Optimal design of sampling sets for least-squares signal recovery with the Frank-Wolfe algorithm

Gilles Chardon

Optimal design of sampling sets for least-squares signal recovery with the Frank-Wolfe algorithm

G. Chardon

***Abstract*—We consider the sensors selection problem in a least-squares setting. The sensors selection is replaced by the relaxed problem of designing a sampling density minimizing the number of samples needed to ensure stability of the recovery, shown to be equivalent to the D-optimal design problem. We propose to use the Frank-Wolfe algorithm to solve this optimization problem, with low space and time computational complexity, linear with respect to the number of possible sensors positions. As the optimal densities are usually sparse, sampling points are drawn from the optimized density using resampling methods. The optimization problem and procedure can be easily modified to account for additional design constraints.**

***Index Terms*—Least-squares, sensor selection, convex optimization, optimal design.**

I. Introduction

The problem of sampling signals is of particular importance in digital signal processing, as one necessary step linking the real, analog world, to the discrete computations of digital devices.

In the classical sampling theorem, the compact support of the Fourier transform of a signal is leveraged to allow its perfect reconstruction from its values on a discrete set of finite density. Starting from low-pass signal models and uniform sampling, this line of work continues through sampling of signal with bandpass spectrum [1] or arbitrary known Fourier support in higher dimensions [2], to sampling of signal with unknown Fourier support of finite measure. In this case, vertices of quasi-crystals are easily built sets that are known to be stable sampling sets [3]. Compressed sensing and its applications put once again the question of sampling set design on the foreground of signal processing and harmonic analysis, by considering the recovery of vectors that are sparse in a given basis with a number of samples lower than the dimension of the basis. [4]. The design of compressed sensing measurement schemes is difficult, and usually relies on randomness, either by using random matrices [5], or by choosing random lines from a deterministic matrix [6].

In this article, we consider the least-squares recovery of a signal f defined on a set X. We use a simple linear model where f is approximated by a signal f_m in a space V_m of finite dimension m spanned by m basis functions $(\phi_i)_{1\leq i\leq m}$. An estimate $\hat{f}_m$ of f is found by matching $\hat{f}_m$ to noisy samples of f in a least-squares sense. A suboptimal sampling set can lead to instability of the recovery, i.e. an estimation error $\|\hat{f}_m - f\|_2^2$ larger than the approximation error $\|f_m - f\|_2^2$ or than the noise level by several orders of magnitude.

These instabilities can be reduced by regularizing the estimation of $\hat{f}_m$. In this approach, priors on the noise and on the signal f are needed, which are not necessarily available. Moreover, regularization reduces the variance of the estimation, but also introduces bias. Our approach will be to design a sampling set such that no regularization is required, i.e. that no prior is necessary apart the knowledge that f can be approximated by $f_m \in V_m$ with sufficient precision.

The problem of sensors placement is an important problem in signal processing [7], [8], [9], but such considerations are no limited to this field. Similar problems are found in numerical analysis, in particular when solving PDEs using the reduced basis method [10], or in optimal design of experiments in statistics [11] (in particular, it will be shown that the problem of D-optimal design is equivalent to the problem of stable least-squares recovery).

A particular difficulty of the design of sampling sets is the non-convexity of the optimization problem. Methods such as convex relaxation of the problem as the optimization of a probability density [7], [12], [9], as well as greedy algorithms [8], [13], [14], can be used to construct approximate solutions to the optimal sampling set problem. We will here use the former approach, also known as *variable density compressive sampling* in sparse recovery [15], [16].

Our main goal is the minimization of the number of samples necessary to ensure stability of the reconstruction. To this end, we use a lower bound on the necessary number of samples needed for stability in function of the probability density used to draw the sampling points, introduced in [17]. We propose a fast algorithm for the design of the measurement density to minimize this bound, as well as a sampling scheme to construct the actual measurement set from the optimized density.

A. Contributions

The problem of designing a sampling set for stable least-squares estimation is shown to be equivalent to the D-optimal design problem, in the sense that optimizing a sampling density to minimize the necessary number a samples for stable recovery is equivalent to maximizing the determinant of the information matrix of a linear regression problem. Furthermore, in generic settings, the optimal densities are discrete.

The application of the Frank-Wolfe algorithm to this problem was introduced in [18]. It is here shown that the complexity in time and space of this algorithm is lower than

Gilles Chardon is with the Laboratoire des Signaux et Systèmes (L2S), CentraleSupélec-CNRS-Université Paris-Sud, Université Paris-Saclay, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France. email: gilles.chardon@centralesupelec.fr.

existing methods, with similar or better results. In particular, the complexity is linear with respect to the number of possible sensor positions. This algorithm, yielding sparse densities, is well adapted to the properties of optimal sampling densities.

The algorithm can take in account additional constraints or modified objective functions. As an example, the joint optimization of a common sampling density for two different spaces V_m is demonstrated. The number of samples necessary for stable approximation in both spaces is also given by the optimization algorithm.

Finally, as the optimal sampling densities are sparse, the use of resampling methods is discussed. These methods yield sampling sets that ensure stability with a number of measurements close to the dimension of the space V_m, with a minimal amount of randomness. The code necessary to reproduce the figures is available online[1].

B. Structure of the article

Results on the stability of least-squares recovery are recalled in section II, as well as connections with D-optimal design and other methods. The optimization problem and algorithm are formulated in section III, with numerical comparisons with existing algorithms. The problem of sampling the optimized density and numerical results on estimation errors are discussed in section IV. Extensions of the method are introduced in section V, and concluding remarks are given in section VI.

II. Stable least-squares recovery and related problems

The standard setting for the least-squares recovery of a signal is as follows. We assume that the signal f to be recovered can be approximated by f_m in a space of finite dimension V_m, and that we observe samples $y_i = f(x_i) + b_i$, where $x_i \in X$ are n fixed sampling points, and the measurement errors b_i are i.i.d. centered random variables of variance σ^2. The space V_m depends on the application. Examples include standard or trigonometric polynomials [17], Fourier-Bessel functions or spherical harmonics [19], [20], cosine basis [8], etc.

The least-squares estimate $\hat{f}_m$ of f is given by

$$\hat{f}_m = \operatorname*{argmin}_{\hat{f}_m \in V_m} \sum_{i=1}^{n} |y_i - \hat{f}_m(x_i)|^2. \tag{1}$$

The Gauss-Markov theorem shows that in the case where $f = f_m$, $\hat{f}_m$ is the best linear unbiased estimate of f.

Our goal is to ensure the stability of the estimation with respect to the measurement noise, as well as the approximation error, i.e. that the estimation error can be bounded:

$$\|\hat{f}_m - f\|_2^2 \leq C_1 \|f_m - f\|_2^2 + C_2 \sigma^2 \tag{2}$$

with C_1 as close as possible to 1, and C_2 as close as possible to 0 for a given number of samples.

Depending on the choice of the sampling points x_i, a large number of samples can be necessary to ensure stability of the estimation, even in noiseless cases. A famous example is known as the Runge phenomenon, where polynomial interpolation of a function with regularly spaced samples fails for a number of samples linear with respect to the dimension m.

Optimizing the sampling points x_i leads to an untractable non-convex optimization problem. We will use a relaxed approach, introduced in [17], where the sampling points are drawn from a probability measure μ. The problem of choosing points x_i is replaced by choosing a density μ. A similar approach is adopted for sparse recovery, where results similar to (2) are obtained through the use of random matrices.

A. Stability of least-squares estimation

In this setting, Cohen et al. [17] give a criterion for the stability of least-squares recovery. With $(L_j)_{1 \leq j \leq m}$ an orthogonal basis of V_m with respect to the measure μ, they define

$$K(x, \mu) = \sum_{j=1}^{m} |L_j(x)|^2 \tag{3}$$

$$K(\mu) = \sup_{x \in X} K(x, \mu). \tag{4}$$

$K(\mu)$ is bounded from below by m. The following theorem shows that the expectation of the estimation error is bounded in a way similar to eq. (2).

Theorem 1. *For a given $r > 0$, $\kappa = (1 - \log 2)/(2 + 2r)$, and $\varepsilon(n) = 4\kappa / \log n$, if the number of measurements n is such that*

$$K(\mu) \leq \kappa \frac{n}{\log n} \tag{5}$$

then the expectation of the estimation error is bounded by

$$E\left(\|\tilde{f}_m - f\|_\mu\right) \leq (1 + 2\varepsilon(n))\|f_m - f\|_\mu^2 + 8M^2 n^{-r} + 8\sigma^2 \frac{m}{n} \tag{6}$$

where M is an upper bound on $|f|$ and

$$\|f\|_\mu^2 = \int_X |f|^2 d\mu \tag{7}$$

$K(\mu)$ essentially measures the number of samples necessary to ensure stability with respect to deterministic approximation errors and random measurement noise. Before introducing the optimization algorithm to design a sampling density μ such that $K(\mu)$ is close to m, we recall alternative state of the art methods that will be tested against our proposed method.

B. D-Optimal design

The criterion $K(\mu)$ also appears in the analysis of D-optimal design [11]. In this setting, the quantity of interest are the parameters α_i of the decomposition of f in the basis $(\phi_i)_{1 \leq i \leq m}$ of V_m.

$$f(x) = \sum_{i=1}^{m} \alpha_i \phi_i(x). \tag{8}$$

Here, deterministic approximation errors are not considered (i.e., $f = f_m$).

[1] https://gilleschardon.fr/fwkm

The parameters α_i are estimated using measurements at sampling points $x_i \in X$, that can be repeated w_i times with independent noise realizations. The problem of D-optimal design is the maximization of the determinant

$$D = \det M \tag{9}$$

where M is the information matrix with coefficients $M_{ij} = \frac{1}{n}\sum_{l=1}^{n} w_n \phi_i(x_l)\phi_j(x_l)$.

This determinant is proportional the inverse of the volume of the uncertainty ellipsoid for the estimation of the parameters α_i. This problem can also be relaxed, by defining $M(\mu)$ with coefficients $M(\mu)_{ij} = \int_X \phi_i(x)\phi_j(x)d\mu$. In this *approximate design problem*, it has been shown that maximizing $D(\mu)$ is equivalent to minimizing $K(\mu)$ [21], in the sense that

$$\operatorname*{argmin}_{\mu} K(\mu) = \operatorname*{argmax}_{\mu} D(\mu). \tag{10}$$

Furthermore, the following bound [22]

$$D(\mu) \geq \exp(m - K(\mu)) \sup_{\mu} D(\mu) \tag{11}$$

shows that the maximization of the determinant can be replaced by the minimization of $K(\mu)$, of which the convergence is easier to control. Indeed, it will be shown that the lower bound m can be reached. Moreover, the variance of the estimation of $f(x)$ is $\sigma^2 K(x, \mu)$.

Different algorithms have been proposed for the design of sampling sets in the D-optimal sense:

- greedy algorithms [13], [14],
- multiplicative algorithms [23], and combination with greedy algorithms [24],
- convex optimization methods, either on a relaxed problem [7], or on a moment problem in the particular case of polynomial approximation [12],...

In [7], instead of optimizing a probability density, Joshi and Boyd constrain the sampling density to sum to the number of desired sensors n. The optimization problem is solved by Newton's method, and sampling points are selected as the n largest values of the density. A local optimization can then refine the choice of the sampling set.

C. FrameSense

The FrameSense algorithm [8], proposed by Ranieri et al., is based on a proxy for the mean squared estimation error when no deterministic measurement errors are present. For a given set of points x_i, the rows of the matrix with coefficients $\phi_{ij} = \phi_i(x_j)$ define a frame of $\mathbf{R}^m$. In the case where this frame has elements with unit norm, it is shown that minimizing the estimation error is equivalent to minimizing the frame potential $\sum_{i,j} |\sum_l \phi_{il}\phi_{jl}|^2$. An argument of sub-modularity is invoked to justify the use of a greedy algorithm to minimize the frame potential.

D. Weighted least-squares

An alternative to the standard least-squares is to use weights in eq. (1) :

$$\hat{f}_m = \operatorname*{argmin}_{\hat{f}_m \in V_m} \sum_{i=1}^{n} w(x_i)|y_i - \hat{f}_m(x_i)|^2. \tag{12}$$

A joint optimization procedure for the sampling density μ and weights $w(x)$ based on a modified version of Theorem 1 is proposed in [25]. This procedure is not well adapted to the case of measurement design, as imposing weights implies that not only the location of the measurements, but also the variance of the noise applied to each measurement can be independently controlled.

III. The Minimization Problem

We aim at designing sampling sets ensuring stability of the estimation with the smallest number of samples. To this end, supported by Theorem 1, we consider the following optimization problem:

$$\mu^\star = \operatorname*{argmin}_{\mu} K(\mu) \tag{13}$$

under the constraint that μ is a probability measure.

As our objective is the numerical optimization of a sampling set, we will limit ourselves to a discretized version of the problem. The space V_m is a subspace of a larger space E of dimension L. V_m is described by a matrix $\mathbf{P}$ containing the basis vectors. The sampling density μ is replaced by a positive vector $\mathbf{w}$ of weights w_i, and the basis $\mathbf{Q}$ is an orthogonal basis of V_m with respect to the weights $\mathbf{w}$ ($\mathbf{Q}^\star \mathrm{diag}(\mathbf{w})\mathbf{Q} = \mathbf{I}$). Its row vectors are denoted $\mathbf{q}_i$ (i.e. each $\mathbf{q}_i$ contains the i-th coefficient of the basis vectors of V_m).

In this discrete setting, the criterion K is defined by

$$K(i, \mathbf{w}) = \|\mathbf{q}_i\|_2^2 \tag{14}$$

$$K(\mathbf{w}) = \max_i K(i, \mathbf{w}) \tag{15}$$

and the optimization problem becomes:

$$\mathbf{w}^\star = \operatorname*{argmin}_{\mathbf{w}} K(\mathbf{w}) \text{ subject to } w_i \geq 0, \sum_{i=1}^{L} w_i = 1 \tag{16}$$

Theorem 2. *The optimization problem (16) is convex and at the optimal density* $\mathbf{w}^\star$

$$K(\mathbf{w}^\star) = m, \tag{17}$$

and

$$w_i^\star = 0 \text{ if } K(i, \mathbf{w}^\star) \neq m. \tag{18}$$

A proof is given in [21] in the continuous case, where the support of $\mu^\star$ is included in the set $\{x | K(x, \mu^\star) = m\}$. An alternative proof is given in appendix A using standard convex analysis tools in the discrete case. An important consequence of the support condition is that in most cases, the optimal density is sparse. Indeed, the functions used to build the space V_m are usually analytic (exponentials, polynomials, cosines, etc.), and so is $K(x, \mu^\star)$. Except in the unlikely cases where $K(x, \mu^\star)$ is constant (e.g. for a Fourier basis), $K(x, \mu^\star) \neq m$ almost everywhere, implying that the support of the measurement density has Lebesgue measure 0.

The optimization problem (16), despite its apparent complexity (cost of evaluation of the objective function which involves the orthogonalization of a large matrix, large number of constraints), can be solved in an efficient way using the Frank-Wolfe algorithm. As will be shown, this algorithm

allows to leverage the particular form of the objective function and of the constraints to efficiently compute a minimizer of (16).

A. Frank-Wolfe algorithm

Input: function f, feasible set D, initialization $\mathbf{x}_0$, number of iterations N
Output: last iterate $\mathbf{x}_N$
$n = 0$
while $n < N$ **do**
  $\gamma = \frac{2}{n+2}$
  $\mathbf{s} = \operatorname{argmin}_{\mathbf{s}\in D} \langle \mathbf{s}, \operatorname{grad} f(\mathbf{x}_n)\rangle$
  $\mathbf{x}_{n+1} = (1-\gamma)\mathbf{x}_n + \gamma s$
  $n = n + 1$
end

Algorithm 1: Frank-Wolfe algorithm

The Frank-Wolfe algorithm, also known as the conditional gradient method, is a convex optimization algorithm introduced in 1956 by Frank and Wolfe [26], [27].

It is similar to a gradient descent, with the difference that the next iterate is not chosen in the direction of the gradient, but in the direction of the minimizer in the feasible set of the linear approximation of the objective function around the current iterate. The domain being convex, the iterates are guaranteed to be feasible, and no projection in the feasible set is necessary. The algorithm is outlined in Alg. 1.

In our case $K(\mathbf{w})$ is the maximum of the smooth functions $K(i, \mathbf{w})$. The Franke-Wolfe algorithm can be extended to such non-smooth functions [28]. The gradient of $K(\mathbf{w})$ is the convex hull of the gradients of the maximal $K(i, \mathbf{w})$, and at most points, the gradient of $K(\mathbf{w})$ is simply the gradient of the maximum $K(i, \mathbf{w})$, with index $i^\star$.

The linear problem to solve at each iteration is:

$$\min_{\mathbf{s}} \langle \mathbf{s}, \operatorname{grad} K(i^\star, \mathbf{w})\rangle \text{ subject to } \sum_{i=1}^{L} s_i = 1, s_i \geq 0. \quad (19)$$

Its solution is a vector with zero coefficients, except at the minimal value of the gradient, where it is one. A straightforward application of the Cauchy-Schwartz inequality to eq. (36) shows that the minimal coordinate of the gradient of $K(i^\star, \mathbf{w})$ is the $i^\star$-th coordinate.

An iteration of the Frank-Wolfe algorithm applied to our optimization problem has thus the simple implementation:

- find the index $i^\star$ of the maximum value of $K(i, \mathbf{w})$,
- add weight on the $i^\star$-th coefficient of the sampling density,
- rescale the sampling density such that $\sum_{i=1}^{L} w_i = 1$.

As at most one point is added to the support at each iterations, the Frank-Wolfe algorithm yields sparse measurement densities.

The results of the Frank-Wolfe algorithm for $K(\mathbf{w})$ minimization, Newton's method for the equivalent problem of determinant maximization[7], and the measurement density for weighted least-squares [25] are plotted on figure 1, for a polynomial basis of size $m = 20$. For Newton's method, the solution of the relaxed problem for $n = 2m$ sensors is plotted. As expected, the obtained density is sparse: $K(i, \mathbf{w}^\star)$ being a polynomial of order $2m-2$, the support of $\mathbf{w}^\star$, where $K(i, \mathbf{w}^\star) = m$, contains at most $2m - 2$ points.

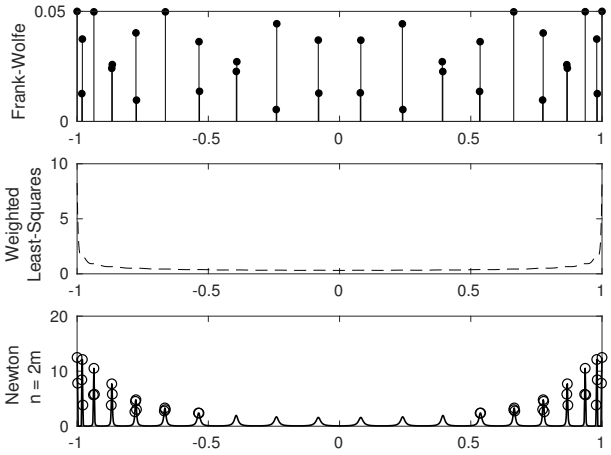

Fig. 1. Measurement densities for a polynomial basis of size $m = 20$. The $2m$ largest values of the output of Newton's method are indicated by circles.

The Frank-Wolfe optimization procedure for $K(\mathbf{w})$ minimization can also be used to solve the approximate D-optimality problem. As shown above, the determinant maximization problem and the $K(\mu)$ minimization problem have the same solutions. Moreover, applying the Frank-Wolfe algorithm to the determinant maximization problem gives the same optimization procedure. Indeed, the Fisher information matrix of the approximate D-optimal design problem (which is also the Gram matrix of $\mathbf{P}$ with respect to the weights $\mathbf{w}$) is given by

$$\mathbf{M} = \mathbf{P}^\star \operatorname{diag}(\mathbf{w}) \mathbf{P} \quad (20)$$

and the determinant lemma shows that the derivative with respect to the weight w_i of the determinant $D = \det \mathbf{M}$ is

$$\frac{\partial D}{\partial w_i} = \mathbf{p}_i \mathbf{M}^{-1} \mathbf{p}_i^\star D = \mathbf{q}_i \mathbf{q}_i^\star D = K(i, \mathbf{w}) D \quad (21)$$

where the orthogonal basis $\mathbf{Q}$ is defined as $\mathbf{Q} = \mathbf{P}\mathbf{M}^{-1/2}$ and $\mathbf{p}_i$ are the row vectors of $\mathbf{P}$. The maximal value of the gradient is at index $i^\star$, identical to the index selected by eq. (19).

B. Time and space complexity

The computational cost of an iteration of the Frank-Wolfe algorithm is dominated by the orthogonalization of the design matrix $\mathbf{P}$ of size $L \times m$. While a standard orthogonalization algorithm would have a complexity in $O(Lm^2)$, the orthogonalization can be accelerated by using the particular way the sampling density is updated. Indeed, an iteration of the algorithm modifies the density at a unique point, and renormalizes the density. As this modification amounts to a rank-1 perturbation of the Gramian matrix, its squared root can be computed using the matrix inversion lemma with complexity $O(Lm)$, which governs the cost of an iteration.

While no theoretical bound on the number of iterations necessary to achieve a value of $K(\mathbf{w})$ close to the optimum

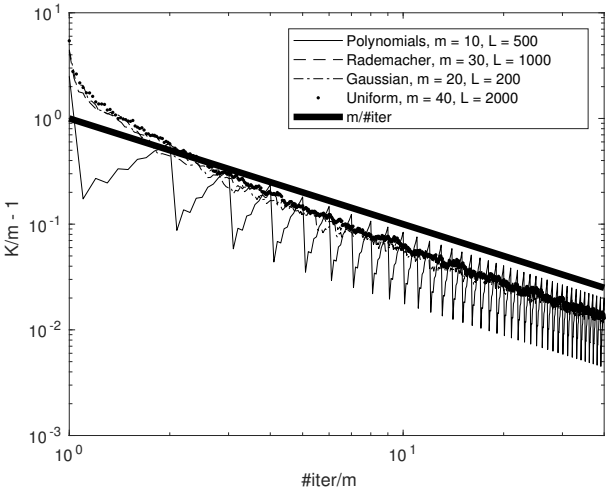

Fig. 2. Convergence of the proposed algorithm in various settings.

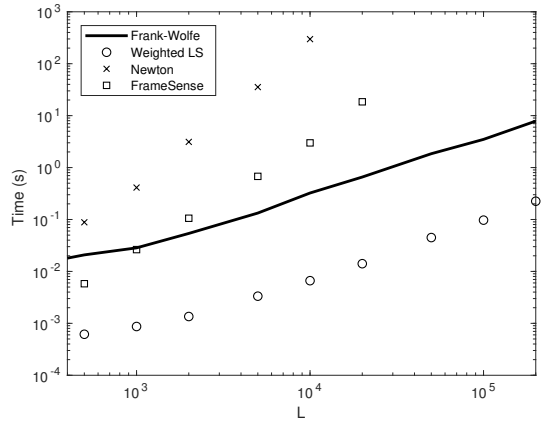

Fig. 3. Computational time of the proposed Frank-Wolfe algorithm, Newton's method [7], FrameSense [8] and weighted least-squares [25], in function of the size of the discretization L, with $m = 20$, for gaussian matrices.

is yet available, numerical experiments show that in various settings, a number of iterations proportional to m is sufficient to reach values of $K(\mathbf{w})$ close to the optimum m. Fig. 2 shows on the same plot the value of $K(\mathbf{w})$ along the iterations, for various dimensions L and m, and various bases (polynomials and random matrices with Rademacher, gaussian, and uniformly distributed independent coefficients). In all settings, the number of iterations necessary to reach a density $\mathbf{w}$ such that $(K(\mathbf{w}) - m) \leq Cm$ is approximately m/C. For a given C, the computation complexity of the algorithm is thus $O(Lm^2)$.

The time and space complexity of the proposed algorithm, Newton's method [7], the FrameSense algorithm [8] and the joint design of a sampling density and weights for weighted least-squares [25] are given in table I. The proposed algorithm, having time and space complexities linear in the number of possible sensor locations, is much more efficient for increasing L and fixed m, and can attain higher dimensions because of its low memory requirements, as can be seen on figure 3 (the computation times are given for a laptop with an Intel Core i7 quad core at 2.10 GHz and 16 GB of RAM, running MATLAB R2017b). The complexity of the joint weights and density design for weighted least-squares is of the same order as the Frank-Wolfe algorithm, but is faster as it is equivalent to its initialization. While the method implies control of the measurement noise variance and does not yield sparse densities, its reduced numerical cost makes it a valuable alternative.

TABLE I
TIME AND SPACE COMPLEXITIES OF THE PROPOSED FRANK-WOLFE ALGORITHM, WEIGHT LEAST-SQUARES, FRAMESENSE, AND THE DETERMINANT MAXIMIZATION.

algorithm	init.	iter.	#iter.	total	space
Frank-Wolfe (prop.)	$O(Lm^2)$	$O(Lm)$	$O(m)$	$O(Lm^2)$	$O(Lm)$
Weigted LS [25]	$O(Lm^2)$	-	-	$O(Lm^2)$	$O(Lm)$
FrameSense [8]	$O(L^2m)$	$O(L)$	$L-n$	$O(L^2m)$	$O(L^2)$
Newton [7]	$O(1)$	$O(L^3)$	≤ 30	$O(L^3)$	$O(L^2)$

IV. SAMPLING THE OPTIMAL DENSITY

Once a density is designed, the measurement points are obtained by sampling this density. After introducing various sampling methods, numerical results on estimation errors are given.

A. Sampling methods

The sampling methods we consider are the following:

a) Maximal values of the density : Joshi and Boyd [7] proposed to select the locations of the n largest values of the weights w_i. While this method works in random settings, it does not produce efficient sampling sets in more typical cases. Indeed, as is visible on figure 1, the n largest values of the optimized density are not necessarily located at all sensor positions necessary for a stable recovery. They also propose a local optimization to improve the sampling set, by testing swaps of a sampling point with a non-selected point.

b) i.i.d. samples: Theorem 1, based on a i.i.d. sampling of the density, is valid for a number of measurements such that

$$K(\mu) \leq \kappa \frac{n}{\log n} \tag{22}$$

This quantity of samples ensures that the empirical Gram matrix of the basis (L_j), with coefficients

$$G^e_{ij} = \frac{1}{n} \sum_{l=1}^{n} L_i(x_l) L_j(x_l) \tag{23}$$

is close enough to the coefficients of the Gram matrix with respect to the sampling measure μ

$$G_{ij} = \int_X L_i(x) L_j(x) d\mu \tag{24}$$

$$= \delta_{ij}. \tag{25}$$

In addition to the large number of samples compared to $K(m)$ and the dimension of the space m, using i.i.d. sampling in the cases where the sampling density is concentrated on few points leads to multiple selections of a given point, which is usually to be avoided in practice.

c) Resampling: The coefficients of the empirical Gram matrix in eq. (23) are a Monte-Carlo evaluation of the integrals (24). Moreover, in most cases, the optimal density is discrete. Sampling of discrete densities is involved in the resampling stage of particle filters, or sequential Monte-Carlo methods. In this setting, the performances of an i.i.d. sampling of a discrete density (also called multinomial sampling in this case) can be improved by considering other resampling methods.

Stratified sampling and systematic sampling are well known simple resampling methods in dimension 1 [29]. With $F(x)$ the cumulative distribution function of the sampling density, the samples are chosen to be

$$x_i = F^{-1}\left(\frac{i-1}{n} + u_i\right) \tag{26}$$

for stratified sampling, and

$$x_i = F^{-1}\left(\frac{i-1}{n} + u\right) \tag{27}$$

for systematic sampling, where $u, u_1, \ldots, u_n$ are i.i.d uniform random variables in the interval $[0, 1/n]$.

To avoid multiple selections of a sampling point, duplicates can be replaced by samples drawn from the uniform density. In higher dimensions, methods based on parametrizing the domain using an Hilbert curve and stratified or systematic sampling along this curve can be used [30], [31].

d) FrameSense as a sampling method: The above sampling methods are based on the sampling density only, and do not take into account the row matrix $\mathbf{q}_i$ associated to a point x_i. In the case when two rows are correlated, selecting both the rows does not add much information. Another less probable but uncorrelated row could yield more information. Removing similar rows is the mode of operation of FrameSense, which can be used as a sampling method after optimization of the density using the proposed method. Here only the support of the optimized density is used.

Conversely, the Frank-Wolfe algorithm can be considered as a pre-processing for FrameSense, selecting important points and lowering the computational complexity. The total cost of the proposed method followed by FrameSense is $O(Lm^2)$, to be compared with the cost $O(L^2m)$ for FrameSense alone.

B. Numerical results

1) Polynomials: Fig 4 shows the mean squared estimation error for different design methods, in the case of polynomial approximation in function of the number of samples. Polynomials of order 19 (i.e. $m = 20$) are used to approximate the function $f(t) = (1 + 25t^2)^{-1}$ in the interval $[-1, 1]$. The mean squared error is estimated by repeating the experiment 10000 times, for 20 to 40 measurements, and the interval is discretized using 2000 points. The FrameSense algorithm, as well as systematic sampling combined with Newton's method algorithm, the proposed Frank-Wolfe algorithm and weighted least-squares, have performances similar to the optimal density (Dirac masses of weight $1/m$ at the extremities of the interval and at zeros of the Legendre polynomial of order $m-2$ [11]). Designing the sampling set using stratified sampling of the optimized density reaches the same performances only when

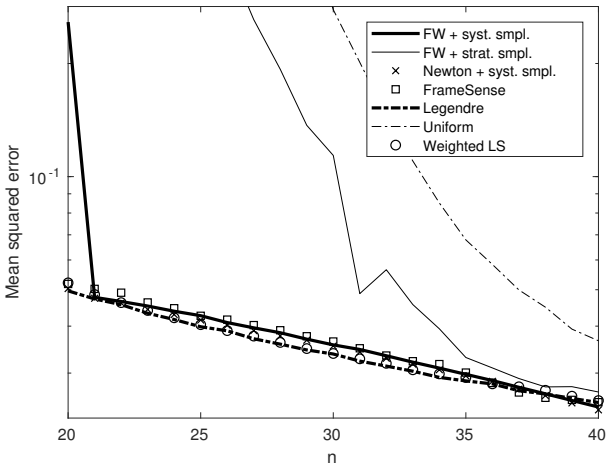

Fig. 4. Mean square error for polynomials approximation, $m = 20$.

$n \approx 2m$ (the increase of the error between $n = 31$ and $n = 32$ is a consequence of the non identical distributions for each sampling point). The performances of a set of regularly spaced points in the interval is also plotted for completeness. The mean squared errors of sets of independent samples of the density optimized by the Frank-Wolfe algorithm, as well as the n largest values of the result of Newton's method and the locally optimized sampling set, are too large to be included in the figure.

2) Fourier-Bessel functions: Acoustical fields can be estimated using the approximation of solutions to the Helmholtz equation $\Delta p + k^2 p = 0$ by so-called generalized harmonic polynomials (in $\mathbf{R}^2$) or spherical harmonics (in $\mathbf{R}^3$) [20], [32]. In polar coordinates:

$$p(r, \theta) \approx \sum_{n=-m'}^{m'} p_n J_n(kr) e^{in\theta} \tag{28}$$

where J_n is the Bessel function of order n, and $m = 2m' + 1$. In the case of a circular domain, it was proven that values of $K(\mu)$ close to the lower bound are obtained with densities that are concentrated on the boundary of the domain [19]. Numerical evidence showed that this was also the case for square domains. Figure 5 shows the optimized density $\mathbf{w}$ in the case of a polygonal domain with a reentrant corner. Here $m' = 10$, $m = 2m' + 1 = 21$, and the square is discretized using $L' = 100$ points on each axis, i.e. $L = 7500$ points in the domain. As expected, most of the optimized density (66%) is concentrated on the boundary of the domain.

The mean square error for the approximation of a plane wave on this domain is plotted on figure 6 for sampling sets obtained by the FrameSense algorithm, as well as set sampled from densities optimized with the Newton and the Frank-Wolfe algorithms. The convex relaxation approach (Frank-Wolfe or Newton) or weighted least-squares with Hilbert parametrization and systematic sampling yields accurate reconstructions. Using FrameSense as a postprocessing of the Frank-Wolfe algorithm improves the results. FrameSense alone is less accurate than the relaxed approached, and selecting the maximal values of the density optimized with Newton's

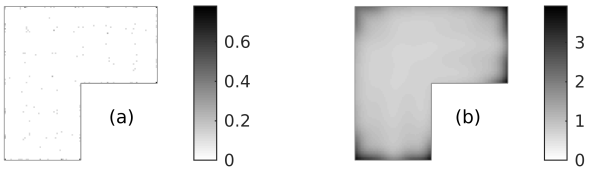

Fig. 5. Optimized density for recovery of a wavefield using Fourier-Bessel functions, $m = 21$. (a) Frank-Wolfe algorithm. The density is here multiplied by m for an easier comparison with the maximal weight $1/m$ of a Dirac mass. (b) Weighted least-squares.

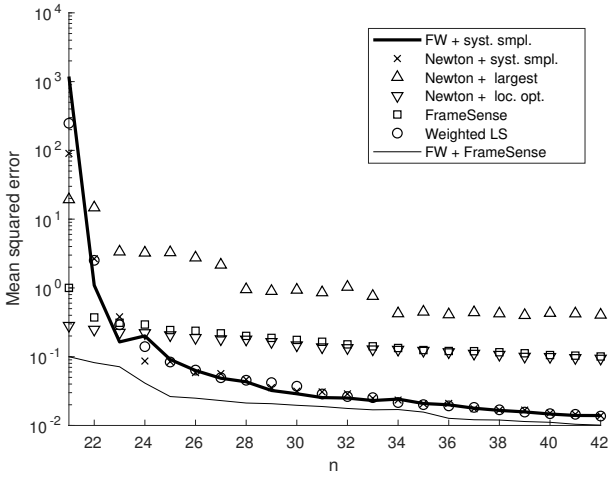

Fig. 6. Mean squared error for recovery using Fourier-Bessel functions, $m = 21$.

method does not yield accurate estimations, but is somewhat improved by the local optimization. The computation times are 0.12s for the proposed Frank-Wolfe method, 1.8s for the FrameSense algorithm, and one minute for Newton's method (with $L' = 250$, $L = 46875$, 2 seconds are necessary for the Frank-Wolfe algorithm, while the FrameSense and Newton's method fails because of too large memory requirements).

V. Extensions

Formulating the sensors selection problem as a convex optimization problem allows to modify the constraints and/or the objective functions to take into account more complex settings. We demonstrate here two simple modifications.

A. Bounding the error

A consequence of the concentration of the optimal sampling density $\mu^\star$ on a set of measure 0 in important settings (e.g. polynomial bases, cosine bases, etc.) is that the bound (6) is unable to control the L_2-norm of the error with respect to the Lebesgue measure λ.

To ensure stability in the L_2-norm, a bound can be applied to the measure, e.g. $\mu(I) \geq \alpha\lambda(I)$ for any subset I of the domain and $\alpha > 0$. The L_2-norm of the error with respect to μ can now be used to bound the error with respect to the Lebesgue measure :

$$\|f\|_\lambda \leq \frac{1}{\alpha}\|f\|_\mu. \tag{29}$$

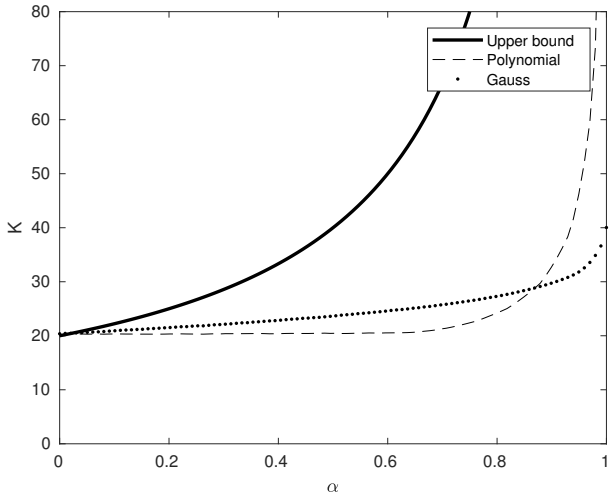

Fig. 7. Optimization of a sampling density for a polynomial basis and a gaussian matrix ensuring an upper bound on the error in the interval, $m = 20$.

A bound on the optimal measure with this additional constraint can be found. The measure $\mu_\alpha^\star = (1-\alpha)\mu^\star + \alpha\lambda$ being feasible, a simple computation (see (35)) shows that

$$K(\mu_\alpha^\star) \leq K((1-\alpha)\mu^\star + \alpha\lambda) \leq \frac{1}{1-\alpha}K(\mu^\star) \tag{30}$$

In the discrete formulation, the optimization problem becomes

$$\hat{\mathbf{w}} = \text{argmin}_{\mathbf{w}} K(\mathbf{w}) \text{ subject to } \sum_{i=1}^{L} w_i = 1, w_i \geq \alpha/L. \tag{31}$$

Algorithm 1 is easily modified to solve this problem. The value of $K(\mathbf{w})$ for a polynomial basis and a gaussian matrix with $L = 500$, $m = 20$ is plotted for α between 0 and 1. The bound (30) appears to be pessimistic in these cases. In the polynomial case in particular, a bound of the error over the entire interval can be obtained at a minimal cost as $K(\mathbf{w})$ remains almost equal to m for $\alpha < 0.6$.

B. Joint optimization for multiple bases

In some applications, a unique sensor array has to be used to sample several signals that are not necessarily described by a unique linear model. An example is soundfield measurement (or more generally, wavefield estimation), where the same microphone array is used at different frequencies. The approched used in [9] for sensor placement in non-linear models involves a similar problem.

In the case of S different subspaces, not necessarily of the same dimension, a quantity $K_s(w)$ can be computed for each subspace. We define

$$K(\mathbf{w}) = \max_s K_s(\mathbf{w}), \tag{32}$$

and the problem (16) with this modified objective function can be solved with a simple application of Alg. 1.

This method is tested for the approximation of a function using polynomials (space V_m^p, with criterion K^p) and a cosine

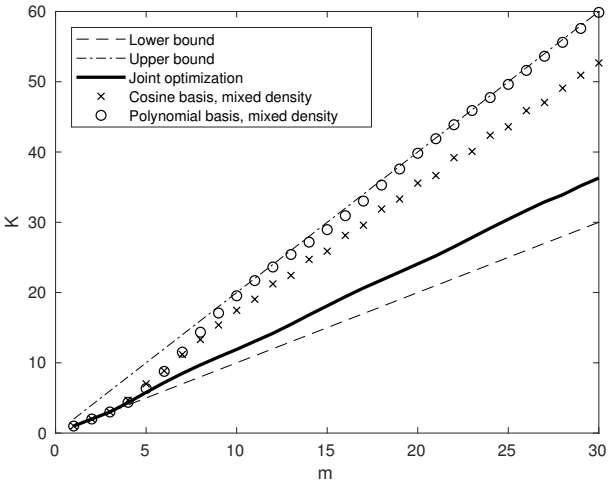

Fig. 8. $K(w)$ for a polynomial basis and a cosine basis, with the jointly optimized density, with comparison to the mixed density and the lower and upper bounds.

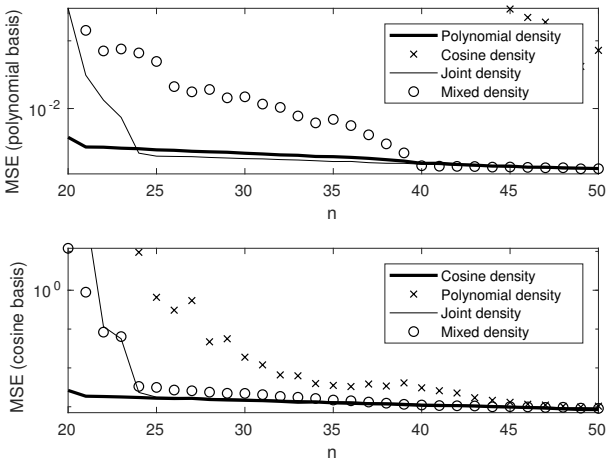

Fig. 9. Mean square error for estimation using polynomial (top) and cosine (bottom) bases.

basis (space V_m^d, with criterion K^d). The following densities are computed:

- $\mathbf{w}_p$, optimized for the polynomial basis,
- $\mathbf{w}_d$, optimized for the cosine basis,
- the mixture $\mathbf{w}_m = (\mathbf{w}_p + \mathbf{w}_d)/2$
- and $\mathbf{w}_j$, jointly optimized for the two bases. In this case, a simple analysis of the optimization problem shows that $K^p(\mathbf{w}_j) = K^d(\mathbf{w}_j)$.

The values of K^p and K^d for the mixed and joint densities are plotted on figure 8 in function of m. While the mixed density yields values of $K^p(\mathbf{w}_m)$ close to the upper bound $2m$, for the jointly optimized density, K^d and K^p can be bounded by $1.3m$, making it possible to construct a sampling set capable of stable recovery in two different spaces with a reasonable number of measurements.

Estimation errors for the different combinations of spaces and sampling densities are plotted on figure 9, using systematic sampling. Here $m = 20$ and $K^p(\mathbf{w}_j) = K^d(\mathbf{w}_j) = 24.1$. As expected, 24 samples are sufficient to ensure the stability of the estimation simultaneously in the polynomial and cosine bases.

VI. Conclusion

The problem of sampling set design for least-squares estimation is here considered in a relaxed setting. The Frank-Wolfe algorithm offers an efficient optimization procedure in terms of computational complexity in time and space. Variants of this algorithm, such as the *away step* variant [33] could improve the speed of convergence of the optimization.

The optimal densities, as well as the Frank-Wolfe iterates, being sparse in generic cases, the sampling points are obtained by resampling techniques. This approach is shown to be more efficient than i.i.d. sampling or selecting the maximal values of the density with local optimization. While efficient in dimension 1 (they allow stable estimation with a number of measurements close to the number of parameters to estimate), the resampling methods considered here do not seem optimal in higher dimensions. The development of sampling methods aware of the rows $\mathbf{q}_i$ associated to a point w_i along to its probability should improve the performances of the sampling sets.

Finally, the combination of the convex optimization formulation with resampling methods offers a wide arrays of possibilities, such as the optimization of sampling densities for non-linear problems, or the optimization of E-criterion in a way similar to [9].

Appendix A
Minimum of $K(w)$

In this appendix, we prove the convexity of $K(\mathbf{w})$ and that the optimal value of $K(\mathbf{w})$ is the dimension m of the approximation space V_m.

We consider a discrete setting. A basis of the space V_m is given by the matrix $\mathbf{P}$ of dimension $L \times m$, with row vectors $\mathbf{p}_i$. We call $\mathbf{G}_\mathbf{w} = \mathbf{P}^\star \mathrm{diag}(\mathbf{w})\mathbf{P}$ the Gram matrix of the basis $\mathbf{P}$ with respect to the set of weights $\mathbf{w}$.

A. Convexity of $K(\mathbf{w})$

As $K(\mathbf{w}) = \max_i K(i, \mathbf{w})$, it is sufficient to prove the convexity of $K(i, \mathbf{w})$. We consider two sets of weights $\mathbf{u}$ and $\mathbf{v}$, and their convex combination $\mathbf{w} = \alpha\mathbf{u} + (1-\alpha)\mathbf{v}$. Then $K(i, \mathbf{w}) = \mathbf{p}_i \mathbf{G}_\mathbf{w}^{-1} \mathbf{p}_i^\star$. By using $\mathbf{G}_\mathbf{w} = \alpha\mathbf{G}_\mathbf{u} + (1-\alpha)\mathbf{G}_\mathbf{v}$ and the derivative of the inverse of a matrix, we find

$$\frac{d^2K(i,\mathbf{w})}{d\alpha^2} = 2\left(\mathbf{p}_i\mathbf{G}_\mathbf{w}^{-1}(\mathbf{G}_\mathbf{u} - \mathbf{G}_\mathbf{v})\right)\mathbf{G}_\mathbf{w}^{-1}\left((\mathbf{G}_\mathbf{u} - \mathbf{G}_\mathbf{v})\mathbf{G}_\mathbf{w}^{-1}\mathbf{p}_i^\star\right) \tag{33}$$

$$= 2\mathbf{r}^\star\mathbf{G}_\mathbf{w}^{-1}\mathbf{r} \geq 0 \tag{34}$$

where $\mathbf{r} = \left((\mathbf{G}_\mathbf{u} - \mathbf{G}_\mathbf{v})\mathbf{G}_\mathbf{w}^{-1}\mathbf{p}_i^\star\right)$ and $\mathbf{G}_\mathbf{w}^{-1}$ is a positive matrix.

The inequality

$$K(\mathbf{w}) \leq \frac{K(\mathbf{u})}{\alpha} \tag{35}$$

can also be proved by using the matrix inequality $\mathbf{G}_\mathbf{w}^{-1} \leq \mathbf{G}_\mathbf{u}^{-1}/\alpha$.

1) Subgradient of $K(\mathbf{w})$: The subgradient of $K(\mathbf{w})$ is the convex combination of the gradients of the $K(i, \mathbf{w})$ such that $K(i, \mathbf{w}) = K(\mathbf{w})$. With similar computations as the previous results and $\mathbf{Q}$ the matrix of a basis orthogonal with respect to $\mathbf{w}$, we find

$$\frac{\partial K(i, \mathbf{w})}{\partial w_j} = -|\langle \mathbf{q}_i, \mathbf{q}_j \rangle|^2 = -\mathbf{J}_{ij} \tag{36}$$

The matrix $\mathbf{J}$, as the Hadamard product of $\mathbf{Q}\mathbf{Q}^\star$ and its conjugate, is semi-definite by the Schur product theorem.

The subgradient of $K(\mathbf{w})$ is the set of vectors $-\mathbf{J}\boldsymbol{\alpha}$, for $\boldsymbol{\alpha}$ a positive vector of sum 1, and $\alpha_i(K(i, \mathbf{w}) - K(\mathbf{w})) = 0$.

B. Karush-Kuhn-Tucker conditions

The Karush-Kuhn-Tucker conditions for the optimization problem (16) implies that there exists a scalar λ and a positive vector $\boldsymbol{\gamma}$ such that the subgradient of $K(\mathbf{w})$ verifies

$$\mathbf{J}\boldsymbol{\alpha} = \lambda \mathbf{1} - \boldsymbol{\gamma} \tag{37}$$

with $w_i \gamma_i = 0$, appropriate conditions on $\boldsymbol{\alpha}$, and $\mathbf{1}$ a vector of L coefficients equal to 1. Combining the definitions of $\mathbf{J}$, $K(\mathbf{w})$, $\mathbf{Q}$, and the properties of $\boldsymbol{\alpha}$, $\mathbf{w}$ and $\boldsymbol{\gamma}$, we find:

- $\boldsymbol{\alpha}^\star \mathbf{J} \mathbf{w} = K(\mathbf{w})$
- $\mathbf{w}^\star \mathbf{J} \boldsymbol{\alpha} = \lambda$
- $\boldsymbol{\alpha}^\star \mathbf{J} \boldsymbol{\alpha} = \lambda - \boldsymbol{\alpha}^\star \boldsymbol{\gamma} = K(\mathbf{w}) - \boldsymbol{\alpha}^\star \boldsymbol{\gamma} \leq K(\mathbf{w})$
- $\mathbf{w}^\star \mathbf{J} \mathbf{w} = m$

Using the symmetry and the positivity of $\mathbf{J}$, we have :

$$|\boldsymbol{\alpha}^\star \mathbf{J} \mathbf{w}|^2 \leq |\mathbf{w}^\star \mathbf{J} \mathbf{w}||\boldsymbol{\alpha}^\star \mathbf{J} \boldsymbol{\alpha}| \tag{38}$$

$$K(\mathbf{w})^2 \leq m K(\mathbf{w}) \tag{39}$$

which implies $K(\mathbf{w}) \leq m$. Combined with the lower bound $K(\mathbf{w}) \geq m$, we have $K(\mathbf{w}) = m$. By remarking that $\sum_i w_i K(i, \mathbf{w}) = \sum_i w_i \mathbf{q}_i \mathbf{q}_i^\star = m$ from the orthogonality of the columns of $\mathbf{Q}$, we also have that $w_i = 0$ if $K(i, \mathbf{w}) < K(\mathbf{w})$.

ACKNOWLEDGMENT

The author thanks Piotr Bojanowski for introducing him to the Frank-Wolfe algorithm.

REFERENCES

[1] R. G. Vaughan, N. L. Scott, and D. R. White, "The theory of bandpass sampling," *IEEE Transactions on Signal Processing*, vol. 39, no. 9, pp. 1973–1984, Sep. 1991.

[2] H. J. Landau, "Necessary density conditions for sampling and interpolation of certain entire functions," *Acta Mathematica*, vol. 117, pp. 37–52, 1967.

[3] B. Matei and Y. Meyer, "Quasicrystals are sets of stable sampling," *Comptes Rendus Mathematique*, vol. 346, no. 23, pp. 1235 – 1238, 2008.

[4] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, ser. Applied and Numerical Harmonic Analysis. New York, NY: Springer New York, 2013. [Online]. Available: http://link.springer.com/10.1007/978-0-8176-4948-7

[5] E. J. Candes and T. Tao, "Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[6] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[7] S. Joshi and S. Boyd, "Sensor Selection via Convex Optimization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 451–462, Feb. 2009.

[8] J. Ranieri, A. Chebira, and M. Vetterli, "Near-Optimal Sensor Placement for Linear Inverse Problems," *IEEE Transactions on Signal Processing*, vol. 62, no. 5, pp. 1135–1146, Mar. 2014.

[9] S. P. Chepuri and G. Leus, "Sparsity-Promoting Sensor Selection for Non-Linear Measurement Models," *IEEE Transactions on Signal Processing*, vol. 63, no. 3, pp. 684–698, Feb. 2015. [Online]. Available: http://ieeexplore.ieee.org/document/6981988/

[10] Y. Maday, N. Nguyen, A. Patera, and S. Pau, "A general multipurpose interpolation procedure: the magic points," *Communications on Pure and Applied Analysis*, vol. 8, no. 1, pp. 383–404, Oct. 2008. [Online]. Available: http://www.aimsciences.org/journals/displayArticles.jsp?paperID=3753

[11] F. Pukelsheim, *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006.

[12] Y. De Castro, F. Gamboa, D. Henrion, R. Hess, and J.-B. Lasserre, "Approximate Optimal Designs for Multivariate Polynomial Regression," 2018, to appear in Annals of Statistics.

[13] H. P. Wynn, "The Sequential Generation of D-Optimum Experimental Designs," *The Annals of Mathematical Statistics*, vol. 41, no. 5, pp. 1655–1664, 1970. [Online]. Available: http://www.jstor.org/stable/2239871

[14] M. Shamaiah, S. Banerjee, and H. Vikalo, "Greedy sensor selection: Leveraging submodularity," in *49th IEEE Conference on Decision and Control (CDC)*, Dec. 2010, pp. 2572–2577.

[15] G. Puy, P. Vandergheynst, and Y. Wiaux, "On Variable Density Compressive Sampling," *IEEE Signal Processing Letters*, vol. 18, no. 10, pp. 595–598, Oct. 2011.

[16] N. Chauffert, P. Ciuciu, J. Kahn, and P. Weiss, "Variable Density Sampling with Continuous Trajectories," *SIAM Journal on Imaging Sciences*, vol. 7, no. 4, pp. 1962–1992, 2014.

[17] A. Cohen, M. A. Davenport, and D. Leviatan, "On the Stability and Accuracy of Least Squares Approximations," *Foundations of Computational Mathematics*, vol. 13, no. 5, pp. 819–834, 2013.

[18] G. Chardon, "Design of variable densities for least-squares approximations," in *2017 International Conference on Sampling Theory and Applications (SampTA)*, Jul. 2017, pp. 566–569.

[19] G. Chardon, A. Cohen, and L. Daudet, "Sampling and Reconstruction of Solutions to the Helmholtz Equation," *Sampling Theory in Signal and Image Processing*, vol. 13, no. 1, pp. 67–89, 2014.

[20] G. Chardon, W. Kreuzer, and M. Noisternig, "Design of Spatial Microphone Arrays for Sound Field Interpolation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 780–790, Aug. 2015.

[21] J. Kiefer and J. Wolfowitz, "The equivalence of two extremum problems," *Canad. J. Math.*, vol. 12, pp. 363–366, 1960.

[22] J. Kiefer, "Optimum Experimental Designs V, with Applications to Systematic and Rotatable Designs," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley, Calif.: University of California Press, 1961, pp. 381–405.

[23] S. D. Silvey, D. H. Titterington, and B. Torsney, "An algorithm for optimal designs on a design space," *Communications in Statistics - Theory and Methods*, vol. 7, no. 14, pp. 1379–1389, 1978.

[24] Y. Yu, "D-optimal designs via a cocktail algorithm," *Statistics and Computing*, vol. 21, no. 4, pp. 475–481, Oct. 2011.

[25] A. Cohen and G. Migliorati, "Optimal weighted least-squares methods," *SMAI-Journal of computational mathematics*, no. 3, pp. 181 – 203, 2017.

[26] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval research logistics quarterly*, vol. 3, no. 1-2, pp. 95–110, 1956.

[27] M. Jaggi, "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization," in *International Conference on Machine Learning*, Feb. 2013, pp. 427–435. [Online]. Available: http://proceedings.mlr.press/v28/jaggi13.html

[28] D. J. White, "Extension of the Frank-Wolfe algorithm to concave nondifferentiable objective functions," *Journal of Optimization Theory and Applications*, vol. 78, no. 2, pp. 283–301, Aug. 1993. [Online]. Available: http://link.springer.com/article/10.1007/BF00939671

[29] R. Douc, O. Cappe, and E. Moulines, "Comparison of resampling schemes for particle filtering," in *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, Sep. 2005, pp. 64–69.

[30] C. Schretter and H. Niederreiter, "A direct inversion method for non-uniform quasi-random point sequences," *Monte Carlo Methods and Applications*, vol. 19, no. 1, pp. 1–9, Jan. 2013.

[Online]. Available: https://www.degruyter.com/view/j/mcma.2013.19.issue-1/mcma-2012-0014/mcma-2012-0014.xml

[31] M. Gerber, N. Chopin, and N. Whiteley, "Negative association, ordering and convergence of resampling methods," *arXiv:1707.01845 [stat]*, Jul. 2017, arXiv: 1707.01845. [Online]. Available: http://arxiv.org/abs/1707.01845

[32] A. Moiola, R. Hiptmair, and I. Perugia, "Plane wave approximation of homogeneous Helmholtz solutions," *Zeitschrift für angewandte Mathematik und Physik*, vol. 62, no. 5, p. 809, Oct. 2011. [Online]. Available: http://link.springer.com/article/10.1007/s00033-011-0147-y

[33] J. Guélat and P. Marcotte, "Some comments on Wolfe's 'away step'," *Mathematical Programming*, vol. 35, no. 1, pp. 110–119, May 1986. [Online]. Available: http://link.springer.com/article/10.1007/BF01589445