Metrika (2023) 86:781–804
https://doi.org/10.1007/s00184-023-00893-6

Robust optimal designs using a model misspecification term

Renata Eirini Tsirpitzi[1] · **Frank Miller**[1,2] · **Carl-Fredrik Burman**[3,4]

Received: 12 March 2022 / Accepted: 6 January 2023 / Published online: 8 February 2023

Abstract

Much of classical optimal design theory relies on specifying a model with only a small number of parameters. In many applications, such models will give reasonable approximations. However, they will often be found not to be entirely correct when enough data are at hand. A property of classical optimal design methodology is that the amount of data does not influence the design when a fixed model is used. However, it is reasonable that a low dimensional model is satisfactory only if limited data is available. With more data available, more aspects of the underlying relationship can be assessed. We consider a simple model that is not thought to be fully correct. The model misspecification, that is, the difference between the true mean and the simple model, is explicitly modeled with a stochastic process. This gives a unified approach to handle situations with both limited and rich data. Our objective is to estimate the combined model, which is the sum of the simple model and the assumed misspecification process. In our situation, the low-dimensional model can be viewed as a fixed effect and the misspecification term as a random effect in a mixed-effects model. Our aim is to predict within this model. We describe how we minimize the prediction error using an optimal design. We compute optimal designs for the full model in different cases. The results confirm that the optimal design depends strongly on the sample size. In low-information situations, traditional optimal designs for models with a small number of parameters are sufficient, while the inclusion of the misspecification term lead to very different designs in data-rich cases.

Renata Eirini Tsirpitzi, Frank Miller and Carl-Fredrik Burman have contributed equally to this work

✉ Renata Eirini Tsirpitzi
renatatsirpi@gmail.com

[1] Department of Statistics, Stockholm University, SE-10691 Stockholm, Sweden

[2] Department of Computer and Information Science, Linköping University, SE-581 83 Linköping, Sweden

[3] Early Biometrics and Statistical Innovation, Data Science and Artificial Intelligence, AstraZeneca, Gothenburg, Sweden

[4] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Keywords Fedorov algorithm · Gaussian process · Mixed-effects model · Optimal experimental design · Statistical modelling

1 Introduction

The emergence of Data Science partly inspires the current work. There is a great overlap between the fields of Statistics and Data Science and there is no consensus view on how exactly to define them or the difference between them. However, one important difference is that Data Science often applies a high-dimensional parameter space to analyze very rich data. At the same time, (traditional) Statistics has often used pre-defined low-parameter models, for situations where information is limited. It is not the size of the sample *per se* that is important but rather the amount of information in relation to the magnitude of the signal in the data. In many applications, one faces both situations with limited and rich information. Therefore, unified approaches could be beneficial.

Statistics has been used extensively both for designed experiments and observational studies, while Data Science typically refers to the latter type of study. In general, while Optimal Design theory has a strong legacy in Statistics, the design of experiments is not a large part of Data Science methodology. Still, there are many situations where experiments can provide large data sets and strong signals. One example is web advertisements, see e.g. Kohavi et al. (2009) and Example 1.6 of Montgomery (2017). Such ads could be displayed to the customer for a different time duration, repeated with different frequency patterns, use different appearances, price deals, etc. The designer of an advertisement might be interested in the conversion rate, the percentage of visits to the website that includes a purchase (Kohavi et al. 2009), or in another performance metric. While the amount of data on customer behavior may be large, in terms of the number of customers and the number of ad showings, the results may have enough uncertainty to call for the optimal design of the experiment. Even small changes in a performance metric like conversion rate can have a huge impact on revenues (Kohavi et al. 2013). Fernandez-Tapia and Guéant (2017) deal with bidding models for ad-inventory auctions using a Poisson process and demonstrate that the Hamilton-Jacobi-Bellman equation describes the optimal bids. Mardanlou et al. (2017) propose an optimal cost design model that presents the connection between a campaign-level control signal and the total cost to the advertiser for the influence they had at each control signal value.

A sub-field of artificial intelligence, active machine learning, has been used over years in data science and is related to optimal experimental design. If a learning algorithm can choose the data it wants to learn from, it can perform better with less data for training. The first applied statistical analyzes of active learning for regression in robot arm kinematics was presented by Cohn (1996) and Cohn et al. (1996). Important areas of application include speech recognition, image or video classification, or information extraction from texts, see Settles (2010) for a review. Nie et al. (2018) focus on active learning for regression models including a model misspecification term. López-Fidalgo and Wiens (2022) recommend methods for binary data for esti-

mation and classification in active learning which are robust in case of both model misspecification and response mislabelling.

When analyzing data, we may either use a multi-purpose model or a scientifically based model. Commonly used multi-purpose models include polynomial models, such as linear and quadratic, and generalized linear models. When possible, it is often useful to base a model on subject matter science. As an example, we may study the dose-dependent effect of drugs. In this context, the Michaelis-Menten model can be derived theoretically for certain simple concentration-response experiments. The Michaelis-Menten is often generalized to an Emax model when studying dose-response based on results from patients. While the Michaelis-Menten model has two parameters for location and maximal response, the standard 4-parameter Emax model includes additional free parameters for the placebo efficacy and the degree of sigmoidicity of the response curve. The Emax model has been very used and useful for clinical trials. However, for several theoretical reasons, it can not be exact. Assume, for example, that individual patient's concentration-response curve follows an Emax model. Late-stage clinical trials often aim at modeling *dose*-response, not measuring plasma concentration. Variations in concentration will therefore distort the population model. Patient heterogeneity in Emax parameters will also lead to distortion. In addition, the Michaelis-Menten model may hold for receptor binding. However, the relation between receptor binding and the clinical endpoint, e.g. blood pressure, glucose, or FEV1, is likely a complex non-linear function. Some of these distortions are possible to include in a more sophisticated scientific model. However, other kinds of distortions, like the relation between receptor binding and response, are likely far too deep to capture in a low-parameter model. Similarly, the relation between price or display time and a customer response like conversion rate will likely not exactly follow a simple model.

We will therefore combine a low-dimensional model with a term that models the misspecification. A simple example can be a linear (straight-line) regression combined with a Brownian bridge as a (modeled) misspecification term. Our objective is to estimate the combined model, including the misspecification term, based on the collected data. The experimental conditions used for data collection should be chosen to optimize the precision of the estimated model.

Optimal designs for models with a stochastic process as error term have been considered e.g. by Sacks and Ylvisaker (1966); Harman and Štulajter (2011); Dette et al. (2016, 2017). However, in contrast to these situations, our stochastic process (the misspecification term) contributes to the regression function of interest and is not an error term.

In our situation, the low-dimensional regression model is a fixed effect, and the misspecification term is a random effect in a mixed-effects model. Since we are here interested in estimating the combination of fixed and random effects, our aim is to predict within the mixed model. Optimal designs for prediction in mixed models have been considered by Prus and Schwabe (2016), Liu et al. (2019), Prus (2020). Fedorov and Jones (2005) describe how a mixed-effects model can be used to determine designs for multicenter clinical trials. If we use in our model a Gaussian process as stochastic process, the model considered here is related to non-parametric Gaussian processes which is used in machine learning (Williams and Rasmussen 2006). A

Bayesian description of these models dates back to the 1970's (Blight and Ott 1975) and also some design optimization has been considered from a Bayesian perspective (O'Hagan 1978). This framework has been used in multiple applications, see e.g. Siivola et al. (2021). The optimal design methodology which we consider is, however, different from the one used by O'Hagan (1978).

The general model set-up and the methods used for prediction and design optimization will be specified in Sect. 2. In Sect. 2.1, we derive how to estimate the response over the whole design space, and the structure of the covariance estimate is presented in Sect. 2.2. We will present the optimized specific model in Sect. 3. In Sect. 4, we will present the optimality criterion used and the results extracted by applying Fedorov's algorithm. It is shown that even if the parameters of the low-dimensional model are known (local optimal design), the optimal design will depend on the sample size, which comes in contrast to standard optimal design theory. In practice, our optimal design will typically be very similar to the traditional optimal design (without a misspecification term) when the sample size is small. However, with increasing sample size, the new optimal design will zoom in on the most important area. Depending on the optimal design criterion, this may e.g. be where the expected response has a certain target value.

2 Inference

2.1 The general model

We assume that we observe N observations Y_i in an experiment which follow

$$Y_i = \mu(x_i) + e_i, \qquad i = 1, \ldots, N,$$

where e_i are uncorrelated random variables with $E(e_i) = 0$ and $Var(e_i) = \sigma^2 > 0$ and x_i are elements of a compact one- or multidimensional design space $\mathcal{X}$.

We consider a simple, low-dimensional approximate model $\nu(x)$ for the mean response. As this model is likely not fully correct, we consider the misspecification function $\mu(x) - \nu(x)$. We will explicitly model the misspecification function with a zero-mean stochastic process $C(X)$, i.e. we assume here that the true mean is $\mu(x) = \nu(x) + C(x)$. The discrepancy $C(x)$ can be seen as a random effect. We can view the model including C as a vehicle to construct robust designs and shed light on the question about when and how much to rely on the approximate simple model ν.

We assume further that $\nu(x) = \mathbf{f}(x)^\top \boldsymbol{\beta}$. Here, $\mathbf{f}(x)^\top \boldsymbol{\beta}$ is a linear regression model with a d-dimensional unknown parameter vector $\boldsymbol{\beta} \in \mathbb{R}^d$ and a d-dimensional known regression function $\mathbf{f} : \mathcal{X} \to \mathbb{R}^d$. So, μ is a mixture of a fixed and random part:

$$\mu(x) = \mathbf{f}(x)^\top \boldsymbol{\beta} + C(x). \tag{1}$$

In the random part, $C(x)$ is assumed to be a stochastic process with zero mean, $E\{C(x)\} = 0, x \in \mathcal{X}$, and existing second moments ($E|C(x)C(y)| < \infty, x, y \in \mathcal{X}$). It is assumed to be independent of $\mathbf{e} = (e_1, \ldots, e_N)^\top$.

The interpretation of (1) is that our model is assumed to be a linear regression model $\mathbf{f}(x)^\top \boldsymbol{\beta}$ plus an additional misspecification term C. We know the functional shape of the model only up to this misspecification term. In this paper, we are interested in predicting the whole $\mu(x)$ on $\mathcal{X}$ and not only the linear regression term, since the misspecification correction part belongs to our model of interest.

Often, one will observe several times at some x. Let $m+1$ be the number of distinct design points and we denote them by $\tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_m \in \mathcal{X}$. If the design space is one-dimensional, we assume that the x_i and $\tilde{x}_j$ are sorted ascendingly and that the number of observations made at $\tilde{x}_j$ is n_j, $j = 0, \ldots, m$, i.e. $\sum_{j=0}^{m} n_j = N$. Our model can be written as mixed-effects model. With $C_j = C(\tilde{x}_j)$ and $\mathbf{Y} = (Y_1, \ldots, Y_N)^\top$, we write

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{e}, \tag{2}$$

where $\mathbf{X} = (\mathbf{f}(x_1), \ldots, \mathbf{f}(x_N))^\top$, $\mathbf{Z} = (z_{ij})_{i=1,\ldots,N,\, j=0,\ldots,m} \in \mathbb{R}^{N\times(m+1)}$ with $z_{ij} = 1$ if observation i is made using design point $\tilde{x}_j$ and $z_{ij} = 0$ otherwise, and $\boldsymbol{\gamma} = (C_0, C_1, \ldots, C_m)^\top$ is a vector of random effects of the mixed model. Let $\mathbf{D}$ be the covariance matrix of $\boldsymbol{\gamma}$ and $\mathbf{R}$ be the covariance matrix of $\boldsymbol{e}$, i.e. $\mathbf{R} = \sigma^2\mathbf{I}_N$ with $\mathbf{I}_N$ being the N-dimensional identity matrix. This implies $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, $Cov(\mathbf{Y}) = \mathbf{Z}\mathbf{D}\mathbf{Z}^\top + \mathbf{R}$.

We write

$$\tilde{\boldsymbol{\mu}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$$

for the N-dimensional mean vector conditional on γ. The vector of conditional means at the $m+1$ points of observation is denoted by

$$\boldsymbol{\mu} = (\mu(\tilde{x}_0), \mu(\tilde{x}_1), \ldots, \mu(\tilde{x}_m))^\top.$$

We have $\tilde{\boldsymbol{\mu}} = \mathbf{Z}\boldsymbol{\mu}$.

Let $\bar{\mathbf{Y}}$ be the $(m+1)$-dimensional vector of mean observations at $\tilde{x}_j$, $j = 0, \ldots m$, i.e. $\bar{\mathbf{Y}} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{Y}$. Writing $\mathbf{W} = (\mathbf{f}(\tilde{x}_0), \ldots, \mathbf{f}(\tilde{x}_m))^\top$, we have then $\mathbf{X} = \mathbf{Z}\mathbf{W}$ and the model for the mean observations is $\boldsymbol{\mu} = \mathbf{W}\boldsymbol{\beta} + \boldsymbol{\gamma}$. Further, we have

$$\bar{\mathbf{Y}} = \mathbf{W}\boldsymbol{\beta} + \boldsymbol{\gamma} + \bar{\mathbf{e}} \tag{3}$$

with $\bar{\mathbf{e}} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{e}$. The components of the vector $\bar{\mathbf{e}}$ have expected value 0 and variance σ^2/n_j, $j = 0, \ldots, m$. Therefore, the distribution of $\bar{\mathbf{Y}}$ depends on σ^2 and n_j only through σ^2/n_j, $j = 0, \ldots, m$. We have $\mathbf{D}_{\bar{\mathbf{Y}}} := Cov(\bar{\mathbf{Y}}) = \mathbf{D} + \sigma^2\mathrm{diag}(n_0^{-1}, \ldots, n_m^{-1})$.

2.2 Inference results of BLUE and BLUP

In mixed-effects models, the interest is to estimate $\boldsymbol{\beta}$ and to predict $\boldsymbol{\gamma}$. A Best Linear Unbiased Estimate (BLUE) for $\boldsymbol{\beta}$ and a Best Linear Unbiased Predictor (BLUP) for $\boldsymbol{\gamma}$ are well known, see Christensen (2002). In the following lemma, we show how they

can be computed specifically for our model. We provide formulae both based on $\mathbf{Y}$ and $\bar{\mathbf{Y}}$.

Let $\mathbf{M} = (\mathbf{Z}\mathbf{D}\mathbf{Z}^\top + \mathbf{R})^{-1}$ which is a symmetric $N \times N$-matrix, and let $\mathbf{I}_d$ be the identity matrix of dimension $d \times d$.

Lemma 2.1 *1. The Best Linear Unbiased Estimator (BLUE) of $\boldsymbol{\beta}$ is*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}\mathbf{Y} = (\mathbf{W}^\top\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}\bar{\mathbf{Y}}.$$

2. The Best Linear Unbiased Predictor (BLUP) is

$$\begin{aligned}\hat{\boldsymbol{\gamma}} &= \mathbf{D}\mathbf{Z}^\top\mathbf{M}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{D}\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}[\mathbf{I}_{m+1} - \mathbf{W}(\mathbf{W}^\top\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}]\bar{\mathbf{Y}}.\end{aligned}$$

3. The Best Linear Unbiased Predictor for $\boldsymbol{\mu}$ is

$$\begin{aligned}\hat{\boldsymbol{\mu}} &= [\mathbf{Z}\mathbf{D}\mathbf{Z}^\top\mathbf{M} + (\mathbf{I}_{m+1} - \mathbf{Z}\mathbf{D}\mathbf{Z}^\top\mathbf{M})\mathbf{X}(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}]\mathbf{Y} \\ &= \left[\mathbf{D}\mathbf{D}_{\bar{\mathbf{Y}}}^{-1} + (\mathbf{I}_{m+1} - \mathbf{D}\mathbf{D}_{\bar{\mathbf{Y}}}^{-1})\mathbf{W}(\mathbf{W}^\top\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}\right]\bar{\mathbf{Y}}.\end{aligned}$$

4. Let $x^* \in [0, 1]$ be an arbitrary point where one wants to predict μ. Define the following functions of x: $\mathbf{W}_*(x) = \mathbf{f}(x)^\top$ and the covariance vector $\mathbf{D}_*(x) = (Cov\{C(\tilde{x}_0), C(x)\}, \ldots, Cov\{C(\tilde{x}_m), C(x)\})^\top$. The Best Linear Unbiased Predictor for $\mu(x^*)$ is

$$\hat{\mu}(x^*) = \mathbf{D}_*^\top(x^*)\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}\bar{\mathbf{Y}} + (\mathbf{W}_*(x^*) - \mathbf{W}\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}\mathbf{D}_*(x^*))^\top\hat{\boldsymbol{\beta}}. \tag{4}$$

The proof is available in the Appendix.

In the case of large variance σ^2 or alternatively in the case of small sample sizes n_i, the random effect part in the mixed model becomes unimportant. Therefore, $\hat{\boldsymbol{\beta}}$ converges to the ordinary least squares estimator $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ and $\hat{\boldsymbol{\gamma}}$ converges to the zero vector $\mathbf{0}_{m+1}$. This can formally be seen based on Lemma 2.1 since $\sigma^2\mathbf{M} \to \mathbf{I}_N$ for $\sigma^2 \to \infty$.

For design optimization, we need expressions for covariance matrices. In Lemma A.1 in the appendix, we show them for the BLUE and the BLUP. Since we aim to optimize the design by minimizing the uncertainty in prediction, the following theorem is important for our elaborations in Sect. 4.

Theorem 2.2 *The covariance of the prediction error $\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}$ is:*

$$\begin{aligned}Cov(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) &= \big(\mathbf{L}\mathbf{X}^\top\mathbf{M}\mathbf{Z} + \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{Z} - \mathbf{I}_{m+1}\big)\mathbf{D}\big(\mathbf{L}\mathbf{X}^\top\mathbf{M}\mathbf{Z} + \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{Z} - \mathbf{I}_{m+1}\big)^\top \\ &\quad + \mathbf{L}\mathbf{X}^\top\mathbf{M}\mathbf{R}\mathbf{M}\mathbf{X}\mathbf{L}^\top + \mathbf{L}\mathbf{X}^\top\mathbf{M}\mathbf{R}\mathbf{M}\mathbf{Z}\mathbf{D} \\ &\quad + \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{R}\mathbf{M}\mathbf{X}\mathbf{L}^\top + \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{R}\mathbf{M}\mathbf{Z}\mathbf{D}\end{aligned}$$

with $\mathbf{L} = (\mathbf{W} - \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{X})(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}$. *The variance of the prediction error* $\hat{\mu}(x^*) - \mu(x^*)$ *at an arbitrary point* $x^* \in [0, 1]$ *is:*

$$\begin{aligned} &Var(\hat{\mu}(x^*) - \mu(x^*)) \\ &\quad = Var(C(x^*)) - \mathbf{D}_*(x^*)^\top\mathbf{D}_{\tilde{\mathbf{Y}}}^{-1}\mathbf{D}_*(x^*) + \mathbf{S}(x^*)^\top(\mathbf{W}^\top\mathbf{D}_{\tilde{\mathbf{Y}}}^{-1}\mathbf{W})^{-1}\mathbf{S}(x^*) \end{aligned}$$

with $\mathbf{S}(x) = \mathbf{W}_*(x) - \mathbf{W}\mathbf{D}_{\tilde{\mathbf{Y}}}^{-1}\mathbf{D}_*(x)$.

The proof can be found in the Appendix. Note that the diagonal entry belonging to x_i in the first expression, $Cov(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$, is equal to the value of the second expression, $Var(\hat{\mu}(x^*) - \mu(x^*))$, at $x^* = \tilde{x}_i$.

3 A specific model as example

3.1 The specific model

To illustrate our model, we use the following specific model. We consider straight line regression on $\mathcal{X} = [0, 1]$, centered at $1/2$: $\mathbf{f}(x) = (1, x - 1/2)^\top$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$. As misspecification term, we use the Brownian bridge $B(x)$ on $\mathcal{X}$, scaled with factor τ, $C(x) = \tau B(x)$. See Ross et al. (1996) and Chow (2009) for definition and properties of the Brownian bridge.

Then,

$$\begin{aligned} \mathbf{X} &= \begin{pmatrix} 1 & \cdots & 1 \\ x_1 - \frac{1}{2} & \cdots & x_N - \frac{1}{2} \end{pmatrix}^\top \\ &= \begin{pmatrix} 1 & \cdots & 1 & \cdots & \cdots & \cdots & 1 & \cdots & 1 \\ \tilde{x}_0 - \frac{1}{2} & \cdots & \tilde{x}_0 - \frac{1}{2} & \cdots & \cdots & \cdots & \tilde{x}_m - \frac{1}{2} & \cdots & \tilde{x}_m - \frac{1}{2} \end{pmatrix}^\top, \end{aligned}$$

and $\mathbf{D} = \tau^2(\tilde{x}_i\tilde{x}_j - \min\{\tilde{x}_i, \tilde{x}_j\})_{i=0,\ldots,m,\, j=0,\ldots,m}$. We will use $\tau^2 = 1$ for several illustrations in the sequel.

The matrix $\mathbf{W}$ in the model (3) for the vector of means $\tilde{\mathbf{Y}}$ is the $(m + 1) \times 2$-dimensional matrix

$$\mathbf{W} = \begin{pmatrix} 1 & \cdots & 1 \\ \tilde{x}_0 - \frac{1}{2} & \cdots & \tilde{x}_m - \frac{1}{2} \end{pmatrix}^\top.$$

3.2 Numerical interpretation of the prediction

In order to better understand the impact of adding more information to the estimators and predictors, we present in this subsection a numerical illustration of them for the case of $m = 2$ and $m = 4$ for equidistantly spaced observations $\tilde{x}_i = i/m, i = 0, \ldots, m$.

In Table 1 and 2, we apply the same sample size for $m = 2$ in all design points, so $n_0 = n_1 = n_2$, and we want to investigate the effect of increasing N. Based on the

Table 1 For $m = 2$ the BLU-estimator $\hat{\beta}$

N	n_0, n_1, n_2	$\hat{\beta}_1$	$\hat{\beta}_2$
0.03	0.01, 0.01, 0.01	$0.3336\bar{Y}_0 + 0.3328\bar{Y}_1 + 0.3336\bar{Y}_2$	$-\bar{Y}_0 + \bar{Y}_2$
3	1, 1, 1	$0.3572\bar{Y}_0 + 0.2856\bar{Y}_1 + 0.3572\bar{Y}_2$	$-\bar{Y}_0 + \bar{Y}_2$
30	10, 10, 10	$0.4375\bar{Y}_0 + 0.125\bar{Y}_1 + 0.4375\bar{Y}_2$	$-\bar{Y}_0 + \bar{Y}_2$
300	100, 100, 100	$0.4906\bar{Y}_0 + 0.0188\bar{Y}_1 + 0.4906\bar{Y}_2$	$-\bar{Y}_0 + \bar{Y}_2$
30000	10000, 10000, 10000	$0.4999\bar{Y}_0 + 0.0002\bar{Y}_1 + 0.4999\bar{Y}_2$	$-\bar{Y}_0 + \bar{Y}_2$

Table 2 For $m = 2$ the BLU-predictor $\hat{\gamma}$

N	n_0, n_1, n_2	$\hat{B}_0$	$\hat{B}_1$	$\hat{B}_2$
0.03	0.01, 0.01, 0.01	0	$-0.0008\bar{Y}_0 + 0.0016\bar{Y}_1 - 0.0008\bar{Y}_2$	0
3	1, 1, 1	0	$-0.0714\bar{Y}_0 + 0.1428\bar{Y}_1 - 0.0714\bar{Y}_2$	0
30	10, 10, 10	0	$-0.3125\bar{Y}_0 + 0.625\bar{Y}_1 - 0.3125\bar{Y}_2$	0
300	100, 100, 100	0	$-0.4717\bar{Y}_0 + 0.9434\bar{Y}_1 - 0.4717\bar{Y}_2$	0
30000	10000, 10000, 10000	0	$-0.4997\bar{Y}_0 + 0.9994\bar{Y}_1 - 0.4997\bar{Y}_2$	0

results of $\hat{\beta}_1$, the intercept of the linear regression, we see that while the weight of all $\bar{Y}_i$ are equal when N is almost 0, the weight of $\bar{Y}_1$ decreases while N increases until the weight is close to 0 when N tends to infinity. Thus, the information is spread equally between the three weights when N is small and we end up with all the information shared by the extreme values when N is large. Regarding the weights for $\hat{B}_1$, we detect that the weight of the middle design point $\bar{Y}_1$ increases while N increases and we have the corresponding weight to be close to 1 when N tends to infinity. Since the sum of weights of $\hat{B}_1$ is equal to 0, the increase of the middle point weight decreases the weight of the extreme points. Thus, by starting with all the weights to be close to 0 when N is almost zero, we get the weight of $\bar{Y}_1$ to be 1 while the weights of $\bar{Y}_0$ and $\bar{Y}_2$ are -0.5 when N is large. Another way to explain the two tables is by seeing the impact of the amount of data in the use of the misspecification term. When there is less data (N is small), we obtain all the information almost exclusively from the simple linear regression. On the other hand, when we have a lot of data (N tends to infinity), the regression includes the Brownian bridge handling the big amount of information.

The results in Table 3 and 4 give us the estimators and the predictors when $m = 4$, so our n-vector consists of 5 elements. We obtain symmetry between the four extreme $\bar{Y}_i$ in the case of the intercept $\hat{\beta}_1$. So the weight of $\bar{Y}_0$ equals the weight of $\bar{Y}_4$ and the one of $\bar{Y}_1$ is the same as the weight of $\bar{Y}_3$. On the other hand, for the slope $\hat{\beta}_2$ we have negative symmetry between $\bar{Y}_0$ and $\bar{Y}_4$ and between $\bar{Y}_1$ and $\bar{Y}_3$. The weights of the intercept sum to 1, while the one of the slope sum to 0.

Since our misspecification term is a Brownian bridge, we get $\hat{B}_0$ and $\hat{B}_4$ equal to zero. There is a symmetry between $\hat{B}_1$ and $\hat{B}_3$. The weight that corresponds to $\bar{Y}_0$ in the case of $\hat{B}_1$ is the same as the weight of $\bar{Y}_4$ for $\hat{B}_3$. In the same way the weight of $\bar{Y}_1$ for $\hat{B}_1$ match with the one of $\bar{Y}_3$ for $\hat{B}_3$. And vice versa, so $\bar{Y}_0$ and $\bar{Y}_1$ for $\hat{B}_3$ equals to

Table 3 For $m = 4$ the BLU-estimator $\hat{\beta}$

N	n_0, n_1, n_2, n_3, n_4	$\hat{\beta}_1$ $\hat{\beta}_2$
10	3, 1, 2, 1, 3	$0.3537\bar{Y}_0 + 0.0813\bar{Y}_1 + 0.13\bar{Y}_2 + 0.0813\bar{Y}_3 + 0.3537\bar{Y}_4$ $-0.931\bar{Y}_0 - 0.1379\bar{Y}_1 + 0\bar{Y}_2 + 0.1379\bar{Y}_3 + 0.931\bar{Y}_4$
100	30, 10, 20, 10, 30	$0.4538\bar{Y}_0 + 0.0359\bar{Y}_1 + 0.0206\bar{Y}_2 + 0.0359\bar{Y}_3 + 0.4538\bar{Y}_4$ $-0.9642\bar{Y}_0 - 0.0714\bar{Y}_1 + 0\bar{Y}_2 + 0.0714\bar{Y}_3 + 0.9642\bar{Y}_4$

Table 4 For $m = 4$ the BLU-predictor $\hat{\gamma}$

N	n_0, n_1, n_2, n_3, n_4	$\hat{B}_0$	$\hat{B}_1$ $\hat{B}_2$ $\hat{B}_3$	$\hat{B}_4$
10	3, 1, 2, 1, 3	0	$-0.1355\bar{Y}_0 + 0.1126\bar{Y}_1 + 0.0975\bar{Y}_2 + 0.0092\bar{Y}_3 - 0.0838\bar{Y}_4$ $-0.1585\bar{Y}_0 + 0.03252\bar{Y}_1 + 0.252\bar{Y}_2 + 0.03252\bar{Y}_3 - 0.1585\bar{Y}_4$ $-0.0838\bar{Y}_0 + 0.0092\bar{Y}_1 + 0.0975\bar{Y}_2 + 0.1126\bar{Y}_3 - 0.1355\bar{Y}_4$	0
100	30, 10, 20, 10, 30	0	$-0.48\bar{Y}_0 + 0.537\bar{Y}_1 + 0.1538\bar{Y}_2 + 0.0013\bar{Y}_3 - 0.2122\bar{Y}_4$ $-0.423\bar{Y}_0 + 0.0512\bar{Y}_1 + 0.7435\bar{Y}_2 + 0.0512\bar{Y}_3 - 0.423\bar{Y}_4$ $-0.2122\bar{Y}_0 + 0.0013\bar{Y}_1 + 0.1538\bar{Y}_2 + 0.537\bar{Y}_3 - 0.48\bar{Y}_4$	0

$\bar{Y}_4$ and $\bar{Y}_3$ respectively for $\hat{B}_1$. The weight of $\bar{Y}_2$ remains the same for both predictors. The middle predictor $\hat{B}_2$ follows the same weight symmetry. As a result, we get the same weight for the pair $\bar{Y}_0$ and $\bar{Y}_4$, and $\bar{Y}_1$, $\bar{Y}_3$. As in the case of $m = 2$, the sum of the weights of the predictions is equal to 0.

In Table 3 and 4, we see also the effect of a 10-fold increase of N on estimators and predictors. The weight of the means $\bar{Y}_0$ and $\bar{Y}_4$ at the extreme points increases when N increases and that has, as a result, the weight of the means $\bar{Y}_1$, $\bar{Y}_2$ and $\bar{Y}_3$ at the middle points to decrease. Therefore, for $N = 100$, we can focus on the boundary points where we have no misspecification due to Brownian bridge. On the other hand, for a smaller sample size, the observations in the edges do not give much information. In contrast, the middle observations gather most of the information, even with the risk of getting bias.

Assume that we made observations with $\bar{\mathbf{Y}} = (0.1, 1.5, 2, 0.6, 2.2)^\top$, see the dots in Figure 1. Assume further that the number of observations was $(n_0, \ldots, n_m) = N \cdot (0.3, 0.1, 0.2, 0.1, 0.3)$ and that $\sigma^2 = 1$. We will illustrate now the prediction of $\mu(x)$, $x \in [0, 1]$ in Figure 1.

In a low information case with N close to 0, the prediction is following the maximum likelihood estimate for a straight line model (without misspecification term). In a high information case with N large, the prediction at $\tilde{x}_i$ becomes $\bar{Y}_i$. Between the observational points $\tilde{x}_i$, the prediction interpolates linearly; this can be seen from formula (4) since $\mathbf{W}_*$ is linear in x and $\mathbf{D}_*$ is linear on each interval between the $\tilde{x}_i$ for our specific example. When we have $N = 10$ or $N = 100$ (like in the previous

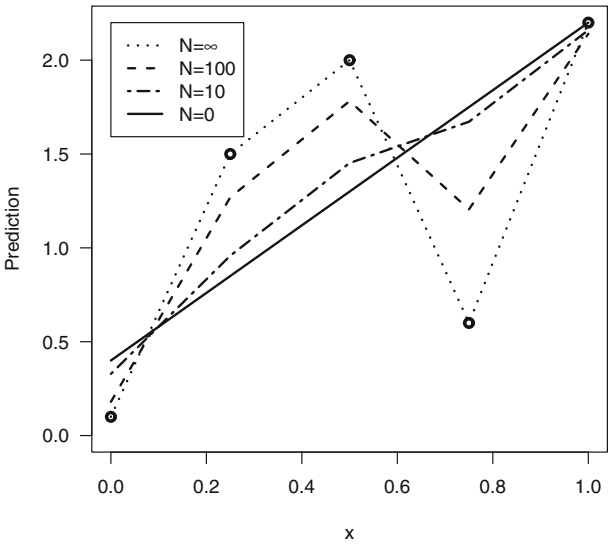

Fig. 1 Prediction of $\mu(x), x \in [0, 1]$ (lines) with observations $\bar{\mathbf{Y}}$ (dots); example: $m = 4$, $(x_0, \ldots, x_m) = (0, 1/4, 1/2, 3/4, 1)$, $(n_0, \ldots, n_m) = N \cdot (0.3, 0.1, 0.2, 0.1, 0.3)$ and $\bar{\mathbf{Y}} = (0.1, 1.5, 2, 0.6, 2.2)^\top$

example in Table 3 and 4), the prediction shrinks the high information curve towards the straight line regression.

4 Optimal designs

Based on the results in Sect. 2, we can predict $\boldsymbol{\mu}$ given data $\mathbf{Y}$. We can base this prediction on the BLUE and BLUP. When we have given data $\mathbf{Y}$, the formulae discussed in Sect. 2 depend on the design for the data collection which is specified by the number of observations $n_j \geq 0$ on each design point $\tilde{x}_j = j/m$, $j = 0, \ldots, m$. Now we consider the planning stage of an experiment where we have the possibility to choose the design, i.e. the numbers n_j. We want to choose a design for the experiment which allows the best possible prediction of the mean vector $\boldsymbol{\mu}$.

4.1 Experimental designs

Following standard notation (Atkinson et al. 2007, for example), a design is written as

$$\xi = \begin{Bmatrix} \tilde{x}_0 & \tilde{x}_1 & \cdots & \tilde{x}_m \\ n_0 & n_1 & \cdots & n_m \end{Bmatrix} \tag{5}$$

with $n_j \geq 0$, $j = 0, \ldots, m$, and $\sum_{j=0}^{m} n_j = N$ for a given total number of observations N. When n_i are only allowed to attain integer values, designs of the form (5) are called exact designs. Dropping this integer requirement and allowing non-negative values for n_j, designs are called approximate designs following the approach

of Kiefer (1974). We will consider later mainly approximate designs but also some exact designs in some examples. Since the quality of experimental designs for fixed effect models usually do not depend on the total number of observations N, designs are often described by weights $w_j = n_j/N$, only, and these weights form a probability measure on the design space. We will however see that the quality of the design depends on N for our model. Therefore, we need the n_j and not only the weights to describe the design even if we drop the integer requirement.

4.2 Optimal designs in fixed and mixed models

In optimal experimental design, we determine a design (5), which minimizes some error of the estimates. In fixed-effects models with a parameter vector $\boldsymbol{\beta}$, one is therefore often interested to minimize the covariance matrix $Cov(\hat{\boldsymbol{\beta}})$. We are here interested not only in the low-dimensional model with the parameter vector β, but consider $\boldsymbol{\mu}$ as the underlying model which we want to estimate. This means that our interest is to reduce the error of $\hat{\boldsymbol{\mu}}$. Since $\boldsymbol{\mu}$ is our underlying model, we should minimize $Var(\hat{\mu}(x) - \mu(x))$ and not $Cov(\hat{\boldsymbol{\mu}})$. Note that $\boldsymbol{\mu}$ is a random parameter in contrast to $\boldsymbol{\beta}$. So while we have $Cov(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = Cov(\hat{\boldsymbol{\beta}})$ in fixed effect models, we cannot remove $\boldsymbol{\mu}$ from the covariance in mixed-effect models. $Var(\hat{\mu}(x) - \mu(x))$ was specified in Theorem 2.2.

4.3 Optimality criteria

While the Loewner ordering could be used to order covariance matrices, defining $\mathbf{A} \geq \mathbf{B}$ if $\mathbf{A} - \mathbf{B}$ is positive semi-definite, this is only a partial order, and minimization of covariance matrices is usually not well defined with this ordering. Instead, following a usual optimal design approach, we have to choose which "aspect" of the covariance matrix to optimize using a criterion function, which maps the space of non-negative definite matrices to real numbers.

When minimizing $Cov(\boldsymbol{\beta})$ in fixed effects models, many optimality criteria have been discussed in literature, see e.g. Atkinson et al. (2007). Popular possibilities are to minimize the determinant of the covariance matrix, $\det(Cov(\hat{\boldsymbol{\beta}}))$, to minimize the average variance, $\text{trace}(Cov(\hat{\boldsymbol{\beta}}))$, or the variance of a linear combination $Var(\mathbf{c}^\top \hat{\boldsymbol{\beta}}) = \mathbf{c}^\top Cov(\hat{\boldsymbol{\beta}})\mathbf{c}$. If prediction is desired in random effects models, these criteria can be applied, too, by applying them on the covariance of the prediction error, $Cov(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$, see Prus and Schwabe (2016); Prus (2020). See also Hooks et al. (2009) for a discussion of optimality criteria for mixed-effects models.

In our case, it is reasonable to focus on the predicted mean function $\mu(x)$, $x \in [0, 1]$. One possibility is to require that the variance of $\hat{\mu}(x) - \mu(x)$, $x \in [0, 1]$ is small and we can consider the integrated variance

$$\int_{\mathcal{X}} \text{Var}[\hat{\mu}(x) - \mu(x)]dx. \tag{6}$$

We will focus in this article on designs minimizing expression (6), where $Var(\hat{\mu}(x) - \mu(x))$ can be found in Theorem 2.2. Prus and Schwabe (2016) called this criterion integrated mean-squared error (IMSE) criterion in the context of prediction and used it as their favorite criterion.

Other alternatives than IMSE-optimality would be possible, of course. First, one could weigh different regions of the design space differently and define a weighted IMSE criterion. One could be concerned with the worst variance in the design space and minimize $\max_{x\in[0,1]}\{\hat{\mu}(x) - \mu(x)\}$. This criterion called G-optimality in fixed-effect models is actually the oldest optimality criterion dating back to the work of Smith (1918). Another alternative is to minimize the prediction error $Cov(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$.

In some contexts, it might be desirable to estimate the value of x which gives a certain response, or the value of x which minimizes some cost-function or maximizes a utility-function, e.g. when the optimal price is of interest. These desires lead to **c**-optimality criteria. See e.g. Tsirpitzi and Miller (2021) for a maximization of a utility in a fixed-effects model.

4.4 Fedorov algorithm for the IMSE-optimal design

Fedorov (1972) introduced an exchange method in which the sum of design weights remains the same. The procedure starts by setting the initial design with the initial weights $n_0, n_1, \cdots n_m$. The main idea of the Fedorov exchange algorithm is to look for the optimal design by exchanging a unit α from n_i to n_j. Thus, an important stage in this algorithm is to identify all possible exchange couples (n_i, n_j). In the classical Fedorov algorithm the α unit equals 1, so the algorithm exchanges one point from n_i to n_j. If we consider not exact but continuous designs, α can be any value in the interval (0, 1] and can differentiate in each iteration. So we can have larger changes in the first iterations and smaller ones in the last when the algorithm approaches the optimal design. In order to find the couple (n_i, n_j) that will trade a unit α, Fedorov considered the interaction between the variance functions of the two weights. He defined the so-called Δ-function, which describes what happens in the optimality criterion when making a small change. Since we consider an IMSE-optimal design, we used the following Δ-function:

$$\Delta(n_i, n_j) = \int_{\mathcal{X}} \mathrm{Var}[\hat{\mu}(x) - \mu(x)]dx - crit$$

where $crit$ is the integral of the prediction error $\hat{\mu} - \mu$ of the initial n-vector.
Thus, the Fedorov algorithm computes the Δ-value for all the possible pairs and chooses the one pair with the smallest Δ-value. If there is more than one couple with the same Δ-value, one will be picked randomly. Since we are looking for the minimum Δ, we already have an improved design compared to the previous one as soon as we have a negative value. This procedure is repeated until there is no other exchange between a couple (n_i, n_j) that will decrease the Δ-value and will improve the optimal design.

Algorithm 1 Fedorov algorithm

```
Set the initial design with the initial values to n-vector
Compute the integral of the prediction error μ̂ − μ
while a negative delta for a couple of weights is found do
  for i from 1 to m + 1 do
    for j from 1 to m + 1 do
      By exchanging α point from n_i to n_j,
      compute the delta function for this couple
    end for
  end for
  Find the couple of weights that has the smallest delta
  among all the combinations
  if there is more than one couple with minimum delta then
    randomly select one couple
  end if
  exchange α point from n_i to n_j
  update n-vector
  reset minimum delta
end while
```

For further information on the exchange algorithm, see e.g. Triefenbach (2008).

4.5 Constrained optimization for approximate designs

If we want to compute optimal approximate designs and not optimal exact designs, we can also use one of many standard optimization algorithms. We can optimize both the number of observations n_i as well as the location $\tilde{x}_i$ of the design points. Optimal design problems are however constraint optimization problems. One equality constraint is that we want to determine the optimal design for a given total sample size, N, i.e. the sum of all n_i needs to be N. We handle this here by setting $n_m = N - \sum_{i=0}^{m-1} n_i$ and dropping n_m from the parameters to be optimized. We have further linear inequality constraints which we handle by applying the barrier method (see e.g. Lange 2013, Chapter 16). The constraints are here $0 \leq \tilde{x}_0, \tilde{x}_{i-1} \leq \tilde{x}_i, i = 1, \ldots, m, \tilde{x}_m \leq 1$, $n_i \geq 0, i = 0, \ldots, m-1, \sum_{i=0}^{m-1} n_i \leq N$. After incorporating a barrier, we can apply standard optimization algorithms. For the numerical calculations in Sect. 4.6.2, we have used the Nelder-Mead algorithm (see e.g. Givens and Hoeting 2013, Chapter 2.2.4) as implemented in the `R`-function `constrOptim`.

4.6 Results of optimal design

This section considers our specific example where the fixed effect model is straight line regression on [0, 1] and the misspecification term is the Brownian bridge. We start in Sect. 4.6.1 with keeping the observational points fixed as $\tilde{x}_i = i/m, i = 0, \ldots, m$, and optimize the weights at these observational points, i.e. we use the discrete design space $\mathcal{X}_m = \{0, 1/m, \ldots, 1\}$. In Sect. 4.6.2, we will optimize both the $\tilde{x}_i$ and the weights and have design space $\mathcal{X} = [0, 1]$ again.

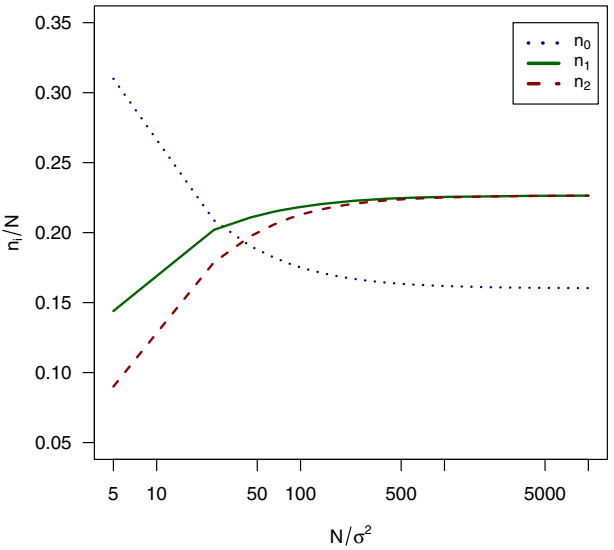

Fig. 2 Optimal design for n_i/N and $m = 4$ for different values of $N/\sigma^2 = \sum n_i/\sigma^2$ by using the Fedorov algorithm with $\alpha = 0.01$ exchange

4.6.1 Optimal weights for fixed observational points for IMSE optimality

In Figures 2, 3, 4, 5, the weights n_i/N of the optimal design are shown in dependence of $N/\sigma^2 = \sum n_i/\sigma^2$ for $m = 4$ and $m = 12$, where σ^2 is set equal to 1 and x-axis is in log scale. Note again that $n_j/N = n_{m-j}/N$, since the structure of the regression function is symmetric around 1/2, the Brownian bridge is tied down and symmetric and the optimality criterion is symmetric, too.

Numerically for $m = 4$, the values of $n_i/N, i = 0, 1, 2$ are shown in Figure 2 depending on the values of $N/\sigma^2 = \sum n_i/\sigma^2$. When we set $\sigma^2 = 1$, the values of N runs from 5 to 10,000 and the exchange α unit in Fedorov's algorithm is set to 0.01. Thus, when N is small the weight goes to the extreme values. While when N is larger than 30, the middle n_i/N are the one with higher weight than the weight of the extreme ones. For any N higher than 145, we see that the weights get stable, and the weight of the extreme points $n_0 = n_4$ is almost 0.17, while the weights of the middle points $n_1 = n_3$ and n_2 is almost 0.22.

In order to better understand how small the values of N should be to get all the weight in the two extreme values n_0, n_4, we created Figure 3. While the log scaled x-axis is in the interval 1 to 10 and $\alpha = 0.001$, the weight of the extreme values moves from 0.5 and they reach 0.25. Accordingly, the two extreme values share the weight as long as N/σ^2 is below to 2. For values bigger or equal to 2, n_0 and n_4 lose the monopoly and they start sharing the weight first with the two less extreme values n_1, n_3 and later with n_2.

In Figure 4 we illustrate two cases in order to show the impact of the exchange unit α in the Fedorov algorithm. The left panel corresponds to the discrete case, where α is 1, while the right is the continuous and $\alpha = 0.001$. While both plots follow the

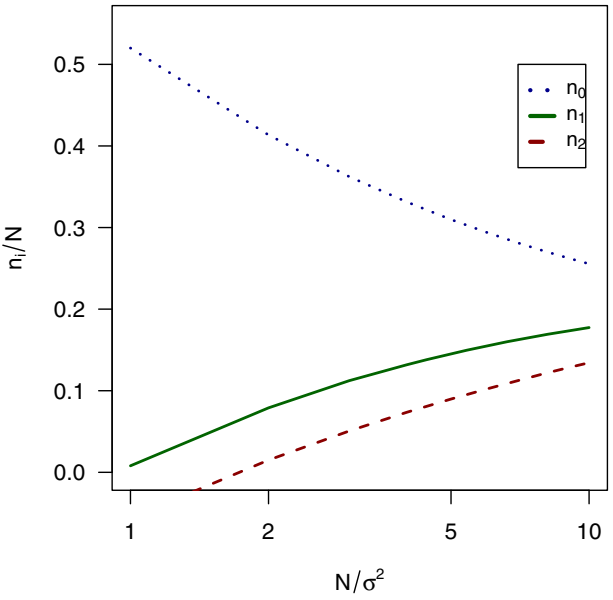

Fig. 3 Optimal design for n_i/N for $m = 4$ and for $N/\sigma^2 = \sum n_i/\sigma^2$ between 1 and 10 by using the Fedorov algorithm with $\alpha = 0.01$ exchange

same pattern, the interesting part in these two cases is the fluctuation of the n_i lines due to discreteness compared to the smoothness that occurs in the continuous. Due to discreteness, the optimal design needs not to be symmetrical and we have therefore not in all cases $n_i = n_{m-i}$; these values can differ by 1. In the continuous case, the computed optimal designs are always symmetrical with $n_i/N = n_{m-i}/N$.

In Figure 5 for $m = 12$, we present the values of $n_i/N, i = 0, 1, \ldots, 6$ on y-axis and the values of $N/\sigma^2 = \sum n_i/\sigma^2$ in log scale on x-axis, while σ^2 is set equal to 1 and α is 0.01. In contrast with the small values of N/σ^2, where the weight is gathered mainly in the extreme values, all middle n_i/N tend to the same value which is almost 0.08 when N/σ^2 gets higher than 1,000.

So in both cases $m = 4$ and $m = 12$, we see the impact of the amount of data in the use of the misspecification part, as we have already mentioned. When we have more and more data (N/σ^2 large), the optimal design estimates the misspecification part, since it is possible to obtain meaningful information on this part when a lot of data can be collected. On the other hand, the misspecification part is ignored by the optimal design if only a little data can be collected (N/σ^2 small) and the optimal design coincides with the one for simple linear regression putting half of the observations in the two endpoints of the design space.

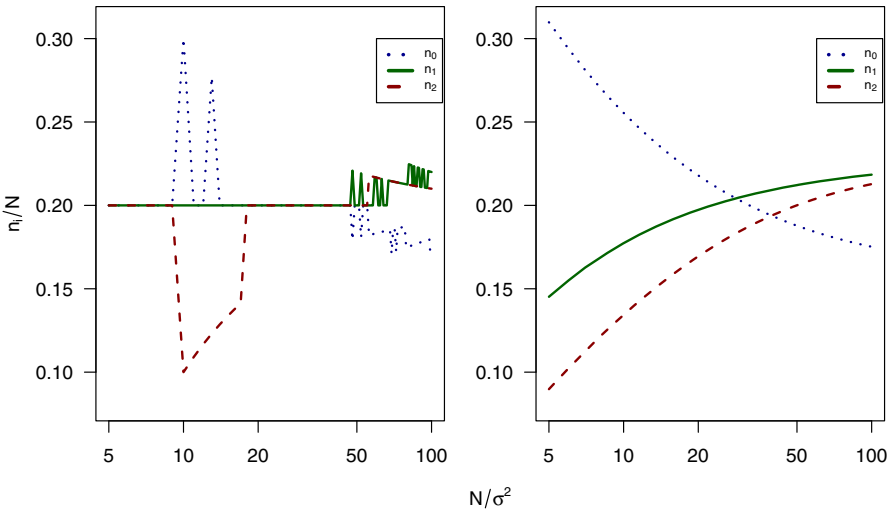

Fig. 4 Optimal design for n_i/N for $m = 4$ and for $N/\sigma^2 = \sum n_i/\sigma^2 = 5, 6, 7, \cdots, 99, 100$ by using the Fedorov algorithm when n_i are discrete ($\alpha = 1$) and when n_i are continuous ($\alpha = 0.001$)

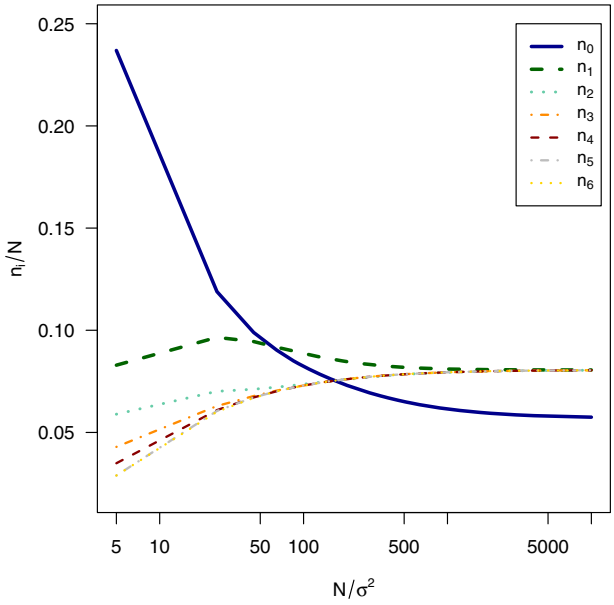

Fig. 5 Optimal design for n_i/N and $m = 12$ for different values of $N/\sigma^2 = \sum n_i/\sigma^2$ by using the Fedorov algorithm with $\alpha = 0.01$ exchange

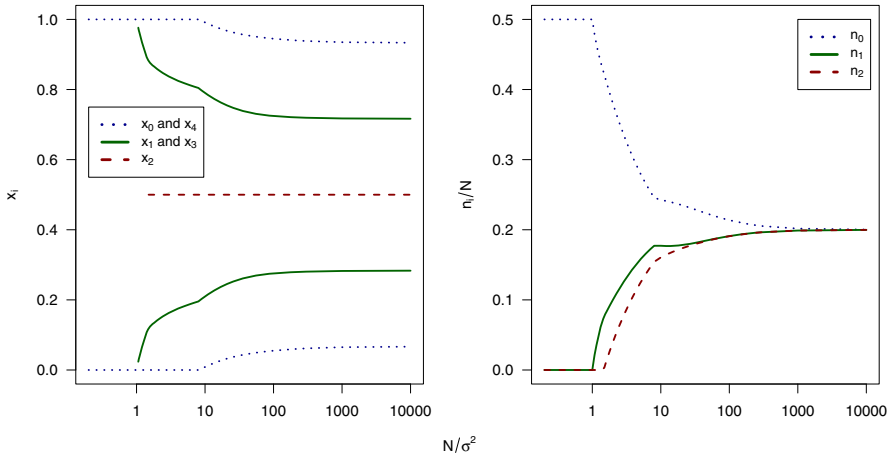

Fig. 6 Optimal design when both, observational points x_i and weights n_i/N are optimized: Left panel shows optimal $\tilde{x}_i$; right panel shows optimal weight n_i/N for $m = 4$

4.6.2 Optimal observational points and weights

For the case $m = 4$, we also allow now the observational points $\tilde{x}_i, i = 0, \ldots, 4$, to be chosen to optimize the IMSE criterion. Figure 6 shows the optimal $\tilde{x}_i$ in the left panel and the corresponding optimal weights n_i/N in the right panel.

For $N/\sigma^2 \leq 1$, the two-point design having 50% weight in 0 and 1 is optimal. For N/σ^2 somewhat larger than 1, two further design points, $\tilde{x}_1, \tilde{x}_3$ close to 0 and 1, respectively, are included with low weights and then a design point $\tilde{x}_2$ is added in the middle at 0.5 as well. With increasing N/σ^2, $\tilde{x}_1$ and $\tilde{x}_3$ tend then towards the middle and increase their weights. For $N/\sigma^2 < 10$, $\tilde{x}_0 = 0, \tilde{x}_4 = 1$ is the optimal choice; while it is better to choose $\tilde{x}_0 > 0$ and $\tilde{x}_4 < 1$, when $N/\sigma^2 \geq 10$.

For large N/σ^2, the optimal design has equal weights on all five observational points. The observational points are equidistantly spaced and $\tilde{x}_0 \approx 0.0668$. We will consider the case $N/\sigma^2 \to \infty$ now formally in Sect. 4.7, where we determine the IMSE-optimal design in the limiting case analytically.

4.7 Asymptotic case for large information

When $N \to \infty$, alternatively $\sigma^2 \to 0$, we get $\mathbf{D}_{\bar{\mathbf{Y}}} \to \mathbf{D}$. This limiting model can be seen as being free from the error term $\mathbf{e}$, i.e.

$$Y_i = \mu(x_i) = \mathbf{f}(x_i)^\top \boldsymbol{\beta} + C(x_i), i = 1, \ldots, N, \quad \text{or} \quad \bar{\mathbf{Y}} = \mathbf{W}\boldsymbol{\beta} + \boldsymbol{\gamma}.$$

It can be seen from our formulas that $\hat{\boldsymbol{\mu}} = \bar{\mathbf{Y}}$. The model without error term $\mathbf{e}$ has been considered for computer experiments, see e.g. Sacks et al. (1989); Williams and Rasmussen (2006).

For the model without error term, it makes no sense for prediction of μ to repeatedly observe at the same $\tilde{x}_i$ and since we have only single observations at each observational point, we can call them here x_i. To optimize the design, we consider therefore no weights. We optimize the choice of the $x_i \in [0, 1], i = 0, \ldots, m$. We derive now the optimal design for our specific example, Brownian bridge with straight line regression.

Theorem 4.1 *Consider the specific example where $C(x)$ is the Brownian bridge and $\mathbf{f}(x) = (1, x)^\top$ is the straight line regression. Let the number of design points $x_0, \ldots, x_m$ be fixed, $m + 1$. For $N \to \infty$ or for $\sigma^2 \to 0$, the IMSE converges to*

$$\frac{1}{6}\left(x_0^2 + \sum_{i=1}^{m}(x_i - x_{i-1})^2 + (1 - x_m)^2\right) + \frac{x_0^2 x_m + (1 - x_m)^2(1 - x_0)}{3(x_m - x_0)}. \quad (7)$$

The IMSE-optimal design for this limiting case is an equidistant design with design points

$$x_i = x_0 + i\frac{1 - 2x_0}{m}, \quad i = 1, \ldots, m, \quad (8)$$

where

$$x_0 = \frac{1}{8m + 8}\left(\frac{m^2}{A} + A + 3m + 4\right) \in (0, 1/m)$$

with $A = \sqrt[3]{-9m^3 - 16m^2 - 8m + 4m(m + 1)\sqrt{5m^2 + 8m + 4}}$.

The proof of Theorem 4.1 can be found in the appendix.

For $m = 4$, $x_0 \approx 0.0668$. Thus, we have analytically confirmed the numeric results for large N/σ^2 obtained in Sect. 4.6.2. For $m = 12$, $x_0 \approx 0.0256$.

We state now the limiting distribution for $m \to \infty$ in the following corollary.

Corollary 4.2 *Consider the IMSE-optimal design for the limiting case $N \to \infty$ or $\sigma^2 \to 0$ from Theorem 4.1. For $m \to \infty$, the IMSE-optimal design converges in probability to the uniform distribution on* $[0, 1]$.

Proof Since x_0 is guaranteed to be below $1/m$ and since the x_i are equidistantly spread, it is obvious that the maximal difference between the empirical distribution function $F_m(x)$ of the optimal design and $F(x) = x, x \in [0, 1]$, converges to 0, i.e. $\sup_{x\in[0,1]}|F_m(x) - F(x)| \to 0$, for $m \to \infty$. □

We have therefore shown that it is best to spread the design points uniformly on the design space when we expect that we can collect data with large information.

In some other cases, IMSE-optimal designs have been computed for the error-free model. As one example model, Mukherjee (2003) considers the Brownian bridge as $C(x)$ without regression function, $\mathbf{f}(x)^\top \boldsymbol{\beta} = 0$; her solution for the IMSE-optimal design is $x_i = i/(m+1), i = 1, \ldots, m$. Abt (1992) has also considered the Brownian bridge $C(x)$ and a constant regression, $\mathbf{f}(x) = 1$. An IMSE-optimal design in his case has distance $1/(m + 1)$ between the design points and the first design point can be chosen freely such that it is at most $1/(m + 1)$.

The case of linear or quadratic regression with the Brownian motion as process has been considered by Harman and Štulajter (2011). They show that equidistant designs are optimal for a large class of optimality criteria (not including IMSE-type criteria).

In addition to the mentioned cases where equidistant designs were optimal, the uniform design has also optimality properties for other models and situations. In the different situation of optimizing a lack of fit test when the alternative hypothesis is a rich class of functions, Wiens (1991) has shown the optimality of the uniform design, see also Biedermann and Dette (2001), Bischoff and Miller (2006), Wiens (2019).

5 Discussion

Model selection is an old area within statistics and is, in specific forms, used in machine learning. The universe of potential models can in itself be viewed as a pre-defined model. However, model selection will traditionally not respond to small or moderate misspecifications of a low-parameter model. A third-order polynomial can be attempted as an alternative to a quadratic. Data mining can cut the covariate space into fragments and offer different models in different intervals. However, an almost linear model with irregular deviations, e.g., may not be caught. However, the idea in this paper of adding a misspecification term could potentially be expanded to such settings, where competing models could be combined with a misspecification term.

The Brownian bridge distortion used in the main example should only be viewed as one possible misspecification function. We do not think that the difference between the true model and the low-parameter model, e.g. a linear model, would be exactly a Brownian bridge. The idea is rather that the misspecification cannot be easily understood or modeled. The Brownian bridge is then one reasonable choice that will result in a more pragmatic design, humbly reflecting that the simple low-parameter model is an approximation. While the Brownian bridge may be a useful misspecification model, an important factor limiting the practical usefulness of the methodology presented in this paper is that the volatility of the Brownian bridge is assumed to be pre-known. The choice of volatility parameter will impact the choice of design. It is, in principle, possible to estimate the volatility from the experimental data. However, that would partly contradict the idea of having a misspecification term to represent unknown model inaccuracies. One solution, in some situations, is to use previous experiments for similar situations to estimate the volatility. Another solution is to guesstimate the volatility before the experiment, and then to check the robustness over a range of plausible sizes for the misspecification. Alternative misspecification models exist that could be used instead of a Brownian bridge but the issue of choosing a volatility parameter, or a similar measure of the size of the misspecification, will remain. An interesting idea is to use spline functions for the misspecification. A drawback, in settings where monotonicity should theoretically be expected, is that the resulting combined model will not be monotone with standard splines. Consequently, the function will not be invertible. The same issue is partly relevant for the Brownian bridge. In a concrete situation, one should consider whether this is a real problem or a useful approximation. One way of distorting e.g. an Emax model while preserving the monotonicity would be to distort (compress/expand) the dose scale (x-axis) rather than the response (y-axis).

The choice of optimality criterion is important when applying optimal design theory. In the example, we have chosen to focus on IMSE-optimality, integrating the variance over the entire design space. Specific situations may call for quite different optimality criteria. For example, optimizing the profit in a simple demand model could imply that the objective is to find $\text{argmax}_x\,((x-k)\cdot f(x))$, where x is the price, k the variable cost per unit, and $f(x)$ is the demand function. Further research could explore optimal designs for this and a multitude of other relevant optimality criteria.

The methodology is relatively computer-intensive. Our example, using only one covariate and two parameters, does not require much computing time. However, it can be expected that computation issues will arise in more complicated models. It would be of interest to build similar models with multiple covariates and bring the methodology closer to machine learning, where many covariates are typically explored. It is straightforward to decrease computing times by using more efficient optimization algorithms. Still, how to optimize these algorithms may be an area for future research.

Acknowledgements We would like to thank Mattias Villani for pointing out the connection to the Gaussian process methodology used in Machine Learning. We are grateful to a reviewer for challenging comments which led to an improved paper.

Funding Open access funding provided by Stockholm University.

Appendix A

Proof of Lemma 2.1 1. and 2.: See e.g. Christensen (2002); the first formula is based on model (2), the second on the model for the means (3).

3.: We put $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ on $\hat{\boldsymbol{\mu}}$ and get:

$$\begin{aligned}
\hat{\boldsymbol{\mu}} &= \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\boldsymbol{\gamma}} \\
&= \mathbf{X}(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}\mathbf{Y} + \mathbf{Z}\mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{Y} - \mathbf{Z}\mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{X}(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}\mathbf{Y} \\
&= [\mathbf{Z}\mathbf{D}\mathbf{Z}^\top\mathbf{M} + (\mathbf{I}_{m+1} - \mathbf{Z}\mathbf{D}\mathbf{Z}^\top\mathbf{M})\mathbf{X}(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}]\mathbf{Y}, \\
\hat{\boldsymbol{\mu}} &= \mathbf{W}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}} \\
&= \mathbf{W}(\mathbf{W}^\top\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}\bar{\mathbf{Y}} \\
&\quad + \mathbf{D}\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}[\mathbf{I}_{m+1} - \mathbf{W}(\mathbf{W}^\top\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}]\bar{\mathbf{Y}} \\
&= [\mathbf{D}\mathbf{D}_{\bar{\mathbf{Y}}}^{-1} + (\mathbf{I}_{m+1} - \mathbf{D}\mathbf{D}_{\bar{\mathbf{Y}}}^{-1})\mathbf{W}(\mathbf{W}^\top\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}]\bar{\mathbf{Y}}.
\end{aligned}$$

4.: This formula can be deduced from Sect. 2.2 and 2.7 of Williams and Rasmussen (2006) using the case of a vague prior for $\boldsymbol{\beta}$, see also O'Hagan (1978). □

Lemma A.1 *1. The covariance matrix for the Best Linear Unbiased Estimator (BLUE) of* $\boldsymbol{\beta}$ *is* $Cov(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}$.

2. The covariance matrix for the Best Linear Unbiased Predictor (BLUP) of $\boldsymbol{\gamma}$ *is* $Cov(\hat{\boldsymbol{\gamma}}) = \mathbf{D}\mathbf{Z}^\top\{\mathbf{I}_N - \mathbf{M}\mathbf{X}(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\}\mathbf{M}\mathbf{Z}\mathbf{D}^\top$.

Proof of Lemma A.1

$$
\begin{aligned}
1.: \mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}\mathbf{X}\{\mathbf{X}^\top\mathbf{M}\mathbf{X}\}^{-1} = (\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}.\\
2.: \mathrm{Cov}(\hat{\boldsymbol{\gamma}}) &= \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{Z}\mathbf{D}^\top - \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{X}(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}\mathbf{Z}\mathbf{D}^\top\\
&= \mathbf{D}\mathbf{Z}^\top\{\mathbf{I}_N - \mathbf{M}\mathbf{X}(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\}\mathbf{M}\mathbf{Z}\mathbf{D}^\top.
\end{aligned}
$$

Proof of Theorem 2.2

$$
\begin{aligned}
&1.: \mathrm{Cov}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\\
&= \mathrm{Cov}(\mathbf{W}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}} - \mathbf{W}\boldsymbol{\beta} - \boldsymbol{\gamma})\\
&= \mathrm{Cov}\big\{\mathbf{W}(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{e}) - \boldsymbol{\gamma}\\
&\quad + \mathbf{D}\mathbf{Z}^\top\mathbf{M}[\mathbf{I}_N - \mathbf{X}(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}](\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{e})\big\}\\
&= \mathrm{Cov}\big\{\mathbf{W}(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}\mathbf{Z}\boldsymbol{\gamma} + \mathbf{W}(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}\boldsymbol{e} - \boldsymbol{\gamma}\\
&\quad + \mathbf{D}\mathbf{Z}^\top\mathbf{M}[\mathbf{I}_N - \mathbf{X}(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}]\mathbf{Z}\boldsymbol{\gamma}\\
&\quad + \mathbf{D}\mathbf{Z}^\top\mathbf{M}[\mathbf{I}_N - \mathbf{X}(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}]\boldsymbol{e}\big\}\\
&= \mathrm{Cov}\big[\{(\mathbf{W} - \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{X})(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}\mathbf{Z} + \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{Z} - \mathbf{I}_{\mathrm{m}+1}\}\boldsymbol{\gamma}\\
&\quad + \{(\mathbf{W} - \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{X})(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M} + \mathbf{D}\mathbf{Z}^\top\mathbf{M}\}\boldsymbol{e}\big].
\end{aligned}
$$

Since $\boldsymbol{\gamma}$ and $\boldsymbol{e}$ are independent, we obtain:

$$
\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) &= \big[(\mathbf{W} - \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{X})(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}\mathbf{Z} + \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{Z} - \mathbf{I}_{m+1}\big]\mathrm{Cov}(\boldsymbol{\gamma})\\
&\quad \big[(\mathbf{W} - \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{X})(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}\mathbf{Z} + \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{Z} - \mathbf{I}_{m+1}\big]^\top\\
&\quad + \big[(\mathbf{W} - \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{X})(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M} + \mathbf{D}\mathbf{Z}^\top\mathbf{M}\big]\mathrm{Cov}(\boldsymbol{e})\\
&\quad \big[(\mathbf{W} - \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{X})(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M} + \mathbf{D}\mathbf{Z}^\top\mathbf{M}\big]^\top\\
&= \big[(\mathbf{W} - \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{X})(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}\mathbf{Z} + \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{Z} - \mathbf{I}_{m+1}\big]\mathbf{D}\\
&\quad \big[(\mathbf{W} - \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{X})(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}\mathbf{Z} + \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{Z} - \mathbf{I}_{m+1}\big]^\top\\
&\quad + (\mathbf{W} - \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{X})(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}\mathbf{R}\mathbf{M}\mathbf{X}(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\\
&\quad (\mathbf{W} - \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{X})^\top\\
&\quad + (\mathbf{W} - \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{X})(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{M}\mathbf{R}\mathbf{M}\mathbf{Z}\mathbf{D}\\
&\quad + \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{R}\mathbf{M}\mathbf{X}(\mathbf{X}^\top\mathbf{M}\mathbf{X})^{-1}(\mathbf{W} - \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{X})^\top + \mathbf{D}\mathbf{Z}^\top\mathbf{M}\mathbf{R}\mathbf{M}\mathbf{Z}\mathbf{D}.
\end{aligned}
$$

2.: This formula can be deduced from Williams and Rasmussen (2006), Sect. 2, for the case of a vague prior for $\boldsymbol{\beta}$. □

Proof of Theorem 4.1 According to Theorem 2.2, the variance for the prediction at x^* is $v_1(x^*) + v_2(x^*)$ with

$$v_1(x^*) = x^*(1 - x^*) - \mathbf{D}_*(x^*)^\top \mathbf{D}_{\bar{\mathbf{Y}}}^{-1} \mathbf{D}_*(x^*),$$
$$v_2(x^*) = \mathbf{S}(x^*)^\top (\mathbf{W}\mathbf{D}_{\bar{\mathbf{Y}}}^{-1}\mathbf{W}^\top)^{-1}\mathbf{S}(x^*).$$

For $\sigma^2 = 0$, the matrix $\mathbf{D}_{\bar{\mathbf{Y}}} = \mathbf{D}$ is not invertible if $x_0 = 0$ or $x_m = 1$, but the limits of the integrated v_1 and v_2 exist when $\sigma^2 \to 0$. I.e. we can do following considerations for the case $x_0 > 0$, $x_m < 1$, but the resulting formula are correct if $x_0 = 0$ or $x_m = 1$ as well. Using the well-known tri-diagonal form of $\mathbf{D}^{-1}$ (see e.g. Abt 1992; Bischoff et al. 2003), it can be shown that

$$\int_0^1 v_1(x^*)\, dx^* \to \frac{1}{6}\left\{x_0^2 + \sum_{i=1}^{m}(x_i - x_{i-1})^2 + (1 - x_m)^2\right\},$$
$$\int_0^1 v_2(x^*)\, dx^* \to \frac{1}{3}\{x_0^2 x_m + (1 - x_m)^2(1 - x_0)\}/(x_m - x_0).$$

To show the optimality of design (8), we compute the derivatives of the IMSE (7) with respect to $x_i, i = 0, \ldots, x_m$ and set them to 0. From the derivatives for $x_i, i = 1, \ldots, m - 1$, we obtain the unique solution $x_i = (x_{i+1} + x_{i-1})/2$ and therefore, the optimal design is equidistant. Setting the derivatives with respect to x_0 and x_m to 0, we obtain that $x_m = 1 - x_0$ and that x_0 solves the third degree equation $h_m(x) = 0$ with

$$h_m(x) = (8m + 8)x^3 - (9m + 12)x^2 + (3m + 6)x - 1. \tag{9}$$

Lemma A.2 shows that $h_m(x) = 0$ has exactly one solution and that this solution is in $(0, 1/m)$. Solving the equation yields the value noted in the theorem. □

Lemma A.2 *For each $m \geq 1$ and for h_m defined by (9), the third degree equation $h_m(x) = 0$ has exactly one solution x_0. This solution is in the interval $(0, 1/m)$.*

Proof of Lemma A.2 It is straightforward to verify that $h_m(0) < 0$, $h_m(1/m) > 0$, and $h_m(1/2) > 0$ for all $m \geq 1$. Further, h_m has a local minimum in $1/2$ since $h_m'(1/2) = 0$ and $h_m''(1/2) > 0$. Therefore the claim in the lemma follows. □

References

Abt M (1992) Some exact optimal designs for linear covariance functions in one dimension. Commun Stat-Theory Methods 21(7):2059–2069

Atkinson A, Donev A, Tobias R (2007) Optimum experimental designs, with SAS. Oxford University Press, Oxford

Biedermann S, Dette H (2001) Optimal designs for testing the functional form of a regression via nonparametric estimation techniques. Stat Prob Lett 52(2):215–224

Bischoff W, Miller F (2006) Optimal designs which are efficient for lack of fit tests. Annals Stat 34(4):2015–2025

Bischoff W, Hashorva E, Hüsler J et al (2003) Exact asymptotics for boundary crossings of the Brownian bridge with trend with application to the Kolmogorov test. Annal Inst Stat Math 55(4):849–864
Blight B, Ott L (1975) A Bayesian approach to model inadequacy for polynomial regression. Biometrika 62(1):79–88
Chow WC (2009) Brownian bridge. Wiley Interdisciplinary Rev: Comput Stat 1(3):325–332
Christensen R (2002) Plane answers to complex questions. Springer, New York
Cohn DA (1996) Neural network exploration using optimal experiment design. Neural Netw 9(6):1071–1083
Cohn DA, Ghahramani Z, Jordan MI (1996) Active learning with statistical models. J Artif Intell Res 4:129–145
Dette H, Pepelyshev A, Zhigljavsky A (2016) Optimal designs in regression with correlated errors. Annals Stat 44:113–152
Dette H, Konstantinou M, Zhigljavsky A (2017) A new approach to optimal designs for correlated observations. Annals Stat 45:1579–1608
Fedorov V, Jones B (2005) The design of multicentre trials. Stat Method Med Res 14(3):205–248
Fedorov VV (1972) Theory of optimal experiments. Academic Press, New York
Fernandez-Tapia J, Guéant O, Lasry JM (2017) Optimal real-time bidding strategies. Appl Math Res Exp 1:142–183
Givens GH, Hoeting JA (2013) Computational statistics, 2nd edn. John Wiley & Sons, Hoboken, New Jersey
Harman R, Štulajter F (2011) Optimality of equidistant sampling designs for the Brownian motion with a quadratic drift. J Stat Plan Infer 141(8):2750–2758
Hooks T, Marx D, Kachman S et al (2009) Optimality criteria for models with random effects. Revista Colombiana de Estadística 32(1):17–31
Kiefer J (1974) General equivalence theory for optimum designs (approximate theory). Ann Stat 2:849–879
Kohavi R, Longbotham R, Sommerfield D et al (2009) Controlled experiments on the web: survey and practical guide. Data Min Knowl Disc 18:140–181
Kohavi R, Deng A, Frasca B, et al (2013) Online controlled experiments at large scale. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 1168–1176, https://exp-platform.com/large-scale/
Lange K (2013) Numerical analysis for statisticians, 2nd edn. Springer, New York
Liu X, Yue RX, Wong WK (2019) D-optimal designs for multi-response linear mixed models. Metrika 82(1):87–98
López-Fidalgo J, Wiens DP (2022) Robust active learning with binary responses. J Stat Plan Infer 220:1–14
Mardanlou V, Karlsson N, Guo J (2017) Statistical plant modeling and simulation in online advertising. In: 2017 American control conference (ACC), IEEE, pp 2176–2181
Montgomery DC (2017) Design and analysis of experiments, 9th edn. Wiley
Mukherjee B (2003) Exactly optimal sampling designs for processes with a product covariance structure. Can J Stat 31(1):69–87
Nie R, Wiens DP, Zhai Z (2018) Minimax robust active learning for approximately specified regression models. Can J Stat 46(1):104–122
O'Hagan A (1978) Curve fitting and optimal design for prediction. J Royal Stat Soc: Series B (Methodological) 40(1):1–24
Prus M (2020) Optimal designs in multiple group random coefficient regression models. Test 29(1):233–254
Prus M, Schwabe R (2016) Optimal designs for the prediction of individual parameters in hierarchical models. J Royal Statist Soc, Series B
Ross SM, Kelly JJ, Sullivan RJ et al (1996) Stochastic processes, vol 2. Wiley, New York
Sacks J, Ylvisaker D (1966) Designs for regression problems with correlated errors. Annals Math Stat 37(1):66–89
Sacks J, Schiller SB, Welch WJ (1989) Designs for computer experiments. Technometrics 31(1):41–47
Settles B (2010) Active learning literature survey. Comput Sci Tech Rep 1648 http://burrsettles.com/pub/settles.activelearning.pdf
Siivola E, Weber S, Vehtari A (2021) Qualifying drug dosing regimens in pediatrics using Gaussian processes. Stat Med 40(10):2355–2372
Smith K (1918) On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. Biometrika 12(1/2):1–85

Triefenbach F (2008) Design of experiments: the D-optimal approach and its implementation as a computer algorithm. Bachelor's Thesis in Information and Communication Technology
Tsirpitzi RE, Miller F (2021) Optimal dose-finding for efficacy-safety-models. Biometric J 63(6):1185–1201
Wiens DP (1991) Designs for approximately linear regression: two optimality properties of uniform designs. Stat Prob Lett 12(3):217–221
Wiens DP (2019) Maximin power designs in testing lack of fit. J Stat Plan Infer 199:311–317
Williams CK, Rasmussen CE (2006) Gaussian processes for machine learning, vol 2. MIT press Cambridge, MA