{1}{n_{\mu,72\epsilon}})$. Let $\bar{\mathbf{c}} = \mathbf{D}^{-1}\mathbf{c}$, $x = \mathbf{Z}^*\bar{\mathbf{c}}$, and the random input be $y = \mathcal{F}_\mu^* x$. That is, $y = \mathcal{F}_\mu^* \mathbf{Z}^* \mathbf{D}^{-1}\mathbf{c} = \mathbf{U}\mathbf{c}$ is a random linear combination of the top eigenfunctions of $\mathcal{K}_\mu$. While, formally, $\mathcal{F}_\mu^* x \in L_2(T)$ is an equivalence class of functions, since our input model requires that y admits pointwise evaluation, we will abuse notation, letting y denote the member of this class with $y(t) = \langle \varphi_t, \mathbf{Z}^*\mathbf{D}^{-1}\mathbf{c}\rangle_\mu = \langle \mathbf{D}^{-1}\mathbf{Z}\varphi_t, \mathbf{c}\rangle$, where $\varphi_t(\xi) = e^{-2\pi i t \xi}$.

We prove that accurately reconstructing y drawn from the hard input distribution yields an accurate reconstruction of the random vector $\mathbf{c}$. Since $\mathbf{c}$ is $n_{\mu,72\epsilon}$ dimensional, this reconstruction requires $\Omega(n_{\mu,72\epsilon})$ samples, giving us a lower bound for accurately reconstructing y.

Claim 19. *For random x distributed as described above, with probability $\geq 5/6$, $\|x\|_\mu^2 \leq \frac{1}{12\epsilon}$.*

Proof.

$$\|x\|_\mu^2 = \langle \mathbf{Z}^*\bar{\mathbf{c}}, \mathbf{Z}^*\bar{\mathbf{c}}\rangle_\mu = \langle \bar{\mathbf{c}}, \mathbf{Z}\mathbf{Z}^*\bar{\mathbf{c}}\rangle = \|\bar{\mathbf{c}}\|_2^2.$$

We then bound $\|\bar{\mathbf{c}}\|_2^2 \leq \|\mathbf{c}\|_2^2 / \lambda_{n_{\mu,72\epsilon}}(\mathcal{K}_\mu) \leq \frac{\|\mathbf{c}\|_2^2}{72\epsilon}$ since $\lambda_{n_{\mu,72\epsilon}}(\mathcal{K}_\mu) \geq 72\epsilon$ by definition. Finally, note that $\|\mathbf{c}\|_2^2$ is a Chi-squared random variable, with $\mathbb{E}[\|\mathbf{c}\|_2^2] = 1$. So loosely, by Markov's inequality, with probability $\geq 5/6$, $\|\mathbf{c}\|_2^2 \leq 6$, which gives the claim. □

From Claim 19 we have:

Claim 20. *Given random input $y = \mathcal{F}_\mu^* x$ generated as described above, with probability $\geq 5/6$, to solve Problem 1, an algorithm must return a representation of $\tilde{y}$ with $\|y - \tilde{y}\|_T^2 \leq \frac{1}{12}$.*

Proof. Solving Problem 1 requires finding a representation of $\tilde{y}$ with $\|y - \tilde{y}\|_T^2 \leq \epsilon\|x\|_\mu^2 + C\|n\|_T^2$. By Claim 19 and the fact that for our input $\|n\|_T^2 = 0$, with probability $\geq 5/6$ one has that $\epsilon\|x\|_\mu^2 + C\|n\|_T^2 \leq \frac{1}{12}$, yielding the claim. □

We next show that finding a $\tilde{y}$ satisfying the condition of Claim 20 is at least as hard as finding an accurate approximation to $\mathbf{c}$.

Claim 21. *For $\tilde{y}$ with $\|y - \tilde{y}\|_T^2 \leq \frac{1}{12}$, $\tilde{\mathbf{c}} = \mathbf{U}^*\tilde{y}$ satisfies $\|\mathbf{c} - \tilde{\mathbf{c}}\|_2^2 \leq \frac{1}{12}$.*

Proof. Recalling that $y = \mathbf{U}\mathbf{c}$, for $\tilde{\mathbf{c}} = \mathbf{U}^*\tilde{y}$ we have:

$$\tilde{\mathbf{c}} = \mathbf{U}^* y + \mathbf{U}^*(\tilde{y} - y) = \mathbf{U}^*\mathbf{U}\mathbf{c} + \mathbf{U}^*(\tilde{y} - y).$$

Recalling that $\mathbf{U}^*\mathbf{U} = \mathbf{I}$ we thus have:

$$\begin{aligned}\|\mathbf{c} - \tilde{\mathbf{c}}\|_2^2 &= \|\mathbf{U}^*(\tilde{y} - y)\|_2^2 \\ &\leq \|\tilde{y} - y\|_T^2 \leq \frac{1}{12}.\end{aligned}$$

The second to last inequality follows since $\mathbf{U}^*\mathbf{U} = \mathbf{I}$ and $\mathbf{U}\mathbf{U}^*$ are finite rank, so are compact and share the same non-zero eigenvalues. Thus, $\mathbf{U}\mathbf{U}^* \preceq \mathcal{I}_T$ [HN01, Lemma 8.26]. This completes the claim. □

Combining Claims 20 and 21 we have:

Claim 22. *If a deterministic algorithm solves Problem 1 with probability $\geq 2/3$ over our random input $y = \mathbf{U}\mathbf{c}$, then with probability $\geq 1/2$, letting $\tilde{y}$ be the output of the algorithm, $\tilde{\mathbf{c}} = \mathbf{U}^*\tilde{y}$ satisfies $\|\mathbf{c} - \tilde{\mathbf{c}}\|_2^2 \leq \frac{1}{12}$.*

Proof. If an algorithm solves Problem 1 probability $\geq 2/3$ then by Claim 20, it returns $\tilde{y}$ with $\|y - \tilde{y}\|_T^2 \leq \frac{1}{12}$ with probability $\geq 2/3 - 1/6 = 1/2$. Thus, by Claim 21, $\tilde{\mathbf{c}}$ satisfies $\|\mathbf{c} - \tilde{\mathbf{c}}\|_2^2 \leq \frac{1}{12}$ with probability $\geq 1/2$. □

Finally, we complete the proof of Theorem 18 by arguing that if $\tilde{y}$ is formed using $o(n_{\mu,72\epsilon})$ queries, then for $\tilde{\mathbf{c}} = \mathbf{U}^*\tilde{y}$, $\|\mathbf{c} - \tilde{\mathbf{c}}\|_2^2 > \frac{1}{12}$ with good probability. Thus the bound of Claim 22 cannot hold and so $\tilde{y}$ cannot be a solution to Problem 1 with good probability.

Assume for the sake of contradiction that there is a deterministic algorithm solving Problem 1 with probability $\geq 2/3$ over the random input $\mathbf{U}\mathbf{c}$ that makes $r = \frac{n_{\mu,72\epsilon}}{20}$ queries on any input (note that if there exists an algorithm that makes fewer queries on some inputs, we can always modify it to make exactly $\frac{n_{\mu,72\epsilon}}{20}$ queries and return the same output.)

As discussed, each query to y is a query to $y(t) = \langle \mathbf{D}^{-1}\mathbf{Z}\varphi_t, \mathbf{c}\rangle$. Consider a deterministic function Q, that is given input $\mathbf{V} \in \mathbb{C}^{i \times n_{\mu,72\epsilon}}$ (for any positive integer i) and outputs $Q(\mathbf{V}) \in \mathbb{C}^{n_{\mu,72\epsilon} \times n_{\mu,72\epsilon}}$ such that $Q(\mathbf{V})$ has orthonormal rows with the first i spanning the i rows of $\mathbf{V}$. For example, Q may run Gram-Schmidt orthogonalization on $\mathbf{V}$ fixing its first $\text{rank}(\mathbf{V}) \leq i$ rows and then fill out the remaining $n_{\mu,72\epsilon} - \text{rank}(\mathbf{V})$ rows using some canonical approach. Letting $\mathbf{D}^{-1}\mathbf{Z}\varphi_{t_1}, \ldots, \mathbf{D}^{-1}\mathbf{Z}\varphi_{t_r}$ denote the queries made by our algorithm on random input $\mathbf{c}$, let $\mathbf{Q}^i = Q([\mathbf{D}^{-1}\mathbf{Z}\varphi_{t_1}, \ldots, \mathbf{D}^{-1}\mathbf{Z}\varphi_{t_i}]^*)$. That is $\mathbf{Q}^i$ is an orthonormal matrix whose first i rows span our first i queries. Note that since our algorithm is deterministic, $\mathbf{Q}^i$ is a deterministic function of the random input $\mathbf{c}$. We have the following claim:

Claim 23. *Conditioned on the queries $y(t_1), \ldots y(t_r)$, for $j > r$, each $[\mathbf{Q}^r\mathbf{c}](j)$ is distributed independently as $\mathcal{N}\left(0, \frac{1}{n_{\mu,72\epsilon}}\right)$.*

Proof. We prove the claim via induction on the number of queries considered. For the base case set $i = 1$. $\mathbf{Q}^1$ is a deterministic matrix (since the choice of our first query is made determinstically before seeing any input) and so by the rotational invariance of the Gaussian distribution, the entries of $\mathbf{Q}^1\mathbf{c}$ are distributed independently as $\mathcal{N}\left(0, \frac{1}{n_{\mu,72\epsilon}}\right)$ (the same as the entries of $\mathbf{c}$). The first row of $\mathbf{Q}^1$ spans our first query, and thus this row is just equal to $\mathbf{D}^{-1}\mathbf{Z}\varphi_{t_1}$ scaled to have unit norm. Thus $y(t_1) = \mathbf{D}^{-1}\mathbf{Z}\varphi_{t_1}\mathbf{c}$ is just a fixed scaling of $[\mathbf{Q}^1\mathbf{c}](1)$. So conditioning on $y(t_1)$, we still have $[\mathbf{Q}^1\mathbf{c}](j)$ for $j > 1$ distributed independently as $\mathcal{N}\left(0, \frac{1}{n_{\mu,72\epsilon}}\right)$.

Now, consider $i > 1$. By the inductive assumption, conditioned on $y(t_1), \ldots y(t_{i-1})$, for $j \geq i$, $[\mathbf{Q}^{i-1}\mathbf{c}](j)$, are distributed independently as $\mathcal{N}\left(0, \frac{1}{n_{\mu,72\epsilon}}\right)$. We can see that both $\mathbf{Q}^{i-1}$ and $\mathbf{Q}^i$ are fixed conditioned on $y(t_1), \ldots y(t_{i-1})$ (since the i^{th} query is chosen deterministically, possibly adaptively as a function of the previously seen queries $y(t_1), \ldots y(t_{i-1})$). Additionally, since they share their first $i-1$ rows, the remaining $n_{\mu,72\epsilon} - i + 1$ rows of $\mathbf{Q}^{i-1}$ and $\mathbf{Q}^i$ have the same rowspans. Thus we can write $\mathbf{Q}^i = [\mathbf{I}; \mathbf{R}]\mathbf{Q}^{i-1}$ where $\mathbf{R} \in \mathbb{C}^{n_{\mu,72\epsilon}-i+1 \times n_{\mu,72\epsilon}-i+1}$ is some fixed rotation with $\mathbf{R}^*\mathbf{R} = \mathbf{I}$. Thus, by the rotational invariance of the Gaussian, for all $j \geq i$, $[\mathbf{Q}^i\mathbf{c}](j)$ are distributed independently as $\mathcal{N}\left(0, \frac{1}{n_{\mu,72\epsilon}}\right)$ (the same as $[\mathbf{Q}^{i-1}\mathbf{c}](j)$). Further conditioning on $y(t_i)$, which is a deterministic function of $[\mathbf{Q}^i\mathbf{c}](i)$ and $y(t_1) \ldots y(t_{i-1})$, we still have that for $j > i$, $[\mathbf{Q}^i\mathbf{c}](j)$ are distributed independently as $\mathcal{N}\left(0, \frac{1}{n_{\mu,72\epsilon}}\right)$. This completes the inductive step and so the claim. □

Armed with Claim 23 we can compute:

$$\begin{aligned}
\Pr\left[\|\mathbf{c}-\tilde{\mathbf{c}}\|_2^2 \le \frac{1}{12}\right] &= \Pr\left[\|\mathbf{Q}^r\mathbf{c}-\mathbf{Q}^r\tilde{\mathbf{c}}\|_2^2 \le \frac{1}{12}\right] && \text{(Since } \mathbf{Q}^r \text{ is orthonormal.)}\\
&\le \Pr\left[\sum_{i=r+1}^{n_{\mu,72\epsilon}} |[\mathbf{Q}^r\mathbf{c}](i)-\mathbf{Q}^r\tilde{\mathbf{c}}(i)|^2 \le \frac{1}{12}\right]\\
&= \mathbb{E}_{y(t_1),\ldots,y(t_r)}\left[\Pr\left[\sum_{i=r+1}^{n_{\mu,72\epsilon}} |[\mathbf{Q}^r\mathbf{c}](i)-\mathbf{Q}^r\tilde{\mathbf{c}}(i)|^2 \le \frac{1}{12} \mid y(t_1)...,y(t_r)\right]\right]\\
&\le \mathbb{E}_{y(t_1),\ldots,y(t_r)}\left[\Pr\left[\sum_{i=r+1}^{n_{\mu,72\epsilon}} |[\mathbf{Q}^r\mathbf{c}](i)|^2 \le \frac{1}{12} \mid y(t_1)...,y(t_r)\right]\right] && (37)
\end{aligned}$$

where the last line follows since, conditioned on $y(t_1)...,y(t_r)$, $\mathbf{Q}^r\tilde{\mathbf{c}}$ is fixed and for $i \ge r+1$, $\mathbf{Q}^r\mathbf{c}(i)$ are distributed independently as Gaussians centered around 0 (by Claim 23). So the probability of the sum of differences being small is only smaller than if we replaced each $\mathbf{Q}^r\tilde{\mathbf{c}}(i)$ by 0.

Now, conditioned on $y(t_1)...,y(t_r)$, $\sum_{i=r+1}^{n_{\mu,72\epsilon}} |[\mathbf{Q}^r\mathbf{c}](i)|^2$ is a Chi-squared random variable with

$$\mathbb{E}\left[\sum_{i=r+1}^{n_{\mu,72\epsilon}} |[\mathbf{Q}^r\mathbf{c}](i)|^2 \mid y(t_1)...,y(t_r)\right] = \frac{n_{\mu,72\epsilon}-r}{n_{\mu,72\epsilon}}.$$

For $r = \frac{n_{\mu,72\epsilon}}{20}$, we thus have $\mathbb{E}\left[\sum_{i=r+1}^{n_{\mu,72\epsilon}} |[\mathbf{Q}^r\mathbf{c}](i)|^2 \mid y(t_1)...,y(t_r)\right] \ge \frac{19}{20}$. We can loosely upper bound the probability in (37), using that for a Chi-squared random variable X with k degrees of freedom, $\Pr[X \le \delta\mathbb{E}[X]] \le (\delta e^{1-\delta})^{k/2} \le (\delta e^{1-\delta})^{1/2}$. So,

$$\Pr\left[\sum_{i=k(\mathbf{c})+1}^{n_{\mu,72\epsilon}} |[\mathbf{Q}^r\mathbf{c}](i)|^2 \le \frac{1}{12} \mid y(t_1)...,y(t_r)\right] \le \left(\frac{20}{19\cdot 12}e^{1-\frac{20}{19\cdot 12}}\right)^{1/2} < \frac{47}{100}.$$

Plugging back into (37) gives:

$$\Pr\left[\|\mathbf{c}-\tilde{\mathbf{c}}\|_2^2 \le \frac{1}{12}\right] \le \mathbb{E}_{y(t_1),\ldots,y(t_r)}\left[\Pr\left[\sum_{i=r+1}^{n_{\mu,72\epsilon}} |[\mathbf{Q}^r\mathbf{c}](i)|^2 \le \frac{1}{12} \mid y(t_1)...,y(t_r)\right]\right] < \frac{47}{100}.$$

However, we have assumed that our algorithm solves Problem 1 with probability $\ge 2/3$, and hence, by Claim 22, $\Pr\left[\|\mathbf{c}-\tilde{\mathbf{c}}\|_2^2 \le \frac{1}{12}\right] \ge \frac{1}{2}$. This is a contradiction, yielding the theorem.

□

6.1 Statistical Dimension Lower Bound

We now use Theorem 18 to prove that the statistical dimension tightly characterizes the sample complexity of solving Problem 1 for any constraint measure μ satisfying a simple condition: we must have $s_{\mu,\epsilon} = O(1/\epsilon^p)$ for some $p < 1$. Note that this assumption holds for all μ considered in this work (including bandlimited, multiband, sparse, Gaussian, and Cauchy-Lorentz), where $s_{\mu,\epsilon}$ either grows as $\log(1/\epsilon)$ or $1/\sqrt{\epsilon}$. Also note that by (5) we can always bound $s_{\mu,\epsilon} \le \mathrm{tr}(\mathcal{K}_\mu)/\epsilon = 1/\epsilon$. So this assumption holds whenever we have a nontrivial upper bound on $s_{\mu,\epsilon}$.

Theorem 24 (Statistical Dimension Lower Bound)**.** *For any probability measure μ, suppose that $s_{\mu,\epsilon} = O(1/\epsilon^p)$ for some constant $p < 1$. Consider any (possibly randomized) algorithm that solves*

Problem 1 with probability $\geq 2/3$ *for any function* y *and any* $\epsilon > 0$ *and makes at most* $r_{\mu,\epsilon}$ *(possibly adaptive) queries on any input. Then* $r_{\mu,\epsilon} = \Omega(s_{\mu,\epsilon})$.[23]

Proof. We simply prove that for this class of measures, $n_{\mu,72\epsilon} = \Omega(s_{\mu,\epsilon})$ and then apply Theorem 18. It suffices to show that $n_{\mu,\epsilon} = \Omega(s_{\mu,c\epsilon})$ for any fixed constant $c \geq 1$ since by (17), $s_{\mu,c\epsilon} \geq \frac{s_{\mu,\epsilon}}{c}$. Thus $n_{\mu,\epsilon} = \Omega(s_{\mu,c\epsilon})$ gives that $n_{\mu,72\epsilon} = \Omega(s_{\mu,72c\epsilon}) = \Omega(s_{\mu,\epsilon})$, giving the theorem.

Let $c_p = 2^{\frac{4}{1-p}} > 1$. Assume for the sake of contradiction that $n_{\mu,\epsilon} = o(s_{\mu,c_p\epsilon})$. By this assumption, there is some fixed ϵ_0 such that,

$$\text{For all } \epsilon \leq \epsilon_0\,, n_{\mu,\epsilon} \leq \frac{s_{\mu,c_p\epsilon}}{2}. \tag{38}$$

We can bound:

$$s_{\mu,c_p\epsilon} = \sum_{i=1}^{\infty} \frac{\lambda_i(\mathcal{K}_\mu)}{\lambda_i(\mathcal{K}_\mu) + c_p\epsilon} \leq n_{\mu,\epsilon} + \sum_{i=n_{\mu,\epsilon}+1}^{\infty} \frac{\lambda_i(\mathcal{K}_\mu)}{c_p\epsilon}$$

and thus by (38) have for any $\epsilon \leq \epsilon_0$:

$$\frac{1}{2} \cdot s_{\mu,c_p\epsilon} \leq \sum_{i=n_{\mu,\epsilon}+1}^{\infty} \frac{\lambda_i(\mathcal{K}_\mu)}{c_p\epsilon}. \tag{39}$$

Now we also have:

$$\begin{aligned} s_{\mu,\epsilon} = \sum_{i=1}^{\infty} \frac{\lambda_i(\mathcal{K}_\mu)}{\lambda_i(\mathcal{K}_\mu) + \epsilon} &\geq \sum_{i=n_{\mu,\epsilon}+1}^{\infty} \frac{\lambda_i(\mathcal{K}_\mu)}{\lambda_i(\mathcal{K}_\mu) + \epsilon} \\ &\geq \sum_{i=n_{\mu,\epsilon}+1}^{\infty} \frac{\lambda_i(\mathcal{K}_\mu)}{2\epsilon} \\ &= \frac{c_p}{2} \cdot \sum_{i=n_{\mu,\epsilon}+1}^{\infty} \frac{\lambda_i(\mathcal{K}_\mu)}{c_p\epsilon}. \end{aligned}$$

Combined with (39) this gives that for any $\epsilon \leq \epsilon_0$:

$$s_{\mu,\epsilon} \geq \frac{c_p}{4} \cdot s_{\mu,c_p\epsilon}. \tag{40}$$

By (40) we in turn have that, for every $\epsilon \leq \epsilon_0$,

$$s_{\mu,\epsilon} \geq s_{\mu,\epsilon_0} \cdot \left(\frac{c_p}{4}\right)^{\lfloor \log_{c_p} \epsilon_0/\epsilon \rfloor}.$$

Using that $\lfloor \log_{c_p} \epsilon_0/\epsilon \rfloor \geq \log_{c_p} \epsilon_0/\epsilon - 1$ and that $c_p = 2^{\frac{4}{1-p}} \geq 16$ we can then bound, for all $\epsilon \leq \epsilon_0$:

$$\begin{aligned} s_{\mu,\epsilon} \geq \left(\frac{c_p}{4}\right)^{\log_{c_p} \epsilon_0 - \log_{c_p} \epsilon - 1} &= \left(\frac{c_p}{4}\right)^{\log_{c_p} \epsilon_0 - 1} \cdot c_p^{\log_{c_p} 1/\epsilon} \cdot \left(\frac{1}{4}\right)^{\log_{c_p} 1/\epsilon} \\ &\geq \left(\frac{c_p}{4}\right)^{\log_{c_p} \epsilon_0 - 1} \cdot \frac{1}{\epsilon} \cdot \epsilon^{\frac{1-p}{2}} \\ &\geq \left(\frac{c_p}{4}\right)^{\log_{c_p} \epsilon_0 - 1} \cdot \frac{1}{\epsilon^{p + \frac{1-p}{2}}}. \end{aligned}$$

[23] Here we follow the Hardy-Littlewood definition [HL14], using $f(\epsilon) = \Omega(g(\epsilon))$ to denote that $\limsup_{x \to \infty} \frac{f(\epsilon)}{g(\epsilon)} > 0$. Thus the lower bound shows that, for some fixed constant $c > 0$, for every ϵ, there is at least some $\epsilon' < \epsilon$ where the number of queries used by any algorithm solving Problem 1 with probability $\geq 2/3$ is at least $c \cdot s_{\mu,\epsilon}$. In other words, the lower bound rules out the possibility that the number of queries is $o(s_{\mu,\epsilon})$.

Note that $\left(\frac{c_p}{4}\right)^{\log_{c_p} \epsilon_0 - 1}$ is a constant independent of ϵ. Thus, the above contradicts the assumption that $s_{\mu,\epsilon} = O(1/\epsilon^p)$, giving the theorem. □

Remark We remark that a similar technique to Theorem 24 can be used to show that $n_{\mu,\epsilon} = \Omega(s_{\mu,\epsilon}/\epsilon^p)$ for any $p > 0$, without any assumptions on $s_{\mu,\epsilon}$.

7 Conclusion and Open Problems

We view our work as the starting point for further exploring the application of techniques from the randomized numerical linear algebra literature (such as leverage score sampling, column based matrix reconstruction, and random projection) in signal processing. We lay out a number of open directions that we consider interesting below:

- The most immediate question is to generalize our results for interpolation over an interval to higher dimensional spaces. Fourier constrained interpolation in two or three dimensions is important in many areas, such as the earth and geosciences [Rip89, Rip05] and image processing [PPV02, RVU06]. Interpolation in even higher dimensions is common in Gaussian process methods in machine learning. We believe that our techniques should extend to higher dimensions in a similar manner to prior related work on kernel approximation [AKM+17].

- We have considered a simple signal reconstruction problem, where we wish to reconstruct a function over a fixed interval given sample access at points in that interval. There are many interesting variations of this problem. For example, can better bounds be achieved if samples can be taken from the interval $[0, T]$, but we only consider reconstruction error over a subset of this interval? In this setting, can uniform sampling give optimal bounds? How can one formulate a similar reconstruction problem and adapt our techniques to the streaming setting, where we hope to estimate a signal at any given point in time using measurements at past samples (and perhaps must limit memory/computation at any given time)? Can our techniques be extended to the setting where the error is averaged using a non-uniform measure in time domain? This question is especially relevant for applications to machine learning, where we may wish to approximate the signal well on average on input points drawn from some non-uniform distribution. In traditional supervised learning, reconstruction would be performed using points drawn from this same distribution. However, in an *active learning* setting, we may be allowed to drawn points from some other distribution, such as the leverage score distribution, which yields better error.

- In our work we have assumed knowledge of the constraint μ. However, as discussed, in the case of sparse and multiband signal reconstruction, it is important to learn μ (i.e., the locations of the frequencies or frequency bands) as part of the reconstruction process. Understanding how to do this, perhaps by combining existing techniques [ME09, Moi15, PS15, CKPS16] with our own is an important direction. More generally, in many applications, μ is derived from the signal itself, by estimating the signal's autocorrelation, which corresponds to our kernel function k_μ. Can our techniques be used to give bounds in this setting?

- Can our techniques be extended to learning signals giving constraints on other transforms such as the short-time Fourier transform (the signal's spectrogram), the wavelet transform, etc.? More generally, can leverage score sampling be used to approximate these transforms and to approximately apply filters or other signal modifications based on them?

- What is the connection between our randomized leverage score sampling method and deterministic 'sampling' methods such as Chebyshev interpolation for low-degree polynomials, uniform sampling for bandlimited signal reconstruction, and non-uniform "multicoset" sampling schemes considered in the signal processing literature [FB96, VB00, Bre08, ME09]. Can our results be made deterministic, perhaps using deterministic sampling methods for operator approximation like those employed in our proof of Lemma 46?

Acknowledgements

We thank Ron Levie for helpful discussions on weak integrals in Hilbert spaces and Zhao Song for discussions on smoothness bounds for sparse Fourier functions. We also thank Yonina Eldar for helpful discussions and pointers to related work.

Haim Avron's work is supported in part by Israel Science Foundation (grant no. 1272/17) and United States-Israel Binational Science Foundation (grant no. 2017698). Michael Kapralov is supported in part by ERC Starting Grant 759471.

References

[AA02] Y.A. Abramovich and C.D. Aliprantis. *An Invitation to Operator Theory*. Graduate studies in mathematics. American Mathematical Society, 2002.

[ACW17] Haim Avron, Kenneth L. Clarkson, and David P. Woodruff. Sharper bounds for regularized data fitting. In *Proceedings of the 21st International Workshop on Randomization and Computation (RANDOM)*, 2017.

[AKM+17] Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random Fourier features for kernel ridge regression: approximation bounds and statistical guarantees. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 253–262, 2017.

[AM15] Ahmed Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 775–783, 2015.

[Bac17] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017.

[BCDH10] Richard G. Baraniuk, Volkin Cevher, Marco F. Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

[BE12] Peter Borwein and Tamás Erdélyi. *Polynomials and polynomial inequalities*, volume 161. Springer Science & Business Media, 2012.

[BM86] Y. Bresler and A. Macovski. Exact maximum likelihood parameter estimation of superimposed exponential signals in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(5):1081–1089, 1986.

[BMD09] Christos Boutsidis, Michael W. Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of*

the 20th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 968–977, 2009.

[Bre08] Yoram Bresler. Spectrum-blind sampling and compressive sensing for continuous-index signals. In *2008 Information Theory and Applications Workshop*, pages 547–554, 2008.

[BSS14] J. Batson, D. Spielman, and N. Srivastava. Twice-Ramanujan sparsifiers. *SIAM Review*, 56(2):315–334, 2014.

[BWvOGD01] Marc Bourgeois, Frank T. A. W. Wajer, Dirk van Ormondt, and Danielle Graveron-Demilly. *Reconstruction of MRI Images from Non-Uniform Sampling and Its Application to Intrascan Motion Correction in Functional MRI*, pages 343–363. Birkhäuser Boston, 2001.

[CDL13] Albert Cohen, Mark A. Davenport, and Dany Leviatan. On the stability and accuracy of least squares approximations. *Foundations of Computational Mathematics*, 13(5):819–834, Oct 2013.

[CKPS16] Xue Chen, Daniel M. Kane, Eric Price, and Zhao Song. Fourier-sparse interpolation without a frequency gap. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 741–750, 2016.

[CLV16] Daniele Calandriello, Alessandro Lazaric, and Michal Valko. Analysis of Nyström method with sequential ridge leverage score sampling. In *Proceedings of the 32nd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 62–71, 2016.

[CM17] Albert Cohen and Giovanni Migliorati. Optimal weighted least-squares methods. *SMAI Journal of Computational Mathematics*, 3:181–203, 2017.

[CMM17] Michael B. Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1758–1777, 2017.

[CNW16] Michael B. Cohen, Jelani Nelson, and David P. Woodruff. Optimal Approximate Matrix Product in Terms of Stable Rank. In Ioannis Chatzigiannakis, Michael Mitzenmacher, Yuval Rabani, and Davide Sangiorgi, editors, *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*, volume 55 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 11:1–11:14, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[CP18] Xue Chen and Eric Price. Active regression via linear-sample sparsification active regression via linear-sample sparsification. *arXiv:1711.10051*, 2018.

[CW13] Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*, pages 81–90, 2013.

[Den11] Chun Yuan Deng. A generalization of the Sherman–Morrison–Woodbury formula. *Applied Mathematics Letters*, 24(9):1561 – 1564, 2011.

[DKM06] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal on Computing*, 36(1):132–157, 2006. Preliminary version in the 42nd Annual IEEE Symposium on Foundations of Computer Science (FOCS).

[DM16] Petros Drineas and Michael W. Mahoney. RandNLA: Randomized numerical linear algebra. *Commun. ACM*, 59(6), 2016.

[Don06] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[dP95] Gaspard Riche de Prony. Essay experimental et analytique: sur les lois de la dilatabilite de fluides elastique et sur celles de la force expansive de la vapeur de l'alcool, a differentes temperatures. *Journal de l'Ecole Polytechnique*, pages 24–76, 1795.

[DR10] Amit Deshpande and Luis Rademacher. Efficient volume sampling for row/column subset selection. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 329–338, 2010.

[DW12] Mark A. Davenport and Michael B. Wakin. Compressive sensing of analog signals using discrete prolate spheroidal sequences. *Applied and Computational Harmonic Analysis*, 33(3):438–472, 2012.

[Eld09] Yonina C. Eldar. Compressed sensing of analog signals in shift-invariant spaces. *IEEE Transactions on Signal Processing*, 57(8):2986–2997, 2009.

[Eld15] Yonina C. Eldar. *Sampling Theory: Beyond Bandlimited Systems*. Cambridge University Press, New York, NY, USA, 1st edition, 2015.

[EM09] Yonina C. Eldar and Moshe Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55(11):5302–5316, 2009.

[ESR18] Environmental Systems Research Institute ESRI. ArcGIS desktop: Release 10, 2018.

[EU06] Yonina C. Eldar and Michael Unser. Nonideal sampling and interpolation from noisy observations in shift-invariant spaces. *IEEE Transactions on Signal Processing*, 54(7):2636–2651, 2006.

[FB96] Ping Feng and Yoram Bresler. Spectrum-blind minimum-rate sampling and reconstruction of multiband signals. In *Proceedings of the 1996 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1688–1691, 1996.

[FJT94] Karl J. Friston, Peter Jezzard, and Robert Turner. Analysis of functional MRI time-series. *Human Brain Mapping*, 1(2):153–171, 1994.

[FMMS16] Roy Frostig, Cameron Musco, Christopher Musco, and Aaron Sidford. Principal component projection without principal component analysis. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 2291–2299, 2016.

[Fud89] Miha Fuderer. Ringing artefact reduction by an efficient likelihood improvement method. In *Science and Engineering of Medical Imaging*, volume 1137, pages 84–90, October 1989.

[Hel69] Gilbert Helmberg. *Introduction to spectral theory in Hilbert space*. North-Holland Pub. Co.; Wiley Amsterdam, London, New York, 1969.

[HIS15] Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt. Approximation algorithms for model-based compressive sensing. *IEEE Transactions on Information Theory*, 61(9):5129–5147, 2015. Preliminary version in the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA).

[HL14] Godfrey Harold Hardy and John Edensor Littlewood. Some problems of Diophantine approximation. *Acta mathematica*, 37(1):155–191, 1914.

[HN01] John K. Hunter and Bruno Nachtergaele. *Applied analysis*. World Scientific Publishing Company, 2001.

[HS93] Mark S. Handcock and Michael L. Stein. A Bayesian analysis of kriging. *Technometrics*, 35(4):403–410, 1993.

[HTF02] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2nd edition, 2002.

[Kot33] Vladimir A. Kotelnikov. On the carrying capacity of the ether and wire in telecommunications. *Material for the First All-Union Conference on Questions of Communication, Izd. Red. Upr. Svyazi RKKA*, 1933.

[KZW+17] Santhosh Karnik, Zhihui Zhu, Michael B. Wakin, Justin Romberg, and Mark A. Davenport. The fast Slepian transform. *Applied and Computational Harmonic Analysis*, 2017.

[Lan67a] H. J. Landau. Sampling, data transmission, and the Nyquist rate. *Proceedings of the IEEE*, 55(10):1701–1706, 1967.

[Lan67b] Henry J. Landau. Necessary density conditions for sampling and interpolation of certain entire functions. *Acta Mathematica*, 17(1):37–52, 1967.

[LH95] Alan H. Lettington and Qi He Hong. Image restoration using a Lorentzian probability model. *Journal of Modern Optics*, 42(7):1367–1376, 1995.

[LH12] Joseph D. Lakey and Jeffrey A. Hogan. On the numerical computation of certain eigenfunctions of time and multiband limiting. *Numerical Functional Analysis and Optimization*, 33(7-9):1095–1111, 2012.

[Lor83] Lee Lorch. Alternative proof of a sharpened form of Bernstein's inequality for Legendre polynomials. *Applicable Analysis*, 14(3):237–240, 1983.

[LP61] Henry J. Landau and Henry O. Pollak. Prolate spheroidal wave functions, Fourier analysis and uncertainty – II. *The Bell System Technical Journal*, 40(1):65–84, 1961.

[LP62] Henry J. Landau and Henry O. Pollak. Prolate spheroidal wave functions, Fourier analysis and uncertainty – III: The dimension of the space of essentially time- and band-limited signals. *The Bell System Technical Journal*, 41(4):1295–1336, 1962.

[LW80] Henry J. Landau and Harold Widom. Eigenvalue distribution of time and frequency limiting. *Journal of Mathematical Analysis and Applications*, 77(2):469–481, 1980.

[MC12] Scott Miller and Donald Childers. *Probability and random processes: With applications to signal processing and communications.* Academic Press, 2012.

[ME09] Moshe Mishali and Yonina C. Eldar. Blind multiband signal reconstruction: Compressed sensing for analog signals. *IEEE Transactions on Signal Processing*, 57(3):993–1009, 2009.

[ME10] Moshe Mishali and Yonina C. Eldar. From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals. *IEEE Journal of Selected Topics in Signal Processing*, 4:375–391, 2010.

[Min17] Stanislav Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics and Probability Letters*, 127:111 – 119, 2017.

[MM17] Cameron Musco and Christopher Musco. Recursive sampling for the Nyström method. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 3833–3845, 2017.

[Moi15] Ankur Moitra. Super-resolution, extremal functions and the condition number of Vandermonde matrices. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 821–830, 2015.

[MW17] Cameron Musco and David P. Woodruff. Sublinear time low-rank approximation of positive semidefinite matrices. *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 672–683, 2017.

[Nev86] Paul Nevai. Géza Freud, orthogonal polynomials and Christoffel functions. A case study. *Journal of Approximation Theory*, 48(1):3–167, 1986.

[Nyq28] Harry Nyquist. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47(2):617–644, 1928.

[Oga88] Hidemitsu Ogawa. An operator pseudo-inversion lemma. *SIAM Journal on Applied Mathematics*, 48(6):1527–1531, 1988.

[OR14] Andrei Osipov and Vladimir Rokhlin. On the evaluation of prolate spheroidal wave functions and associated quadrature rules. *Applied and Computational Harmonic Analysis*, 36(1):108–142, 2014.

[PBV18] Edouard Pauwels, Francis Bach, and Jean-Philippe Vert. Relating leverage scores and density using regularized Christoffel functions. In *Advances in Neural Information Processing Systems 31 (NIPS)*, 2018.

[Pet38] B. J. Pettis. On integration in vector spaces. *Transactions of the American Mathematical Society*, 44(2):277–304, 1938.

[Pis73] Vladilen F. Pisarenko. The retrieval of harmonics from a covariance function. *Geophysical Journal International*, 33(3):347–366, 1973.

[PPV02] Béatrices Pesquet-Popescu and Jacques L. Vehel. Stochastic fractal models for image processing. *IEEE Signal Processing Magazine*, 19(5):48–62, 2002.

[PS15] Eric Price and Zhao Song. A robust sparse Fourier transform in the continuous setting. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 583–600, 2015.

[Ras04] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

[Rip89] Brian D. Ripley. *Statistical Inference for Spatial Processes*. Cambridge University Press, 1989.

[Rip05] Brian D. Ripley. *Spatial statistics*. John Wiley & Sons, 2005.

[RvdVU05] Sathish Ramani, Dimitri van de Ville, and Michael Unser. Sampling in practice: is the best reconstruction space bandlimited? In *IEEE International Conference on Image Processing*, 2005.

[RVU06] Sathish Ramani, Dimitri Van De Ville, and Michael Unser. Non-ideal sampling and adapted reconstruction using the stochastic Matern model. In *Proceedings of the 2006 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[RW06] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[Sar06] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 143–152, 2006.

[Sch86] Ralph O. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[Sha49] Claude E. Shannon. Communication in the presence of noise. *Proceedings of the Institute of Radio Engineers*, 37(1):10–21, 1949.

[SP61] David Slepian and Henry O. Pollak. Prolate spheroidal wave functions, Fourier analysis and uncertainty – I. *The Bell System Technical Journal*, 40(1):43–63, 1961.

[SS11] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011. Preliminary version in the 40th Annual ACM Symposium on Theory of Computing (STOC).

[STC04] John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

[STR06] Yoel Shkolnisky, Mark Tygert, and Vladimir Rokhlin. Approximation of bandlimited functions. *Applied and Computational Harmonic Analysis*, 21(3):413–420, 2006.

[Tot00] Vilmos Totik. Asymptotics for Christoffel functions for general measures on the real line. *Journal d'Analyse Mathématique*, 81(1):283–303, 2000.

[VB00] Raman Venkataramani and Yoram Bresler. Perfect reconstruction formulas and bounds on aliasing error in sub-Nyquist nonuniform sampling of multiband signals. *IEEE Transactions on Information Theory*, 46(6):2173–2183, 2000.

[Wai18] Martin J Wainwright. High-dimensional statistics: A non-asymptotic viewpoint. *Cambridge Series in Statistical and Probabilistic Mathematics*, 2018.

[Whi15] Edmund T. Whittaker. On the functions which are represented by the expansions of the interpolation theory. *Proceedings of the Royal Society of Edinburgh*, 35:181–194, 1915.

[Wil12] Virginia Vassilevska Williams. Multiplying matrices faster than Coppersmith-Winograd. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC)*, pages 887–898. ACM, 2012.

[WMN+96] Keith J. Worsley, Sean Marrett, Peter Neelin, Alain C. Vandal, Karl J. Friston, and Alan C. Evans. A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4(1):58–73, 1996.

[Wou66] A. Wouk. A note on square roots of positive operators. *SIAM Review*, 8(1):100–102, 1966.

[Xia01] Hong Xiao. *Prolate spheroidal wavefunctions, quadrature, interpolation, and asymptotic formulae*. PhD thesis, Yale University, 2001.

[XRY01] Hong Xiao, Vladimir Rokhlin, and Norman Yarvin. Prolate spheroidal wavefunctions, quadrature and interpolation. *Inverse Problems*, 17(4):805–838, 2001.

[Zha05] Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.

A Prior work on Fourier constrained interpolation

As mentioned in Section 1, constrained interpolation problems similar to Problem 1 have been studied for decades in a number of different communities, often with widely varying computational models, assumptions, and goals. We discuss the most relevant prior work here.

Bandlimited functions. The most well studied special case of Problem 1 is when μ is uniform on an interval $[-F, F]$, which corresponds to reconstructing a bandlimited function from discrete samples. Work on this problem goes back to famous results of Nyquist, Shannon, and others [Whi15, Nyq28, Kot33, Sha49], who showed that it suffices to sample time uniformly with frequency $O(1/F)$. While this rate roughly suggests that $O(FT)$ samples should be required to reconstruct a signal on $[0, T]$, this does not follow directly: common reconstruction methods like Whittaker-Shannon sinc intetrpolation infer $y(t)$ from an *infinite* sum of past and future samples from y. It is possible to perform approximate reconstruction by truncating this sum, however the number of samples required to gives error ϵ will be large: $\Omega(1/\epsilon)$. See Example 25 at the end of the section.

Progress on the finite time reconstruction question beyond truncated Whittaker-Shannon began with the pioneering work Slepain, Landau, and Pollak, who study the operator $\mathcal{K}_\mu$ for uniform, bandlimited measures μ [SP61, LP61, LP62]. They bound the number of eigenvalues of $\mathcal{K}_\mu$ above ϵ, a quantity that is at most a constant factor larger than our $s_{\mu,\epsilon}$. Using this bound, it is possible to argue that Problem 1 can be solved via regression onto at most $O(s_{\mu,\epsilon})$ *prolate spheroidal wave functions (PSWFs).*

While the prolate spheroidal wave functions cannot be explicitly represented and used directly in a regression algorithm, later work presents practical methods for working with them using quadrature rules and a finite number of time samples [XRY01, STR06, OR14]. For the noiseless version of Problem 1, that work, combined with the statistical dimension bound of Landau and Widom [LW80], yields algorithms that take roughly $\tilde{O}\left(FT + \log(1/\epsilon)\right)$ samples and $\tilde{O}\left((FT + \log(1/\epsilon))^\omega\right)$ time, matching our results up to log factors.[24]

We note that existing quadrature methods access the function $f(t)$ at a pre-determined set of time domain points. Thus, they are inherently not robust to noise, since the noise function $n(t)$ of Problem 1 can place arbitrarily bad corruptions on $\mathcal{F}_\mu x$ at the pre-determined sample points. To the best of our knowledge our work is the first to solve Problem 1 for bandlimited signals in the adversarial noise setting.

Sparse functions. Signal interpolation has also been studied extensively when y is assumed to have a sparse Fourier transform: this is the basic problem of compressed sensing and sparse recovery. While most results in the compressed sensing literature are for discrete functions and address sparsity in the *discrete Fourier transform*, there has been interest in extending that work to the continuous case [DW12]. Furthermore, there are a number of results on the continuous problem that predate compressed sensing: reader's may be familiar with Prony's method [dP95], Pisarenko's method [Pis73], the matrix pencil method [BM86], or the MUSIC algorithm [Sch86].

While these methods do not provide direct guarantees for Problem 1, recently Chen, Kane, Price, and Song study a formulation of the sparse signal interpolation problem that closely matches our formulation [CKPS16]. Follow-up work in [CP18] achieves a sample complexity of $\tilde{O}(k \log^2 k)$, exactly matching our bounds. In fact, our proofs for general constraint measures rely directly on two essentially lemmas on the smoothness of Fourier-sparse functions from [CKPS16] and [CP18].

We note that most compressed sensing type results, including those of [CKPS16, CP18], are distinguished from our work in that they can also learn the support of μ – our methods assume

[24] Landau and Widom's bounds can be used to show that $s_{\mu,\epsilon} = \tilde{\Theta}(FT + \log(1/\epsilon))$, however their result only holds asymptotically as FT goes to infinity. Ours holds for all values of F, T, ϵ – see Theorem 48.

that this support is known *a priori*. We believe that our methods can be combined with existing techniques for learning the Fourier support and view this is an interesting open direction.

Multiband functions. Due to applications in radio, radar, medical imaging, and many other areas, there has been substantial interest in sample efficient algorithms for reconstructing multiband functions [Eld15]. Landau [Lan67a, Lan67b] was the first to characterize the sample complexity of reconstructing such functions in the sense of the Nyquist-Shannon sampling theorem, showing that to recover a signal with s frequency bands of widths $F_1, \ldots, F_s$, the average sampling rate must be at least $1/\sum_i F_i$. Unlike bandlimited interpolation, it is not obvious how to construct sampling schemes that achieve this optimal rate, and doing so has been the subject of a rich line of work on non-uniform "multicoset" sampling schemes [FB96, VB00, Bre08, ME09].

As in the bandlimited setting, the rate of the infinite time-horizon problem suggests, but does not imply, that the finite-time problem can be solved with roughly $\sum_i F_i T$ samples. Via a direct analysis of $\mathcal{K}_\mu$, results on prolate spheroidal wave functions can be used to upper bound the Fourier statistical dimension for a multiband support by $O\left(\sum_i F_i T + s\log(1/\epsilon)\right)$ [LW80], matching our bound in Theorem 53. However, we are unaware of existing work that solves Problem 1 with a number of samples matching this statistical dimension bound. We suspect that, as in the bandlimited case, in the noiseless setting, our methods could be matched via a combination of numerical quadrature and PSWF regression. There has been some initial work in that direction [LH12].

General constraints. Beyond the three standard settings discussed above, there has been an effort to understand the complexity of approximately reconstructing functions with more general Fourier transform constraints. In the discrete setting, *model based compressed sensing* has proven to be a powerful framework [BCDH10, HIS15]. Similar ideas have been extended to continuous functions [EM09, Eld09]. However, the constraints considered in prior work do not correspond with those captured by Problem 1. We are interested in a more refined understanding of how sample complexity depends on a the complexity of a function's representation in the Fourier basis. Model based compressed sensing focuses on functions that can be sparsely represented in other bases.

Leverage score sampling. Finally, we note that, beyond widespread applications in randomized numerical linear algebra, there has been prior work studying leverage score sampling schemes for discretizing continuous operators, which is the approach we take to solving Problem 1. See for example [CM17, Bac17, PBV18]. Our main contribution is demonstrating how to actually upper bound the leverage scores for operators of interest, which is the missing ingredient that typically prevents such sampling results from being algorithmic. We think that the tools presented in this paper offer a powerful approach to discretization in general, with significant potential for future research. We use similar methods in our recent work on randomized approximation schemes for Gaussian kernel matrices [AKM+17].

Example 25. *Truncated Whittaker-Shannon interpolation requires $\Omega(1/\epsilon)$ samples to approximate $y(t)$ with bandlimit $F = 1/2$ on $[0,1]$ up to error ϵ (i.e., to solve Problem 1, outputting $\tilde{y}(t)$ with $\|y - \tilde{y}\|_{[0,1]}^2 \leq \epsilon \|\hat{y}\|_\mu^2$ where μ is the uniform probability measure on $[-1/2, 1/2]$ and $\hat{y}$ is the Fourier transform of y with $y = \mathcal{F}_\mu^* \hat{y}$.)*

Proof. Let $\mathcal{E}$ be the set of even integers in $[\lfloor 1/2\epsilon \rfloor, \lfloor 1/\epsilon \rfloor]$. Note that $|\mathcal{E}| = \Theta(1/\epsilon)$. Let y be a sum of $\Theta(1/\epsilon)$ standard sinc functions centered at the points in $\mathcal{E}$:

$$y(t) = \sum_{k \in \mathcal{E}} y_k(t) \text{ where } y_k(t) = \frac{\sin(\pi \cdot (t-k))}{\pi \cdot (t-k)}.$$

The Fourier transform $\hat{y}_k(\xi)$ is the box on $[-1/2, 1/2]$ multiplied by $e^{-2\pi i k}$. Thus the Fourier transform $\hat{y}(\xi) = \sum_{k \in \mathcal{E}} \hat{y}_k(\xi)$ is supported on $[-1/2, 1/2]$ and so the Nyquist rate is 1 and Whittaker-

Shannon interpolation reconstructs $y(t)$ as a sum of sinc functions centered at the integers:

$$y(t) = \sum_{k=-\infty}^{\infty} y(k) \cdot \frac{\sin(\pi \cdot (t-k))}{\pi \cdot (t-k)}.$$

We can see that this reconstruction is exact since $y(k) = 0$ for all integer k except $y(k) = 1$ for $k \in \mathcal{E}$. However, if we approximate $y(t)$ on the range $[0,1]$ by truncating the Whittaker-Shannon sum to $\leq \lfloor 1/\epsilon \rfloor$ samples centered at 0, we will not include the terms corresponding to $k \in [\lfloor 1/2\epsilon \rfloor, \lfloor 1/\epsilon \rfloor] \supseteq \mathcal{E}$ and so will have $\tilde{y}(t) = 0$ and so $\|y - \tilde{y}\|^2_{[0,1]} = \|y\|^2_{[0,1]}$. Since $\mathcal{E}$ is the set of even integers in $[\lfloor 1/2\epsilon \rfloor, \lfloor 1/\epsilon \rfloor]$:

$$\begin{aligned} \|y\|^2_{[0,1]} = \int_0^1 y(t)^2 \, dt &= \int_0^1 \left(\sum_{k \in \mathcal{E}} \frac{\sin(\pi \cdot (t-k))}{\pi \cdot (t-k)} \right)^2 dt \\ &= \Omega\left(\epsilon^2\right) \int_0^1 \left(\sum_{k \in \mathcal{E}} \sin(\pi \cdot t) \right)^2 dt = \Omega\left(1\right). \end{aligned} \tag{41}$$

Finally we note that the for $j \neq k$ $\langle \hat{y}_k, \hat{y}_j \rangle_\mu = \int_{-1/2}^{1/2} e^{-2\pi i (j-k)\xi} d\xi = 0$. Thus

$$\|\hat{y}\|^2_\mu = \sum_{k \in \mathcal{E}} \|\hat{y}_k\|^2_\mu = \Theta(1/\epsilon).$$

Combined with (41) this gives:

$$\|y - \tilde{y}\|^2_{[0,1]} = \|y\|^2_{[0,1]} = \Omega\left(1\right) = \Omega(\epsilon) \cdot \|\hat{y}\|^2_\mu$$

which completes the lower bound. □

B Operator theory preliminaries

Throughout the paper, we use the term *operator* for linear transformation between two Hilbert spaces. In this section we discuss and prove basic results on operators that we use throughout the paper.

B.1 Basic definitions and the Loewner partial ordering

Consider two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ with inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}_1}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}_2}$. We denote by $\mathbb{B}(\mathcal{H}_1, \mathcal{H}_2)$ the set of bounded operators from $\mathcal{H}_1$ to $\mathcal{H}_2$, and abbreviate $\mathbb{B}(\mathcal{H})$ if $\mathcal{H}_1 = \mathcal{H}_2 = \mathcal{H}$. We denote by $\mathbb{B}_{TC}(\mathcal{H})$ and $\mathbb{B}_{HS}(\mathcal{H})$ the set of trace-class and Hilbert-Schmidt operators (respectively) on $\mathcal{H}$ (i.e. from $\mathcal{H}$ to $\mathcal{H}$). Note that $\mathbb{B}_{TC}(\mathcal{H}) \subset \mathbb{B}_{HS}(\mathcal{H}) \subset \mathbb{B}(\mathcal{H})$. Recall that for operators, boundedness is equivalent to continuity. The open mapping theorem states that if $\mathcal{A}$ is invertible, then $\mathcal{A}^{-1}$ is bounded. This implies that a compact operator is not invertible. If $\mathcal{A} \in \mathbb{B}(\mathcal{H})$ and $\mathcal{B} \in \mathbb{B}_{TC}(\mathcal{H})$ then $\mathcal{AB}, \mathcal{BA} \in \mathbb{B}_{TC}(\mathcal{H})$ and $\mathrm{tr}(\mathcal{AB}) = \mathrm{tr}(\mathcal{BA})$.

We call self-adjoint $\mathcal{A}$ *positive semidefinite* (or simply *positive*) and write $\mathcal{A} \succeq 0$ if $\langle x, \mathcal{A}x \rangle_{\mathcal{H}} \geq 0$ for all $x \in \mathcal{H}$. We write $\mathcal{A} \succnsim 0$ if $\mathcal{A}$ is *positive definite*, i.e. $\langle x, \mathcal{A}x \rangle > 0$ for all $x \in \mathcal{H}$. We denote $\mathcal{A} \succ 0$ is $\mathcal{A}$ if $\mathcal{A}$ is *strictly positive*, i.e. there exist a $c > 0$ such that $\mathcal{A} \succnsim c \cdot \mathcal{I}_{\mathcal{H}}$ where $\mathcal{I}_{\mathcal{H}}$ is the identity operator on $\mathcal{H}$. Note that for operators on finite dimensional Hilbert spaces, $\mathcal{A} \succnsim 0$ if and only if $\mathcal{A} \succ 0$, but this is not always the case for infinite dimensional Hilbert spaces. The notation for $\mathcal{A} \succeq \mathcal{B}$, $\mathcal{A} \succnsim \mathcal{B}$, and $\mathcal{A} \succ \mathcal{B}$ follow in the standard way.

If $\mathcal{A} \succeq 0$ is self-adjoint and bounded, then it possesses a unique self-adjoint bounded square root $\mathcal{A}^{1/2} \succeq 0$ [Wou66]. Furthermore, if $\mathcal{A}$ is strictly positive then so is $\mathcal{A}^{1/2}$. This implies that if $\mathcal{A}$ is strictly positive and bounded, then $\mathcal{A}$ is bounded below and that the inverse of of the square root of $\mathcal{A}$ is $\mathcal{A}^{-1/2} \stackrel{\text{def}}{=} (\mathcal{A}^{-1})^{1/2}$. Lidskii's theorem states the that trace of a trace-class operator is the sum of its eigenvalues.

Many of the following claims are well known of matrices, and the proofs in most cases, but not all, mirror the matrix case. However, for the operator case we need to be more careful with the conditions due to the aforementioned distinction between $\succsim$ and $\succ$.

Claim 26. *Suppose that $\mathcal{A}$ is a self-adjoint bounded positive semidefinite operator on an Hilbert space $\mathcal{H}$. For every $\epsilon > 0$, the operator $\mathcal{A} + \epsilon\mathcal{I}_{\mathcal{H}}$ is bounded, strictly positive and invertible, and the inverse is bounded.*

Proof. The operator $\mathcal{A} + \epsilon\mathcal{I}_{\mathcal{H}}$ is the sum of two bounded operators, and so it is bounded. It is also clearly bounded below, since $\mathcal{A} + \epsilon\mathcal{I}_{\mathcal{H}} \succeq \epsilon\mathcal{I}_{\mathcal{H}} \succ 0$. A continuous (i.e., bounded) bounded-below operator is invertible, so $\mathcal{A} + \epsilon\mathcal{I}_{\mathcal{H}}$ is invertible. The inverse is bounded due to the open mapping theorem. □

Claim 27. *Suppose that $0 \prec \mathcal{A} \preceq \mathcal{I}_{\mathcal{H}}$ for a self-adjoint operator $\mathcal{A}$. Then, $\mathcal{A}^{-1} \succeq \mathcal{I}_{\mathcal{H}}$.*

Proof. For every $x \in \mathcal{H}$ we have $\langle x, \mathcal{A}x\rangle_{\mathcal{H}} \leq \langle x, x\rangle_{\mathcal{H}}$. Given y, let $x = \mathcal{A}^{-1/2}y$. Then $\langle y, y\rangle_{\mathcal{H}} = \langle \mathcal{A}^{1/2}x, \mathcal{A}^{1/2}x\rangle_{\mathcal{H}} = \langle x, \mathcal{A}x\rangle_{\mathcal{H}} \leq \langle x, x\rangle_{\mathcal{H}} = \langle \mathcal{A}^{-1/2}y, \mathcal{A}^{-1/2}y\rangle_{\mathcal{H}} = \langle y, \mathcal{A}^{-1}y\rangle_{\mathcal{H}}$ so $\mathcal{A}^{-1} \succeq \mathcal{I}_{\mathcal{H}}$. □

Claim 28. *Suppose that $\mathcal{A} \in \mathbb{B}(\mathcal{H})$ and that $\mathcal{B} \succeq 0$ is self-adjoint trace-class operator. Then, $\mathcal{B}^{1/2}\mathcal{A}\mathcal{B}^{1/2}$ is trace-class, and $\operatorname{tr}(\mathcal{B}^{1/2}\mathcal{A}\mathcal{B}^{1/2}) = \operatorname{tr}(\mathcal{A}\mathcal{B})$.*

Proof. Since $\mathcal{B}$ is trace-class, $\mathcal{B}^{1/2} \in \mathbb{B}_{HS}(\mathcal{H})$. This implies that $\mathcal{A}\mathcal{B}^{1/2}$ is also Hilbert-Schmidt. Thus, $\mathcal{B}^{1/2}\mathcal{A}\mathcal{B}^{1/2}$ is the product of two Hilbert-Schmidt operators, so it is trace-class. The trace equality follows from the cyclic property of the trace. □

Claim 29. *Suppose that $\mathcal{A} \succ 0$ is a self-adjoint bounded operator, and that $\mathcal{B} \succeq 0$ is self-adjoint trace-class operator, both on a separable Hilbert space $\mathcal{H}$. Suppose we have $\operatorname{tr}(\mathcal{A}\mathcal{B}) \leq 1$. Then, $\mathcal{B} \preceq \mathcal{A}^{-1}$.*

Proof. Due to the cyclicity of the trace $\operatorname{tr}(\mathcal{A}^{1/2}\mathcal{B}\mathcal{A}^{1/2}) \leq 1$. The operator $\mathcal{A}^{1/2}\mathcal{B}\mathcal{A}^{1/2}$ is positive semidefinite, so due to Lidskii's theorem it's largest eigenvalue ≤ 1. For $\mathcal{A}^{1/2}\mathcal{B}\mathcal{A}^{1/2}$, the largest eigenvalue is equal to the operator norm , so for any y,

$$\langle y, \mathcal{A}^{1/2}\mathcal{B}\mathcal{A}^{1/2}y\rangle_{\mathcal{H}} \leq \langle y, y\rangle_{\mathcal{H}}.$$

Since $\mathcal{A}^{1/2}$ is invertible, with inverse $\mathcal{A}^{-1/2}$, the conclusion of the claim follows. □

Claim 30. *Let $\mathcal{A}, \mathcal{B}$ be two self-adjoint, bounded, strictly positive operators. If $\mathcal{A} \preceq \mathcal{B}$ then $\mathcal{A}^{-1} \succeq \mathcal{B}^{-1}$.*

Proof. Since $\mathcal{B}$ is bounded and strictly positive, then it is invertible and has an invertible square root. For any $y \in \mathcal{H}$ let $x = \mathcal{B}^{-1/2}y$. We have

$$\begin{aligned}
\langle y, \mathcal{B}^{-1/2}\mathcal{A}\mathcal{B}^{-1/2}y\rangle_{\mathcal{H}} &= \langle \mathcal{B}^{-1/2}y, \mathcal{A}\mathcal{B}^{-1/2}y\rangle_{\mathcal{H}} \\
&= \langle x, \mathcal{A}x\rangle_{\mathcal{H}} \\
&\leq \langle x, \mathcal{B}x\rangle_{\mathcal{H}} \\
&= \langle y, y\rangle_{\mathcal{H}}.
\end{aligned}$$

So $\mathcal{B}^{-1/2}\mathcal{A}\mathcal{B}^{-1/2} \preceq \mathcal{I}_{\mathcal{H}}$. Since both $\mathcal{A}$ and $\mathcal{B}$ are strictly positive, then $\mathcal{B}^{-1/2}\mathcal{A}\mathcal{B}^{-1/2}$ is also strictly positive. Thus, according to Claim 27, $\mathcal{B}^{1/2}\mathcal{A}^{-1}\mathcal{B}^{1/2} \succeq \mathcal{I}_{\mathcal{H}}$, from which the claim easily follows. □

Claim 31. *Suppose that $\mathcal{A} \succ 0$ and $\mathcal{A} \succeq \mathcal{B}$. Then for any $0 \leq c < 1$ we have $\mathcal{A} - c\mathcal{B} \succ 0$.*

Proof. Suppose by contradiction that $\mathcal{A} - c\mathcal{B} \not\succ 0$. Then for any $\epsilon > 0$ there exists a x with unit norm ($\langle x, x\rangle_{\mathcal{H}} = 1$) such that $\langle x, (\mathcal{A} - c\mathcal{B})x\rangle_{\mathcal{H}} \leq \epsilon$. We have $\langle x, \mathcal{B}x\rangle_{\mathcal{H}} \geq (\langle x, \mathcal{A}x\rangle_{\mathcal{H}} - \epsilon)/c$, and since $\langle x, \mathcal{A}x\rangle_{\mathcal{H}}$ is bounded away from zero and $c < 1$, for sufficiently small ϵ we have $\langle x, \mathcal{B}x\rangle_{\mathcal{H}} > \langle x, \mathcal{A}x\rangle_{\mathcal{H}}$ so $\langle x, (\mathcal{A} - \mathcal{B})x\rangle_{\mathcal{H}} < 0$ which contradicts the assumption that $\mathcal{A} \succeq \mathcal{B}$. □

Definition 5. *Given $x \in \mathcal{H}_1$ and $y \in \mathcal{H}_2$, we define the operator $x \otimes y : \mathcal{H}_2 \to \mathcal{H}_1$ by*

$$(x \otimes y)z \stackrel{\text{def}}{=} \langle y, z\rangle_{\mathcal{H}_2} x.$$

Claim 32. *Let $\mathcal{H}$ be a separable Hilbert space, and assume that $\mathcal{A} \in \mathbb{B}(\mathcal{H})$ and $v \in \mathcal{H}$. Then, $\langle v, \mathcal{A}v\rangle_{\mathcal{H}} = \operatorname{tr}(\mathcal{A}(v \otimes v))$. (We remark that $\mathcal{A}(v \otimes v)$ is trace-class since $v \otimes v$ has finite-rank and $\mathcal{A}$ is bounded.)*

Proof. Let $e_1, e_2, \ldots$ be an orthonormal basis for $\mathcal{H}$. Write $v = \sum_{i=1}^{\infty} \alpha_i e_i$. On one hand we have

$$\begin{aligned}\langle v, \mathcal{A}v\rangle_{\mathcal{H}} &= \left\langle \sum_{i=1}^{\infty} \alpha_i e_i, \mathcal{A}v \right\rangle_{\mathcal{H}} \\ &= \frac{1}{T}\sum_{i=1}^{\infty} \alpha_i^* \langle e_i, \mathcal{A}v\rangle_{\mathcal{H}}.\end{aligned}$$

On the other we have

$$\begin{aligned}\operatorname{tr}(\mathcal{A}(v \otimes v)) &= \sum_{i=1}^{\infty} \langle e_i, \mathcal{A}(v \otimes v)e_i\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^{\infty} \langle e_i, \mathcal{A}\langle v, e_i\rangle_{\mathcal{H}} v\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^{\infty} \langle v, e_i\rangle_{\mathcal{H}} \langle e_i, \mathcal{A}v\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^{\infty} \alpha_i^* \langle e_i, \mathcal{A}v\rangle_{\mathcal{H}},\end{aligned}$$

so the two terms are equal. □

B.2 Weak integrals of operators

We are going to work with operator-valued random variables. To reason about the expected value, we need a notion of an integral of operator-valued functions. We use a generalization of the concept of weak integrals (also called Pettis integral) of vector-valued functions [Pet38].

Definition 6. *Let $\mathcal{H}_1, \mathcal{H}_2$ be two separable Hilbert spaces, G a measurable space and μ a measure on G, and consider a mapping $\mathcal{A} : G \to \mathbb{B}(\mathcal{H}_1, \mathcal{H}_2)$. If the mapping $(x, z) \mapsto \int_G \langle x, \mathcal{A}(\xi)z\rangle_{\mathcal{H}_2} \, d\mu(\xi)$*

is a bounded sesquilinear map in x, z, then we say that $\mathcal{A}$ is a weakly integrable operator valued function *and the* weak operator integral *is defined to be the unique bounded operator*

$$\int_G \mathcal{A}(\xi)\, d\mu(\xi) \in \mathbb{B}(\mathcal{H}_1, \mathcal{H}_2)$$

such that for all x and z we have

$$\int_G \langle x, \mathcal{A}(\xi) z\rangle_{\mathcal{H}_2}\, d\mu(\xi) = \left\langle x, \left(\int_G \mathcal{A}(\xi)\, d\mu(\xi)\right) z\right\rangle_{\mathcal{H}_2}.$$

The existence and uniqueness of such an operator is guaranteed by the Riesz representation theorem for sesquilinear maps [Hel69, Page 92, Theorem 5].[25]

Claim 33. *Suppose that $\mathcal{A} : G \to \mathbb{B}(\mathcal{H}_1, \mathcal{H}_2)$ is weakly integrable operator valued function, and $\mathcal{S} \in \mathbb{B}(\mathcal{H}_2), \mathcal{T} \in \mathbb{B}(\mathcal{H}_1)$. Then $\xi \mapsto \mathcal{T}\mathcal{A}(\xi)\mathcal{S}$ is also a weakly integrable operator valued function and*

$$\int_G \mathcal{T}\mathcal{A}(\xi)\mathcal{S}\, d\mu(\xi) = \mathcal{T}\left(\int_G \mathcal{A}(\xi)\, d\mu(\xi)\right)\mathcal{S}.$$

Proof. Recall that $(x, z) \mapsto \int_G \langle x, \mathcal{A}(\xi) z\rangle_{\mathcal{H}_2}\, d\mu(\xi)$ is bounded, so there exists a γ such that for every $x \in \mathcal{H}_2, z \in \mathcal{H}_1$ we have

$$\left|\int_G \langle x, \mathcal{A}(\xi) z\rangle_{\mathcal{H}_2}\, d\mu(\xi)\right| \le \gamma \|x\|_{\mathcal{H}_2}\|z\|_{\mathcal{H}_1}$$

Let $x \in \mathcal{H}_2, z \in \mathcal{H}_1$. We have

$$\left|\int_G \langle x, \mathcal{T}\mathcal{A}(\xi)\mathcal{S}z\rangle_{\mathcal{H}_2} d\mu(\xi)\right| = \left|\int_G \langle \mathcal{T}^* x, \mathcal{A}(\xi)\mathcal{S}z\rangle_{\mathcal{H}_2} d\mu(\xi)\right| \le \gamma \|\mathcal{T}^* x\|_{\mathcal{H}_2}\|\mathcal{S}z\|_{\mathcal{H}_1} \le \gamma \|\mathcal{T}\|_{\mathrm{op}}\|\mathcal{S}\|_{\mathrm{op}}\|x\|_{\mathcal{H}_2}\|z\|_{\mathcal{H}_1}$$

where we used the fact that both $\mathcal{S}$ and $\mathcal{T}$ are bounded. So the mapping $(x, z) \mapsto \int_G \langle x, \mathcal{T}\mathcal{A}(\xi)\mathcal{S}z\rangle_{\mathcal{H}_2} d\mu(\xi)$ is bounded and $\xi \mapsto \mathcal{T}\mathcal{A}(\xi)\mathcal{S}$ is weakly integrable.

We now show that the value of the integral is $\mathcal{T}\left(\int_G \mathcal{A}(\xi) d\mu(\xi)\right)\mathcal{S}$. Again, for any $x \in \mathcal{H}_2, z \in \mathcal{H}_1$:

$$\left\langle x, \mathcal{T}\left(\int_G \mathcal{A}(\xi) d\mu(\xi)\right)\mathcal{S}z\right\rangle_{\mathcal{H}_2} = \left\langle \mathcal{T}^* x, \left(\int_G \mathcal{A}(\xi) d\mu(\xi)\right)\mathcal{S}z\right\rangle_{\mathcal{H}_2}$$

By definition of $\int_G \mathcal{A}(\xi) d\mu(\xi)$ we have

$$\left\langle \mathcal{T}^* x, \left(\int_G \mathcal{A}(\xi) d\mu(\xi)\right)\mathcal{S}z\right\rangle_{\mathcal{H}_2} = \int_G \langle \mathcal{T}^* x, \mathcal{A}(\xi)\mathcal{S}z\rangle_{\mathcal{H}_2} = \int_G \langle x, \mathcal{T}\mathcal{A}(\xi)\mathcal{S}z\rangle_{\mathcal{H}_2}$$

so indeed $\int_G \mathcal{T}\mathcal{A}(\xi)\mathcal{S} d\mu(\xi) = \mathcal{T}\left(\int_G \mathcal{A}(\xi) d\mu(\xi)\right)\mathcal{S}$. □

Claim 34. *Let ρ, μ be two, possibly different, probability measures, on $\mathbb{R}$, and let $\mathcal{A} \in \mathbb{B}(L_2(\rho))$ be self-adjoint and positive semi-definite, and let $\mathcal{B} \in \mathbb{B}_{TC}(L_2(\rho))$. Assume that there exists an orthonormal basis for $L_2(\rho)$ consisting of eigenvectors of $\mathcal{A}$. Given a mapping $\eta \in \mathbb{R} \mapsto v_\eta \in L_2(\rho)$ such that $\mathcal{B} = \int_{\mathbb{R}} (v_\eta \otimes v_\eta) d\mu(\eta)$ we have:*

$$\int_{\mathbb{R}} \langle v_\eta, \mathcal{A} v_\eta\rangle_\rho d\mu(\eta) = \mathrm{tr}(\mathcal{A}\mathcal{B})$$

[25] We remark that [Hel69, Page 92, Theorem 5] is stated and proved only for sesquilinear forms on the same Hilbert space (i.e., $\mathcal{H}_1 = \mathcal{H}_2$). However, it is easy to verify that the result also holds for sesquilinear forms between two Hilbert spaces.

Proof. Let $e_1, e_2, \ldots$ be an orthonormal basis for $L_2(\rho)$ consisting of eigenvectors of $\mathcal{A}$. Using Claim 32, we have

$$\begin{aligned}
\int_{\mathbb{R}} \langle v_\eta, \mathcal{A} v_\eta \rangle_\rho d\mu(\eta) &= \int_{\mathbb{R}} \mathrm{tr}(\mathcal{A}(v_\eta \otimes v_\eta)) d\mu(\eta) \\
&= \int_{\mathbb{R}} \sum_{i=1}^{\infty} \langle e_i, \mathcal{A}(v_\eta \otimes v_\eta) e_i \rangle_\rho d\mu(\eta) \\
&= \sum_{i=1}^{\infty} \int_{\mathbb{R}} \langle e_i, \mathcal{A}(v_\eta \otimes v_\eta) e_i \rangle_\rho d\mu(\eta) \\
&= \sum_{i=1}^{\infty} \langle e_i, \int_{\mathbb{R}} \mathcal{A}(v_\eta \otimes v_\eta) d\mu(\eta) e_i \rangle_\rho \\
&= \sum_{i=1}^{\infty} \langle e_i, \mathcal{A} \int_{\mathbb{R}} (v_\eta \otimes v_\eta) d\mu(\eta) e_i \rangle_\rho \\
&= \sum_{i=1}^{\infty} \langle e_i, \mathcal{A}\mathcal{B} e_i \rangle_\mu \\
&= \mathrm{tr}(\mathcal{A}\mathcal{B})
\end{aligned}$$

where the exchange of the integral and infinite sum in the third equality is justified by Tonelli's Theorem. In order to apply Tonelli's theorem we need to show that $\langle e_i, \mathcal{A}(v_\eta \otimes v_\eta) e_i \rangle_\rho \geq 0$ for every i and η. This is indeed the case since $\langle e_i, \mathcal{A}(v_\eta \otimes v_\eta) e_i \rangle_\rho = \langle \mathcal{A} e_i, (v_\eta \otimes v_\eta) e_i \rangle_\rho = \lambda_i \langle e_i, (v_\eta \otimes v_\eta) e_i \rangle_\rho \geq 0$ where λ_i is the eigenvalue corresponding to e_i. Note that since $\mathcal{A}$ is self-adjoint and positive semi-definite, λ_i is real and non-negative. We also used the immediate fact that $v_\eta \otimes v_\eta$ is positive semi-definite. □

Remark: One way to guarantee that there is an orthonormal basis of eigenvectors of $\mathcal{A}$ is to require $\mathcal{A}$ to be compact. However, it is quite possible for $\mathcal{A}$ not to be compact, and still have an orthonormal basis of eigenvectors. In fact, we primarily apply Claim 34 to operators of the form $(\mathcal{C} + \epsilon \mathcal{I})^{-1}$ where $\mathcal{C}$ is compact, and such operators have an orthonormal basis of eigenvectors (since they share eigenvectors with $\mathcal{C}$).

We say that a weakly integrable $\mathcal{A}(\cdot)$ is *self-adjoint* if $\mathcal{A}(\xi)$ is self-adjoint for all ξ. It is easy to verify that if $\mathcal{A}(\cdot)$ is self-adjoint, then $\int_G A(\xi) d\mu(\xi)$ is self-adjoint as well.

Claim 35. *Suppose that $\mathcal{A}, \mathcal{B} : G \to \mathbb{B}(\mathcal{H})$ are two self-adjoint weakly integrable operator valued functions. If, with respect to a measure μ on G, $\mathcal{A}(\xi) \preceq \mathcal{B}(\xi)$ almost everywhere, then $\int_G \mathcal{A}(\xi) d\mu(\xi) \preceq \int_G \mathcal{B}(\xi) d\mu(\xi)$.*

Proof. For every $x \in \mathcal{H}$,

$$\left\langle x, \int_G \mathcal{A}(\xi) d\mu(\xi) x \right\rangle_{\mathcal{H}} = \int_G \langle x, \mathcal{A}(\xi) x \rangle_{\mathcal{H}} d\mu(\xi) \leq \int_G \langle x, \mathcal{B}(\xi) x \rangle_{\mathcal{H}} d\mu(\xi) = \left\langle x, \int_G \mathcal{B}(\xi) d\mu(\xi) x \right\rangle_{\mathcal{H}}$$

so indeed $\int_G \mathcal{A}(\xi) d\mu(\xi) \preceq \int_G \mathcal{B}(\xi) d\mu(\xi)$. □

Claim 36. *Suppose that $\mathcal{B} : G \to \mathbb{B}(\mathcal{H})$ is a self-adjoint weakly integrable operator valued function. Consider another self-adjoint operator valued function $\mathcal{A} : G \to \mathbb{B}(\mathcal{H})$. If for every $\xi \in G$ we have $0 \preceq \mathcal{A}(\xi) \preceq \mathcal{B}(\xi)$, then $\mathcal{A}$ is weakly integrable and $\int_G \mathcal{A}(\xi) d\mu(\xi) \preceq \int_G \mathcal{B}(\xi) d\mu(\xi)$.*

Proof. We need to prove only that $\mathcal{A}$ is weakly integrable, since the integral bound follows from Claim 35. A sesquilinear form is bounded if and only if the associated quadratic form is bounded [Hel69, Page 92, Theorem 3], so we need to show that the integral of the quadratic form associated with $\mathcal{A}$ is bounded. Since $\mathcal{A}(\xi)$ is always positive semidefinite, for any x

$$\left|\int_G \langle x, \mathcal{A}(\xi)x\rangle_{\mathcal{H}} d\mu(\xi)\right| = \int_G \langle x, \mathcal{A}(\xi)x\rangle_{\mathcal{H}} d\mu(\xi) \leq \int_G \langle x, \mathcal{B}(\xi)z\rangle_{\mathcal{H}} d\mu(\xi) = \left|\int_G \langle x, \mathcal{B}(\xi)x\rangle_{\mathcal{H}} d\mu(\xi)\right|$$

and since the integral of the quadratic form associated with $\mathcal{B}$ is bounded (since $\mathcal{B}$ is weakly integrable) we conclude that integral quadratic form associated with $\mathcal{A}$ is bounded, so indeed $\mathcal{A}$ is weakly integrable. □

B.3 Concentration of random operators

Let $\mathcal{A} : G \to \mathbb{B}(\mathcal{H})$ be a weakly integrable operator valued function. If the underlying measure μ is a probability measure, then we shall call $\mathcal{A}$ a *random operator*, and write

$$\mathbb{E}(\mathcal{A}) = \int_G \mathcal{A}(\xi) d\mu(\xi).$$

Certain matrix concentration results can be generalized to the case that $\mathcal{A}$ is a random operator which takes only self-adjoint Hilbert-Schmidt values. The underlying reason is that Hilbert-Schmidt operators can be well-approximated using finite rank operators. The basic technique is outlined in [Min17, Section 3.2]. We use this technique to prove the following lemma.

Lemma 37. *Suppose that $\mathcal{H}$ is a separable Hilbert space, and that $\mathcal{B}$ is a fixed self-adjoint Hilbert-Schmidt operator on $\mathcal{H}$. Let $\mathcal{R}$ be a self-adjoint Hilber-Schmidt random operator that satisfies*

$$\mathbb{E}(\mathcal{R}) = \mathcal{B} \quad \textit{and} \quad \|\mathcal{R}\|_{\mathrm{op}} \leq L$$

Let $\mathcal{M}$ be another self-adjoint trace-class operator such that $\mathbb{E}(\mathcal{R}^2) \preceq \mathcal{M}$. Form the operator sampling estimator

$$\bar{\mathcal{R}}_n = \frac{1}{n}\sum_{k=1}^{n} \mathcal{R}_k$$

where each $\mathcal{R}_k$ is an independent copy of $\mathcal{R}$. Then, for all $t > \sqrt{\|\mathcal{M}\|_{\mathrm{op}}/n} + 2L/3n$,

$$\Pr\left(\|\bar{\mathcal{R}}_n - \mathcal{B}\|_{\mathrm{op}} > t\right) \leq \frac{8\,\mathrm{tr}(\mathcal{M})}{\|\mathcal{M}\|_{\mathrm{op}}} \exp\left(\frac{-nt^2/2}{\|\mathcal{M}\|_{\mathrm{op}} + 2Lt/3}\right). \tag{42}$$

Proof. Let $e_1, e_2, \ldots$ be the eigenvectors of $\mathcal{M}$, ordered according to the magnitude of the corresponding eigenvalue, and let $\mathcal{P}_j$ be the orthogonal projector on the span of $e_1, e_2, \ldots, e_j$. Consider the finite-rank operators $\mathcal{R}^{(j)} = \mathcal{P}_j\mathcal{R}\mathcal{P}_j$, $\mathcal{R}_k^{(j)} = \mathcal{P}_j\mathcal{R}_k\mathcal{P}_j$, $\bar{\mathcal{R}}_n^{(j)} = \mathcal{P}_j\bar{\mathcal{R}}_n\mathcal{P}_j$, $\mathcal{B}^{(j)} = \mathcal{P}_j\mathcal{B}\mathcal{P}_j$ and $\mathcal{M}^{(j)} = \mathcal{P}_j\mathcal{M}\mathcal{P}_j$. We will apply on these operator sequences the matrix version of the current lemma [AKM+17][26]

Due to linearity of weak operator integrals we have $\mathbb{E}(\mathcal{R}^{(j)}) = \mathcal{P}_j\mathcal{B}^{(j)}\mathcal{P}_j$. We can bound the operator norm of $\mathcal{R}^{(j)}$: $\|\mathcal{R}^{(j)}\|_{\mathrm{op}} \leq \|\mathcal{P}_j\mathcal{R}\mathcal{P}_j\|_{\mathrm{op}} \leq \|\mathcal{P}_j\|_{\mathrm{op}}^2\|\mathcal{R}\|_{\mathrm{op}} \leq L$ since the operator norm of a projection operator is 1. Using the fact that $\mathcal{P}_j \preceq \mathcal{I}_{\mathcal{H}}$ and so $\mathcal{R}\mathcal{P}_j\mathcal{R} \preceq \mathcal{R}^2$ we have

$$\mathbb{E}((\mathcal{R}^{(j)})^2) = \mathcal{P}_j\mathbb{E}(\mathcal{R}\mathcal{P}_j\mathcal{R})\mathcal{P}_j \preceq \mathcal{P}_j\mathbb{E}(\mathcal{R}^2)\mathcal{P}_j \preceq \mathcal{M}^{(j)}$$

[26] The lemma in [AKM+17] is stated as a bound on $\Pr\left(\|\bar{\mathcal{R}}_n - \mathcal{B}\|_{\mathrm{op}} \geq t\right)$, while for operators strict inequality is necessary. It is easy to verify that the matrix version of the Lemma continues to hold for $\Pr\left(\|\bar{\mathcal{R}}_n - \mathcal{B}\|_{\mathrm{op}} > t\right)$.

Now applying the aforementioned matrix version of the current lemma[27] we find that

$$\Pr\left(\|\bar{\mathcal{R}}_n^{(j)} - \mathcal{B}^{(j)}\|_{\text{op}} \geq t\right) \leq \frac{8\,\text{tr}(\mathcal{M}^{(j)})}{\|\mathcal{M}^{(j)}\|_{\text{op}}} \exp\left(\frac{-nt^2/2}{\|\mathcal{M}^{(j)}\|_{\text{op}} + 2Lt/3}\right). \tag{43}$$

Due to the way we constructed $\mathcal{P}_j$, and $\mathcal{M}$ being trace-class, we have $\text{tr}(\mathcal{M}^{(j)}) \to \text{tr}(\mathcal{M})$ as $j \to \infty$. Furthermore, since $\mathcal{M}$ is trace-class, $\mathcal{P}_j\mathcal{M} \to \mathcal{M}$ uniformly [HN01, Theorem 9.21], and so $\mathcal{M}^{(j)} \to \mathcal{M}$ while implies that $\|\mathcal{M}^{(j)}\|_{\text{op}} \to \|\mathcal{M}\|_{\text{op}}$. Thus, the entire right side of (43) converges to the right side of (42), so

$$\liminf_{j\to\infty} \Pr\left(\|\bar{\mathcal{R}}_n^{(j)} - \mathcal{B}^{(j)}\|_{\text{op}} > t\right) \leq \frac{8\,\text{tr}(\mathcal{M})}{\|\mathcal{M}\|_{\text{op}}} \exp\left(\frac{-nt^2/2}{\|\mathcal{M}\|_{\text{op}} + 2Lt/3}\right).$$

Let G and μ denote the underlying probability space and probability measure. Let f_j now denote the indicator function for the event $\|\bar{\mathcal{R}}_n^{(j)} - \mathcal{B}^{(j)}\|_{\text{op}} > t$, and f the indicator for the event $\|\bar{\mathcal{R}}_n - \mathcal{B}\|_{\text{op}} > t$. Again, due to the fact that $\bar{\mathcal{R}}_n - \mathcal{B}$ is Hilbert-Schmidt we have $\bar{\mathcal{R}}_n^{(j)} - \mathcal{B}^{(j)} \to \bar{\mathcal{R}}_n - \mathcal{B}$, while implies that that for any $\xi \in G$, $f(\xi) = \liminf_{j\to\infty} f_j(\xi)$. Now due to Fatou's lemma:

$$\begin{aligned}
\Pr\left(\|\bar{\mathcal{R}}_n - \mathcal{B}\|_{\text{op}} > t\right) &= \int_G f(\xi) d\mu(\xi) \\
&= \int_G \liminf_{j\to\infty} f_j(\xi) d\mu(\xi) \\
&\leq \liminf_{j\to\infty} \int_G f_j(\xi) d\mu(\xi) \\
&= \liminf_{j\to\infty} \Pr\left(\|\bar{\mathcal{R}}_n^{(j)} - \mathcal{B}^{(j)}\|_{\text{op}} > t\right) \\
&\leq \frac{8\,\text{tr}(\mathcal{M})}{\|\mathcal{M}\|_{\text{op}}} \exp\left(\frac{-nt^2/2}{\|\mathcal{M}\|_{\text{op}} + 2Lt/3}\right).
\end{aligned}$$

□

C Properties of the ridge leverage scores

C.1 Basic facts about leverage scores

In this section we prove Theorem 5, giving fundamental properties of the ridge leverage scores that we use both in bounding these scores and proving that ridge leverage score sampling can be used to solve the regularized regression problem of (10) (and hence Problem 1 by Claim 4).

Theorem 5 (Leverage Function Properties)**.** *Letting $\tau_{\mu,\epsilon}(t)$ be the ridge leverage function of Definition 3, that is*

$$\tau_{\mu,\epsilon}(t) = \frac{1}{T} \cdot \max_{\{\alpha \in L_2(\mu):\, \|\alpha\|_\mu > 0\}} \frac{|[\mathcal{F}_\mu^* \alpha](t)|^2}{\|\mathcal{F}_\mu^* \alpha\|_T^2 + \epsilon \|\alpha\|_\mu^2}, \tag{44}$$

and let $\varphi_t \in L_2(\mu)$ be defined by $\varphi_t(\xi) = e^{-2\pi i t \xi}$, we have the following basic properties:

[27] Technically, the aforementioned concentration result is for *matrices*, while here we deal with abstract operators on finite dimensional subspaces. We can address this issue by using the corresponding transformation matrices, but we find that to be tedious details.

- *The leverage scores integrate to the statistical dimension:*

$$\int_0^T \tau_{\mu,\epsilon}(t)dt = s_{\mu,\epsilon} \stackrel{\text{def}}{=} \operatorname{tr}(\mathcal{K}_\mu(\mathcal{K}_\mu + \epsilon\mathcal{I}_T)^{-1}). \tag{45}$$

- *Inner Product characterization:*

$$\tau_{\mu,\epsilon}(t) = \frac{1}{T} \cdot \langle \varphi_t, (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t \rangle_\mu. \tag{46}$$

- *Minimization Characterization:*

$$\tau_{\mu,\epsilon}(t) = \frac{1}{T} \cdot \min_{\beta \in L_2(T)} \frac{\|\mathcal{F}_\mu \beta - \varphi_t\|_\mu^2}{\epsilon} + \|\beta\|_T^2. \tag{47}$$

Proof. Recall, that given $t \in [0,T]$, we defined $\varphi_t(\xi) \stackrel{\text{def}}{=} e^{-2\pi i t\xi}$ ($\varphi_t \in L_2(\mu)$). It is easy to verify that:

$$\mathcal{G}_\mu = \frac{1}{T}\int_0^T (\varphi_t \otimes \varphi_t)dt. \tag{48}$$

To prove the equality between Equations (44), (46), and (47), we first show that the right hand side of (46) is equal to the right hand side of (47) and then show that the right hand side of (46) is equal to the right hand side of (44).

First, we need an auxiliary lemma regarding the solution of regularized least squares problems. If $\mathcal{A}$ is matrix with full column rank or a one-to-one linear operator between finite-dimensional Hilbert spaces, and b some vector, then $F(x) = \|\mathcal{A}x - b\|^2$ has a unique minimizer. In infinite dimension spaces, this remains true if only the co-domain of $\mathcal{A}$ is infinite dimensional. However, if both the domain and co-domain are infinite dimensional there might not be a minimizer even if the $\mathcal{A}$ is bounded: the range of $\mathcal{A}$ might not be closed, so it is possible that $\|\mathcal{A}x - b\| > 0$ for every x, but also that there exists a series $\{x_n\}$ such that $\|\mathcal{A}x_n - b\| \to 0$ as $n \to \infty$. However, once we introduce a ridge term (i.e., minimize $F(x) = \|\mathcal{A}x - b\|^2 + \lambda\|x\|^2$ for some $\lambda > 0$) there is always a unique minimizer (as long as $\mathcal{A}$ is bounded), due to the extreme value theorem (since we can bound the search domain). Furthermore, we can write an analytic expression for the minimizer in an analogous way to the finite dimensional case, as the following lemma shows.

Lemma 38 (Regularized Least Squares Minimizer)**.** *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two Hilbert spaces, and $\mathcal{A} : \mathcal{H}_1 \to \mathcal{H}_2$ be a bounded linear operator. Let $b \in \mathcal{H}_2$ and $\lambda > 0$. The function*

$$F(x) = \|\mathcal{A}x - b\|_{\mathcal{H}_2}^2 + \lambda\|x\|_{\mathcal{H}_1}^2$$

has a unique minimizer which is $x^\star = \mathcal{A}^*(\mathcal{A}\mathcal{A}^* + \lambda\mathcal{I}_{\mathcal{H}_2})^{-1}b$.

Proof. Consider the Hilbert space $\mathcal{H}_1 \times \mathcal{H}_2$ equipped with the inner product

$$\langle (a_1,a_2),(b_1,b_2)\rangle_{\mathcal{H}_1\times\mathcal{H}_2} \stackrel{\text{def}}{=} \langle a_1,b_1\rangle_{\mathcal{H}_1} + \langle a_2,b_2\rangle_{\mathcal{H}_2}.$$

Define the operator $\mathcal{T} : \mathcal{H}_1 \to \mathcal{H}_1 \times \mathcal{H}_2, \quad \mathcal{T}(x) = (\sqrt{\lambda}x, \mathcal{A}x)$. Let $y = (0,b) \in \mathcal{H}_1 \times \mathcal{H}_2$. We have $F(x) = \|\mathcal{T}x - y\|_{\mathcal{H}_1\times\mathcal{H}_2}^2$. Thus, we need to show that there is a unique point $\tilde{y} \in \text{range}(\mathcal{T})$ that minimizes $\|\tilde{y} - y\|_{\mathcal{H}_1\times\mathcal{H}_2}^2$ and that $\tilde{y} = \mathcal{T}x^\star$ for $x^\star = \mathcal{A}^*(\mathcal{A}\mathcal{A}^* + \lambda\mathcal{I}_{\mathcal{H}_2})^{-1}b$.

The operator $\mathcal{T}$ is a bounded linear operator, so it is continuous. We also have that for every $x \in \mathcal{H}_1$, $\|\mathcal{T}x\|_{\mathcal{H}_1\times\mathcal{H}_2}^2 \geq \lambda\|x\|_{\mathcal{H}_1}^2$ where $\lambda > 0$, so $\mathcal{T}$ is bounded from below. So $\mathcal{T}$ has a closed

range [AA02, Theorem 2.5]. Thus, there is a unique $\tilde{y} \in \text{range}(\mathcal{T})$ that minimizes $\|\tilde{y} - y\|^2_{\mathcal{H}_1 \times \mathcal{H}_2}$, and that $\tilde{y}$ is the unique element of $\text{range}(\mathcal{T})$ with the property $y - \tilde{y} \perp \text{range}(\mathcal{T})$ [HN01, Theorem 6.13]. So it suffices to show that for every $x \in \mathcal{H}_1$ we have $y - \mathcal{T}x^\star \perp \mathcal{T}x$. We compute:

$$\begin{aligned}
\langle y - \mathcal{T}x^\star, \mathcal{T}x\rangle_{\mathcal{H}_1 \times \mathcal{H}_2} &= \langle(-\sqrt{\lambda}\mathcal{A}^*(\mathcal{A}\mathcal{A}^* + \lambda\mathcal{I}_{\mathcal{H}_2})^{-1}b, b - \mathcal{A}\mathcal{A}^*(\mathcal{A}\mathcal{A}^* + \lambda\mathcal{I}_{\mathcal{H}_2})^{-1}b), (\sqrt{\lambda}x, \mathcal{A}x)\rangle_{\mathcal{H}_1 \times \mathcal{H}_2} \\
&= \langle(-\sqrt{\lambda}\mathcal{A}^*(\mathcal{A}\mathcal{A}^* + \lambda\mathcal{I}_{\mathcal{H}_2})^{-1}b, \lambda(\mathcal{A}\mathcal{A}^* + \lambda\mathcal{I}_{\mathcal{H}_2})^{-1}b), (\sqrt{\lambda}x, \mathcal{A}x)\rangle_{\mathcal{H}_1 \times \mathcal{H}_2} \\
&= -\lambda\langle\mathcal{A}^*(\mathcal{A}\mathcal{A}^* + \lambda\mathcal{I}_{\mathcal{H}_2})^{-1}b, x\rangle_{\mathcal{H}_1} + \lambda\langle(\mathcal{A}\mathcal{A}^* + \lambda\mathcal{I}_{\mathcal{H}_2})^{-1}b, \mathcal{A}x\rangle_{\mathcal{H}_2} \\
&= -\lambda\langle\mathcal{A}^*(\mathcal{A}\mathcal{A}^* + \lambda\mathcal{I}_{\mathcal{H}_2})^{-1}b, x\rangle_{\mathcal{H}_1} + \lambda\langle\mathcal{A}^*(AA^* + \lambda\mathcal{I}_{\mathcal{H}_2})^{-1}b, x\rangle_{\mathcal{H}_1} = 0.
\end{aligned}$$

So indeed, for every $x \in \mathcal{H}_1$ we have $y - \mathcal{T}x^\star \perp \mathcal{T}x$ and $x^\star$ is the unique minimizer. □

Using Lemma 38 we now proceed with the proof of Theorem 5.

Corollary 39. *Let*

$$\beta^\star = \mathcal{F}^*_\mu(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t.$$

Then,

$$\frac{1}{T} \cdot \left(\frac{\|\mathcal{F}_\mu\beta^\star - \varphi_t\|^2_\mu}{\epsilon} + \|\beta^\star\|^2_T \right) = \frac{1}{T} \cdot \min_{\beta \in L_2(T)} \frac{\|\mathcal{F}_\mu\beta - \varphi_t\|^2_\mu}{\epsilon} + \|\beta\|^2_T.$$

Claim 40. *We have*

$$\langle\varphi_t, (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t\rangle_\mu = \frac{\|\mathcal{F}_\mu\beta^\star - \varphi_t\|^2_\mu}{\epsilon} + \|\beta^\star\|^2_T$$

so the right hand side of (46) is equal to the right hand side of (47).

Proof. We compute:

$$\begin{aligned}
\|\beta^\star\|^2_T &= \langle\mathcal{F}^*_\mu(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t, \mathcal{F}^*_\mu(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t\rangle_\mu \\
&= \langle(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t, \mathcal{G}_\mu(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t\rangle_\mu \\
&= \langle(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t, (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu - \epsilon\mathcal{I}_\mu)(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t\rangle_\mu \\
&= \langle\varphi_t, (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t\rangle_\mu - \epsilon\langle\varphi_t, (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-2}\varphi_t\rangle_\mu
\end{aligned}$$

and

$$\begin{aligned}
\|\mathcal{F}_\mu\beta^\star - \varphi_t\|^2_\mu &= \|\mathcal{F}_\mu\mathcal{F}^*_\mu(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t - \varphi_t\|^2_\mu \\
&= \|\left(\mathcal{G}_\mu(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1} - \mathcal{I}_\mu\right)\varphi_t\|^2_\mu \\
&= \|\left((\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu - \epsilon\mathcal{I}_\mu)(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1} - \mathcal{I}_\mu\right)\varphi_t\|^2_\mu \\
&= \|\epsilon(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t\|^2_\mu \\
&= \epsilon^2\langle\varphi_t, (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-2}\varphi_t\rangle_\mu
\end{aligned}$$

Summing the last equalities completes the proof. □

Claim 41. *We have*

$$\langle\varphi_t, (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t\rangle_\mu = \max_{\{\alpha \in L_2(\mu):\, \|\alpha\|_\mu > 0\}} \frac{|[\mathcal{F}^*_\mu\alpha](t)|^2}{\|\mathcal{F}^*_\mu\alpha\|^2_T + \epsilon\|\alpha\|^2_\mu}$$

so the right hand side of (46) is equal to the right hand side of (44).

Proof. We can reformulate the previous claim as :

$$\begin{aligned}\langle \varphi_t, (\mathcal{G}_\mu + \epsilon \mathcal{I}_\mu)^{-1} \varphi_t \rangle_\mu = \text{minimum} \quad & \|\beta\|_\mu^2 + \|u\|_T^2 \\ \beta \in L_2(\mu); \quad & u \in L_2(T) \\ \text{subject to:} \quad & \mathcal{F}_\mu \beta + \sqrt{\epsilon} u = \varphi_t.\end{aligned}$$

Let the optimal solution be $\beta^\star$ and $u^\star$. We have $\varphi_t = \mathcal{F}_\mu \beta^\star + \sqrt{\epsilon} u^\star$, hence for any $0 \neq \alpha \in L_2(\mu)$:

$$\begin{aligned}|[\mathcal{F}_\mu^* \alpha](t)| &= |\langle \varphi_t, \alpha \rangle_\mu| \\ &= |\langle \alpha, \varphi_t \rangle_\mu| \\ &= |\langle \alpha, \mathcal{F}_\mu \beta^\star + \sqrt{\epsilon} u^\star \rangle_\mu| \\ &\leq |\langle \alpha, \mathcal{F}_\mu \beta^\star \rangle_\mu| + |\langle \alpha, \sqrt{\epsilon} u^\star \rangle_\mu| \\ &= |\langle \mathcal{F}_\mu^* \alpha, \beta^\star \rangle_T| + |\langle \alpha, \sqrt{\epsilon} u^\star \rangle_\mu| \\ &\leq \|(\mathcal{F}_\mu^* \alpha)\|_T \cdot \|\beta^\star\|_T + \sqrt{\epsilon} \|\alpha\|_\mu \cdot \|u^\star\|_\mu\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz inequality. Using Cauchy-Schwarz again:

$$\begin{aligned}|[\mathcal{F}_\mu^* \alpha](t)|^2 &\leq \left(\|(\mathcal{F}_\mu^* \alpha)\|_T \cdot \|\beta^\star\|_T + \sqrt{\epsilon} \|\alpha\|_\mu \cdot \|u^\star\|_\mu \right)^2 \\ &\leq \left(\|\mathcal{F}_\mu^* \alpha\|_T^2 + \epsilon \|\alpha\|_\mu^2 \right) \cdot \left(\|\beta^\star\|_T^2 + \|u^\star\|_\mu^2 \right)\end{aligned}$$

So for every $0 \neq \alpha \in L_2(\mu)$:

$$\frac{|[\mathcal{F}_\mu^* \alpha](t)|^2}{\|\mathcal{F}_\mu^* \alpha\|_T^2 + \epsilon \|\alpha\|_\mu^2} \leq \|\beta^\star\|_T^2 + \|u^\star\|_\mu^2 = \langle \varphi_t, (\mathcal{G}_\mu + \epsilon \mathcal{I}_\mu)^{-1} \varphi_t \rangle_\mu$$

We conclude by showing that the maximum value is attained. Let $\alpha^\star = (\mathcal{G}_\mu + \epsilon \mathcal{I}_\epsilon)^{-1} \varphi_t$. We have

$$\|\mathcal{F}_\mu^* \alpha^\star\|_T^2 + \epsilon \|\alpha^\star\|_\mu^2 = \langle \alpha^\star, (\mathcal{G}_\mu + \epsilon \mathcal{I}_\epsilon) \alpha^\star \rangle = \langle \varphi_t, (\mathcal{G}_\mu + \epsilon \mathcal{I}_\epsilon)^{-1} \varphi_t \rangle_\mu$$

and finally,

$$\frac{|[\mathcal{F}_\mu^* \alpha^\star](t)|^2}{\|\mathcal{F}_\mu^* \alpha^\star\|_T^2 + \epsilon \|\alpha\|_\mu^2} = \frac{|\langle \varphi_t, \alpha^\star \rangle_\mu|^2}{\langle \varphi_t, (\mathcal{G}_\mu + \epsilon \mathcal{I}_\epsilon)^{-1} \varphi_t \rangle_\mu} = \langle \varphi_t, (\mathcal{G}_\mu + \epsilon \mathcal{I}_\mu)^{-1} \varphi_t \rangle_\mu.$$

□

We now turn to showing that the leverage function integrates to the statistical dimension.

Claim 42.

$$\int_0^T \tau_{\mu,\epsilon}(t) dt = s_{\mu,\epsilon}.$$

Proof. It follows from Eq. (48) and Claim 34 that $\int_0^T \tau_{\mu,\epsilon}(t) = \operatorname{tr}((\mathcal{G}_\mu + \epsilon \mathcal{I}_\mu)^{-1} \mathcal{G}_\mu)$. The claim follows by noting that $\mathcal{K}_\mu$ and $\mathcal{G}_\mu$ have the same eigenvalues (both operators are compact self adjoint operators, so their spectrum consists of only eigenvalues, and it is easy to verify that if x is an eigenvector $\mathcal{K}_\mu$ then $\mathcal{F}_\mu x$ is eigenvector of $\mathcal{G}_\mu$).

□

We thus have completed the proof of Theorem 5. □

C.2 Operator Approximation via Leverage Score Sampling

Analogs of the following concentration result are well known for matrices. Accordingly, the proof is an adaptation of standard proofs for finite matrix approximation by leverage score sampling, where matrix concentration results are replaced with operator concentration results. A similar strategy was employed in [Bac17].

Lemma 43. *Consider the conditions of Theorem 6, and denote* $\widehat{\mathcal{G}}_\mu = \mathbf{F}\mathbf{F}^*$. *Let* $\Delta \le 1/2$ *and* $\epsilon \le \|\mathcal{G}_\mu\|_{\mathrm{op}}$. *If* $s \ge \frac{8}{3}\Delta^{-2}\tilde{s}_{\mu,\epsilon}\ln(16\tilde{s}^2_{\mu,\epsilon}/\delta)$, *then*

$$(1-\Delta)(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu) \preceq \widehat{\mathcal{G}}_\mu + \epsilon\mathcal{I}_\mu \preceq (1+\Delta)(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu) \tag{49}$$

with probability of at least $1-\delta$.

Proof. The condition (49) is equivalent to

$$\mathcal{G}_\mu - \Delta(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu) \preceq \widehat{\mathcal{G}}_\mu \preceq \mathcal{G}_\mu + \Delta(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)$$

By composing with $(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}$ on the left and right, we find that the condition is equivalent to

$$\|(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}(\widehat{\mathcal{G}}_\mu - \mathcal{G}_\mu)(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}\|_{\mathrm{op}} \le \Delta. \tag{50}$$

Noticing that

$$\mathbf{F}g = \sum_{j=1}^{s} w_j g(j)\varphi_{t_j}$$

and that

$$[\mathbf{F}^* g](j) = w_j\langle\varphi_{t_j}, g\rangle_\mu$$

we understand that $\widehat{\mathcal{G}}_\mu = \sum_{j=1}^s w_j^2(\varphi_{t_j}\otimes\varphi_{t_j})$. Let

$$\mathcal{X}_j \stackrel{\mathrm{def}}{=} s w_j^2(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}(\varphi_{t_j}\otimes\varphi_{t_j})(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}.$$

Note that $\mathcal{X}_j$ is self-adjoint and Hilbert-Schmidt (since it has finite rank). We have

$$(\mathcal{G}_\mu + \mathcal{I}_\mu)^{-1/2}\widehat{\mathcal{G}}_\mu(\mathcal{G}_\mu + \mathcal{I}_\mu)^{-1/2} = \frac{1}{s}\sum_{j=1}^{s}\mathcal{X}_j.$$

Since the time samples are drawn randomly, $\mathcal{X}_1,\ldots,\mathcal{X}_s$ are random operators. We also have, using Claim 33,

$$\mathbb{E}_{t_j\propto\tilde{\tau}_{\mu,\epsilon}}[\mathcal{X}_j] = (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}\mathbb{E}_{t_j\propto\tilde{\tau}_{\mu,\epsilon}}\left[s w_j^2(\varphi_{t_j}\otimes\varphi_{t_j})\right](\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}.$$

Write $w(t) = \sqrt{\frac{\tilde{s}_{\mu,\epsilon}}{T\cdot\tilde{\tau}_{\mu,\epsilon}(t)}}$. For every $x, z \in L_2(\mu)$,

$$\int_0^T \left\langle x, w(t)^2\cdot(\varphi_t\otimes\varphi_t)z\right\rangle_\mu \cdot(\tilde{\tau}_{\mu,\epsilon}(t)/\tilde{s}_{\mu,\epsilon})dt = \frac{1}{T}\int_0^T\langle x,(\varphi_t\otimes\varphi_t)z\rangle_\mu\,dt = \langle x,\mathcal{G}_\mu z\rangle_\mu$$

which shows that

$$\mathbb{E}_{t_j\propto\tilde{\tau}_{\mu,\epsilon}}\left[s w_j^2(\varphi_{t_j}\otimes\varphi_{t_j})\right] = \mathcal{G}_\mu$$

so,

$$\mathbb{E}_{t_j \propto \tilde{\tau}_{\mu,\epsilon}}[\mathcal{X}_j] = (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}\mathcal{G}_\mu(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}. \tag{51}$$

Next, we bound the operator norm of $\mathcal{X}_j$. The random operator only takes values that are both positive semidefinite and rank one, so the operator norm of $\mathcal{X}_j$ is equal to the trace of the operator. Thus, we have

$$\begin{aligned}
\|\mathcal{X}_j\|_{\text{op}} &= s w_j^2 \operatorname{tr}\left((\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}(\varphi_{t_j} \otimes \varphi_{t_j})(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}\right) \\
&= \frac{\tilde{s}_{\mu,\epsilon}}{\tilde{\tau}_{\mu,\epsilon}(t_j)} \cdot \frac{1}{T}\operatorname{tr}\left((\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}(\varphi_{t_j} \otimes \varphi_{t_j})\right) \\
&= \frac{\tilde{s}_{\mu,\epsilon}}{\tilde{\tau}_{\mu,\epsilon}(t_j)} \cdot \tau_{\mu,\epsilon}(t_j) \qquad \text{(via Theorem 5, equation (46).)} \\
&\le \tilde{s}_{\mu,\epsilon}
\end{aligned}$$

where the last line follows since $\tilde{\tau}_{\mu,\epsilon}(t_j) \ge \tau_{\mu,\epsilon}(t_j)$ by assumption.

The final ingredient for applying Lemma 37 is to bound $\mathcal{X}_j^2$. Again, using the fact that $\tilde{\tau}_{\mu,\epsilon}(t_j) \ge \tau_{\mu,\epsilon}(t_j)$ we have:

$$\begin{aligned}
\mathcal{X}_j^2 =& \frac{\tilde{s}_{\mu,\epsilon}^2}{T^2 \cdot \tilde{\tau}_{\mu,\epsilon}(t_j)^2}(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}(\varphi_{t_j} \otimes \varphi_{t_j})(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}(\varphi_{t_j} \otimes \varphi_{t_j})(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2} \\
=& \frac{\tilde{s}_{\mu,\epsilon}^2 \cdot \langle \varphi_{t_j}, (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_{t_j}\rangle_\mu}{T^2 \cdot \tilde{\tau}_{\mu,\epsilon}(t_j)^2}(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}(\varphi_{t_j} \otimes \varphi_{t_j})(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2} \\
=& \frac{\tilde{s}_{\mu,\epsilon}^2 \cdot \tau_{\mu,\epsilon}(t_j)}{T \cdot \tilde{\tau}_{\mu,\epsilon}(t_j)^2}(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}(\varphi_{t_j} \otimes \varphi_{t_j})(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2} \\
=& \frac{\tilde{s}_{\mu,\epsilon} \cdot \tau_{\mu,\epsilon}(t_j)}{\tilde{\tau}_{\mu,\epsilon}(t_j)}\mathcal{X}_j \preceq \tilde{s}_{\mu,\epsilon}\mathcal{X}_j.
\end{aligned}$$

So, using Claim 36,

$$\mathbb{E}_{t_j \propto \tilde{\tau}_{\mu,\epsilon}}[\mathcal{X}_j^2] \preceq \mathbb{E}_{t_j \propto \tilde{\tau}_{\mu,\epsilon}}[\tilde{s}_{\mu,\epsilon}\mathcal{X}_j] = \tilde{s}_{\mu,\epsilon}(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}\mathcal{G}_\mu(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2} \stackrel{\text{def}}{=} \mathcal{M}.$$

Noticing that $\operatorname{tr}(\mathcal{M}) = \tilde{s}_{\mu,\epsilon} \cdot s_{\mu,\epsilon}$ and $\|\mathcal{M}\|_{\text{op}} = \frac{\|\mathcal{G}_\mu\|_{\text{op}}}{\|\mathcal{G}_\mu\|_{\text{op}}+\epsilon} \ge 1/2$ by our assumption that $\epsilon \le \|\mathcal{G}_\mu\|_{\text{op}}$, and Lemma 37 we have:

$$\begin{aligned}
\Pr\left(\|(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}(\widehat{\mathcal{G}}_\mu - \mathcal{G}_\mu)(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1/2}\|_{\text{op}} \ge \Delta\right) \le& \frac{8\operatorname{tr}(\mathcal{M})}{\|\mathcal{M}\|_{\text{op}}}\exp\left(\frac{-s\Delta^2/2}{\|\mathcal{M}\|_{\text{op}} + 2\tilde{s}_{\mu,\epsilon}\Delta/3}\right) \\
\le& 16\tilde{s}_{\mu,\epsilon} \cdot s_{\mu,\epsilon} \cdot \exp\left(\frac{-s\Delta^2}{2\tilde{s}_{\mu,\epsilon}(1 + 2\Delta/3)}\right) \\
\le& 16\tilde{s}_{\mu,\epsilon}^2 \cdot \exp\left(\frac{-3s\Delta^2}{8\tilde{s}_{\mu,\epsilon}}\right) \le \delta.
\end{aligned}$$

□

C.3 Approximate Discretization via Leverage Score Sampling

With the operator approximation bound of Lemma 43 in place, we can prove Theorem 6, which shows that we can approximately solve the regression problem of (10) (and by Claim 4 solve Problem 1) by sampling time domain points via over-approximations to their ridge leverage scores.

Theorem 6 (Approximate Regression via Leverage Score Sampling). *Assume $\epsilon \leq \|\mathcal{K}_\mu\|_{\text{op}}$ and consider a measurable $\tilde{\tau}_{\mu,\epsilon}(t)$ with $\tilde{\tau}_{\mu,\epsilon}(t) \geq \tau_{\mu,\epsilon}(t)$ for all t and let $\tilde{s}_{\mu,\epsilon} = \int_0^T \tilde{\tau}_{\mu,\epsilon}(t)dt$. Let $s = c \cdot (\tilde{s}_{\mu,\epsilon} \cdot [\log(\tilde{s}_{\mu,\epsilon}) + 1/\delta])$ for sufficiently large fixed constant c and let $t_1, \ldots, t_s$ be time points selected by drawing each randomly from $[0,T]$ with probability proportional to $\tilde{\tau}_{\mu,\epsilon}(t)$. For $j \in 1, \ldots, s$ let $w_j = \sqrt{\frac{1}{sT} \cdot \frac{\tilde{s}_{\mu,\epsilon}}{\tilde{\tau}_{\mu,\epsilon}(t_j)}}$. Let $\mathbf{F} : \mathbb{C}^s \to L_2(\mu)$ be the operator defined by:*

$$[\mathbf{F}\,\mathbf{x}](\xi) = \sum_{j=1}^{s} w_j \cdot \mathbf{x}(j) \cdot e^{-2\pi i \xi t_j} \tag{52}$$

and $\mathbf{y}, \mathbf{n} \in \mathbb{R}^s$ be the vector with $\mathbf{y}(j) = w_j \cdot y(t_j)$ and $\mathbf{n}(j) = w_j \cdot n(t_j)$. With probability $\geq 1 - \delta$:

- *For any $\beta \geq 0$, if $\tilde{g} \in L_2(\mu)$ satisfies* [28]

$$\|\mathbf{F}^*\tilde{g} - (\mathbf{y} + \mathbf{n})\|_2^2 + \epsilon\|\tilde{g}\|_\mu^2 \leq (1 + \delta\beta) \cdot \min_{g \in L_2(\mu)} \left[\|\mathbf{F}^* g - (\mathbf{y} + \mathbf{n})\|_2^2 + \epsilon\|g\|_\mu^2\right], \tag{53}$$

  then

$$\|\mathcal{F}_\mu^*\tilde{g} - (y + n)\|_T^2 + \epsilon\|\tilde{g}\|_\mu^2 \leq 3(1 + 2\beta) \cdot \min_{g \in L_2(\mu)} \left[\|\mathcal{F}_\mu^* g - (y + n)\|_T^2 + \epsilon\|g\|_\mu^2\right]. \tag{54}$$

So $\tilde{g}$ provides an approximate solution to (10) and by Claim 4, $\tilde{y} = \mathcal{F}_\mu^*\tilde{g}$ solves Problem 1 with parameter $\Theta(\epsilon)$.

Proof. Throughout the proof we will let $\bar{y} = y + n$ and $\bar{\mathbf{y}} = \mathbf{y} + \mathbf{n}$. Let

$$g^\star \stackrel{\text{def}}{=} \arg\min_{g \in L_2(\mu)} \left[\|\mathcal{F}_\mu^* g - \bar{y}\|_T^2 + \epsilon\|g\|_\mu^2\right].$$

By Lemma 38, $g^\star = \mathcal{F}_\mu(\mathcal{K}_\mu + \lambda\mathcal{I}_T)^{-1}\bar{y}$. Denote the optimal error as $b^\star \stackrel{\text{def}}{=} \mathcal{F}_\mu^* g^\star - \bar{y}$ and the optimal cost as $B^\star \stackrel{\text{def}}{=} \|\mathcal{F}_\mu^* g^\star - \bar{y}\|_T^2 + \epsilon\|g^\star\|_\mu^2$.

Reduction to Affine Embedding

We prove that, for all $g \in L_2(\mu)$, ridge leverage score sampling lets us approximate the value of the objective function of (10) when evaluated at g. In the randomized linear algebra literature, this is known as an *affine embedding guarantee*. Specifically, we show that, with probability $\geq 1 - \delta$, for all $g \in L_2(\mu)$,

$$\frac{1}{2}\left(\|\mathcal{F}_\mu^* g - \bar{y}\|_T^2 + \epsilon\|g\|_\mu^2\right) \leq \|\mathbf{F}^* g - \bar{\mathbf{y}}\|_2^2 + \epsilon\|g\|_\mu^2 + \alpha \leq \frac{3}{2}\left(\|\mathcal{F}_\mu^* g - \bar{y}\|_T^2 + \epsilon\|g\|_\mu^2\right) \tag{55}$$

where α is some fixed value independent of g (but which depends on $\mathbf{F}$ and $\bar{\mathbf{y}}$) with $|\alpha| \leq \frac{1}{\delta} \cdot B^\star$.

It is not hard to see that (55) gives the theorem. For any $\tilde{g} \in L_2(\mu)$ satisfying:

$$\|\mathbf{F}^*\tilde{g} - \bar{\mathbf{y}}\|_2^2 + \epsilon\|\tilde{g}\|_\mu^2 \leq (1 + \delta C) \cdot \min_{g \in L_2(\mu)} \left[\|\mathbf{F}^* g - \bar{\mathbf{y}}\|_2^2 + \epsilon\|g\|_\mu^2\right], \tag{56}$$

[28] We can see that the adjoint $\mathbf{F}^* : L_2(\mu) \to \mathbb{C}^s$ is given by $[\mathbf{F}^*\,g](j) = w_j \cdot \int_{\mathbb{R}} g(\xi) e^{2\pi i \xi t_j}\, d\mu(\xi)$.

we can apply (55) to give the main claim of the theorem:

$$\begin{aligned}
\|\mathcal{F}_\mu^*\tilde{g} - \bar{y}\|_T^2 + \epsilon\|\tilde{g}\|_\mu^2 &\le 2\left(\|\mathbf{F}^*\tilde{g} - \bar{\mathbf{y}}\|_2^2 + \epsilon\|\tilde{g}\|_\mu^2 + \alpha\right) && \text{(applying lower bound of (55))}\\
&\le 2(1+\delta C)\cdot \min_{g\in L_2(\mu)} \left(\|\mathbf{F}^*g - \bar{\mathbf{y}}\|_2^2 + \epsilon\|g\|_\mu^2\right) + 2\alpha && \text{(by assumption of (56))}\\
&\le 2(1+\delta C)\cdot \left(\|\mathbf{F}^*g^\star - \bar{\mathbf{y}}\|_2^2 + \epsilon\|g^\star\|_\mu^2\right) + 2\alpha\\
&= 2(1+\delta C)\cdot \left(\|\mathbf{F}^*g^\star - \bar{\mathbf{y}}\|_2^2 + \epsilon\|g^\star\|_\mu^2 + \alpha\right) - 2\delta C\alpha\\
&\le 3(1+\delta C)\cdot \left(\|\mathcal{F}_\mu^*g^\star - \bar{y}\|_F^2 + \epsilon\|g^\star\|_\mu^2\right) - 2\delta C\alpha && \text{(upper bound of (55))}\\
&\le [3(1+\delta C) + 2C]\cdot \left(\|\mathcal{F}_\mu^*g^\star - \bar{y}\|_F^2 + \epsilon\|g^\star\|_\mu^2\right) && \text{(since } |\alpha| \le \tfrac{B^\star}{\delta}\text{)}\\
&\le 3(1+2C)\cdot \min_{g\in L_2(\mu)} \left[\|\mathcal{F}_\mu^*g - \bar{y}\|_T^2 + \epsilon\|g\|_\mu^2\right]. && \text{(definition of } g^\star \text{ as optimum)}
\end{aligned}$$

Thus, we focus our attention to proving that the affine embedding guarantee of (55) holds with probability $\ge 1-\delta$.

Expression of Error in Terms of $g - g^\star$

We begin by showing how, for any $g \in L_2(\mu)$, the cost $\|\mathcal{F}_\mu^*g - \bar{y}\|_T^2 + \epsilon\|g\|_\mu^2$ can be written as a function of the deviation from the optimum: $g - g^\star$.

Claim 44 (Expression for Excess Cost)**.** *For any $g \in L_2(\mu)$:*

$$\|\mathcal{F}_\mu^*g - \bar{y}\|_T^2 + \epsilon\|g\|_\mu^2 = \|\mathcal{F}_\mu^*(g-g^\star)\|_T^2 + \epsilon\|g-g^\star\|_\mu^2 + B^\star,$$

recalling that $B^\star \stackrel{\text{def}}{=} \|\mathcal{F}_\mu^*g^\star - \bar{y}\|_T^2 + \epsilon\|g^\star\|_\mu^2$ is the optimum cost of the ridge regression problem.

Proof. Following Lemma 38 we define $\mathcal{T} : L_2(\mu) \to L_2(\mu)\times L_2(T)$, $\mathcal{T}g = (\sqrt{\epsilon}g, \mathcal{F}_\mu^*g)$. For any g, $\|\mathcal{F}_\mu^*g - \bar{y}\|_T^2 + \epsilon\|g\|_\mu^2 = \|\mathcal{T}g - (0,\bar{y})\|_{L_2(\mu)\times L_2(T)}^2$. Again, as in Lemma 38 we know $g^\star$ is the unique minimizer of this function with the property that $(0,y) - \mathcal{T}g^\star \perp \text{range}(\mathcal{T})$ [HN01, Theorem 6.13]. We can thus decompose:

$$\begin{aligned}
\|\mathcal{F}_\mu^*g - \bar{y}\|_T^2 + \epsilon\|g\|_\mu^2 &= \|\mathcal{T}g - (0,\bar{y})\|_{L_2(\mu)\times L_2(T)}^2\\
&= \|\mathcal{T}g^\star - (0,\bar{y}) + (\mathcal{T}g - \mathcal{T}g^\star)\|_{L_2(\mu)\times L_2(T)}^2\\
&= \|\mathcal{T}g^\star - (0,\bar{y})\|_{L_2(\mu)\times L_2(T)}^2 + \|\mathcal{T}(g-g^\star)\|_{L_2(\mu)\times L_2(T)}^2\\
&= B^\star + \|\mathcal{F}_\mu^*(g-g^\star)\|_T^2 + \epsilon\|g-g^\star\|_\mu^2
\end{aligned}$$

which gives the claim. □

Bounding The Sampling Error

We now show that Claim 44 holds approximately, even after sampling. This almost immediately yields the affine embedding bound of (55).

Let $\tilde{B} \stackrel{\text{def}}{=} \|\mathbf{F}^*g^\star - \bar{\mathbf{y}}\|_2^2 + \epsilon\|g^\star\|_\mu^2$ be the error of the optimal solution in our randomly discretized regression problem. We can write the discretized objective function value for any $g \in L_2(\mu)$ as:

$$\begin{aligned}
\|\mathbf{F}^*g - \bar{\mathbf{y}}\|_2^2 + \epsilon\|g\|_\mu^2 &= \|\mathbf{F}^*(g-g^\star) + \mathbf{F}^*g^\star - \bar{\mathbf{y}}\|_2^2 + \epsilon\|g^\star + (g-g^\star)\|_\mu^2\\
&= \tilde{B} + \|\mathbf{F}^*(g-g^\star)\|_2^2 + \epsilon\|g-g^\star\|_\mu^2 + 2\Re(\langle \mathbf{F}^*(g-g^\star), \mathbf{F}^*g^\star - \bar{\mathbf{y}}\rangle) + 2\epsilon\Re(\langle (g-g^\star), g^\star\rangle_\mu).
\end{aligned} \tag{57}$$

Let $\bar{\mathcal{F}}_\mu : L_2(T) \times L_2(\mu) \to L_2(\mu)$ be the operator $\bar{\mathcal{F}}_\mu(f,g) = \mathcal{F}_\mu f + \sqrt{\epsilon}\cdot g$. We can see that $\bar{\mathcal{F}}_\mu^* : L_2(\mu) \to L_2(T)\times L_2(\mu)$ is given by $\bar{\mathcal{F}}_\mu^* g = (\mathcal{F}_\mu^* g, \sqrt{\epsilon}\cdot g)$. Further, we see that $\bar{\mathcal{F}}_\mu \bar{\mathcal{F}}_\mu^* = \mathcal{G}_\mu + \epsilon\mathcal{I}_\mu$. We can write:

$$\bar{\mathcal{F}}_\mu^* = \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu) = \bar{\mathcal{P}}_\mu \bar{\mathcal{F}}_\mu^*$$

where $\bar{\mathcal{P}}_\mu = \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\bar{\mathcal{F}}_\mu$. Note that $\bar{\mathcal{P}}_\mu$ is self adjoint. Correspondingly, let $\bar{\mathbf{F}} : \mathbb{C}^s \times L_2(\mu) \to L_2(\mu)$ be given by $\bar{\mathbf{F}}(f,g) = \mathbf{F}f + \sqrt{\epsilon}\cdot g$. We have $\bar{\mathbf{F}}^* g = (\mathbf{F}^* g, \sqrt{\epsilon}\cdot g)$. We can also write $\bar{\mathbf{P}} = \bar{\mathbf{F}}^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\bar{\mathcal{F}}_\mu$, and observe that $\bar{\mathbf{F}}^* = \bar{\mathbf{P}}\bar{\mathcal{F}}_\mu^*$.

With this notation in place we can rewrite the last term of (57) as:

$$\begin{aligned}
\langle \mathbf{F}^*(g-g^\star), \mathbf{F}^* g^\star - \bar{\mathbf{y}}\rangle + \epsilon\langle (g-g^\star), g^\star\rangle_\mu &= \langle \bar{\mathbf{F}}^*(g-g^\star), (\mathbf{F}^* g^\star - \bar{\mathbf{y}}, \sqrt{\epsilon}g^\star)\rangle_{\mathbb{C}^s\times L_2(\mu)} \\
&= \langle \bar{\mathbf{P}}\bar{\mathcal{F}}_\mu^*(g-g^\star), (\mathbf{F}^* g^\star - \bar{\mathbf{y}}, \sqrt{\epsilon}g^\star)\rangle_{\mathbb{C}^s\times L_2(\mu)} \\
&= \langle \bar{\mathcal{F}}_\mu^*(g-g^\star), \bar{\mathbf{P}}^*(\mathbf{F}^* g^\star - \bar{\mathbf{y}}, \sqrt{\epsilon}g^\star)\rangle_{L_2(T)\times L_2(\mu)}. \qquad (58)
\end{aligned}$$

Using the fact that $\Re(z) \le |z|$ for all $z \in \mathbb{C}$, and applying Cauchy-Schwarz to (58) and plugging back into (57) we have:

$$\begin{aligned}
\|\mathbf{F}^* g - \bar{\mathbf{y}}\|_2^2 + \epsilon\|g\|_\mu^2 \in \tilde{B} &+ \|\mathbf{F}^*(g-g^\star)\|_2^2 + \epsilon\|g-g^\star\|_\mu^2 \\
&\pm 2(\|\mathcal{F}_\mu^*(g-g^\star)\|_T + \epsilon\|g-g^\star\|_\mu)\cdot\|\bar{\mathbf{P}}^*(\mathbf{F}^* g^\star - \bar{\mathbf{y}}, \sqrt{\epsilon}g^\star)\|_{L_2(T)\times L_2(\mu)}. \qquad (59)
\end{aligned}$$

We now bound $\|\bar{\mathbf{P}}^*(\mathbf{F}^* g^\star - \bar{\mathbf{y}}, \sqrt{\epsilon}g^\star)\|_{L_2(T)\times L_2(\mu)}$. *If we had not sampled*, this would equal:

$$\|\bar{\mathcal{P}}_\mu(\mathcal{F}_\mu^* g^\star - \bar{y}, \sqrt{\epsilon}g^\star)\|_{L_2(T)\times L_2(\mu)} = \|\bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\bar{\mathcal{F}}_\mu\left[\bar{\mathcal{F}}_\mu^* g^\star - (\bar{y},0)\right]\|_{L_2(T)\times L_2(\mu)} = 0 \qquad (60)$$

since $g^\star$ is the optimum of $\|\bar{\mathcal{F}}_\mu^* g - (\bar{y},0)\|_{L_2(T)\times L_2(\mu)}$ and thus $\bar{\mathcal{F}}_\mu^* g^\star - (\bar{y},0)$ is orthogonal to range$(\bar{\mathcal{F}}_\mu^*)$. We will show that after sampling, while the norm no longer equals 0, it is still small. The bound we give is analogous to standard approximate matrix multiplication results for finite dimensional matrices. Specifically, our proof follows that of Lemma 4 in [DKM06].

Claim 45 (Approximate Operator Application)**.** *With probability* $\ge 1-\delta$*:*

$$\|\bar{\mathbf{P}}^*(\mathbf{F}^* g^\star - \bar{\mathbf{y}}, \sqrt{\epsilon}g^\star)\|_{L_2(T)\times L_2(\mu)} \le \frac{1}{64}\cdot B^\star.$$

Proof. For conciseness let $\mathcal{H}$ denote the space $L_2(T)\times L_2(\mu)$. Let $\varphi_t \in L_2(\mu)$ be given by $\varphi_t(\xi) = e^{-2\pi i t\xi}$. Let $b^\star \stackrel{\text{def}}{=} \mathcal{F}_\mu^* g^\star - \bar{y}$ and $\mathbf{b}^\star \in \mathbb{C}^s$ be given by $\mathbf{b}^\star \stackrel{\text{def}}{=} \mathbf{F}^* g^\star - \bar{\mathbf{y}}$. We can see that $\mathbf{b}^\star(j) = w_j\cdot[\langle\varphi_{t_j}, g^\star\rangle_\mu - \bar{y}(t_j)]$. We have:

$$\begin{aligned}
\mathbb{E}\left[\|\bar{\mathbf{P}}^*(\mathbf{F}^* g^\star - \bar{\mathbf{y}}, \sqrt{\epsilon}g^\star)\|_{\mathcal{H}}^2\right] &= \mathbb{E}\left[\|\bar{\mathbf{P}}^*(\mathbf{b}^\star, \sqrt{\epsilon}g^\star)\|_{\mathcal{H}}^2\right] \\
&= \mathbb{E}\left[\|\bar{\mathbf{P}}^*(\mathbf{b}^\star, \sqrt{\epsilon}g^\star) - \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\bar{\mathcal{F}}_\mu\left[\bar{\mathcal{F}}_\mu^* g^\star - (\bar{y},0)\right]\|_{\mathcal{H}}^2\right] \\
&\qquad \text{(Since by (60), } \|\bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\bar{\mathcal{F}}_\mu\left[\bar{\mathcal{F}}_\mu^* g^\star - (\bar{y},0)\right]\|_{\mathcal{H}} = 0.) \\
&= \mathbb{E}\left[\|\bar{\mathbf{P}}^*(\mathbf{b}^\star, \sqrt{\epsilon}g^\star) - \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\bar{\mathcal{F}}_\mu\left(b^\star, \sqrt{\epsilon}g^\star\right)\|_{\mathcal{H}}^2\right] \\
\text{(Since } \bar{\mathcal{F}}_\mu^* g^\star = (\mathcal{F}_\mu^* g, \sqrt{\epsilon}g) &\text{ and since by definition } b^\star = \mathcal{F}_\mu^* g^\star - \bar{y}, \text{ giving } \left[\bar{\mathcal{F}}_\mu^* g^\star - (\bar{y},0)\right] = (b^\star, \sqrt{\epsilon}g^\star).) \\
&= \mathbb{E}\left[\|\bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\left(\bar{\mathbf{F}}(\mathbf{b}^\star, \sqrt{\epsilon}g^\star) - \bar{\mathcal{F}}_\mu\left(b^\star, \sqrt{\epsilon}g^\star\right)\right)\|_{\mathcal{H}}^2\right] \\
&\qquad \text{(Factoring } \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1} \text{ out of } \bar{\mathbf{P}}^* = \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\bar{\mathbf{F}}.) \\
&= \mathbb{E}\left[\|\bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\left(\mathbf{F}\mathbf{b}^\star - \mathcal{F}_\mu b^\star\right)\|_{\mathcal{H}}^2\right] \\
&\qquad \text{(Recalling that } \bar{\mathbf{F}}(f,g) = \mathbf{F}f + \sqrt{\epsilon}g \text{ and similarly } \bar{\mathcal{F}}_\mu(f,g) = \mathcal{F}_\mu f + \sqrt{\epsilon}g.) \\
&= \mathbb{E}\left[\left\|\bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\sum_{i=1}^s\left(\varphi_{t_j}\cdot w_j\cdot\mathbf{b}^\star(j) - \frac{1}{s}\mathcal{F}_\mu b^\star\right)\right\|_{\mathcal{H}}^2\right], \qquad (61)
\end{aligned}$$

where the last equality follows since by (52), for any $\mathbf{x} \in \mathbb{C}^s$, $\mathbf{F}\mathbf{x} = \sum_{j=1}^{s} \varphi_{t_j} \cdot w_j \cdot \mathbf{x}(j)$. To simplify (61) we first bound, for any $g \in L_2(\mu)$, $\mathbb{E}\left[\langle g, \varphi_{t_j} \cdot w_j \cdot \mathbf{b}^\star(j)\rangle_\mu\right]$, recalling that $\mathbf{b}^\star(j) = w_j \cdot [\langle \varphi_{t_j}, g^\star\rangle_\mu - \bar{y}(t_j)]$. Let $p(t) = \frac{\tilde{\tau}_{\mu,\epsilon}(t)}{\tilde{s}_{\mu,\epsilon}}$ be the density with which we sample our time points $t_1, \ldots, t_s$ and $w(t) = \sqrt{\frac{1}{sT \cdot p(t)}}$ be the reweighting factor we apply if we sample time t (so $w_j = w(t_j)$).

First we argue that we can apply Fubini's theorem to switch the order of the double integration in $\mathbb{E}\left[\langle g, \varphi_{t_j} \cdot w_j \cdot \mathbf{b}^\star(j)\rangle_\mu\right]$ (over random instantiations of $\varphi_{t_j} \cdot w_j \cdot \mathbf{b}^\star(j)$ and within the inner product). Letting for $z \in L_2(\mu)$, $|z| \in L_2(\mu)$ be given by $|z|(\eta) = |z(\eta)|$ we have:

$$\mathbb{E}\left[\langle |g|, |\varphi_{t_j} \cdot w_j \cdot \mathbf{b}^\star(j)|\rangle_\mu\right] \leq \|g\|_\mu \cdot \mathbb{E}\left[\|\varphi_{t_j} w_j \mathbf{b}^\star(j)\|_\mu\right],$$

which, noting that $\|\varphi_{t_j}\|_\mu = 1$ gives:

$$\begin{aligned}
\mathbb{E}\left[\langle g, \varphi_{t_j} \cdot w_j \cdot \mathbf{b}^\star(j)\rangle_\mu\right] &\leq \|g\|_\mu \cdot \mathbb{E}\left[|w_j \mathbf{b}^\star(j)|\right] \\
&= \|g\|_\mu \cdot \int_0^T |\langle \varphi_t, g^\star\rangle_\mu - y(t)| w(t)^2 \cdot p(t)\, dt \\
&= \|g\|_\mu \cdot \frac{1}{sT} \int_0^T |\langle \varphi_t, g^\star\rangle_\mu - y(t)|\, dt \\
&< \infty
\end{aligned}$$

where the last line follows since $g \in L_2(\mu)$ so $\|g\|_\mu < \infty$ and since $\frac{1}{T}\int_0^T |\langle \varphi_t, g^\star\rangle_\mu - y(t)|\, dt \leq \frac{1}{T}\int_0^T \left(|\langle \varphi_t, g^\star\rangle_\mu - y(t)|^2 + 1\right) dt = \|\mathcal{F}_\mu^* g^\star - y\|_T^2 + T \leq \|y\|_T^2 < \infty$. Since we have established that $\mathbb{E}\left[\langle |g|, |\varphi_{t_j} \cdot w_j \cdot \mathbf{b}^\star(j)|\rangle_\mu\right]$ is finite we can apply Fubini's theorem to compute:

$$\begin{aligned}
\mathbb{E}\left[\langle g, \varphi_{t_j} \cdot w_j \cdot \mathbf{b}^\star(j)\rangle_\mu\right] &= \int_0^T [\langle \varphi_t, g^\star\rangle_\mu - y(t)] w(t)^2 \cdot \langle g, \varphi_t\rangle_\mu \cdot p(t)\, dt \\
&= \frac{1}{sT} \int_0^T \left(b^\star(t) \cdot \int_{\xi \in \mathbb{R}} g(\xi)^* e^{-2\pi i \xi t} d\mu(\xi)\right) dt \\
&= \frac{1}{s} \int_{\xi \in \mathbb{R}} \left(g(\xi)^* \cdot \frac{1}{T} \int_0^T e^{-2\pi i \xi t} b^\star(t) dt\right) d\mu(\xi) \\
&= \frac{1}{s} \langle g, \mathcal{F}_\mu b^\star\rangle_\mu. \qquad (62)
\end{aligned}$$

This in turn gives that

$$\mathbb{E}\left[\left\langle g, \varphi_{t_j} \cdot w_j \cdot \mathbf{b}^\star(j) - \frac{1}{s}\mathcal{F}_\mu b^\star\right\rangle_\mu\right] = 0$$

and so for any $g \in L_2(\mu)$:

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1} g, \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\left(\varphi_{t_j} \cdot w_j \cdot \mathbf{b}^\star(j) - \frac{1}{s}\mathcal{F}_\mu b^\star\right)\right\rangle_{\mathcal{H}}\right] = \\
&\qquad \mathbb{E}\left[\left\langle (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1} \bar{\mathcal{F}}_\mu \bar{\mathcal{F}}_\mu^* (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1} g, \varphi_{t_j} \cdot w_j \cdot \mathbf{b}^\star(j) - \frac{1}{s}\mathcal{F}_\mu b^\star\right\rangle_\mu\right] = 0. \qquad (63)
\end{aligned}$$

Further, since $t_1, \ldots, t_s$ are independent, the above gives that for $j \neq k$:

$$\mathbb{E}\left[\left\langle \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\left(\varphi_{t_j} \cdot w_j \mathbf{b}^\star(j) - \frac{1}{s}\mathcal{F}_\mu b^\star\right), \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\left(\varphi_{t_t k} \cdot w_k \mathbf{b}^\star(k) - \frac{1}{s}\mathcal{F}_\mu b^\star\right)\right\rangle_{\mathcal{H}}\right] = 0. \quad (64)$$

We can apply (63) and (64) to expand out (61), giving:

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\mathbf{P}}^*(\mathbf{F}^* g^\star - \bar{\mathbf{y}}, \sqrt{\epsilon} g^\star)\|_{\mathcal{H}}^2\right] = \\
&\quad \sum_{j=1}^{s}\sum_{k=1}^{s} \mathbb{E}\left[\left\langle \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\left(\varphi_{t_j}\cdot w_j \cdot \mathbf{b}^\star(j) - \frac{1}{s}\mathcal{F}_\mu b^\star\right), \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\left(\varphi_{t_k}\cdot w_k \cdot \mathbf{b}^\star(k) - \frac{1}{s}\mathcal{F}_\mu b^\star\right)\right\rangle_{\mathcal{H}}\right] \\
&\quad = \sum_{j=1}^{s} \mathbb{E}\left[\left\langle \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\left(\varphi_{t_j}\cdot w_j \cdot \mathbf{b}^\star(j) - \frac{1}{s}\mathcal{F}_\mu b^\star\right), \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\left(\varphi_{t_j}\cdot w_j \cdot \mathbf{b}^\star(j) - \frac{1}{s}\mathcal{F}_\mu b^\star\right)\right\rangle_{\mathcal{H}}\right] \\
&\qquad\qquad \text{(since cross terms are 0 via (64))} \\
&\quad = \sum_{j=1}^{s} \mathbb{E}\left[\left\langle \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\left(\varphi_{t_j}\cdot w_j \cdot \mathbf{b}^\star(j) - \frac{1}{s}\mathcal{F}_\mu b^\star\right), \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\left(\varphi_{t_j}\cdot w_j \cdot \mathbf{b}^\star(j)\right)\right\rangle_{\mathcal{H}}\right] \\
&\qquad\qquad \text{(applying (63) to } -\tfrac{1}{s}\mathcal{F}_\mu b^\star) \\
&\quad = \sum_{i=1}^{s} \mathbb{E}\big[\left\|\bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_{t_j}\cdot w_j \cdot \mathbf{b}^\star(j)\right\|_{\mathcal{H}}^2 - \frac{1}{s}\left\langle \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_{t_j}\cdot w_j \cdot \mathbf{b}^\star(j), \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\mathcal{F}_\mu b^\star\right\rangle_{\mathcal{H}}\big] \\
&\quad = \sum_{i=1}^{s} \mathbb{E}\big[\left\|\bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_{t_j}\cdot w_j \cdot \mathbf{b}^\star(j)\right\|_{\mathcal{H}}^2 - \frac{1}{s^2}\|\bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\mathcal{F}_\mu b^\star\|_{\mathcal{H}}^2 \\
&\quad \le \sum_{i=1}^{s} \mathbb{E}\big[\left\|\bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_{t_j}\cdot w_j \cdot \mathbf{b}^\star(j)\right\|_{\mathcal{H}}^2 \qquad (65)
\end{aligned}$$

where the second to last line follows from (62) which gives

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_{t_j}\cdot w_j \cdot \mathbf{b}^\star(j), \bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\mathcal{F}_\mu b^\star\right\rangle_{\mathcal{H}}\right] \\
&\qquad = \mathbb{E}\left[\left\langle \varphi_{t_j}\cdot w_j \cdot \mathbf{b}^\star(j), (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\bar{\mathcal{F}}_\mu\bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\mathcal{F}_\mu b^\star\right\rangle_{\mu}\right] \\
&\qquad = \frac{1}{s}\left\langle \mathcal{F}_\mu b^\star, (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\bar{\mathcal{F}}_\mu\bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\mathcal{F}_\mu b^\star\right\rangle_{\mu} \\
&\qquad = \frac{1}{s}\|\bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\mathcal{F}_\mu b^\star\|_{\mathcal{H}}^2.
\end{aligned}$$

Given the bound of (65) we can now expand out, using the fact that time t is sampled with probability proportional to $\tilde{\tau}_{\mu,\epsilon}(t)$:

$$\begin{aligned}
\mathbb{E}\left[\|\bar{\mathbf{P}}^*(\mathbf{F}^* g^\star - \bar{\mathbf{y}}, \sqrt{\epsilon} g^\star)\|_{\mathcal{H}}^2\right] &\le s \cdot \int_{t=0}^{T} \frac{\tilde{\tau}_{\mu,\epsilon}(t)}{\tilde{s}_{\mu,\epsilon}} \cdot \left\|\bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t \cdot \frac{(\langle\varphi_t, g^\star\rangle_\mu - \bar{y}(t))\cdot\tilde{s}_{\mu,\epsilon}}{sT\cdot\tilde{\tau}_{\mu,\epsilon}(u)}\right\|_{\mathcal{H}}^2 dt \\
&= \frac{1}{sT^2}\cdot\int_{t=0}^{T}\frac{\tilde{s}_{\mu,\epsilon}\cdot b^\star(t)^2}{\tilde{\tau}_{\mu,\epsilon}(t)}\cdot\left\|\bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t\right\|_{\mathcal{H}}^2 dt \\
&= \frac{1}{sT^2}\cdot\int_{t=0}^{T}\frac{\tilde{s}_{\mu,\epsilon}\cdot b^\star(t)^2}{\tilde{\tau}_{\mu,\epsilon}(t)}\cdot\langle\bar{\mathcal{F}}_\mu\bar{\mathcal{F}}_\mu^*(\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t, (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t\rangle_\mu^2 dt \\
&= \frac{1}{sT^2}\cdot\int_{t=0}^{T}\frac{\tilde{s}_{\mu,\epsilon}\cdot b^\star(t)^2}{\tilde{\tau}_{\mu,\epsilon}(t)}\cdot\langle\varphi_t, (\mathcal{G}_\mu + \epsilon\mathcal{I}_\mu)^{-1}\varphi_t\rangle_\mu^2 dt \\
&\qquad\qquad \text{(since } \bar{\mathcal{F}}_\mu\bar{\mathcal{F}}_\mu^* = \mathcal{G}_\mu + \epsilon\mathcal{I}_\mu) \\
&= \frac{1}{sT}\cdot\int_{t=0}^{T}\frac{\tilde{s}_{\mu,\epsilon}\cdot b^\star(t)^2\cdot\tau_{\mu,\epsilon}(t)}{\tilde{\tau}_{\mu,\epsilon}(t)} \qquad \text{(Theorem 5, (29))} \\
&\le \frac{\tilde{s}_{\mu,\epsilon}\cdot\|b^\star\|_T^2}{s}. \qquad \text{(since by assumption } \tilde{\tau}_{\mu,\epsilon}(t) \ge \tau_{\mu,\epsilon}(t))
\end{aligned}$$

Since $s = \Omega\left(\frac{\tilde{s}_{\mu,\epsilon}}{\delta}\right)$ we thus have via Markov's inequality, with probability $\geq 1-\delta$,

$$\|\bar{\mathbf{P}}^*(\mathbf{F}^* g^\star - \bar{\mathbf{y}}, \sqrt{\epsilon} g^\star)\|_{\mathcal{H}}^2 \leq \frac{1}{64} \cdot \|b^\star\|_T^2 \leq \frac{1}{64} \cdot B^\star$$

which completes the claim. Note that 64 is an arbitrarily chosen constant, which can be made as small as we want by increasing the sample size s by a constant factor. □

Plugging Claim 45 back into (59) gives:

$$\begin{aligned}
\|\mathbf{F}^* g - \bar{\mathbf{y}}\|_2^2 + \epsilon\|g\|_\mu^2 &\in \tilde{B} + \|\mathbf{F}^*(g-g^\star)\|_2^2 + \epsilon\|g-g^\star\|_\mu^2 \pm \frac{1}{4}(\|\mathcal{F}_\mu^*(g-g^\star)\|_T + \epsilon\|g-g^\star\|_\mu) \cdot \sqrt{B^\star} \\
&\in \tilde{B} + \|\mathbf{F}^*(g-g^\star)\|_2^2 + \epsilon\|g-g^\star\|_\mu^2 \pm \frac{1}{8}(\|\mathcal{F}_\mu^*(g-g^\star)\|_T + \epsilon\|g-g^\star\|_\mu)^2 \pm \frac{1}{8}B^\star \\
&\in \tilde{B} + \|\mathbf{F}^*(g-g^\star)\|_2^2 + \epsilon\|g-g^\star\|_\mu^2 \pm \frac{1}{4}(\|\mathcal{F}_\mu^*(g-g^\star)\|_T^2 + \epsilon\|g-g^\star\|_\mu^2) \pm \frac{1}{8}B^\star.
\end{aligned}$$

Applying the operator approximation bound of Lemma 43 with error $\Delta = 1/4$ then gives:

$$\|\mathbf{F}^* g - \bar{\mathbf{y}}\|_2^2 + \epsilon\|g\|_\mu^2 \in \tilde{B} + \left(1 \pm \frac{1}{2}\right)\left(\|\mathcal{F}_\mu^*(g-g^\star)\|_2^2 + \epsilon\|g-g^\star\|_\mu^2\right) \pm \frac{1}{8}B^\star.$$

Finally, applying Claim 44 gives:

$$\|\mathbf{F}^* g - \bar{\mathbf{y}}\|_2^2 + \epsilon\|g\|_\mu^2 \in (\tilde{B} - B^\star) + \|\mathcal{F}_\mu^* g - \bar{y}\|_T^2 + \epsilon\|g\|_\mu^2 \pm \frac{1}{2}\left(\|\mathcal{F}_\mu^* g - \bar{y}\|_T^2 + \epsilon\|g\|_\mu^2\right).$$

Note that $\mathbb{E}[\tilde{B}] = B^\star$. So writing $\alpha = \tilde{B} - B^\star$ we have $|\alpha| \leq \frac{1}{\delta} \cdot B^\star$ with probability $1-\delta$. This completes the theorem. □

C.4 Frequency Subset Selection

We now prove the frequency subset selection guarantee Theorem 9 used in Section 5.1 to bound the leverage scores for general constraints μ, by showing that $\mathcal{F}_\mu^*$ can be well approximated by an operator whose columns are spanned by just $O(s_{\mu,\epsilon})$ frequencies.

Theorem 9 (Frequency Subset Selection)**.** *For some $s \leq \lceil 36 \cdot s_{\mu,\epsilon} \rceil$ there exists a set of distinct frequencies $\xi_1, \ldots, \xi_s \in \mathbb{C}$ such that, letting $\mathbf{C}_s : L_2(T) \to \mathbb{C}^s$ be defined by:*

$$[\mathbf{C}_s g](j) = \frac{1}{T}\int_0^T g(t) e^{-2\pi i \xi_j t}\, dt,$$

and $\mathbf{Z} = (\mathbf{C}_s \mathbf{C}_s^*)^{-1}\mathbf{C}_s \mathcal{F}_\mu^*$, for $\varphi_t \in L_2(\mu), \boldsymbol{\phi}_t \in \mathbb{C}^s$ with $\varphi_t(\xi) = e^{-2\pi i t \xi}$ and $\boldsymbol{\phi}_t(j) = \varphi_t(\xi_j)$:

$$\frac{1}{T}\int_{t\in[0,T]} \|\varphi_t - \mathbf{Z}^* \boldsymbol{\phi}_t\|_\mu^2\, dt \leq 4\epsilon \cdot s_{\mu,\epsilon}. \tag{66}$$

Our proof relies on the following spectral error bound for weighted frequency subset selection:

Lemma 46 (Frequency Subset Selection – Direct Spectral Approximation)**.** *For some $s \leq \lceil 36 \cdot s_{\mu,\epsilon} \rceil$ there exists a set of distinct frequencies $\xi_1, \ldots, \xi_s \in \mathbb{R}$ and positive weights $w_1, \ldots, w_s \in \mathbb{R}$ such that letting $\bar{\mathbf{C}}_s : L_2(T) \to \mathbb{C}^s$ be given by:*

$$[\bar{\mathbf{C}}_s g](j) = \frac{1}{T}\int_0^T g(t) w_j e^{-2\pi i \xi_j t}\, dt,$$

and letting $\widehat{\mathcal{K}}_\mu = \bar{\mathbf{C}}_s^* \bar{\mathbf{C}}_s$*, we have*

$$\frac{1}{2} \cdot (\mathcal{K}_\mu + \epsilon \mathcal{I}_T) \preceq \widehat{\mathcal{K}}_\mu + \epsilon \mathcal{I}_T \preceq \frac{3}{2} \cdot (\mathcal{K}_\mu + \epsilon \mathcal{I}_T). \tag{67}$$

Proof. We prove a more general statement, in which we are given $0 < \Delta < 1$ and we select $s = \lceil 9 s_{\mu,\epsilon} / \Delta^2 \rceil$ frequencies $\xi_1, \dots, \xi_s \in \mathbb{R}$ and weights $w_1, \dots, w_s \in \mathbb{R}$ such that

$$(1 - \Delta)(\mathcal{K}_\mu + \epsilon \mathcal{I}_T) \preceq \widehat{\mathcal{K}}_\mu + \epsilon \mathcal{I}_T \preceq (1 + \Delta)(\mathcal{K}_\mu + \epsilon \mathcal{I}_T).$$

The claim follows by setting $\Delta = 1/2$. We can assume that $\xi_1, \dots, \xi_s$ are distinct, since if ξ_i, ξ_j are equal, we can simply remove ξ_j and update $w_i \leftarrow \sqrt{w_i^2 + w_j^2}$, leaving $\widehat{\mathcal{K}}_\mu$ unchanged and only decreasing s.

The last condition is equivalent to

$$\mathcal{K}_\mu - \Delta(\mathcal{K}_\mu + \epsilon \mathcal{I}_T) \preceq \widehat{\mathcal{K}}_\mu \preceq \mathcal{K}_\mu + \Delta(\mathcal{K}_\mu + \epsilon \mathcal{I}_T).$$

Multiplying with $(\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1/2}$ on the left and right, we find that the condition is equivalent to:

$$-\Delta \mathcal{I}_T \preceq (\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1/2} \widehat{\mathcal{K}}_\mu (\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1/2} - (\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1/2} \mathcal{K}_\mu (\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1/2} \preceq \Delta \mathcal{I}_T.$$

To shorten notation, we write $\mathcal{Z} = (\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1/2} \mathcal{K}_\mu (\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1/2}$ and $\widehat{\mathcal{Z}} = (\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1/2} \widehat{\mathcal{K}}_\mu (\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1/2}$. Given $\xi \in \mathbb{R}$, we define $\vartheta_\xi(t) \stackrel{\text{def}}{=} e^{2\pi i t \xi}$ ($\vartheta_\xi \in L_2(T)$). It is easy to verify that

$$\mathcal{K}_\mu = \int_{\mathbb{R}} (\vartheta_\xi \otimes \vartheta_\xi) d\mu(\xi)$$

and

$$\widehat{\mathcal{K}}_\mu = \sum_{i=1}^{s} w_i^2 (\vartheta_{\xi_i} \otimes \vartheta_{\xi_i}).$$

Further define $\bar{\vartheta}_\xi \stackrel{\text{def}}{=} (\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1/2} \vartheta_\xi$. Since $(\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1/2}$ is self-adjoint and bounded, we have

$$\mathcal{Z} = \int_{\mathbb{R}} (\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi) d\mu(\xi)$$

and

$$\widehat{\mathcal{Z}} = \sum_{i=1}^{s} w_i^2 (\bar{\vartheta}_{\xi_i} \otimes \bar{\vartheta}_{\xi_i}).$$

We prove the existence of $\xi_1, \dots, \xi_s$ and $w_1, \dots, w_s$ using the deterministic selection process known as "BSS" [BSS14].[29] In particular, we use a process that in essence is the same as the one described in [CNW16, Theorem 5 (arxiv version)]. Indeed, since $\|\mathcal{Z}\|_{\text{op}} \leq 1$ and $\text{tr}(\mathcal{Z}) = s_{\mu,\epsilon}$ the aforementioned results would suffice if we were dealing with matrices instead of operators. The rest of the proof extends these results to the operator case. Let

$$\delta_u \stackrel{\text{def}}{=} \Delta/3 + 2\Delta^2/9, \quad \delta_l \stackrel{\text{def}}{=} \Delta/3 - 2\Delta^2/9$$

[29] We remark that unlike the process described in [BSS14], our existence proof does not trivially translate to an algorithm, since it involves a search over an infinite domain. Nevertheless, for our needs, existence suffices.

and for $j = 0, 1, \ldots, s$,

$$\mathcal{X}_l^{(j)} \stackrel{\text{def}}{=} j\delta_l \cdot \mathcal{Z} - s_{\mu,\epsilon} \cdot \mathcal{I}_T, \quad \mathcal{X}_u^{(j)} \stackrel{\text{def}}{=} j\delta_u \cdot \mathcal{Z} + s_{\mu,\epsilon} \cdot \mathcal{I}_T.$$

The process we shall describe iteratively selects $\xi_1, \xi_2, \ldots$ and unscaled weights $\tilde{w}_1, \tilde{w}_2, \ldots$ such that if we define $\widehat{\mathcal{Z}}^{(j)} \stackrel{\text{def}}{=} \sum_{i=1}^{j} \tilde{w}_i (\bar{\vartheta}_{\xi_i} \otimes \bar{\vartheta}_{\xi_i})$ the invariant

$$\mathcal{X}_l^{(j)} \prec \widehat{\mathcal{Z}}^{(j)} \prec \mathcal{X}_u^{(j)} \tag{68}$$

is held. Let us write $s = \lceil 9 s_{\mu,\epsilon}/\Delta^2 \rceil$, so $s = C s_{\mu,\epsilon}/\Delta^2$ for $C \geq 9$. If indeed we are able to select the frequencies and weights for s steps such that this invariant holds, we shall have

$$\frac{C s_{\mu,\epsilon}}{3\Delta} \cdot \mathcal{Z} - (1 + 2C/9) \cdot s_{\mu,\epsilon} \cdot \mathcal{I}_T \preceq \widehat{\mathcal{Z}}^{(s)} \preceq \frac{C s_{\mu,\epsilon}}{3\Delta} \cdot \mathcal{Z} + (1 + 2C/9) \cdot s_{\mu,\epsilon} \cdot \mathcal{I}_T$$

where we used the fact that $\mathcal{Z} \preceq \mathcal{I}_T$. Since $C \geq 9$ we have

$$-\Delta \cdot \mathcal{I}_T \preceq \frac{3\Delta}{C s_{\mu,\epsilon}} \widehat{\mathcal{Z}}^{(s)} - \mathcal{Z} \preceq \Delta \cdot \mathcal{I}_T$$

so by defining $w_i = \sqrt{\frac{3\Delta}{C s_{\mu,\epsilon}}} \tilde{w}_i$ for $i = 1, \ldots, s$ we shall then have $\widehat{\mathcal{Z}} = \frac{3\Delta}{C s_{\mu,\epsilon}} \widehat{\mathcal{Z}}^{(s)}$ thereby establishing the desired bound.

Thus, it suffices to show that we can select frequencies and weights iteratively so that (68) is maintained. In fact, the iterative selection process will maintain two additional invariants:

$$\int_{\mathbb{R}} \langle \bar{\vartheta}_\xi, (\mathcal{X}_u^{(j)} - \widehat{\mathcal{Z}}^{(j)})^{-1} \bar{\vartheta}_\xi \rangle_T d\mu(\xi) \leq 1$$
$$\int_{\mathbb{R}} \langle \bar{\vartheta}_\xi, (\widehat{\mathcal{Z}}^{(j)} - \mathcal{X}_l^{(j)})^{-1} \bar{\vartheta}_\xi \rangle_T d\mu(\xi) \leq 1$$

All the invariants hold for $j = 0$. Eq. (68) trivially holds for $j = 0$. As for the integral,

$$\begin{aligned}
\int_{\mathbb{R}} \langle \bar{\vartheta}_\xi, (\mathcal{X}_u^{(0)} - \widehat{\mathcal{Z}}^{(0)})^{-1} \bar{\vartheta}_\xi \rangle_T d\mu(\xi) &= \int_{\mathbb{R}} \langle \bar{\vartheta}_\xi, s_{\mu,\epsilon}^{-1} \bar{\vartheta}_\xi \rangle_T d\mu(\xi) \\
&= s_{\mu,\epsilon}^{-1} \int_{\mathbb{R}} \langle (\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1/2} \vartheta_\xi, (\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1/2} \vartheta_\xi \rangle_T d\mu(\xi) \\
&= s_{\mu,\epsilon}^{-1} \int_{\mathbb{R}} \langle \vartheta_\xi, (\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1} \vartheta_\xi \rangle_T d\mu(\xi) \\
&= s_{\mu,\epsilon}^{-1} \operatorname{tr} \left((\mathcal{K}_\mu + \epsilon \mathcal{I}_T)^{-1} \mathcal{K}_T \right) = 1
\end{aligned}$$

and similarly for the second invariant. In the above, the last equality is due to Claim 34.

Suppose by induction that the invariants for j. We prove that it is possible to pick a frequency ξ and weight $w > 0$ such that if we set $\xi_{j+1} = \xi$ and $\tilde{w}_{j+1} = w$ then the invariants will hold for $j + 1$.

Fix j. For $t \geq 0$, let us denote

$$M_u(t) = \left(\mathcal{X}_u^{(j)} + t\mathcal{Z} - \widehat{\mathcal{Z}}^{(j)} \right)^{-1}$$
$$M_l(t) = \left(\widehat{\mathcal{Z}}^{(j)} - \mathcal{X}_l^{(j)} - t\mathcal{Z} \right)^{-1}$$

where M_u is defined for any t (since the inverted operator is strictly positive and bounded, so invertible), and M_l is defined for $t < 1$. We can define M_l for $t < 1$ since $\widehat{\mathcal{Z}}^{(j)} - \mathcal{X}_l^{(j)} - t\mathcal{Z} \succ 0$ for $t < 1$ as we now show. Due to Claim 34:

$$\mathrm{tr}((\widehat{\mathcal{Z}}^{(j)} - \mathcal{X}_l^{(j)})^{-1}\mathcal{Z}) = \int_{\mathbb{R}} \langle \bar{\vartheta}_\xi, (\widehat{\mathcal{Z}}^{(j)} - \mathcal{X}_l^{(j)})^{-1}\bar{\vartheta}_\xi\rangle_T d\mu(\xi) \leq 1.$$

Since $\widehat{\mathcal{Z}}^{(j)} - \mathcal{X}_l^{(j)} \succ 0$ (induction assumption), $(\widehat{\mathcal{Z}}^{(j)} - \mathcal{X}_l^{(j)})^{-1}$ is bounded so according to Claim 29, $\mathcal{Z} \preceq \widehat{\mathcal{Z}}^{(j)} - \mathcal{X}_l^{(j)}$, and then Claim 31 implies that $\widehat{\mathcal{Z}}^{(j)} - \mathcal{X}_l^{(j)} - t\mathcal{Z} \succ 0$.

Consider some fixed ξ. We first claim that for $w < 1/\langle \bar{\vartheta}_\xi, M_u(\delta_u)\bar{\vartheta}_\xi\rangle_T$ we have $M_u(\delta_u)^{-1} - w(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi) \succ 0$. Obviously, the last statement holds for $w = 0$, and due to continuity of $w \mapsto \langle x, (M_u(\delta_u)^{-1} - w(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi))x\rangle_T$ with respect to w, it will also hold for some interval around 0. Let $w^\star$ be the maximal value such that for all $w \in [0, w^\star)$ we have $M_u(\delta_u)^{-1} - w(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi) \succ 0$. Our goal is to show that $w^\star \geq 1/\langle \bar{\vartheta}_\xi, M_u(\delta_u)\bar{\vartheta}_\xi\rangle_T$. Assume by contradiction that $w^\star < 1/\langle \bar{\vartheta}_\xi, M_u(\delta_u)\bar{\vartheta}_\xi\rangle_T$. For every $w \in [0, w^\star)$, the operator $M_u(\delta_u)^{-1} - w(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi)$ is invertible, and we can apply a operator pseudo-inversion lemma due to Deng [Den11, Theorem 2.1] to find that

$$(M_u(\delta_u)^{-1} - w(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_T))^{-1} = M_u(\delta_u) + \frac{w}{1 - w \cdot \langle \bar{\vartheta}_\xi, M_u(\delta_u)\bar{\vartheta}_\xi\rangle_T} M_u(\delta_u)(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi)M_u(\delta_u).$$

Since we assumed $w^\star < 1/\langle \bar{\vartheta}_\xi, M_u(\delta_u)\bar{\vartheta}_\xi\rangle_T$, clearly, there exists a K such that for all $w \in [0, w^\star)$ we have:

$$(M_u(\delta_u)^{-1} - w(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_T))^{-1} \leq K \cdot \mathcal{I}_T.$$

Note that $M_u(\delta_u)^{-1} - w^\star(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi)$ is not strictly positive for otherwise due to continuity we could have extended the interval, so there exists a x with norm 1 such that $\langle x, (M_u(\delta_u)^{-1} - w^\star(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi))x\rangle < 1/2K$. Let $w_1, w_2, \ldots$ be a sequence which converges to $w^\star$, and let $y_i = (M_u(\delta_u)^{-1} - w_i(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi)^{1/2})x$. We now have $\langle y_i, y_i\rangle_T = \langle x, (M_u(\delta_u)^{-1} - w_i(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi)x\rangle_T \to \langle x, (M_u(\delta_u)^{-1} - w^\star(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi))x\rangle_T < 1/2K$ as $i \to \infty$. However $\langle y_i, (M_u(\delta_u)^{-1} - w_i(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi))^{-1}y_i\rangle_T = \langle x_i, x_i\rangle_T = 1$ which contradicts the bound on $(M_u(\delta_u)^{-1} - w_i(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi))^{-1}$.

Thus, if we picked ξ and $w < 1/\langle \bar{\vartheta}_\xi, M_u(\delta_u)\bar{\vartheta}_\xi\rangle_T$ for the step, we shall have $\widehat{\mathcal{Z}}^{(j+1)} - \mathcal{X}_l^{(j)} = M_u(\delta_u)^{-1} - w(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi) \succ 0$ as required, and the upper invariant will translate to

$$\int_{\mathbb{R}} \left\langle \bar{\vartheta}_\eta, \left(M_u(\delta_u)^{-1} - w(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi)\right)^{-1} \bar{\vartheta}_\eta \right\rangle_T d\mu(\eta) \leq 1,$$

which is equivalent to

$$\int_{\mathbb{R}} \langle \bar{\vartheta}_\eta, M_u(\delta_u)\bar{\vartheta}_\eta\rangle_T d\mu(\eta) + \frac{w \cdot \int_{\mathbb{R}} \left\langle \bar{\vartheta}_\eta, M_u(\delta_u)(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi)M_u(\delta_u)\bar{\vartheta}_\eta \right\rangle_T d\mu(\eta)}{1 - w \cdot \langle \bar{\vartheta}_\xi, M_u(\delta_u)\bar{\vartheta}_\xi\rangle_T} \leq 1.$$

The induction hypothesis is

$$\int_{\mathbb{R}} \langle \bar{\vartheta}_\eta, M_u(0)\bar{\vartheta}_\eta\rangle_T d\mu(\eta) \leq 1.$$

so the upper invariant is held if

$$\int_{\mathbb{R}} \langle \bar{\vartheta}_\eta, M_u(\delta_u)\bar{\vartheta}_\eta\rangle_T d\mu(\eta) - \int_{\mathbb{R}} \langle \bar{\vartheta}_\eta, M_u(0)\bar{\vartheta}_\eta\rangle_T d\mu(\eta) + \frac{w \cdot \int_{\mathbb{R}} \left\langle \bar{\vartheta}_\eta, M_u(\delta_u)(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi)M_u(\delta_u)\bar{\vartheta}_\eta \right\rangle_T d\mu(\eta)}{1 - w \cdot \langle \bar{\vartheta}_\xi, M_u(\delta_u)\bar{\vartheta}_\xi\rangle_T} \leq 0. \tag{69}$$

Consider any $\eta \in \mathbb{R}$, and let $f_\eta(y) \stackrel{\text{def}}{=} \langle \bar{\vartheta}_\eta, M_u(y)\bar{\vartheta}_\eta\rangle_T$. Using the operator inversion formula, we have for any $t_2 \geq t_1$:

$$M_u(t_2) = M_u(t_1) - (t_2 - t_1)M_u(t_1)\mathcal{Z}^{1/2}\left(\mathcal{I}_T + (t_2 - t_1)\mathcal{Z}^{1/2}M_u(t_1)\mathcal{Z}^{1/2}\right)^{-1}\mathcal{Z}^{1/2}M_u(t_1).$$

From this equation we see that

$$f'_\eta(y) = \langle \bar{\vartheta}_\eta, M_u(y)\mathcal{Z}M_u(y)\bar{\vartheta}_\eta\rangle_T.$$

Furthermore, since for $t_2 > t_1$ we have $\mathcal{I}_T + t_2\mathcal{Z}^{1/2}M_u(t_1)\mathcal{Z}^{1/2} \succeq \mathcal{I}_T + t_1\mathcal{Z}^{1/2}M_u(t_1)\mathcal{Z}^{1/2}$ and both operators are strictly positive and bounded, then $(\mathcal{I}_T + t_1\mathcal{Z}^{1/2}M_u(t_1)\mathcal{Z}^{1/2})^{-1} \preceq (\mathcal{I}_T + t_2\mathcal{Z}^{1/2}M_u(t_1)\mathcal{Z}^{1/2})^{-1}$, and we can easily verify that f_η is convex. Thus,

$$f_\eta(\delta_u) - f_\eta(0) \leq -\delta_u\langle \bar{\vartheta}_\eta, M_u(y)\mathcal{Z}M_u(y)\bar{\vartheta}_\eta\rangle_T.$$

After integrating on both sides, we have the bound

$$\int_{\mathbb{R}}\langle \bar{\vartheta}_\eta, M_u(\delta_u)\bar{\vartheta}_\eta\rangle_T d\mu(\eta) - \int_{\mathbb{R}}\langle \bar{\vartheta}_\eta, M_u(0)\bar{\vartheta}_\eta\rangle_T d\mu(\eta) \leq -\delta_u\int_{\mathbb{R}}\langle \bar{\vartheta}_\eta, M_u(\delta_u)\mathcal{Z}M_u(\delta_u)\bar{\vartheta}_\eta\rangle_T d\mu(\eta).$$

Using this bound in (69) and rearranging, we find that for any ξ, the upper invariant is held if we select w such that

$$\frac{1}{w} > \frac{\int_{\mathbb{R}}\left\langle \bar{\vartheta}_\eta, M_u(\delta_u)(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi)M_u(\delta_u)\bar{\vartheta}_\eta\right\rangle_T d\mu(\eta)}{\delta_u\int_{\mathbb{R}}\langle \bar{\vartheta}_\eta, M_u(\delta_u)\mathcal{Z}M_u(\delta_u)\bar{\vartheta}_\eta\rangle_T d\mu(\eta)} + \langle \bar{\vartheta}_\xi, M_u(\delta_u)\bar{\vartheta}_\xi\rangle_T. \tag{70}$$

Note that if this is held, we also have $w < 1/\langle \bar{\vartheta}_\xi, M_u(\delta_u)\bar{\vartheta}_\xi\rangle_T$, as previously required.

We now consider the lower invariants. If we picked ξ and $w > 0$ for the step, then $\widehat{\mathcal{Z}}^{(j+1)} - \mathcal{X}_l^{(j)} = M_l(\delta_l)^{-1} + w(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi) \succeq M_l(\delta_l)^{-1} \succ 0$ as long $\delta_l < 1$ which holds for our choice of δ_l. So the left part of (68) will hold regardless of how we choose ξ and $w > 0$. As for the lower trace bound, it translates to

$$\int_{\mathbb{R}}\left\langle \bar{\vartheta}_\eta, \left(M_l(\delta_l)^{-1} + w(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi)\right)^{-1}\bar{\vartheta}_\eta\right\rangle_T d\mu(\eta) \leq 1.$$

Applying another variant of operator pseudo-inversion lemma [Oga88, Theorem 2], we find that the last condition is equivalent to

$$\int_{\mathbb{R}}\langle \bar{\vartheta}_\eta, M_l(\delta_l)\bar{\vartheta}_\eta\rangle_T d\mu(\eta) - \frac{w\cdot\int_{\mathbb{R}}\left\langle \bar{\vartheta}_\eta, M_l(\delta_l)(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi)M_l(\delta_l)\bar{\vartheta}_\eta\right\rangle_T d\mu(\eta)}{1 + w\cdot\langle \bar{\vartheta}_\xi, M_l(\delta_l)\bar{\vartheta}_\xi\rangle_T} \leq 1.$$

The induction hypothesis is

$$\int_{\mathbb{R}}\langle \bar{\vartheta}_\eta, M_l(0)\bar{\vartheta}_\eta\rangle_T d\mu(\eta) \leq 1$$

so the lower invariant is held if

$$\int_{\mathbb{R}}\langle \bar{\vartheta}_\eta, M_l(\delta_l)\bar{\vartheta}_\eta\rangle_\mu d\mu(\eta) - \int_{\mathbb{R}}\langle \bar{\vartheta}_\eta, M_l(0)\bar{\vartheta}_\eta\rangle_T d\mu(\eta) - \frac{w\cdot\int_{\mathbb{R}}\left\langle \bar{\vartheta}_\eta, M_l(\delta_l)(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi)M_l(\delta_l)\bar{\vartheta}_\eta\right\rangle_T d\mu(\eta)}{1 + w\cdot\langle \bar{\vartheta}_\xi, M_l(\delta_l)\bar{\vartheta}_\xi\rangle_T} \leq 0. \tag{71}$$

Similarly to before, by using the convexity of each integrand, we can bound

$$\int_{\mathbb{R}}\langle \bar{\vartheta}_\eta, M_l(\delta_l)\bar{\vartheta}_\eta\rangle_T d\mu(\eta) - \int_{\mathbb{R}}\langle \bar{\vartheta}_\eta, M_l(0)\bar{\vartheta}_\eta\rangle_T d\mu(\eta) \leq \delta_l\int_{\mathbb{R}}\langle \bar{\vartheta}_\eta, M_l(\delta_l)\mathcal{Z}M_u(\delta_l)\bar{\vartheta}_\eta\rangle_T d\mu(\eta).$$

Using this bound in (71) and rearranging, we find that for any ξ, the lower invariant is held if we select w such that

$$\frac{1}{w} \leq \frac{\int_{\mathbb{R}} \left\langle \bar{\vartheta}_\eta, M_l(\delta_l)(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi) M_l(\delta_l) \bar{\vartheta}_\eta \right\rangle_T d\mu(\eta)}{\delta_l \int_{\mathbb{R}} \langle \bar{\vartheta}_\eta, M_l(\delta_l) \mathcal{Z} M_u(\delta_l) \bar{\vartheta}_\eta \rangle_T d\mu(\eta)} - \langle \bar{\vartheta}_\xi, M_l(\delta_l) \bar{\vartheta}_\xi \rangle_T. \tag{72}$$

Thus, we need to show that there exists a ξ and w such that both (70) and (72) hold. However, for a given ξ, such a w will surely exist if

$$\begin{aligned}
&\frac{\int_{\mathbb{R}} \left\langle \bar{\vartheta}_\eta, M_u(\delta_u)(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi) M_u(\delta_u) \bar{\vartheta}_\eta \right\rangle_T d\mu(\eta)}{\delta_u \int_{\mathbb{R}} \langle \bar{\vartheta}_\eta, M_u(\delta_u) \mathcal{Z} M_u(\delta_u) \bar{\vartheta}_\eta \rangle_T d\mu(\eta)} + \langle \bar{\vartheta}_\xi, M_u(\delta_u) \bar{\vartheta}_\xi \rangle_T \\
&< \frac{\int_{\mathbb{R}} \left\langle \bar{\vartheta}_\eta, M_l(\delta_l)(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi) M_l(\delta_l) \bar{\vartheta}_\eta \right\rangle_T d\mu(\eta)}{\delta_l \int_{\mathbb{R}} \langle \bar{\vartheta}_\eta, M_l(\delta_l) \mathcal{Z} M_u(\delta_l) \bar{\vartheta}_\eta \rangle_T d\mu(\eta)} - \langle \bar{\vartheta}_\xi, M_l(\delta_l) \bar{\vartheta}_\xi \rangle_T.
\end{aligned}$$

Thus, it it suffices to show that there exists a ξ for which the last inequality holds. To show that such a ξ exists, we will show that the inequality holds for the integral of both sides with respect to μ measure. This will guarantee the existence of such a ξ since the Lebesgue integral is strictly positive for non-negative functions. We compute:

$$\begin{aligned}
&\int_{\mathbb{R}} \int_{\mathbb{R}} \left\langle \bar{\vartheta}_\eta, M_u(\delta_u)(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi) M_u(\delta_u) \bar{\vartheta}_\eta \right\rangle_T d\mu(\eta) d\mu(\xi) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \left\langle \bar{\vartheta}_\eta, M_u(\delta_u)(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi) M_u(\delta_u) \bar{\vartheta}_\eta \right\rangle_T d\mu(\xi) d\mu(\eta) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \left\langle M_u(\delta_u) \bar{\vartheta}_\eta, (\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi) M_u(\delta_u) \bar{\vartheta}_\eta \right\rangle_T d\mu(\xi) d\mu(\eta) \\
&= \int_{\mathbb{R}} \left\langle M_u(\delta_u) \bar{\vartheta}_\eta, \mathcal{Z} M_u(\delta_u) \bar{\vartheta}_\eta \right\rangle_T d\mu(\eta) \\
&= \int_{\mathbb{R}} \langle \bar{\vartheta}_\eta, M_u(\delta_u) \mathcal{Z} M_u(\delta_u) \bar{\vartheta}_\eta \rangle_T d\mu(\eta).
\end{aligned}$$

Similarly,

$$\int_{\mathbb{R}} \int_{\mathbb{R}} \left\langle \bar{\vartheta}_\eta, M_l(\delta_l)(\bar{\vartheta}_\xi \otimes \bar{\vartheta}_\xi) M_l(\delta_l) \bar{\vartheta}_\eta \right\rangle_T d\mu(\eta) d\mu(\xi) = \int_{\mathbb{R}} \langle \bar{\vartheta}_\eta, M_l(\delta_l) \mathcal{Z} M_l(\delta_l) \bar{\vartheta}_\eta \rangle_T d\mu(\eta).$$

$\mathcal{Z}$ is self-adjoint and positive definite, so the operator pseudo-inversion lemma [Oga88, Theorem 2] implies that $M_u(\delta_u) \preceq M_u(0)$, so by the induction hypothesis

$$\int_{\mathbb{R}} \langle \bar{\vartheta}_\xi, M_u(\delta_u) \bar{\vartheta}_\xi \rangle_T d\mu(\xi) \leq \int_{\mathbb{R}} \langle \bar{\vartheta}_\xi, M_u(0) \bar{\vartheta}_\xi \rangle_T d\mu(\xi) \leq 1.$$

We now consider the lower invariant. We already showed that $\mathcal{Z} \preceq \widehat{\mathcal{Z}}^{(j)} - \mathcal{X}_l^{(j)}$, so as long as $\delta_l \leq 1/2$ we will have:

$$\int_{\mathbb{R}} \langle \bar{\vartheta}_\xi, M_l(\delta_l) \bar{\vartheta}_\xi \rangle_T d\mu(\xi) = \int_{\mathbb{R}} \langle \bar{\vartheta}_\xi, (M_l(0)^{-1} - \delta_l \mathcal{Z})^{-1} \bar{\vartheta}_\xi \rangle_T d\mu(\xi) \leq 2 \int_{\mathbb{R}} \langle \bar{\vartheta}_\xi, M_l(0) \bar{\vartheta}_\xi \rangle_T d\mu(\xi) \leq 2$$

where we used Claim 30. So there will be a gap in the value of the integrals (as desired), if

$$\frac{1}{\delta_u} + 1 < \frac{1}{\delta_l} - 2,$$

which is the case for our selection of δ_l and δ_u. □

From Lemma 46 we can prove a stronger spectral error bound for the projection onto the range of $\bar{\mathbf{C}}_s$.

Lemma 47 (Frequency Subset Selection – Projection Based Spectral Approximation)**.** *For some $s \le \lceil 36 \cdot s_{\mu,\epsilon} \rceil$ there exists a set of distinct frequencies $\xi_1, \dots, \xi_s \in \mathbb{C}$ such that letting $\mathbf{C}_s : L_2(T) \to \mathbb{C}^s$ and $\mathbf{Z} : L_2(\mu) \to \mathbb{C}^s$ be defined as in Theorem 9 and $\widehat{\mathcal{G}}_\mu = \mathbf{Z}^*\mathbf{C}_s\mathbf{C}_s^*\mathbf{Z}$,*

$$\widehat{\mathcal{G}}_\mu \preceq \mathcal{G}_\mu \preceq \widehat{\mathcal{G}}_\mu + \epsilon\mathcal{I}_\mu. \tag{73}$$

Proof. Let $\xi_1, \dots, \xi_s \in \mathbb{C}$ and $w_1, \dots, w_s \in \mathbb{R}$ be the frequencies and weights shown to exist in Lemma 46 and let $\bar{\mathbf{C}}_s$ be as defined in that lemma (note that $\bar{\mathbf{C}}_s$ is identical to $\mathbf{C}_s$ except with its rows weighted by $w_1, \dots, w_s$.) First note that for any $g \in L_2(\mu)$,

$$\langle g, \widehat{\mathcal{G}}_\mu g\rangle_\mu = \|\mathbf{C}_s^*\mathbf{Z}g\|_\mu^2 = \|\mathbf{C}_s^*(\mathbf{C}_s\mathbf{C}_s^*)^{-1}\mathbf{C}_s\mathcal{F}_\mu^* g\|_\mu^2 \le \|\mathcal{F}_\mu^* g\|_\mu^2 = \langle g, \mathcal{G}_\mu g\rangle_\mu$$

where the inequality follows from observing that $\mathbf{C}_s^*(\mathbf{C}_s\mathbf{C}_s^*)^{-1}\mathbf{C}_s$ is an orthogonal projection. Thus $\widehat{\mathcal{G}}_\mu \preceq \mathcal{G}_\mu$. It remains to show that $\mathcal{G}_\mu \preceq \widehat{\mathcal{G}}_\mu + \epsilon\mathcal{I}_\mu$. Let $\bar{\mathcal{P}} = \mathcal{I}_T - \mathbf{C}_s^*(\mathbf{C}_s\mathbf{C}^*)^{-1}\mathbf{C}_s$ be the projection to the orthogonal complement of $\mathbf{C}_s^*$'s range and let $\widehat{\mathcal{K}}_\mu = \bar{\mathbf{C}}_s^*\bar{\mathbf{C}}_s$ be as defined in Lemma 46. Rearranging the guarantee of Lemma 46 gives

$$\mathcal{K}_\mu \preceq 2 \cdot \widehat{\mathcal{K}}_\mu + \epsilon\mathcal{I}_T$$

which immediately gives

$$\bar{\mathcal{P}}\mathcal{K}_\mu\bar{\mathcal{P}} \preceq 2 \cdot \bar{\mathcal{P}}\widehat{\mathcal{K}}_\mu\bar{\mathcal{P}} + \epsilon\bar{\mathcal{P}}\mathcal{I}_T\bar{\mathcal{P}}.$$

Note that $\bar{\mathbf{C}}_s\bar{\mathcal{P}} = 0$ (since $\bar{\mathcal{P}}$ is an orthogonal projection onto $\ker(\mathbf{C}_s) = \ker(\bar{\mathbf{C}}_s)$) and so $\bar{\mathcal{P}}\widehat{\mathcal{K}}_\mu\bar{\mathcal{P}} = 0$, giving:

$$\bar{\mathcal{P}}\mathcal{K}_\mu\bar{\mathcal{P}} \preceq \epsilon\bar{\mathcal{P}}\mathcal{I}_T\bar{\mathcal{P}} \preceq \epsilon\mathcal{I}_T. \tag{74}$$

Note that $\bar{\mathcal{P}}\mathcal{K}_\mu\bar{\mathcal{P}} = \bar{\mathcal{P}}\mathcal{F}_\mu^*\mathcal{F}_\mu\bar{\mathcal{P}}$ and

$$\mathcal{G}_\mu - \widehat{\mathcal{G}_\mu} = \mathcal{F}_\mu\mathcal{F}_\mu^* - \mathbf{Z}^*\mathbf{C}_s\mathbf{C}_s^*\mathbf{Z} = \mathcal{F}_\mu\bar{\mathcal{P}}\mathcal{F}_\mu^*.$$

Thus by (74) we also have $\mathcal{G}_\mu - \widehat{\mathcal{G}_\mu} \preceq \epsilon\mathcal{I}_\mu$ (since the norm of an operator and its adjoint are the same so $\bar{\mathcal{P}}\mathcal{K}_\mu\bar{\mathcal{P}} \preceq \epsilon\mathcal{I}_T \implies \mathcal{F}_\mu\bar{\mathcal{P}}\mathcal{F}_\mu^* \preceq \epsilon\mathcal{I}_\mu$), which completes the lemma. □

Finally, from Lemma 47 we can prove the frequency subset selection guarantee of Theorem 9.

Proof of Theorem 9. We consider the same set of frequencies $\xi_1, \dots, \xi_s$ shown to exist in Lemma 47 and the corresponding operators $\mathbf{C}_s$, $\mathbf{Z}$. We show that these frequencies satisfy the guarantee of Theorem 9. First, we note that

$$\mathbf{K} \stackrel{\text{def}}{=} \mathbf{C}_s\mathbf{C}_s^* = \frac{1}{T}\int_0^T (\boldsymbol{\phi}_t \otimes \boldsymbol{\phi}_t)dt$$

(In the above, we abuse notation and use $\boldsymbol{\phi}_t$ to denote both the vector defined in the Theorem statement, and the operator $x \in \mathbb{C} \mapsto x\boldsymbol{\phi}_t$). From Claim 34:

$$\begin{aligned}
\frac{1}{T}\int_{t\in[0,T]} \|\varphi_t - \mathbf{Z}^*\boldsymbol{\phi}_t\|_\mu^2\, dt &= \operatorname{tr}\left(\frac{1}{T}\int_{t\in[0,T]} (\varphi_t - \mathbf{Z}^*\boldsymbol{\phi}_t) \otimes (\varphi_t - \mathbf{Z}^*\boldsymbol{\phi}_t)\, dt\right) \\
&= \operatorname{tr}\left(\frac{1}{T}\int_{t\in[0,T]} \varphi_t \otimes \varphi_t\, dt\right) + \operatorname{tr}\left(\frac{1}{T}\int_{t\in[0,T]} \mathbf{Z}^*\boldsymbol{\phi}_t \otimes \mathbf{Z}^*\boldsymbol{\phi}_t\, dt\right) \\
&\quad - \operatorname{tr}\left(\frac{1}{T}\int_{t\in[0,T]} \mathbf{Z}^*\boldsymbol{\phi}_t \otimes \varphi_t\, dt\right) - \operatorname{tr}\left(\frac{1}{T}\int_{t\in[0,T]} \varphi_t \otimes \mathbf{Z}^*\boldsymbol{\phi}_t\, dt\right)
\end{aligned}$$

We have,

$$\frac{1}{T}\int_{t\in[0,T]} \varphi_t \otimes \varphi_t \, dt = \mathcal{G}_\mu \,,$$

From Claim 33:

$$\frac{1}{T}\int_{t\in[0,T]} \mathbf{Z}^*\boldsymbol{\phi}_t \otimes \mathbf{Z}^*\boldsymbol{\phi}_t \, dt = \mathbf{Z}^*\left(\frac{1}{T}\int_{t\in[0,T]} \boldsymbol{\phi}_t \otimes \boldsymbol{\phi}_t \, dt\right)\mathbf{Z} = \mathbf{Z}^*\mathbf{K}\mathbf{Z} = \widehat{\mathcal{G}_\mu}$$

Next, consider $\frac{1}{T}\int_0^T \boldsymbol{\phi}_t \otimes \varphi_t \, dt$. For any α,

$$\frac{1}{T}\left(\int_0^T \boldsymbol{\phi}_t \otimes \varphi_t \, dt\right)\alpha = \frac{1}{T}\int_0^T \langle \varphi_t, \alpha\rangle_\mu \boldsymbol{\phi}_t \, dt$$

where the integral on the left is a weak vector integral. Since for every $g \in L_2(T)$,

$$\mathbf{C}_s g = \frac{1}{T}\int_0^T g(t)\boldsymbol{\phi}_t \, dt$$

and for every $\alpha \in L_2(\mu)$, $[\mathcal{F}_\mu^*\alpha](t) = \langle \varphi_t, \alpha\rangle_\mu$, we have $\frac{1}{T}\int_0^T \boldsymbol{\phi}_t \otimes \varphi_t \, dt = \mathbf{C}_s\mathcal{F}_\mu^*$, so

$$\frac{1}{T}\int_{t\in[0,T]} \mathbf{Z}^*\boldsymbol{\phi}_t \otimes \varphi_t \, dt = \mathbf{Z}^*\left(\frac{1}{T}\int_{t\in[0,T]} \boldsymbol{\phi}_t \otimes \varphi_t \, dt\right) = \mathbf{Z}^*\mathbf{C}_s\mathcal{F}_\mu^* = \mathbf{Z}^*\mathbf{K}\mathbf{Z} = \widehat{\mathcal{G}_\mu}\,.$$

Combining the previous observations, we find that

$$\frac{1}{T}\int_{t\in[0,T]} \|\varphi_t - \mathbf{Z}^*\boldsymbol{\phi}_t\|_\mu^2 \, dt = \mathrm{tr}(\mathcal{G}_\mu - \widehat{\mathcal{G}_\mu}).$$

Let $v_1, \ldots, v_{2s_{\mu,\epsilon}} \in L_2(\mu)$ be the eigenfunctions of $\mathcal{G}_\mu$ corresponding to its top $2s_{\mu,\epsilon}$ eigenvalues. Define $\mathbf{X} : L_2(\mu) \to \mathbb{C}^{2s_{\mu,\epsilon}}$ as: for $g \in L_2(\mu)$, $[\mathbf{X}g](j) = \langle v_j, g\rangle_\mu$. Note that

$$\begin{aligned}\mathrm{tr}(\widehat{\mathcal{G}_\mu} - \mathbf{X}^*\mathbf{X}\widehat{\mathcal{G}_\mu}\mathbf{X}^*\mathbf{X}) &= \mathrm{tr}(\mathbf{Z}^*\mathbf{C}_s\mathbf{C}_s\mathbf{Z} - \mathbf{X}^*\mathbf{X}\mathbf{Z}^*\mathbf{C}_s\mathbf{C}_s\mathbf{Z}\mathbf{X}^*\mathbf{X}) \\ &= \mathrm{tr}(\mathbf{C}_s\mathbf{Z}\mathbf{Z}^*\mathbf{C}_s - \mathbf{C}_s\mathbf{Z}\mathbf{X}^*\mathbf{X}\mathbf{Z}^*\mathbf{C}_s) \geq 0\end{aligned}$$

since $\mathbf{C}_s\mathbf{Z}\mathbf{Z}^*\mathbf{C}_s \succeq \mathbf{C}_s\mathbf{Z}\mathbf{X}^*\mathbf{X}\mathbf{Z}^*\mathbf{C}_s$ ($\mathbf{X}^*\mathbf{X}$ is a projection, so $\mathbf{X}^*\mathbf{X} \preceq \mathcal{I}_\mu$). So we can bound:

$$\begin{aligned}\frac{1}{T}\int_{t\in[0,T]} \|\varphi_t - \mathbf{Z}^*\boldsymbol{\phi}_t\|_\mu^2 \, dt = \mathrm{tr}(\mathcal{G}_\mu - \widehat{\mathcal{G}_\mu}) &\leq \mathrm{tr}(\mathcal{G}_\mu - \widehat{\mathcal{G}_\mu}) + \mathrm{tr}(\widehat{\mathcal{G}_\mu} - \mathbf{X}^*\mathbf{X}\widehat{\mathcal{G}_\mu}\mathbf{X}^*\mathbf{X}) \\ &= \mathrm{tr}(\mathcal{G}_\mu - \mathbf{X}^*\mathbf{X}\mathcal{G}_\mu\mathbf{X}^*\mathbf{X}) + \mathrm{tr}(\mathbf{X}^*\mathbf{X}(\mathcal{G}_\mu - \widehat{\mathcal{G}_\mu})\mathbf{X}^*\mathbf{X}). \quad (75)\end{aligned}$$

Let i_ϵ be the smallest i with $\lambda_i(\mathcal{G}_\mu) \leq \epsilon$. We have:

$$s_{\mu,\epsilon} = \sum_{i=1}^{\infty} \frac{\lambda_i(\mathcal{G}_\mu)}{\lambda_i(\mathcal{G}_\mu) + \epsilon} \geq \sum_{i=1}^{i+\epsilon} \frac{\lambda_i(\mathcal{G}_\mu)}{\lambda_i(\mathcal{G}_\mu) + \epsilon} \geq \frac{i_\epsilon}{2}.$$

Thus we can bound $\mathrm{tr}(\mathcal{G}_\mu - \mathbf{X}^*\mathbf{X}\mathcal{G}_\mu\mathbf{X}^*\mathbf{X})$ as:

$$\mathrm{tr}(\mathcal{G}_\mu - \mathbf{X}^*\mathbf{X}\mathcal{G}_\mu\mathbf{X}^*\mathbf{X}) = \sum_{i=2s_{\mu,\epsilon}+1}^{\infty} \lambda_i(\mathcal{G}_\mu) \leq \sum_{i=i_\epsilon+1}^{\infty} \lambda_i(\mathcal{G}_\mu) \leq 2\epsilon s_{\mu,\epsilon}. \quad (76)$$

where the last bound follows from the fact that $s_{\mu,\epsilon} \geq \sum_{i=i_\epsilon+1}^{\infty} \frac{\lambda_i(\mathcal{G}_\mu)}{\lambda_i(\mathcal{G}_\mu)+\epsilon} \geq \sum_{i=i_\epsilon+1}^{\infty} \frac{\lambda_i(\mathcal{G}_\mu)}{2\epsilon}$.

We can also bound $\operatorname{tr}(\mathbf{X}^*\mathbf{X}(\mathcal{G}_\mu - \widehat{\mathcal{G}_\mu})\mathbf{X}^*\mathbf{X})$ using Lemma 47. Since $\mathcal{G}_\mu \leq \widehat{\mathcal{G}_\mu} + \epsilon\mathcal{I}_\mu$ we have:

$$\operatorname{tr}(\mathbf{X}^*\mathbf{X}(\mathcal{G}_\mu - \widehat{\mathcal{G}_\mu})\mathbf{X}^*\mathbf{X}) \leq \epsilon \operatorname{tr}(\mathbf{X}^*\mathbf{X}\mathbf{X}^*\mathbf{X}) = \epsilon s_{\mu,\epsilon}. \tag{77}$$

Plugging (76) and (77) back into (75) we have:

$$\frac{1}{T}\int_{t\in[0,T]} \|\varphi_t - \mathbf{Z}^*\boldsymbol{\phi}_t\|_\mu^2 \, dt \leq 4\epsilon \cdot s_{\mu,\epsilon},$$

which completes the theorem. □

D Tight Statistical Dimension Bound for Bandlimited Functions

In Section 5 we demonstrate, perhaps surprisingly, that a simple function $\tilde{\tau}_{\mu,\epsilon}(t)$ (defined in Theorem 17) exists for *any* μ that upper bounds $\tau_{\mu,\epsilon}(t)$ and has $\tilde{s}_{\mu,\epsilon} = \tilde{O}(s_{\mu,\epsilon})$. Combined with Theorem 7 this yields our main algorithmic result Theorem 3, which shows that we can achieve $O\left(s_{\mu,\epsilon}\log^2(s_{\mu,\epsilon}\right)$ sample complexity with just $\tilde{O}\left(s_{\mu,\epsilon}^{\omega}\right)$ runtime.

Instantiating Theorem 3 using the approximate ridge leverage function of Theorem 17 requires an upper bound on $s_{\mu,\epsilon}$. In this section we show how to bound $s_{\mu,\epsilon}$ when μ is uniform measure on some interval – i.e., when our interpolation problem is over bandlimited functions. In Section E we leverage this result to bound $s_{\mu,\epsilon}$ for a number of other important priors, including for multiband, Gaussian, and Cauchy-Lorentz.

Beyond letting us upper bound $s_{\mu,\epsilon}$ to apply Theorem 3, our proof for bandlimited functions is constructive, giving a simple upper bound for $\tau_{\mu,\epsilon}(t)$ for any t. This upper bound can be plugged directly into Algorithm 1 and Theorem 7 to give a tightening of Theorem 3 by a logarithmic factor in the bandlimited case. Like our general result, the proof is based on the definition of leverage scores given in (11). This definition makes it clear that, to upper bound $\tau_{\mu,\epsilon}(t)$, it suffices to show that a function with Fourier support controlled by μ cannot "spike" too extremely at time t.

For bandlimited functions, we obtain a smoothness bound by introducing and applying a Bernstein type smoothness bound for low-degree polynomials and relying on the fact that any bandlimited function is well approximated by a low-degree polynomial. This approach mirrors the general proof in Section 5, which uses a more sophisticated smoothness bound for sparse Fourier functions.

Our result for bandlimited function is as follows:

Theorem 48. *Let μ be the uniform measure on $[-F, F]$. Let $q = \lceil 16\pi eFT + 2\log(1/\epsilon) + 11\rceil$. For all $t \in [0, T]$, let the approximate ridge leverage function $\tilde{\tau}_{\mu,\epsilon}$ equal:*

$$\tilde{\tau}_{\mu,\epsilon}(t) = \frac{1}{T}\left(4 + \frac{q}{\sqrt{\min(t, T-t)/T}}\right).$$

For any $\epsilon \leq 1, F, T$, $\tilde{\tau}_{\mu,\epsilon}(t)$ satisfies:

1. $\tilde{\tau}_{\mu,\epsilon}(t) \geq \tau_{\mu,\epsilon}(t)$.
2. $\int_0^T \tilde{\tau}_{\mu,\epsilon}(t)dt \stackrel{\text{def}}{=} \tilde{s}_{\mu,\epsilon} = O\left(FT + \log(1/\epsilon)\right).$

Thus we have $s_{\mu,\epsilon} \leq s_{\tilde{\mu},\epsilon} = O\left(FT + \log(1/\epsilon)\right)$.

Combined with Theorem 7, Theorem 48 immediately gives:

Corollary 49. *Let μ be the uniform measure on $[-F, F]$. Using $\tilde{\tau}_{\mu,\epsilon}$ as defined in Theorem 48, Algorithm 1 returns $t_1, \ldots, t_s \in [0, T]$ and $\mathbf{z} \in \mathbb{C}^s$ such that $\tilde{y}(t) = \sum_{i=1}^{s} \mathbf{z}(i) \cdot k_\mu(t_i, t)$ satisfies with probability $\geq 1 - \delta$:*

$$\|\tilde{y} - y\|_T^2 \leq 6\epsilon \|x\|_\mu^2 + 7\|n\|_T^2.$$

The algorithm queries $y + n$ at s points and runs in $O(s^\omega)$ time where $s = O([FT + \log(1/\epsilon)] \cdot [\log(FT + \log(1/\epsilon)) + 1/\delta])$. The output $\tilde{y}(t)$ can be evaluated using Algorithm 2 in $O(s)$ time.

Proof. The corollary follows immediately from Theorem 7 after noting that

- $Z = O(1)$ since, as shown in Appendix F, $k_\mu(t_1, t_2) = \frac{\sin(2\pi F(t_1 - t_2))}{2\pi F(t_1 - t_2)}$ and so can be computed in $O(1)$ arithmetic operations.
- $W = O(1)$ since to sample points proportional to $\tilde{\tau}_{\mu,\epsilon}(t)$, we must sample a mixture of the uniform distribution and the distribution with density proportional to $\frac{1}{\sqrt{\min(t, T-t)/T}}$. It suffices to show that we can sample from the later in $O(1)$ time, and in fact that we can sample $t \in [0, 1/2]$ with probability proportional to $\frac{1}{\sqrt{t}}$ in $O(1)$ time, since we can then symmetrize and scale this distribution. We can accomplish this with inverse transform sampling. Our density is $p(t) = \frac{1}{2\sqrt{2t}}$ and so its cumulative distribution function is $C(t) = \sqrt{t/2}$. Thus we can sample z uniformly in $[0, 1]$ and return $C^{-1}(z) = 2z^2$, which will be a sample from the desired distribution. This can be done with $O(1)$ arithmetic computations.

□

D.1 Smoothness bounds for polynomials

As mentioned our main techniques tool is a Bernstein type smoothness bounds for low-degree polynomials. In general, low-degree polynomials are smoother than high-degree polynomials, and thus cannot spike as sharply. There are a number of ways to formalize this statement. The well known Markov brother's inequality and Bernstein inequality bound the maximum derivative of a polynomial by a function of the polynomial's degree and it's maximum value on an interval.

To bound leverage scores, we are interested in a slightly different metric of smoothness. In particular, we need to bound the maximum squared value of a polynomial by its average squared value on $[0, T]$. We can use standard properties of the Legendre polynomials to prove: s

Claim 50. *For any degree d polynomial $p(\cdot)$ with complex coefficients and $t \in [0, T]$, let $r = \frac{\min(t, T-t)}{T}$. Then:*

$$|p(t)|^2 \leq \frac{d+1}{\sqrt{r}} \cdot \frac{1}{T} \int_0^T |p(t)|^2 \, dt.$$

This bound is tighter for points near the center of the interval $[0, T]$ and goes to infinity near the edges. Using the Markov brother's inequality, it's possible to obtain a fixed up bound of $O(d^2)$, which is tighter for small values of r. However, this won't be necessary for our purposes. We note that, when $t = T/2$, the upper bound on $p(t)^2$ improves to $O(d)$ times the average squared value of p, quadratically better than an $O(d^2)$ bound. This improvement is nearly optimal: the upper bound of Claim 50 is matched up to a logarithmic factor by an appropriately scaled and shifted Chebyshev polynomial of the first kind applied to $[T/2 - t]^2$ (see e.g. [FMMS16] for a construction).

Proof of Claim 50. The claim follows from properties of the standard orthogonal Legendre polynomials, which are denoted $P_0, P_1, \ldots$ and defined via the recurrence relation:

$$\begin{aligned} P_0(x) &= 1 \\ P_1(x) &= x \\ &\vdots \\ P_k(x) &= \frac{2k-1}{k} x \cdot P_{k-1}(x) - \frac{n-1}{n} \cdot P_{k-2}(x). \end{aligned}$$

The Legendre polynomials are orthogonal over the interval $[-1, 1]$ with respect to the constant weight function. In particular, they satisfy

$$\int_{-1}^{1} P_j(x) P_k(x)\, dx = \frac{2}{2j+1} \delta_{j,k}, \tag{78}$$

where $\delta_{m,n}$ is the Kronecker delta function. Additionally, for $x \in [-1, 1]$, $|P_j(x)| \leq 1$ for all j.

Using these facts we can show that for any degree d polynomial $p(\cdot)$, interval $[a, b]$, and $x \in [a, b]$:

$$|p(x)|^2 \leq \frac{d+1}{\sqrt{r}} \cdot \frac{\int_a^b |p(t)|^2\, dt}{(b-a)},$$

where $r = \frac{\min(|a-x|, |b-x|)}{(b-a)}$. Setting $a = 0$ and $b = T$ gives the claim.

We begin by noting that, without loss of generality, we can assume that $a = -1$ and $b = 1$. In particular, shift and stretch $p(x)$ by defining $g(x) = p\left(\frac{2(x-a)}{b-a} - 1\right)$. g has degree d and the maximum of $|g(x)|^2$ for $x \in [-1, 1]$ is the same as the maximum of $|p(x)|^2$ for $x \in [a, b]$. Additionally, $\frac{\int_{-1}^{1} |g(t)|^2\, dt}{2} = \frac{\int_a^b |p(t)|^2\, dt}{(b-a)}$. Accordingly, to prove the claim it suffices to prove that, for any degree d polynomial g,

$$\max_{x \in [-1,1]} |g(x)|^2 \leq \frac{d+1}{\sqrt{r}} \cdot \frac{\int_{-1}^{1} |g(t)|^2\, dt}{2}. \tag{79}$$

Our proof depends on a Bernstein type inequality for Legendre polynomials, which can be found in [Lor83]. Specifically, for all $j = 0, 1, 2, \ldots$ and any $x \in [-1, 1]$ it holds that:

$$P_j(x)^2 \leq \frac{2}{\pi(j + 1/2)} \frac{1}{\sqrt{1 - x^2}}. \tag{80}$$

Writing g in the Legendre basis:

$$g(x) = \sum_{j=0}^{d} c_j P_j(x),$$

we have from (80) that

$$|g(x)| \leq \sum_{j=0}^{d} |c_j| \left(\frac{2}{\pi(j + 1/2)} \frac{1}{\sqrt{1 - x^2}} \right)^{1/2}$$

and thus

$$
\begin{aligned}
|g(x)|^2 &\le (d+1)\sum_{j=0}^{d}|c_j|^2\frac{2}{\pi(j+1/2)}\frac{1}{\sqrt{1-x^2}} \\
&= \frac{2}{\pi}\frac{(d+1)}{\sqrt{1-x^2}}\sum_{j=0}^{d}|c_j|^2\frac{2}{2j+1} \\
&= \frac{2}{\pi}\frac{(d+1)}{\sqrt{1-x^2}}\int_{-1}^{1}|g(t)|^2\,dt. \qquad (81)
\end{aligned}
$$

The last equality step follows from (78). Finally, let $q = \min\left(|-1-x|, |1-x|\right)$ and note that

$$
\frac{1}{\sqrt{1-x^2}} = \frac{1}{\sqrt{1-(1-q)^2}} \le \frac{1}{\sqrt{q}}.
$$

As defined, $r = q/2$ Plugging into (81) we have a final bound of

$$
|g(x)|^2 \le \frac{4}{\pi}\frac{(d+1)}{\sqrt{2r}}\frac{\int_{-1}^{1}|g(t)|^2\,dt}{2} < \frac{(d+1)}{\sqrt{r}}\frac{\int_{-1}^{1}|g(t)|^2\,dt}{2},
$$

which establishes (79) and thus the claim. □

D.2 Smoothness bounds for bandlimited functions

With Claim 50 in place, we are now ready to prove our main result for bandlimited functions.

Proof of Theorem 48. Following Definition 3, our goal is to choose $\tilde{\tau}_{\mu,\epsilon}$ to satisfy:

$$
\tilde{\tau}_{\mu,\epsilon}(t) \ge \frac{1}{T}\cdot\frac{|[\mathcal{F}_\mu\alpha](t)|^2}{\|\mathcal{F}_\mu\alpha\|_T^2 + \epsilon\|\alpha\|_\mu^2}. \qquad (82)
$$

for any α. Let $z = \mathcal{F}_\mu\alpha$. Expanding $e^{-2i\pi\xi t}$ using its Maclaurin series and letting d be some degree parameter that we will fix later, we write z as the sum of two functions, a and b:

$$
\begin{aligned}
z(t) &= \frac{1}{2F}\int_{-F}^{F}\alpha(\xi)e^{-2i\pi\xi t}\,d\xi \\
&= \sum_{j=0}^{\infty}\frac{1}{2F}\int_{-F}^{F}\alpha(\xi)\frac{(-2\pi i\xi)^j}{j!}t^j\,d\xi \\
&= \sum_{j=0}^{d}\left(\frac{1}{2F}\int_{-F}^{F}\alpha(\xi)\frac{(-2\pi i\xi)^j}{j!}\,d\xi\right)t^j + \sum_{j=d+1}^{\infty}\frac{1}{2F}\int_{-F}^{F}\alpha(\xi)\frac{(-2\pi i\xi)^j}{j!}t^j\,d\xi \\
&\stackrel{\text{def}}{=} a(t) + b(t). \qquad (83)
\end{aligned}
$$

Note that a is a degree d polynomial with complex coefficients. So by Claim 50,

$$
|a(t)|^2 \le \frac{d+1}{\sqrt{\min(t, T-t)/T}}\cdot\|a\|_T^2. \qquad (84)
$$

Turning our attention to b, we see that:

$$|b(t)| = \left| \sum_{j=d+1}^{\infty} \frac{1}{2F} \int_{-F}^{F} \alpha(\xi) \frac{(-2\pi i \xi)^j}{j!} t^j \, d\xi \right| \leq \sum_{j=d+1}^{\infty} \frac{(2\pi FT)^j}{j!} \left| \frac{1}{2F} \int_{-F}^{F} \alpha(\xi) \, d\xi \right|$$
$$\leq \sum_{j=d+1}^{\infty} \frac{(2\pi FT)^j}{j!} \sqrt{\frac{1}{2F} \int_{-F}^{F} 1 \, d\xi} \sqrt{\|\alpha\|_\mu^2} = \sum_{j=d+1}^{\infty} \frac{(2\pi FT)^j}{j!} \cdot \|\alpha\|_\mu. \tag{85}$$

The second to last step uses Cauchy-Schwarz inequality. Finally using that for all j, $j! \geq (j/e)^j$, for any $d \geq 4\pi eFT$:

$$\begin{aligned} \sum_{j=d+1}^{\infty} \frac{(2\pi FT)^j}{j!} &\leq \sum_{j=d+1}^{\infty} \left(\frac{2\pi eFT}{j} \right)^j \\ &\leq \sum_{j=d+1}^{\infty} \left(\frac{2\pi eFT}{d+1} \right)^j \\ &\leq \sum_{j=d+1}^{\infty} \left(\frac{1}{2} \right)^j = \frac{1}{2^d}. \end{aligned} \tag{86}$$

So, if we take $d = \lceil 4\pi eFT + \log(1/\epsilon)/2 + 1 \rceil$, it follows from (85) and (86) that

$$|b(t)| \leq \frac{1}{2^d} \cdot \|\alpha\|_\mu \leq \frac{1}{2^{\lceil \log(1/\epsilon)/2 \rceil + 1}} \cdot \|\alpha\|_\mu \leq \frac{\sqrt{\epsilon}}{2} \cdot \|\alpha\|_\mu.$$

It follows that $\|b\|_T \leq \frac{\sqrt{\epsilon}}{2} \|\alpha\|_\mu$. Using the decomposition of (83) and the fact that for any real nonnegative c, d, $c^2 + d^2 \leq (c+d)^2$, and for any complex e, f, $|e+f|^2 \leq 2|e|^2 + 2|f|^2$:

$$\begin{aligned} \frac{|z(t)|^2}{\|z\|_T^2 + \epsilon \|\alpha\|_\mu^2} &\leq \frac{|a(t) + b(t)|^2}{(\|a\|_T - \|b\|_T)^2 + \epsilon \|\alpha\|_\mu^2} \\ &\leq \frac{2|a(t)|^2 + 2|b(t)|^2}{\frac{1}{2}(\|a\|_T - \|b\|_T + \sqrt{\epsilon}\|\alpha\|_\mu)^2} \\ &\leq \frac{4|a(t)|^2 + 4|b(t)|^2}{(\|a\|_T + \frac{\sqrt{\epsilon}}{2}\|\alpha\|_\mu)^2} \\ &\leq \frac{4|a(t)|^2 + \epsilon \|\alpha\|_\mu^2}{\|a\|_T^2 + \frac{\epsilon}{4}\|\alpha\|_\mu^2}. \end{aligned}$$

It follows from (84) that:

$$\begin{aligned} \frac{|z(t)|^2}{\|z\|_T^2 + \epsilon \|\alpha\|_\mu^2} &\leq \max\left(\frac{4|a(t)|^2}{\|a\|_T^2}, 4 \right) \\ &\leq \frac{4(d+1)}{\sqrt{\min(t, T-t)/T}} + 4. \end{aligned}$$

In Theorem 48 we set $q = \lceil 16\pi eFT + 2\log(1/\epsilon) + 11 \rceil$. We have $q \geq 4 \cdot \lceil 4\pi eFT + \log(1/\epsilon)/2 + 2 \rceil = 4(d+1)$ since, for any x, $\lceil 4x + 3 \rceil \geq 4\lceil x \rceil$. Recalling that $z = \mathcal{F}_\mu \alpha$, it follows $\tilde{\tau}_{\mu,\epsilon}$ defined in that

theorem satisfies (82) for any α. It remains to bound the total measure of our approximate ridge leverage function, $\tilde{s}_{\mu,\epsilon}$. To do so, note that:

$$\tilde{s}_{\mu,\epsilon} = \frac{2}{T}\int_0^{T/2} \frac{q}{\sqrt{t/T}} + 4\, dt.$$

We can compute:

$$\frac{2}{T}\int_0^{T/2} \frac{q}{\sqrt{t/T}} + 4\, dt = 2\int_0^{1/2} \frac{q}{\sqrt{t}} + 4\, dt = 2\sqrt{2}q + 4 = O(FT + \log(1/\epsilon)).$$

This bound establishes the theorem. □

E Statistical dimension for common Fourier constraints

In this section we leverage Theorem 48 to give upper bounds on the statistical dimensions of a number common priors μ used for Fourier constrained interpolation, including multiband, Gaussian, and Cauchy-Lorentz priors. We start by giving two simple lemmas that we use to translate our bound for bandlimited functions to these more general priors.

Lemma 51 (Statistical Dimension of Sum of Measures)**.** *Let $\mu_1, \mu_2, \cdots \mu_s$ be finite measures on $\mathbb{R}$. Let μ be a probability measure defined by $\mu = \mu_1 + \mu_2 + \cdots + \mu_s$.*

$$s_{\mu,\epsilon} \leq \sum_{i=1}^{s} s_{\mu_i,\epsilon}.$$

Proof. We can see from Definition 2 that for $\mu = \mu_1 + \ldots + \mu_s$ the kernel operator $\mathcal{K}_\mu$ satisfies $\mathcal{K}_\mu = \sum_{i=1}^s \mathcal{K}_{\mu_i}$. We can thus bound:

$$\begin{aligned} s_{\mu,\epsilon} = \mathrm{tr}(\mathcal{K}_\mu(\mathcal{K}_\mu + \epsilon\mathcal{I}_T)^{-1}) &= \sum_{i=1}^{s} \mathrm{tr}(\mathcal{K}_{\mu_i}(\mathcal{K}_\mu + \epsilon\mathcal{I}_T)^{-1}) \\ &\leq \sum_{i=1}^{s} \mathrm{tr}(\mathcal{K}_{\mu_i}(\mathcal{K}_{\mu_i} + \epsilon\mathcal{I}_T)^{-1}) \\ &= \sum_{i=1}^{s} s_{\mu_i,\epsilon}. \end{aligned}$$

The second to last inequality follows since $0 \preceq \mathcal{K}_{\mu_i} \preceq \mathcal{K}_\mu$, so $0 \prec \mathcal{K}_{\mu_i} + \epsilon\mathcal{I}_T \preceq \mathcal{K}_\mu + \epsilon\mathcal{I}_T$ and $(\mathcal{K}_\mu + \epsilon\mathcal{I}_T)^{-1} \preceq (\mathcal{K}_{\mu_i} + \epsilon\mathcal{I}_T)^{-1}$ by Claim 30. Letting e_1, e_2 be an orthonormal basis for $L_2(T)$, we thus have:

$$\begin{aligned} \mathrm{tr}(\mathcal{K}_{\mu_i}(\mathcal{K}_{\mu_i} + \epsilon\mathcal{I}_T)^{-1}) &= \mathrm{tr}(\mathcal{K}_{\mu_i}^{1/2}(\mathcal{K}_{\mu_i} + \epsilon\mathcal{I}_T)^{-1}\mathcal{K}_{\mu_i}^{1/2}) \quad \text{(By cyclic property of the trace, Claim 28.)} \\ &= \sum_{i=1}^{\infty} \langle \mathcal{K}_{\mu_i}^{1/2} e_i, (\mathcal{K}_{\mu_i} + \epsilon\mathcal{I}_T)^{-1}\mathcal{K}_{\mu_i}^{1/2} e_i\rangle_T \\ &\geq \sum_{i=1}^{\infty} \langle \mathcal{K}_{\mu_i}^{1/2} e_i, (\mathcal{K}_{\mu} + \epsilon\mathcal{I}_T)^{-1}\mathcal{K}_{\mu_i}^{1/2} e_i\rangle_T \\ &= \mathrm{tr}(\mathcal{K}_{\mu_i}^{1/2}(\mathcal{K}_{\mu} + \epsilon\mathcal{I}_T)^{-1}\mathcal{K}_{\mu_i}^{1/2}) \\ &= \mathrm{tr}(\mathcal{K}_{\mu_i}(\mathcal{K}_{\mu} + \epsilon\mathcal{I}_T)^{-1}). \end{aligned}$$

This completes the lemma. □

Lemma 52 (Statistical Dimension of Scaled Measures)**.** *Let μ be a measure on $\mathbb{R}$. For any parameter $\gamma > 0$, we have:*

$$s_{\mu,\epsilon} = s_{(\mu/\gamma),(\epsilon/\gamma)}.$$

Proof. From Definition 2, we can see that $\mathcal{K}_{(\mu/\gamma)} = \frac{1}{\gamma}\mathcal{K}_\mu$ and thus has eigenvalues equal to $\lambda_1(\mathcal{K}_\mu)/\gamma, \lambda_2(\mathcal{K}_\mu)/\gamma, \ldots$ We can thus compute:

$$\begin{aligned} s_{(\mu/\gamma),(\epsilon/\gamma)} &= \sum_{i=1}^{\infty} \frac{\lambda_i(\mathcal{K}_{(\mu/\gamma)})}{\lambda_i(\mathcal{K}_{(\mu/\gamma)}) + \epsilon/\gamma} \\ &= \sum_{i=1}^{\infty} \frac{\lambda_i(\mathcal{K}_\mu)/\gamma}{\lambda_i(\mathcal{K}_\mu)/\gamma + \epsilon/\gamma} \\ &= \sum_{i=1}^{\infty} \frac{\lambda_i(\mathcal{K}_\mu)}{\lambda_i(\mathcal{K}_\mu) + \epsilon} \\ &= s_{\mu,\epsilon}. \end{aligned}$$

□

We now use Lemmas 51 and 52 to prove our statistical dimension bounds. We first start with multiband Fourier constraints, showing that the statistical dimension is roughly proportional to the total length of all the frequency bands times the time domain window size, intuitively matching the Landau rate for asymptotic recovery of multiband functions [Lan67a].

Theorem 53 (Multiband Statistical Dimension)**.** *Consider a set of s disjoint frequency bands, $I_1, I_2, \cdots, I_s$, and suppose that the length of the band I_i is denoted by F_i. Let μ be the measure which induces a uniform probability density on $I_1 \cup I_2 \cup \cdots \cup I_s$. We have:*

$$s_{\mu,\epsilon} = O\left(\sum_{i=1}^{s} F_i T + s\log(1/\epsilon)\right).$$

Proof. For every i, let μ_i be the measure defined by $\mu_i(A) = \mu(A \cap I_i)$. Note that we have $\mu = \sum_i \mu_i$ and so can invoke Lemma 51, giving:

$$s_{\mu,\epsilon} \leq \sum_{i=1}^{s} s_{\mu_i,\epsilon}. \tag{87}$$

If μ_i gave a uniform probability measure on frequency band I_i (i.e., if we had $\mu_i(\mathbb{R}) = 1$), we could use the result of Theorem 48 to bound $s_{\mu_i,\epsilon} = O(F_i T + \log(1/\epsilon))$. This is not the case, but we can instead let $\gamma_i \stackrel{\text{def}}{=} \mu_i(\mathbb{R}) \leq 1$. By Lemma 52,

$$s_{\mu_i,\epsilon} = s_{(\mu_i/\gamma_i),(\epsilon/\gamma_i)}.$$

Now μ_i/γ_i is a uniform probability measure on I_i, so we can invoke Theorem 48 giving:

$$s_{\mu_i,\epsilon} = s_{(\mu_i/\gamma_i),(\epsilon/\gamma_i)} = O\left(F_i T + \log(\gamma_i/\epsilon)\right).$$

Plugging this bound in (87) and using that $\gamma_i \leq 1$ we obtain:

$$s_{\mu,\epsilon} = O\left(\sum_{i=1}^{s} F_i T + \log(\gamma_i/\epsilon)\right) = O\left(\sum_{i=1}^{s} F_i T + s\log(1/\epsilon)\right),$$

completing the theorem. □

We next bound the statistical dimension of Gaussian measure.

Theorem 54 (Gaussian Statistical Dimension)**.** *Let μ induce the Gaussian probability distribution with standard deviation F defined by $d\mu(\xi) = \frac{1}{\sqrt{2\pi F^2}} e^{-\xi^2/2F^2} d\xi$. We have:*

$$s_{\mu,\epsilon} = O\left(FT\sqrt{\log(1/\epsilon)} + \log(1/\epsilon)\right).$$

Proof. Let I_h be the interval defined by $I_h = \{\xi \in \mathbb{R} : |\xi| \leq F\sqrt{\log(1/\epsilon)}\}$. We decompose μ into two measures μ_h and μ_t as follows:

$$\mu_h(A) = \mu(A \cap I_h) \text{ and } \mu_t(A) = \mu(A - A \cap I_h).$$

We can see that $\mu = \mu_h + \mu_t$ and so by Lemma 51, $s_{\mu,\epsilon} \leq s_{\mu_t,\epsilon} + s_{\mu_h,\epsilon}$. For μ_t we have:

$$\begin{aligned}\operatorname{tr}(\mathcal{K}_{\mu_t}) = \mu_t(\mathbb{R}) &= \frac{1}{\sqrt{2\pi F^2}} \int_{|\xi| > F\sqrt{\log(1/\epsilon)}} e^{-\xi^2/2F^2}\, d\xi \\ &= 1 - \operatorname{erf}(\sqrt{\log(1/\epsilon)}) \leq 2\epsilon,\end{aligned}$$

where the last bound follows from a Chernoff bound, giving $1 - \operatorname{erf}(x) \leq 2e^{-x^2}$ [Wai18]. This lets us crudely bound:

$$s_{\mu_t,\epsilon} = \operatorname{tr}(\mathcal{K}_{\mu_t}(\mathcal{K}_{\mu_t} + \epsilon\mathcal{I}_T)^{-1}) \leq \operatorname{tr}(\mathcal{K}_{\mu_t})/\epsilon \leq 2, \tag{88}$$

where the first ineguality is because $\|(\mathcal{K}_{\mu_t} + \epsilon\mathcal{I}_T)^{-1}\|_{\text{op}} \leq 1/\epsilon$.

We next bound the statistical dimension of μ_h. Let $\tilde{\mu}_h$ be a uniform measure on I_h, with $d\mu(\xi) = \frac{1}{\sqrt{2\pi F^2}} d\xi$ for all $\xi \in I_h$. Note that $d\tilde{\mu}_h(\xi) \geq d\mu_h(\xi)$ for all $\xi \in I_h$ which gives that $K_{\mu_h} \preceq K_{\tilde{\mu}_h}$ and so $s_{\mu_h,\epsilon} \leq s_{\tilde{\mu}_h,\epsilon}$.

Let $\gamma \stackrel{\text{def}}{=} \tilde{\mu}_h(\mathbb{R}) = \sqrt{\frac{2\log(1/\epsilon)}{\pi}}$. By Lemma 52, $s_{\tilde{\mu},\epsilon} = s_{(\tilde{\mu}/\gamma),(\epsilon/\gamma)}$. Since $\tilde{\mu}/\gamma$ is a uniform probability measure on I_h, we can invoke Theorem 48 to give:

$$\begin{aligned}s_{\mu_h,\epsilon} \leq s_{\tilde{\mu}_h,\epsilon} = s_{(\tilde{\mu}_h/\gamma),(\epsilon/\gamma)} &= O\left(FT\sqrt{\log(1/\epsilon)} + \log(\gamma/\epsilon)\right) \\ &= O\left(FT\sqrt{\log(1/\epsilon)} + \log(1/\epsilon)\right),\end{aligned} \tag{89}$$

where the last equality follows from the fact that $\gamma = O(\sqrt{\log(1/\epsilon)})$. Combining (88) and (89) and applying Lemma 51 we have:

$$\begin{aligned}s_{\mu,\epsilon} &\leq s_{\mu_t,\epsilon} + s_{\mu_h,\epsilon} \\ &= 2 + O\left(FT\sqrt{\log(1/\epsilon)} + \log(1/\epsilon)\right) \\ &= O\left(FT\sqrt{\log(1/\epsilon)} + \log(1/\epsilon)\right),\end{aligned}$$

which completes the theorem. □

Finally, we bound the statistical dimension of the Cauchy-Lorentz measure.

Theorem 55. *Let μ induce the Cauchy-Lorentz probability distribution with scale parameter F defined by $d\mu(\xi) = \frac{1}{\pi F\left[1+\left(\frac{\xi}{F}\right)^2\right]} d\xi$. We have:*

$$s_{\mu,\epsilon} = O\left(\frac{FT}{\sqrt{\epsilon}} + \frac{1}{\sqrt{\epsilon}}\right).$$

Proof. As in the proof of Theorem 54 we define I_h to be the interval $I_h = \{\xi \in \mathbb{R} : |\xi| \leq F/\sqrt{\epsilon}\}$. We decompose μ into two measures μ_h and μ_t as follows:

$$\mu_h(A) = \mu(A \cap I_h) \text{ and } \mu_t(A) = \mu(A - A \cap I_h).$$

We have $\mu = \mu_h + \mu_t$ by Lemma 51, $s_{\mu,\epsilon} \leq s_{\mu_t,\epsilon} + s_{\mu_h,\epsilon}$. For μ_t we have:

$$\begin{aligned}
\operatorname{tr}(\mathcal{K}_{\mu_t}) = \mu_t(\mathbb{R}) &= \frac{1}{\pi F} \int_{|\xi|>F/\sqrt{\epsilon}} \frac{1}{1 + \left(\frac{\xi}{F}\right)^2} d\xi \\
&= \frac{2}{\pi} \int_{1/\sqrt{\epsilon}}^{\infty} \frac{1}{1+\xi^2} d\xi \\
&\leq \frac{2}{\pi} \int_{1/\sqrt{\epsilon}}^{\infty} \frac{1}{\xi^2} d\xi = \frac{2\sqrt{\epsilon}}{\pi}.
\end{aligned}$$

As in (88) we can thus bound:

$$s_{\mu_t,\epsilon} \leq \operatorname{tr}(\mathcal{K}_{\mu_t})/\epsilon = O(1/\sqrt{\epsilon}). \tag{90}$$

We next bound the statistical dimension of μ_h. Let $\tilde{\mu}_h$ be a uniform measure on I_h with $d\mu(\xi) = \frac{1}{\pi F}$ for all $\xi \in I_h$. As in the proof of Theorem 54, $d\tilde{\mu}_h(\xi) \geq d\mu_h(\xi)$ for all $\xi \in I_h$ which gives that $K_{\mu_h} \preceq K_{\tilde{\mu}_h}$ and so $s_{\mu_h,\epsilon} < s_{\tilde{\mu}_h,\epsilon}$.

Let $\gamma \stackrel{\text{def}}{=} \tilde{\mu}_h(\mathbb{R}) = \frac{2}{\pi\sqrt{\epsilon}}$. By Lemma 52, $s_{\tilde{\mu},\epsilon} = s_{(\tilde{\mu}/\gamma),(\epsilon/\gamma)}$. Since $\tilde{\mu}/\gamma$ is a uniform probability measure on I_h, we can invoke Theorem 48 to give:

$$\begin{aligned}
s_{\mu_h,\epsilon} \leq s_{\tilde{\mu}_h,\epsilon} = s_{(\tilde{\mu}_h/\gamma),(\epsilon/\gamma)} &= O\left(\frac{FT}{\sqrt{\epsilon}} + \log(\gamma/\epsilon)\right) \\
&= O\left(\frac{FT}{\sqrt{\epsilon}} + \log(1/\epsilon)\right),
\end{aligned} \tag{91}$$

where the last equality follows from the fact that $\gamma = O(1/\sqrt{\epsilon})$. Combining (90) and (91) and applying Lemma 51 we have:

$$s_{\mu,\epsilon} \leq s_{\mu_t,\epsilon} + s_{\mu_h,\epsilon} = O\left(\frac{1}{\sqrt{\epsilon}} + \frac{FT}{\sqrt{\epsilon}} + \log(1/\epsilon)\right) = O\left(\frac{FT}{\sqrt{\epsilon}} + \frac{1}{\sqrt{\epsilon}}\right),$$

which completes the theorem. □

F Kernel computation for common Fourier constraints

Algorithm 1 and the corresponding Theorem 3 assumes the ability to compute the kernel function $k_\mu(t_1, t_2) = \int_{\xi\in\mathbb{R}} e^{-2\pi i(t_1-t_2)} d\mu(\xi)$. In this section we give close forms for the kernel functions of popular measures μ, including all those whose statistical dimension we bound in Appendix E.

Bandlimited Fourier Constraint: When μ is the uniform measure on frequencies in $[-F, F]$,

$$\begin{aligned}
k_\mu(t_1, t_2) &= \int_{\xi\in\mathbb{R}} e^{-2\pi i(t_1-t_2)\xi} d\mu(\xi) \\
&= \frac{1}{2F} \int_{-F}^{F} e^{-2\pi i(t_1-t_2)\xi} d\xi \\
&= \frac{\sin(2\pi F(t_1-t_2))}{2\pi F(t_1-t_2)}.
\end{aligned}$$

So, k_μ is the *sinc kernel*.

Multiband Fourier Constraint: Consider a set of s disjoint frequency bands, $I_1, I_2, \cdots, I_s$, where $I_j = [c_j - F_j, c_j + F_j]$. Let μ be the uniform measure on $I_1 \cup I_2 \cup \cdots \cup I_s$. Then we have:

$$\begin{aligned} k_\mu(t_1, t_2) &= \int_{\xi \in \mathbb{R}} e^{-2\pi i(t_1 - t_2)\xi} d\mu(\xi) \\ &= \frac{1}{2\sum_j F_j} \cdot \sum_j e^{-2\pi i c_j (t_1 - t_2)\xi} \int_{-F_j}^{F_j} e^{-2\pi i(t_1 - t_2)\xi} d\xi \\ &= \frac{1}{2\pi \sum_j F_j (t_1 - t_2)} \sum_j e^{-2\pi i c_j (t_1 - t_2)} \cdot \sin(2\pi F_j (t_1 - t_2)). \end{aligned}$$

Gaussian Fourier Constraint: When μ induces the Gaussian probability distribution with standard deviation F defined by $d\mu(\xi) = \frac{1}{\sqrt{2\pi F^2}} e^{-\xi^2/2F^2} d\xi$, then

$$\begin{aligned} k_\mu(t_1, t_2) &= \int_{\xi \in \mathbb{R}} e^{-2\pi i(t_1 - t_2)\xi} d\mu(\xi) \\ &= \frac{1}{\sqrt{2\pi F^2}} \cdot \int_{\xi \in \mathbb{R}} e^{-2\pi i(t_1 - t_2)\xi} e^{-\xi^2/2F^2} d\xi \\ &= e^{-2\pi^2 F^2 (t_1 - t_2)^2}. \end{aligned}$$

So, k_μ is the *Gaussian kernel*.

Cauchy-Lorentz Fourier Constraint: When μ induces the Cauchy-Lorentz probability density with scale parameter F defined by $d\mu(\xi) = \frac{1}{\pi F\left[1 + \left(\frac{\xi}{F}\right)^2\right]} \cdot d\xi$, we have:

$$\begin{aligned} k_\mu(t_1, t_2) &= \int_{\xi \in \mathbb{R}} e^{-2\pi i(t_1 - t_2)} d\mu(\xi) \\ &= \int_{\xi \in \mathbb{R}} e^{-2\pi i(t_1 - t_2)} \frac{1}{\pi F\left[1 + \left(\frac{\xi}{F}\right)^2\right]} d\xi \\ &= e^{-2\pi F |t_1 - t_2|}. \end{aligned}$$

In the machine learning literature, k_μ is known as the *Laplacian kernel*.

G Signal Reconstruction as Bayesian Estimation

In this section, we show how, as an alternative to Problem 1, we can formulate signal fitting as a Bayesian estimation problem, where the signal y is a stationary stochastic process and the measure μ (which we assume to be symmetric about 0 throughout this section so that $k_\mu(t_1, t_2)$ is real valued) corresponds to a prior on y's power spectral density. This form of prior is commonly used in statistical signal processing, kriging, and machine learning applications [HS93, Rip05, RW06]: We first define a stationary Gaussian process:

Definition 7 (Stationary Gaussian Process [RW06])**.** *A stochastic process $y : \mathbb{R} \to \mathbb{R}$ is a* Gaussian process *if for any finite collection $t_1, \ldots t_s \in \mathbb{R}$, $y(t_1), \ldots, y(t_s)$ is distributed as a multivariate Gaussian. y is a* stationary Gaussian process *if the mean $\mathbb{E}[y(t)]$ is independent of t and the autocorrelation $\mathbb{E}[y(t_1) \cdot y(t_2)]$ depends only on $t_1 - t_2$.*

We now define the specific Gaussian process prior we consider:

Definition 8 (Gaussian Process Prior)**.** *Consider a symmetric probability density function $p_\mu : \mathbb{R} \to \mathbb{R}^+$ and the associated measure μ corresponding to p_μ. We say that a stochastic process $y : \mathbb{R} \to \mathbb{R}$ is distributed according to $\mathcal{D}_\mu$ if y is distributed as a stationary Gaussian process (Definition 7) with with mean $\mathbb{E}[y(t)] = 0$ and autocorrelation function $\mathbb{E}[y(t_1) \cdot y(t_2)] = k_\mu(t_1, t_2)$ for any t_1, t_2, where k_μ is defined in (6).*

As discussed, the prior of Definition 8 amounts to a prior on the power spectral density of y, with the expected power spectral density given by p_μ. Formally:

Claim 56 (Equivalent Power Spectral Density Prior)**.** *Consider y distributed as in Definition 8. Suppose that p_μ is bounded. For every $T > 0$, let $\hat{y}_T : \mathbb{R} \to \mathbb{R}$ be the truncated Fourier transform of y, a.k.a. the amplitude spectral density of y:*

$$\hat{y}_T(\xi) \stackrel{\text{def}}{=} \frac{1}{\sqrt{T}} \int_{-T/2}^{T/2} y(t) e^{-2\pi i t \xi} \, dt.$$

For every T, $\hat{y}_T$ is a Gaussian process with $\mathbb{E}[\hat{y}(t)] = 0$. Also as T goes to infinity, the covariance of $\hat{y}_T$ converges to a diagonal covariance given by p_μ. That is, for any $\xi_1, \ldots, \xi_s \in \mathbb{R}$, $\lim_{T\to\infty} [\hat{y}_T(\xi_1), \ldots, \hat{y}_T(\xi_s)] \sim \mathcal{N}(0, \mathbf{P})$ where $\mathbf{P}$ is a diagonal matrix with $\mathbf{P}_{i,i} = p_\mu(\xi_i)$.

Proof. $\hat{y}$ is a Gaussian process since it is a linear transformation of a Gaussian process, y [Ras04]. We first check that for every T, the mean of this random process is zero at every point ξ.

$$\begin{aligned}
\mathbb{E}[\hat{y}_T(\xi)] &= \mathbb{E}\left[\frac{1}{\sqrt{T}} \int_{-T/2}^{T/2} y(t) e^{-2\pi i t \xi} \, dt\right] \\
&= \frac{1}{\sqrt{T}} \int_{-T/2}^{T/2} \mathbb{E}[y(t)] e^{-2\pi i t \xi} \, dt \\
&= 0,
\end{aligned}$$

where the application of Fubini's theorem in second line above is valid because $k_\mu(0) = 1$ and hence for every $t \in \mathbb{R}$, $\mathbb{E}[|y(t)|] < \infty$. Now in order to show that the covariance of $\hat{y}_T$ converges to being diagonal we check the covariance of $\hat{y}_T$ at two arbitrary points $\xi_1 \neq \xi_2$, as T goes to infinity,

$$\lim_{T\to\infty} \mathbb{E}[\hat{y}_T(\xi_1)^* \hat{y}_T(\xi_2)] = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T} \int_{-T/2}^{T/2} \int_{-T/2}^{T/2} y(t_1) e^{2\pi i t_1 \xi_1} y(t_2) e^{-2\pi i t_2 \xi_2} \, dt_1 dt_2\right]$$

Now note that for every fixed T,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T} \int_{-T/2}^{T/2} \int_{-T/2}^{T/2} y(t_1) e^{2\pi i t_1 \xi_1} y(t_2) e^{-2\pi i t_2 \xi_2} \, dt_1 dt_2\right] &= \frac{1}{T} \int_{-T/2}^{T/2} \int_{-T/2}^{T/2} \mathbb{E}[y(t_1) y(t_2)] e^{2\pi i t_1 \xi_1} e^{-2\pi i t_2 \xi_2} \, dt_1 dt_2 \\
&= \frac{1}{T} \int_{-T/2}^{T/2} \int_{-T/2}^{T/2} k_\mu(t_1, t_2) e^{2\pi i t_1 \xi_1} e^{-2\pi i t_2 \xi_2} \, dt_1 dt_2
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lim_{T\to\infty}\mathbb{E}[\hat{y}_T(\xi_1)^*\hat{y}_T(\xi_2)] &= \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2}\int_{-T/2}^{T/2} k_\mu(t_1,t_2)e^{2\pi i t_1\xi_1}e^{-2\pi i t_2\xi_2}\,dt_1dt_2\\
&= \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2}\int_{-T/2-t}^{T/2-t} k_\mu(\tau)e^{2\pi i t\xi_1}e^{-2\pi i(t+\tau)\xi_2}\,d\tau dt\\
&= \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} e^{2\pi i t(\xi_1-\xi_2)}\int_{-T/2-t}^{T/2-t} k_\mu(\tau)e^{-2\pi i\tau\xi_2}\,d\tau dt\\
&= \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} e^{2\pi i t(\xi_1-\xi_2)}\int_{-\infty}^{\infty}\widehat{k}_\mu(\xi+\xi_2)\left(i\frac{e^{-2\pi i(T/2+t)\xi}-e^{2\pi i(T/2-t)\xi}}{2\pi\xi}\right)d\xi dt,
\end{aligned}$$

where the last equality above used Plancherel theorem. Now we switch the order of two integrals,

$$\begin{aligned}
\mathbb{E}[\hat{y}_T(\xi_1)^*\hat{y}_T(\xi_2)] &= \frac{1}{T}\int_{-T/2}^{T/2} e^{2\pi i t(\xi_1-\xi_2)}\int_{-\infty}^{\infty} p_\mu(\xi+\xi_2)\left(i\frac{e^{-2\pi i(T/2+t)\xi}-e^{2\pi i(T/2-t)\xi}}{2\pi\xi}\right)d\xi dt\\
&= i\int_{-\infty}^{\infty}\frac{p_\mu(\xi+\xi_2)}{2\pi\xi}\left(e^{-2\pi i(T/2)\xi}-e^{2\pi i(T/2)\xi}\right)\int_{-T/2}^{T/2}\frac{e^{2\pi i t(-\xi+\xi_1-\xi_2)}}{T}dtd\xi\\
&= \int_{-\infty}^{\infty} p_\mu(\xi+\xi_2)\frac{\sin(\pi T\xi)}{\pi\xi}\cdot\frac{\sin(\pi T(\xi-\xi_1+\xi_2))}{\pi T(\xi-\xi_1+\xi_2)}d\xi\\
&= \int_{-\infty}^{\infty} T\,\mathrm{sinc}\,(T(\xi-\xi_1+\xi_2))\mathrm{sinc}\,(T\xi)\,p_\mu(\xi+\xi_2)d\xi
\end{aligned}$$

Let $\tilde{\mu}$ be the measure which induces the probability density $p_\mu(\cdot+\xi_2)$, hence it satisfies $d\tilde{\mu}(\xi) = p_\mu(\xi+\xi_2)d\xi$. Now if we take the limit of covariance as $T\to\infty$ we get that,

$$\lim_{T\to\infty}\mathbb{E}[\hat{y}_T(\xi_1)^*\hat{y}_T(\xi_2)] = \lim_{T\to\infty}\int_{-\infty}^{\infty} T\,\mathrm{sinc}\,(T(\xi-\xi_1+\xi_2))\cdot\mathrm{sinc}\,(T\xi)\,d\tilde{\mu}(\xi)$$

For ease of notation we call the integrand in above $f_T(\xi) = T\,\mathrm{sinc}\,(T(\xi-\xi_1+\xi_2))\mathrm{sinc}\,(T\xi)$. Remember, the assumption is that $\xi_1\neq\xi_2$. The sequence $\{f_T(\xi)\}$ converges pointwise to zero for all $\xi\in\mathbb{R}\setminus\{\xi_1,\xi_2\}$. On points ξ_1,ξ_2 it is also bounded by $\frac{1}{|\xi_2-\xi_1|}$. Therefore, the sequence $\{f_T(\xi)\}$ converges pointwise to zero $\tilde{\mu}$-almost everywhere. Also the seguence $\{f_T\}$ is $\tilde{\mu}$-almost dominated by an integrable function g in the sense that for all $T\geq 1$,

$$|f_T(\xi)|\leq g(\xi)$$

g exists since $|f_T(\xi)|\leq T\cdot\frac{2}{T|\xi|+1}\cdot\frac{2}{T|\xi-\xi_1+\xi_2|+1}\overset{\text{def}}{=} h_T(\xi)$ for every $\xi\in\mathbb{R}$ and $h_T(\xi)$ is monotonely converging to zero for $\tilde{\mu}$-almost every ξ and $h_T(\xi)$ is integrable for every value of T. Therefore by Lebesgue's dominated convergence theorem we have,

$$\begin{aligned}
\lim_{T\to\infty}\mathbb{E}[\hat{y}_T(\xi_1)^*\hat{y}_T(\xi_2)] &= \lim_{T\to\infty}\int_{-\infty}^{\infty} f_T(\xi)\,d\tilde{\mu}\\
&= \int_{-\infty}^{\infty}\lim_{T\to\infty} f_T(\xi)\,d\tilde{\mu}\\
&= 0.
\end{aligned}$$

Finally note that the limit of the diagonal entries of the covariance, $\lim_{T\to\infty}\mathbb{E}[|\hat{y}_T(\xi)|^2] = p_\mu(\xi)$ for every $\xi\in\mathbb{R}$ by the Wiener-Khintchine-Einstein Theorem [MC12]. □

It is well known [RW06] that for y distributed as in Definition 8, the posterior distribution of y given samples $t_1, \ldots t_s \in [0,T]$ is also a Gaussian process. Its mean (the Bayes MMSE estimator) and its mode (the MAP estimator) coincide and are given by:

Theorem 57 (Gaussian Process Prior Signal Estimation – Finite Samples)**.** *Consider y distributed as in Definition 8 and noise n distributed as a Gaussian process covariance $\epsilon \cdot \mathbf{I}$. Given $t_1, \ldots, t_s \in [0,T]$, let $\mathbf{y}, \mathbf{n} \in \mathbb{R}^s$ be given by $\mathbf{y}(i) = y(t_i)$ and $\mathbf{n}(t) = n(t_i)$. Let $\mathbf{F} : \mathbb{C}^s \to L_2(\mu)$ be the operator defined by $[\mathbf{F}g](\xi) = \sum_{j=1}^{s} g(j) e^{-2\pi i \xi t_j}$. Both the MAP and MMSE estimates for y are given by $\tilde{y} = \mathcal{F}_\mu^* \tilde{g}$ where:*

$$\tilde{g} = \arg\min_{g \in L_2(\mu)} \left[\frac{1}{s} \|\mathbf{F}^* g - (\mathbf{y} + \mathbf{n})\|_2^2 + \epsilon \|g\|_\mu^2 \right].$$

Proof. Letting $\mathbf{K} = \mathbf{F}^*\mathbf{F}$, so $\mathbf{K}(i,j) = k_\mu(t_i, t_j)$, it is well known [RW06] that the posterior distribution of y given $t_1, \ldots, t_s$ is a Gaussian process with mean $\tilde{y}(t)$ given by:

$$\tilde{y}(t) = \mathbf{k}_t^* (\mathbf{K} + \epsilon \mathbf{I})^{-1} (\mathbf{y} + \mathbf{n}).$$

where $\mathbf{k}_t \in \mathbb{R}^n$ is given by $\mathbf{k}_t(i) = k_\mu(t_i, t)$. It can be shown, analogously to the proof of Theorem 7, that $\tilde{y} = \mathcal{F}_\mu \tilde{g}$ where

$$\tilde{g} = \arg\min_{g \in L_2(\mu)} \left[\frac{1}{s} \|\mathbf{F}^* g - (\mathbf{y} + \mathbf{n})\|_2^2 + \epsilon \|g\|_\mu^2 \right].$$

Further, since $\tilde{y}$ is the mean of the posterior distribution, it gives the Bayes MMSE estimator, and since this posterior distribution is a Gaussian process, also gives the MAP estimator. □

We can see that the least squares problem (10) roughly corresponds to a limit of the finite sample optimization problem of Theorem 57 as the number of samples goes to infinity. Via Theorem 3, this optimization problem can be solved approximately with $\tilde{O}(s_{\mu,\epsilon})$ samples using Algorithm 1 and the universal sampling distribution of Theorem 17. Via Claim 4 one can see that the lower bound of Section 6 (Theorem 24) extends to solving (10), even approximately, and thus our sample complexity is nearly optimal.