SMAI-JCM
SMAI Journal of Computational Mathematics

Weighted least-squares approximation with determinantal point processes and generalized volume sampling

Anthony Nouy & Bertrand Michel

Volume 11 (2025), p. 1-36.
https://doi.org/10.5802/smai-jcm.117

The SMAI Journal of Computational Mathematics is a member of the Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/
Submissions at https://smai-jcm.centre-mersenne.org/ojs/submission
e-ISSN: 2426-8399

SMAI Journal of Computational Mathematics
Vol. 11, 1-36 (2025)

Weighted least-squares approximation with determinantal point processes and generalized volume sampling

Anthony Nouy [1]
Bertrand Michel [2]

[1] Nantes Université, Centrale Nantes, Laboratoire de Mathématiques Jean Leray, CNRS UMR 6629, France
E-mail address: anthony.nouy@ec-nantes.fr
[2] Nantes Université, Centrale Nantes, Laboratoire de Mathématiques Jean Leray, CNRS UMR 6629, France
E-mail address: bertrand.michel@ec-nantes.fr.

Abstract. We consider the problem of approximating a function from L^2 by an element of a given m-dimensional space V_m, associated with some feature map $\boldsymbol{\varphi}$, using evaluations of the function at random points $x_1, \dots, x_n$. After recalling some results on optimal weighted least-squares using independent and identically distributed points, we consider weighted least-squares using projection determinantal point processes (DPP) or volume sampling. These distributions introduce dependence between the points that promotes diversity in the selected features $\boldsymbol{\varphi}(x_i)$. We first provide a generalized version of volume-rescaled sampling yielding quasi-optimality results in expectation with a number of samples $n = O(m \log(m))$, that means that the expected L^2 error is bounded by a constant times the best approximation error in L^2. Also, further assuming that the function is in some normed vector space H continuously embedded in L^2, we further prove that the approximation error in L^2 is almost surely bounded by the best approximation error measured in the H-norm. This includes the cases of functions from L^∞ or reproducing kernel Hilbert spaces. Finally, we present an alternative strategy consisting in using independent repetitions of projection DPP (or volume sampling), yielding similar error bounds as with i.i.d. or volume sampling, but in practice with a much lower number of samples. Numerical experiments illustrate the performance of the different strategies.

Keywords. Weighted least-squares, Optimal sampling, Determinantal point process, Volume sampling.

1. Introduction

We consider the problem of approximating a function f by an element of a given m-dimensional space V_m using point evaluations of the function. The function is defined on a set $\mathcal{X}$ equipped with a positive measure μ and the error is assessed in the natural norm in $L^2_\mu(\mathcal{X})$ defined by

$$\|f\|^2 = \int_{\mathcal{X}} |f(x)|^2 d\mu(x).$$

$\mathcal{X}$ can, for example, be a subset of $\mathbb{R}^d$ but more general Polish spaces can be considered as well. The best approximation error that can be achieved by elements of V_m is

$$\inf_{v \in V_m} \|f - g\| = \|f - P_{V_m} f\|$$

This project is funded by the ANR-DFG project COFNET (ANR-21-CE46-0015). This work was partially conducted within the France 2030 framework programme, Centre Henri Lebesgue ANR-11-LABX-0020-01.
https://doi.org/10.5802/smai-jcm.117

where $P_{V_m} f$ is the orthogonal projection of f onto V_m. An approximation $\hat{f}_m$ can be obtained by a weighted least-squares projection of f defined as the minimizer of

$$\min_{g \in V_m} \frac{1}{n} \sum_{i=1}^{n} w(x_i) |f(x_i) - g(x_i)|^2 \tag{1.1}$$

where $w : \mathcal{X} \to \mathbb{R}$ is a positive weight function and the $x_1, \ldots, x_n$ are points in $\mathcal{X}$. The approximation $\hat{f}_m$ is said quasi-optimal if

$$\|f - \hat{f}_m\| \leq C \inf_{g \in V_m} \|f - g\|,$$

with a constant C independent of m. When using random points, it is said quasi-optimal in expectation whenever

$$\mathbb{E}(\|f - \hat{f}_m\|^2)^{1/2} \leq C \inf_{g \in V_m} \|f - g\|,$$

which guarantees that the averaged error $\mathbb{E}(\|f - \hat{f}_m\|^2)^{1/2}$ converges as least as fast as the best approximation error $e_m(f)_{L^2} := \inf_{g \in V_m} \|f - g\|$. A fundamental problem is to select points and weights that achieve quasi-optimality with a number of points as close as possible to the dimension m of V_m. The weighted least-squares approximation $\hat{f}_m$ defined by (1.1) is such that

$$\|f - \hat{f}_m\|_n = \min_{g \in V_m} \|f - g\|_n \tag{1.2}$$

where $\|\cdot\|_n$ is the empirical (discrete) semi-norm defined by

$$\|f\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} w(x_i) f(x_i)^2. \tag{1.3}$$

The function $\hat{f}_m$ is the orthogonal projection $\hat{P}_{V_m} f$ of f onto V_m with respect to the empirical semi-norm, and the quality of the approximation is related to how close $\|\cdot\|_n$ is from the norm $\|\cdot\|$.

We assume that we are given an orthonormal basis $\varphi_1, \ldots, \varphi_m$ of V_m, and we let $\boldsymbol{\varphi} : \mathcal{X} \to \mathbb{R}^m$ be the associated feature map defined by $\boldsymbol{\varphi}(x) = (\varphi_1(x), \ldots, \varphi_m(x))^T$. Then for any g in V_m, where $g(x) = \boldsymbol{\varphi}(x)^T \mathbf{a}$ for some $\mathbf{a} \in \mathbb{R}^m$, it holds $\|g\|^2 = \|\mathbf{a}\|_2^2$ and $\|g\|_n^2 = \mathbf{a}^T \mathbf{G}^w \mathbf{a}$, where $\mathbf{G}^w$ is the empirical Gram matrix

$$\mathbf{G}^w = \mathbf{G}^w(x_1, \ldots, x_m) := \frac{1}{n} \sum_{i=1}^{n} w(x_i) \boldsymbol{\varphi}(x_i) \boldsymbol{\varphi}(x_i)^T, \tag{1.4}$$

so that

$$\lambda_{min}(\mathbf{G}^w) \|g\|^2 \leq \|g\|_n^2 \leq \lambda_{max}(\mathbf{G}^w) \|g\|^2, \quad \forall g \in V_m, \tag{1.5}$$

which is known as a Marcinkiewicz–Zygmund inequality in sampling discretization [19]. The quality of the projection is therefore related to how much the spectrum of $\mathbf{G}^w$ deviates from one. In particular, it holds

$$\|f - \hat{f}_m\|^2 \leq \|f - P_{V_m} f\|^2 + \lambda_{min}(\mathbf{G}^w)^{-1} \|f - P_{V_m} f\|_n^2.$$

A control of the minimal eigenvalue of $\mathbf{G}^w$ is therefore necessary to achieve quasi-optimality. A control of the highest eigenvalue of $\mathbf{G}^w$ is also needed for numerical stability reasons, so that quasi-optimality can be achieved in finite precision arithmetic. The choice of optimal points (and weights) is a classical problem of design of experiments [27]. A classical approach, called E-optimal design, consists in selecting points (and weights) that maximize $\lambda_{min}(\mathbf{G}^w)$. Variants of this problem consist in maximizing the trace of the inverse of $\mathbf{G}^w$, which is called A-optimal design, or maximizing the determinant $\det(\mathbf{G}^w)$, which is called D-optimal design. The latter is related to Fekete points for polynomial interpolation or more general kernel based interpolation [8, 18]. It is also related to maximum volume concept in linear algebra [15, 16]. However, these optimization problems are in general intractable.

The above mentioned approaches are deterministic. Here, we follow a probabilistic avenue, where the points $x_1, \ldots, x_n$ are drawn from a suitable distribution allowing a control of the spectrum of

the empirical Gram matrix. When the points x_i are drawn from a distribution ν with density w^{-1} with respect to μ, the empirical Gram matrix is an unbiased estimate of the identity. Provided the points are independent and identically distributed (i.i.d.), the empirical Gram matrix almost surely converges to the identity and matrix concentration inequalities allow to analyze how fast is this convergence. An optimization of the convergence rate over all possible distributions yields an optimal density $w_m(x)^{-1} = \frac{1}{m}\|\boldsymbol{\varphi}(x)\|_2^2$, that is known as the inverse Christoffel function for polynomial spaces V_m [9]. The measure $\nu_m = w_m^{-1}\mu$ is also known as leverage score distribution in statistics and machine learning. Sampling from this distribution guarantees that the event $S_\delta = \{\lambda_{min}(\mathbf{G}^{w_m}) \geq 1-\delta\}$ is satisfied with a controlled probability $1-\eta$ provided the number of samples $n = O(\delta^{-2} m \log(m\eta^{-1}))$, where the dependence in m is known to be optimal for i.i.d. sampling [28]. A similar control in probability is obtained for the maximum eigenvalue, and $\lambda_{max}(\mathbf{G}^{w_m}) \leq m$ even holds almost surely with the particular choice of weight function w_m. By drawing i.i.d. samples from ν_m and conditioning to the event S_δ (that can be achieved by a rejection sampling with controlled rejection probability), it holds $\mathbb{E}(\|\cdot\|_n^2) \leq \beta\|\cdot\|^2$ for some constant β, and the resulting least-squares projection $\hat{f}_m$ is quasi-optimal in expectation. If we further assume that the target function f is in a subspace H continuously embedded in $L^2_\mu(\mathcal{X})$ and $L^\infty_{\mu,h^{-1/2}}(\mathcal{X})$ (the space of functions f defined on $\mathcal{X}$ such that $h^{-1/2}f$ is uniformly bounded), with h a probability density with respect to μ, and if we choose for $\nu = w^{-1}\mu$ a mixture of the optimal sampling distribution $\nu_m = w_m^{-1}\mu$ and $h\mu$, we prove that quasi-optimality still holds in expectation and we also prove that $\|\cdot\|_n$ is almost surely bounded by $\|\cdot\|_H$ (up to a constant), which ensures that it holds almost surely

$$\|f - \hat{f}_m\| \leq C \inf_{g\in V_m} \|f-g\|_H, \tag{1.6}$$

which we will call $H \to L^2_\mu$ quasi-optimality. Examples of such spaces H are $L^\infty_\mu(\mathcal{X})$ (with μ a finite measure and $h = \mu(\mathcal{X})^{-1}$), or reproducing kernel Hilbert spaces. Note that the idea of using a mixture between ν_m and μ to control the discrete norm by the L^∞_μ-norm is not new, see, e.g., [2, 26]. The inequality (1.6) ensures that the approximation error in L^2-norm is upper bounded by the best approximation error in H-norm $e_m(f)_H := \inf_{g\in V_m}\|f-g\|_H$. Of course, further assumptions on f and a suitable choice of V_m are required to guarantee some decay of $e_m(f)_H$ (which converges slower than $e_m(f)_{L^2}$ in general). In this paper, we are not concerned with the choice of V_m (which is assumed to be given) and the analysis of the convergence of best approximation errors in V_m (in L^2 or H norms), but only with the construction of algorithms yielding quasi-optimal approximations (in expectation, with high probability or almost surely).

In practice, the number of i.i.d. samples n needed for a stable projection may be large and far from the dimension m. In order to further reduce the sampling complexity, various subsampling approaches have been recently proposed. They start with a set of points that guarantee that the spectrum of $\mathbf{G}^w$ is contained in some interval $[a,b]$, and then extract a subset of points that guarantee that the spectrum of the empirical Gram matrix (up to a possible reweighting) is still contained in some prescribed interval $[a',b']$. The approach proposed in [13, 14] yields quasi-optimality in expectation with a number of samples $n = O(m)$. The algorithm is a randomized version of algorithms provided in [23, 24] for the solution of the Kadinson–Singer problem. This algorithm is unfortunately intractable. However, it is interesting from a theoretical perspective since it allows to prove that quasi-optimality in expectation can be achieved with a number of samples linear in m, therefore showing that sampling numbers in a randomized setting and Kolmogorov widths are comparable for compact sets in $L^2_\mu(\mathcal{X})$. A greedy subsampling algorithm with polynomial complexity has been proposed in [17], that reaches in practice a number of samples n close (and sometimes equal) to m. However, it provides a suboptimal guaranty in expectation, that is $\mathbb{E}(\|f-\hat{f}_m\|^2)^{1/2} \leq C\log(m)^{1/2}\|f - P_{V_m}f\|$, and no theoretical guaranty to extract a set of samples of size $n = O(m)$. Another tractable approach has been proposed in [2],

which allows to reach a number of samples $n = O(m)$. Yet, this algorithm does not provide quasi-optimality in expectation. These conditioning and subsampling approaches all yield a set of points with a dependence structure that is not given explicitly. They require to start with a rather large set of samples and suffer from the complexity of subsampling algorithms, which is polynomial in the initial number of samples.

Another route is to leave the i.i.d. setting from the start and sample from a distribution that introduces a dependence between the samples. An algorithm which achieves quasi-optimality in expectation with $n = O(m)$ samples has been proposed in [12]. It is a randomized variant of the subsampling algorithm from [2]. Another prominent approach is to rely on volume sampling, first introduced in a discrete setting in [1, 11], and then extended to more general settings in [25]. Volume sampling has found many applications in machine learning. For classical (non weighted) least-squares, it consists in drawing samples $\mathbf{x} = (x_1, \ldots, x_n)$ from a distribution γ_n over $\mathcal{X}^n$ having a density proportional to $\det(\Phi(\mathbf{x})^T\Phi(\mathbf{x}))$, with $\Phi(\mathbf{x}) = (\boldsymbol{\varphi}(x_1), \ldots, \boldsymbol{\varphi}(x_n))^T$. The distribution γ_m, for $n = m$, corresponds to a projection determinantal point processes (DPP) [21]. The density drops down to zero whenever two vectors $\boldsymbol{\varphi}(x_i)$ get collinear, hence this distribution introduces a repulsion between the points and promotes diversity in the selected features $\boldsymbol{\varphi}(x_i)$. For $n > m$, up to a random permutation of points, this distribution corresponds to m points from a projection DPP and an independent set of $n-m$ i.i.d. samples from μ (provided μ is a probability measure). The associated empirical Gram matrix (with weight $w = 1$) has bad concentration properties. Here we consider a generalized volume sampling distribution γ_n^ν for weighted least-squares, which has a density proportional to $\det(\mathbf{G}^w(\mathbf{x}))$ with respect to a product measure $\nu^{\otimes n}$ (the measure μ is no more required to be a probability measure). This introduces a compromise between promoting a high likelihood with respect to the reference measure ν and promoting a high determinant of the empirical Gram matrix. For $\nu = \nu_m$, $\gamma_n^{\nu_m}$ corresponds to the volume-rescaled sampling distribution introduced in [10]. This distribution yields quasi-optimality in expectation, without the need of conditioning. Moreover, this distribution has the very nice property of providing an unbiased approximation, i.e. $\mathbb{E}(\hat{f}_m) = P_{V_m} f$, which allows to perform an averaging of estimators for improving quasi-optimality constant.

Our first main contribution is to consider a general version of volume-rescaled sampling distribution, with a measure $\nu = w^{-1}\mu$ allowing to obtain not only quasi-optimality in expectation but also an almost sure $H \to L^2_\mu$ quasi-optimality for functions from subspaces H described above. Despite the many advantages of volume sampling compared to i.i.d. optimal sampling, the number of samples to ensure stability of the empirical Gram matrix with high probability is essentially of the same order as for i.i.d. sampling, i.e. $n = O(m\delta^{-2}\log(m\eta^{-1}))$.

Our second contribution is to propose an alternative that consists in using r independent samples from the projection DPP distribution γ_m, or from the volume sampling distribution γ_n^ν with a suitable mixture distribution ν. Using conditioning, the former allows to obtain quasi-optimality in expectation, while the latter allows to achieve $H \to L^2_\mu$ quasi-optimality almost surely for a subspace of functions H. These results are similar to optimal i.i.d. sampling (with suitable mixture measures) or to our general version of volume sampling. We can prove that stability S_δ is achieved with probability $1-\eta$ under a suboptimal condition $n = O(m^2\delta^{-2}\log(m\eta^{-1}))$, or a better condition $n = O(m\delta^{-2}\log(m\eta^{-1}))$ (similar to i.i.d. and volume sampling) under a conjecture (only checked numerically) on the properties of the empirical Gram matrix associated with projection DPP or volume sampling. Although this theoretical guaranty does not show any advantage of this new sampling strategy, we observe in practice a much better concentration of the empirical Gram matrix, hence a much lower number of samples needed for obtaining the stability condition S_δ with high probability.

Although they are not directly in line with our setting (the approximation of a function in an arbitrary subspace V_m), we would like to mention related works [4, 5] using determinantal point

processes for the approximation of functions from reproducing kernel Hilbert spaces H. In these works, the sampling distribution is related to the kernel of H.

In this paper, we only provide upper bounds of the error in L^2-norm in terms of errors of best approximation in L^2 or H norms. Obtaining a control of the error in other norms, e.g. L^∞ or a some RKHS norm, would certainly be of interest but this in general requires to modify the projection or the sampling methods, see the recent works [20, 31] in this direction.

The outline of the paper is as follows. In Section 2, we provide some preliminary results on weighted least-squares projections. In Section 3, we recall some classical results on optimal weighted least-squares with i.i.d. sampling, with quasi-optimality results in expectation and also $H \to L^2_\mu$ quasi-optimality results for a large class of function spaces, that extend previous results [17] to a more general setting. In Section 4, we introduce DPP and more general volume sampling distributions, and analyze the properties of corresponding weighted least-squares projections. In particular we obtain quasi-optimality in expectation and almost sure $H \to L^2_\mu$ quasi-optimality when using our general volume sampling distribution with a suitable weight function. In Section 5, we present the alternative strategy consisting in using independent repetitions of DPP (or volume sampling), and obtain similar quasi-optimality results. In Section 6, we provide numerical evidence of the efficiency of the strategy based on independent repetitions of DPP, compared to optimal i.i.d. or volume-rescaled sampling.

2. Preliminary results on weighted least-squares approximation

Here, we provide some preliminary results on weighted least-squares approximation. We start with a control of the bias of the empirical semi-norm, provided a condition on the weight function w that needs to be related to the sampling distribution.

Lemma 2.1. *Assuming that the points are drawn from a distribution over $\mathcal{X}^n$ with marginals all equal to $\tilde{\nu} = \tilde{w}\mu$, and assuming that the weight function w is such that $w \leq \beta\tilde{w}$, it holds for all $f \in L^2_\mu$*

$$\mathbb{E}(\|f\|_n^2) \leq \beta \|f\|^2$$

with equality when $\tilde{w} = w$.

Proof. It holds $\mathbb{E}(\|f\|_n^2) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{x_i\sim\tilde{\nu}}(w(x_i)f(x_i)^2) \leq \beta \int f(x)^2 d\mu(x) = \beta\|f\|^2$. ∎

Assuming $\mathbf{G}^w$ invertible, the projection error satisfies

$$\|f - \hat{f}_m\|^2 = \|f - P_{V_m} f\|^2 + \|\hat{P}_{V_m}(f - P_{V_m} f)\|^2 \tag{2.1}$$

from which we deduce

$$\|f - \hat{f}_m\|^2 \leq \|f - P_{V_m} f\|^2 + \lambda_{min}(\mathbf{G}^w)^{-1}\|f - P_{V_m} f\|_n^2,$$

and the following result.

Lemma 2.2. *Assume that the points $(x_1, \ldots, x_n)$ are drawn from a distribution over $\mathcal{X}^n$ with marginals all equal to $\tilde{\nu} = \tilde{w}^{-1}\mu$ and we use weighted least-squares with a weight function w such that $w \leq \beta\tilde{w}$. Letting $S_\delta = \{\lambda_{min}(\mathbf{G}^w) \geq 1 - \delta\}$, it holds for any $f \in L^2_\mu$*

$$\mathbb{E}(\|\hat{P}_{V_m} f\|^2 | S_\delta) \leq \mathbb{P}(S_\delta)^{-1}(1-\delta)^{-1}\beta\|f\|^2,$$

and

$$\mathbb{E}(\|f - \hat{P}_{V_m} f\|^2 | S_\delta) \leq (1 + \mathbb{P}(S_\delta)^{-1}(1-\delta)^{-1}\beta) \inf_{g\in V_m} \|f - g\|^2.$$

Proof. From (1.5), we obtain

$$\mathbb{E}(\|\hat{P}_{V_m} f\|^2 | S_\delta) \leq (1-\delta)^{-1} \mathbb{E}(\|\hat{P}_{V_m} f\|_n^2 | S_\delta) \leq (1-\delta)^{-1} \mathbb{P}(S_\delta)^{-1} \mathbb{E}(\|\hat{P}_{V_m} f\|_n^2),$$

and since $\hat{P}_{V_m}$ is an orthogonal projection with respect to the inner product $\|\cdot\|_n$, it holds $\mathbb{E}(\|\hat{P}_{V_m} f\|_n^2) \leq \mathbb{E}(\|f\|_n^2) \leq \beta \|f\|^2$, where the last inequality results from Lemma 2.1. The second inequality then follows from (2.1). ∎

In order to obtain error bounds with high probability or even almost surely, we introduce additional assumptions on the target function, and choose a weight function accordingly. For some strictly positive function h, we let $L^\infty_{\mu,h^{-1/2}}(\mathcal{X})$ be the space of functions defined on $\mathcal{X}$ such that $fh^{-1/2}$ is in $L^\infty_\mu(\mathcal{X})$. Let H be a normed vector space of functions defined on $\mathcal{X}$, continuously embedded in both L^2_μ and $L^\infty_{\mu,h^{-1/2}}$. That means respectively that for any $f \in H$,

$$\|f\| \leq C_H \|f\|_H, \tag{2.2}$$

and

$$\|f\|_{L^\infty_{\mu,h^{-1/2}}} = \operatorname*{ess\,sup}_{x \in \mathcal{X}} h(x)^{-1/2} |f(x)| \leq \|f\|_H. \tag{2.3}$$

Example 2.3. When μ is a probability measure, the properties (2.2) and (2.3) hold for $H = L^\infty_\mu$, with $h = 1$, and embedding constant $C_H = 1$.

Example 2.4. The properties (2.2) and (2.3) hold for H a reproducing kernel Hilbert space of functions with kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ having finite trace $\int K(x,x) d\mu(x) < \infty$. H is compactly embedded in L^2_μ with embedding constant $C_H^2 = \int K(x,x) d\mu(x)$, and continuously embedded in $L^\infty_{\mu,h^{-1/2}}$ with $h(x) = K(x,x)$. The kernel admits a Mercer decomposition $K(x,x) = \sum_{i=1}^M \lambda_i \psi_i(x) \psi_j(y)$ with $M \in \mathbb{N} \cup \{+\infty\}$, where the ψ_i form an orthonormal system in L^2_μ and the $\lambda_i > 0$ are such that $\sum_{i=1}^M \lambda_i = C_H^2$. The kernel can be rescaled such that $C_H = 1$, in which case h is a probability density with respect to μ. In the case when μ is itself a probability measure and $\mathcal{X}$ has a group structure with $K(x,y) = k(x-y)$, then $h(x) = k(0)$ is a constant function, and with the previously mentioned rescaling, $C_H = 1$ and $h = 1$, and H is continuously embedded in L^∞_μ. More generally, when h is uniformly bounded, H is continuously embedded in L^∞_μ. However, there are some interesting cases of RKHS for which h is not uniformly bounded, e.g. the Sobolev space $H^1_\nu(\mathbb{R})$ with $\nu = \mathcal{N}(0,1)$ the standard Gaussian measure, whose kernel has diagonal $h(x) = \sqrt{\pi/2} \exp(x^2)(1 - \operatorname{erf}(x/\sqrt{2})^2)$, and which is continuously embedded in L^2_μ for $\mu \sim \mathcal{N}(0,a)$, for $a < 1$. We refer to [6] for an introduction to RKHS.

Noting that for any $g \in V_m$, it holds

$$\|f - \hat{f}_m\| \leq \|f - g\| + \lambda_{min}(\mathbf{G}^w)^{-1/2} \|f - g\|_n,$$

we can deduce another useful lemma, provided some condition on the sampling measure.

Lemma 2.5. *Assume $f \in H$ with H satisfying (2.2) and (2.3). If the weight function w is such that $w \geq \zeta^{-1} h$, it holds $\|g\|_n^2 \leq \zeta \|g\|_H^2$ almost surely, for any $g \in H$. If we further assume that $\mathbf{G}^w$ is almost surely invertible, it holds almost surely*

$$\|f - \hat{f}_m\| \leq \left(C_H + \lambda_{min}(\mathbf{G}^w)^{-1/2} \zeta^{1/2} \right) \inf_{g \in V_m} \|f - g\|_H.$$

3. Least-squares with independent and identically distributed samples

We here consider the classical setting where the $x_1, \dots, x_n$ are i.i.d. samples from a distribution $\nu = w^{-1}\mu$ with density w^{-1} with respect to μ. The empirical Gram matrix can be written

$$\mathbf{G}^w = \frac{1}{n}\sum_{i=1}^n \mathbf{A}_i, \quad \mathbf{A}_i = w(x_i)\boldsymbol{\varphi}(x_i)\boldsymbol{\varphi}(x_i)^T.$$

The $\mathbf{A}_i$ are i.i.d. rank-one matrices, with expectation $\mathbb{E}(\mathbf{A}_i) = \mathbf{I}$ and spectral norm satisfying almost surely

$$\|\mathbf{A}_i\| = w(x_i)\|\boldsymbol{\varphi}(x_i)\|_2^2.$$

From matrix Chernoff inequality (recalled in Theorem A.1), we then deduce the following result from [9].

Lemma 3.1. *Assume the points $(x_1, \dots, x_n)$ are i.i.d. samples from $\nu = w^{-1}\mu$, with w such that*

$$K_{w,m} = \sup_{x\in\mathcal{X}} w(x)\|\boldsymbol{\varphi}(x)\|_2^2 < \infty.$$

Then for any $0 < \delta < 1$, it holds

$$\mathbb{P}(\lambda_{min}(\mathbf{G}^w) < 1-\delta) \leq m\exp(-nc_\delta/K_{w,m})$$

with $c_\delta = \delta + (1-\delta)\log(1-\delta)$ such that $\delta^2/2 \leq c_\delta \leq \delta^2$. Then it holds

$$\mathbb{P}(\lambda_{min}(\mathbf{G}^w) < 1-\delta) \leq \eta \quad \text{if} \quad n \geq c_\delta^{-1}K_{w,m}\log(m\eta^{-1}).$$

Since

$$K_{w,m} \geq \mathbb{E}_{x\sim\nu}(w(x)\|\boldsymbol{\varphi}(x)\|_2^2) = \int \|\boldsymbol{\varphi}(x)\|_2^2 d\mu(x) = m,$$

we deduce that $K_{w,m} \geq m$. The optimal sampling measure that minimizes the upper bound of the matrix Chernoff inequality is therefore given by

$$\nu_m = w_m^{-1}\mu,$$

where the density w_m^{-1} with respect to μ is given by

$$w_m^{-1}(x) := \frac{1}{m}\sum_{i=1}^m \varphi_i(x)^2 = \frac{1}{m}\|\boldsymbol{\varphi}(x)\|_2^2,$$

that provides an optimal constant $K_{w_m,m} = m$. This optimal distribution for i.i.d. sampling is also known as leverage score distribution. Choosing a function w such that $w^{-1} \geq \alpha w_m^{-1}$ for some $\alpha > 0$ yields a constant

$$K_{w,m} \leq \alpha^{-1}K_{w_m,m} = \alpha^{-1}m, \tag{3.1}$$

and we have $\mathbb{P}(\lambda_{min}(\mathbf{G}^w) < 1-\delta) \leq \eta$ provided $n \geq c_\delta^{-1}\alpha^{-1}m\log(m\eta^{-1})$.

We next provide a useful lemma on the stability of the empirical least-squares projection.

Lemma 3.2. *Assume that $(x_1, \dots, x_n)$ is drawn from $\nu^{\otimes n}$ with $\nu = w^{-1}\mu$. Let $S_\delta = \{\lambda_{min}(\mathbf{G}^w) \geq 1-\delta\}$ with $0 < \delta < 1$. Then for any $f \in L^2_\mu$, it holds*

$$\mathbb{E}(\|\hat{P}_{V_m}f\|^2|S_\delta) \leq \mathbb{P}(S_\delta)^{-1}(1-\delta)^{-1}\|f\|^2,$$

and

$$\mathbb{E}(\|\hat{P}_{V_m}f\|^2|S_\delta) \leq \mathbb{P}(S_\delta)^{-1}(1-\delta)^{-2}\left(\|P_{V_m}f\|^2 + \frac{K_{w,m}}{n}\|f\|^2\right).$$

Proof. The first inequality directly comes from Lemma 2.2 with $\beta = 1$. For the proof of the second inequality, let $\mathbf{G} := \mathbf{G}^w$, and first note that $\hat{P}_{V_m} f(x) = \boldsymbol{\varphi}(x)^T \mathbf{c}$, with $\mathbf{c} = \mathbf{G}^{-1}\mathbf{b}$ and $\mathbf{b} = \frac{1}{n}\sum_{i=1}^n w(x_i)\boldsymbol{\varphi}(x_i)f(x_i)$. Therefore

$$\|\hat{P}_{V_m} f\|^2 = \|\mathbf{c}\|_2^2 = \|\mathbf{G}^{-1}\mathbf{b}\|_2^2 \leq \lambda_{min}(\mathbf{G})^{-2}\|\mathbf{b}\|_2^2,$$

and

$$\mathbb{E}(\|\hat{P}_{V_m} f\|^2 | S_\delta) \leq (1-\delta)^{-2}\mathbb{P}(S_\delta)^{-1}\mathbb{E}(\|\mathbf{b}\|_2^2).$$

Then we have

$$\begin{aligned}
\mathbb{E}(\|\mathbf{b}\|_2^2) &= \frac{1}{n^2}\sum_{i,j=1}^n \mathbb{E}(w(x_i)\boldsymbol{\varphi}(x_i)^T f(x_i) w(x_j)\boldsymbol{\varphi}(x_j)^T f(x_j)) \\
&= \frac{1}{n}\mathbb{E}_{x\sim\nu}(f(x)^2 w(x)^2 \|\boldsymbol{\varphi}(x)\|_2^2) + \frac{n-1}{n}\|\int f(x)\boldsymbol{\varphi}(x)d\mu(x)\|_2^2 \\
&\leq \frac{K_{w,m}}{n}\|f\|^2 + \frac{n-1}{n}\|P_{V_m} f\|^2,
\end{aligned}$$

which ends the proof. ∎

Theorem 3.3. *Assume that $(x_1, \ldots, x_n)$ is drawn from $\nu^{\otimes n}$ with $\nu = w^{-1}\mu$ such that $w^{-1} \geq \alpha w_m^{-1}$ for some $\alpha > 0$. Further assume that*

$$n \geq c_\delta^{-1}\alpha^{-1} m \log(m\eta^{-1}),$$

with $0 < \delta < 1$. Then the event $S_\delta = \{\lambda_{min}(\mathbf{G}^w) \geq 1-\delta\}$ is such that $\mathbb{P}(S_\delta) \geq 1-\eta$ and it holds

$$\mathbb{E}(\|f - \hat{f}_m\|^2 | S_\delta) \leq (1 + (1-\eta)^{-1}(1-\delta)^{-1}) \inf_{g\in V_m} \|f-g\|^2,$$

and

$$\mathbb{E}(\|f - \hat{f}_m\|^2 | S_\delta) \leq (1 + \alpha^{-1}\frac{m}{n}(1-\eta)^{-1}(1-\delta)^{-2}) \inf_{g\in V_m} \|f-g\|^2,$$

Proof. The first inequality comes from Lemma 2.2 and Lemma 3.1, while the second inequality follows from (2.1), Lemma 3.1 and Lemma 3.2, noting that $P_{V_m}(f - P_{V_m} f) = 0$. ∎

The next theorem provides a control of error in probability, provided that the target function f is in a space H satisfying (2.2) and (2.3), with h a probability density w.r.t. μ, and we use a sampling distribution $\nu = w^{-1}\mu$ with

$$w^{-1} = \alpha w_m^{-1} + (1-\alpha)h. \tag{3.2}$$

The measure ν is a mixture between $\nu_m = w_m^{-1}\mu$ and the measure $h\mu$, with respective weights α and $1-\alpha$.

Theorem 3.4. *Assume $f \in H$, with H satisfying (2.2) and (2.3), with h a probability density with respect to μ. Assume $(x_1, \ldots, x_n)$ is drawn from $\nu^{\otimes n}$ with $\nu = w^{-1}\mu$ and $w^{-1} = \alpha w_m^{-1} + (1-\alpha)h$. Then provided*

$$n \geq c_\delta^{-1}\alpha^{-1} m \log(m\eta^{-1}),$$

with $0 < \delta < 1$, the event $S_\delta = \{\lambda_{min}(\mathbf{G}^w) \geq 1-\delta\}$ is such that $\mathbb{P}(S_\delta) \geq 1-\eta$ and it holds

$$\|f - \hat{f}_m\| \leq \left(C_H + (1-\delta)^{-1/2}(1-\alpha)^{-1/2}\right) \inf_{g\in V_m} \|f-g\|_H$$

with probability greater than $1-\eta$.

Proof. It is directly deduced from Lemma 2.5 and Lemma 3.1. ∎

Remark 3.5 (Sampling from the mixture ν)**.** Sampling from the mixture $\nu = \alpha\nu_m + (1-\alpha)h\mu$ can be performed by sampling from ν_m with probability α and from $h\mu$ with probability $1-\alpha$. When $H = L^\infty_\mu$ and μ is a probability measure (see Example 2.3), $h = 1$ and it requires sampling from the reference measure μ. When H is a RKHS (see Example 2.4) with kernel K such that $\int K(x,x)d\mu(x) = 1$, it requires sampling from $h\mu$ with $h(x) = K(x,x)$. If K is known from its Mercer decomposition $K(x,y) = \sum_{i=1}^M \lambda_i\psi_i(x)\psi(y)$, with $M \in \mathbb{N}\cup\{+\infty\}$, then $h(x) = \sum_{i=1}^M \lambda_i\psi_i(x)^2$ and $h\mu$ can be sampled as a mixture of distributions $\psi_i^2\mu$ with weights λ_i. An alternative is to sample independent Bernoulli variables $B_i \sim B(\lambda_i)$, and then sample from the distribution $h_I\mu$ with $h_I(x) = \frac{1}{\#I}\sum_{i\in I}\psi_i(x)^2$ with $I = \{i : B_i = 1\}$.

4. Least-squares with determinantal point processes and volume sampling

4.1. Projection determinantal point process

A projection determinantal point process $\mathrm{DPP}_\mu(V_m)$ associated with the space V_m and the reference measure μ (not necessarily a finite measure) is a distribution over $\mathcal{X}^m$ defined by

$$d\gamma_m(\mathbf{x}) = \frac{1}{m!}\det(\Phi(\mathbf{x})^T\Phi(\mathbf{x}))d\mu^{\otimes m}(\mathbf{x}), \quad \mathbf{x}\in\mathcal{X}^m,$$

where $\Phi(\mathbf{x}) \in \mathbb{R}^{n\times m}$ is the matrix whose i-th row is $\boldsymbol{\varphi}(x_i)^T$. It is a determinantal point process with projection kernel $K(x,y) = \boldsymbol{\varphi}(x)^T\boldsymbol{\varphi}(y)$ and reference measure μ. Sampling from γ_m tends to select at random a set of features $(\boldsymbol{\varphi}(x_1),\dots,\boldsymbol{\varphi}(x_m))$ with high volume in $\mathbb{R}^m$. The density $\frac{1}{m!}\det(\Phi(\mathbf{x})^T\Phi(\mathbf{x}))$ is equal to zero when two points x_i and x_j are equal, or more generally when two features $\boldsymbol{\varphi}(x_i)$ and $\boldsymbol{\varphi}(x_j)$ are collinear for $i \neq j$. It is a particular class of repulsive point processes. The following result indicates that the marginals of γ_m are all equal to the optimal sampling measure for i.i.d. sampling for V_m, and provides a factorization of the distribution in terms of conditional distributions.

Proposition 4.1 ([21, Theorem 2.7])**.** *Let $(x_1,\dots,x_m) \sim \gamma_m$. Each x_k has for marginal distribution $\nu_m = w_m^{-1}\mu$, with $w_m^{-1}(x) = \frac{1}{m}\|\boldsymbol{\varphi}(x)\|_2^2$. For $2 \le k \le m$, the conditional distribution of x_k knowing $x_1,\dots,x_{k-1}$ has for probability density with respect to μ the function*

$$p_k(x_k) := \frac{1}{m-k+1}\|\boldsymbol{\varphi}(x_k) - P_{W_{k-1}}\boldsymbol{\varphi}(x_k)\|_2^2$$

where $P_{W_{k-1}}$ is the orthogonal projection onto the space $W_{k-1} = \mathrm{span}\{\boldsymbol{\varphi}(x_1),\dots,\boldsymbol{\varphi}(x_{k-1})\}$ in $\mathbb{R}^m$.

Proof. See Appendix B. ∎

From the previous result, we deduce a sequential procedure to draw a sample $(x_1,\dots,x_m)$ from the distribution $\gamma_m = \mathrm{DPP}_\mu(V_m)$. The first point x_1 is obtained by drawing a sample from ν_m. Then given the points $(x_1,\dots,x_{k-1})$, the point x_k is drawn from the probability measure $p_k(x)d\mu(x)$.

Example 4.2. Consider $\mathcal{X} = [0,1]$ equipped with the uniform measure μ and the space V_m of piecewise constant functions on a uniform partition of $[0,1]$ with m intervals. An orthogonal basis is given by $\varphi_j(x) = \sqrt{m}\mathbf{1}_{x\in[(j-1)/m,j/m)}$. Here $\boldsymbol{\varphi}(x_i) = \sqrt{m}\mathbf{e}_i$, where $\mathbf{e}_i$ is the i-th canonical vector in $\mathbb{R}^m$. Then the density of γ_m is 0 once two points or more are in the same interval, and equal to $m^m/m!$ if there is exactly one point in each interval. The marginals are all equal to μ. The conditional density p_k is equal to 0 on the intervals containing the points $(x_1,\dots,x_{k-1})$, and equal to $\frac{m}{m-k+1}$ elsewhere.

Remark 4.3. Letting $\mathbf{v}_1, \dots, \mathbf{v}_m$ be the orthonormal basis of $\mathbb{R}^m$ such that $\mathbf{v}_i \propto \boldsymbol{\varphi}(x_i) - P_{W_{i-1}}\boldsymbol{\varphi}(x_i)$, we have that the functions $\psi_i(x) = \mathbf{v}_i^T\boldsymbol{\varphi}(x)$ form an L^2_μ-orthonormal basis of V_m, and

$$p_k(x)d\mu(x) = \frac{1}{m-k+1}\left(\sum_{i=k}^m \psi_i(x)^2\right)d\mu(x),$$

that is the optimal sampling distribution for the space $\operatorname{span}\{\psi_k, \dots, \psi_m\}$, which is the orthogonal complement of $\operatorname{span}\{\psi_1, \dots, \psi_{k-1}\}$ in V_m.

Remark 4.4. When replacing the random draw $x_{k+1} \sim \frac{1}{m-k}\|\boldsymbol{\varphi}(x) - P_{W_k}\boldsymbol{\varphi}(x)\|_2^2 d\mu(x)$ by a deterministic selection

$$x_{k+1} \in \arg\max_{x\in\mathcal{X}} \|\boldsymbol{\varphi}(x) - P_{W_k}\boldsymbol{\varphi}(x)\|_2^2,$$

the resulting algorithm corresponds to a deterministic greedy algorithm for the construction of a hierarchical sequence of spaces $W_1 \subset \ldots \subset W_m$ for the approximation of the manifold $\mathcal{M} = \{\boldsymbol{\varphi}(x) : x \in \mathcal{X}\}$ [7, 22]. It also coincides with the sequential design in Gaussian process interpolation, using a kernel $K(x,y) = \boldsymbol{\varphi}(x)^T\boldsymbol{\varphi}(y)$. Indeed, in this case, $W_k = \operatorname{span}\{K(x_1,\cdot), \dots, K(x_k,\cdot)\}$ and the interpolation of a function at points $(x_1, \dots, x_k)$ is the orthogonal projection onto W_k with respect to the RKHS associated with the kernel K. The variance at point x of this interpolation given $(x_1, \dots, x_k)$ is $\|\boldsymbol{\varphi}(x) - P_{W_k}\boldsymbol{\varphi}(x)\|_2^2$. Therefore, the selected point x_{k+1} is where the interpolation has maximum uncertainty.

For any probability density w^{-1} w.r.t. μ, we let $\boldsymbol{\varphi}^w : \mathcal{X} \to \mathbb{R}^m$ be the weighted feature map such that $\boldsymbol{\varphi}^w(x) = (\varphi_1^w(x), \dots, \varphi_m^w(x))^T = w(x)^{1/2}\boldsymbol{\varphi}(x)$ and $\Phi^w(\mathbf{x})$ be the matrix in $\mathbb{R}^{n\times m}$ whose i-th row is $\boldsymbol{\varphi}^w(x_i)^T$. We have the following straightforward property.

Proposition 4.5. *For any distribution $\nu = w^{-1}\mu$, it holds*

$$d\gamma_m(\mathbf{x}) = \frac{1}{m!}\det(\Phi^w(\mathbf{x})^T\Phi^w(\mathbf{x}))d\nu^{\otimes m}(\mathbf{x}), \quad \mathbf{x} \in \mathcal{X}^m.$$

The functions $\varphi_1^w, \dots, \varphi_m^w$ form an orthonormal basis of a subspace V_m^w in $L^2_\nu(\mathcal{X})$, and the distribution $\mathrm{DPP}_\mu(V_m)$ *coincides with* $\mathrm{DPP}_\nu(V_m^w)$.

From the above, we deduce that for $\nu = w^{-1}\mu$,

$$d\gamma_m(\mathbf{x}) = \frac{m^m}{m!}\det(\mathbf{G}^w(\mathbf{x}))d\nu^{\otimes m}(\mathbf{x}), \quad \mathbf{x} \in \mathcal{X}^m.$$

Therefore, sampling from γ_m tends to favor points $\mathbf{x} \in \mathcal{X}^m$ leading simultaneously to a high likelihood with respect to the product measure $\nu^{\otimes m}$ and a high value of the determinant of $\mathbf{G}^w(\mathbf{x})$. This tends to favor empirical Gram matrices with high eigenvalues.

Remark 4.6 (Complexity)**.** Let us assume that μ is a discrete measure with N atoms. The cost of sampling a projection DPP is $O(m^3 + Nm^2)$. This can be improved to $O(m^3 + Nm)$ (up to log factors) using rejection sampling, see [3].

4.2. **Volume sampling**

The volume sampling distribution $\mathrm{VS}_\mu^n(V_m)$ is the distribution over $\mathcal{X}^n$ defined by

$$d\gamma_n(\mathbf{x}) = \frac{(n-m)!}{n!}\det(\Phi(\mathbf{x})^T\Phi(\mathbf{x}))d\mu^{\otimes n}(\mathbf{x}),$$

for $n \geq m$. For $n = m$, the volume sampling distribution $\mathrm{VS}_\mu^m(V_m)$ coincides with the projection determinantal point process $\mathrm{DPP}_\mu(V_m)$. For $n > m$, provided μ is a probability measure, a sample from γ_n is composed by m samples from the projection determinantal process $\mathrm{DPP}_\mu(V_m)$ and $n-m$ i.i.d. samples from the measure μ, to which is applied a random permutation, as stated in the next proposition.

Theorem 4.7 ([10, Theorem 2.4])**.** *Assume that μ is a probability measure. If $(x_1, \dots, x_n) \sim \gamma_m \otimes \mu^{\otimes(n-m)}$ and σ is an independent permutation drawn uniformly at random over the set of permutations of $\{1, \dots, n\}$, then $(x_{\sigma(1)}, \dots, x_{\sigma(n)}) \sim \gamma_n$. The marginals of the distribution γ_n are all equal to the mixture*

$$\frac{m}{n}\nu_m + \frac{n-m}{n}\mu = \left(\frac{m}{n}w_m + \frac{n-m}{n}\right)\mu.$$

Given a probability measure $\nu = w^{-1}\mu$ (μ is no more required to be a probability measure), for $n \geq m$ we can define another volume sampling distribution $\mathrm{VS}_\nu^n(V_m^w)$ over $\mathcal{X}^n$ defined by

$$d\gamma_n^\nu(\mathbf{x}) = \frac{(n-m)!}{n!}\det(\Phi^w(\mathbf{x})^T\Phi^w(\mathbf{x}))d\nu^{\otimes n}(\mathbf{x}) = n^m\frac{(n-m)!}{n!}\det(\mathbf{G}^w(\mathbf{x}))d\nu^{\otimes n}(\mathbf{x}).$$

Sampling from γ_n^ν tends to favor points $\mathbf{x} \in \mathcal{X}^m$ leading simultaneously to a high likelihood with respect to the product measure $\nu^{\otimes n}$ and a high value of the determinant of $\mathbf{G}^w(\mathbf{x})$. As a corollary of Theorem 4.7, we have the following result.

Theorem 4.8. *If $(x_1, \dots, x_n) \sim \gamma_m \otimes \nu^{\otimes(n-m)}$, with $\nu = w^{-1}\mu$ a probability measure, and σ is an independent permutation drawn uniformly at random over the set of permutations of $\{1, \dots, n\}$, then $(x_{\sigma(1)}, \dots, x_{\sigma(n)}) \sim \gamma_n^\nu$. The marginals of the distribution γ_n^ν are all equal to the mixture*

$$\tilde{\nu} = \tilde{w}^{-1}\mu \quad \textit{with} \quad \tilde{w}^{-1} = \frac{m}{n}w_m^{-1} + \frac{n-m}{n}w^{-1}.$$

If $w_m \geq \alpha w$, then $\tilde{w}$ satisfies

$$\left(1 - \frac{m}{n}\right)w^{-1} \leq \tilde{w}^{-1} \leq \left(1 + (\alpha^{-1} - 1)\frac{m}{n}\right)w^{-1}.$$

Proof. Since $\boldsymbol{\varphi}^w$ form an orthonormal basis of the space V_m^w in L^2_ν, we deduce from Theorem 4.7 and Proposition 4.5 that up to a random permutation, a sample from γ_n^ν is composed by m points drawn from $\mathrm{DPP}(V_m^w, \nu) = \mathrm{DPP}(V_m, \mu)$ (with marginals ν_m) and $n - m$ i.i.d. samples from the measure ν. The expression of the marginals is a direct consequence. ∎

Taking $\nu = \mu$ (provided μ is a probability measure), we have $\gamma_n^\mu = \gamma_n$, that is the classical volume sampling distribution $\mathrm{VS}_\mu^n(V_m)$. Taking $\nu = \nu_m$, we obtain the distribution

$$d\gamma_n^{\nu_m}(\mathbf{x}) = \frac{(n-m)!}{n!}\det(\Phi^{w_m}(\mathbf{x})^T\Phi^{w_m}(\mathbf{x}))d\nu_m^{\otimes n}(\mathbf{x})$$

which corresponds to the volume-rescaled sampling distribution from [10, Section 3], whose marginals are all equal to the optimal sampling measure ν_m (leverage score sampling). Up to a random permutation, this consists of m samples from γ_m and $n - m$ i.i.d. samples from the optimal sampling measure ν_m. Considering γ_n^ν with $\nu \neq \nu_m$ will further allow us to obtain $H \to L^2_\mu$ quasi-optimality result in probability.

4.3. **Properties of least-squares projection**

In this section, we consider weighted least-squares projection based on volume sampling with reference probability measure $\nu = w^{-1}\mu$. The case $\nu = \nu_m$ corresponds to volume-rescaled sampling and enjoy favorable properties for the error in expectation. However, as we will see, taking ν as a mixture allows us to obtain a control of errors with high probability.

We first state some results on the minimal eigenvalue of the Gram matrix when using volume sampling distribution γ_n^ν. This is a straightforward extension of Theorem 2.9 from [10].

Lemma 4.9. *Assume* $\mathbf{x}$ *is drawn from the distribution* γ_n^ν *with* $\nu = w^{-1}\mu$ *a probability measure. It holds*

$$\mathbb{E}((\mathbf{G}^w)^{-1}) \preceq \frac{n}{n-m+1}\mathbf{I},$$

where the Loewner ordering $\preceq$ *is replaced by an equality whenever the matrix* $\Phi^w(\mathbf{y})$ *for* $\mathbf{y} \sim \nu^{\otimes n}$ *has rank* m *almost surely, and*

$$\mathbb{E}(\lambda_{min}(\mathbf{G}^w)^{-1}) \leq \frac{nm}{n-m+1}.$$

Proof. See Appendix C. ∎

Provided a condition on a minimal number of samples, the next result improves the above upper bound by exploiting a matrix concentration inequality.

Lemma 4.10. *Assume* $\mathbf{x}$ *is drawn from the distribution* γ_n^ν *with* $\nu = w^{-1}\mu$ *and* $w^{-1} \geq \alpha w_m^{-1}$. *Then*

$$\mathbb{P}\left(\lambda_{min}(\mathbf{G}^w)^{-1} > (1-\delta)^{-1}\frac{n}{n-m}\right) \leq m\exp\left(-\frac{c_\delta(n-m)\alpha}{m}\right).$$

Moreover, if

$$n \geq m + mc_\delta^{-1}\alpha^{-1}\log(nm^2)$$

it holds

$$\mathbb{E}(\lambda_{min}(\mathbf{G}^w)^{-1}) \leq 1 + \frac{n}{n-m}(1-\delta)^{-1}.$$

Proof. See Appendix C. ∎

Proposition 4.11. *Let* $\mathbf{x} = (x_1, \ldots, x_n)$ *be drawn from the distribution* γ_n^ν *with* $\nu = w^{-1}\mu$ *and* $w^{-1} \geq \alpha w_m^{-1}$. *Assume we use weighted least-squares with weight function* w. *Then for any function* f, *letting* $S_t = \{\lambda_{min}(\mathbf{G}^w(\mathbf{x}))^{-1} \leq t\}$, *it holds*

$$\mathbb{E}(\|\hat{P}_{V_m}f\|^2|S_t) \leq \mathbb{P}(S_t)^{-1}t\left(1-\frac{m}{n}\right)^{-1}\|f\|^2,$$

and

$$\mathbb{E}(\|\hat{P}_{V_m}f\|^2|S_t) \leq \mathbb{P}(S_t)^{-1}t^2\left(\frac{m}{n}\alpha^{-1}(\beta + \xi m\alpha^{-1})\|f\|^2 + \|P_{V_m}f\|^2\right),$$

with $\beta = 1 + (\alpha^{-1}-1)\frac{m}{n}$, $\xi = 0$ *if* $\nu = \nu_m$, $\xi = 1$ *in the case* $\nu \neq \nu_m$. *If* $n \geq m + c_\delta^{-1}\alpha^{-1}m\log(m\eta^{-1})$ *and* $t = (1-\delta)^{-1}\frac{n}{n-m}$, *then* $\mathbb{P}(S_t) \geq 1-\eta$.

Proof. For the first inequality, we note that

$$\mathbb{E}(\|\hat{P}_{V_m}f\|^2|S_t) \leq t\mathbb{P}(S_t)^{-1}\mathbb{E}(\|\hat{P}_{V_m}f\|_n^2) \leq t\mathbb{P}(S_t)^{-1}\mathbb{E}(\|f\|_n^2),$$

where we have used the fact that $\hat{P}_{V_m}$ is an orthogonal projection with respect to $\|\cdot\|_n$. Then since the marginals of $\mathbf{x}$ are $\tilde{w}^{-1}\mu$ with $\tilde{w}^{-1} \geq (1-\frac{m}{n})w^{-1}$ (Theorem 4.8), we deduce from Lemma 2.1 that $\mathbb{E}(\|f\|_n^2) \leq (1-\frac{m}{n})^{-1}\|f\|^2$. For the second inequality, we note that $\|\hat{P}_{V_m}f\| = \|\mathbf{c}\|_2^2$ with $\mathbf{c} = \mathbf{G}^w(\mathbf{x})^{-1}\mathbf{b}$ and $\mathbf{b} = \frac{1}{n}\Phi^w(\mathbf{x})^T f^w(\mathbf{x})$, where $f^w = fw^{1/2}$. Then noting that $\|\mathbf{c}\|_2 \leq \|\mathbf{G}^w(\mathbf{x})^{-1}\|_2\|\mathbf{b}\|_2 = \lambda_{min}(\mathbf{G}^w(\mathbf{x}))^{-1}\|_2\|\mathbf{b}\|_2$, we have

$$\mathbb{E}(\|\mathbf{c}\|_2^2|S_t) \leq t^2\mathbb{E}(\|\mathbf{b}\|_2^2|S_t) \leq \mathbb{P}(S_t)^{-1}t^2\mathbb{E}(\|\mathbf{b}\|_2^2),$$

and the result follows from Lemma C.3 and Lemma 4.10. ∎

Theorem 4.12. *Assume* $\mathbf{x}$ *is drawn from the distribution* γ_n^ν *with* $\nu = w^{-1}\mu$ *and* $w^{-1} \geq \alpha w_m^{-1}$*, and assume we use weighted least-squares with weight function* w*. If*

$$n \geq m + c_\delta^{-1}\alpha^{-1} m \log(m\eta^{-1}),$$

with $0 < \delta < 1$*, then the event* $S = \{\lambda_{min}(\mathbf{G}^w) \geq (1-\delta)\frac{n-m}{n})\}$ *is such that* $\mathbb{P}(S) \geq 1-\eta$*, and it holds*

$$\mathbb{E}(\|f - \hat{f}_m\|^2 | S) \leq \left(1 + (1-\eta)^{-1}(1-\delta)^{-1}\left(1 - \frac{m}{n}\right)^{-1}\beta\right) \inf_{g \in V_m} \|f - g\|^2,$$

and

$$\mathbb{E}(\|f - \hat{f}_m\|^2 | S) \leq \left(1 + (1-\eta)^{-1}(1-\delta)^{-2}\left(1 - \frac{m}{n}\right)^{-2}\frac{m}{n}\alpha^{-1}(\beta + \xi m \alpha^{-1})\right) \inf_{g \in V_m} \|f - g\|^2,$$

with $\beta = 1 + (\alpha^{-1} - 1)\frac{m}{n}$ *and* $\xi = 1$ *if* $\nu \neq \nu_m$ *or* $\xi = 0$ *if* $\nu = \nu_m$*.*

Proof. Lemma 4.10 implies $\mathbb{P}(S) \geq 1 - \eta$. The marginal distributions are all equal to $\tilde{w}^{-1}\mu$ with $\tilde{w}^{-1} \leq \beta w^{-1}$. Then the first inequality follows from Lemma 2.2, and the second inequality follows from Proposition 4.11. ∎

We next provide a result in probability and another result in expectation (without conditioning) under the assumption that the target function f is in some subspace H of L^2_μ.

Theorem 4.13. *Assume that* $f \in H$*, with* H *satisfying* (2.2) *and* (2.3)*, with* h *a probability density with respect to* μ*. Assume that* $(x_1, \ldots, x_n)$ *is drawn from* γ_n^ν *with* $\nu = w^{-1}\mu$ *and* $w^{-1} = \alpha w_m^{-1} + (1-\alpha)h$*, and we use weighted least-squares with weight function* w*. Then it holds*

$$\|f - \hat{f}_m\| \leq \left(C_H + (1-\delta)^{-1/2}(1-\alpha)^{-1/2}\left(1 - \frac{m}{n}\right)^{-1/2}\right) \inf_{g \in V_m} \|f - g\|_H$$

with probability greater than $1 - m\exp(-\frac{c_\delta(n-m)\alpha}{m})$*, and if*

$$n \geq m + c_\delta^{-1}\alpha^{-1} m \log(nm^2)$$

it holds

$$\mathbb{E}(\|f - \hat{f}_m\|^2) \leq \left(2C_H^2 + 2(1-\alpha)^{-1}\left(1 + (1-\delta)^{-1}\left(1 - \frac{m}{n}\right)^{-1}\right)\right) \inf_{g \in V_m} \|f - g\|_H^2$$

Proof. It is directly deduced from Lemma 4.10 with $\zeta = (1-\alpha)^{-1}$ and Lemma 2.5. ∎

Note that the result in expectation from Theorem 4.13 does not require to use conditioning for ensuring the stability of the Gram matrix.

Remark 4.14. Quasi-optimality guarentees from Theorems 4.12 and 4.13 are obtained under the condition $n \gtrsim m\log(m)$ on the sampling complexity, which is similar to the results from Theorems 3.3 and 3.4, respectively for i.i.d. sampling. The numerical experiments will confirm that the requirement on the number of samples required to obtain stability is similar for volume-rescaled and i.i.d. sampling. However, in terms of approximation errors, we observe in practice that volume-rescaled sampling outperforms i.i.d. sampling.

Unbiased projection and aggregation of projections. We next state a remarkable result, proven in [10, Theorem 3.1] for classical and volume-rescaled sampling, showing that with such sampling, the projection $\hat{f}_m = \hat{P}_{V_m} f$ is an unbiased estimation of the element of best approximation $f_m = P_{V_m} f$. The result is here stated for the distribution γ_n^ν with a general probability measure ν.

Theorem 4.15. *Assume $(x_1, \dots, x_n)$ is drawn from the distribution γ_n^ν with probability measure $\nu = w^{-1}\mu$ and we use weighted least-squares with weight function w. Then for any $f \in L^2_\mu$, it holds*

$$\mathbb{E}(\hat{P}_{V_m} f) = P_{V_m} f.$$

Proof. We have $\hat{P}_{V_m} f(\cdot) = \boldsymbol{\varphi}(\cdot)^T \Phi^w(\mathbf{x})^\dagger f^w(\mathbf{x})$ with $f^w = f w^{1/2}$. Then using Lemma C.2, we obtain $\mathbb{E}(\hat{P}_{V_m} f(\cdot)) = \boldsymbol{\varphi}(\cdot)^T \mathbb{E}(\Phi^w(\mathbf{x})^\dagger f^w(\mathbf{x})) = \boldsymbol{\varphi}(\cdot)^T \int \boldsymbol{\varphi}(y) f(y) d\mu(y) = P_{V_m} f$. ∎

The next result shows a stability of empirical projection in expectation, and hence a quasi-optimality in expectation, which does not require a conditioning to ensure stability of the Gram matrix. It extends [10, Theorem 3.1] to volume sampling with general reference measure ν.

Theorem 4.16. *Assume $(x_1, \dots, x_n)$ is drawn from the distribution γ_n^ν with $\nu = w^{-1}\mu$ such that $w^{-1} \geq \alpha w_m^{-1}$ and we use weighted least-squares with weight function w. Provided $n \geq 2m+2$ and $n \geq 2m\alpha^{-1} c_\delta^{-1} \log(\zeta^{-1} m^2 n)$, it holds*

$$\mathbb{E}(\|\hat{P}_{V_m} g\|^2) \leq \left(4\frac{m}{n}(1-\delta)^{-2}(\beta + \xi m \alpha^{-1}) + \alpha^{-1}\zeta\right)\|g\|^2 + 4(1-\delta)^{-2}\|P_{V_m} g\|^2$$

for any $g \in L^2_\mu$, where $\xi = 0$ for $\nu = \nu_m$ or $\xi = 1$ for $\nu \neq \nu_m$, and $\beta = 1 + (\alpha^{-1} - 1)\frac{m}{n}$. Then provided

$$n \geq C(m \log(\epsilon^{-1} m) + m\epsilon^{-1})$$

for a sufficiently large C, it holds

$$\mathbb{E}(\|f - \hat{P}_{V_m} f\|^2) \leq (1 + \epsilon(1 + \xi m))\|f - P_{V_m} f\|^2 \tag{4.1}$$

with $\xi = 0$ for $\nu = \nu_m$ or $\xi = 1$ for $\nu \neq \nu_m$.

Proof. We have

$$\mathbb{E}(\|f - \hat{P}_{V_m} f\|^2) = \|f - P_{V_m}\|^2 + \mathbb{E}(\|\hat{P}_{V_m}(f - P_{V_m} f)\|^2).$$

Let $g = f - P_{V_m} f$. Note that $\mathbb{E}(\|\hat{P}_{V_m} g\|^2) = \mathbb{E}(\|\Phi^w(\mathbf{x})^\dagger g^w(\mathbf{x})\|_2^2)$. Then using Lemma C.4, we show that provided $n \geq 2m+2$ and $n \geq 2m\alpha^{-1} c_\delta^{-1} \log(\zeta^{-1} m^2 n)$, it holds

$$\mathbb{E}(\|\hat{P}_{V_m} g\|^2) \leq \left(4\frac{m}{n}(1-\delta)^{-2}(\beta + \xi m \alpha^{-1}) + \alpha^{-1}\zeta\right)\|g\|^2$$

with $\beta = 1 + (\alpha^{-1} - 1)\frac{m}{n}$, and $\xi = 0$ if $\nu = \nu_m$ or $\xi = 1$ if $\nu \neq \nu_m$. The condition $n \geq 2m\alpha^{-1}\delta^{-2}\log(\zeta^{-1} m^2 n)$ can be converted into $n \geq C' m \log(\zeta^{-1} m)$ for some C'. Therefore, provided $n \geq C(m\log(\epsilon^{-1} m) + m\epsilon^{-1})$ with a sufficiently large C, it holds

$$\mathbb{E}(\|f - \hat{P}_{V_m} f\|^2) \leq (1 + \epsilon(1 + \xi m))\|f - P_{V_m} f\|^2$$

with $\xi = 0$ for $\nu = \nu_m$ or $\xi = 1$ for $\nu \neq \nu_m$. ∎

The above results allow to analyze the property of an aggregation of r independent least-squares projections based on volume sampling, that yields a quasi-optimality result in expectation (without conditioning), and a convergence to best approximation when $r \to \infty$.

Corollary 4.17. *Let $r \in \mathbb{N}$. Let $\hat{f}^{(1)}, \dots, \hat{f}^{(r)}$ be r independent least-squares projections constructed from independent samples $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(r)}$ drawn from γ_n^ν, with $\nu = w^{-1}\mu$ such that $w^{-1} \geq \alpha w_m^{-1}$, and*

using weighted least-squares with weight w. Then provided $n \geq C(m\log(\epsilon^{-1}m) + m(1+\xi m)\epsilon^{-1})$ with sufficiently large C, the averaged estimator $\bar{f}^r = \frac{1}{r}\sum_{k=1}^r \hat{f}^{(k)}$ satisfies

$$\mathbb{E}(\|f - \bar{f}^r\|^2) \leq \left(1 + \frac{1}{r}(1+\xi m)\epsilon\right)\|f - P_{V_m}f\|^2,$$

with $\xi = 0$ for $\nu = \nu_m$ or $\xi = 1$ for $\nu \neq \nu_m$.

Proof. The estimators $\hat{f}^{(k)}$ are independent and follow the distribution of an estimator $\hat{P}_{V_m}f$ constructed with samples drawn from γ_n^ν. From Theorem 4.15, we have that $\mathbb{E}(\hat{f}^{(k)}) = P_{V_m}f$ for all k. Then using the independence of the $\hat{f}^{(k)}$ and Theorem 4.16, we obtain

$$\begin{aligned}\mathbb{E}(\|f - \bar{f}^r\|^2) &= \|f - P_{V_m}f\|^2 + \mathbb{E}(\|P_{V_m}f - \bar{f}^r\|^2) \\ &\leq \|f - P_{V_m}f\|^2 + \frac{1}{r}\mathbb{E}(\|P_{V_m}f - \hat{P}_{V_m}f\|^2) \\ &\leq \left(1 + \frac{1}{r}(1+\xi m)\epsilon\right)\|f - P_{V_m}f\|^2.\end{aligned}$$

provided $n \geq C(m\log(\epsilon^{-1}m) + m\epsilon^{-1})$ with sufficiently large C. ∎

The quasi-optimality constant $(1 + \frac{1}{r}(1+\xi m)\epsilon)$ is optimal when $\xi = 0$, i.e. $\nu = \nu_m$. When $\nu \neq \nu_m$, having quasi-optimality requires either $\epsilon \sim m^{-1}$ for fixed r, or $r \sim m$ for fixed ϵ, both cases yielding a condition on the total number of samples in $nr \sim m^2$, which is suboptimal compared to the case $\nu = \nu_m$ and even compared with i.i.d. sampling. However, no conditioning is required. Also, there is an interest in using the volume sampling distribution γ_n^ν with $\nu \neq \nu_m$ in order to obtain simultaneously a guaranty in expectation (yet suboptimal) and a guaranty in probability for functions from a specific function space H. Indeed, as a corollary of Theorem 4.13, and under the assumptions of these theorems, we obtain using a simple union bound that

$$\|f - \bar{f}^r\| \leq \left(C_H + (1-\delta)^{-1/2}(1-\alpha)^{-1/2}\left(1 - \frac{m}{n}\right)^{-1}\right)\inf_{g\in V_m}\|f - g\|_H$$

with probability greater than $1 - rm\exp\left(-\frac{c_\delta(n-m)\alpha}{m}\right)$.

5. Least-squares with independent repetitions of volume sampling

We now consider approximation methods relying on independent repetitions from the volume sampling distribution. We will first consider repetitions of projection DPP distribution γ_m and prove some results in expectation. Next we will consider repetitions of general volume sampling distribution γ_n^ν, which allows to obtain results both in expectation and in probability for some specific function spaces, with a suitable choice of the measure ν.

5.1. Independent repetitions of DPP distribution γ_m

We consider $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_r)$ where the $\mathbf{x}_k = (x_{1,k}, \ldots, x_{m,k})$ are i.i.d. samples from γ_m, and the corresponding weighted least-squares projection using $n = mr$ points. We consider least-squares with the optimal weight function w_m. The empirical Gram matrix can be written

$$\mathbf{G}^{w_m}(\mathbf{x}) = \frac{1}{r}\sum_{k=1}^r \mathbf{G}^{w_m}(\mathbf{x}_k),$$

with

$$\mathbf{G}^{w_m}(\mathbf{x}_k) = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\varphi}^{w_m}(x_{i,k})\boldsymbol{\varphi}^{w_m}(x_{i,k})^T = \sum_{i=1}^{m} \frac{\boldsymbol{\varphi}(x_{i,k})\boldsymbol{\varphi}(x_{i,k})^T}{\|\boldsymbol{\varphi}(x_{i,k})\|_2^2}$$

where the $\mathbf{G}^{w_m}(\mathbf{x}_k)$ are i.i.d. matrices with expectation $\mathbf{I}$ and spectral norm bounded by m. Using conditioning, we have the following result.

Theorem 5.1. *Assume* $\mathbf{x}$ *is drawn from the distribution* $\gamma_m^{\otimes r}$*, and assume we use weighted least-squares with weight function* w_m*. Letting* $S_\delta = \{\lambda_{min}(\mathbf{G}^{w_m}(\mathbf{x})) \geq 1-\delta\}$*, it holds*

$$\mathbb{E}(\|f - \hat{f}_m\|^2 | S_\delta) \leq (1 + \mathbb{P}(S_\delta)^{-1}(1-\delta)^{-1}) \inf_{g\in V_m} \|f-g\|^2,$$

and

$$\mathbb{E}(\|f - \hat{f}_m\|^2 | S_\delta) \leq (1 + \mathbb{P}(S_\delta)^{-1}(1-\delta)^{-2}r^{-1}) \inf_{g\in V_m} \|f-g\|^2.$$

Proof. This is a particular case of Theorem 5.8 with $\bar{n} = m$ and $\nu = \nu_m$, where $\gamma_m^\nu = \gamma_m$, $\alpha = 1$ and $\xi = 0$. ∎

It remains to control the probability of the event $S_\delta = \{\lambda_{min}(\mathbf{G}^{w_m}(\mathbf{x})) \geq 1-\delta\}$. From Matrix Chernoff inequality (Lemma 3.1), we deduce that

$$\mathbb{P}(\lambda_{min}(\mathbf{G}^{w_m}(\mathbf{x})) < 1-\delta) \leq m \exp\left(-\frac{rc_\delta}{m}\right).$$

and we conclude that

$$\mathbb{P}(\lambda_{min}(\mathbf{G}^{w_m}(\mathbf{x})) < 1-\delta) \leq \eta$$

provided $n = rm \geq c_\delta^{-1} m^2 \log(m\eta^{-1})$. This result is suboptimal compared to i.i.d. sampling from ν_m, but it does not exploit the properties of DPP, which may yield to matrices $\mathbf{G}^{w_m}(\mathbf{x}_k)$ with spectral norm (much) lower than m with high probability. We have to better analyse the distribution of the random matrix

$$\mathbf{A}(\mathbf{x}) := \mathbf{G}^{w_m}(\mathbf{x}) = \sum_{i=1}^{m} \frac{\boldsymbol{\varphi}(x_i)\boldsymbol{\varphi}(x_i)^T}{\|\boldsymbol{\varphi}(x_i)\|_2^2}, \quad \mathbf{x} = (x_1, \ldots, x_m) \sim \gamma_m.$$

In particular, if the distribution γ_m is such that the $\boldsymbol{\varphi}(x_1), \ldots, \boldsymbol{\varphi}(x_m)$ are close to orthogonal with high probability, then with high probability, $\mathbf{A}(\mathbf{x})$ is close to identity and the least-squares problem is well conditioned.

Example 5.2. An ideal situation occurs when V_m is the space of piecewise constant functions on a uniform partition of $[0,1]$ with m intervals, where $\varphi_j(x) = \sqrt{m}\mathbf{1}_{x\in[(j-1)/m, j/m)}$, $w_m = 1$, and $\boldsymbol{\varphi}(x_i) = \sqrt{m}\mathbf{e}_i$, where $\mathbf{e}_i$ is the i-th canonical vector in $\mathbb{R}^m$. Here the vectors $\boldsymbol{\varphi}(x_1), \ldots, \boldsymbol{\varphi}(x_m)$ are orthogonal almost surely, and $\mathbf{A}(\mathbf{x}) = \mathbf{I}$ almost surely.

Recall that we have

$$\gamma_m(\mathbf{x}) = \frac{1}{m!}\det(\Phi(\mathbf{x})^T\Phi(\mathbf{x}))\mu^{\otimes m} = \frac{m^m}{m!}\det(\mathbf{A}(\mathbf{x}))\nu_m^{\otimes m}(\mathbf{x})$$

so that with $\mathbf{x} \sim \gamma_m$ and $\mathbf{y} \sim \nu_m^{\otimes m}$, it is more likely to have matrices $\mathbf{A}(\mathbf{x})$ with higher determinant than $\mathbf{A}(\mathbf{y})$, and hence higher eigenvalues. This leads us to make the following conjecture.

Conjecture 1. *The distribution* γ_m *satisfies*

$$\mathbb{P}_{\mathbf{z}\sim\gamma_m}(F(\mathbf{A}(\mathbf{z})) > t) \leq \mathbb{P}_{\mathbf{y}\sim\nu_m^{\otimes m}}(F(\mathbf{A}(\mathbf{y})) > t) \tag{5.1}$$

for $t > 0$ *and* F *a real-valued positive, convex and decreasing function in the Loewner order.*

If the distribution γ_m satisfies the property of Conjecture 1, we obtain the following result.

Proposition 5.3. *Let* $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_r) \sim \gamma_m^{\otimes r}$ *with* γ_m *a distribution over* $\mathcal{X}^m$ *satisfying* (5.1). *Then*

$$\mathbb{P}(\lambda_{min}(\mathbf{G}^{w_m}(\mathbf{x})) < 1 - \delta) \leq m \exp\left(-\frac{c_\delta n}{m}\right).$$

Proof. We have $\mathbb{P}(\lambda_{min}(\mathbf{G}^{w_m}(\mathbf{x})) < 1 - \delta) = \mathbb{P}(\lambda_{min}(\mathbf{G}^{w_m}(\mathbf{x}))^{-1} > t)$ with $t := (1-\delta)^{-1}$. The function $F := \mathbf{B} \mapsto \lambda_{min}(\mathbf{B})^{-1}$ is a positive convex and monotonically decreasing function in the Loewner order. For any fixed symmetric positive semi-definite matrix $\mathbf{H}$, $\mathbf{A} \mapsto \lambda_{min}(\mathbf{H} + \mathbf{A}/r)^{-1}$ is also a positive convex and monotonically decreasing function in the Loewner order. Letting $\mathbf{G}^{w_m}(\mathbf{x}) = \frac{1}{r}\sum_{i=1}^m \mathbf{A}(\mathbf{x}_i) := \mathbf{H}(\mathbf{x}_1, \dots, x_{r-1}) + \mathbf{A}(\mathbf{x}_r)/r$, we then have

$$\begin{aligned}
\mathbb{P}(F(\mathbf{G}^{w_m}(\mathbf{x})) > t) &= \mathbb{E}(\mathbf{1}_{F(\frac{1}{r}\sum_{i=1}^r \mathbf{A}(\mathbf{x}_i)) > t}) \\
&= \mathbb{E}(\mathbb{E}(\mathbf{1}_{F(\mathbf{H}(\mathbf{x}_1,\dots,\mathbf{x}_{r-1}) + \mathbf{A}(\mathbf{x}_r)/r) > t} | \mathbf{x}_1, \dots, \mathbf{x}_{r-1})) \\
&\leq \mathbb{E}(\mathbb{E}(\mathbf{1}_{F(\mathbf{H}(\mathbf{x}_1,\dots,\mathbf{x}_{r-1}) + \mathbf{A}(\mathbf{y}_r)/r) > t} | \mathbf{x}_1, \dots, \mathbf{x}_{r-1})) \\
&= \mathbb{E}(\mathbf{1}_{F(\mathbf{H}(\mathbf{x}_1,\dots,\mathbf{x}_{r-1}) + \mathbf{A}(\mathbf{y}_r)/r) > t}),
\end{aligned}$$

where $\mathbf{y}_r \sim \nu_m^{\otimes m}$. Letting $\mathbf{y}_1, \dots, \mathbf{y}_r$ be i.i.d. samples from $\nu_m^{\otimes m}$, and successively conditioning on $(\mathbf{x}_1, \dots, \mathbf{x}_{r-2}, \mathbf{y}_r)$, $(\mathbf{x}_1, \dots, \mathbf{x}_{r-3}, \mathbf{y}_{r-1}, \mathbf{y}_r)$,..., $(\mathbf{y}_2 \dots, \mathbf{y}_r)$, we obtain

$$\mathbb{P}(F(\mathbf{G}^{w_m}(\mathbf{x})) > t) \leq \mathbb{E}(\mathbf{1}_{F(\frac{1}{r}\sum_{i=1}^r \mathbf{A}(\mathbf{y}_i)) > t}) = \mathbb{P}(F(\mathbf{G}^{w_m}(\mathbf{y})) > t),$$

where $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_r) \sim \nu_m^{\otimes n}$, $n = rm$. Then it holds $\mathbb{P}(\lambda_{min}(\mathbf{G}^{w_m}(\mathbf{x})) < 1-\delta) \leq \mathbb{P}(\lambda_{min}(\mathbf{G}^{w_m}(\mathbf{y})) < 1 - \delta)$, and we conclude using Lemma 3.1. ∎

The above result ensures that stability is controlled in probability with a number of samples which is at most the number of samples required by i.i.d. sampling from the optimal distribution ν_m. In numerical experiments, we observe that this number of samples is in fact much lower than with i.i.d. sampling. To be understood, this would require other tools for analyzing the concentration of $\mathbf{G}^{w_m}$.

Remark 5.4. The proof of Proposition 5.3 exploits the assumption (5.1) to prove that

$$\mathbb{P}_{\mathbf{z}\sim\gamma_m}(\lambda_{min}(\mathbf{H} + \mathbf{A}(\mathbf{z})/r)^{-1} > t) \leq \mathbb{P}_{\mathbf{y}\sim\nu_m^{\otimes m}}(\lambda_{min}(\mathbf{H} + \mathbf{A}(\mathbf{y})/r)^{-1} > t) \tag{5.2}$$

for any fixed p.s.d. matrix $\mathbf{H}$. The assumption (5.2) on γ_m would be sufficient to obtain the result of Proposition 5.3.

Remark 5.5. In order to obtain the result of Proposition 5.3, an alternative assumption on γ_m would be that

$$\mathbb{E}_{\mathbf{z}\sim\gamma_m}(G(e^{s\mathbf{A}(\mathbf{z})})) \leq \mathbb{E}_{\mathbf{y}\sim\nu_m^{\otimes m}}(G(e^{s\mathbf{A}(\mathbf{y})})) \tag{5.3}$$

for any $s < 0$ and G a real-valued positive, concave and monotonically increasing function in the Loewner order. Under this assumption, we have to follow the proof of matrix Chernoff. The first steps of the proof of matrix Chernoff inequality (Theorem A.1) yield

$$\mathbb{P}\left(\lambda_{min}\left(\frac{1}{r}\sum_{i=1}^r \mathbf{A}(\mathbf{x}_i)\right) < t\right) \leq \inf_{\theta<0} e^{-\theta t}\mathbb{E}\left(\operatorname{tr}\exp\left(\sum_{i=1}^r \theta\mathbf{A}(\mathbf{x}_i)/r\right)\right).$$

Letting $\frac{1}{r}\sum_{i=1}^r \mathbf{A}(\mathbf{x}_i) := \mathbf{H} + \theta\mathbf{A}(\mathbf{x}_r)/r$, we have

$$\operatorname{tr}\exp\left(\sum_{i=1}^r \theta\mathbf{A}(\mathbf{x}_i)/r\right) = G(e^{\mathbf{A}(\mathbf{x}_r)\theta/r})$$

with $G := \mathbf{X} \mapsto \operatorname{tr}\exp(\mathbf{H} + \log(\mathbf{X}))$ a concave and increasing function in the Loewner order. The assumption then implies that

$$\mathbb{E}\left(\operatorname{tr}\exp\left(\sum_{i=1}^{r}\theta\mathbf{A}(\mathbf{x}_i)/r\right)\right) \leq \mathbb{E}(\operatorname{tr}\exp(\mathbf{H} + \theta\mathbf{A}(\mathbf{y}_r)/r))$$

with $\mathbf{y}_r \sim \nu_m^{\otimes m}$. Then by successive conditioning (as in the proof of Proposition 5.3), we obtain

$$\mathbb{P}\left(\lambda_{min}\left(\frac{1}{r}\sum_{i=1}^{r}\mathbf{A}(\mathbf{x}_i)\right) < t\right) \leq \inf_{\theta<0} e^{-\theta t}\mathbb{E}\left(\operatorname{tr}\exp\left(\sum_{i=1}^{r}\theta\mathbf{A}(\mathbf{y}_i)/r\right)\right),$$

where the $\mathbf{y}_i$ are i.i.d. samples from $\nu_m^{\otimes m}$, and we proceed with the classical proof of matrix Chernoff inequality for sums of i.i.d. matrices (Theorem A.1).

Remark 5.6 (Complexity)**.** Let us assume that μ is a discrete measure with N atoms. From Remark 4.6, we know that getting r independent samples from the DPP distribution costs $O(r(m^3+Nm))$ (up to log factors). With $r \sim \log(m)$, this results in a cost in $O(m^3 + Nm)$ (up to log factors). On the other hand, getting n i.i.d. samples from ν_m costs $O(Nn)$. Then the subsampling algorithm from [2] to obtain a subsample of size $O(m)$ costs $O(nm^3)$. With $n \sim m\log(m)$, this yields a total cost in $O(m^4 + Nm)$ (up to log factors). This shows the advantage of using repeated DPP to directly obtain a sample of size $O(m)$, compared to using i.i.d. sampling and subsampling.

5.2. Independent repetitions of volume sampling $\gamma_{\bar{n}}^{\nu}$

We here consider weighted least-squares projection using a set of samples gathering independent samples from the volume sampling distribution $\gamma_{\bar{n}}^{\nu}$ with $\bar{n} \geq m$ and $\nu = w^{-1}\nu$ with $w^{-1} = \alpha w_m^{-1} + (1-\alpha)h$, where the probability density h is chosen according to some prior assumption on the target function class.

We consider r i.i.d. samples $\mathbf{x}_k = (x_{1,k}, \ldots, x_{\bar{n},k}) \in \mathcal{X}^{\bar{n}}$ from $\gamma_{\bar{n}}^{\nu}$ and the corresponding weighted least-squares minimization with $n = \bar{n}r$ points. The empirical Gram matrix can be written

$$\mathbf{G}^w = \frac{1}{r}\sum_{k=1}^{r}\mathbf{G}^w(\mathbf{x}_k)$$

where the $\mathbf{G}^w(\mathbf{x}_k)$ are i.i.d. matrices with expectation $\mathbf{I}$ and spectral norm bounded by $\alpha^{-1}m$. We start by providing results in expectation.

Proposition 5.7. *Let $\mathbf{x} = (x_1, \ldots, x_n)$ be drawn from the distribution $(\gamma_{\bar{n}}^{\nu})^{\otimes r}$ with $\nu = w^{-1}\mu$ and $w^{-1} \geq \alpha w_m^{-1}$. Assume we use weighted least squares with weight function w. Let $S_\delta = \{\lambda_{min}(\mathbf{G}^w) \geq 1-\delta\}$ with $0 < \delta < 1$. Then for any $f \in L^2_\mu$, it holds*

$$\mathbb{E}(\|\hat{P}_{V_m}f\|^2|S_\delta) \leq \mathbb{P}(S_\delta)^{-1}(1-\delta)^{-1}\beta\|f\|^2,$$

and

$$\mathbb{E}(\|\hat{P}_{V_m}f\|^2|S_\delta) \leq \mathbb{P}(S_\delta)^{-1}(1-\delta)^{-2}\left(\frac{m}{n}\alpha^{-1}(\beta + \xi m\alpha^{-1} + \beta\xi n)\|f\|^2 + (1-\xi+\xi/r)\|P_{V_m}f\|^2\right),$$

with $\beta = 1 + (\alpha^{-1}-1)\frac{m}{\bar{n}}$, $\xi = 0$ if $\nu = \nu_m$, or $\xi = 1$ in the case $\nu \neq \nu_m$.

Proof. The marginal distributions of $\mathbf{x}$ are all equal to $\tilde{\nu} = \tilde{w}^{-1}\mu$ with $\tilde{w}^{-1} \leq \beta w^{-1}$. Then the first inequality directly follows from Lemma 3.2. Then following the proof of Proposition 4.11, we obtain

$$\mathbb{E}(\|\hat{P}_{V_m}f\|_2^2|S_\delta) \leq \mathbb{P}(S_\delta)^{-1}(1-\delta)^{-2}\mathbb{E}(\|\mathbf{b}\|_2^2),$$

with $\mathbf{b} = \frac{1}{n}\Phi^w(\mathbf{x})^T f^w(\mathbf{x}) = \frac{1}{r}\sum_{k=1}^r \mathbf{b}(\mathbf{x}_k)$, where $\mathbf{b}(\mathbf{x}_k) = \frac{1}{\bar{n}}\Phi^w(\mathbf{x}_k)^T f^w(\mathbf{x}_k)$ and $\mathbf{x}_k \sim \gamma_{\bar{n}}^{\nu}$. The $\mathbf{b}(\mathbf{x}_k)$ being i.i.d., it holds

$$\mathbb{E}(\|\mathbf{b}\|_2^2) = \frac{1}{r}\mathbb{E}(\|\mathbf{b}(\mathbf{z})\|_2^2) + \frac{r-1}{r}\|\mathbb{E}(\mathbf{b}(\mathbf{z}))\|_2^2$$

with $\mathbf{z} \sim \gamma_{\bar{n}}^{\nu}$. Using Lemma C.3, we have

$$\mathbb{E}(\|\mathbf{b}(\mathbf{z})\|_2^2) \leq \frac{m}{\bar{n}}\alpha^{-1}(\beta + \xi m\alpha^{-1})\|f\|^2 + \|P_{V_m}f\|^2,$$

with $\beta = 1 + (\alpha^{-1}-1)\frac{m}{\bar{n}}$, $\xi = 0$ if $\nu = \nu_m$ and $\xi = 1$ in the case $\nu \neq \nu_m$. When $\nu = \nu_m$, it holds $\|\mathbb{E}(\mathbf{b}(\mathbf{z}))\|_2^2 = \|\mathbb{E}_{x\sim\nu_m}(\boldsymbol{\varphi}(x)f(x)w_m(x))\|_2^2 = \|P_{V_m}f\|^2$. When $\nu \neq \nu_m$, we have

$$\begin{aligned}\|\mathbb{E}(\mathbf{b}(\mathbf{z}))\|_2^2 &= \|\mathbb{E}_{x\sim\tilde{\nu}}(\boldsymbol{\varphi}(x)f(x)w(x))\|_2^2 \\ &\leq \mathbb{E}_{x\sim\tilde{\nu}}(\|\boldsymbol{\varphi}(x)\|_2^2 f(x)^2 w(x)^2) \\ &\leq m\alpha^{-1}\beta\int f(x)^2 d\mu(x) = m\alpha^{-1}\beta\|f\|^2.\end{aligned}$$

Gathering the above results, we obtain

$$\|\mathbb{E}(\mathbf{b}(\mathbf{z}))\|_2^2 \leq \frac{m}{n}\alpha^{-1}(\beta + \xi m\alpha^{-1})\|f\|^2 + (1-\xi+\xi/r)\|P_{V_m}f\|^2 + m\alpha^{-1}\beta\xi\|f\|^2,$$

which ends the proof. ∎

Theorem 5.8. *Let $r \in \mathbb{N}$, $\bar{n} \geq m$ and $n = \bar{n}r$. Assume $\mathbf{x}$ is drawn from the distribution $(\gamma_{\bar{n}}^{\nu})^{\otimes r}$ with $\nu = w^{-1}\mu$ such that $w^{-1} \geq \alpha w_m^{-1}$, and assume we use weighted least-squares with weight function w. Letting $S_\delta = \{\lambda_{min}(\mathbf{G}^{w_m}(\mathbf{x})) \geq 1-\delta\}$, it holds*

$$\mathbb{E}(\|f - \hat{f}_m\|^2 | S_\delta) \leq (1 + \mathbb{P}(S_\delta)^{-1}(1-\delta)^{-1}\beta) \inf_{g\in V_m}\|f-g\|^2,$$

and

$$\mathbb{E}(\|f - \hat{f}_m\|^2 | S_\delta) \leq \left(1 + \mathbb{P}(S_\delta)^{-1}(1-\delta)^{-2}\left(\frac{m}{n}\alpha^{-1}(\beta + \xi m\alpha^{-1} + \beta\xi n)\right)\right)\inf_{g\in V_m}\|f-g\|^2.$$

with $\beta = 1 + (\alpha^{-1}-1)\frac{m}{\bar{n}}$, $\xi = 0$ if $\nu = \nu_m$, or $\xi = 1$ in the case $\nu \neq \nu_m$.

Proof. This simply results from Lemma 2.2 and Proposition 5.7. ∎

We next provide a result in probability and another result in expectation (without conditioning) under the assumption that the target function f in some subspace H.

Theorem 5.9. *Assume that $f \in H$, with H satisfying (2.2) and (2.3), with h a probability density with respect to μ. Assume that $\mathbf{x} = (x_1, \ldots, x_n)$ is drawn from $(\gamma_{\bar{n}}^{\nu})^{\otimes r}$ with $\nu = w^{-1}\mu$ and $w^{-1} = \alpha w_m^{-1} + (1-\alpha)h$, and we use weighted least-squares with weight function w. Then it holds*

$$\|f - \hat{f}_m\| \leq \left(C_H + (1-\delta)^{-1/2}(1-\alpha)^{-1/2}\right)\inf_{g\in V_m}\|f-g\|_H$$

with probability at least $\mathbb{P}(S_\delta)$, where $S_\delta = \{\lambda_{min}(\mathbf{G}^w(\mathbf{x})) \geq 1-\delta\}$. If

$$\bar{n} \geq m + c_\delta^{-1}\alpha^{-1}m\log(nm^2)$$

it holds

$$\mathbb{E}(\|f - \hat{f}_m\|^2) \leq \left(2C_H^2 + 2(1-\alpha)^{-1}r\left(1 + \left(1 - \frac{m}{\bar{n}}\right)^{-1}(1-\delta)^{-1}\right)\right)\inf_{g\in V_m}\|f-g\|_H^2.$$

Proof. The first inequality is deduced from Lemma 2.5 with $\zeta = (1-\alpha)^{-1}$. For the second inequality, we use Lemma 2.5 with $\zeta = (1-\alpha)^{-1}$ and the fact that for any $1 \leq k \leq r$, it holds

$$\mathbb{E}(\lambda_{min}(\mathbf{G}^w(\mathbf{x}))^{-1}) \leq r\mathbb{E}(\lambda_{min}(\mathbf{G}^w(\mathbf{x}_k))^{-1}) \leq r\left(1 + \left(1 - \frac{m}{\bar{n}}\right)^{-1}(1-\delta)^{-1}\right),$$

where the last inequality comes from Lemma 4.10. ∎

The first statement of Theorem 5.9 is a $H \to L^2_\mu$ quasi-optimality result in probability. The second statement is a $H \to L^2_\mu$ quasi-optimality property in expectation, without conditioning the sample to satisfy the event S_δ.

Again it remains to control the probability of the event S_δ. In fact, the distribution of $\mathbf{G}^w$ is the one of the average of r independent Gram matrices associated with γ_m, and $r(\bar{n}-m)$ rank-one matrices associated with i.i.d. samples from ν. All these $r(\bar{n}-m+1)$ matrices have spectral norm bounded by $\alpha^{-1}m$. Therefore, from matrix Chernoff inequality (Theorem A.1), we deduce that

$$\mathbb{P}(\lambda_{min}(\mathbf{G}^w) > 1-\delta) \leq m\exp\left(-\frac{r(\bar{n}-m+1)c_\delta^{-1}\alpha}{m}\right),$$

and we can conclude that

$$\mathbb{P}(\lambda_{min}(\mathbf{G}^w) > 1-\delta) \leq \eta$$

whenever $r(\bar{n}-m+1) \geq c_\delta^{-1}m\alpha^{-1}\log(m\eta^{-1})$, or the condition $n \geq \frac{\bar{n}}{\bar{n}-m+1}c_\delta^{-1}m\alpha^{-1}\log(m\eta^{-1})$ on the total number of samples. For, e.g., $\bar{n} = 2m$, we obtain a condition $n \geq 2c_\delta^{-1}m\alpha^{-1}\log(m\eta^{-1})$, that is suboptimal (by a factor 2) compared to i.i.d. sampling. Here, we obtain a complexity in $O(m\log(m))$ similar to i.i.d. but this is essentially due to the presence in $\mathbf{x}$ of mr i.i.d. samples from ν. Again, this analysis does not really exploit the properties of volume sampling.

A better understanding of the distribution of matrices $\mathbf{G}^w(\mathbf{x}_k)$ allows to improve the above results. As for the case of determinantal point processes, we conjecture that

$$\mathbb{P}(F(\mathbf{G}^w(\mathbf{x}_k)) > t) \leq \mathbb{P}_{\mathbf{y}\sim\nu^{\otimes\bar{n}}}(F(\mathbf{G}^w(\mathbf{y})) > t) \tag{5.4}$$

for $t > 0$ and F a real-valued positive, convex and monotonically decreasing function in the Loewner order. Under this conjecture, following the proof of Proposition 5.3, we would obtain the following concentration result, similar to the case of $n = m\bar{n}$ i.i.d. sampling from $\nu = w\mu$.

Proposition 5.10. *Let $r \in \mathbb{N}$, $\bar{n} \geq m$ and $n = \bar{n}r$. Assume $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_r) \sim (\gamma_{\bar{n}}^\nu)^{\otimes r}$ with $\gamma_{\bar{n}}^\nu$ a distribution over $\mathcal{X}^{\bar{n}}$ satisfying* (5.4)*. Then it holds*

$$\mathbb{P}(\lambda_{min}(\mathbf{G}^w(\mathbf{x}) < 1-\delta) \leq m\exp\left(-\frac{c_\delta n}{m}\right).$$

Remark 5.11. The assumption (5.4) in Proposition 5.10 could be replaced by a weaker condition of the form (5.2), or an alternative condition of the form (5.3).

We can avoid any conjecture on volume sampling and still obtain a result similar to Proposition 5.10 by assuming that the DPP distribution γ_m satisfies the conjecture (5.1) (or one of the two conjectures (5.2) or (5.3)), and further assuming that $w_m \leq \xi_m w$ for some constant ξ_m (possibly depending on m).

Proposition 5.12. *Let $\bar{n} \geq m$, $r \in \mathbb{N}$ and $n = \bar{n}m$. Let $\mathbf{y} \sim \nu^{\otimes(n-mr)}$ with $\nu = w^{-1}\mu$ such that $w_m \leq \xi_m w$. Let $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_r) \sim \gamma_m^{\otimes r}$ where γ_m is a distribution over $\mathcal{X}^m$ satisfying either* (5.1)*,* (5.2) *or* (5.3)*. Letting $\mathbf{x} = (\mathbf{y}, \mathbf{z})$, it holds*

$$\mathbb{P}(\lambda_{min}(\mathbf{G}^w(\mathbf{x})) < (1-\delta)\xi_m^{-1}\left(\frac{m}{\bar{n}}\right) \leq m\exp(-c_\delta r).$$

Proof. We have

$$\mathbf{G}^w(\mathbf{x}) = \frac{mr}{n}\mathbf{G}^w(\mathbf{z}) + \frac{n-mr}{n}\mathbf{G}^w(\mathbf{y}) \succeq \frac{mr}{n}\mathbf{G}^w(\mathbf{z}) \succeq \frac{mr}{n}\xi_m^{-1}\mathbf{G}^{w_m}(\mathbf{z}).$$

Therefore, using Proposition 5.3 with assumption (5.1) or the alternative assumptions (5.2) or (5.3) (see Remarks 5.4 and 5.5), it holds

$$\mathbb{P}(\lambda_{min}(\mathbf{G}^w(\mathbf{x})) < (1-\delta)\xi_m^{-1}\left(\frac{mr}{n}\right) \leq \mathbb{P}(\lambda_{min}(\mathbf{G}^{w_m}(\mathbf{z})) < 1-\delta) \leq m\exp(-c_\delta r). \qquad \blacksquare$$

Provided $n = \bar{n}r \geq \bar{n}c_\delta^{-1}\log(m\eta^{-1})$, it then holds $\mathbb{P}(\lambda_{min}(\mathbf{G}^w(\mathbf{x}))^{-1} < (1-\delta')^{-1}) \geq 1-\eta$ with $(1-\delta')^{-1} = (1-\delta)^{-1}\xi_m\frac{\bar{n}}{m}$. Theorem 5.8 then gives

$$\mathbb{E}(\|f-\hat{f}_m\|^2|S_{\delta'}) \leq \left(1+(1-\eta)^{-1}\beta\xi_m\frac{\bar{n}}{m}\right)\inf_{v\in V_m}\|f-v\|^2,$$

which is a quasi-optimality result only if ξ_m is uniformly bounded with m.

Remark 5.13. Note that if V_m contains the constant function and $w^{-1} = \alpha w_m^{-1} + (1-\alpha)h$ with $h = 1$, then $w_m^{-1} \geq \frac{1}{m}$ and therefore, $w_m \leq \xi_m w$ with $\xi_m = \alpha + (1-\alpha)m \leq m$.

Note that the above theoretical results only show that repeated volume sampling is not worse than i.i.d. sampling, but numerical experiments reveal that repeated volume sampling clearly outperforms i.i.d. sampling. To be understood, this would again require other tools for analyzing the concentration of $\mathbf{G}^w$.

6. Numerical experiments

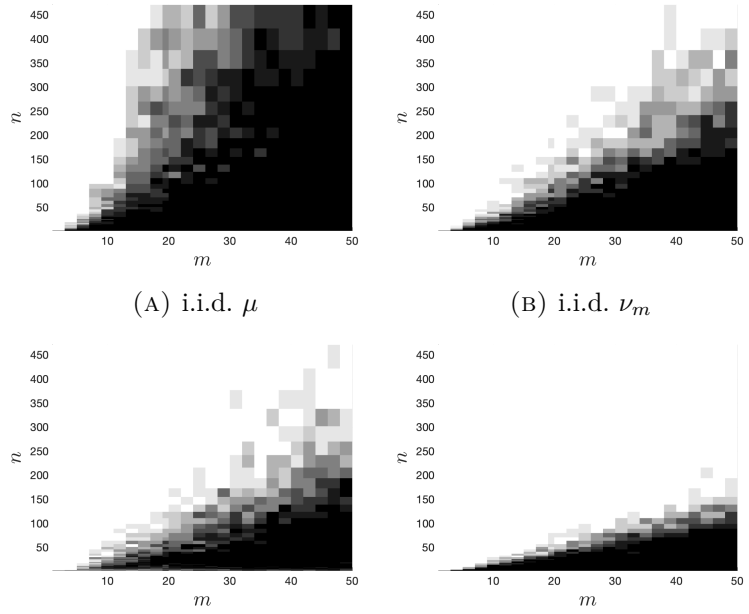
We consider two simple cases of polynomial approximation where V_m is the space of polynomials of degree $m-1$, with either $\mathcal{X} = [-1,1]$ equipped with the uniform measure $\mu \sim U(-1,1)$, or $\mathcal{X} = \mathbb{R}$ equipped with the standard gaussian measure $\mu \sim \mathcal{N}(0,1)$. We compare (weighted) least-squares methods using (i) n i.i.d. samples from μ, (ii) n i.i.d. samples from $\nu_m = w_m^{-1}\mu$, (iii) n samples drawn from volume-rescaled sampling distribution $\gamma_n^{\nu_m}$ (that is equivalent to m samples from γ_m and $n-m$ i.i.d. samples from ν_m), and (iv) n samples from independent repetitions of projection DPP distribution γ_m. In the latter case, we perform $r = \lceil n/m \rceil$ i.i.d. samples from γ_m and keep the first n points.

Concerning sampling, we systematically approximated measures μ by discrete measures with $N = 2000$ atoms, and then relied on exact samplers for discrete distributions. This approximation has no impact on the qualitative results below.

For the case of the uniform distribution over $[-1,1]$, Figure 6.1 shows estimations of the probability to satisfy $S_\delta = \{\lambda_{min}(\mathbf{G}^w) \geq 1-\delta\}$ for the different sampling methods, as a function of m and n. We observe a clear superiority of optimal i.i.d. sampling compared to classical i.i.d. sampling. We also observe that volume-rescaled sampling brings only a small improvement over optimal i.i.d. sampling. Finally, we observe that using independent repetitions of γ_m clearly outperforms volume-rescaled sampling. For i.i.d. optimal sampling and volume-rescaled sampling, we can observe from the figure the condition $n \sim m\log(m)$ to ensure that S_δ is satisfied with probability $1/2$. For repeated DPP, we observe a condition better than $n \sim m\log(m)$, or at least with a much better constant than with other approaches, that is unfortunately not explained by our theoretical findings.

Figure 6.2 illustrates the same quantities for the case of the standard gaussian measure μ over $\mathcal{X} = \mathbb{R}$. We draw essentially the same conclusions. We notice in this case the very poor performance of classical sampling from μ.

From now on, we consider the approximation of the function $f(x) = (1+2x^2)^{-1}$ on $\mathbb{R}$ equipped with the standard Gaussian measure μ, and V_m the space of polynomials of degree $m-1$. On Figure 6.3,



(A) i.i.d. μ (B) i.i.d. ν_m

(C) $\gamma_n^{\nu_m}$ (D) $\gamma_m^{\otimes \lceil n/m \rceil}$

FIGURE 6.1. For V_m the space of polynomials of degree $m-1$ and μ the uniform measure on $[-1,1]$, we plot $\mathbb{P}(\lambda_{min}(\mathbf{G}^w) \geq 1/4)$ as a function of dimension m and number of samples n. Probability estimates go from 0 (black) to 1 (white).

we plot the histograms of the logarithm of the L^2_μ relative error, $\log(\|f - \hat{f}_m\|/\|f\|)$, for the different sampling strategies and using $n = rm$ for different r. For small oversampling ($r = 2$), we observe the benefit of volume-rescaled sampling over i.i.d. sampling, which shifts the distribution towards small values. We also observe a clear superiority of repeated DPP $\gamma_m^{\otimes r}$, which is further improved by conditioning to satisfy the event S_δ with $\delta = 3/4$. For large oversampling ($r = 10$), the histograms are roughly similar. We only observe a slight benefit of conditioned repeated DPP over the other methods, even over i.i.d. optimal sampling.

Table 6.1 shows the expected relative error $\mathbb{E}(\|f - \hat{f}_m\|^2)^{1/2}/\|f\|$ and the quantile of $\|f - \hat{f}_m\|/\|f\|$ of level 95%. We first observe the catastrophic results for classical i.i.d. sampling from μ. For small oversampling ($n = 2m$), we observe on both criteria a clear benefit of volume-rescaled and repeated DPP over i.i.d. optimal sampling. In terms of expected error, we observe a slight improvement of repeated DPP compared to volume-rescaled sampling. However, concerning the quantile, we observe a clear superiority of repeated DPP over volume-rescaled sampling. For the same number of samples, this quantile is divided by up to a factor 2. For larger oversampling ($n = 5m$), i.i.d. performs much better and gets closer to the performance of volume-rescaled sampling and repeated DPP.

The above numerical experiments illustrate the superiority of repeated DPP distribution $\gamma_m^{\otimes r}$ over i.i.d. optimal sampling, but also over volume-rescaled sampling, in the small oversampling regime. For large oversampling, the different sampling strategies yield similar results. The interest of repeated DPP distribution is that for very small oversampling, stability S_δ can be achieved with a reasonable probability. This allows for sampling conditioned to S_δ, which further improves the quality of the least-squares projection.

We observe in Table 6.1 that i.i.d. sampling with conditioning yields almost the same performance as repeated DPP with conditioning. This proves the interest of conditioning. However, we were not able to generate an i.i.d. sample of size $n = 2m$ in reasonable time for $m \geq 30$ (crosses in Table 1).

m	best	i.i.d. μ	i.i.d. ν_m	i.i.d. ν_m (cond.)	$\gamma_n^{\nu_m}$	$\gamma_m^{\otimes n/m}$	$\gamma_m^{\otimes n/m}$ (cond.)
10	$1.3e-01$	$8.4e+02$	$4.7e-01$	$1.7e-01$	$2.0e-01$	$1.7e-01$	$1.6e-01$
20	$5.1e-02$	$3.7e+07$	$1.5e-01$	$6.4e-02$	$8.2e-02$	$6.7e-02$	$6.3e-02$
30	$2.6e-02$	$3.3e+10$	$2.5e-01$	$\times$	$4.1e-02$	$3.5e-02$	$3.2e-02$
40	$1.4e-02$	$1.2e+10$	$7.2e-02$	$\times$	$2.3e-02$	$1.8e-02$	$1.7e-02$
50	$7.9e-03$	$4.9e+09$	$3.4e-02$	$\times$	$1.3e-02$	$1.1e-02$	$9.6e-03$

(A) $n = 2m$. Expected relative error $\mathbb{E}(\|f - \hat{f}_m\|^2)^{1/2}/\|f\|$.

m	best	i.i.d. μ	i.i.d. ν_m	i.i.d. ν_m (cond.)	$\gamma_n^{\nu_m}$	$\gamma_m^{\otimes n/m}$	$\gamma_m^{\otimes n/m}$ (cond.)
10	$1.3e-01$	$2.8e+03$	$1.2e+00$	$2.3e-01$	$3.4e-01$	$2.5e-01$	$2.2e-01$
20	$5.1e-02$	$1.7e+08$	$4.2e-01$	$8.0e-02$	$1.4e-01$	$9.6e-02$	$8.2e-02$
30	$2.6e-02$	$9.3e+10$	$3.2e-01$	$0.0e+00$	$6.6e-02$	$5.5e-02$	$4.0e-02$
40	$1.4e-02$	$4.2e+10$	$1.6e-01$	$0.0e+00$	$4.3e-02$	$2.4e-02$	$2.2e-02$
50	$7.9e-03$	$1.5e+10$	$1.1e-01$	$0.0e+00$	$2.3e-02$	$1.5e-02$	$1.2e-02$

(B) $n = 2m$. Quantile of $\|f - \hat{f}_m\|/\|f\|$ of level 95%.

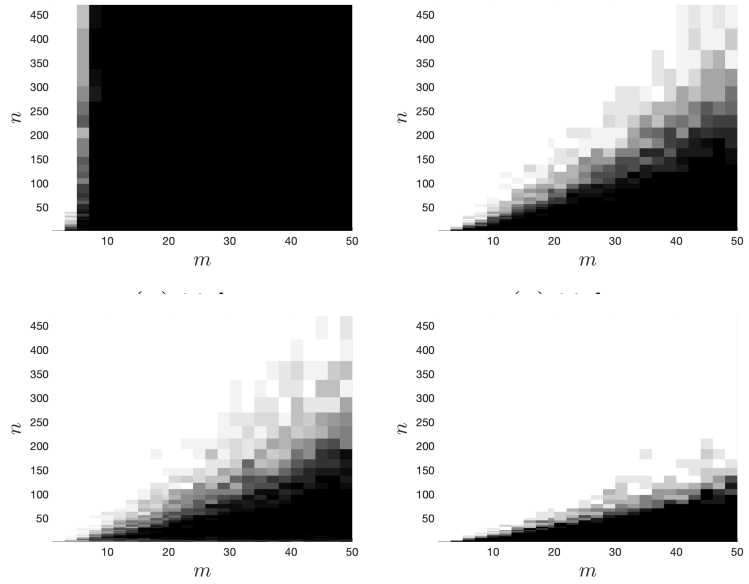
m	best	i.i.d. μ	i.i.d. ν_m	i.i.d. ν_m (cond.)	$\gamma_n^{\nu_m}$	$\gamma_m^{\otimes n/m}$	$\gamma_m^{\otimes n/m}$ (cond.)
10	$1.3e-01$	$9.0e+01$	$1.5e-01$	$1.5e-01$	$1.5e-01$	$1.4e-01$	$1.4e-01$
20	$5.1e-02$	$8.7e+05$	$5.8e-02$	$5.6e-02$	$5.7e-02$	$5.4e-02$	$5.4e-02$
30	$2.4e-02$	$1.3e+08$	$2.9e-02$	$2.8e-02$	$2.8e-02$	$2.6e-02$	$2.6e-02$
40	$1.3e-02$	$4.6e+07$	$1.6e-02$	$1.5e-02$	$1.5e-02$	$1.4e-02$	$1.4e-02$
50	$7.8e-03$	$2.1e+08$	$9.2e-03$	$8.8e-03$	$9.0e-03$	$8.4e-03$	$8.4e-03$

(C) $n = 5m$. Expected relative error $\mathbb{E}(\|f - \hat{f}_m\|^2)^{1/2}/\|f\|$.

m	best	i.i.d. μ	i.i.d. ν_m	i.i.d. ν_m (cond.)	$\gamma_n^{\nu_m}$	$\gamma_m^{\otimes n/m}$	$\gamma_m^{\otimes n/m}$ (cond.)
10	$1.32e-01$	$3.4e+02$	$1.9e-01$	$1.7e-01$	$1.8e-01$	$1.6e-01$	$1.6e-01$
20	$5.1e-02$	$3.8e+06$	$7.1e-02$	$6.5e-02$	$6.9e-02$	$6.1e-02$	$6.0e-02$
30	$2.4e-02$	$4.6e+08$	$3.9e-02$	$3.3e-02$	$3.2e-02$	$3.0e-02$	$2.9e-02$
40	$1.3e-02$	$1.9e+08$	$1.9e-02$	$1.8e-02$	$1.8e-02$	$1.6e-02$	$1.5e-02$
50	$7.8e-03$	$9.7e+08$	$1.1e-02$	$1.0e-02$	$1.1e-02$	$9.2e-03$	$9.2e-03$

(D) $n = 5m$. Quantile of $\|f - \hat{f}_m\|/\|f\|$ of level 95%.

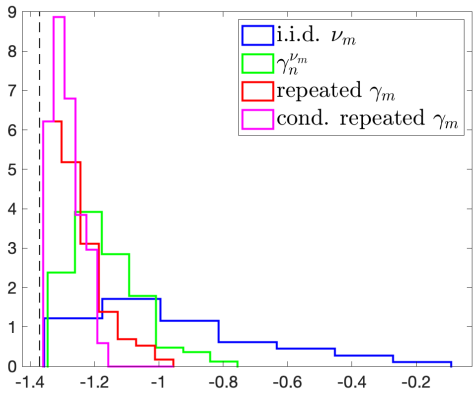
TABLE 6.1. For V_m the space of polynomials of degree $m-1$ and μ the standard gaussian measure on $\mathbb{R}$, we indicate the expected relative error or the quantile of the relative error of level 95%, using $n = rm$ samples from $\nu_m^{\otimes n}$, the volume-rescaled distribution $\gamma_n^{\nu_m}$, repeated DPP $\gamma_m^{\otimes r}$, or repeated DPP conditioned to S_δ with $\delta = 3/4$. The column "best" indicates the best approximation error in V_m.



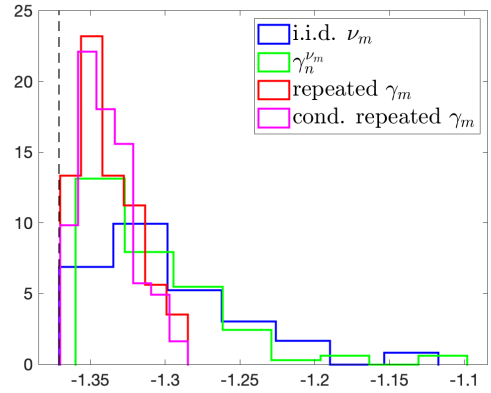
(c) $\gamma_m + n - m$ i.i.d. ν_m (d) repeated γ_m

FIGURE 6.2. For V_m the space of polynomials of degree $m-1$ and μ the standard gaussian measure on $\mathbb{R}$, we plot $\mathbb{P}(\lambda_{min}(\mathbf{G}^w) \geq 1/4)$ as a function of dimension m and number of samples n. Probability estimates go from 0 (black) to 1 (white).

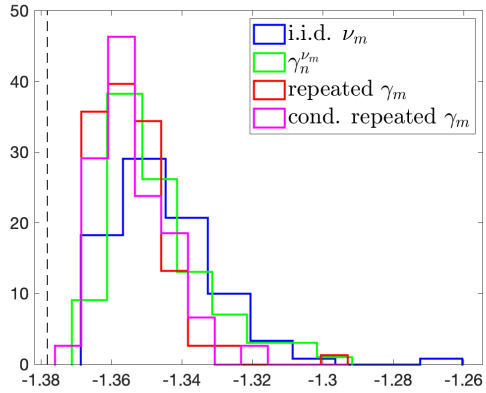
This is explained by the fact that using i.i.d. sampling requires a high number of samples to satisfy S_δ with a reasonable probability. This proves the advantage of using repeated DPP, which allows to use a conditioning technique with a very small sample size (even $2m$).



(A) $n = 2m$



(B) $n = 5m$



(C) $n = 10m$

FIGURE 6.3. For V_m the space of polynomials of degree $m-1$ and μ the standard gaussian measure on $\mathbb{R}$, we plot the histogram of $\log(\|f - \hat{f}_m\|/\|f\|)$ using $n = rm$ samples from $\nu_m^{\otimes n}$ (blue), the volume-rescaled sampling distribution $\gamma_n^{\nu_m}$ (green), repeated DPP $\gamma_m^{\otimes r}$ (red), or repeated DPP conditioned to satisfy S_δ with $\delta = 3/4$ (magenta).

Appendix A. Some known results on random matrices

Theorem A.1 (Matrix Chernoff inequality [29]). *Let $\mathbf{A}_1, \dots \mathbf{A}_n$ be independent random symmetric matrices of size m-by-m such that for all i, $0 \leq \lambda_{min}(\mathbf{A}_i)$ and $\lambda_{max}(\mathbf{A}_i) \leq L$. Then*

$$\mathbb{P}\left(\lambda_{max}\left(\sum_{i=1}^n \mathbf{A}_i\right) \geq (1+\delta)\mu_{max}\right) \leq m \exp(-d_\delta \mu_{max}/L) \quad \textit{for } \epsilon \geq 0, \quad \textit{and} \tag{A.1}$$

$$\mathbb{P}\left(\lambda_{min}\left(\sum_{i=1}^n \mathbf{A}_i\right) \leq (1-\delta)\mu_{min}\right) \leq m \exp(-c_\delta \mu_{min}/L) \quad \textit{for } \epsilon \in [0,1), \tag{A.2}$$

where $\mu_{min} = \lambda_{min}(\mathbb{E}(\sum_{i=1}^n \mathbf{A}_i))$, $\mu_{max} = \lambda_{max}(\mathbb{E}(\sum_{i=1}^n \mathbf{A}_i))$, $c_\delta = \delta + (1-\delta)\log(1-\delta)$ and $d_\delta = -\delta + (1+\delta)\log(1+\delta)$. It holds $\frac{5}{13}\delta^2 \leq d_\delta \leq \delta^2/2 \leq c_\delta \leq \delta^2$.

Proof. We provide a sketch of the proof of Tropp [30] for the bound on the minimal eigenvalue. Let $\mathbf{B} = \sum_{i=1}^n \mathbf{A}_i$. For any $\theta < 0$, it holds

$$\mathbb{P}(\lambda_{min}(\mathbf{B}) \leq t) = \mathbb{P}(e^{\theta \lambda_{min}(\mathbf{B})} \geq e^{\theta t}) = \mathbb{P}(e^{\lambda_{min}(\theta \mathbf{B})} \geq e^{\theta t}) \leq e^{-\theta t}\mathbb{E}(e^{\lambda_{min}(\theta \mathbf{B})}),$$

where the last inequality is given by Markov inequality. Then using $e^{\lambda_{min}(\theta \mathbf{B})} = \lambda_{min}(e^{\theta \mathbf{B}}) \leq \mathrm{tr}(e^{\theta \mathbf{B}})$, we obtain

$$\mathbb{P}(\lambda_{min}(\mathbf{B}) \leq t) \leq \inf_{\theta < 0} e^{-\theta t}\mathbb{E}(\mathrm{tr} e^{\theta \mathbf{B}}).$$

For positive-definite matrix $\mathbf{H}$, the map $\mathbf{A} \mapsto \mathrm{tr} e^{\mathbf{H} + \log(\mathbf{A})}$ is concave on the positive cone of positive definite matrices. Letting $\mathbf{X}_i := e^{\theta \mathbf{A}_i}$, a sequential application of Jensen's inequality gives

$$\mathbb{E}(\mathrm{tr} e^{\theta \mathbf{B}}) = \mathbb{E}\left(\mathrm{tr}\exp\left(\sum_{i=1}^n \log(\mathbf{X}_i)\right)\right) \leq \mathrm{tr}\exp\left(\sum_{i=1}^n \log(\mathbb{E}(\mathbf{X}_i))\right) = \mathrm{tr}\exp\left(\sum_{i=1}^n \log(\mathbb{E}(e^{\theta \mathbf{A}_i}))\right).$$

Also it holds $\log \mathbb{E}(e^{\theta \mathbf{A}_i}) \preceq g(\theta)\mathbb{E}(\mathbf{A}_i)$ where $g(\theta) = L^{-1}(e^{\theta L} - 1)$, so that

$$\mathrm{tr}\exp\left(\sum_{i=1}^n \log(\mathbb{E}(e^{\theta \mathbf{A}_i}))\right) \leq \mathrm{tr}\exp\left(g(\theta)\mathbb{E}\left(\sum_{i=1}^n \mathbf{A}_i\right)\right) \leq n e^{g(\theta)\mu_{min}}.$$

Therefore,

$$\mathbb{P}(\lambda_{min}(\mathbf{B}) \leq t) \leq \inf_{\theta < 0} n e^{-\theta t} e^{g(\theta)\mu_{min}}.$$

Taking $t = (1-\delta)\mu_{min}$, the infimum is attained at $\theta = L^{-1}\log(1-\delta)$, which gives the desired result. ∎

Lemma A.2 (Lemma 2.3 in [10]). *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times m}$ be two random matrices whose row vectors are drawn as an i.i.d. sequence of n pairs of random vectors $(\mathbf{a}_i, \mathbf{b}_i)$. Then*

$$n^m \mathbb{E}(\det(\mathbf{A}^T\mathbf{B})) = \frac{n!}{(n-m)!}\det(\mathbb{E}(\mathbf{A}^T\mathbf{B})) \quad \textit{for any } n \geq m, \textit{ and}$$

$$n^{m-1}\mathbb{E}(\mathrm{adj}(\mathbf{A}^T\mathbf{B})) = \frac{n!}{(n-m+1)!}\mathrm{adj}(\mathbb{E}(\mathbf{A}^T\mathbf{B})) \quad \textit{for any } n \geq m-1.$$

Lemma A.3 (Lemma 2.11 in [10]). *For any matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ with $n > m$, it holds*

$$\det(\mathbf{A}^T\mathbf{A})\mathbf{A}^\dagger = \frac{n}{n-m}\sum_{i=1}^n \det(\mathbf{A}^T(\mathbf{I} - \mathbf{e}_i\mathbf{e}_i^T)\mathbf{A})((\mathbf{I} - \mathbf{e}_i\mathbf{e}_i^T)\mathbf{A})^\dagger.$$

Appendix B. Properties of projection determinental point processes

Proof of Proposition 4.1. This is a standard result on projection determinantal processes, see e.g. [21]. We here provide a short proof with our notations. The fact that all marginals are the same comes from the invariance to permutations of the distribution γ_m. From the classical "base times height formula" for a determinant, we have

$$\det(\Phi(\mathbf{x})) = \det((\boldsymbol{\varphi}(x_1), \dots \boldsymbol{\varphi}(x_m))) = \|\boldsymbol{\varphi}(x_1)\|_2 \|P_{W_1^\perp}\boldsymbol{\varphi}(x_2)\|_2 \dots \|P_{W_{m-1}^\perp}\boldsymbol{\varphi}(x_m)\|_2$$

where $P_{W_k^\perp} = I_m - P_{W_k}$ is the orthogonal projection onto the orthogonal complement of $W_k = \mathrm{span}\{\boldsymbol{\varphi}(x_1), \dots, \boldsymbol{\varphi}(x_k)\}$ in $\mathbb{R}^m$. Therefore, the density of γ_m with respect to $\mu^{\otimes m}$ has the following expression

$$\frac{1}{m!}\det(\Phi(\mathbf{x})^T\Phi(\mathbf{x})) = \prod_{k=1}^m p_k(x_k), \quad p_k(x_k) := \frac{1}{m-k+1}\|P_{W_{k-1}^\perp}\boldsymbol{\varphi}(x_k)\|_2^2,$$

with the convention $W_0 = \{0\}$. The function p_k depends on $(x_1, \dots, x_{k-1})$ and is a probability density since

$$\begin{aligned}\int p_k(x)d\mu(x) &= \frac{1}{m-k+1}\int \boldsymbol{\varphi}(x)^T P_{W_{k-1}^\perp}\boldsymbol{\varphi}(x)d\mu(x) \\ &= \frac{1}{m-k+1}\mathrm{tr}(P_{W_{k-1}^\perp}\int \boldsymbol{\varphi}(x)\boldsymbol{\varphi}(x)^T d\mu(x)) \\ &= \frac{1}{m-k+1}\mathrm{tr}(P_{W_{k-1}^\perp}) = 1,\end{aligned}$$

where we have used the fact that $\mathrm{tr}(P_{W_{k-1}^\perp}) = \dim(W_{k-1}^\perp) = m-k+1$. That provides a factorization of γ_m in terms of the marginal $p_1(x_1)d\mu(x_1) = \frac{1}{m}\|\boldsymbol{\varphi}(x_1)\|_2^2 d\mu$ and the conditional distributions $p_k(x_k)d\mu(x_k)$ of x_k knowing $(x_1, \dots, x_{k-1})$, which ends the proof. ∎

The next result provides the distribution of all marginal distributions of γ_m. This is a standard result on DPPs (see, e.g., [10, Lemma 3.3]). For completeness, we here provide a proof with our notations.

Proposition B.1. *Let $\mathbf{x} = (x_1, \dots, x_m) \sim \gamma_m$. For a nonempty tuple $T \subset [m]$, $\mathbf{x}_T = (x_i)_{i\in T}$ has for distribution*

$$\frac{(m-|T|)!}{m!}\det(\Phi(\mathbf{y})\Phi(\mathbf{y})^T)d\mu^{\otimes|T|}(\mathbf{y}), \quad \mathbf{y} \in \mathcal{X}^{|T|}.$$

Proof. Because of the symmetry of the distribution γ_m, it is sufficient to consider sets $T = [k]$ and $\mathbf{x}_T = \mathbf{x}_{[k]} = (x_1, \dots, x_k)$. Let denote $p_{[k]}\mu^{\otimes k}$ the distribution of $\mathbf{x}_{[k]}$. The result is true for $k = 1$. Then we proceed by induction. From Proposition 4.1, we know that the conditional distribution of x_{k+1} knowing $\mathbf{x}_{[k]}$ is

$$\frac{1}{m-k}(\|\boldsymbol{\varphi}(x_{k+1})\|_2^2 - \|\Phi(\mathbf{x}_{[k]})(\Phi(\mathbf{x}_{[k]})\Phi(\mathbf{x}_{[k]})^T)^{-1}\Phi(\mathbf{x}_{[k]})^T\boldsymbol{\varphi}(x_{k+1})\|_2^2)d\mu.$$

Assuming the distribution of $\mathbf{x}_{[k]}$ is

$$\frac{(m-k)!}{m!}\det(\Phi(\mathbf{x}_{[k]})\Phi(\mathbf{x}_{[k]})^T)d\mu^{\otimes k}(\mathbf{x}_{[k]}),$$

we deduce that the distribution of $\mathbf{x}_{[k+1]} = (\mathbf{x}_{[k]}, x_{k+1})$ admits as density with respect to $\mu^{\otimes(k+1)}$

$$\begin{aligned}
&\frac{(m-k-1)!}{m!}\det(\Phi(\mathbf{x}_{[k]})\Phi(\mathbf{x}_{[k]})^T)(\|\boldsymbol{\varphi}(x_{k+1})\|_2^2 \\
&\quad - \|\Phi(\mathbf{x}_{[k]})(\Phi(\mathbf{x}_{[k]})\Phi(\mathbf{x}_{[k]})^T)^{-1}\Phi(\mathbf{x}_{[k]})^T\boldsymbol{\varphi}(x_{k+1})\|_2^2) \\
&= \frac{(m-k-1)!}{m!}\det(\Phi(\mathbf{x}_{[k]})\Phi(\mathbf{x}_{[k]})^T)(\boldsymbol{\varphi}(x_{k+1})^T\boldsymbol{\varphi}(x_{k+1}) \\
&\quad - \boldsymbol{\varphi}(x_{k+1})^T\Phi(\mathbf{x}_{[k]})(\Phi(\mathbf{x}_{[k]})\Phi(\mathbf{x}_{[k]})^T)^{-1}\Phi(\mathbf{x}_{[k]})^T\boldsymbol{\varphi}(x_{k+1})) \\
&= \frac{(m-k-1)!}{m!}\det(\Phi(\mathbf{x}_{[k+1]})\Phi(\mathbf{x}_{[k+1]}^T)),
\end{aligned}$$

which ends the proof. ∎

Appendix C. Properties of volume sampling

We first provide a straightforward result.

Lemma C.1. *Assume $\mathbf{x} = (x_1, \dots, x_n)$ is drawn from the distribution γ_n^ν with $\nu = w^{-1}\mu$ a probability measure. For any measurable function $g : \mathcal{X}^n \to \mathbb{R}$,*

$$\mathbb{E}_{\mathbf{x}\sim\gamma_n^\nu}(g(\mathbf{x})) = \mathbb{E}_{\mathbf{y}\sim\nu^{\otimes n}}\left(\frac{(n-m)!}{n!}\det(\Phi^w(\mathbf{y})^T\Phi^w(\mathbf{y}))g(\mathbf{y})\right).$$

We next provide a generalization of [10, Theorem 2.10] which is fundamental to prove the unbiasedness of weighted least-squares projections based on volume sampling with general reference measures.

Lemma C.2. *Assume $\mathbf{x} = (x_1, \dots, x_n)$ is drawn from the distribution γ_n^ν with $\nu = w^{-1}\mu$ a probability measure. Then for any function f, it holds*

$$\mathbb{E}(\Phi^w(\mathbf{x})^\dagger f^w(\mathbf{x})) = \int \boldsymbol{\varphi}(y)f(y)d\mu(y),$$

where $f^w = fw^{1/2}$.

Proof. First consider the case $n = m$, where $\mathbf{x} = (x_1, \dots, x_m)$ is drawn from $\gamma_m = \mathrm{DPP}_\mu(V_m)$. In this case, $\Phi(\mathbf{x})$ is a square matrix, almost surely invertible, and $\Phi^w(\mathbf{x})^\dagger f^w(\mathbf{x}) = \Phi(\mathbf{x})^\dagger f(\mathbf{x})$. We obtain using Cramer's rule[1]

$$\begin{aligned}
\mathbb{E}(\Phi(\mathbf{x})^\dagger f(\mathbf{x}))_i &= \frac{1}{m!}\int \left(\det(\Phi(\mathbf{y})^T\Phi(\mathbf{y}))\Phi(\mathbf{y})^\dagger f(\mathbf{y})\right)_i d\mu(y) \\
&= \frac{1}{m!}\int \det(\Phi(\mathbf{y})^T)\det(\Phi(\mathbf{y}) + (f(\mathbf{y}) - \Phi(\mathbf{y})\mathbf{e}_i)\mathbf{e}_i^T)d\mu(y) \\
&= \frac{1}{m!}\int \det(\Phi(\mathbf{y})^T(\Phi(\mathbf{y}) + (f(\mathbf{y}) - \Phi(\mathbf{y})\mathbf{e}_i)\mathbf{e}_i^T))d\mu(y) \\
&\overset{(*)}{=} \det\left(\frac{1}{m}\int \Phi(\mathbf{y})^T(\Phi(\mathbf{y}) + (f(\mathbf{y}) - \Phi(\mathbf{y})\mathbf{e}_i)\mathbf{e}_i^T)d\mu(y)\right) \\
&= \det\left(\mathbf{I} + \left(\int \boldsymbol{\varphi}(y)f(y)d\mu(y) - \mathbf{e}_i\right)\mathbf{e}_i^T\right) = \left(\int \boldsymbol{\varphi}(y)f(y)d\mu(y)\right)_i,
\end{aligned}$$

[1]For any matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and vector $\mathbf{b} \in \mathbb{R}^n$, $\det(\mathbf{A})(\mathbf{A}^\dagger\mathbf{b})_i = \det(\mathbf{A} + (\mathbf{b} - \mathbf{A}\mathbf{e}_i)\mathbf{e}_i^T)$

where $(*)$ is deduced from Lemma A.2. For the case $n > m$, we proceed by induction. Letting $\mathbf{y} \sim \nu^{\otimes n}$ and using Lemma A.3, we have

$$\begin{aligned}
\mathbb{E}(\Phi^w(\mathbf{x})^\dagger f^w(\mathbf{x})) &= \frac{(n-m)!}{n!}\mathbb{E}(\det(\Phi^w(\mathbf{y})^T\Phi^w(\mathbf{y}))\Phi^w(\mathbf{y})^\dagger f^w(\mathbf{y})) \\
&= \frac{(n-m)!}{n!}\frac{n}{n-m}\sum_{i=1}^n \mathbb{E}(\det(\Phi^w(\mathbf{y})^T(\mathbf{I}-\mathbf{e}_i\mathbf{e}_i^T)\Phi^w(\mathbf{y}))((\mathbf{I}-\mathbf{e}_i\mathbf{e}_i^T)\Phi^w(\mathbf{y}))^\dagger f^w(\mathbf{y})) \\
&= \frac{(n-m-1)!}{(n-1)!}\sum_{i=1}^n \mathbb{E}(\det(\Phi^w(\mathbf{y}_{-i})^T\Phi^w(\mathbf{y}_{-i}))(\Phi^w(\mathbf{y}_{-i}))^\dagger f^w(\mathbf{y}_{-i})) \\
&= \mathbb{E}(\Phi^w(\tilde{\mathbf{x}})^{-1}f^w(\tilde{\mathbf{x}})),
\end{aligned}$$

with $\tilde{\mathbf{x}} \sim \gamma_{n-1}^\nu$, and $\mathbf{y}_{-i}$ is the vector $\mathbf{y}$ without the i-th component. We then deduce $\mathbb{E}(\Phi^w(\mathbf{x})^\dagger f^w(\mathbf{x})) = \int \boldsymbol{\varphi}(y)f(y)d\mu(y)$. ∎

Lemma C.3. *Assume $(x_1,\dots,x_n)$ is drawn from the distribution γ_n^ν with $\nu = w^{-1}\mu$ and $w^{-1} \le \alpha w_m^{-1}$. Then for any function f and $f^w = fw^{1/2}$, it holds*

$$\mathbb{E}\left(\left\|\frac{1}{n}\Phi^w(\mathbf{x})^T f^w(\mathbf{x})\right\|^2\right) \le \frac{m}{n}\alpha^{-1}(\beta + \xi m\alpha^{-1})\|f\|^2 + \|P_{V_m}f\|^2$$

with $\beta = 1 + (\alpha^{-1}-1)\frac{m}{n}$ and $\xi = 0$ if $\nu = \nu_m$ or $\xi = 1$ if $\nu \neq \nu_m$. In the case where $\boldsymbol{\varphi}(x_i)f(x_i) \ge 0$ almost surely, it holds

$$\mathbb{E}\left(\left\|\frac{1}{n}\Phi^w(\mathbf{x})^T f^w(\mathbf{x})\right\|^2\right) \le \frac{m}{n}\alpha^{-1}\beta\|f\|^2 + \left(1 + 2\alpha^{-2}\frac{m}{n}\right)\|P_{V_m}f\|^2.$$

Proof. Let $\mathbf{b} = \frac{1}{n}\Phi^w(\mathbf{x})^T f^w(\mathbf{x}) = \frac{1}{n}\sum_{k=1}^n \boldsymbol{\varphi}(x_k)^T f(x_k)w(x_k)$. Letting $\mathbf{a}(x) := \boldsymbol{\varphi}(x)f(x)w(x)$, we have

$$\mathbb{E}(\|\mathbf{b}\|_2^2) = \frac{1}{n^2}\sum_{k,l=1}^n \mathbb{E}((\mathbf{a}(x_k),\mathbf{a}(x_l))) = \frac{1}{n^2}\sum_{k=1}^n \mathbb{E}(\|\mathbf{a}(x_k)\|^2) + \frac{1}{n^2}\sum_{\substack{k,l=1\\k\neq l}}^n \mathbb{E}((\mathbf{a}(x_k),\mathbf{a}(x_l))).$$

The marginal distribution of γ_n^ν is $\tilde{\nu} = \tilde{w}^{-1}\mu$ with $\tilde{w}^{-1} = \frac{n-m}{n}w^{-1} + \frac{m}{n}w_m^{-1}$ such that $\tilde{w}^{-1} \le \beta w^{-1}$ with $\beta = 1 + (\alpha^{-1}-1)\frac{m}{n}$. Then,

$$\begin{aligned}
\frac{1}{n^2}\sum_{k=1}^n \mathbb{E}(\|\mathbf{a}(x_k)\|^2) &= \frac{1}{n}\mathbb{E}_{x\sim\tilde{\nu}}(\|\boldsymbol{\varphi}(x)\|_2^2 f(x)^2 w(x)^2) \\
&\le \frac{m}{n}\alpha^{-1}\beta\,\mathbb{E}_{x\sim\tilde{\nu}}(f(x)^2\tilde{w}(x)) \\
&= \frac{m}{n}\alpha^{-1}\beta\|f\|^2.
\end{aligned}$$

Up to a permutation, we can now consider that $(x_1,\dots,x_m) \sim \gamma_m$ and $(x_{m+1},\dots,x_n) \sim \nu^{\otimes(n-m)}$ are independent. Letting $z \sim \nu$, we have

$$\begin{aligned}
\mathbb{E}\left(\sum_{\substack{k,l=1\\k\neq l}}^n (\mathbf{a}(x_k),\mathbf{a}(x_l))\right) &= m(m-1)\mathbb{E}((\mathbf{a}(x_1),\mathbf{a}(x_2))) \\
&\quad + 2m(n-m)(\mathbb{E}(\mathbf{a}(x_1)),\mathbb{E}(\mathbf{a}(z))) + (n-m)(n-m-1)\|\mathbb{E}(\mathbf{a}(z))\|_2^2.
\end{aligned}$$

We have

$$\|\mathbb{E}(\mathbf{a}(z))\|_2 = \|\mathbb{E}(\boldsymbol{\varphi}(z)f(z)w(z))\|_2 = \|\int \boldsymbol{\varphi}(y)f(y)d\mu(y)\|_2^2 = \|P_{V_m}f\|_2.$$

Letting $(y_1, y_2) \sim \mu^{\otimes 2}$, and using Proposition B.1, we obtain

$$\begin{aligned} m(m-1)&\mathbb{E}((\mathbf{a}(x_1), \mathbf{a}(x_2))) \\ &= \int (\mathbf{a}(y_1), \mathbf{a}(y_2)) \det(\Phi(y_1, y_2)\Phi(y_1, y_2)^T) d\mu(y_1) d\mu(y_2) \\ &= \int (\mathbf{a}(y_1), \mathbf{a}(y_2))(\|\boldsymbol{\varphi}(y_1)\|^2 \|\boldsymbol{\varphi}(y_2)\|^2 - (\boldsymbol{\varphi}(y_1), \boldsymbol{\varphi}(y_2))^2) d\mu(y_1) d\mu(y_2) \\ &\leq \left\| \int \mathbf{a}(y_1) \|\boldsymbol{\varphi}(y_1)\|^2 d\mu(y_1) \right\|_2^2 - \int (\boldsymbol{\varphi}(y_1), \boldsymbol{\varphi}(y_2))^3 f(y_1) f(y_2) w(y_1) w(y_2) d\mu(y_1) d\mu(y_2). \end{aligned}$$

The second term in the above upper bound can be written

$$\int \sum_l g_l(y_1) g_l(y_2) d\mu(y_1) d\mu(y_2) = \sum_l \int g_l(y))^2 d\mu(y)$$

for some functions g_l, so that

$$m(m-1)(\mathbb{E}(\mathbf{a}(x_1)), \mathbb{E}(\mathbf{a}(x_2))) \leq m^2 \|\mathbb{E}(\mathbf{a}(x))\|_2^2,$$

with $x \sim \nu_m$. Gathering the above results, we get

$$\begin{aligned} \mathbb{E}(\|\mathbf{b}\|_2^2) \leq \frac{m}{n} \alpha^{-1} \beta \|f\|^2 + \frac{m^2}{n^2} \|\mathbb{E}(\mathbf{a}(x))\|_2^2 \\ + \frac{2m(n-m)}{n^2} \|\mathbb{E}(\mathbf{a}(x))\|_2 \|P_{V_m} f\| + \frac{(n-m)(n-m-1)}{n^2} \|P_{V_m} f\|^2. \end{aligned}$$

If $w = w_m$, then $\alpha = \beta = 1$ and $\|\mathbb{E}(\mathbf{a}(x))\|_2 = \|P_{V_m} f\|$, and therefore

$$\begin{aligned} \mathbb{E}(\|\mathbf{b}\|_2^2) &\leq \frac{m}{n} \|f\|^2 + \frac{m^2 + 2m(n-m) + (n-m)(n-m-1)}{n^2} \|P_{V_m} f\|^2 \\ &\leq \frac{m}{n} \|f\|^2 + \frac{n^2 - n + m}{n^2} \|P_{V_m} f\|^2 \leq \frac{m}{n} \|f\|^2 + \left(1 - \frac{n-m}{n^2}\right) \|P_{V_m} f\|^2. \end{aligned}$$

If $w^{-1} \geq \alpha w_m^{-1}$, we have

$$\|\mathbb{E}(\mathbf{a}(x))\|_2 \leq \mathbb{E}(\|\mathbf{a}(x)\|_2^2)^{1/2} = m^{1/2} \left(\int f(y)^2 w(y)^2 w_m(y)^{-2} d\mu(y) \right)^{1/2} \leq m^{1/2} \alpha^{-1} \|f\|,$$

and therefore, letting $\xi_m = m^{1/2}$,

$$\begin{aligned} \mathbb{E}(\|\mathbf{b}\|_2^2) &\leq \frac{m}{n} \alpha^{-1} \beta \|f\|^2 + \frac{m^2}{n^2} \xi_m^2 \alpha^{-2} \|f\|^2 \\ &\quad + \frac{2m(n-m)}{n^2} \xi_m \alpha^{-1} \|f\| \|P_{V_m} f\| + \frac{(n-m)(n-m-1)}{n^2} \|P_{V_m} f\|^2 \\ &\leq \left(\frac{m}{n} \alpha^{-1} \beta + \frac{m}{n} \xi_m^2 \alpha^{-2} \right) \|f\|^2 + \frac{(n-m)(n-m-1) + m(n-m)}{n^2} \|P_{V_m} f\|^2 \\ &\leq \frac{m}{n} \alpha^{-1} (\beta + \xi_m^2 \alpha^{-1}) \|f\|^2 + \|P_{V_m} f\|^2. \end{aligned}$$

If the particular case where $\boldsymbol{\varphi}(y) f(y) \geq 0$ almost surely, then

$$\|\mathbb{E}(\mathbf{a}(x))\|_2 = \left\| \int \boldsymbol{\varphi}(y) f(y) w(y) w_m(y)^{-1} d\mu(y) \right\|_2 \leq \alpha^{-1} \left\| \int \boldsymbol{\varphi}(y) f(y) d\mu(y) \right\|_2 = \alpha^{-1} \|P_{V_m} f\|,$$

and we get

$$\begin{aligned} \mathbb{E}(\|\mathbf{b}\|_2^2) &\leq \frac{m}{n} \alpha^{-1} \beta \|f\|^2 + \left(\frac{m^2}{n^2} \alpha^{-2} + \frac{2m(n-m)}{n^2} \alpha^{-1} + \frac{(n-m)(n-m-1)}{n^2} \right) \|P_{V_m} f\|^2 \\ &\leq \frac{m}{n} \alpha^{-1} \beta \|f\|^2 + \left(1 + 2 \frac{m}{n} \alpha^{-2} \right) \|P_{V_m} f\|^2. \end{aligned}$$

■

Proof of Lemma 4.9. The first statement results from [10, Theorem 2.9]. We here provide the proof for completeness. First note that $\mathbf{G}^w(\mathbf{x})$ is invertible almost surely. Letting $\mathbf{y} \sim \nu^{\otimes n}$, we have

$$\begin{aligned}\mathbb{E}(\mathbf{G}^w(\mathbf{x})^{-1}) = \mathbb{E}(\mathbf{G}^w(\mathbf{x})^\dagger) &= n^m \frac{(n-m)!}{n!}\mathbb{E}(\mathbf{G}^w(\mathbf{y})^\dagger \det(\mathbf{G}^w(\mathbf{y}))) \\ &\preceq n^m \frac{(n-m)!}{n!}\mathbb{E}(\mathrm{adj}(\mathbf{G}^w(\mathbf{y}))) \\ &\overset{(*)}{=} n^m \frac{(n-m)!}{n!}\frac{n!}{(n-m+1)!n^{m-1}}\mathrm{adj}(\mathbb{E}(\mathbf{G}^w(\mathbf{y}))) \\ &= \frac{n}{n-m+1}\,\mathrm{adj}(\mathbf{I}) = \frac{n}{n-m+1}\,\mathbf{I},\end{aligned}$$

where $(*)$ is obtained from Lemma A.2, and where the Loewner ordering $\preceq$ can be replaced by an equality when $\Phi^w(\mathbf{y})$ has rank m almost surely, which implies that $\mathbf{G}^w(\mathbf{y})^\dagger = \mathbf{G}^w(\mathbf{y})^{-1}$. We deduce from the first statement that $\mathbf{G} := \mathbf{G}^w(\mathbf{x})$ satisfies $\mathbb{E}(\lambda_{min}(\mathbf{G})^{-1}) = \mathbb{E}(\lambda_{max}(\mathbf{G}^{-1})) \leq \mathbb{E}(\mathrm{tr}(\mathbf{G}^{-1})) = \mathrm{tr}(\mathbb{E}(\mathbf{G}^{-1})) \leq \frac{nm}{n-m+1}$. ∎

Proof of Lemma 4.10. The distribution of $\mathbf{G}^w$ is the same as the Gram matrix associated with m samples from γ_m and $n-m$ i.i.d. samples from ν, independent from the first m samples. Then write $\mathbf{G}^w = \frac{m}{n}\mathbf{G}^w_{[m]} + \frac{n-m}{n}\mathbf{G}^w_{[m]^c}$, where $\mathbf{G}^w_{[m]^c}$ is the Gram matrix associated with $n-m$ i.i.d. samples from ν, and $\mathbf{G}^w_{[m]}$ is the Gram matrix associated with m points from the distribution γ_m. Matrices $\mathbf{G}^w_{[m]}$ and $\mathbf{G}^w_{[m]^c}$ are independent. First note that for $\mathbf{A}$ and $\mathbf{B}$ symmetric positive definite, $\lambda_{min}(\mathbf{A}+\mathbf{B})^{-1} \leq \lambda_{min}(\mathbf{A})^{-1}$. We then deduce that

$$\lambda_{min}(\mathbf{G}^w)^{-1} \leq \frac{n}{m}\lambda_{min}(\mathbf{G}^w_{[m]})^{-1} \quad \text{and} \quad \lambda_{min}(\mathbf{G}^w)^{-1} \leq \frac{n}{n-m}\lambda_{min}(\mathbf{G}^w_{[m]^c})^{-1}.$$

Noting that $K_{m,w} \leq \alpha^{-1}m$, we have from Lemma 3.1 that the event $S = \{\lambda_{min}(\mathbf{G}^w_{[m]^c}) < 1-\delta\}$ satisfies $\mathbb{P}(S) \leq m\exp(-\frac{c_\delta(n-m)\alpha}{m}) := \eta(n-m,m)$. We deduce that

$$\mathbb{P}\left(\lambda_{min}(\mathbf{G}^w)^{-1} > (1-\delta)^{-1}\frac{n}{n-m}\right) \leq \mathbb{P}\left(\lambda_{min}(\mathbf{G}^w_{[m]^c})^{-1} > (1-\delta)^{-1}\right) \leq \eta(n-m,m),$$

that is the second statement. For the final statement, we have that

$$\begin{aligned}\mathbb{E}(\lambda_{min}(\mathbf{G}^w)^{-1}) &\leq \mathbb{E}(\lambda_{min}(\mathbf{G}^w)^{-1}|S)\eta(n-m,m) + \mathbb{E}(\lambda_{min}(\mathbf{G}^w)^{-1}|S^c) \\ &\leq \frac{n}{m}\mathbb{E}(\lambda_{min}(\mathbf{G}^w_{[m]})^{-1})\eta(n-m,m) + \frac{n}{n-m}\mathbb{E}(\lambda_{min}(\mathbf{G}^w_{[m]^c})^{-1}|S^c) \\ &\leq nm\eta(n-m,m) + \frac{n}{n-m}(1-\delta)^{-1}\end{aligned}$$

where we have used the independence of $\mathbf{G}^w_{[m]}$ and S, and the second statement of Lemma 4.9 applied to $\mathbf{G}^w_{[m]}$. Then it holds

$$\mathbb{E}(\lambda_{min}(\mathbf{G}^w)^{-1}) \leq nm^2\exp\left(-\frac{c_\delta(n-m)\alpha}{m}\right) + \frac{n}{n-m}(1-\delta)^{-1}$$

which concludes the proof. ∎

Lemma C.4. *Let $\mathbf{x} \sim \gamma_n^\nu$ with $\nu = w^{-1}\mu$ and $w^{-1} \geq \alpha w_m^{-1}$. For an arbitrary function g, provided $n \geq 2m+2$ and $n \geq 2m\alpha^{-1}c_\delta^{-1}\log(\zeta^{-1}m^2n)$, it holds*

$$\mathbb{E}(\|\Phi^w(\mathbf{x})^\dagger g^w(\mathbf{x})\|_2^2) \leq \left(4\frac{m}{n}(1-\delta)^{-2}(\beta + \xi m\alpha^{-1}) + \alpha^{-1}\zeta\right)\|g\|^2 + 4(1-\delta)^{-2}\|P_{V_m}g\|^2$$

with $\beta = 1 + (\alpha^{-1}-1)\frac{m}{n}$, and $\xi = 0$ if $\nu = \nu_m$ or $\xi = 1$ if $\nu \neq \nu_m$.

Proof. First note that $\Phi^w(\mathbf{x})^\dagger g^w(\mathbf{x}) = \mathbf{G}^w(\mathbf{x})^{-1}\mathbf{b}(\mathbf{x})$ with $\mathbf{b}(\mathbf{x}) = \frac{1}{n}\Phi^w(\mathbf{x})^T g^w(\mathbf{x})$. Up to a reordering, assume that $(x_1, \dots, x_n) \sim \gamma_m \otimes \nu^{\otimes n-m}$. Let $m \le s \le n$. Then write $\mathbf{G}^w(\mathbf{x}) := \frac{s}{n}\mathbf{G}^w_{[s]} + \frac{n-s}{n}\mathbf{G}^w_{[s]^c}$, where $\mathbf{G}^w_{[s]^c}$ is the Gram matrix associated with $n-s$ i.i.d. samples from ν, and $\mathbf{G}^w_{[s]}$ is the Gram matrix associated with s points from the distribution γ^ν_s. Matrices $\mathbf{G}^w_{[s]}$ and $\mathbf{G}^w_{[s]^c}$ are independent. Let $A = \{\lambda_{min}(\mathbf{G}^w_{[s]^c}) \ge (1-\delta)\frac{n-s-1}{n-s}\}$. We have

$$\mathbb{E}(\|\Phi^w(\mathbf{x})^\dagger g^w(\mathbf{x})\|_2^2) = \mathbb{E}(\|\Phi^w(\mathbf{x})^\dagger g^w(\mathbf{x})\|_2^2|A)\mathbb{P}(A) + \mathbb{E}(\|\Phi^w(\mathbf{x})^\dagger g^w(\mathbf{x})\|_2^2|A^c)\mathbb{P}(A^c)$$

For the first term, we have

$$\begin{aligned}\mathbb{E}(\|\Phi^w(\mathbf{x})^\dagger g^w(\mathbf{x})\|_2^2|A)\mathbb{P}(A) &\le \mathbb{E}(\|\mathbf{G}^w(\mathbf{x})^{-1}\|_2^2\|\mathbf{b}\|_2^2|A)\mathbb{P}(A)\\ &\le \frac{n^2}{(n-s)^2}\mathbb{E}(\lambda_{min}(\mathbf{G}^w_{[s]^c}(\mathbf{x}))^{-2}\|\mathbf{b}\|_2^2|A)\mathbb{P}(A)\\ &\le \frac{n^2}{(n-s-1)^2}(1-\delta)^{-2}\mathbb{E}(\|\mathbf{b}\|_2^2),\end{aligned}$$

and using Lemma C.3, we obtain

$$\mathbb{E}(\|\Phi^w(\mathbf{x})^\dagger g^w(\mathbf{x})\|_2^2|A)\mathbb{P}(A) \le \frac{n^2}{(n-s-1)^2}(1-\delta)^{-2}\left(\frac{m}{n}\alpha^{-1}(\beta + \xi m\alpha^{-1})\|g\|^2 + \|P_{V_m}g\|^2\right)$$

with $\beta = 1 + (\alpha^{-1}-1)\frac{m}{n}$, and $\xi = 0$ if $\nu = \nu_m$ or $\xi = 1$ if $\nu \ne \nu_m$.
For the second term, noting that

$$\|\Phi^w(\mathbf{x})^\dagger\|_2^2 = \|(\Phi^w(\mathbf{x})^T\Phi^w(\mathbf{x}))^{-1}\|_2 = n^{-1}\|\mathbf{G}^w(\mathbf{x})^{-1}\|_2 \le s^{-1}\lambda_{min}(\mathbf{G}^w_{[s]})^{-1},$$

we have

$$\begin{aligned}\mathbb{E}(\|\Phi^w(\mathbf{x})^\dagger g^w(\mathbf{x})\|_2^2|A^c) &\le s^{-1}\mathbb{E}(\lambda_{min}(\mathbf{G}^w_{[s]})^{-1}\|g^w(\mathbf{x})\|_2^2|A^c)\\ &= s^{-1}\mathbb{E}(\lambda_{min}(\mathbf{G}^w_{[s]})^{-1}\|g^w(\mathbf{x}_{[s]})\|_2^2) + s^{-1}\mathbb{E}(\lambda_{min}(\mathbf{G}^w_{[s]})^{-1})\mathbb{E}(\|g^w(\mathbf{x}_{[s]^c})\|_2^2|A^c)\\ &\le \mathbb{E}(\lambda_{min}(\mathbf{G}^w_{[s]})^{-1}g^w(x_1)^2) + \frac{m}{s-m+1}\mathbb{E}(\|g^w(\mathbf{x}_{[s]^c})\|_2^2|A^c),\end{aligned}$$

where we have the invariance through permutations of γ_s and the independence of $\mathbf{G}^w_{[s]}$ and $g^w(\mathbf{x}_{[s]^c})$. Letting $B = \{\lambda_{\min}(\mathbf{G}^w_{[s+1,n-1]^c}) \ge (1-\delta)\} \subset A$, we have

$$\begin{aligned}\mathbb{E}(\|g^w(\mathbf{x}_{[s]^c})\|_2^2|A^c) &= \sum_{i=s+1}^n \mathbb{E}(g^w(x_i)^2|A^c) = (n-s)\mathbb{E}(g^w(x_n)^2|A^c)\\ &\le (n-s)\mathbb{E}(g^w(x_n)^2|B^c)\frac{\mathbb{P}(B^c)}{\mathbb{P}(A^c)} = (n-s)\mathbb{E}(g^w(x_n)^2)\frac{\mathbb{P}(B^c)}{\mathbb{P}(A^c)}\end{aligned}$$

so that

$$\mathbb{E}(\|g^w(\mathbf{x}_{[s]^c})\|_2^2|A^c)\mathbb{P}(A^c) \le (n-s)\|g\|^2 m\exp\left(-\frac{c_\delta\alpha(n-s-1)}{m}\right)$$

Finally, using Lemma C.5, we obtain

$$\begin{aligned}\mathbb{E}(\lambda_{min}(\mathbf{G}^w_{[s]})^{-1}g^w(x_1)^2) &\le \mathbb{E}(\mathrm{tr}((\Phi^w(\mathbf{x}_{[s]})^T\Phi^w(\mathbf{x}_{[s]}))^{-1})g^w(x_1)^2)\\ &\le \alpha^{-1}\frac{m}{s-m+1}\mathbb{E}_{x\sim\nu}(g^w(x)^2)\\ &= \alpha^{-1}\frac{m}{s-m+1}\|g\|^2.\end{aligned}$$

Gathering the above results, we have

$$\begin{aligned}\mathbb{E}(\|\Phi^w(\mathbf{x})^\dagger g^w(\mathbf{x})\|_2^2) &\leq \frac{n^2}{(n-s-1)^2}(1-\delta)^{-2}\left(\frac{m}{n}\alpha^{-1}(\beta+\xi m\alpha^{-1})\|g\|^2+\|P_{V_m}g\|^2\right)\\ &\quad+\frac{m^2}{s-m+1}((n-s)+\alpha^{-1})\|g\|^2\exp\left(-\frac{c_\delta\alpha(n-s-1)}{m}\right)\\ &\leq C\|g\|^2+D\|P_{V_m}\|^2\end{aligned}$$

with

$$D=\frac{n^2}{(n-s-1)^2}(1-\delta)^{-2}$$

and

$$C=\frac{n^2}{(n-s-1)^2}(1-\delta)^{-2}\frac{m}{n}\alpha^{-1}(\beta+\xi m\alpha^{-1})+\alpha^{-1}\frac{m^2(n-s+1)}{s-m+1}\exp\left(-\frac{c_\delta\alpha(n-s-1)}{m}\right)$$

Taking $m\leq s\leq n/2-1$, we have

$$D\leq 4(1-\delta)^{-2}$$

and

$$C\leq 4\frac{m}{n}(1-\delta)^{-2}(\beta+\xi m\alpha^{-1})+\alpha^{-1}m^2n\exp(-\frac{c_\delta\alpha n}{2m}).$$

Therefore, provided $n\geq 2m+2$ and

$$n\geq 2m\alpha^{-1}c_\delta^{-1}\log(\zeta^{-1}m^2n),$$

it holds

$$C\leq 4\frac{m}{n}(1-\delta)^{-2}(\beta+\xi m\alpha^{-1})+\alpha^{-1}\zeta.$$ ∎

Lemma C.5. *Let $\mathbf{x}\sim\gamma_n^\nu$ with $\nu=w^{-1}\mu$ satisfying $w^{-1}\geq\alpha w_m^{-1}$. Then for any function f, it holds*

$$\mathbb{E}(f(x_1)\mathrm{tr}((\Phi^w(\mathbf{x})^T\Phi^w(\mathbf{x}))^{-1}))\leq\left(\frac{m}{n}+\alpha^{-1}\frac{m(m-1)}{n(n-m+1)}\right)\mathbb{E}_{x\sim\nu}(f(x)),$$

with an equality if $\Phi^w(\mathbf{y})$ is almost surely of rank m for $\mathbf{y}\sim\nu^{\otimes n}$. In the particular case where $\nu=\nu_m$, it holds

$$\mathbb{E}(f(x_1)\mathrm{tr}((\Phi^{w_m}(\mathbf{x})^T\Phi^{w_m}(\mathbf{x}))^{-1}))\leq\frac{m}{n-m+1}\mathbb{E}_{x\sim\nu}(f(x)),$$

with an equality if $\Phi^{w_m}(\mathbf{x})$ is almost surely of rank m for $\mathbf{y}\sim\nu_m^{\otimes m}$.

Proof. The proof follows the one of [10, Lemma 3.4] for the particular case $\nu=\nu_m$. For completeness, we here detail the proof for a general case. Letting $\mathbf{y}\sim\nu^{\otimes n}$ and $\mathbf{A}(\mathbf{x}):=\Phi^w(\mathbf{x})$, we have

$$\begin{aligned}\mathbb{E}(f(x_1)\mathrm{tr}((\Phi^w(\mathbf{x})^T\Phi^w(\mathbf{x}))^{-1}))&=\mathbb{E}(f(x_1)\mathrm{tr}((\mathbf{A}(\mathbf{x})^T\mathbf{A}(\mathbf{x}))^\dagger))\\&\leq\frac{(n-m)!}{n!}\mathbb{E}(f(y_1)\mathrm{tr}(\mathrm{adj}(\mathbf{A}(\mathbf{y})^T\mathbf{A}(\mathbf{y})))),\end{aligned}$$

where the inequality becomes an equality when $\Phi^w(\mathbf{y})$ is almost surely full rank. From the Cauchy–Binet formula, we have

$$\begin{aligned}\mathbb{E}(f(y_1)\mathrm{tr}(\mathrm{adj}(\mathbf{A}(\mathbf{y})^T\mathbf{A}(\mathbf{y}))))&=\frac{1}{n-m+1}\sum_{i=1}^n\mathbb{E}(f(y_1)\mathrm{tr}(\mathrm{adj}(\mathbf{A}(\mathbf{y}_{-i})^T\mathbf{A}(\mathbf{y}_{-i}))))\\&=\frac{1}{n-m+1}\mathbb{E}(f(y_1))\mathbb{E}(\mathrm{tr}(\mathrm{adj}(\mathbf{A}(\mathbf{z})^T\mathbf{A}(\mathbf{z}))))+\frac{n-1}{n-m+1}\mathbb{E}(f(z_1)\mathrm{tr}(\mathrm{adj}(\mathbf{A}(\mathbf{z})^T\mathbf{A}(\mathbf{z}))))\end{aligned}$$

with $\mathbf{z} \sim \nu^{\otimes(n-1)}$. Using Lemma A.2, we have

$$\begin{aligned}\mathbb{E}(\mathrm{tr}(\mathrm{adj}(\mathbf{A}(\mathbf{z})^T\mathbf{A}(\mathbf{z})))) &= \frac{(n-1)!}{(n-1)^{m-1}(n-m)!}\mathrm{tr}(\mathrm{adj}(\mathbb{E}(\mathbf{A}(\mathbf{z})^T\mathbf{A}(\mathbf{z}))))\\ &= \frac{(n-1)!}{(n-1)^{m-1}(n-m)!}\mathrm{tr}(\mathrm{adj}((n-1)\mathbf{I}_m))\\ &= \frac{(n-1)!}{(n-1)^{m-1}(n-m)!}\mathrm{tr}((n-1)^{m-1}\mathbf{I}_m)\\ &= \frac{(n-1)!}{(n-m)!}m.\end{aligned}$$

Letting $B_n := \mathbb{E}_{\mathbf{y}\sim\nu^{\otimes n}}(f(y_1)\mathrm{tr}(\mathrm{adj}(\mathbf{A}(\mathbf{y})^T\mathbf{A}(\mathbf{y}))))$, we have found

$$B_n = \frac{(n-1)!}{(n-m+1)!}m\mathbb{E}_{x\sim\nu}(f(x)) + B_{n-1} = \frac{(n-1)!}{(n-m)!}m\mathbb{E}_{x\sim\nu}(f(x)) + \binom{n-1}{m-2}B_{m-1}$$

where the last equality is obtained by induction. It remains to evaluate B_{m-1}. Let now $\mathbf{z} \sim \nu^{\otimes(m-1)}$. Letting $\mathbf{A}^{-j}(\mathbf{z})$ being the matrix $\mathbf{A}(\mathbf{z})$ without the j-th column, and letting $\mathbf{a}^{-j}(z_1)$ being the first row vector of $\mathbf{A}^{-j}(\mathbf{z})$, that is the vector $\boldsymbol{\varphi}^w(z_1)$ without the j-th entry, we have

$$\begin{aligned}\mathrm{tr}(\mathrm{adj}(\mathbf{A}(\mathbf{z})^T\mathbf{A}(\mathbf{z}))) &= \sum_{j=1}^m \mathrm{adj}(\mathbf{A}(\mathbf{z})^T\mathbf{A}(\mathbf{z}))_{jj} = \sum_{j=1}^m \det(\mathbf{A}^{-j}(\mathbf{z})^T\mathbf{A}^{-j}(\mathbf{z}))\\ &= \sum_{j=1}^m \det(\mathbf{A}^{-j}(\mathbf{z}_{-1})^T\mathbf{A}^{-j}(\mathbf{z}_{-1}) + \mathbf{a}^{-j}(z_1)\mathbf{a}^{-j}(z_1)^T)\\ &= \sum_{j=1}^m \mathbf{a}^{-j}(z_1)^T\mathrm{adj}(\mathbf{A}^{-j}(\mathbf{z}_{-1})^T\mathbf{A}^{-j}(\mathbf{z}_{-1}))\mathbf{a}^{-j}(z_1),\end{aligned}$$

where we have used $\det(\mathbf{A}^{-j}(\mathbf{z}_{-1})^T\mathbf{A}^{-j}(\mathbf{z}_{-1})) = 0$. Then using Lemma A.2, we obtain

$$\begin{aligned}B_{m-1} &= \mathbb{E}(f(z_1)\mathrm{tr}(\mathrm{adj}(\mathbf{A}(\mathbf{z})^T\mathbf{A}(\mathbf{z}))))\\ &= \sum_{j=1}^m \mathbb{E}(f(z_1)\mathbf{a}^{-j}(z_1)^T\mathbb{E}(\mathrm{adj}(\mathbf{A}^{-j}(\mathbf{z}_{-1})^T\mathbf{A}^{-j}(\mathbf{z}_{-1})))\mathbf{a}^{-j}(z_1))\\ &= \frac{(m-2)!}{(m-2)^{m-2}}\sum_{j=1}^m \mathbb{E}(f(z_1)\mathbf{a}^{-j}(z_1)^T\mathrm{adj}(\mathbb{E}(\mathbf{A}^{-j}(\mathbf{z}_{-1})^T\mathbf{A}^{-j}(\mathbf{z}_{-1})))\mathbf{a}^{-j}(z_1))\\ &= \frac{(m-2)!}{(m-2)^{m-2}}\sum_{j=1}^m \mathbb{E}(f(z_1)\mathbf{a}^{-j}(z_1)^T\mathrm{adj}((m-2)\mathbf{I}_{m-1})\mathbf{a}^{-j}(z_1))\\ &= (m-2)!\sum_{j=1}^m \mathbb{E}(f(z_1)\|\mathbf{a}^{-j}(z_1)\|_2^2)\\ &= (m-2)!\sum_{j=1}^m \mathbb{E}(f(z_1)(\|\boldsymbol{\varphi}^w(z_1)\|_2^2 - \varphi_j^w(z_1)^2))\\ &= (m-1)!\,\mathbb{E}(f(z_1)\|\boldsymbol{\varphi}^w(z_1)\|_2^2)\\ &= (m-1)!\,\mathbb{E}(f(z_1)w(z_1)mw_m(z_1)^{-1})\\ &\leq m!\alpha^{-1}\mathbb{E}(f(z_1))\end{aligned}$$

where we have used $\|\boldsymbol{\varphi}^w(z)\|_2^2 = w(z)\|\boldsymbol{\varphi}(z)\|_2^2 \leq \alpha^{-1} w_m(z)\|\boldsymbol{\varphi}(z)\|_2^2 = m$. We finally obtain

$$\begin{aligned}\mathbb{E}(f(x_1)\mathrm{tr}((\Phi^w(\mathbf{x})^T\Phi^w(\mathbf{x}))^{-1})) &\leq \frac{m}{n}\mathbb{E}_{x\sim\nu}(f(x)) + \alpha^{-1}\frac{(n-m)!}{n!}m!\binom{n-1}{m-2}\mathbb{E}_{x\sim\nu}(f(x)) \\ &= \left(\frac{m}{n} + \alpha^{-1}\frac{m(m-1)}{n(n-m+1)}\right)\mathbb{E}_{x\sim\nu}(f(x)).\end{aligned}$$

■

References

[1] Haim Avron and Christos Boutsidis. Faster subset selection for matrices and applications. *SIAM J. Matrix Anal. Appl.*, 34(4):1464–1499, 2013.

[2] Felix Bartel, Martin Schäfer, and Tino Ullrich. Constructive subsampling of finite frames with applications in optimal function recovery. *Appl. Comput. Harmon. Anal.*, 65:209–248, 2023.

[3] Simon Barthelmé, Nicolas Tremblay, and Pierre-Olivier Amblard. A Faster Sampler for Discrete Determinantal Point Processes. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5582–5592. PMLR, 2023.

[4] Ayoub Belhadji, Rémi Bardenet, and Pierre Chainais. Kernel interpolation with continuous volume sampling. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 725–735. PMLR, 2020.

[5] Ayoub Belhadji, Rémi Bardenet, and Pierre Chainais. Signal reconstruction using determinantal sampling. `https://arxiv.org/abs/2310.09437`, 2023.

[6] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer, 2011.

[7] Peter Binev, Albert Cohen, Wolfgang Dahmen, Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk. Convergence rates for greedy algorithms in reduced basis methods. *SIAM J. Math. Anal.*, 43(3):1457–1472, 2011.

[8] L. Bos, S. De Marchi, A. Sommariva, and M. Vianello. Computing Multivariate Fekete and Leja Points by Numerical Linear Algebra. *SIAM J. Numer. Anal.*, 48(5):1984–1999, 2010.

[9] Albert Cohen and Giovanni Migliorati. Optimal weighted least-squares methods. *SMAI J. Comput. Math.*, 3:181–203, 2017.

[10] Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu. Unbiased estimators for random design regression. *J. Mach. Learn. Theory*, 23(1):7539–7584, 2022.

[11] Amit Deshpande, Luis Rademacher, Santosh S Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. *Theory Comput.*, 2(1):225–247, 2006.

[12] Matthieu Dolbeault and Moulay Abdellah Chkifa. Randomized Least-Squares with Minimal Oversampling and Interpolation in General Spaces. *SIAM J. Numer. Anal.*, 62(4):1515–1538, 2024.

[13] Matthieu Dolbeault and Albert Cohen. Optimal pointwise sampling for $L2$ approximation. *J. Complexity*, 68: article no. 101602 (12 pages), 2022.

[14] Matthieu Dolbeault, David Krieg, and Mario Ullrich. A sharp upper bound for sampling numbers in $L2$. *Appl. Comput. Harmon. Anal.*, 63:113–134, 2023.

[15] Alexander Fonarev, Alexander Mikhalev, Pavel Serdyukov, Gleb Gusev, and Ivan Oseledets. Efficient rectangular maximal-volume algorithm for rating elicitation in collaborative filtering. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 141–150. IEEE, 2016.

[16] Sergei A. Goreinov and Eugene E. Tyrtyshnikov. The maximal-volume concept in approximation by low-rank matrices. In *Structured matrices in mathematics, computer science, and engineering I*, volume 280 of *Contemporary Mathematics*, pages 47–52. American Mathematical Society, 2001.

[17] Cécile Haberstich, Anthony Nouy, and Guillaume Perrin. Boosted optimal weighted least-squares. *Math. Comput.*, 91(335):1281–1315, 2022.

[18] Toni Karvonen, Simo Särkkä, and Ken'ichiro Tanaka. Kernel-based interpolation at approximate Fekete points. *Numer. Algorithms*, 87(1):445–468, 2021.

[19] Boris Kashin, Egor Kosov, Irina Limonova, and Vladimir Temlyakov. Sampling discretization and related problems. *J. Complexity*, 71: article no. 101653 (55 pages), 2022.

[20] David Krieg, Kateryna Pozharska, Mario Ullrich, and Tino Ullrich. Sampling projections in the uniform norm. `https://arxiv.org/abs/2401.02220`, 2024.

[21] Frédéric Lavancier, Jesper Møller, and Ege Rubak. Determinantal point process models and statistical inference. *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, 77(4):853–877, 2015.

[22] Yvon Maday, Ngoc Cuong Nguyen, Anthony T. Patera, and S. H. Pau. A general multipurpose interpolation procedure: the magic points. *Commun. Pure Appl. Anal.*, 8:383–404, 2008.

[23] Adam W. Marcus, Daniel A. Spielman, and Nikhil Srivastava. Interlacing families II: Mixed characteristic polynomials and the Kadison–Singer problem. *Ann. Math.*, 182(1):327–350, 2015.

[24] Shahaf Nitzan, Alexander Olevskii, and Alexander Ulanovskii. Exponential frames on unbounded sets. *Proc. Am. Math. Soc.*, 144(1):109–118, 2016.

[25] Arnaud Poinas and Rémi Bardenet. On proportional volume sampling for experimental design in general spaces. *Stat. Comput.*, 33(1): article no. 29 (22 pages), 2022.

[26] Kateryna Pozharska and Tino Ullrich. A Note on Sampling Recovery of Multivariate Functions in the Uniform Norm. *SIAM J. Numer. Anal.*, 60(3):1363–1384, 2022.

[27] Friedrich Pukelsheim. *Optimal design of experiments*, volume 50 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, 2006.

[28] Mathias Sonnleitner and Mario Ullrich. On the power of iid information for linear approximation. `https://arxiv.org/abs/2310.12740`, 2023.

[29] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

[30] Joel A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1-2):1–230, 2015.

[31] Philipp Trunschke and Anthony Nouy. Almost-sure quasi-optimal approximation in reproducing kernel Hilbert spaces. `https://arxiv.org/abs/2407.06674`, 2024.