Mathematical Programming
https://doi.org/10.1007/s10107-019-01421-z

FULL LENGTH PAPER

Series A

Multidimensional sum-up rounding for integer programming in optimal experimental design

Jing Yu[1] · Mihai Anitescu[2]

Received: 16 December 2017 / Accepted: 6 August 2019

Abstract

We present a numerical method for approximating the solution of convex integer programs stemming from optimal experimental design. The statistical setup consists of a Bayesian framework for linear inverse problems for which the direct relationship is described by a discretized integral equation. Specifically, we aim to find the optimal sensor placement from a set of candidate locations where data are collected with measurement error. The convex objective function is a measure of the uncertainty, described here by the trace or log-determinant of the posterior covariance matrix, for the discretized linear inverse problem solution. The resulting convex integer program is relaxed, producing a lower bound. An upper bound is obtained by extending the sum-up rounding approach to multiple dimensions. For this extension, we analyze its accuracy as a function of the discretization mesh size for a rectangular domain. We show asymptotic optimality of the integer solution defining the upper bound for different experimental design criteria (A- and D-optimal), by proving the convergence to zero of the gap between the upper and lower bounds as the mesh size goes to zero. The technique is illustrated on a two-dimensional gravity surveying problem for both A-optimal and D-optimal sensor placement where our designs yield better results compared with a thresholding rounding approach.

Keywords Optimal experimental design · Sum-up rounding · Integer programming · Bayesian inverse problem

Also, Preprint ANL/MCS 9032-1218.

Government License: The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

✉ Mihai Anitescu
anitescu@mcs.anl.gov

Extended author information available on the last page of the article

Published online: 06 September 2019

Mathematics Subject Classification 62K05 · 90C10 · 15A29

1 Introduction

Design of experiments (DOE) is an important endeavor of statistics. It aims to determine experimental settings that yield accurate results for statistical model parameters. In traditional DOE, one selects suitable treatments (explanatory variables), assigns the treatments to experimental units under statistically optimal conditions (usually to minimize the variance of parameter estimates), and observes treatment effects by measuring response variables (see [9]). Another important branch of DOE seeks to determine the optimal sampling locations given a set of available measurement points (see [8, §7.5] and [25, §9, §12]). In [8], the goal is to select m regression vectors with replacement from a prescribed set of p regression vectors, so as to obtain best ordinary least squares (OLS) estimates. The optimality criteria are based on the trace, log-determinant, or maximum eigenvalue of the covariance matrix of OLS estimates. This is an integer programming problem, and it is generally NP-hard [30]. One tractable approach is to first solve the convex problem obtained from relaxing the integrality constraints, and then round the solution off to an integer one. In [25], the setting is also linear, where measurements are selected from an infinite set of regression vectors, allowing for repeated measurements. Several efficient rounding-to-integrality procedures are proposed and an analysis of asymptotic performance loss is given. A common feature of all these approaches is that the analysis is done with respect to a fixed number of model parameters.

Our focus of investigation is related to such previous endeavors but takes a different direction. Instead of a linear relationship between response (output) and parameters (input) in fixed and finite dimensions, our measurement of response is determined by the discretization of an integral functional of distributed parameters. The unknown quantity is a function that belongs to an infinite-dimensional space, which is approximated by discretization on increasingly fine meshes. Here, we aim to understand the asymptotics of the rounding procedure in the limit of the mesh size going to zero. As a result, the inverse Fisher information matrix we try to minimize (with respect to a given design criterion, such as its trace) increases in size with the number of discretization points, which makes analysis with common design criteria difficult (Sect. 2.5). We are not aware of prior theoretical work on the convergence analysis of discretized design of experiments with a number of sites that can grow unboundedly. Moreover, we assume here—as would be the case in many physical settings—that each data site is measured only once, so repeated measurements (as in [8,25]) are not allowed. This would be the case, for example, if the problem is time dependent and thus a certain point in space cannot be revisited at the same instant in time or if the sensor error is constant in time but has mean zero over the sensor population, as is typical of physical sensors [17, §34.3].

Since we aim to determine the optimal sensor locations starting from a relaxed problem, the construction of an integer solution with appropriate rounding strategies of the relaxed version is a critical endeavor. Numerous rounding heuristics are given in the literature (see [6,22,23]), and some specifically aim for binary variables (see

[4,15,32]). In [6], the author studied the optimal rounding by recording and comparing empirical success rates, defined as the percentage of "roundable relaxation" optima (in the words of [6]), for different types of optimization problems (mixed-integer quadratically constrained program, mixed-integer nonlinear program, etc.) among the existing rounding strategies. Classical mixed-integer techniques have been used specifically for sensor placement aiming at detecting contamination in water networks (see [5,21,29]) but focusing mainly on a fixed-sized discretization without investigation of limiting properties. Closer to the continuously indexed (in the limit) framework in this paper, sensor placement for systems governed by partial differential equations has been studied using a Bayesian framework [1]. In that case, the discrete nature of sensor placement problems was recovered by seeking sparsity in the solution of the relaxed problems by means of an l_0 penalty that is approximated by a sequence of smooth functions. This approach can be applied to infinite-dimensional problems, but the numerical results can be unstable, and they depend on the choice of various tuning parameters. All the rounding approaches described in this paragraph have shown good performance for certain classes of problems, including the type studied here, but their asymptotic properties have not been investigated theoretically.

Since we are interested in problems that can be continuously indexed, we investigate an extension of *sum-up rounding* (SUR), a recently proposed technique that was first used in the context of continuous-time mixed-integer optimal control problems (MIOCPs) [27]. Sum-up rounding for binary variables, as we also pursue here, has been shown in temporally indexed problems to have the desirable asymptotic property of being arbitrarily close to an integer solution as long as the discretization mesh is sufficiently fine [26,27]. In [27], the authors not only clarify the role of SUR in MIOCPs but also obtain a guaranteed bound on the performance loss, depending on the size of discretization mesh. In [26], a specific structure in one dimension is considered where the objective is a function of either the Fisher information matrix or its inverse, and the optimality gap converges to zero. Recently we used SUR as a heuristic for the sensor placement problem in natural gas pipelines governed by systems of nonlinear hyperbolic differential equations. We observed convergence of the integrality gap as the spatial mesh was progressively refined [33]; but since the spatial problem had a different nature from [27], we did not have theory to justify that observation. That was one of the main motivators for this work.

Here, we investigate DOE based on a Bayesian framework for parameter estimation [1], and we minimize functions of the posterior covariance matrix based on common experimental design criteria [25]. Our parameter to the observations map is based on an integral equation, as opposed to the solution of a partial differential equation as in [1], although the two are conceptually equivalent if one considers the Green function resolvent with the prior interpreted as a regularization term [11]. The resulting DOE problem after spatial discretization is a convex mixed-integer program; see Sect. 2.5. After solving the relaxed problem, we define and employ a multidimensional SUR procedure inspired by the one-dimensional procedure proposed in [26,27]. Our main objective is to investigate whether the integrality gap between the DOE criteria at the rounded solution and relaxed solution converges to zero in the limit of zero mesh size, as was observed for MIOCPs in [26,27]. Our contributions consist of proposing an extension of the SUR rounding procedure in multiple dimensions and proving that, for common experimental design criteria, the integrality gap converges to zero as the mesh size shrinks to zero. The techniques we employ to this end are related to the spectral theory of self-adjoint integral operators [3]. We emphasize that questions about the asymptotic quality of DOE solutions over varying design space size have not been investigated in classical DOE theory [25].

While inspired from the idea of SUR in [27] and using it as a building block, this work is different in several respects. First, applying it in a multidimensional setting allows for a larger number of rounding options and our theory covers a fairly general setup based on what we call compatible two-level domain decomposition schemes. Also, while the SUR technique itself works for rectangular domains, (which in effect, we argue in the construction at the end of Sect. 3.2), the proof in [27] relies on the convergence of one-dimensional integrals which would not directly apply to more than one dimension. While in the end, for implementation simplicity, our examples are for rectangular domains as well, the theoretical framework itself allows in principle a broad set of domain shapes and other rounding techniques, another example of which we give in Sect. A.1. Second, the functions we optimize here, which define the objective of the experimental design, depend on the posterior covariance matrix, whereas the entries in the precision matrix (the inverse of the covariance matrix) are the ones related to an integral quantity for which the typical SUR analysis applies. To carry out the gap convergence analysis for experimental design requires the investigation of SUR effects on the eigenvalues of the precision and covariance matrices. Moreover, the sizes of these matrices go to infinity, which poses additional obstacles to the convergence analysis as we discuss in Sect. 4, whereas results in [26] primarily address a fixed dimensional parameter space, and thus, covariance matrix.

The paper is organized as follows. In Sect. 2, we define the parameter-to-observable map, we quantify the estimation error using the posterior covariance matrix in the Bayesian framework, we formulate the original mixed-integer nonlinear program and the relaxed problems, and we make a connection to integral operators. In Sect. 3, we define a SUR procedure based on a two-level meshing framework, and we prove the SUR approximation properties in multiple dimensions. In Sect. 4, we show convergence of the integrality gap based on SUR for different experimental design criteria. In Sect. 5, we give simulation results on two-dimensional gravity surveying and compare them with thresholding designs. In Sect. 6, we draw conclusions, discuss limitations, and propose future work for our approach.

2 Estimation framework

While the contribution of this work concerns primarily the behavior of the SUR-induced integrality gap, some of the assumptions we make stem from the estimation framework itself. In particular, our results are tied to a common but specific choice of the covariance matrices as well as to a limiting interpretation in terms of a certain integral operator. In the latter case, the integer programming relaxation needs to be interpreted in an extended output space. We thus describe the estimation framework that we use to define our DOE problem. The setup is based primarily on [1].

2.1 Parameter-to-observable map

Consider the input domain $\Omega_{in} \subset \mathbb{R}^Q$ and output domain $\Omega_{out} \subset \mathbb{R}^P$, both of which are compact sets. Suppose the output without measurement error depends on the input through an integral equation:

$$u(x) = \int_{\Omega_{in}} f(x, y) u_0(y)\, \mathrm{d}y, \quad x \in \Omega_{out}, \tag{1}$$

where $f(x, y)$ is prescribed by the physical constraints in the setup; we thus assume it is known. The output $u(x)$ can be measured at selected points but is affected by measurement error. Our goal is to infer the parameter vector u_0 from the observation vector u. Equation (1) defines a parameter-to-observable map.

To create a finite-dimensional approximation we now discuss a simple discretization strategy. More advanced discretization approaches as in [19] could easily be incorporated but would complicate the presentation whose focus is on the SUR approximation properties for DOE. We divide $D = \Omega_{in}$ (or an approximation of Ω_{in}) into m subdomains $D_1, D_2, \ldots, D_m$ with equal size $\mu(D_i) = \Delta_y = \mu(\Omega_{in})/m$ for $i = 1, 2, \ldots, m$ (as is done, e.g., for versions of Nyström's method in [28]). Then, we select a representation point y_i in each D_i and represent the input function u_0 as the finite-dimensional vector $\hat{u}_0 = \big(u_0(y_1), u_0(y_2), \ldots, u_0(y_m)\big)$. Similarly we divide $V = \Omega_{out}$ into n subdomains $V_1, V_2, \ldots, V_n$ with equal size $\mu(V_j) = \Delta_x = \mu(\Omega_{out})/n$ for $j = 1, 2, \ldots, n$ and select a representation point x_j for each V_j. Then we represent the continuous output u as the vector $\hat{u} = \big(u(x_1), u(x_2), \ldots, u(x_n)\big)$. These $x_1, x_2, \ldots, x_n$ points are also the candidate locations to place sensors. We approximate the integral from (1) by the Riemann sum:

$$u(x_j) = \int_{\Omega_{in}} f(x_j, y) u_0(y)\, \mathrm{d}y \approx \sum_i f(x_j, y_i) u_0(y_i) \Delta_y.$$

To write it in matrix form, we define $F \in \mathbb{R}^{n \times m}$ with $F(j, i) = f(x_j, y_i)\Delta_y$, and then $\hat{u} = F\hat{u}_0$.

We note that in applications the function $f(x, y)$ in (1) may not always be continuous. For example, when the function f encapsulates wave dynamics, it is represented by a Dirac functional $f\big((x, t), y\big) \equiv \delta(y, x - at)$, where a is the wave speed. For the remainder of this work, we assume f to be continuous. Another restriction in (1) is that $u(x)$ depends linearly on $u_0(x)$, which is not the case in nonlinear relationships, such as for pipeline gas dynamics [33]. In that case, the target problem can be approximated in the framework of (1) by linearization, as was done in [1,33].

In the rest of this work, we use $\delta(x)$ to denote the Kronecker δ symbol:

$$\delta(x) = \begin{cases} 1, & \text{if } x = 0, \\ 0, & \text{otherwise.} \end{cases}$$

2.2 Bayesian estimation framework

Our goal is to estimate the parameter vector $\hat{u}_0$ as a proxy for the unknown function u_0. We consider a Bayesian framework where $\hat{u}_0$ is the parameter vector to be estimated and the measurements $\hat{u}$ are data perturbed by noise. Similar to [1,33], we assume that both the parameter prior and the measurements distributions are Gaussian:

$$\begin{aligned}&\hat{u}_0 \sim N(u_{pri}, \Gamma_{pri}),\\ &\hat{u} = F\hat{u}_0 + \eta, \text{ where } \eta \sim N(0, \Gamma_{noise}).\end{aligned}$$

Here, Γ_{pri} and Γ_{noise} represent the prior and measurement noise covariance matrices, respectively, whereas u_{pri} is the prior mean. We assume the measurement error to be unbiased conditional on the realization of u_0, and thus η has mean 0. From Bayes' rule, the posterior distribution of $\hat{u}_0$ is also Gaussian and has (up to a constant) the following density:

$$\begin{aligned}\pi_{post}(\hat{u}_0|\hat{u}) \propto \exp\Bigg\{&-\frac{1}{2}(\hat{u} - F\hat{u}_0)^T \Gamma_{noise}^{-1}(\hat{u} - F\hat{u}_0)\\ &-\frac{1}{2}(\hat{u}_0 - u_{pri})^T \Gamma_{pri}^{-1}(\hat{u}_0 - u_{pri})\Bigg\}.\end{aligned}$$

We now quantify the sensor placement effect in the posterior. We achieve this by creating a weight vector $w = (w_1, w_2, \ldots, w_n) \in \{0, 1\}^n$ where the jth component w_j corresponds to candidate location x_j in the output domain. A sensor is placed at location x_j if $w_j = 1$ and is not placed if $w_j = 0$, so there is a one-to-one mapping between sensor placement and weight vectors. Let W be a diagonal matrix with weight vector w on its diagonal. The w-weighted posterior likelihood, conditional on the data u and weight vector w, is

$$\begin{aligned}\pi_{post}(\hat{u}_0|\hat{u}, w) \propto \exp\Big\{&-\tfrac{1}{2}(\hat{u} - F\hat{u}_0)^T W^{1/2}\Gamma_{noise}^{-1} W^{1/2}(\hat{u} - F\hat{u}_0)\\ &-\tfrac{1}{2}(\hat{u}_0 - u_{pri})^T \Gamma_{pri}^{-1}(\hat{u}_0 - u_{pri})\Big\}.\end{aligned}$$

One can immediately verify that for any integer-valued vector w, the posterior distribution is exactly the one for Bayesian least squares with data measured for indices of $u(x)$ where $w_i = 1$, in (1) for $i = 1, 2, \ldots, n$.

Under these assumptions and accounting for the prior distribution, we can compute the posterior $\hat{u}_0$, which is the normal distribution $N(u_{post}, \Gamma_{post})$, where

$$u_{post} = \Gamma_{post}\left(F^T\Gamma_{noise}^{-1}\hat{u} + \Gamma^{-1}u_{pri}\right), \quad \Gamma_{post} = \left(F^T W^{1/2}\Gamma_{noise}^{-1} W^{1/2} F + \Gamma_{pri}^{-1}\right)^{-1}$$

are the posterior mean and covariance matrix, respectively. We point out that in this estimation model the posterior covariance matrix does not depend on data $\hat{u}$. In other words, the optimal sensor placement is determined by the parameter-to-observable map and two Γ matrices.

2.3 Choice of covariance matrices

We assume that, conditional on the true $\hat{u}$, the measurement errors are independent. In most physical processes and sensor systems this is a reasonable assumption [10]. Consequently, the matrix Γ_{noise} is diagonal and commutes with W and all its positive powers, resulting in the expression

$$u_{post} = \Gamma_{post}\left(F^T \Gamma_{noise}^{-1}\hat{u} + \Gamma^{-1}u_{pri}\right),\ \Gamma_{post} = \left(F^T \Gamma_{noise}^{-1} W F + \Gamma_{pri}^{-1}\right)^{-1}.$$

In particular, the precision matrix (the inverse of the covariance matrix) becomes *linear* in W, which considerably simplifies our calculations and analysis. We assume identical sensors, and therefore $\Gamma_{noise} = \sigma_{noise} I_n$ for some prescribed sensor noise standard deviation σ_{noise}. The other covariance matrix that needs to be selected is the one corresponding to the prior distribution. Here we use a multiple of the identity $\Gamma_{pri} = \sigma_{pri} I_m$. This choice can be interpreted as ridge regression [13] or Tikhonov regularization of an inverse problem [18]. While for some setups our choice is not the ideal prior [18] it is one of the most common choices, at least before significant collection of data.

Our analysis is tied significantly to these choices, and particularly so for the prior where other reasonable choices may be available. On the other hand, this is one of the most common choices in statistical analysis of inverse problems [18]; therefore our setup does represent many problems of interest.

2.4 Connection to integral operators

With the covariance choices specified in Sect. 2.3, the precision matrix, the inverse of the posterior matrix Γ_{post}, becomes

$$\Gamma_{post}^{-1} = \sigma_{noise}^{-1} F^T W F + \sigma_{pri}^{-1} I_m.$$

Note that the (i, j)th entry in Γ_{post}^{-1} is

$$\Gamma_{post}^{-1}(i, j) = (\Delta_y)^2 \sigma_{noise}^{-1} \sum_{k=1}^{n} f(x_k, y_i) w_k^n f(x_k, y_j) + \sigma_{pri}^{-1} \cdot \delta(x_i - x_j), \qquad (2)$$

with w_k^n being the weights from the diagonal of W. With reference to the notations from Sect. 2.1, we denote by $w^n(x)$ the piecewise constant function defined as $w^n(x) = w_k^n$, $x \in D_k$, which is the discretized area corresponding to kth candidate location in Ω_{in}. Assume that there is a measurable function $w(x) : \Omega_{out} \to [0, 1]$ such that $w^n(x) \to w(x)$ in L^1. For purposes of illustration we assume that $w^n(x)$ converges in this subsection; that will not be required in our results in Sect. 4. Then, if $\Delta_x, \Delta_y \to 0$ with Δ_y/Δ_x constant, the first term in (2) will converge to

$$\Delta_y \left(\frac{\Delta_y}{\Delta_x}\right) \sigma_{noise}^{-1} \int_{\Omega_{out}} f(x, y_i) w(x) f(x, y_j)\, \mathrm{d}x. \tag{3}$$

This quantity relates to the discretization of an integral operator

$$\mathcal{L}u_0(z) = \left(\frac{\Delta_y}{\Delta_x}\right) \sigma_{noise}^{-1} \iint_{\Omega_{out} \times \Omega_{in}} f(x, z) w(x) f(x, s) u_0(s)\, \mathrm{d}x \mathrm{d}s, \quad z \in \Omega_{out}. \tag{4}$$

Note that if $\Delta_x = \Delta_y$, then (3) is one coefficient of the discretization of (4) along the input variable s. If $w(x)$ is nonnegative, then the eigenvalues of $\mathcal{L}$ are nonnegative. Because $\mathcal{L}$ is a compact operator [3], it has a countable spectrum with 0 its only accumulation point. Moreover, because of its integral form, its trace is finite [31]. This prompts the hypothesis that the spectrum of Γ_{post}^{-1} is related to the spectrum of $\mathcal{L}$ and σ_{pri}. Specifically, eigenvalues of $\sigma_{noise}^{-1} F^T W F$ approach eigenvalues of $\mathcal{L}$ [31] in the limit of Δ_x, Δ_y going to 0 at a fixed ratio. This indicates that the eigenvalues of Γ_{post} will approximately be $1/(\lambda + \sigma_{pri}^{-1})$, where λ are eigenvalues of $\mathcal{L}$. This insight, with mathematical statements that will be made more rigorous in Sect. 4, allows the analysis of optimization problems whose objectives are functions of the spectrum of Γ_{post}, as is the case for the DOE problems described in Sect. 2.5.

2.5 Design of experiments problems

We are ready to formulate our DOE problem that addresses the issue of optimal sensor placement. We aim to minimize the estimation error of the parameter $\hat{u}_0$, which can be quantified by using its posterior covariance matrix, $\phi(\Gamma_{post})$. The three most widely used criteria in experimental design to measure the size of this error are [25]

- A-optimal design: $\phi(\Gamma_{post}) = tr(\Gamma_{post})$;
- D-optimal design: $\phi(\Gamma_{post}) = det(\Gamma_{post})$;
- E-optimal design: $\phi(\Gamma_{post}) = \lambda_{max}(\Gamma_{post})$.

Lemma 1 *$tr(\Gamma_{post})$, $\log\det(\Gamma_{post})$ and $\lambda_{max}(\Gamma_{post})$ are convex functions in the weight vector w.*

Proof The posterior matrix can be written as

$$\Gamma_{post}(w) = \left(\sigma_{noise}^{-1} \sum_{i=1}^{n} w_i F_i F_i^T + \sigma_{pri}^{-1} I_m\right)^{-1},$$

where F_i is the ith column of F^T. The desired results follow because $tr(X^{-1})$, $\log\det(X^{-1})$ and $\lambda_{max}(X^{-1})$ are all convex in X [8, Exercise 3.26], and the fact that X is affine in w. □

We formulate the DOE problem as follows (ϕ represents one of the three criteria, and we use $logdet$ for D-optimal design):

$$\begin{aligned} &\min\ \phi(\Gamma_{post}(w)) \\ &\text{ s.t. } w_i \in \{0,1\},\ \textstyle\sum_{i=1}^{n} w_i = n_0, \end{aligned} \tag{5}$$

where n_0 is the number of sensors. To avoid the complexity of integer programming, we start by examining the relaxed problem obtained by relaxing the integer constraint,

$$\begin{aligned} &\min\ \phi(\Gamma_{post}(w)) \\ &\text{ s.t. } 0 \leq w_i \leq 1, i = 1, 2, \ldots, n,\quad \textstyle\sum_{i=1}^{n} w_i = n_0, \end{aligned} \tag{6}$$

whose solution we denote by w_{rel}. Problem (6) is convex from Lemma 1. It can be solved, after using some standard semidefinite programming reformulations, by interior-point algorithms [8]. The relaxed solution w_{rel} provides a lower bound to the optimal objective of the convex integer program (5).

Our results will apply for any n_0 (and its value could also change with the number of discretization domains n), but they would be most meaningful in certain ranges. An examination of (2) indicates that if f is bounded by C, then the trace of the discretization of the integral operator is nonnegative and upper bounded by $n_0 n C^2 \Delta_y^2$. We must have $n\Delta y = O(1)$ since $n\Delta y$ must be the volume of the initial set V. Therefore, for the estimation problem to carry information comparable to the prior, we need to have $n_0 \Delta_y = O(1)$; that is, n_0 must be of comparable order with n. Otherwise the contribution from ϕ would originate in the limit exclusively from the prior. In other words, a meaningful asymptotics is the one where the number of sensors is in a fixed ratio with the number of mesh domains. This is the corresponding constraint to the one in [26] whereby the measurement time is proportional to the considered time range $[0, T]$.

3 A sum-up rounding procedure

In this section we describe a sum-up rounding procedure that maps the fractional vector w_{rel} solution of (6) into an integer vector w_{SUR} in a way that ensures the spectrum of $\Gamma_{post}(w_{rel})$ and $\Gamma_{post}(w_{SUR})$ are not too far from each other. In turn, this will ensure that the gap $\phi(\Gamma_{post}(w_{SUR})) - \phi(\Gamma_{post}(w_{rel}))$ stays small.

Our procedure is presented here for rectangular domains V (i.e., Ω_{out}, but the same construction can be applied to Ω_{in}), divided into n subdomains $V_1, V_2, \ldots, V_n$ of equal size $\mu(V_k) = \Delta_x = \frac{\mu(V)}{n}$. Given the function $w^n(x) : V \to [0, 1]$, which is constant on each V_i, we construct a 0–1 valued function $\tilde{w}^n(x)$ that is also constant on each V_i such that the two sums

$$S_1^n = \sum_{k=1}^{n} f(x_k) w^n(x_k) \Delta_x, \qquad S_2^n = \sum_{k=1}^{n} f(x_k) \tilde{w}^n(x_k) \Delta_x \tag{7}$$

are arbitrarily close to each other as long as n is large enough. Our analysis is centered around estimating the variation in the entry i, j of Γ_{post}^{-1} following the SUR procedure. The bounding technique will end up being uniform in i, j. To simplify our exposition, we ignore in the rest of the analysis the argument y of f in (3) since it has no effect on our approach.

Note that the function f need not be the same as the one defining the integral Eq. (1), and it can be any function defined on Ω_{out} satisfying certain continuity conditions. If $V \subset R$, this is essentially a one-dimensional time domain problem that has already been studied in [27]. In multiple dimensions, we can flatten the multidimensional vector and apply the basic sum-up rounding. However, the integration-by-part technique in the proof of [27, Theorem 2] becomes problematic in multiple dimensions, and this is why we resort to a two-level decomposition which also covers the basic one-dimensional case. It is worth mentioning that depending on the ordering of entries, we can obtain different integer vectors. In this section, we discuss the basic sum-up rounding strategy in Sect. 3.1 where Lemma 2 is an analogue to [27, Theorem 3]. The multidimensional strategy and its properties are given in Sects. 3.2 and 3.3 respectively, and Theorem 2 in Sect. 3.3 is an extension of [27, Theorem 2].

3.1 Basic sum-up rounding strategy

We denote $\tilde{w}_i^n$ (w_1^n) as the value of $\tilde{w}^n(x)$ ($w^n(x)$) in V_i and construct the binary function $\tilde{w}^n(x)$ from $w^n(x)$ as follows.

(1) Compute $I_1 = w_1^n \cdot \mu(V_1)$, and set $\tilde{w}_1^n$ to

$$\tilde{w}_1^n = \begin{cases} 0, & \text{if } I_1 \le \frac{1}{2}\mu(V_1), \\ 1, & \text{otherwise.} \end{cases}$$

(2) For $i = 2, 3, \ldots, n$, compute

$$I_i = \sum_{k=1}^{i} w^n(x_k)\mu(V_k) \quad \text{and} \quad \tilde{I}_{i-1} = \sum_{k=1}^{i-1} \tilde{w}^n(x_k)\mu(V_k),$$

where $\tilde{w}_i^n$ is given by

$$\tilde{w}_i^n = \begin{cases} 0, & \text{if } I_i - \tilde{I}_{i-1} \le \frac{1}{2}\mu(V_i), \\ 1, & \text{otherwise.} \end{cases}$$

We call this strategy basic *sum-up rounding*, in reference to the name of the one-dimensional technique introduced in [26,27] which inspired this approach. The basic idea is that each element is scanned sequentially and is rounded to either 0 or 1 determined by the difference in the accumulated sum of elements that are already processed. The strategy has the property that for large n, $w^n(x)$ and $\tilde{w}^n(x)$ get close to each other for all partial sums, which is stated in the following lemma.

Lemma 2 *The function $\tilde{w}^n(x)$ has the following property: For any $i = 1, 2, \ldots, n$,*

$$|I_i - \tilde{I}_i| = \left| \sum_{k=1}^{i} \left(w^n(x_k) - \tilde{w}^n(x_k) \right) \Delta_x \right| \leq \frac{1}{2n}\mu(V),$$

where V is the rectangular output domain with fixed size.

Proof We prove this result by induction. For $i = 1$, we have the following.

– When $I_1 \leq \frac{1}{2}\mu(V_1) = \frac{1}{2n}\mu(V)$, we have $\tilde{w}_1^n = 0$ and $\tilde{I}_1 = 0$, and therefore

$$|I_1 - \tilde{I}_1| = I_1 \leq \frac{1}{2n}\mu(V).$$

– When $I_1 > \frac{1}{2n}\mu(V)$, we have $\tilde{w}_1^n = 1$. Since $w^n(x) \leq 1$, we get

$$\frac{1}{2n}\mu(V) < I_1 \leq \frac{1}{n}\mu(V), \quad |I_1 - \tilde{I}_1| = \frac{1}{n}\mu(V) - I_1 \leq \frac{1}{2n}\mu(V).$$

By the induction hypothesis, assume $|I_i - \tilde{I}_i| \leq \frac{1}{2n}\mu(V)$ is true for $i = k$. We show it for $i = k+1$ as follows.

– When $0 \leq I_k - \tilde{I}_k \leq \frac{1}{2n}\mu(V)$, note that $I_k \leq I_{k+1}$. We discuss two cases.

(a) If $0 \leq I_{k+1} - \tilde{I}_k \leq \frac{1}{2n}\mu(V)$, then $\tilde{w}_{k+1}^n = 0$ from the rounding rule, and thus $\tilde{I}_{k+1} = \tilde{I}_k$. Therefore

$$0 \leq I_{k+1} - \tilde{I}_{k+1} \leq \frac{1}{2n}\mu(V),$$

and the induction hypothesis is satisfied.

(b) If $\frac{1}{2n}\mu(V) < I_{k+1} - \tilde{I}_k$, which implies

$$\frac{1}{2n}\mu(V) < I_{k+1} - \tilde{I}_k \leq I_k - \tilde{I}_k + \frac{w_{k+1}^n}{n}\mu(V) \leq \frac{1}{2n}\mu(V) + \frac{1}{n}\mu(V), \quad (8)$$

then from the rounding rule we have that $\tilde{w}_{k+1}^n = 1$, and we obtain

$$\tilde{I}_{k+1} - \tilde{I}_k = \mu(V_{i+1}) = \frac{1}{n}\mu(V). \quad (9)$$

Subtracting the equality (9) from the inequality (8) gives the desired result:

$$-\frac{1}{2n}\mu(V) < I_{k+1} - \tilde{I}_{k+1} = I_{k+1} - \tilde{I}_k - \frac{1}{n}\mu(V) \leq \frac{1}{2n}\mu(V).$$

- When $-\frac{1}{2n}\mu(V) \le I_k - \tilde{I}_k \le 0$, since $I_{k+1} = I_k + w^n_{k+1}\frac{\mu(V)}{n}$, we also have that $I_{k+1} \ge I_k$, and thus $-\frac{1}{2n}\mu(V) \le I_{k+1} - \tilde{I}_k$. We discuss two cases in a similar way.
  (a) If $-\frac{1}{2n}\mu(V) \le I_{k+1} - \tilde{I}_k \le \frac{1}{2n}\mu(V)$, then $\tilde{w}^n_{k+1} = 0$ from the rounding rule, and thus $\tilde{I}_{k+1} = \tilde{I}_k$. Hence

$$-\frac{1}{2n}\mu(V) \le I_{k+1} - \tilde{I}_{k+1} \le \frac{1}{2n}\mu(V).$$

  (b) If $\frac{1}{2n}\mu(V) < I_{k+1} - \tilde{I}_k$, then

$$\frac{1}{2n}\mu(V) < I_{k+1} - \tilde{I}_k \le I_k - \tilde{I}_k + \frac{w^n_{k+1}}{n}\mu(V) \le 0 + \frac{1}{n}\mu(V). \qquad (10)$$

  In turn, from the rounding rule this implies that $\tilde{w}^n_{k+1} = 1$. As a result, we have

$$\tilde{I}_{k+1} = \tilde{I}_k + \mu(V_{i+1}) = \tilde{I}_k + \frac{1}{n}\mu(V). \qquad (11)$$

  Replacing the identity (11) in the inequality (10), we obtain

$$-\frac{1}{2n}\mu(V) < I_{k+1} - \tilde{I}_{k+1} = I_{k+1} - \tilde{I}_k - \frac{1}{n}\mu(V) \le 0.$$

Inspecting the consequences of these four branches, we have completed the proof for $i = k+1$, namely, $|I_{k+1} - \tilde{I}_{k+1}| \le \frac{1}{2n}\mu(V)$. Therefore the statement is true for $i = 1, 2, \ldots, n$ and the proof is complete. □

We now have a rounding strategy, and before we apply it, it is important to check feasibility of the resulting integer vector. The lemma below states that sum-up rounding always provides a feasible vector for our main optimization problem (5).

Lemma 3 *With the basic sum-up rounding strategy, if* $\sum_{k=1}^n w^n(x_k) = n_0$ *is an integer, then*

$$\sum_{k=1}^n \tilde{w}^n(x_k) = \sum_{k=1}^n w^n(x_k) = n_0.$$

Proof In Lemma 2 we have that $\Delta_x = \frac{\mu(V)}{n}$, and the conclusion for $i = n$ can be rewritten as

$$\left|\sum_{k=1}^n w^n(x_k) - \sum_{k=1}^n \tilde{w}^n(x_k)\right| \le \frac{1}{2}.$$

Since both $\sum_{k=1}^n w^n(x_k)$ and $\sum_{k=1}^n \tilde{w}^n(x_k)$ are integers, they have to be equal. □

3.2 Sum-up rounding algorithms

We showed in Sect. 3.1 that $w^n(x)$ and $\tilde{w}^n(x)$ are close to each other, but our goal is to prove that the two sums in (7) are close. Suppose $V = [l_1^1, l_2^1] \times [l_1^2, l_2^2] \times \cdots \times [l_1^P, l_2^P] \subset \mathbb{R}^P$, and each $[l_1^i, l_2^i]$ is divided into n_i intervals $\mathcal{I}_{i,1}, \mathcal{I}_{i,2}, \ldots, \mathcal{I}_{i,n_i}$ (script letters represent one-dimensional intervals) of equal length. Then there are $n = n_1 n_2 \cdots n_P$ unit rectangles of the form

$$\prod_{\substack{i=1,2,\ldots,P,\\ j_i \in \{1,2,\ldots,n_i\}}} \mathcal{I}_{i,j_i}.$$

They all have the same size $\mu(V)/n$, and we call them $R_1, R_2, \ldots, R_n$. In addition, we assume that there exist two positive constants c_1, c_2 such that

$$c_1 \leq \frac{\max_{i=1,2,\ldots,P} n_i}{\min_{i=1,2,\ldots,P} n_i} \leq c_2. \tag{12}$$

This implies that $n_i = \mathcal{O}(n^{1/P})$ for any $i \in \{1, 2, \ldots, P\}$ and that each rectangle R_i is not far from a "unit box."

Definition 1 We call a *compatible two-level decomposition scheme* a domain decomposition setup of a compact domain V with the following properties. The rectangles R_i, $i = 1, 2, \ldots, n$, are grouped in subdomains V_j, $j = 1, 2, \ldots, \tilde{k}(n)$, for which the first $k(n)$ subdomains contain an equal number of rectangles, $r(n)$. The intersections between the interiors of each two subdomains V_j is empty, moreover the subdomains V_j need not cover the entire domain V, and we denote the remainder by $V_{rem} = V - \cup_{j=1}^{\tilde{k}(n)} V_j$. We denote by $\rho(V_j)$ the diameter of the subdomains, $j = 1, 2, \ldots, k(n)$. Subsequently, we *reindex* the rectangles such that their ordering respects the subdomains ordering, that is, $R_{i_1} \in V_{j_1}, R_{i_2} \in V_{j_2}, j_1 \leq j_2 \Rightarrow i_1 \leq i_2$. Our *sum-up rounding approach* consists of applying the basic method from Sect. 3.1 to the rectangles R_i in their modified ordering.

To obtain the approximation properties, it would be sufficient to apply the basic method from Sect. 3.1 to each subdomain V_i. The extra steps of reordering and the application to the entire rectangle set ensure that we preserve the total sum of the weights, and thus that we satisfy the constraints from (5).

To achieve a vanishing integrality gap, we will be interested in compatible two-level decompositions that satisfy in the limit the following properties:

$$\lim_{n\to\infty} \max_{1\leq j\leq k(n)} \rho(V_j) = 0, \quad k(n), r(n) \overset{n\to\infty}{\longrightarrow} \infty, \quad \frac{r(n)k(n)}{n} \overset{n\to\infty}{\longrightarrow} 1, \quad \mu(V_{rem}) \to 0. \tag{13}$$

For many domains V such compatible two-level decompositions can be easily obtained based on algorithms for hexahedral meshing [34] that are commonly used in spectral element methods [24]. Note that our problem is easier than most in that sense, since the mesh need not be conformal [7], that is, we allow $V_{rem} \neq \emptyset$. Even in that case, however, a rigorous proof of (13) for a wide class of domains is non-trivial and significantly beyond the scope of the paper. The theoretical existence of such decompositions, however, seems clear as similar techniques are central to Riemann sums convergence arguments.

We thus demonstrate how to create compatible two-level decompositions for rectangular domains only, as follows.

(i) We divide V into $n = n_1 n_2 \cdots n_P$ small rectangles of the form (3.2) as before, and we list them as $R_1, R_2, \ldots, R_n$.

(ii) We order the unit rectangles $R_1, R_2, \ldots, R_n$, as follows:

$$\begin{aligned}
R_1 &= \mathcal{I}_{1,1} \times \mathcal{I}_{2,1} \times \cdots \times \mathcal{I}_{P,1}\\
R_2 &= \mathcal{I}_{1,2} \times \mathcal{I}_{2,1} \times \cdots \times \mathcal{I}_{P,1}\\
&\vdots\\
R_{n_1} &= \mathcal{I}_{1,n_1} \times \mathcal{I}_{2,1} \times \cdots \times \mathcal{I}_{P,1}\\
R_{n_1+1} &= \mathcal{I}_{1,1} \times \mathcal{I}_{2,2} \times \cdots \times \mathcal{I}_{P,1}\\
R_{n_1+2} &= \mathcal{I}_{1,2} \times \mathcal{I}_{2,2} \times \cdots \times \mathcal{I}_{P,1}\\
&\vdots\\
R_n &= \mathcal{I}_{1,n_1} \times \mathcal{I}_{2,n_2} \times \cdots \times \mathcal{I}_{P,n_P}.
\end{aligned}$$

They are ordered "line by line" according to the first dimension. Denoting $k_1(n_1) \overset{\Delta}{=} \lfloor\sqrt{n_1}\rfloor$, we now build the subdomains V_j as follows.

(a) On $[l_1^1, l_2^1]$ we group the first $k_1(n_1)$ intervals $\{\mathcal{I}_{i,j}\}_{j=1}^{k_1(n_1)}$ as $\mathcal{G}_{1,1}$, group the next $k_1(n_1)$ intervals $\{\mathcal{I}_{1,j}\}_{k_1(n_1)+1}^{2k_1(n_1)}$ as $\mathcal{G}_{1,2}$, and so forth until we get $\mathcal{G}_{1,k_1(n_1)}$. The remaining intervals $\{\mathcal{I}_{1,j}\}_{j=k_1(n_1)^2+1}^{n}$ are grouped as $\mathcal{G}_{1,last}$, and the number of intervals in $\mathcal{G}_{1,last}$ equals $n_1 - \lfloor\sqrt{n_1}\rfloor^2$.

(b) The subdomain V_j has the following form:

$$\mathcal{G}_{1,j_1} \times \mathcal{I}_{2,j_2} \times .. \times \mathcal{I}_{P,j_P},$$

where $j_1 \in \{1, 2, \ldots, k_1(n_1), last\}$, $j_i \in \{1, 2, \ldots, n_i\}$ for $i \geq 2$.

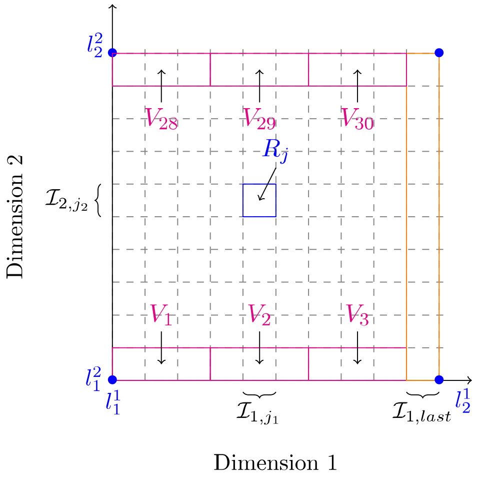
This decomposition has the following parameters and properties, in reference to Definition 1.

$$k(n) = \lfloor\sqrt{n_1}\rfloor \prod_{i=2}^{P} n_i,\ \ \tilde{k}(n) = \lceil\sqrt{n_1}\rceil \prod_{i=2}^{P} n_i,\ \ r(n) = \lfloor\sqrt{n_1}\rfloor \tag{14}$$

$$\rho(V_j) = \sqrt{\left(\frac{(l_2^1 - l_1^1)}{\lfloor\sqrt{n_1}\rfloor}\right)^2 + \sum_{i=2}^{P}\left(\frac{(l_2^i - l_1^i)}{n_i}\right)^2},\quad j = 1, 2, \ldots, k(n) \tag{15}$$

With these definitions, sum-up rounding is applied as described in Definition 1. We note that many other compatible two-level decompositions are possible, another one is presented in Sect. A.1.

The following simple example illustrates the idea of two-level decomposition on a square domain in $\mathbb{R}^2$. There are 10 unit intervals evenly spaced on each side ($n_1 = n_2 = 10$), and the number of unit rectangles R_j is $n = 100$; we group $\lfloor\sqrt{10}\rfloor$, which is 3, unit intervals $\mathcal{I}_{1,j_1}$ as $\mathcal{G}_{1,j_1}$ on Dimension 1, and then form 30 subdomains V_j; the basic sum-up rounding strategy is applied to each V_j. As the construction is repeated for increasing n, the remainder area (yellow in color rendering) will diminish compared to the full domain since $1 - \frac{\lfloor\sqrt{n_1}\rfloor^2}{n_1} \to 0$, and its effect on the difference between the sums in (7) and the corresponding integral will vanish.



- domain: $[l_1^1, l_2^1] \times [l_1^2, l_2^2]$
- discretization parameter: $n_1 = n_2 = 10$, $n = n_1 n_2 = 100$
- $k_1(n_1) = \lfloor\sqrt{10}\rfloor = 3$, $k(n) = 30, \tilde{k}(n) = 40$
- $R_j = \mathcal{I}_{1,j_1} \times \mathcal{I}_{2,j_2}$ where $j = (j_2 - 1)n_1 + j_1$
- $\mathcal{G}_{1,1} = \mathcal{I}_{1,1} \cup \mathcal{I}_{1,2} \cup \mathcal{I}_{1,3}$
  $\mathcal{G}_{1,2} = \mathcal{I}_{1,4} \cup \mathcal{I}_{1,5} \cup \mathcal{I}_{1,6}$
  $\mathcal{G}_{1,3} = \mathcal{I}_{1,7} \cup \mathcal{I}_{1,8} \cup \mathcal{I}_{1,9}$
  $\mathcal{G}_{1,last} = \mathcal{I}_{1,last}$
- subdomain:
  $V_1 = \mathcal{G}_{1,1} \times \mathcal{I}_{2,1}$
  $V_2 = \mathcal{G}_{1,2} \times \mathcal{I}_{2,1}$
  $V_3 = \mathcal{G}_{1,3} \times \mathcal{I}_{2,1}$
  …

We will characterize essential features of this approach in the next subsection.

3.3 Properties of sum-up rounding

For our results, we use the notation $\|x\| = \|x\|_2$ for the norm of a vector $x \in \mathbb{R}^n$.

Theorem 1 *Assume that V is a compact domain in $\mathbb{R}^P$ and that $f(x)$ is Lipschitz continuous on V with Lipschitz constant L: for any $x, y \in V$,*

$$|f(x) - f(y)| \leq L\|x - y\|.$$

Consider a compatible two-level domain decomposition and let $\tilde{w}^n(x)$ be the binary function from a sum-up rounding algorithm as described in Definition 1. Let x_k be a point in R_k, $k = 1, 2, \ldots, n$. Then we have

$$\Big|\sum_{k=1}^{n} f(x_k)\left(w^n(x_k)-\tilde{w}^n(x_k)\right)\Delta_x\Big| \le \max_{x\in V}|f(x)|\frac{\mu(V-V_{rem})}{r(n)}\frac{k(n)r(n)}{n}$$
$$+\max_{j=1,2,\ldots,k(n)}\rho(V_j)2L\mu(V-V_{rem})\frac{k(n)r(n)}{n}$$
$$+2\max_{x\in V}|f(x)|\mu(V-V_{rem})\left(1-\frac{k(n)r(n)}{n}\right).$$

Moreover, if $\sum_{k=1}^{n} w^n(x_k)=n_0$ *is an integer, then* $\sum_{k=1}^{n}\tilde{w}^n(x_k)=n_0$.

Proof We prove first the result for the case where $k(n)=\tilde{k}(n)$ and $V_{rem}=\emptyset$ (that is, all subdomains V_j have the same size and properties and they exactly cover the domain V). In this case Lemma 2 gives

$$\Big|\sum_{x_k\in V_1\cup..\cup V_j}\left(w^n(x_k)-\tilde{w}^n(x_k)\right)\Delta_x\Big| \le \frac{1}{2j\cdot r(n)}\mu(V_1\cup..\cup V_j)=\frac{1}{2r(n)}\mu(V_j).$$

This implies

$$\begin{aligned}\Big|\sum_{x_k\in V_j}&\left(w^n(x_k)-\tilde{w}^n(x_k)\right)\Delta_x\Big|\\ &\le\Big|\sum_{x_k\in V_1\cup..\cup V_j}\left(w^n(x_k)-\tilde{w}^n(x_k)\right)\Delta_x\Big|\\ &\quad+\Big|\sum_{x_k\in V_1\cup..\cup V_{j-1}}\left(w^n(x_k)-\tilde{w}^n(x_k)\right)\Delta_x\Big|\\ &\le\frac{1}{2r(n)}\mu(V_j)+\frac{1}{2r(n)}\mu(V_j)\\ &=\frac{1}{r(n)}\mu(V_j).\end{aligned}\tag{16}$$

Let y_j be any point in subdomain V_j, and define

$$\Upsilon=\sum_{k=1}^{n} f(x_k)\left(w^n(x_k)-\tilde{w}^n(x_k)\right)\Delta_x,$$
$$\Psi=\sum_{j=1}^{k(n)} f(y_j)\sum_{x_k\in V_j}\left(w^n(x_k)-\tilde{w}^n(x_k)\right)\Delta_x.$$

A bound on $|\Psi|$ is given as

$$|\Psi|\le\sum_{j=1}^{k(n)}|f(y_j)|\Big|\sum_{x_k\in V_j}\left(w^n(x_k)-\tilde{w}^n(x_k)\right)\Delta_x\Big|$$

$$
\begin{aligned}
&\overset{(16)}{\leq} \max_{x\in V} |f(x)| \sum_{j=1}^{k(n)} \frac{\mu(V_j)}{r(n)} \\
&= \frac{\mu(V)}{r(n)} \max_{x\in V} |f(x)|. \qquad (17)
\end{aligned}
$$

Lipschitz continuity implies $|f(x) - f(y)| \leq L\|x - y\|$ for any $x, y \in V_j$ and

$$
\begin{aligned}
|\Upsilon - \Psi| &= \left| \Delta_x \sum_{j=1}^{k(n)} \sum_{x_k\in V_j} \left(f(x_k) - f(y_j)\right)\left(w^n(x_k) - \tilde{w}^n(x_k)\right) \right| \\
&\leq \Delta_x \sum_{j=1}^{k(n)} \sum_{x_k\in V_j} \left| \left(f(x_k) - f(y_j)\right)\left(w^n(x_k) - \tilde{w}^n(x_k)\right) \right| \\
&\leq \Delta_x \sum_{j=1}^{k(n)} \sum_{x_k\in V_j} 2L\|x_k - y_j\| \\
&\leq \sum_{j=1}^{k(n)} 2L\rho(V_j) \sum_{x_k\in V_j} \Delta_x \\
&= 2L \sum_{j=1}^{k(n)} \rho(V_j)\mu(V_j) \\
&\leq 2L\mu(V) \max_j \rho(V_j). \qquad (18)
\end{aligned}
$$

Therefore we obtain from (17) and (18) that

$$
\begin{aligned}
\left| \sum_{k=1}^{n} f(x_k)\left(w^n(x_k) - \tilde{w}^n(x_k)\right)\Delta_x \right| &= |\Upsilon| \\
&\leq |\Psi| + |\Upsilon - \Psi| \\
&\leq \max_{x\in V} |f(x)| \frac{\mu(V)}{r(n)} + 2L\mu(V) \max_j \rho(V_j).
\end{aligned}
$$

When $\tilde{k}(n) > k(n)$ and $V_{rem} \neq \emptyset$, we divide $V - V_{rem}$ into two disjoint domains $V_{main} = \bigcup_{j=1}^{k(n)} V_j$ and $V_{last} = \bigcup_{j=k(n)+1}^{\tilde{k}(n)} V_j$. We apply the results in the case $k(n) = \tilde{k}(n)$ to V_{main} to obtain

$$
\begin{aligned}
\left| \sum_{x_k\in V_{main}} f(x_k)\left(w^n(x_k) - \tilde{w}^n(x_k)\right)\Delta_x \right| &\leq \max_{x\in V} |f(x)| \frac{\mu(V_{main})}{r(n)} \\
&\quad + 2L\mu(V_{main}) \max_j \rho(V_j), \qquad (19)
\end{aligned}
$$

For the remaining part of the sum, using the fact that the components of w and $\tilde{w}$ are bounded between 0 and 1, we have

$$\left| \sum_{x_k \in V_{last}} f(x_k) \left(w^n(x_k) - \tilde{w}^n(x_k) \right) \Delta_x \right| \leq 2 \max_{x \in V} |f(x)| \mu(V_{last}). \tag{20}$$

Because each unit rectangle R_k has the same size, we have

$$\begin{aligned} \mu(V_{main}) &= \frac{k(n)r(n)}{n} \mu(V - V_{rem}), \\ \mu(V_{last}) &= \mu(V) - \mu(V_{main}) = \mu(V - V_{rem}) \left(1 - \frac{k(n)r(n)}{n} \right). \end{aligned}$$

Applying these identities to the inequalities (19) and (20), we obtain the inequality claimed in the proof. The equality is a consequence of applying the basic sum-up rounding rule from Sect. 3.1 to the set of all rectangles as described in Definition 1, in conjunction with Lemma 3. The proof is complete. □

The preceding result gives us the following immediate corollary.

Corollary 1 *With the assumptions of Theorem 1, further assume that a sequence of compatible two-level domain decompositions satisfy (13). We then obtain that*

$$\lim_{n \to \infty} \left| \sum_{k=1}^{n} f(x_k) \left(w^n(x_k) - \tilde{w}^n(x_k) \right) \Delta_x \right| = 0.$$

and that, if $\sum_{k=1}^{n} w^n(x_k) = n_0$ is an integer, then $\sum_{k=1}^{n} \tilde{w}^n(x_k) = n_0$. In other words the gap between the relaxation and our sum-up rounded integer solution goes to zero, and the integer solution is feasible for the original problem (5).

As discussed following the definition of compatible two-level domain decompositions, Definition 1, this result can be used to show the vanishing integrality gap of our approach for many types of domains. A complete analysis of when (13) holds appears extensive, though cases such as unions of rectangles or polyhedral sets do not seem to require particularly deep analysis. Given our focus on consequences for optimization, we focus exclusively on the rectangular domain case. For that situation, we can strengthen (13) and Corollary 1 by giving a bound on the rate of convergence as $n \to \infty$ (also note that $V_{rem} = \emptyset$ in that case).

Theorem 2 *Under the assumptions of Theorem 1, there exists a C such that our sum-up rounding construction satisfies*

$$\left| \sum_{k=1}^{n} f(x_k) \left(w^n(x_k) - \tilde{w}^n(x_k) \right) \Delta_x \right| \leq \frac{C}{n^{1/2P}}.$$

Proof We use the inequalities

$$\frac{(\lfloor\sqrt{n}\rfloor)^2}{n} \geq 1 - \frac{2}{\sqrt{n}}, \forall n \in \mathbb{N}, \qquad \frac{(\lfloor\sqrt{n}\rfloor)^2}{n} \geq \frac{1}{2}, \forall n > 3, \tag{21}$$

and

$$c_1 n^{\frac{1}{P}} \leq \min_{i=1,2,\ldots,P} n_i \leq n^{\frac{1}{P}}, \quad n^{\frac{1}{P}} \leq \max_{i=1,2,\ldots,P} n_i \leq c_2 n^{\frac{1}{P}} \tag{22}$$

that follow from (12).

We use the definitions of the sum-up rounding scheme parameters (14)–(15) to infer the following inequalities.

$$\frac{1}{\sqrt{n_1}} \leq c_1^{-\frac{1}{2}} n^{-\frac{1}{2P}}; \quad \frac{r(n)k(n)}{n} \leq 1; \quad \frac{1}{r(n)} = \frac{1}{\lfloor\sqrt{n_1}\rfloor} \leq \frac{\sqrt{2}}{\sqrt{n_1}} \tag{23}$$

For the maximum diameter of V_j we obtain from (15) and (21)

$$\begin{aligned}\max_{j=1,2,\ldots,k(n)} \rho(V_j) &\leq \sqrt{P}\frac{\max_{i=1,2,\ldots,P}(l_2^i - l_1^i)}{\frac{1}{2}\min_{i=1,2,\ldots,P}\sqrt{n_i}} \\ &\overset{(22)}{\leq} \sqrt{P}\frac{\max_{i=1,2,\ldots,P}(l_2^i - l_1^i)}{\frac{1}{2}\sqrt{c_1}} n^{-\frac{1}{2P}}.\end{aligned} \tag{24}$$

We also obtain

$$1 - \frac{k(n)r(n)}{n} = 1 - \frac{\lfloor\sqrt{n_1}\rfloor^2}{n_1} \overset{(21)}{\leq} 1 - \left(1 - \frac{2}{\sqrt{n_1}}\right) \overset{(23)}{\leq} 2c_1^{-\frac{1}{2}} n^{-\frac{1}{2P}}. \tag{25}$$

We now use Theorem 1 along with (23), (24), and (25) to obtain the statement of this theorem with the choice

$$\begin{aligned}C = {} & \max_{x\in V}|f(x)|\mu(V)\sqrt{2}c_1^{-\frac{1}{2}} + 4L\mu(V)\sqrt{P}\max_{i=1,2,\ldots,P}(l_2^i - l_1^i)c_1^{-\frac{1}{2}} \\ & + 4\max_{x\in V}|f(x)|\mu(V)c_1^{-\frac{1}{2}}.\end{aligned}$$

This completes the proof. □

We note that other compatible two-level relaxations observe similar bounds when used for sum-up rounding; see Sect. A.1.

4 Approximation of functions of the covariance matrix

We rely on the convergence of the sum-up rounding strategy to prove the main results on functions of covariance matrices. We keep the ratio Δ_y/Δ_x (or n/m) constant, say

α, in (3). We define

$$G_m^n = \Delta_y \cdot \{g^{w^n}(y_i, y_j)\}_{i,j=1}^m \quad \text{and} \quad \tilde{G}_m^n = \Delta_y \cdot \{g^{\tilde{w}^n}(y_i, y_j)\}_{i,j=1}^m,$$

where

$$g^{w^n}(y_i, y_j) = \alpha\sigma_{noise}^{-1} \sum_{k=1}^n f(x_k, y_i) w^n(x_k) f(x_k, y_j) \Delta_x$$
$$g^{\tilde{w}^n}(y_i, y_j) = \alpha\sigma_{noise}^{-1} \sum_{k=1}^n f(x_k, y_i) \tilde{w}^n(x_k) f(x_k, y_j) \Delta_x.$$

Here w^n is the solution to the relaxed optimization problem (6) with the discretization parameter n, and we construct $\tilde{w}^n$ from the SUR technique in Sect. 3. The quantities G_m^n, $\tilde{G}_m^n$, and Γ_{post} satisfy the following relationships

$$\Gamma_{post}(w^n) = \left(G_m^n + \sigma_{pri}^{-1} I_m\right)^{-1}, \quad \Gamma_{post}(\tilde{w}^n) = \left(\tilde{G}_m^n + \sigma_{pri}^{-1} I_m\right)^{-1}. \tag{26}$$

The assumption of Lipschitz continuity we make on $f(x, y)$ is

$$\left| f(x_1, y_1) f(x_1, y_2) - f(x_2, y_1) f(x_2, y_2) \right| \le L \|x_1 - x_2\|,$$

where $y_1, y_2 \in \Omega_{in}$ and L is independent of y_1 and y_2. This is not a stringent assumption, since we can let L depend on y_1, y_2 first and then take $L := \max_{y_1, y_2 \in \Omega_{in}} L(y_1, y_2)$ (note that Ω_{in} is bounded and closed, thus ensuring $L < \infty$). Theorem 2 then implies that

$$\forall i, j = 1, 2, \ldots, m, \quad |g^{w^n}(y_i, y_j) - g^{\tilde{w}^n}(y_i, y_j)| \le \tilde{\epsilon}_n \to 0, \quad \text{as } n \to \infty. \tag{27}$$

Here $\tilde{\epsilon}_n$ is the bound from Theorem 2. By definition of the Frobenius norm,

$$\|G_m^n - \tilde{G}_m^n\|_F \le \Delta_y \sqrt{m^2 \epsilon_n^2} = \mu(\Omega_{in}) \tilde{\epsilon}_n \to 0.$$

Since $\mu(\Omega_{in})$ is constant, we can introduce a new sequence $\{\epsilon_n\} \to 0$, $\epsilon_n = \max\{1, \mu(\Omega)\tilde{\epsilon}_n\}$. With this notation we have

$$|g^{w^n}(y_i, y_j) - g^{\tilde{w}^n}(y_i, y_j)| \le \epsilon_n \quad \text{and} \quad \|G_m^n - \tilde{G}_m^n\|_F \le \epsilon_n. \tag{28}$$

Denote eigenvalues of G_m^n and $\tilde{G}_m^n$ as

$$\lambda_1^n \ge \lambda_2^n \ge \cdots \ge \lambda_m^n \ge 0$$
$$\tilde{\lambda}_1^n \ge \tilde{\lambda}_2^n \ge \cdots \ge \tilde{\lambda}_m^n \ge 0.$$

Note the number of eigenvalues for both G_m^n and $\tilde{G}_m^n$ is m, which changes and rises up to infinity. We will show the kth eigenvalues of G_m^n and $\tilde{G}_m^n$ are arbitrarily close for any fixed $k \in \mathbb{Z}_+$.

Lemma 4 *If λ_k^n and $\tilde{\lambda}_k^n$ are the kth eigenvalues of G_m^n and $\tilde{G}_m^n$, respectively, then*

$$|\lambda_k^n - \tilde{\lambda}_k^n| \leq 2 \cdot \epsilon_n. \tag{29}$$

Proof From the Courant–Fischer theorem for real-valued symmetric matrices [16, Theorem 4.2.11], the kth largest eigenvalue of G_m^n can be computed as

$$\lambda_k^n = \sup_{dim(S)=k} \inf \left\{ \frac{\|G_m^n \cdot u\|}{\|u\|} : u \in S,\ u \neq 0 \right\}. \tag{30}$$

From this, we know there exists a subspace S of dimension k in $\mathbb{R}^m$ such that

$$\frac{\|G_m^n \cdot u\|}{\|u\|} \geq \lambda_k^n - \epsilon_n$$

for any $u \in S$, $u \neq 0$. We apply (28); and using the relationship $\|A\| \leq \|A\|_F$, we obtain

$$\begin{aligned}
\inf_{u \in S, u \neq 0} \frac{\|\tilde{G}_m^n \cdot u\|}{\|u\|} &\geq \frac{\|G_m^n \cdot u\| - \|G_m^n - \tilde{G}_m^n\|_F \|u\|}{\|u\|} \\
&\geq \frac{\|G_m^n \cdot u\| - \epsilon_n \|u\|}{\|u\|} \\
&\geq \lambda_k^n - 2\epsilon_n.
\end{aligned}$$

Again from (30), we get

$$\tilde{\lambda}_k^n \geq \lambda_k^n - 2\epsilon_n.$$

Switching $\tilde{G}_m^n$ and G_m^n and using similar arguments, we obtain the reverse inequality

$$\lambda_k^n \geq \tilde{\lambda}_k^n - 2\epsilon_n.$$

Then (29) follows directly. □

Lemma 4 can directly be used to show convergence of the gap for E-optimality, since in that case, the difference between the objectives is

$$\left| \frac{1}{\sigma + \lambda_n^n} - \frac{1}{\sigma + \tilde{\lambda}_n^n} \right| \leq \frac{|\lambda_n^n - \tilde{\lambda}_n^n|}{\sigma^2} \leq \frac{\epsilon_n}{\sigma^2}.$$

On the other hand, for integral operators with continuous kernels it can be shown that λ_n approaches zero, therefore any design will produce the same result in the limit which makes this criterion uninteresting in our setup. For the A- and D-optimality case, however, the objective function can be seen as the sum of eigenvalues or logarithm of eigenvalues of the covariance matrix, and the number of its terms goes to infinity. In that case, the objective functions may not even be bounded as $n \to \infty$, as we discuss in (54) and (55). Therefore directly invoking Lemma 4 would not prove convergence. As a simple example, consider the situation where $\lambda_k^n = 1 + \frac{k}{n\sqrt{n}}$, and $\tilde{\lambda}_k^n = 1$, $k = 1, 2, \ldots, n,$. For any k we have that $|\lambda_k^n - \tilde{\lambda}_k^n| \leq n^{-\frac{1}{2}}$ and thus the two eigenvalue sequences satisfy a relationship as the one in the conclusion of Lemma 4. On the other hand the difference between the A-optimal criteria would be

$$\sum_{k=1}^{n} \left(\frac{1}{\sigma + 1 + \frac{k}{n\sqrt{n}}} - \frac{1}{\sigma + 1} \right) \leq -\frac{\sum_{k=1}^{n} k}{n\sqrt{n}(\sigma + 1)(\sigma + 2)} \to -\infty.$$

A proof of a zero gap between function of a matrix and its SUR version will require more results beyond Lemma 4. In the following two theorems, we provide rigorous proofs on convergence for A- and D-optimal design criteria respectively.

Theorem 3 *Let $M_m^n = \left(\sigma I_m + G_m^n\right)^{-1}$ and $\tilde{M}_m^n = \left(\sigma I_m + \tilde{G}_m^n\right)^{-1}$, where $\sigma = \sigma_{pri}^{-1}$. Then*

$$tr(M_m^n) - tr(\tilde{M}_m^n) \to 0$$

as $m, n \to \infty$ and with $n/m = \alpha$ constant.

The proof is based on the fact that from Lemma 4, the spectra of G_m^n and of $\tilde{G}_m^n$ are close to each other. From the definition of M_m^n, its spectra can be inferred from that of G_m^n through $\lambda_M = 1/(\sigma + \lambda_G)$, where λ_G is an eigenvalue of G_m^n and λ_M is an eigenvalue of M_m^n. The key is to exploit this relationship to show that the spectra of M_m^n and $\tilde{M}_m^n$ are also close, combined with the consequences of the compactness of the integral operator.

Proof Since w^n and $\tilde{w}^n$ are between 0 and 1, then $g^{w^n}(y, y)$ and $g^{\tilde{w}^n}(y, y)$ are absolutely integrable.

$$\begin{aligned}
0 < \sum_{k=1}^{m} \lambda_k^n = tr(G_m^n) &= \Delta_y \cdot \sum_{i=1}^{m} g^{w^n}(y_i, y_i) \\
&\leq \Delta_y \cdot \sum_{i=1}^{m} |g^{w^n=1}(y_i, y_i)| \rightarrow \int_{\Omega_{in}} |g^{w=1}(y, y)| \, \mathrm{d}y \\
0 < \sum_{k=1}^{m} \tilde{\lambda}_k^n = tr(\tilde{G}_m^n) &= \Delta_y \cdot \sum_{i=1}^{m} g^{\tilde{w}^n}(y_i, y_i) \\
&\leq \Delta_y \cdot \sum_{i=1}^{m} |g^{\tilde{w}^n=1}(y_i, y_i)| \rightarrow \int_{\Omega_{in}} |g^{w=1}(y, y)| \, \mathrm{d}y
\end{aligned}$$

The inequality holds because $g^{w^n}(y_i, y_j)$ depends linearly on w^n. Since convergent sequences are uniformly bounded, there exists a constant $C > 0$ such that for any $n > 0$,

$$0 < \sum_{k=1}^{m} \lambda_k^n \leq C, \qquad 0 < \sum_{k=1}^{m} \tilde{\lambda}_k^n \leq C. \tag{31}$$

We also have that

$$\begin{aligned}
\Big|\sum_{k=1}^{m} \lambda_k - \sum_{k=1}^{m} \tilde{\lambda}_k^n\Big| &= \Big|\Delta_y \sum_{i=1}^{m} \Big(g^{w^n}(y_i, y_i) - g^{\tilde{w}^n}(y_i, y_i)\Big)\Big| \\
&\leq \Delta_y \sum_{i=1}^{m} \Big|g^{w^n}(y_i, y_i) - g^{\tilde{w}^n}(y_i, y_i)\Big| \\
&\leq \Delta_y \cdot m \cdot \epsilon_n = \mu(\Omega_{in})\epsilon_n,
\end{aligned}$$

where the last inequality follows from (27). Since $\mu(\Omega_{in})$ does not depend on n, and similar to the way we defined $\{\epsilon_n\}$ in (28), we can redefine the sequence $\{\epsilon_n\} \rightarrow 0$ (for example as $\epsilon_n \leftarrow \max\{1, \mu(\Omega_{in})\}\epsilon_n$) such that the following inequalities hold simultaneously

$$|g^{w^n}(y_i, y_j) - g^{\tilde{w}^n}(y_i, y_j)| \leq \epsilon_n, \ \|G_m^n - \tilde{G}_m^n\|_F \leq \epsilon_n, \ \Big|\sum_{k=1}^{m} \lambda_k^n - \sum_{k=1}^{m} \tilde{\lambda}_k^n\Big| \leq \epsilon_n. \tag{32}$$

We now show that for any small $\epsilon > 0$, there exists an integer $N > 0$ such that for any $n > N$, we have

$$\Big|S\Big| \leq D \cdot \epsilon, \quad S \triangleq \sum_{k=1}^{m} \frac{1}{\sigma + \lambda_k^n} - \sum_{k=1}^{m} \frac{1}{\sigma + \tilde{\lambda}_k^n} \tag{33}$$

with some positive constant D. Note that $n/m=\alpha$, so m is determined by n and they increase at the same rate. We fix $\epsilon > 0$. From the upper bound in (31), there are at most $N_0 = \lceil C/\epsilon \rceil$ eigenvalues satisfying $\lambda_k > \epsilon$, or equivalently, when $k > N_0$, $\lambda_k^n < \epsilon$ for any n, and, from similar reasoning, $\tilde{\lambda}_k^n < \epsilon$. From (29), there exists $N_1 > 0$ such that for any $n > N_1$, $|\lambda_k^n - \tilde{\lambda}_k^n| < \epsilon^2$ for all $k = 1, 2, \ldots, n$. We choose $n > \max\{N_0, N_1\}$ and split the sum in (33) into two parts:

$$S = \sum_{k \le N_0} \left(\frac{1}{\sigma + \lambda_k^n} - \frac{1}{\sigma + \tilde{\lambda}_k^n} \right) + \sum_{k > N_0} \left(\frac{1}{\sigma + \lambda_k^n} - \frac{1}{\sigma + \tilde{\lambda}_k^n} \right).$$

For the first part, we note that

$$\Big| \sum_{k \le N_0} \left(\frac{1}{\sigma + \lambda_k^n} - \frac{1}{\sigma + \tilde{\lambda}_k^n} \right) \Big| \le \sum_{k \le N_0} \frac{|\lambda_k^n - \tilde{\lambda}_k^n|}{(\sigma + \lambda_k^n)(\sigma + \tilde{\lambda}_k^n)} \le N_0 \cdot \frac{\epsilon^2}{\sigma^2} \le \frac{C}{\sigma^2} \cdot \epsilon. \tag{34}$$

For the second part, we know $\lambda_k^n,\ \tilde{\lambda}_k^n < \epsilon$, and we discuss two cases.

(1) When $\tilde{\lambda}_k^n > \lambda_k^n$ and $k > N_0$,

$$0 < \frac{1}{(\sigma + \epsilon)^2}(\tilde{\lambda}_k^n - \lambda_k^n) \le \frac{1}{\sigma + \lambda_k^n} - \frac{1}{\sigma + \tilde{\lambda}_k^n} = \frac{\tilde{\lambda}_k^n - \lambda_k^n}{(\sigma + \lambda_k^n)(\sigma + \tilde{\lambda}_k^n)} \le \frac{\tilde{\lambda}_k^n - \lambda_k^n}{\sigma^2}. \tag{35}$$

(2) When $\tilde{\lambda}_k^n < \lambda_k^n$ and $k > N_0$,

$$\frac{\tilde{\lambda}_k^n - \lambda_k^n}{\sigma^2} \le \frac{1}{\sigma + \lambda_k^n} - \frac{1}{\sigma + \tilde{\lambda}_k^n} = \frac{\tilde{\lambda}_k^n - \lambda_k^n}{(\sigma + \lambda_k^n)(\sigma + \tilde{\lambda}_k^n)} \le \frac{\tilde{\lambda}_k^n - \lambda_k^n}{(\sigma + \epsilon)^2} < 0. \tag{36}$$

So we have

$$\begin{aligned}
&\sum_{k > N_0} \left(\frac{1}{\sigma + \lambda_k^n} - \frac{1}{\sigma + \tilde{\lambda}_k^n} \right) \\
&\quad \le \sum_{\tilde{\lambda}_k^n > \lambda_k^n,\ k > N_0} \frac{\tilde{\lambda}_k^n - \lambda_k^n}{\sigma^2} + \sum_{\tilde{\lambda}_k^n < \lambda_k^n,\ k > N_0} \frac{\tilde{\lambda}_k^n - \lambda_k^n}{(\sigma + \epsilon)^2} \\
&\quad = \sum_{k > N_0} \frac{\tilde{\lambda}_k^n - \lambda_k^n}{\sigma^2} + \sum_{\tilde{\lambda}_k^n < \lambda_k^n,\ k > N_0} \left(\frac{1}{(\sigma + \epsilon)^2} - \frac{1}{\sigma^2} \right) (\tilde{\lambda}_k^n - \lambda_k^n) \\
&\quad = \frac{1}{\sigma^2} \sum_{k > N_0} (\tilde{\lambda}_k^n - \lambda_k^n) + \frac{\epsilon(2\sigma + \epsilon)}{\sigma^2(\sigma + \epsilon)^2} \sum_{\tilde{\lambda}_k^n < \lambda_k^n,\ k > N_0} (\lambda_k^n - \tilde{\lambda}_k^n).
\end{aligned}$$

With a similar use of (35) and (36) we obtain

$$\begin{aligned}
&\sum_{k>N_0}\left(\frac{1}{\sigma+\lambda_k^n}-\frac{1}{\sigma+\tilde{\lambda}_k^n}\right)\\
&\quad\geq \sum_{\tilde{\lambda}_k^n>\lambda_k^n,\ k>N_0}\frac{\tilde{\lambda}_k^n-\lambda_k^n}{(\sigma+\epsilon)^2}+\sum_{\tilde{\lambda}_k^n<\lambda_k^n,\ k>N_0}\frac{\tilde{\lambda}_k^n-\lambda_k^n}{\sigma^2}\\
&\quad= \sum_{\tilde{\lambda}_k^n>\lambda_k^n,\ k>N_0}\left(\frac{1}{(\sigma+\epsilon)^2}-\frac{1}{\sigma^2}\right)(\tilde{\lambda}_k^n-\lambda_k^n)+\sum_{k>N_0}\frac{\tilde{\lambda}_k^n-\lambda_k^n}{\sigma^2}\\
&\quad= \frac{1}{\sigma^2}\sum_{k>N_0}(\tilde{\lambda}_k^n-\lambda_k^n)+\frac{\epsilon(2\sigma+\epsilon)}{\sigma^2(\sigma+\epsilon)^2}\sum_{\tilde{\lambda}_k^n>\lambda_k^n,\ k>N_0}(\lambda_k^n-\tilde{\lambda}_k^n).
\end{aligned}$$

From the last two inequalities, we obtain

$$\begin{aligned}
&\Big|\sum_{k>N_0}\left(\frac{1}{\sigma+\lambda_k^n}-\frac{1}{\sigma+\tilde{\lambda}_k^n}\right)\Big|\leq\frac{1}{\sigma^2}\Big|\sum_{k>N_0}(\lambda_k^n-\tilde{\lambda}_k^n)\Big|\\
&\quad+\frac{\epsilon(2\sigma+\epsilon)}{\sigma^2(\sigma+\epsilon)^2}\max\left\{\sum_{\tilde{\lambda}_k<\lambda_k,\ k>N_0}(\lambda_k^n-\tilde{\lambda}_k^n),\sum_{\tilde{\lambda}_k>\lambda_k,\ k>N_0}(\tilde{\lambda}_k^n-\lambda_k^n)\right\}.
\end{aligned}\tag{37}$$

In order to bound $\sum_{k>N_0}(\tilde{\lambda}_k^n-\lambda_k^n)$, recall (32). From it, there exists $N_2>0$ such that for any $n>N_2$ we have

$$\Big|\sum_{k=1}^{m}(\lambda_k^n-\tilde{\lambda}_k^n)\Big|<\epsilon.\tag{38}$$

Choose $n>\max\{N_0,N_1,N_2\}$. Because $n\geq N_1$, we have

$$\Big|\sum_{k\leq N_0}(\lambda_k^n-\tilde{\lambda}_k^n)\Big|\leq N_0\cdot\epsilon^2=C\epsilon,\tag{39}$$

and thus from (38), (39) and the triangle inequality we get

$$\Big|\sum_{k>N_0}(\lambda_k^n-\tilde{\lambda}_k^n)\Big|\leq\Big|\sum_{k=1}^{m}(\lambda_k^n-\tilde{\lambda}_k^n)\Big|+\Big|\sum_{k\leq N_0}(\lambda_k^n-\tilde{\lambda}_k^n)\Big|\leq(C+1)\epsilon.\tag{40}$$

Note that if we let $\epsilon<\sigma$ and use (31), we obtain

$$0<\frac{\epsilon(2\sigma+\epsilon)}{\sigma^2(\sigma+\epsilon)^2}\sum_{\tilde{\lambda}_k<\lambda_k,\ k>N_0}(\lambda_k^n-\tilde{\lambda}_k^n)\leq\frac{3\epsilon}{\sigma^3}\sum_{k=1}^{m}\lambda_k^n\leq\frac{3C}{\sigma^3}\epsilon\tag{41}$$

$$0 < \frac{\epsilon(2\sigma+\epsilon)}{\sigma^2(\sigma+\epsilon)^2} \sum_{\tilde{\lambda}_k > \lambda_k,\ k>N_0} (\tilde{\lambda}_k^n - \lambda_k^n) \le \frac{3\epsilon}{\sigma^3}\sum_{k=1}^{m} \tilde{\lambda}_k^n \le \frac{3C}{\sigma^3}\epsilon. \tag{42}$$

Combining (37), (40), (41), and (42), we get

$$\Big|\sum_{k>N_0}\left(\frac{1}{\sigma+\lambda_k^n} - \frac{1}{\sigma+\tilde{\lambda}_k^n}\right)\Big| \le \frac{C+1}{\sigma^2}\epsilon + \frac{3C}{\sigma^3}\epsilon = \left(\frac{C+1}{\sigma^2} + \frac{3C}{\sigma^3}\right)\epsilon \tag{43}$$

According to (34) and (43), we get

$$\begin{aligned}\Big|\sum_{k=1}^{m}\left(\frac{1}{\sigma+\lambda_k^n} - \frac{1}{\sigma+\tilde{\lambda}_k^n}\right)\Big| \le & \Big|\sum_{k\le N_0}\left(\frac{1}{\sigma+\lambda_k^n} - \frac{1}{\sigma+\tilde{\lambda}_k^n}\right)\Big| \\ & + \Big|\sum_{k>N_0}\left(\frac{1}{\sigma+\lambda_k^n} - \frac{1}{\sigma+\tilde{\lambda}_k^n}\right)\Big| \\ & \le \frac{C}{\sigma^2}\epsilon + \left(\frac{C+1}{\sigma^2} + \frac{3C}{\sigma^3}\right)\epsilon = \left(\frac{2C+1}{\sigma^2} + \frac{3C}{\sigma^3}\right)\epsilon.\end{aligned}$$

Let $D = \frac{2C+1}{\sigma^2} + \frac{3C}{\sigma^3}$. Then for any $\epsilon > 0$ smaller than σ, there exists $N = \max\{N_0, N_1, N_2\}$ such that for any $n > N$,

$$\Big|\sum_{k=1}^{m}\left(\frac{1}{\sigma+\lambda_k^n} - \frac{1}{\sigma+\tilde{\lambda}_k^n}\right)\Big| \le D\cdot\epsilon.$$

By definition of limit, as $m, n \to \infty$ and $n/m = \alpha$,

$$\Big|\sum_{k=1}^{m}\left(\frac{1}{\sigma+\lambda_k^n} - \frac{1}{\sigma+\tilde{\lambda}_k^n}\right)\Big| \to 0. \tag{44}$$

Given that the first quantity in (44) is $tr(M_m^n)$ and the second is $tr(\tilde{M}_m^n)$, the conclusion follows. □

Theorem 4 *$logdet(M_m^n) - logdet(\tilde{M}_m^n) \to 0$, or equivalently*

$$\sum_{k=1}^{m}\log\frac{1}{\sigma+\lambda_k^n} - \sum_{k=1}^{m}\log\frac{1}{\sigma+\tilde{\lambda}_k^n} \to 0.$$

Here, M_m^n and $\tilde{M}_m^n$ are the matrices from Theorem 3.

Proof First note that using the mean value theorem and the monotonicity of the log function and its derivative, we have that, if $0 < c_1 < x < y < c_2$, then

$$0 < \frac{1}{c_2}(y-x) \le \log\frac{1}{x} - \log\frac{1}{y} \le \frac{1}{c_1}(y-x). \tag{45}$$

Again we show that for any $\epsilon > 0$, there exists an integer $N > 0$ such that for any $n > N$,

$$\Big| \sum_{k=1}^{m} \log \frac{1}{\sigma + \lambda_k^n} - \sum_{k=1}^{m} \log \frac{1}{\sigma + \tilde{\lambda}_k^n} \Big| \leq D \cdot \epsilon, \tag{46}$$

with some positive constant D. First, from (31) we choose N_0 such that when $k > N_0$, $\lambda_k^n < \epsilon$ and $\tilde{\lambda}_k^n < \epsilon$ for any n, using a similar argument in the proof of Theorem 3. Second, from (29), we can find $N_1 > 0$ such that for any $n > N_1$, $|\lambda_k^n - \tilde{\lambda}_k^n| < \epsilon^2$ for all $k = 1, 2, \ldots, n$. Third, from (32) there exists $N_2 > 0$ such that for any $n > N_2$, $\Big| \sum_{k=1}^{m} (\lambda_k^n - \tilde{\lambda}_k^n) \Big| < \epsilon$. We then split the sum in (46) into two parts:

$$\sum_{k \leq N_0} \left(\log \frac{1}{\sigma + \lambda_k^n} - log \frac{1}{\sigma + \tilde{\lambda}_k^n} \right) + \sum_{k > N_0} \left(\log \frac{1}{\sigma + \lambda_k^n} - \log \frac{1}{\sigma + \tilde{\lambda}_k^n} \right).$$

For the first part, we apply (45) to obtain

$$\Big| \sum_{k \leq N_0} \left(\log \frac{1}{\sigma + \lambda_k^n} - \log \frac{1}{\sigma + \tilde{\lambda}_k^n} \right) \Big| \leq \sum_{k \leq N_0} \frac{|\lambda_k^n - \tilde{\lambda}_k^n|}{\sigma} \leq N_0 \cdot \frac{\epsilon^2}{\sigma} = \frac{C}{\sigma} \cdot \epsilon. \tag{47}$$

For the second part, $0 \leq \lambda_k^n,\ \tilde{\lambda}_k^n < \epsilon$, and we discuss two cases.

(1) When $\tilde{\lambda}_k^n > \lambda_k^n$ and $k > N_0$,

$$0 < \frac{1}{\sigma + \epsilon} (\tilde{\lambda}_k^n - \lambda_k^n) \leq \log \frac{1}{\sigma + \lambda_k^n} - \log \frac{1}{\sigma + \tilde{\lambda}_k^n} \leq \frac{\tilde{\lambda}_k^n - \lambda_k^n}{\sigma}.$$

(2) When $\tilde{\lambda}_k^n < \lambda_k^n$ and $k > N_0$,

$$\frac{\tilde{\lambda}_k^n - \lambda_k^n}{\sigma} \leq \log \frac{1}{\sigma + \lambda_k^n} - \log \frac{1}{\sigma + \tilde{\lambda}_k^n} \leq \frac{\tilde{\lambda}_k^n - \lambda_k^n}{\sigma + \epsilon} < 0.$$

Therefore, we have

$$\begin{aligned} &\sum_{k > N_0} \left(\log \frac{1}{\sigma + \lambda_k^n} - \log \frac{1}{\sigma + \tilde{\lambda}_k^n} \right) \\ &\quad \leq \sum_{\tilde{\lambda}_k^n > \lambda_k^n,\ k > N_0} \frac{\tilde{\lambda}_k^n - \lambda_k^n}{\sigma} + \sum_{\tilde{\lambda}_k^n < \lambda_k^n,\ k > N_0} \frac{\tilde{\lambda}_k^n - \lambda_k^n}{\sigma + \epsilon} \end{aligned}$$

$$\begin{aligned}
&= \sum_{k>N_0} \frac{\tilde{\lambda}_k^n - \lambda_k^n}{\sigma} + \sum_{\tilde{\lambda}_k^n<\lambda_k^n,\; k>N_0} \left(\frac{1}{\sigma+\epsilon} - \frac{1}{\sigma}\right)(\tilde{\lambda}_k^n - \lambda_k^n)\\
&= \frac{1}{\sigma}\sum_{k>N_0}(\tilde{\lambda}_k^n - \lambda_k^n) + \frac{\epsilon}{\sigma(\sigma+\epsilon)} \sum_{\tilde{\lambda}_k^n<\lambda_k^n,\; k>N_0} (\lambda_k^n - \tilde{\lambda}_k^n)
\end{aligned}$$

and similarly

$$\begin{aligned}
&\sum_{k>N_0}\left(\log\frac{1}{\sigma+\lambda_k^n} - \log\frac{1}{\sigma+\tilde{\lambda}_k^n}\right)\\
&\quad\geq \sum_{\tilde{\lambda}_k^n>\lambda_k^n,\; k>N_0} \frac{\tilde{\lambda}_k^n - \lambda_k^n}{\sigma+\epsilon} + \sum_{\tilde{\lambda}_k^n<\lambda_k^n,\; k>N_0} \frac{\tilde{\lambda}_k^n - \lambda_k^n}{\sigma}\\
&\quad= \sum_{\tilde{\lambda}_k^n>\lambda_k^n,\; k>N_0} \left(\frac{1}{\sigma+\epsilon} - \frac{1}{\sigma}\right)(\tilde{\lambda}_k^n - \lambda_k^n) + \sum_{k>N_0}\frac{\tilde{\lambda}_k^n - \lambda_k^n}{\sigma}\\
&\quad= \frac{1}{\sigma}\sum_{k>N_0}(\tilde{\lambda}_k^n - \lambda_k^n) + \frac{\epsilon}{\sigma(\sigma+\epsilon)} \sum_{\tilde{\lambda}_k^n>\lambda_k^n,\; k>N_0} (\lambda_k^n - \tilde{\lambda}_k^n).
\end{aligned}$$

From these two inequalities, we get

$$\begin{aligned}
&\Big|\sum_{k>N_0}\left(\log\frac{1}{\sigma+\lambda_k^n} - \log\frac{1}{\sigma+\tilde{\lambda}_k^n}\right)\Big| \leq \frac{1}{\sigma}\Big|\sum_{k>N_0}(\lambda_k^n - \tilde{\lambda}_k^n)\Big|\\
&\quad+ \frac{\epsilon}{\sigma(\sigma+\epsilon)}\max\left\{\sum_{\tilde{\lambda}_k<\lambda_k,\; k>N_0}(\lambda_k^n - \tilde{\lambda}_k^n),\ \sum_{\tilde{\lambda}_k>\lambda_k,\; k>N_0}(\tilde{\lambda}_k^n - \lambda_k^n)\right\}. \qquad (48)
\end{aligned}$$

Using the same rationale that led us to (40), we have

$$\Big|\sum_{k>N_0}(\lambda_k^n - \tilde{\lambda}_k^n)\Big| \leq \Big|\sum_{k=1}^{m}(\lambda_k^n - \tilde{\lambda}_k^n)\Big| + \Big|\sum_{k\leq N_0}(\lambda_k^n - \tilde{\lambda}_k^n)\Big| \leq (C+1)\epsilon. \qquad (49)$$

Moreover, using (31) and the nonnegativity of the eigenvalues of M^n and $\tilde{M}^n$, we obtain

$$0 < \frac{\epsilon}{\sigma(\sigma+\epsilon)}\sum_{\tilde{\lambda}_k<\lambda_k,\; k>N_0}(\lambda_k^n - \tilde{\lambda}_k^n) \leq \frac{\epsilon}{\sigma^2}\sum_{k=1}^{m}\lambda_k^n \leq \frac{C}{\sigma^2}\epsilon \qquad (50)$$

$$0 < \frac{\epsilon}{\sigma(\sigma+\epsilon)}\sum_{\tilde{\lambda}_k>\lambda_k,\; k>N_0}(\tilde{\lambda}_k^n - \lambda_k^n) \leq \frac{\epsilon}{\sigma^2}\sum_{k=1}^{m}\tilde{\lambda}_k^n \leq \frac{C}{\sigma^2}\epsilon. \qquad (51)$$

Combining (48), (49), (50) and (51), we get

$$\Big|\sum_{k>N_0}\left(\log\frac{1}{\sigma+\lambda_k^n}-\log\frac{1}{\sigma+\tilde{\lambda}_k^n}\right)\Big| \le \frac{C+1}{\sigma}\epsilon+\frac{C}{\sigma^2}\epsilon=\left(\frac{C+1}{\sigma}+\frac{C}{\sigma^2}\right)\epsilon \tag{52}$$

Using the bounds (47) and (52), we get

$$\begin{aligned}&\Big|\sum_{k=1}^{m}\left(\log\frac{1}{\sigma+\lambda_k^n}-\log\frac{1}{\sigma+\tilde{\lambda}_k^n}\right)\Big|\\ &\quad\le\Big|\sum_{k\le N_0}\left(\log\frac{1}{\sigma+\lambda_k^n}-\log\frac{1}{\sigma+\tilde{\lambda}_k^n}\right)\Big|+\Big|\sum_{k>N_0}\left(\log\frac{1}{\sigma+\lambda_k^n}-\log\frac{1}{\sigma+\tilde{\lambda}_k^n}\right)\Big|\\ &\quad\le\frac{C}{\sigma}\epsilon+\left(\frac{C+1}{\sigma}+\frac{C}{\sigma^2}\right)\epsilon=\left(\frac{2C+1}{\sigma}+\frac{C}{\sigma^2}\right)\epsilon.\end{aligned}$$

Let $D=\frac{2C+1}{\sigma}+\frac{C}{\sigma^2}$. We conclude that for any $\epsilon>0$, there exists $N=\max\{N_0, N_1, N_2\}$ such that for any $n>N$,

$$\Big|\sum_{k=1}^{m}\left(\log\frac{1}{\sigma+\lambda_k^n}-\log\frac{1}{\sigma+\tilde{\lambda}_k^n}\right)\Big|\le D\cdot\epsilon.$$

By definition of limit, as $m, n\to\infty$ and $n/m=\alpha$,

$$\Big|\sum_{k=1}^{m}\left(\log\frac{1}{\sigma+\lambda_k^n}-\log\frac{1}{\sigma+\tilde{\lambda}_k^n}\right)\Big|\to 0.$$

Given that the first quantity is the logarithm of the determinant of M_m^n and the second is the logarithm of the determinant of $\tilde{M}_m^n$, this proves the claim. □

Given the relation of Γ_{post} and G_m^n in (26), Theorem 3 proves that the lower bound of the A-optimal design, which is given by the relaxed optimization problem (6), can be achieved by using the sum-up rounding strategy. Theorem 4 does the same for the D-optimal design. The E-optimal design, where we aim to minimize the largest eigenvalue of Γ_{post}, is actually trivial in this framework because the smallest eigenvalue of G_m^n goes to 0 and the largest eigenvalue of $\Gamma_{post}(w^n)$ converges to σ_{pri}, which is also true for $\Gamma_{post}(\tilde{w}^n)$. This argument also shows that the E-optimal result is trivial for this case since virtually any design will then be E-optimal; hence we do not emphasize it in this paper. To conclude, with the sum-up rounding strategy described in Sect. 4, we are able to find sensor locations that are asymptotically optimal for A and D design criteria.

While our proofs include several restrictions, they can be extended in several ways. To include more general domains or sum-up rounding patterns would require proving results such as Theorem 2 and, subsequently, the critical property (27) needed to show the shrinking gap for a given design strategy. General domains are not difficult to include, but the resulting proofs would be extensive, involving computational geometry technicalities. However, the two-level strategy presented in Definition 1 resembles the spectral element philosophy [24] that is widely used for quite complex domains. Moreover, the within-subdomain ordering in Definition 1 is entirely open, which would allow experimentation with various strategies such as space-filling curves. While our results are proved for linear operators only, we note that as a first step to extending our results to the case where the nonlinear parameter-to-observable map F is nonlinear, one could use the Laplace approximation as was done in [2,33].

5 Numerical experiments

We now present numerical experiments based on the model problem of gravity surveying (see Example 1.5 in [20]) in our simulation. Suppose mass is distributed at depth d below the surface where sensors can be deployed, in a unit square $[0, 1] \times [0, 1]$ indexed by the two-dimensional variable y, and we want to estimate the mass density function $g_0(y)$. Measurements are carried out on a unit square in a plane indexed by the two-dimensional variable x, and we can measure the vertical component of gravitational field $g(x)$ but with error. By Newton's law of universal gravitation, the integral equation of $g(x)$ for $x \in [0, 1] \times [0, 1]$ is

$$g(x) = \int_{[0,1]\times[0,1]} f(x, y) g_0(y)\, \mathrm{d}y, \qquad f(x, y) = \frac{d}{(d^2 + \|x - y\|^2)^{3/2}},$$

where $\|x - y\|$ is the Euclidean distance between points x and y. In this problem, $\Omega_{in} = \Omega_{out}$, and we use the same discretization for the two domains. We divide $[0, 1] \times [0, 1]$ into n^2 small squares with equal size $1/n^2$. On each side, there are n points $0 < x_1 < x_2 < \cdots < x_n < 1$ ($x_i = i/n + 0.5$) and $\Delta_x = 1/n$. We have n^2 candidate locations, and $w = (w_1, w_2, \ldots, w_{n^2})$ is the corresponding weight vector. Let $F \in \mathbb{R}^{n^2 * n^2}$ be the discretization of the above integral operator, and order the candidate locations as $z_1, z_2, \ldots, z_{n^2}$. Then

$$F(i, j) = \frac{d}{(d^2 + \|z_i - z_j\|^2)^{3/2}} \cdot (\Delta_x)^2,$$

for $i, j = 1, 2, \ldots, n^2$. Let $W = \mathrm{diag}(w)$. The relaxed problem is

$$\begin{aligned} \min_{w} \quad & \phi\left(\left(F^T W F + \sigma I_{n^2}\right)^{-1}\right) \\ s.t. \quad & 0 \le w_i \le 1, \quad \sum_i w_i = \left\lfloor r n^2 \right\rfloor \quad (0 < r < 1), \end{aligned} \qquad (53)$$

where σ is not a variance but the ratio of σ_{noise} and σ_{pri}. We keep the number of sensors in a proportion r to the number of candidate locations, as discussed at the end of Sect. 2.5.

Using the solver `Ipopt` in Julia, we compute w_{rel} and then construct a feasible integer vector w_{int} via the sum-up round approach we developed in this paper. Our experiments are run on a laptop with one CPU processor (1.6 GHz Intel Core i5) and 4 GB of memory, and we provide the Hessian of the objective and the relevant objective and constraint gradients. By far the most expensive part of the computation is the Hessian. For example, for the case where $\phi(\cdot) = tr(\cdot)$, the entry ij in the Hessian is proportional to $tr(\Gamma^{-1} f_i f_i^T \Gamma^{-1} f_j f_j^T \Gamma^{-1})$, where $\Gamma = \left(F^T W F + \sigma I_{n^2}\right)$ and f_i is the ith column of F^T. Here $1 \leq i < j \leq n^2$. Note that Γ is a dense matrix, and for the rest of discussion in this paragraph, n represents the size of Γ. While the computation can be streamlined to carry out the factorization of Γ once per iteration, followed by solving n linear systems of equations with f_i, then computing $\approx \frac{n^2}{2}$ inner products, each of these operations is $O(n^3)$. The largest problem we solve has $n = 3600$ (a 60×60 two-dimensional grid) and `Ipopt` takes about 3 hours to produce a solution for it, though our code is far from optimized. Interestingly, note that computing even one entry in the gradient, whose $i$the entry is $-tr(\Gamma^{-1} f_i f_i^T \Gamma^{-1})$ would still take $O(n^3)$ as at least one linear system with Γ needs to be solved. For this reason it is doubtful one can do much better, as most convex integer programming solvers need gradients of the objective. In any case, we had difficulties comparing with other approaches, as most of the ones we had reasonably easy access to required the function to be expressible in a modeling environment such as JuMP or AMPL. This does not occur for matrix functions, as they cannot atomically be expressed in terms of standard libraries. An alternative was to reformulate the problem (53) as a semidefinite program with integer variables, which we aimed to do with `Pajarito`. However, solving the $n = 50$ case (in one dimension) took one hour to achieve a gap of less than 1%. Therefore this did not appear to be an easy way to go either. Solving larger problems will probably require reaching towards other ideas, such as perhaps exploiting the (approximate) hierarchical off diagonal low rank structure, as we recently proposed in [14].

In any case, results for D-optimal and A-optimal designs using `Ipopt` as described above are demonstrated below. The E-optimal design is not considered because the largest eigenvalue is extremely close to $1/\sigma$ irrespective of w and there is not much difference in objective values for different designs.

We compare our sum-up rounding design with a thresholding heuristic: let $w = (w_1, w_2, \cdots, w_{n^2})$ be the relaxed solution and its components are ordered by $w_{i_1} \geq w_{i_2} \geq \cdots \geq w_{i_{n^2}}$. The thresholding integer solution $\tilde{w}$ is given by

$$\tilde{w}_j = \begin{cases} 1, & \text{if } j \in \left\{i_1, i_2, \ldots, i_{\lfloor rn^2 \rfloor}\right\}; \\ 0, & \text{otherwise.} \end{cases}$$

In other words, we set elements to 1 if they have the largest values in the relaxation, up to the available budget of sensors. We will compare the performance of two strategies by measuring integrality gap.

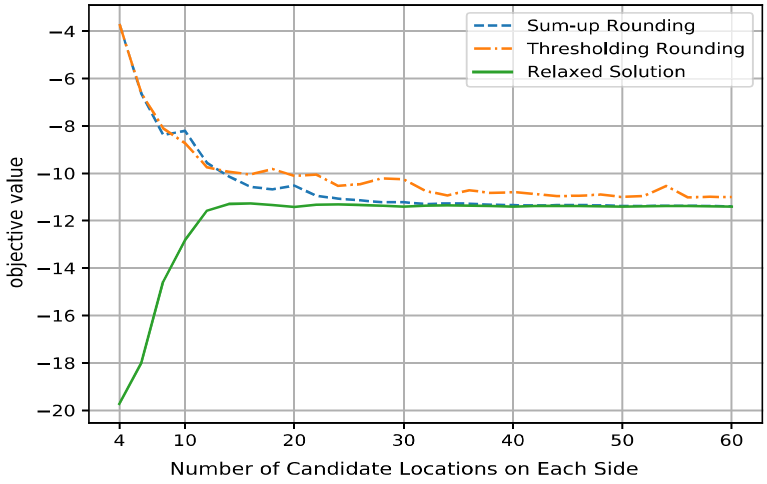


Fig. 1 Objective value, D-optimal design

5.1 D-optimal design

The parameters we choose are $\sigma = 1$, $d = 0.1$, $r = 0.1$. Figure 1 shows the objective value (i.e. log determinant) with the continuous relaxation, sum-up rounding and thresholding strategy as n increases from 4 to 60. For the thresholding heuristic, it does not seem to converge at $n = 40$, or at least its gap decreases more slowly than sum-up rounding. We note that this validates the result of Theorem 3. One point we want to add is the objective value in Fig. 1 converges to a fixed number (around -11.3), which is related to our choice $\sigma = 1$. Notice, when $\sigma = 1$, that

$$logdet(\Gamma_{post}) = \sum_{k=1}^{n^2} \log \frac{1}{\sigma + \lambda_k} \approx \sum_{k=1}^{n^2} (-\lambda_k) \tag{54}$$

and $\sum \lambda_k$ is finite, see (31). For other values of σ, the objective value will approach infinity, but the gap will still converge to zero as proved by our theorem.

We also plot the absolute and relative gaps for the two rounding strategies in Fig. 2, in logarithmic scale. The relative gap is defined as the ratio of absolute gap and the lower bound from the relaxation. We observe that sum-up rounding has a relative gap below 1% at $n = 40$, compared with 5% for the thresholding heuristic.

Figures 3, 4 and 5 give the relaxed solution, the sum-up rounding solution and thresholding solution, respectively, when $n = 40$ (there are 1600 variables). The design is symmetric since both $f(x, y)$ and the output domain $[0, 1] \times [0, 1]$ are symmetric. Sensors are placed toward the boundary and also in the interior. We note that the design highly depends on d: When d goes to zero or infinity, the relaxed solution tends to be uniform. Therefore, if we hope to observe interesting designs, d should be neither too big nor too small. For the thresholding heuristic, a common feature is that sensors tend to be placed together when values in the relaxation change smoothly,

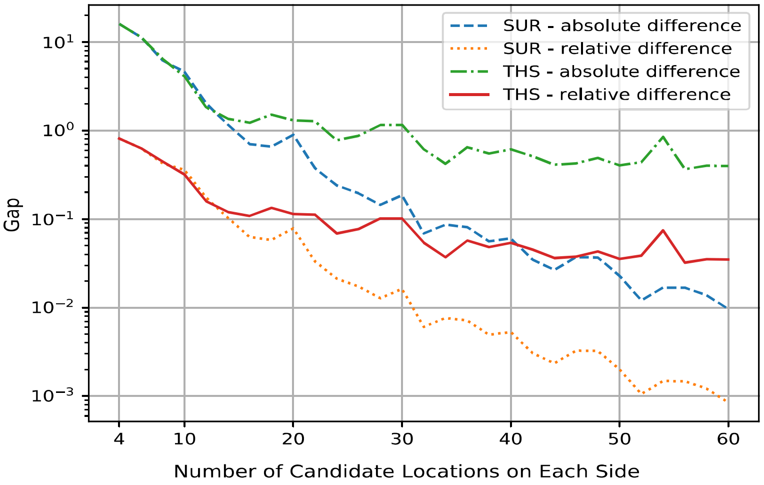
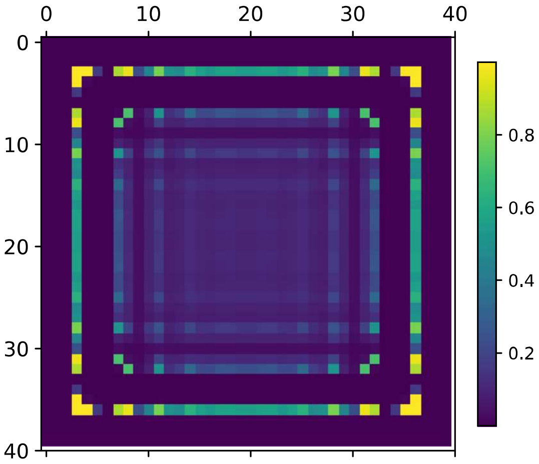


Fig. 2 Integrality gap, D-optimal design (SUR = sum-up rounding; THS = thresholding rounding)

Fig. 3 Relaxation, D-optimal design



and we do not see sensors placed near the center. Sum-up rounding, however, has the property that the 0 or 1 value in the relaxation will remain the same in the integer solution, and the sensor placement is less concentrated than for the thresholding heuristic.

5.2 A-optimal design

We investigate the A-optimal design with the same setting and parameters as in the D-optimal design case: $\sigma = 1$, $d = 0.1$, $r = 0.1$, and n starting at 4 and ending at 50. We observe in Fig. 6 a similar decaying trend as in the D-optimal design case, which validates the finding of Theorem 4. We would like to mention that in the trace case,

Fig. 4 SUR solution, D-optimal design

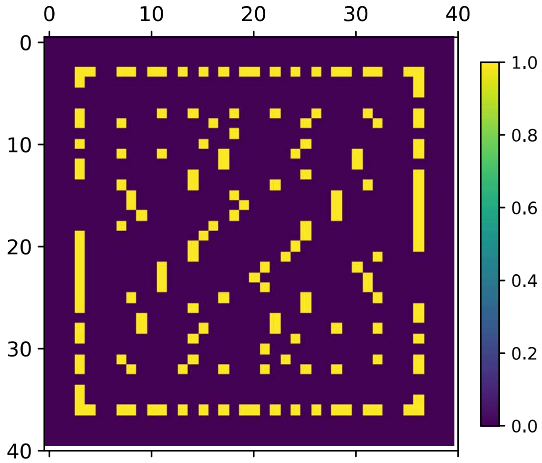
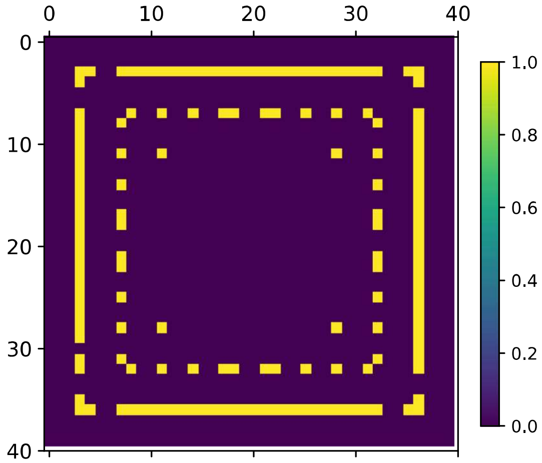


Fig. 5 THS solution, D-optimal design



$$tr(\Gamma_{post}) = \sum_{k=1}^{n^2} \frac{1}{\sigma + \lambda_k} = O(n^2), \tag{55}$$

so the optimal objective value increases about linearly with respect to the number of candidate locations. However, both the absolute and relative gaps between the upper bound induced by sum-up rounding and the lower bound obtained from the relaxation approach zero for large n, as shown in Fig. 6 and as claimed in Sect. 4.

The designs in Figs. 7, 8 and 9 also have patterns similar to those in Figs. 3, 4 and 5, although they are slightly more centered. It is worth mentioning that, as indicated by Figs. 2 and 6, monotonicity with n is unlikely. Indeed, kinks at $n = 20, 30, \ldots$ are related to the particularities of sum-up rounding design. When n reaches those values, there is a change in shape which induces a small increase in the gap, but the gap will be under control and eventually go to zero.

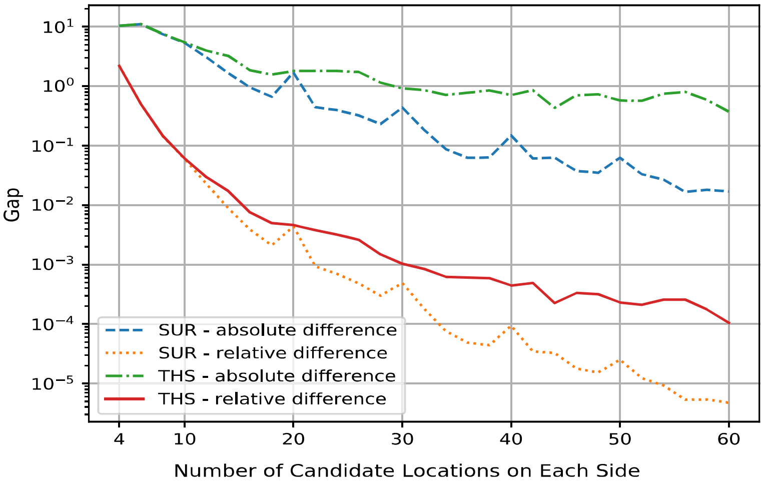
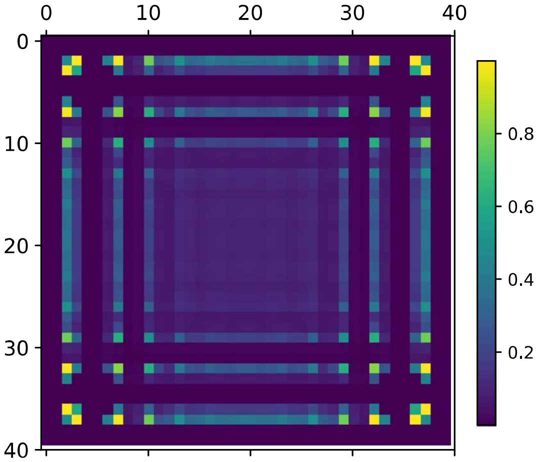


Fig. 6 Integrality gap, A-optimal design (SUR = sum-up rounding; THS = thresholding rounding)

Fig. 7 Relaxation, A-optimal design



5.3 Discussion

In practice, we normally do not wish to see clusters of sensors because data are usually informative of other data nearby, while sum-up rounding tends to place sensors close to each other because of smoothness in the relaxed solutions. One can request the sensor density not to exceed a given value in any region. An alternative is to use a space-filling curve approach for the sum-up rounding path to "randomize" the choices of 1. For this initial study, we note the significant improvement in the objective, and we leave such issues to further research.

Fig. 8 SUR solution, A-optimal design

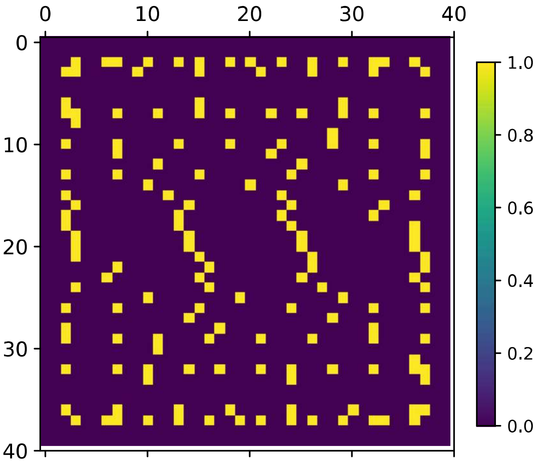
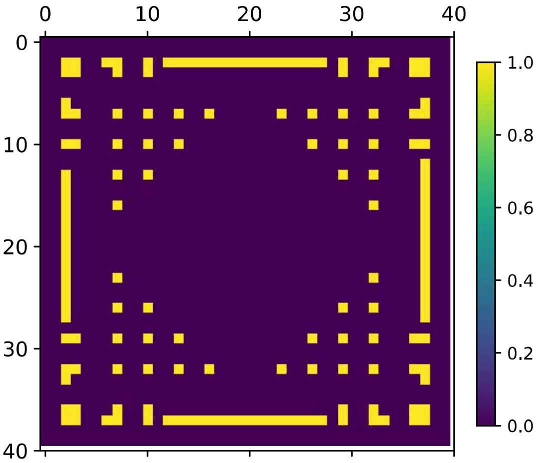


Fig. 9 THS solution, A-optimal design



6 Conclusions

Using a Bayesian estimation framework, we propose a multidimensional sum-up rounding strategy to compute asymptotically optimal sensor locations for systems where the output depends on input through an integral equation. The approach is an extension of recent ideas by Sager et al. that were proposed in the context of mixed-integer optimal control [26,27]. Our method can be used on systems for which the input to output relationship is linear or can be well approximated by one. The optimization problems obtained by relaxing the binary site selection constraints are convex and can be solved efficiently with interior-point algorithms. Our main result is that for different optimal experimental design criteria (called A-optimal and D-optimal in DOE terminology), the integrality gap between the objective value of the relaxed solution and the rounded-up solution shrinks to zero in the limit of increasingly fine meshes on which the integral equation is approximated. We validate this finding with two-dimensional numerical experiments from the gravity surveying problem.

This initial work has several limitations. We give complete proofs only for a rectangular domain, but the extensions to many other domains seem straightforward, if perhaps technically complex from a geometrical content standpoint. The two-level nature of our sum-up rounding allows inclusion of more sophisticated strategies in the subdomains, such as those based on space-filling curves. These may result in better rounding schemes, particularly at avoiding the excess clustering that is clearly suboptimal for small numbers of sensors. In this work we assumed that the relationship between input and output is given by an explicit linear integral operator. Many real systems are governed by nonlinear partial differential equations, and we may be unable to write an explicit formula for the dependence of output on input. For nonlinear systems, computing posterior distributions is hard even under the assumption of Gaussian prior and Gaussian measurement error; this can in a first stage be fixed with a Laplacian approximation [2]. Another limitation is our assumption that the prior covariance matrix is a multiple of the identity (which we need in order to obtain a nondegenerate posterior density). While this choice, the original Tikhonov regularization, is a common one in inverse problems [12,18], in many cases one would use a different prior or regularization operator, such as the discrete Laplacian [12]. Future work will address these limitations.

Acknowledgements This material was based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR) under Contract DE-AC02-06CH11347. We acknowledge partial NSF funding through awards FP061151-01-PR and CNS-1545046. We thank Noemi Petra and Victor Zavala for feedback on this topic. We also thank the two anonymous referees whose comments have helped significantly improve the paper.

A Other rounding strategies

A.1 Another sum-up rounding procedure for rectangular domains

We present the *sum-up rounding algorithm II* based on the following compatible two-level decomposition, with concepts defined in Definition 1. We use the notation

$$k_1(n_i) = \lfloor \sqrt{n_i} \rfloor, \quad \text{and} \quad k(n) = k_1(n_1)k_1(n_2)..k_1(n_P).$$

(i) On $[l_1^i, l_2^i]$ for $i = 1, 2, \ldots, P$, group the first $k_1(n_i)$ intervals $\{\mathcal{I}_{i,j}\}_{j=1}^{k_1(n_i)}$ as $\mathcal{G}_{i,1}$, group the next $k_1(n_i)$ intervals $\{\mathcal{I}_{i,j}\}_{j=k_1(n_i)+1}^{2k_1(n_i)}$ as $\mathcal{G}_{i,2}$, and so forth until we get $\mathcal{G}_{i,k_1(n_i)}$. The remaining intervals $\{\mathcal{I}_{i,j}\}_{j=k_1(n_i)^2+1}^{n}$ are grouped as $\mathcal{G}_{i,last}$, and the number of intervals in $\mathcal{G}_{i,last}$ equals $n_i - k_1(n_i)^2$. Note that

$$\sqrt{n_i} - 1 < k_1(n_i) = \lfloor \sqrt{n_i} \rfloor \le \sqrt{n_i}.$$

We can bound the number of intervals in the last group by

$$n_i - (\sqrt{n_i})^2 \le n_i - k_1(n_i)^2 < n_i - (\sqrt{n_i} - 1)^2$$
$$0 \le n_i - k_1(n_i)^2 < 2\sqrt{n_i},$$

so the cardinality of $\mathcal{G}_{i,.}$ is $\mathcal{O}(\sqrt{n_i})$, and its size is $\mathcal{O}(1/\sqrt{n_i})$.

(ii) Consider a subdomain V_j of the form

$$\prod_{\substack{i=1,2,\ldots,P \\ j_i \in \{1,2,\ldots,k_1(n_i),last\}}} \mathcal{G}_{i,j_i}.$$

This decomposition has the following parameters and properties, in reference to Definition 1.

$$k(n) = \prod_{i=1}^{P} \lfloor \sqrt{n_i} \rfloor, \;\; \tilde{k}(n) = \prod_{i=1}^{P} \lceil \sqrt{n_i} \rceil, \;\; r(n) = k(n), \tag{56}$$

$$\rho(V_j) = \sqrt{\sum_{i=1}^{P} \left(\frac{(l_2^i - l_1^i)}{\lfloor \sqrt{n_i} \rfloor} \right)^2}, \quad j = 1, 2, \ldots, k(n) \tag{57}$$

Theorem 5 *Under the assumptions of Theorem 1, there exists a C such that the sum-up rounding algorithm II construction satisfies*

$$\left| \sum_{k=1}^{n} f(x_k) \left(w^n(x_k) - \tilde{w}^n(x_k) \right) \Delta_x \right| \le \frac{C}{n^{1/2P}}.$$

Proof We use the definitions of the sum-up rounding procedure parameters (56)–(57), and the inequalities (21)–(22) to infer the following inequalities:

$$\frac{1}{\sqrt{n_i}} \le c_1^{-\frac{1}{2}} n^{-\frac{1}{2P}}, \; i = 1, 2, \ldots, P; \quad \frac{1}{r(n)} = \prod_{i=1}^{P} \frac{1}{\lfloor \sqrt{n_i} \rfloor} \overset{(21)}{\le} \frac{2^{\frac{P}{2}}}{\sqrt{n}}. \tag{58}$$

For the maximum diameter of V_j we obtain from (57) and (21)

$$\max_{j=1,2,\ldots,k(n)} \rho(V_j) \le \sqrt{P} \frac{\max_{i=1,2,\ldots,P} (l_2^i - l_1^i)}{\frac{1}{2} \min_{i=1,2,\ldots,P} \sqrt{n_i}}$$
$$\overset{(22)}{\le} \sqrt{P} \frac{\max_{i=1,2,\ldots,P} (l_2^i - l_1^i)}{\frac{1}{2} \sqrt{c_1}} n^{-\frac{1}{2P}}. \tag{59}$$

We also obtain

$$1-\frac{k(n)r(n)}{n}=1-\prod_{i=1}^{P}\frac{\lfloor\sqrt{n_i}\rfloor^2}{n_i}\le 1-\prod_{i=1}^{P}\left(1-\frac{2}{\sqrt{n_i}}\right)\overset{(22)}{\le} 1-\left(1-2c_1^{-\frac{1}{2}}n^{-\frac{1}{2P}}\right)^{P}.$$

In turn, from the mean value theorem applied to $(1-x)^P$ for $x\in[0,1]$ and the last inequality, we have

$$1-(1-x)^P\le Px,\ \forall x\in[0,1]\Rightarrow 1-\frac{k(n)r(n)}{n}\le 2Pc_1^{-\frac{1}{2}}n^{-\frac{1}{2P}}. \tag{60}$$

We now use Theorem 1 along with (21)–(22), (58), (59), and (60) to obtain the statement of this theorem for the *sum-up rounding algorithm* II with the choice

$$C=\max_{x\in V}|f(x)|\mu(V)2^{\frac{P}{2}}+2L\mu(V)\sqrt{P}\frac{\max_{i=1,2,\dots,P}(l_2^i-l_1^i)}{\frac{1}{2}\sqrt{c_1}}$$
$$+4\max_{x\in V}|f(x)|\mu(V)Pc_1^{-\frac{1}{2}}.$$

References

1. Alexanderian, A., Petra, N., Stadler, G., Ghattas, O.: A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems with regularized l_0-sparsification. SIAM J. Sci. Comput. **36**, A2122–A2148 (2014)
2. Alexanderian, A., Petra, N., Stadler, G., Ghattas, O.: A fast and scalable method for a-optimal design of experiments for infinite-dimensional Bayesian nonlinear inverse problems. SIAM J. Sci. Comput. **38**(1), A243–A272 (2016)
3. Atkinson, K., Han, W.: Theoretical Numerical Analysis, vol. 39. Springer, New York (2005)
4. Balas, E., Ceria, S., Dawande, M., Margot, F., Pataki, G.: Octane: a new heuristic for pure 0–1 programs. Operat. Res. **49**, 207–225 (2001)
5. Berry, J., Hart, W.E., Phillips, C.A., Uber, J.G., Watson, J.P.: Sensor placement in municipal water networks with temporal integer programming models. J. Water Resour. Plan. Manag. **132**, 218–224 (2006)
6. Berthold, T.: RENS–the optimal rounding. Math. Prog. Comput. **6**, 33–54 (2014)
7. Blacker, T.: Meeting the challenge for automated conformal hexahedral meshing. In: 9th International Meshing Roundtable, pp. 11–20 (2000)
8. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
9. Cox, D.R., Reid, N.: The Theory of the Design of Experiments. Chapman & Hall/CRC, Boca Raton (2000)
10. Cressie, N., Wikle, C.K.: Statistics for Spatio-Temporal Data. Wiley, New York (2015)
11. Drăgănescu, A.: Multigrid preconditioning of linear systems for semi-smooth newton methods applied to optimization problems constrained by smoothing operators. Optim. Methods Softw. **29**(4), 786–818 (2014)
12. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems, vol. 375. Springer, New York (1996)
13. Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning, vol. 1. Springer, New York (2001)
14. Geoga, C.J., Anitescu, M., Stein, M.L.: Scalable Gaussian process computations using hierarchical matrices (2018). arXiv preprint, arXiv:1808.03215
15. Hammer, P.L., Johnson, E.L., Peled, U.N.: Facet of regular 0–1 polytopes. Math. Program. **8**, 179–206 (1975)

16. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1990)
17. Iyengar, S.S., Brooks, R.R.: Distributed Sensor Networks, Second Edition: Sensor Networking and Applications. Chapman and Hall/CRC, Boca Raton (2016)
18. Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems, vol. 160. Springer, New York (2006)
19. Kendall, E.A.: The Numerical Solution of Integral Equations of the Second Kind. Cambridge University Press, Cambridge (1997)
20. Kirsch, A.: An Introduction to the Mathematical Theory of Inverse Problems. Springer, New York (2011)
21. Krause, A., Leskovec, J., Guestrin, C., VanBriesen, J., Faloutsos, C.: Efficient sensor placement optimization for securing large water distribution networks. J. Water Resour. Plan. Manag. **134**, 516–526 (2008)
22. Lodi, A., Bonami, P., Cornuéjols, G., Margot, F.: A feasibility pump for mixed integer nonlinear programs. Math. Program. **119**, 331–352 (2009)
23. Nannicini, G., Belotti, P.: Rounding-based heuristics for nonconvexminlps. Math. Program. Comput. **4**, 1–31 (2012)
24. Patera, A.T.: A Spectral Element Method for Fluid Dynamics: Laminar Flow in a Channel Expansion, vol. 54. Elsevier, Amsterdam (1984)
25. Pukelsheim, F.: Optimal Design of Experiments. Classics in Applied Mathematics, vol. 50. SIAM (2006)
26. Sager, S.: Sampling decision in optimum experimental design in the light of Pontryagin's maximum principle. SIAM J. Control Optim. **51**, 3181–3207 (2013)
27. Sager, S., Bock, H.G., Diehl, M.: The integer approximation error in mixed-integer optimal control. Math. Program. Ser. A **133**, 1–23 (2012)
28. Wang, Y., Yagola, A.G., Yang, C.: Computational Methods for Applied Inverse Problems. Higher Education Press, Beijing (2012)
29. Watson, J.P., Greenberg, H.J., Hart, W.E.: A multiple-objective analysis of sensor placement optimization in water networks. In: Proceedings of the World Water and Environment Resources Congress. American Society of Civil Engineers (2004)
30. Welch, W.J.: Algorithmic complexity: three NP-hard problems in computational statistics. J. Stat. Comput. Simul. **15**(1), 17–25 (1982)
31. Wielandt, H.: Error bounds for eigenvalues of symmetric integral equations. Proc. Sympos. Appl. Math **6**, 261–282 (1956)
32. Wolsey, L.A.: Faces for a linear inequality in 0–1 variables. Math. Program. **8**, 165–178 (1975)
33. Yu, J., Zavala, V.M., Anitescu, M.: A scalable design of experiments framework for optimal sensor placement. J. Process Control **67**, 44–55 (2017)
34. Zhang, Yongjie, Bajaj, Chandrajit: Adaptive and quality quadrilateral/hexahedral meshing from volumetric data. Comput. Methods Appl. Mech. Eng. **195**(9–12), 942–960 (2006)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Jing Yu[1] **· Mihai Anitescu**[2]

Jing Yu
jingyu@galton.uchicago.edu

[1] Department of Statistics, Physical Sciences Division, The University of Chicago,
5747 S. Ellis Ave., Chicago, IL 60637, USA

[2] Mathematics and Computer Science Division, Argonne National Laboratory,
9700 S. Cass Ave., Lemont, IL 60439, USA