Frame-based Optimal Design

Sebastian Mair, Yannick Rudolph, Vanessa Closius, and Ulf Brefeld

Leuphana University of Lüneburg, Germany,
{mair,brefeld}@leuphana.de

Abstract. Optimal experimental design (OED) addresses the problem of selecting an optimal subset of the training data for learning tasks. In this paper, we propose to efficiently compute OED by leveraging the geometry of data: We restrict computations to the set of instances lying on the border of the convex hull of all data points. This set is called the frame. We (i) provide the theoretical basis for our approach and (ii) show how to compute the frame in kernel-induced feature spaces. The latter allows us to sample optimal designs for non-linear hypothesis functions without knowing the explicit feature mapping. We present empirical results showing that the performance of frame-based OED is often on par or better than traditional OED approaches, but its solution can be computed up to twenty times faster.

Keywords: Active Learning · Fast Approximation · Frame · Optimal Experimental Design · Regression

1 Introduction

Consider a supervised learning task with n unlabeled data points $\mathcal{X}$. Obtaining labels for all instances is prohibitive, but there is a budget k that allows to label $k \ll n$ points. The goal is to select the best subset of $\mathcal{X}$ of size k such that the learned model is optimal with respect to some optimality measure. This problem is known as Optimal Experimental Design (OED) [13].

In the classical setting of OED, the learning task is a linear regression $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$ with target vector $\mathbf{y}$, design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, parameters $\mathbf{w}$, and i.i.d. Gaussian noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$. There are many optimality criteria that can be employed for optimal designs. A common choice is to minimize the covariance of the parameter estimation given by

$$\mathrm{Cov}[\mathbf{w}] = \sigma^2 \left(\sum_{\substack{\mathbf{x} \in S \\ |S| = k}} \mathbf{x}\mathbf{x}^\top \right)^{-1} .$$

Minimizing the above quantity is equivalent to maximizing the confidence of the learned parameters. However, finding the subset S of size k that actually minimizes the covariance turns into a combinatorial problem that, depending on n and k, is often infeasible. There are two scenarios to cope with the situation.

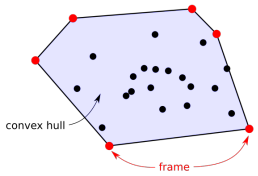

Fig. 1. Illustration of the frame.

The first one builds upon the assumption that experiments can be repeated many times such that the design matrix contains duplicate rows; hence, this approach requires obtaining multiple outcomes for the same experiment. The second and more relevant scenario does not allow for repeating the experiment.

Surrogates have been suggested to quantify the optimality of a subset. Popular choices exploit the determinant (D-optimality), the spectral-norm (E-optimality), or the trace (A-optimality) of the covariance matrix of the k points [23]. Nevertheless, intrinsically the problem remains combinatorial and very demanding and the only remedy being a pre-selection of promising candidate points to reduce the complexity of the task.

In this paper, we exploit the geometry of the data and propose to use the frame as such a candidate set of points. The frame is the smallest subset of the data that realizes the same convex hull as all data points. Thus, the frame consists of the extreme points of the data set. Figure 1 shows an example. We show that restricting the optimization problem to the frame yields competitive results in terms of optimality and predictive performance but comes with a much smaller computational cost. To leverage OED for non-linear problems, we devise a novel approach to compute the frame in kernel-induced feature spaces; this allows us to sample random designs for non-linear regression models without knowing the explicit feature mapping. Our approach of computing the frame can be seen as a transposed LASSO that selects data points instead of features. We discuss the relation to LASSO [27] in greater detail and also address the connection to active learning.

The remainder is structured as follows: Section 2 contains the main contribution of frame-based optimal experimental design and Section 3 reports on empirical results. Section 4 reviews related work and Section 5 concludes.

2 Frame-based Optimal Design

2.1 Preliminaries

We consider a discrete input set $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ consisting of n data points in d dimensions. The convex hull $\text{conv}(\mathcal{X})$ is the intersection of all convex sets containing $\mathcal{X}$. Furthermore, $\text{conv}(\mathcal{X})$ is the set of all convex combinations of points in $\mathcal{X}$. A central concept of this paper is the frame that is introduced in the following definition.

Definition 1. *Let $\mathcal{X}$ be a discrete input set. The minimal cardinality subset of $\mathcal{X}$, which produces the same convex hull as $\mathcal{X}$, is called the frame $\mathcal{F}$, i.e.,* $\mathrm{conv}(\mathcal{F}) = \mathrm{conv}(\mathcal{X})$.

Hence, the frame consists of the extreme points of $\mathcal{X}$. Those points cannot be represented as convex combinations of other points rather than themselves. By $q = |\mathcal{F}|$, we refer to the size of the frame and we call the portion of points in $\mathcal{X}$ belonging to the frame $\mathcal{F}$ the *frame density* q/n.

2.2 Optimal Experimental Design

In the classical setting of OED, the task is a linear regression

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y}$ is the vector of targets $y_i \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the pool of n experiments $\mathbf{x}_i \in \mathbb{R}^d$, $\mathbf{w} \in \mathbb{R}^d$ are the model parameters and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ is a vector of i.i.d. Gaussian noise. The maximum likelihood estimate of the parameters $\mathbf{w}$ has a closed-form and is given by

$$\mathbf{w}^\star = \operatorname*{argmin}_{\mathbf{w}} \sum_{i=1}^{n} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

The goal of OED is to choose a subset S of size k out of the n points for which the estimation of $\mathbf{w}$ is optimal in some sense. As common, we require $d \leq k \ll n$. Optimality can be measured in several ways. One idea is to increase the confidence of learning the parameters by minimizing the covariance of the parameter estimation For the regression problem stated above, the covariance matrix is given by

$$\mathrm{Cov}_S[\mathbf{w}] = \sigma^2 \left(\sum_{\mathbf{x} \in S} \mathbf{x}\mathbf{x}^\top \right)^{-1},$$

where $S \subset \mathcal{X}$ is the selected subset with $|S| = k$. This leads to a combinatorial optimization problem as follows

$$\min_{\boldsymbol{\lambda}} f\left(\sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^\top \right) \text{ s. t. } \sum_{i=1}^{n} \lambda_i \leq k \text{ and } \lambda_i \in \{0, 1\} \ \forall i. \tag{2}$$

Here, $f : \mathbb{S}_d^+ \to \mathbb{R}$ is an optimality criterion that assigns a real number to every feasible experiment (positive semi-definite matrices). The setting can be seen as maximizing the information we obtain from executing the experiment with fixed effort. The most popular choices for f are D-, E-, and A-optimality [23] given by

$$\begin{aligned} f_D(\Sigma) &= (\det(\Sigma))^{-1/d} && \text{(D-optimality)} \\ f_E(\Sigma) &= \|\Sigma^{-1}\|_2 && \text{(E-optimality)} \\ f_A(\Sigma) &= d^{-1} \operatorname{tr}(\Sigma^{-1}) && \text{(A-optimality)} \end{aligned}$$

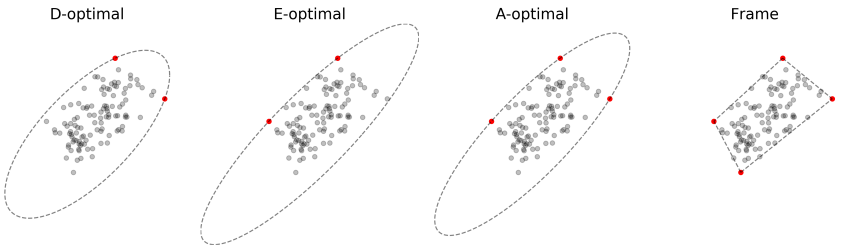

Fig. 2. Example of D-, E-, A-optimal designs and the frame on toy data.

Unfortunately, the combinatorial optimization problem in Equation (2) cannot be solved efficiently. A remedy is to use a continuous relaxation, which is efficiently solvable:

$$\min_{\boldsymbol{\lambda}} \; f\left(\sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^\top\right) \text{ s. t. } \sum_{i=1}^{n} \lambda_i \leq k \text{ and } \lambda_i \in [0,1] \; \forall i. \tag{3}$$

The following lemma characterizes the solution of the optimization problem above.

Lemma 1. *Let $\boldsymbol{\lambda}^\star$ be the optimal solution of Problem* (3)*. Then $\|\boldsymbol{\lambda}^\star\|_1 = k$.*

However, the support of $\boldsymbol{\lambda}^\star$ is usually much larger than k and the solution needs to be sparsified in order to end up with k experiments. Approaches therefor include pipage rounding schemes [1], sampling [28], regret minimization [2], and greedy removal strategies [19,28].

2.3 Restricting Optimal Experimental Design to the Frame

D-optimal design minimizes the determinant of the error covariance matrix. Its dual problem is known as Minimum Volume Confidence Ellipsoid [10]. Geometrically, the optimal solution is an ellipsoid that encloses the data with minimum volume. For E-optimality, the dual problem can be interpreted as minimizing the diameter of the confidence ellipsoid [5]. In A-optimal design the goal is to find the subset of points that optimizes the total variance of parameter estimation. Figure 2 depicts the support of the optimal solution $\boldsymbol{\lambda}^\star$ for D-, E-, and A-optimal designs as well as their confidence ellipsoids derived from their dual problems. The right hand figure shows the frame of the same data. The confidence ellipsoids clearly touch the points at the border of the data while the interior points are enclosed. Hence, we propose to discard all interior points entirely in the optimization and restrict the optimization to the frame, that is, to the points lying on the border of the convex hull.

Non-linear regression can be done by applying a feature mapping $\phi : \mathcal{X} \to \mathcal{X}'$ to the data. The model then becomes $y_i = \mathbf{w}^\top \phi(\mathbf{x}_i)$, which is still linear in parameters. Considering the dual of the regression problem we can employ

kernels that implicitly do a feature mapping. However, the regression is still a linear model, but in feature space $\mathcal{X}'$. Knowing the frame in $\mathcal{X}'$ would allow us to sample random designs rendering a naive version of non-linear or kernelized OED possible. In Subsection 2.5, we show how to compute the frame in kernel-induced feature spaces.

2.4 Computing the Frame

As outlined in Definition 1, the frame $\mathcal{F}$ is the minimum cardinality subset of the data which yields the same convex hull as the data $\mathcal{X}$. Hence, it is given as a solution of the following problem:

$$\begin{aligned} \mathcal{F} = \underset{\{\mathbf{z}_1,\ldots,\mathbf{z}_m\}\subseteq\mathcal{X}}{\operatorname{argmin}} \quad & |\{\mathbf{z}_1,\ldots,\mathbf{z}_m\}| \\ \text{s.t.} \quad & \forall \mathbf{x}\in\mathcal{X} : \mathbf{x} = \sum_j \mathbf{s}_j\mathbf{z}_j \text{ with } \mathbf{s}^\top\mathbf{1} = 1 \text{ and } \mathbf{s}_j \geq 0. \end{aligned} \tag{4}$$

We briefly review prior work [18] employing quadratic programming to compute a representation of every data point using only points from the frame. The representation of $\mathbf{x}_i$ is given by a convex combination of frame points as

$$\begin{aligned} \underset{\mathbf{s}}{\text{solve}} \quad & \mathbf{X}^\top\mathbf{s} = \mathbf{x}_i \\ \text{s.t.} \quad & \mathbf{s}_j \geq 0 \quad \forall j \\ & \mathbf{s}^\top\mathbf{1} = 1 \\ & \mathbf{s}_j \neq 0 \Rightarrow \mathbf{x}_j \in \mathcal{F} \quad \forall j. \end{aligned} \tag{5}$$

Equation (5) can be rewritten as a non-negative least-squares problem with an additional condition that only points on the frame are allowed to contribute to the solution $\mathbf{s}$. [18] also shows that the NNLS algorithm of Lawson and Hanson [15] solves the resulting optimization problem. After computing the representation $\mathbf{s}$ for every point, the frame is recovered by unifying the support of every $\mathbf{s}$. This yields the full frame since every frame point can only be represented by itself.

2.5 Computing the Frame in Kernel-induced Feature Spaces

Let $\phi : \mathbb{R}^d \to \mathbb{R}^D$ be a feature mapping, $\boldsymbol{\Phi} \in \mathbb{R}^{D\times n}$ be the mapped design matrix, and $\mathbf{K}$ be the kernel matrix induced by a kernel $k(\mathbf{x},\mathbf{z}) = \phi(\mathbf{x})^\top\phi(\mathbf{z})$. As before, the idea is to solve a linear system subject to the constraints that the solution $\mathbf{s}$ is non-negative, sums up to one, and uses only points from the frame; however, this time, we aim to solve the problem in feature space spanned by ϕ. We obtain

$$\begin{aligned} \underset{\mathbf{s}}{\text{solve}} \quad & \boldsymbol{\Phi}\mathbf{s} = \phi(\mathbf{x}_i) \\ \text{s.t.} \quad & \mathbf{s}_j \geq 0 \quad \forall j \\ & \mathbf{1}^\top\mathbf{s} = 1 \\ & \mathbf{s}_j \neq 0 \Rightarrow \phi(\mathbf{x}_j) \in \mathcal{F} \quad \forall j. \end{aligned}$$

The constraint $\mathbf{1}^\top \mathbf{s} = 1$ can be incorporated into the system of linear equations by augmenting Φ with a row of ones and $\phi(\mathbf{x})$ with a static 1. Let $\psi(\mathbf{x}) = (\phi(\mathbf{x})^\top, 1)^\top$ and $\Psi = (\psi(\mathbf{x}_1), \ldots, \psi(\mathbf{x}_n)) \in \mathbb{R}^{(D+1)\times n}$, we obtain

$$\text{solve } \Psi\mathbf{s} = \psi(\mathbf{x}_i) \text{ s.t. } \mathbf{s}_j \geq 0 \ \wedge \ \mathbf{s}_j \neq 0 \Rightarrow \phi(\mathbf{x}_j) \in \mathcal{F}.$$

The approach can be kernelized by multiplying from the left with $\Psi^\top$:

$$\Psi^\top\Psi\mathbf{s} = \Psi^\top\psi(\mathbf{x}_i) \text{ s.t. } \mathbf{s}_j \geq 0 \ \wedge \ \mathbf{s}_j \neq 0 \Rightarrow \phi(\mathbf{x}_j) \in \mathcal{F}.$$

Since there is always a solution [18], we can equivalently solve the non-negative least squares problem

$$\begin{aligned} \underset{\mathbf{s}\geq 0}{\operatorname{argmin}} \quad & \frac{1}{2}\|\Psi^\top\Psi\mathbf{s} - \Psi^\top\psi(\mathbf{x}_i)\|_2^2 \\ \text{s.t.} \quad & \mathbf{s}_j \neq 0 \Rightarrow \phi(\mathbf{x}_j) \in \mathcal{F}. \end{aligned} \tag{6}$$

A kernel can now be applied by exploiting the relationship between Ψ, Φ, and $\mathbf{K}$ as follows

$$\Psi^\top\Psi = \Phi^\top\Phi + \mathbb{1}_{nn} = \mathbf{K} + \mathbb{1}_{nn} =: \mathbf{L} \tag{7}$$

$$\Psi^\top\psi(\mathbf{x}_i) = \Phi^\top\phi(\mathbf{x}_i) + \mathbb{1}_{n1} = \mathbf{K}_{\cdot i} + \mathbb{1}_{n1} = \mathbf{L}_{\cdot i}, \tag{8}$$

where $\mathbb{1}_{nm} \in \mathbb{R}^{n\times m}$ denotes the matrix of ones. The resulting problem becomes

$$\begin{aligned} \underset{\mathbf{s}\geq 0}{\operatorname{argmin}} \quad & \frac{1}{2}\|\mathbf{L}\mathbf{s} - \mathbf{L}_{\cdot i}\|_2^2 \\ \text{s.t.} \quad & \mathbf{s}_i \neq 0 \Rightarrow \phi(\mathbf{x}_i) \in \mathcal{F}. \end{aligned} \tag{9}$$

A standard non-negative least squares problem can be solved, for example, by the algorithm of Lawson and Hanson [15]. Bro and De Jong [6] increase the efficiency by caching the quantities in Equation (7). This renders the problem in Equation (9) feasible. To demonstrate this, we first show that whenever an inner product between two points is maximized, one of the points is an extreme point and thus belongs to the frame of the data.

Lemma 2. *Let $\mathcal{X}$ be a finite set of discrete points, then*

$$\forall \mathbf{x} \in \mathcal{X}: \quad \underset{\mathbf{x}'\in\mathcal{X}}{\operatorname{argmax}} \ \langle \mathbf{x}, \mathbf{x}' \rangle \in \mathcal{F}.$$

Proof. Linearity and convexity of the inner product imply that its maximum is realized by an extreme point of the domain. Since the domain is $\mathcal{X}$, the maximum belongs to its frame $\mathcal{F}$. □

Theorem 1. *The active-set method from Bro and De Jong [6] solves the problem in Equation* (9)*.*

Algorithm 1 Kernel-Frame

Input: kernel matrix $\mathbf{K}$
Output: indices of ext. points $\mathcal{E}$
$\mathbf{L} = \mathbf{K} + \mathbb{1}_{nn}$
$\mathcal{E} = \emptyset$
for $i = 1, 2, \ldots, n$ **do**
  $\mathbf{s}_i = \text{bro-dejong}(\mathbf{L}, \mathbf{L}[:, i])$
  $\mathcal{P}_i = \{\ j \in \{1, 2, \ldots, n\} \mid (\mathbf{s}_i)_j > 0\ \}$
  $\mathcal{E} = \mathcal{E} \cup \mathcal{P}_i$
end for

Proof. The algorithm selects points that contribute to the solution $\mathbf{s}$ by maximizing the negative gradient of the objective. The selection is implemented by the criterion $j = \text{argmax}_j\ [\mathbf{L}_{\cdot i} - \mathbf{L}\mathbf{s}]_j$, where j is the index of the selected point. Thus, the selection process is maximizing a linear function and Lemma 2 assures that this point belongs to the frame. □

Solving problem (6) for the i-th data point $\mathbf{x}_i$ yields either its index i in case $\mathbf{x}_i$ is a point on the frame or, if $\mathbf{x}_i$ is an interior point, the solution is the index set $\mathcal{P}_i$ of points on the frame that recover $\mathbf{x}_i$ as a convex combination. The entire frame is recovered by solving Equation (6) for all points in $\mathcal{X}$ as stated in Corollary 1 and depicted in Algorithm 1.

Corollary 1. *Let $k(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^\top \phi(\mathbf{z})$ be a kernel and $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a data set. Then Algorithm 1 yields the frame of $\{\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)\}$.*

Proof. Algorithm 1 computes the solution $\mathbf{s}_i$ for every mapped data point $\phi(\mathbf{x}_i)$. Theorem 1 ensures that the positive positions of every $\mathbf{s}_i$ $(i = 1, 2, \ldots, n)$ refers to points on the frame. Hence, taking the union of those positions recovers the frame indices $\mathcal{E}$. □

Note first that the frame in kernel-induced feature space can be found without knowing the explicit feature map ϕ and second that the *for*-loop in Algorithm 1 can be trivially parallelized.

2.6 Frame Densities for Common Kernels

To analyze the frame sizes in kernel-induced feature spaces, we focus on rbf and polynomial kernels. The former is given by $k(\mathbf{x}, \mathbf{z}) = \exp(-\gamma \|\mathbf{x} - \mathbf{z}\|_2^2)$, where $\gamma > 0$ is a scaling parameter. The induced feature mapping ϕ of the rbf kernel has an infinite dimensionality. Corollary 2 shows, that this kernel always yields a full frame, that is: every point belongs to the frame and the frame density is consequently equal to one.

Corollary 2. *Let $\mathcal{X}$ be the data set of distinct points and k be the rbf kernel with parameter $\gamma \neq 0$. Then every point belongs to the frame $\mathcal{F}$ in feature space.*

Proof. Gaussian gram matrices have full rank [24]. Hence, the images $\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)$ in feature space are linearly independent. Thus, every image can only be represented by itself and, every point belongs to the frame $\mathcal{F}$. □

The polynomial kernel is given by $k(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\top \mathbf{z} + c)^p$ with degree $p \in \mathbb{N}$ and constant $c \in \mathbb{R}_0^+$. A feature of the polynomial kernel is an explicit representation of the implicit feature mapping ϕ. E.g., for the homogeneous polynomial kernel with $c = 0$, we have

$$\phi_{\mathbf{m}}(\mathbf{x}) = \sqrt{\frac{p!}{\prod_{i=1}^n m_i!}} \prod_{i=1}^n x_i^{m_i}$$

for all multi-indices $\mathbf{m} = (m_1, \ldots, m_n) \in \mathbb{N}^n$ satisfying $\sum_{i=1}^n m_i = p$. That is, new features consist of monomials of the input features x_i, while the multi-indices $\mathbf{m}$ denote their respective degrees. The condition $\sum_{i=1}^n m_i = p$ assures that all possible combinations are uniquely accounted for and leads to a feature space dimension of size

$$D = \binom{p+d-1}{p} = \frac{(p+d-1)!}{p!(d-1)!}.$$

For the explicit mapping corresponding to the heterogeneous kernel (where $c \neq 0$) that realizes a feature space with dimensionality

$$D = \binom{p+d}{p},$$

as well as for more details, we refer to [25] and [24]. For the polynomial kernel we obtain a full frame if the dimension D of the feature space exceeds the number of data points n.

Corollary 3. *Let $\mathcal{X}$ be the normalized and distinct data set of size n in d dimensions and k be the polynomial kernel with degree p and offset $c = 0$. If $n \leq \frac{(p+d-1)!}{p!(d-1)!}$, then every point belongs to the frame $\mathcal{F}$ in feature space.*

Proof. The polynomial feature map yields linearly independent feature vectors of size $\frac{(p+d-1)!}{p!(d-1)!}$ for a data set with unique observations. Hence, if the number of data points is lower than the dimensionality of the mapping, all points belong to the frame $\mathcal{F}$. □

Although a formal proof regarding the influence of the degree p of the homogeneous polynomial kernel is missing, we would like to provide some intuition: We empirically apply a homogeneous polynomial kernel to a synthetic data set with $n = 2500$ points in $d = 5$ dimensions with an initial frame density of 1%. Figure 3 shows the resulting frame densities. For odd degrees, the frame density is growing with increasing values of p. This is due to the increasing dimensionality in feature space. However, for even degrees the frame is always full. We conclude with the following conjecture:

Conjecture 1. Let $\mathcal{X}$ be the normalized and distinct data set of size n in d dimensions and k be the polynomial kernel with degree p and offset $c = 0$. If p is even, then every point belongs to the frame $\mathcal{F}$ in feature space.

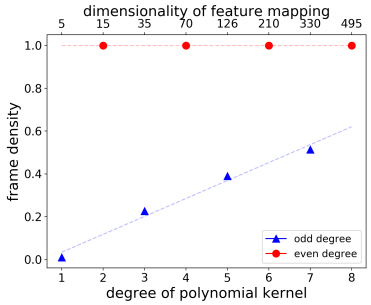

Fig. 3. Frame density for various polynomial degrees on synthetic data of size $n = 2500$ in $d = 5$ dimensions. The initial frame density is 1%. The data set is introduced in Section 3.

2.7 Computing the Frame and LASSO

LASSO [27] solves regression tasks by combining a squared loss with an ℓ_1-regularizer on the parameters. Thus, LASSO simultaneously performs a regression and variable selection such that the influence of redundant variables is set to zero and a sparse parameter vector is obtained. The corresponding optimization problem for a regression scenario as in Equation (1) is given by

$$\min_{\mathbf{w}} \ \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1,$$

where $\lambda \geq 0$ is a trade-off parameter. A special case is obtained by restricting the parameters to be positive, yielding a *non-negative LASSO*:

$$\min_{\mathbf{w}\geq 0} \ \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1 \quad \Longleftrightarrow \quad \min_{\mathbf{w}\geq 0} \ \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\mathbf{1}^\top\mathbf{w}.$$

Computing the frame can be seen as a transposed version of the LASSO problem in which not variables but data points are selected. The following proposition shows that the problem in Equation (9) is equivalent to a non-negative LASSO, if one ignores the constraint that elicits only frame points to contribute to the solution.

Proposition 1. *Problem* (9) *solved with the active-set method from Bro and De Jong is equivalent to a non-negative LASSO with trade-off parameter* $\lambda = n$.

Proof. By using the identities $\mathbf{L} = \mathbf{K} + \mathbb{1}_{nn}$ and $\mathbf{L}_{\cdot i} = \mathbf{K}_{\cdot i} + \mathbf{1} = \mathbf{k} + \mathbf{1}$, we rewrite the objective of the optimization problem in Equation (9) as follows:

$$\begin{aligned}
\|\mathbf{L}\mathbf{s} - \mathbf{L}_{\cdot i}\|_2^2 &= \|(\mathbf{K} + \mathbb{1})\mathbf{s} - (\mathbf{k} + \mathbf{1})\|_2^2 = \|\mathbf{K}\mathbf{s} - \mathbf{k}\|_2^2 + \|\mathbb{1}\mathbf{s} - \mathbf{1}\|_2^2 \\
&= \|\mathbf{K}\mathbf{s} - \mathbf{k}\|_2^2 + n\|\mathbf{1}^\top\mathbf{s} - 1\|_2^2 = \|\mathbf{K}\mathbf{s} - \mathbf{k}\|_2^2 + n\mathbf{1}^\top\mathbf{s} - n \\
&\equiv \|\mathbf{K}\mathbf{s} - \mathbf{k}\|_2^2 + n\mathbf{1}^\top\mathbf{s} = \|\mathbf{K}\mathbf{s} - \mathbf{k}\|_2^2 + n\|\mathbf{s}\|_1.
\end{aligned}$$

Hence, the objective is an ℓ_1-regularized least-squares problem. In combination with the non-negativity constraint, we obtain a non-negative LASSO. □

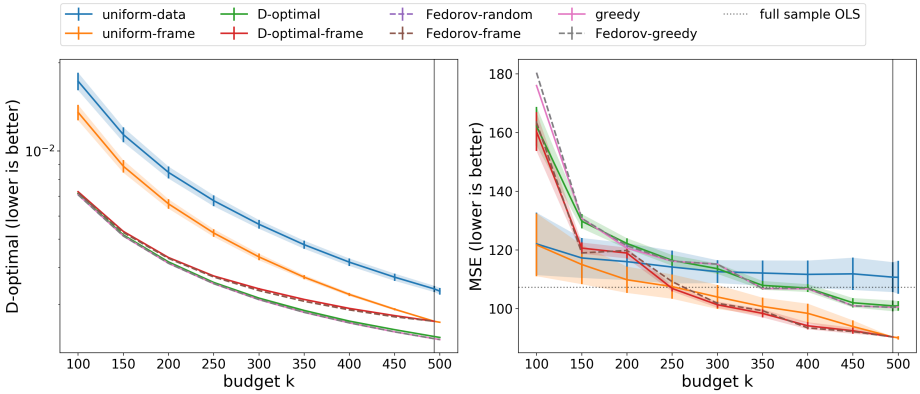

Fig. 4. Results for D-optimal designs on Concrete.

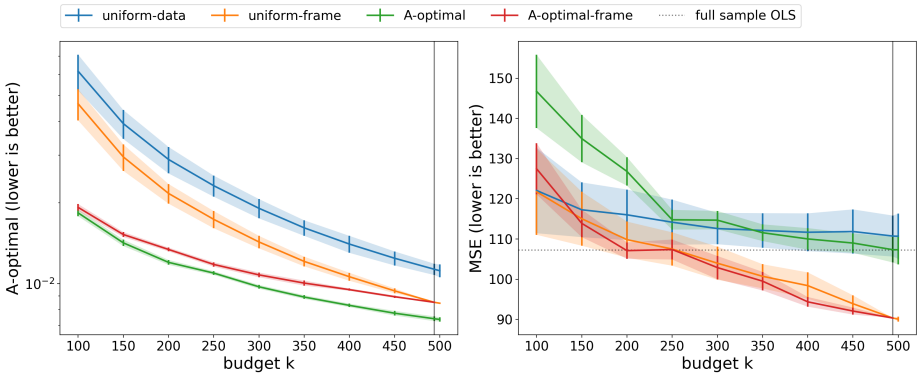

Fig. 5. Results for A-optimal designs on Concrete.

3 Experiments

In this section, we empirically investigate frame-based optimal experimental design. Throughout this section, we compare the performance of the following different approaches. *Uniform-data* samples the subset S uniformly at random without replacement from all data points $\mathcal{X}$. A second approach *uniform-frame* uses the same strategy but samples points from the frame $\mathcal{F}$ instead of $\mathcal{X}$. If the size of $|S|$ exceeds the size of the frame, *uniform-frame* always draws the full frame and randomly selects the remaining points from $\mathcal{X} \setminus \mathcal{F}$. The *greedy* baseline chooses the points in S one after another according to their contribution to the objective of D-optimal design. The baselines $\{D,E,A\}$*-optimal* use the continuous relaxations of the {D,E,A}-optimal design criteria, respectively. After solving the optimization problem, we sample the subset S according to a strategy outlined by

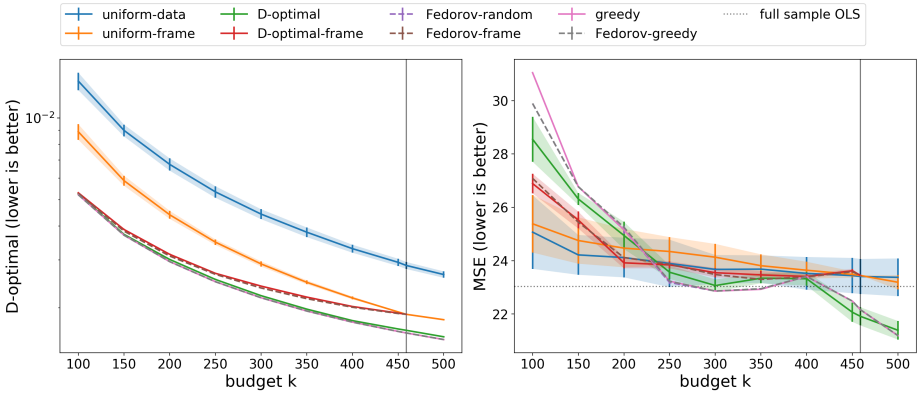

Fig. 6. Results for D-optimal designs on Airfoil.

Wang et al. [28]. Analogously, $\{D,E,A\}$-*optimal-frame* restricts the computation of the previous three baselines to the frame. Finally, the *Fedorov* baseline selects S according to the Fedorov Exchange algorithm [13] and optimizes D-optimality. We initialize this baseline using random samples from $\mathcal{X}$, random samples from $\mathcal{F}$, and with the output of *greedy*.

The continuous relaxations are optimized using sequential quadratic programming [20]; the number of iterations is limited to 250. We report on average performances over 100 repetitions, error bars indicate one standard deviation and a vertical line denotes the frame size when included. The greedy algorithm is executed once and we conduct only 10 repetitions for every Fedorov Exchange initialization due to its extensive runtime. We want to shed light on the following list of questions.

Is the restriction to the frame competitive in terms of performance? The first experiment thus studies the performance of optimal designs of the proposed approaches on the real-world data set Concrete [29]. Concrete consists of a design pool of $n = 1030$ instances with $d = 8$ dimensions and has a frame density of 48%. The task is to predict the compressive strength of different types of concrete.

We measure the performance in terms of the D-optimality criterion as well as the mean squared error (MSE), given by $\text{MSE} = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$. For the latter, we train an ordinary least squares regression on the selected points and evaluate on the remaining $n - k$ points.

Figure 4 (left) shows the results with respect to the D-optimality criterion. Sampling uniformly from the frame (*uniform-frame*) performs consistently better than sampling from all data (*uniform-data*). Thus, exploiting the frame allows to sample better designs without solving any optimization problem other than computing the frame. The situation changes once the designs are

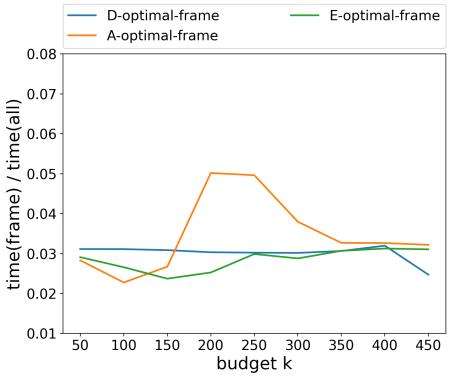

Fig. 7. Timing results on Airfoil.

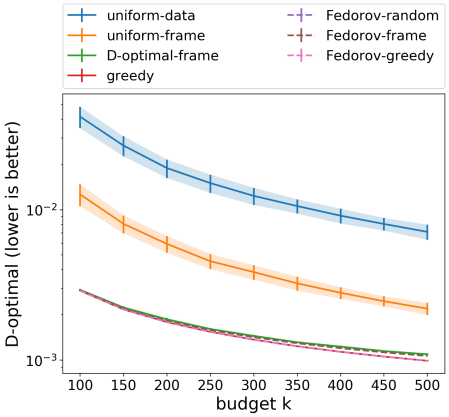

Fig. 8. Results on California Housing.

optimized in addition. Frame-based approaches (**-optimal-frame*) are close to their competitors computed on all data (**-optimal*) but no longer better. Interior points thus do contribute, if only marginally, to the optimization.

However, Figure 4 (right) shows that the slight improvement in the objective function does not carry over for the predictive performance. By contrast, the frame-based approaches (**-optimal-frame*) consistently outperform the other approaches and lead to significantly lower MSEs. For comparison, the MSE trained and evaluated on all data points is shown as a dashed horizontal line. Training only on a few points of the frame already leads to more accurate models than using all available data.

We obtain similar pictures for evaluating against the A- and E-optimality criteria. Due to their similar performance we only report on the results for A-optimal designs in Figure 5. Once again, the frame-based optimization is only sightly worse in terms of the optimization objective (left) but clearly outperforms the traditional approaches in predictive performance (right).

We additionally experiment on the Airfoil data [7]. The task is to predict the self-noise of airfoil blades of different designs and the data comes with $n = 1503$ experiments describing tests in a wind tunnel with $d = 5$ attributes and the data has a frame density of 31%.

The results for D-optimal designs are shown in Figure 6. Once again, the frame-based approaches perform slightly worse or on par in terms of the optimality criterion. However, the predictive performance measured in MSE is no longer superior. The errors are similar to those using uniform samples of the data. Thus, the dataset shows that even though the optimality criterion is well approximated, an error reduction is not guaranteed. However, this does not pose as a limitation to our approach as D-optimal design does not guarantee a reduction either.

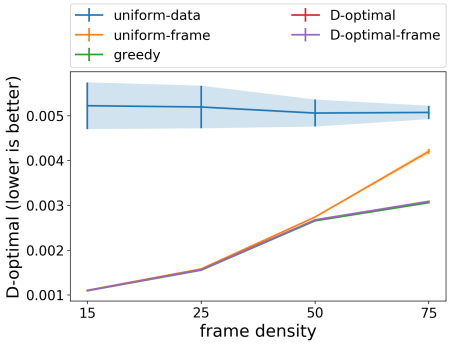

Fig. 9. Effect of the frame size on synthetic data.

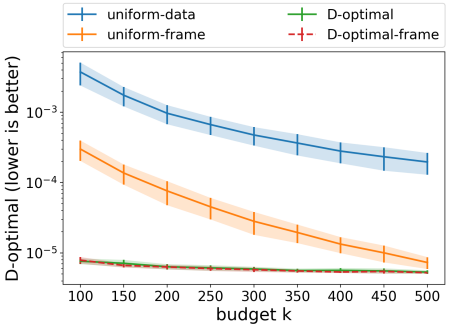

Fig. 10. Results on synthetic data using a polynomial kernel of degree $p = 3$.

Is the restriction to the frame efficient? We now report on the efficiency of our approach on Airfoil. Figure 7 illustrates the relative time of the frame-based approaches in comparison to their traditional analogues that are computed on all data. The y-axis thus shows *time(frame)/time(all)*. We can report a drastically faster computation taking only 2-5% of the time of the traditional variants. We credit this finding to Airfoil's frame density of 31%. That is, restricting the data to the frame already discards 69% of the data and the resulting optimization problems are much smaller.

Naturally, the smaller the frame size the faster the computation as we leave out more and more interior points. We thus experiment on the California Housing data [22] where the task is to estimate the median housing prices for different census blocks. This data comes with $n = 20,640$ instances in $d = 8$ dimensions but possesses a frame density of only 8%.

Figure 8 depicts the result with respect to the D-optimal criterion. The figure again shows that naively sampling from the frame (*uniform-frame*) is significantly better than a drawing random samples from all data (*uniform-data*). All other tested algorithms perform even better and realize almost identical curves. The D-optimal baseline could not be computed in a reasonable amount of time due to the size of the data. Only restricting the computations to the frame rendered the computation feasible.

What is the impact of the frame density? We already mentioned that the frame density q/n influences the efficiency of frame-based approaches. A frame density of z implies that the $(1 - z)$-th part of the data are interior points and can thus be ignored in subsequent computations.

To show this influence empirically, we control the frame density on synthetic data from López [17]. The data we use consists of $n = 2,500$ instances in $d = 5$ dimensions and comes in five different sets realizing frame densities of 1%, 15%, 25%, 50% and 75% respectively. Figure 9 shows the resulting D-optimality criteria for the different frame densities. Up to a frame density of

50%, randomly sampling from the frame (*uniform-frame*) performs on par with all other approaches, thus showing the efficiency of our proposal. For higher frame densities the performance of *uniform-frame* diverges towards *uniform-data*. Nevertheless, restricting D-OED to the frame stays on par with its peers. This experiment suggests, that the smaller the frame density the better the competitiveness of frame-based OED.

Does sampling in kernel-induced feature spaces work? In our last set of experiments, we consider sampling random designs on synthetic data for non-linear regression problems. We use synthetic data as described above with a frame density of 1%. We employ a homogeneous ($c = 0$) polynomial kernel with a degree of $p = 3$ that allows for obtaining the explicit feature mapping ϕ which is needed for all approaches except *uniform-**.

Figure 10 illustrates the results. Approaches optimizing the D-optimal design criterion (*D-**) perform equally well, irrespectively of whether they sample from the frame or not. This result confirms the competitiveness of restricting OED to the frame. However, both approaches rely on the explicit feature map.

Strategies that are purely based on sampling (*uniform-**) do not need an explicit mapping. Sampling at random from all data (*uniform-data*) trivially does not rely on anything but a list of indexes. Finally, sampling from the frame (*uniform-frame*) uses the proposed kernel frame algorithm (Algorithm 1) to sample in feature space. The figure shows that our approach samples much better designs from the frame which is only 23% in feature space. The larger the sample size, the less relevant becomes an explicit mapping.

4 Related Work

Optimal Experimental Design is a well-studied problem in statistics [13,23]. Recent work focuses on efficiency and performance and aims to devise approximation guarantees for relaxations of the combinatorial problems. For example, Wang et al. [28] consider A-optimal designs and propose sampling strategies (for the settings with and without replacement) with statistical efficiency bounds as well as a greedy removal approach. Allen-Zhu et al. [2] propose a regret-minimization strategy for the setting without replacement which works for most optimality criteria. Mariet and Sra [19] use elementary symmetric polynomials (ESP) for OED and introduce ESP-design, a interpolation between A- and D-optimal design that includes both as special cases. They provide approximation guarantees for sampling and greedy removal strategies.

OED has close ties to many other problems. D-optimality, for example, is related to volume sampling [3,9,16] and determinantal point processes [14]; both are used in many applications to sample informative and diverse subsets.

The problem setting we consider is moreover related to active learning [26,8]. Common active learning strategies sequentially select data points based on some uncertainty criterion or heuristic. Data points are for instance selected based on the confidence of the model to an assigned label or according to the

maximal model update in the worst case. Usually active learning iteratively selects instances and then re-trains to include the newly gained label into the model generation. In contrast to such iterative active learning scenarios with feedback, OED corresponds to selecting a single optimal batch prior to labeling and learning.

The frame can be straight forwardly obtained by convex hull algorithms. However, many of them are motivated and limited to two- or three-dimensional settings. Quickhull [4] works in higher dimensionalities but quickly becomes infeasible. If the enumeration of vertices is dropped, convex hull algorithms can be turned into methods that directly (and only) compute the frame. Common approaches for examples include linear programming to test whether a point is part of the frame or not [11,21,12]. Recent methods use quadratic programming to efficiently compute the frame [18].

5 Conclusion

We proposed to leverage the geometry of the data to efficiently compute optimal designs. Our contribution was motivated by the observation that traditional OED variants optimize enclosing ellipsoids that are supported by extreme data points. Hence, we proposed to restrict the computations to the frame which is the smallest subset of the data that yields the same convex hull as all data. We devised an optimization problem to compute the frame to sample random designs in kernel-induced feature spaces and provided a theoretical foundation for the eligibility of different kernel functions. Our contribution can be viewed as a transposed version of LASSO that selects data points instead of features.

Empirically, we showed that restricting optimal design to the frame yields competitive designs with respect to D-, E-, and A-optimality criteria on several real-world data sets. Interior data points are ignored by our frame-based approaches and we observed computational speed-ups of up to a factor of twenty. Our contribution rendered OED problems feasible on data at large scales for moderate frame densities.

References

1. Ageev, A.A., Sviridenko, M.I.: Pipage rounding: A new method of constructing algorithms with proven performance guarantee. Journal of Combinatorial Optimization **8**(3), 307–328 (2004)
2. Allen-Zhu, Z., Li, Y., Singh, A., Wang, Y.: Near-optimal design of experiments via regret minimization. In: International Conference on Machine Learning. pp. 126–135 (2017)
3. Avron, H., Boutsidis, C.: Faster subset selection for matrices and applications. SIAM Journal on Matrix Analysis and Applications **34**(4), 1464–1499 (2013)
4. Barber, C.B., Dobkin, D.P., Huhdanpaa, H.: The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software (TOMS) **22**(4), 469–483 (1996)

5. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge university press (2004)
6. Bro, R., De Jong, S.: A fast non-negativity-constrained least squares algorithm. Journal of Chemometrics **11**(5), 393–401 (1997)
7. Brooks, T.F., Pope, D.S., Marcolini, M.A.: Airfoil self-noise and prediction (1989)
8. Chaudhuri, K., Kakade, S.M., Netrapalli, P., Sanghavi, S.: Convergence rates of active learning for maximum likelihood estimation. In: Advances in Neural Information Processing Systems. pp. 1090–1098 (2015)
9. Dereziński, M., Warmuth, M.K.: Subsampling for ridge regression via regularized volume sampling. arXiv preprint arXiv:1710.05110 (2017)
10. Dolia, A.N., De Bie, T., Harris, C.J., Shawe-Taylor, J., Titterington, D.M.: The minimum volume covering ellipsoid estimation in kernel-defined feature spaces. In: European Conference on Machine Learning. pp. 630–637. Springer (2006)
11. Dulá, J.H., Helgason, R.V.: A new procedure for identifying the frame of the convex hull of a finite collection of points in multidimensional space. European Journal of Operational Research **92**(2), 352–367 (1996)
12. Dulá, J.H., López, F.J.: Competing output-sensitive frame algorithms. Computational Geometry **45**(4), 186–197 (2012)
13. Fedorov, V.V.: Theory of optimal experiments. Elsevier (1972)
14. Kulesza, A., Taskar, B., et al.: Determinantal point processes for machine learning. Foundations and Trends® in Machine Learning **5**(2–3), 123–286 (2012)
15. Lawson, C.L., Hanson, R.J.: Solving least squares problems, vol. 15. SIAM (1995)
16. Li, C., Jegelka, S., Sra, S.: Polynomial time algorithms for dual volume sampling. In: Advances in Neural Information Processing Systems. pp. 5045–5054 (2017)
17. Lopez, F.J.: Generating random points (or vectors) controlling the percentage of them that are extreme in their convex (or positive) hull. Journal of Mathematical Modelling and Algorithms **4**(2), 219–234 (2005)
18. Mair, S., Boubekki, A., Brefeld, U.: Frame-based data factorizations. In: International Conference on Machine Learning. pp. 2305–2313 (2017)
19. Mariet, Z.E., Sra, S.: Elementary symmetric polynomials for optimal experimental design. In: Advances in Neural Information Processing Systems. pp. 2136–2145 (2017)
20. Nocedal, J., Wright, S.J.: Sequential quadratic programming. Springer (2006)
21. Ottmann, T., Schuierer, S., Soundaralakshmi, S.: Enumerating extreme points in higher dimensions. Nordic Journal of Computing **8**(2), 179–192 (2001)
22. Pace, R.K., Barry, R.: Sparse spatial autoregressions. Statistics & Probability Letters **33**(3), 291–297 (1997)
23. Pukelsheim, F.: Optimal Design of Experiments. SIAM (2006)
24. Schölkopf, B., Smola, A.J.: Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press (2002)
25. Smola, A.J., Schölkopf, B., Müller, K.R.: The connection between regularization operators and support vector kernels. Neural Networks **11**(4), 637–649 (1998)
26. Sugiyama, M., Nakajima, S.: Pool-based active learning in approximate linear regression. Machine Learning **75**(3), 249–274 (2009)
27. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288 (1996)
28. Wang, Y., Yu, A.W., Singh, A.: On computationally tractable selection of experiments in measurement-constrained regression models. Journal of Machine Learning Research **18**(143), 1–41 (2017)
29. Yeh, I.C.: Modeling of strength of high-performance concrete using artificial neural networks. Cement and Concrete research **28**(12), 1797–1808 (1998)