SIAM J. CONTROL OPTIM.
Vol. 51, No. 4, pp. 3181–3207

SAMPLING DECISIONS IN OPTIMUM EXPERIMENTAL DESIGN IN THE LIGHT OF PONTRYAGIN'S MAXIMUM PRINCIPLE*

SEBASTIAN SAGER†

Abstract. Optimum experimental design (OED) problems are optimization problems in which an experimental setting and decisions on when to measure—the so-called sampling design—are to be determined such that a follow-up parameter estimation yields accurate results for model parameters. In this paper we use the interpretation of OED as optimal control problems with a very particular structure for the analysis of optimal sampling decisions. We introduce the information gain function, motivated by an analysis of necessary conditions of optimality. We highlight differences between problem formulations and propose to use a linear penalization of sampling decisions to overcome the intrinsic ill-conditioning of OED. The results of this paper are independent from the actual numerical method to compute the solution to the OED problem and of the question of local and global optima.

Key words. optimal design of experiments, optimal control, mixed integer programming

AMS subject classifications. 62K05, 49J15, 90C11

DOI. 10.1137/110835098

1. Introduction.

Modeling, simulation, and optimization has become an indispensable tool in science, complementary to theory and experiment. It builds on detailed mathematical models that are able to represent real world behavior of complex processes. In addition to correct equations, problem specific *model parameters*, such as masses, reaction velocities, or mortality rates, need to be estimated. The methodology optimum experimental design (OED) helps to design experiments that yield as much information on these model parameters as possible.

OED has a long tradition in statistics and practice; compare the textbook [18]. References to some algorithmic approaches are given, e.g., in [1, 23]. Algorithms for OED of nonlinear dynamic processes are usually based on the works of [3, 11, 12]. As investigated in [14], derivative based optimization strategies are state of the art. The methodology has been extended in [12] to cope with the need for robust designs. In [13] a reformulation is proposed that allows an application of Bock's direct multiple shooting method. An overview of model-based design of experiments can be found in [6]. Applications of OED to process engineering are given in [2, 24].

OED of dynamic processes is a nonstandard optimal control problem in the sense that the objective function is a function of either the Fisher information matrix, or of its inverse, the variance-covariance matrix. The Fisher matrix can be formulated as the time integral over derivative information. This results in an optimal control problem with a very specific structure. In this paper we analyze this structure to shed light on the question under which circumstances it is optimal to measure.

Notation. When analyzing OED problems with the maximum principle, one encounters one notational challenge. We have an objective function that is a function of a matrix; however, the necessary conditions are usually formulated for vector-valued variables. We have two options: either we redefine matrix operations as the inverse, trace or determinant for vectors, or we need to interpret matrices as vectors and

*Received by the editors May 23, 2011; accepted for publication (in revised form) June 17, 2013; published electronically August 6, 2013.
http://www.siam.org/journals/sicon/51-4/83509.html
†Otto-von-Guericke-Universität Magdeburg, 39106 Magdeburg, Germany (sebastian.sager@ovgu.de, http://mathopt.de).

define a scalar product for matrix-valued variables that allows us to multiply them with Lagrange multipliers and obtain a map to the real numbers. We decided to use the second option. In the interest of better readability we use bold symbols for all matrices. (In)equalities are always meant to hold componentwise, also for matrices.

DEFINITION 1.1 (scalar product of matrices). *The map* $\langle \cdot, \cdot \rangle : (\boldsymbol{\lambda}, \boldsymbol{A}) \mapsto \langle \boldsymbol{\lambda}, \boldsymbol{A} \rangle \in \mathbb{R}$ *with matrices* $\boldsymbol{\lambda}$ *and* $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ *is defined as*

$$\langle \boldsymbol{\lambda}, \boldsymbol{A} \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} \lambda_{i,j} \; A_{i,j}.$$

Partial derivatives are often written as subscripts, e.g. $\mathcal{H}_\lambda = \frac{\partial \mathcal{H}}{\partial \lambda}$. In our analysis we encounter the necessity to calculate directional derivatives of matrix functions with respect to matrices. In order to conveniently write them, we define a map analogously to the case in $\mathbb{R}^n$.

DEFINITION 1.2 (matrix-valued directional derivatives). *Let a differentiable map* $\Phi : \mathbb{R}^{n \times n} \mapsto \mathbb{R}^{n \times n}$ *be given, and let* $\boldsymbol{A}, \boldsymbol{\Delta A} \in \mathbb{R}^{n \times n}$. *Then the directional derivative is denoted by*

$$\left(\frac{\partial \Phi(\boldsymbol{A})}{\partial \boldsymbol{A}} \cdot \boldsymbol{\Delta A} \right)_{k,l} := \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\partial \Phi(A)_{k,l}}{\partial A_{i,j}} \, \Delta A_{i,j} = \lim_{h \to 0} \frac{\Phi(A + h \Delta A)_{k,l} - \Phi(A)_{k,l}}{h}$$

for $1 \leq k, l \leq n$, *hence* $\frac{\partial \Phi(\boldsymbol{A})}{\partial \boldsymbol{A}} \cdot \boldsymbol{\Delta A} \in \mathbb{R}^{n \times n}$.

Let a differentiable map $\phi : \mathbb{R}^{n \times n} \mapsto \mathbb{R}$ *be given, and let* $\boldsymbol{A}, \boldsymbol{\Delta A} \in \mathbb{R}^{n \times n}$. *Then the directional derivative* $\lim_{h \to 0} \frac{\phi(A + h \Delta A) - \phi(A)}{h}$ *is denoted by*

$$\left\langle \frac{\partial \phi(\boldsymbol{A})}{\partial \boldsymbol{A}}, \boldsymbol{\Delta A} \right\rangle \; = \; \frac{\partial \phi(\boldsymbol{A})}{\partial \boldsymbol{A}} \cdot \boldsymbol{\Delta A} \; := \; \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\partial \phi(\boldsymbol{A})}{\partial A_{i,j}} \, \Delta A_{i,j},$$

hence $\langle \frac{\partial \phi(\boldsymbol{A})}{\partial \boldsymbol{A}}, \boldsymbol{\Delta A} \rangle \; = \frac{\partial \phi(\boldsymbol{A})}{\partial \boldsymbol{A}} \cdot \boldsymbol{\Delta A} \in \mathbb{R}$.

In the following we use the map $\Phi(\cdot)$ for the inverse operation, and the map $\phi(\cdot)$ for either trace, determinant, or maximum eigenvalue function.

Outline. The paper is organized as follows. In section 2 we revise results from optimal control theory. In section 3 we formulate the OED problem as an optimal control problem. We apply the maximum principle to OED in section 4, and derive conclusions from our analysis. Two numerical examples are presented in section 5, before we summarize in section 6. Useful lemmas are provided for convenience in the appendix.

2. Indirect approach to optimal control. The basic idea of indirect approaches is *first optimize, then discretize.* In other words, first necessary conditions for optimality are applied to the optimization problem in function space, and in a second step the resulting boundary value problem is solved by an adequate discretization, such as multiple shooting. The necessary conditions for optimality are given by the famous maximum principle of Pontryagin. Assume we want to solve the optimal control problem of Bolza type

$$\begin{aligned} &\min_{y,u} \quad \Phi(y(t_\mathrm{f})) + \textstyle\int_{\mathcal{T}} L(y(\tau), u(\tau)) \, \mathrm{d}\tau \\ &\text{subject to} \\ &\qquad \dot{y}(t) \; = f(y(t), u(t), p), \quad t \in \mathcal{T}, \\ &\qquad u(t) \; \in \mathcal{U}, \qquad\qquad\quad t \in \mathcal{T}, \\ &\qquad 0 \quad \leq c(y(t_\mathrm{f})), \\ &\qquad y(0) = y_0 \end{aligned} \tag{2.1}$$

on a fixed time horizon $\mathcal{T} = [0, t_\mathrm{f}]$ with differential states $y : \mathcal{T} \mapsto \mathbb{R}^{n_y}$, fixed model parameters $p \in \mathbb{R}^{n_p}$, a bounded feasible set $\mathcal{U} \in \mathbb{R}^{n_u}$ for the control functions $u : \mathcal{T} \mapsto \mathbb{R}^{n_u}$, and sufficiently smooth functions $\Phi(\cdot), L(\cdot), f(\cdot), c(\cdot)$. To state the maximum principle we need the concept of the Hamiltonian.

DEFINITION 2.1 (Hamiltonian, adjoint states, end-point Lagrangian). *The Hamiltonian of optimal control problem* (2.1) *is given by*

$$\mathcal{H}(y(t), u(t), \lambda_0, \lambda(t), p) := -\lambda_0 L(x(t), u(t)) + \lambda(t)^T\, f(y(t), u(t), p) \tag{2.2}$$

with variables $\lambda_0 \in \mathbb{R}$ *and* $\lambda : \mathcal{T} \mapsto \mathbb{R}^{n_y}$ *called adjoint variables. The end–point Lagrangian function* ψ *is defined as*

$$\psi(y(t_\mathrm{f}), \mu) := \Phi(y(t_\mathrm{f})) - \mu^T c(y(t_\mathrm{f})) \tag{2.3}$$

with nonnegative Lagrange multipliers $\mu \in \mathbb{R}^{n_c}_+$.

The *maximum principle* in its basic form, also sometimes referred to as *minimum principle*, goes back to the early fifties and the works of Hestenes, Boltyanskii, Gamkrelidze, and of course Pontryagin. Although we refer to it as the maximum principle for historic reasons, we chose to use a formulation with a minimization term which is more standard in the optimization community. Precursors of the maximum principle as well as of the Bellman equation can already be found in Carathéodory's book of 1935; compare [16] for details.

The maximum principle states the existence of adjoint variables $\lambda^*(\cdot)$ that satisfy adjoint differential equations and transversality conditions. The optimal control $u^*(\cdot)$ is characterized as an implicit function of the states and the adjoint variables—a minimizer $u^*(\cdot)$ of problem (2.1) also minimizes the Hamiltonian subject to additional constraints.

THEOREM 2.2 (maximum principle). *Let problem* (2.1) *have a feasible optimal solution* $(y^*, u^*)(\cdot)$. *Then there exist* $\lambda_0^* \in \mathbb{R}$, *adjoint variables* $\lambda^*(\cdot)$, *with* $(\lambda_0^*, \lambda^*(\cdot)) \neq 0$, *and Lagrange multipliers* $\mu^* \in \mathbb{R}^{n_c}_+$ *such that*

$$\dot{y}^*(t) = \mathcal{H}_\lambda(y^*(t), u^*(t), \lambda_0^*, \lambda^*(t), p) = f(y^*(t), u^*(t), p), \tag{2.4a}$$

$$\dot{\lambda}^{*T}(t) = -\mathcal{H}_y(y^*(t), u^*(t), \lambda_0^*, \lambda^*(t), p), \tag{2.4b}$$

$$y^*(0) = y_0, \tag{2.4c}$$

$$\lambda^{*T}(t_\mathrm{f}) = -\psi_y(y^*(t_\mathrm{f}), \mu^*), \tag{2.4d}$$

$$u^*(t) \in \arg\min_{u \in \mathcal{U}} \mathcal{H}(y^*(t), u, \lambda_0^*, \lambda^*(t), p), \tag{2.4e}$$

$$0 \leq c(y^*(t_\mathrm{f})), \tag{2.4f}$$

$$0 \leq \mu^*, \tag{2.4g}$$

$$0 = \mu^{*T} c(y^*(t_\mathrm{f})) \tag{2.4h}$$

for $t \in \mathcal{T}$ *almost everywhere.*

For a proof of the maximum principle see [17, 8]. Further references can be found, e.g., in [5]. Although formulation (2.1) is not the most general formulation of an optimal control problem, it covers the experimental design optimization task as we formulate it in the next section. However, one may also be interested in the case where measurements are not performed continuously over time, but rather at discrete points in time. To include such discrete events on a given time grid, we need to extend

(2.1) to

$$
\begin{aligned}
&\min_{y,u,w} \quad \Phi(y(t_\mathrm{f})) + \textstyle\int_{\mathcal{T}} L(y(\tau),u(\tau))\,\mathrm{d}\tau + \sum_{k=1}^{n_m} L^{\mathrm{tr}}(w_k)\\
&\text{subject to}\\
&\qquad \begin{aligned}
\dot{y}(t) &= f(y(t),u(t),p), && t\in\mathcal{T}^k,\\
y(t_k^+) &= f^{\mathrm{tr}}(y(t_k^-),w_k,p), && k=1\dots n_m,\\
u(t) &\in \mathcal{U}, && t\in\mathcal{T},\\
w_k &\in \mathcal{W}, && k=1,\dots,n_m,\\
0 &\le c(y(t_\mathrm{f})),\\
y(0) &= y_0
\end{aligned}
\end{aligned}
\tag{2.5}
$$

on fixed time horizons $\mathcal{T}^k = [t_k, t_{k+1}]$, $k = 0,\dots,n_m-1$ with $t_0 = 0$ and $t_{n_m} = t_\mathrm{f}$. In addition to (2.1) we have variables $w = (w_1,\dots,w_{n_m})$ with $w_k \in \mathcal{W} \subset \mathbb{R}$, a second smooth Lagrange term function $L^{\mathrm{tr}}(\cdot)$, and a smooth transition function $f^{\mathrm{tr}}(\cdot)$ that causes jumps in some of the differential states.

The boundary value problem (2.4) needs to be modified by additional jumps in the adjoint variables, e.g., for $k = 1\dots n_m$,

$$
\lambda^{*T}(t_k^+) = \lambda^{*T}(t_k^-) - \mathcal{H}_y^{\mathrm{tr}}(y^*(t_k^-), w_k^*, p, \lambda_0^{\mathrm{tr}}, \lambda^*(t_k^+)),
\tag{2.6}
$$

$$
w_k^* \in \arg\min_{w_k\in\mathcal{W}} \mathcal{H}^{\mathrm{tr}}(y(t_k^-), w_k, p, \lambda_0^{\mathrm{tr}}, \lambda^*(t_k^+))
\tag{2.7}
$$

with the discrete time Hamiltonian

$$
\mathcal{H}^{\mathrm{tr}}(y(t_k^-), w_k, p, \lambda^*(t_k^+)) := -\lambda_0^{\mathrm{tr}} L^{\mathrm{tr}}(w_k) + \lambda^T(t_k^+)\, f^{\mathrm{tr}}(y(t_k^-), w_k, p).
\tag{2.8}
$$

A derivation and examples for the (purely) discrete time maximum principle can be found, e.g., in [25]. Please note that there are many different versions and proofs of the maximum principle. Unfortunately, we are not aware of one that includes the exact conditions with a proof for problem (2.5). As it is beyond the scope of this paper to prove them, we assume the conditions and our choice of $\lambda_0 = 1$ to be correct on a heuristic basis.

One interesting aspect about the global maximum principle (2.4) is that the constraint $u \in \mathcal{U}$ has been transferred towards the inner minimization problem (2.4e). This is done on purpose, so no assumptions need to be made on the feasible control domain $\mathcal{U}$. The maximum principle also applies to nonconvex and disjoint sets $\mathcal{U}$, such as, e.g., $\mathcal{U} = \{0,1\}$ in mixed-integer optimal control. For a disjoint set $\mathcal{U}$ of moderate size the pointwise minimization of (2.4e) can be performed by enumeration between the different choices, implemented as switching functions that determine changes in the minimum. This approach, the *competing Hamiltonians* approach, has to our knowledge first been successfully applied to the optimization of operation of subway trains with discrete acceleration stages in New York by [4].

In this study we are not interested in applying the maximum principle directly to the disjoint set $\mathcal{U}$, but rather to its convex hull. We are interested in the question when the solutions of the two problems coincide, and which exact problem formulations are favorable in this sense. Having analyzed problem structures with the help of the maximum principle, we switch to direct, *first-discretize–then-optimize* approaches to actually solve OED problems. Using the convex hull simplifies the usage of modern gradient-based optimization strategies.

3. Optimum experimental design problems. In this section we formulate the problem classes of experimental design problems we are interested in.

3.1. Problem formulation: Discrete time. We are interested in optimal parameter values for a model-measurements fit. Assuming an experimental setup is given by means of control functions $u^i(\cdot)$ and sampling decisions w^i that indicate whether a measurement is performed or not for n_{exp} experiments, we formulate this parameter estimation problem as

$$\begin{aligned} \min_{x,p} \quad & \frac{1}{2}\sum_{i=1}^{n_{\text{exp}}}\sum_{k=1}^{n_h^i}\sum_{j=1}^{n_t^i} w_{k,j}^i \frac{(\eta_{k,j}^i - h_k^i(x^i(t_j^i)))^2}{{\sigma_{k,j}^i}^2} \\ \text{s.t.} \quad & \dot{x}^i(t) = f(x^i(t), u^i(t), p), \quad t \in \mathcal{T}, \\ & x^i(0) = x_0^i. \end{aligned} \tag{3.1}$$

Here $n_{\text{exp}}, n_h^i, n_t^i$ indicate the number of independent experiments, number of different measurement functions per experiment, and number of time points for possible measurements per experiment, respectively. The $n_{\text{exp}} \cdot n_{\text{x}}$ dimensional differential state vector $(x^i)_{(i=1,\dots,n_{\text{exp}})}$ with $x^i : \mathcal{T} \mapsto \mathbb{R}^{n_x}$ is evaluated on a finite time grid $\{t_j^i\}$. The states $x^i(\cdot)$ of experiment i enter the model response functions $h_k^i : \mathbb{R}^{n_x} \mapsto \mathbb{R}^{n_{h_k^i}}$. The variances are denoted by $\sigma_{k,j}^i \in \mathbb{R}$, the sampling decisions $w_{k,j}^i \in \Omega$ denote how many measurements are taken at time t_j^i. If only one measurement is possible then $\Omega = \{0, 1\}$. We are also interested in the possibility of multiple measurements, then we have $\Omega = \{0, 1, \dots, w^{\max}\}$. The measurement errors leading to the measurement values $\eta_{k,j}^i$ are assumed to be random variables free of systematic errors, independent from one another, attributed with constant variances, distributed around a mean of zero, and distributed according to a common probability density function. All these assumptions lead to this special form of least squares minimization.

In the interest of a clearer presentation we neglect time-independent control values, such as initial values, consider only an unconstrained parameter estimation problem, assume we only do have one single measurement function per experiment, $n_h = n_h^i = 1$, and define all variances to be one, $\sigma_{k,j}^i = 1$. We need the following definitions.

DEFINITION 3.1 (solution of variational differential equations). *The matrix-valued maps* $\boldsymbol{G^i}(\cdot) = \frac{\mathbf{d}\boldsymbol{x^i}}{\mathbf{d}\boldsymbol{p}}(\cdot) : \mathcal{T} \mapsto \mathbb{R}^{n_x \times n_p}$ *are defined as the solutions of the variational differential equations*

$$\dot{\boldsymbol{G}}^{\boldsymbol{i}}(t) = \boldsymbol{f_x}(x^i(t), u^i(t), p)\boldsymbol{G^i}(t) + \boldsymbol{f_p}(x^i(t), u^i(t), p), \quad \boldsymbol{G^i}(0) = \mathbf{0}, \tag{3.2}$$

obtained from differentiating $x^i(t) = x_0^i + \int_{\mathcal{T}} f(x^i(\tau), u^i(\tau), p)\ \mathrm{d}\tau$ *with respect to time and parameters* $p \in \mathbb{R}^{n_p}$. *As they denote the dependency of differential states upon parameters, we also refer to* $\boldsymbol{G^i}(\cdot)$ *as* sensitivities. *Note that throughout the paper the ordinary differential equations are meant to hold componentwise for the matrices on both sides of the equation.*

DEFINITION 3.2 (Fisher information matrix). *The matrix* $\boldsymbol{F} = \boldsymbol{F}(t_{\text{f}}) \in \mathbb{R}^{n_p \times n_p}$ *defined by*

$$\boldsymbol{F}(t_{\text{f}}) = \sum_{i=1}^{n_{\text{exp}}}\sum_{j=1}^{n_t^i} w_j^i \left(\boldsymbol{h_x^i}(x^i(t_j^i))\boldsymbol{G^i}(t_j^i)\right)^T \boldsymbol{h_x^i}(x^i(t_j^i))\boldsymbol{G^i}(t_j^i)$$

is called the (discrete) Fisher information matrix.

DEFINITION 3.3 (covariance matrix). *The matrix* $\boldsymbol{C} = \boldsymbol{C}(t_\mathrm{f}) \in \mathbb{R}^{n_p \times n_p}$ *defined by*

$$\boldsymbol{C}(t_\mathrm{f}) = \boldsymbol{F}^{-1}(t_\mathrm{f})$$

is called the (discrete) covariance matrix of the unconstrained parameter estimation problem (3.1).

We assume that we have n_{exp} experiments for which we can determine control functions $u^i(\cdot)$ and sampling decisions w^i in the interest of optimizing a performance index, which is related to information gain with respect to the parameter estimation problem (3.1). As formulated in the groundbreaking work of [11], the optimum experimental design task is then to optimize over $u(\cdot)$ and w. The performance index is a function $\phi(\cdot)$ of either the Fisher information matrix $\boldsymbol{F}(t_\mathrm{f})$ or of it's inverse, the covariance matrix $\boldsymbol{C}(t_\mathrm{f})$.

DEFINITION 3.4 (objective OED functions). *We call*

- $\phi_A^F(\boldsymbol{F}(t_\mathrm{f})) := -\frac{1}{n_p} \operatorname{trace}(\boldsymbol{F}(t_\mathrm{f}))$ *the Fisher A-criterion,*
- $\phi_D^F(\boldsymbol{F}(t_\mathrm{f})) := -(\det(\boldsymbol{F}(t_\mathrm{f})))^{\frac{1}{n_p}}$ *the Fisher D-criterion,*
- $\phi_E^F(\boldsymbol{F}(t_\mathrm{f})) := -\min\{\lambda : \lambda \text{ is eigenval of } \boldsymbol{F}(t_\mathrm{f})\}$ *the Fisher E-criterion,*
- $\phi_A^C(\boldsymbol{F}(t_\mathrm{f})) := \frac{1}{n_p} \operatorname{trace}(\boldsymbol{F}^{-1}(t_\mathrm{f}))$ *the A-criterion,*
- $\phi_D^C(\boldsymbol{F}(t_\mathrm{f})) := (\det(\boldsymbol{F}^{-1}(t_\mathrm{f})))^{\frac{1}{n_p}}$ *the D-criterion,*
- $\phi_E^C(\boldsymbol{F}(t_\mathrm{f})) := \max\{\lambda : \lambda \text{ is eigenval of } \boldsymbol{F}(t_\mathrm{f})\}$ *the E-criterion,*

and write $\phi(\boldsymbol{F}(t_\mathrm{f}))$ *for any one of them in the following. If* $\phi \in \{\phi_A^F, \phi_D^F, \phi_E^F\}$ *we speak of a* Fisher objective function; *otherwise, if* $\phi \in \{\phi_A^C, \phi_D^C, \phi_E^C\}$, *of a* covariance objective function.

Note that maximizing a function (which we want to do for the Fisher information matrix) is equivalent to minimizing its negative. Additionally there are typically constraints on state and control functions, plus restrictions on the sampling decisions, such as a maximum number of measurements per experiment.

In this paper we follow the alternative formulation of [13], in which the sensitivities $\boldsymbol{G^i}(\cdot)$ and the Fisher information matrix function $\boldsymbol{F}(\cdot)$ are included as states in one structured optimal control problem. The performance index $\phi(\cdot)$ then has the form of a standard Mayer-type functional. The optimal control problem reads

$$\begin{aligned}
&\min_{x^i, \boldsymbol{G^i}, \boldsymbol{F}, z^i, u^i, w^i} \ \phi(\boldsymbol{F}(t_\mathrm{f})) \\
&\text{subject to} \\
&\quad \dot{x}^i(t) = f(x^i(t), u^i(t), p), \\
&\quad \dot{\boldsymbol{G}}^{\boldsymbol{i}}(t) = \boldsymbol{f_x}(x^i(t), u^i(t), p)\boldsymbol{G^i}(t) + \boldsymbol{f_p}(x^i(t), u^i(t), p), \\
&\quad \boldsymbol{F}(t_j^i) = \boldsymbol{F}(t_{j-1}^i) + \sum_{i=1}^{n_{\mathrm{exp}}} w_j^i \left(\boldsymbol{h_x^i}(x^i(t_j^i))\boldsymbol{G^i}(t_j^i)\right)^T \left(\boldsymbol{h_x^i}(x^i(t_j^i))\boldsymbol{G^i}(t_j^i)\right), \\
&\quad z^i(t_j^i) = z^i(t_{j-1}^i) + w_j^i, \\
&\quad x^i(0) = x_0, \\
&\quad \boldsymbol{G^i}(0) = \boldsymbol{0}, \\
&\quad \boldsymbol{F}(0) = \boldsymbol{0}, \\
&\quad z^i(0) = 0, \\
&\quad u^i(t) \in \mathcal{U}, \\
&\quad w_j^i \in \mathcal{W}, \\
&\quad 0 \leq M^i - z^i(t_\mathrm{f})
\end{aligned} \tag{3.3}$$

for experiment number $i = 1 \dots n_{\text{exp}}$, time index $j = 1 \dots n_t^i$, and $t \in \mathcal{T}$ almost everywhere. Note that the Fisher information matrix $\boldsymbol{F}(t_\text{f})$ is calculated as a discrete time state, just like the measurement counters $z^i(\cdot)$. The values $M^i \in \mathbb{R}$ give an upper bound on the possible number of measurements per experiment. Of course other problem formulations, e.g., a penalization of measurements via costs in the objective function, are also possible. In our study we exemplarily treat the case of an explicitly given upper bound.

The set $\mathcal{W}$ is either $\mathcal{W} = \Omega$ or its convex hull $\mathcal{W} = \operatorname{conv} \Omega$, i.e., either $\mathcal{W} = \{0, \dots, w^{\max}\}$ or $\mathcal{W} = [0, w^{\max}]$. In the first setting we refer to (3.3) as a mixed-integer optimal control problem (MIOCP). In the second case we use the term *relaxed* optimal control problem. It is the main aim of this paper to shed more light on the question under which circumstances the optimal solution of the relaxed problem (which is the outcome of most numerical approaches) is identical to the one of the MIOCP.

3.2. Problem formulation: Continuous measurements. It is interesting to also look at the case in which measurements are not performed at a single point in time, but over a whole interval. The continuous data flow would result in a slightly modified parameter estimation problem

$$\begin{aligned} \min_{x,p} \quad & \frac{1}{2} \sum_{i=1}^{n_{\text{exp}}} \int_0^{t_\text{f}} w^i(t) \cdot \frac{(\eta^i(t) - h^i(x^i(t)))^2}{\sigma^i(t)^2} \, \mathrm{d}t \\ \text{s.t.} \quad & \dot{x}^i(t) = f(x^i(t), u^i(t), p), \quad t \in \mathcal{T}, \\ & x^i(0) = x_0^i. \end{aligned} \tag{3.4}$$

This results in a modified definition of the Fisher information matrix.

DEFINITION 3.5 (Fisher information matrix). *The matrix* $\boldsymbol{F} = \boldsymbol{F}(t_\text{f}) \in \mathbb{R}^{n_p \times n_p}$ *defined by*

$$\boldsymbol{F}(t_\text{f}) = \sum_{i=1}^{n_{\text{exp}}} \int_0^{t_\text{f}} w^i(t) \left(\boldsymbol{h}_{\boldsymbol{x}}^{\boldsymbol{i}}(x^i(t)) \boldsymbol{G}^{\boldsymbol{i}}(t) \right)^T \boldsymbol{h}_{\boldsymbol{x}}^{\boldsymbol{i}}(x^i(t)) \boldsymbol{G}^{\boldsymbol{i}}(t) \, \mathrm{d}t$$

is called the (continuous) Fisher information matrix.

All other definitions from section 3.1 are identical. This allows us to formulate the OED problem as

$$\begin{aligned} \min_{x^i, \boldsymbol{G}^{\boldsymbol{i}}, \boldsymbol{F}, z^i, u^i, w^i} \quad & \phi(\boldsymbol{F}(t_\text{f})) \\ \text{subject to} \quad & \\ & \dot{x}^i(t) = f(x^i(t), u^i(t), p), \\ & \dot{\boldsymbol{G}}^{\boldsymbol{i}}(t) = \boldsymbol{f}_{\boldsymbol{x}}(x^i(t), u^i(t), p) \boldsymbol{G}^{\boldsymbol{i}}(t) + \boldsymbol{f}_{\boldsymbol{p}}(x^i(t), u^i(t), p), \\ & \dot{\boldsymbol{F}}(t) = \sum_{i=1}^{n_{\text{exp}}} w^i(t) \left(\boldsymbol{h}_{\boldsymbol{x}}^{\boldsymbol{i}}(x^i(t)) \boldsymbol{G}^{\boldsymbol{i}}(t) \right)^T \left(\boldsymbol{h}_{\boldsymbol{x}}^{\boldsymbol{i}}(x^i(t)) \boldsymbol{G}^{\boldsymbol{i}}(t) \right), \\ & \dot{z}^i(t) = w^i(t), \\ & x^i(0) = x_0, \\ & \boldsymbol{G}^{\boldsymbol{i}}(0) = \boldsymbol{0}, \\ & \boldsymbol{F}(0) = \boldsymbol{0}, \\ & z^i(0) = 0, \\ & u^i(t) \in \mathcal{U}, \\ & w^i(t) \in \mathcal{W}, \\ & 0 \leq M^i - z^i(t_\text{f}). \end{aligned} \tag{3.5}$$

Comparing (3.5) to the formulation (3.3) with measurements on the discrete time grid, one observes that now the states $\boldsymbol{F}(\cdot)$ and $z^i(\cdot)$ are specified by means of ordinary differential equations instead of difference equations, and the finite-dimensional control vector w now is a time-dependent integer control function $w(\cdot)$.

The two formulations have the advantage that they are separable, and hence accessible for the direct multiple shooting method [13]. In addition, they fall into the general optimal control formulations (2.5) and (2.1), respectively, and allow for an application of the maximum principle.

4. Analyzing relaxed sampling decisions. An observation in practice is that the optimized relaxed samplings $w^i(t) \in \operatorname{conv} \Omega$ are almost always $w^i(t) \in \Omega$. To get a better understanding of what is going on, we apply the maximum principle from Theorem (2.2). We proceed with the continuous case of the control problem (3.5). The vector of differential states of the general problem (2.1) is then given by

$$y(\cdot) = \begin{pmatrix} x^i(\cdot) \\ \boldsymbol{G^i}(\cdot) \\ VEC\boldsymbol{F}(\cdot) \\ z^i(\cdot) \end{pmatrix}_{(i=1\ldots n_{\exp})}$$

with $i = 1 \ldots n_{\exp}$. Hence $y(\cdot)$ is a map $y : \mathcal{T} \mapsto \mathbb{R}^{n_y}$ with dimension $n_y = n_{\exp} n_x + n_{\exp} n_x n_p + n_p n_p + n_{\exp}$. Note that some components of this vector are matrices that need to be "flattened" in order to write y as a vector. We define the right-hand-side function $\tilde{f} : \mathbb{R}^{n_y \times n_{\exp} n_u \times n_{\exp} \times n_p} \mapsto \mathbb{R}^{n_y}$ as

$$\tilde{f}(y(t), u(t), w(t), p) := \begin{pmatrix} f(x^i(t), u^i(t), p) \\ \boldsymbol{f_x}(x^i(t), u^i(t), p)\boldsymbol{G^i}(t) + \boldsymbol{f_p}(x^i(t), u^i(t), p) \\ \sum_{i=1}^{n_{\exp}} w^i(t)\big(\boldsymbol{h_x^i}(x^i(t))\boldsymbol{G^i}(t)\big)^T \big(\boldsymbol{h_x^i}(x^i(t))\boldsymbol{G^i}(t)\big) \\ w^i(t) \end{pmatrix} \tag{4.1}$$

again with multiple entries for all $i = 1 \ldots n_{\exp}$. We define $\lambda_{x^i}, \boldsymbol{\lambda_{G^i}}, \boldsymbol{\lambda_F}, \lambda_{z^i}$ to be corresponding adjoint variables with dimensions n_x, $n_x \times n_p$, $n_p \times n_p$, and 1, respectively, and λ as the compound of these variables. Note that $\boldsymbol{\lambda_{G^i}}$ and $\boldsymbol{\lambda_F}$ are treated as matrices, just like their associated states $\boldsymbol{G^i}$ and $\boldsymbol{F}$. The Hamiltonian is then given as

$$\begin{aligned} \mathcal{H}(y(t), u(t), w(t), \lambda(t), p) &= \Big\langle \lambda(t), \tilde{f}(y(t), u(t), w(t), p) \Big\rangle \\ &= \sum_{i=1}^{n_{\exp}} \lambda_{x^i}^T f^i(\cdot) + \sum_{i=1}^{n_{\exp}} \Big\langle \boldsymbol{\lambda_{G^i}}, \boldsymbol{f_x^i}(\cdot)\boldsymbol{G^i} + \boldsymbol{f_p^i}(\cdot) \Big\rangle \\ &\quad + \left\langle \boldsymbol{\lambda_F}, \sum_{i=1}^{n_{\exp}} w^i \left(\boldsymbol{h_x^i}(\cdot)\boldsymbol{G^i}\right)^T \left(\boldsymbol{h_x^i}(\cdot)\boldsymbol{G^i}\right) \right\rangle + \sum_{i=1}^{n_{\exp}} \lambda_{z^i} w^i, \end{aligned} \tag{4.2}$$

where we are omitting the time arguments (t) and argument lists of f and h. Note that Definition 1.1 of the scalar product allows us to use the matrices $\boldsymbol{\lambda_{G^i}} \in \mathbb{R}^{n_x \times n_p}$ and $\boldsymbol{\lambda_F} \in \mathbb{R}^{n_p \times n_p}$ in a straightforward way.

COROLLARY 4.1 (maximum principle for OED problems). *Let problem* (3.5) *have a feasible optimal solution* (y^*, u^*, w^*). *Then there exist adjoint variables* $\lambda^*(\cdot)$ *and Lagrange multipliers* $\mu^* \in \mathbb{R}^{n_{\exp}}$ *such that for* $t \in \mathcal{T}$ *it holds almost everywhere*

$$\dot{y}^*(t) = \tilde{f}(y^*(t), u^*(t), w^*(t), p), \tag{4.3a}$$

$$\dot{\lambda_{x^i}}^{*T}(t) = \lambda_{x^i}^{*T} \boldsymbol{f_x^i}(\cdot) + \frac{\partial}{\partial x^i}\left(\left\langle \boldsymbol{\lambda_{G^i}}^*, \boldsymbol{f_x^i}(\cdot)\boldsymbol{G^{i*}} + \boldsymbol{f_p^i}(\cdot)\right\rangle\right)^T + \frac{\partial}{\partial x^i}\left(\left\langle \boldsymbol{\lambda_F}^*, w^{i*}\left(\boldsymbol{h_x^i}(\cdot)\boldsymbol{G^{i*}}\right)^T\left(\boldsymbol{h_x^i}(\cdot)\boldsymbol{G^{i*}}\right)\right\rangle\right)^T, \tag{4.3b}$$

$$\dot{\boldsymbol{\lambda}}_{\boldsymbol{G^i}}^{\boldsymbol{*T}}(t) = \left\langle \boldsymbol{\lambda_{G^i}}^*, \boldsymbol{f_x^i}(\cdot)\right\rangle + \frac{\partial}{\partial \boldsymbol{G^i}}\left(w^{i*}\left\langle \boldsymbol{\lambda_F}^*, \left(\boldsymbol{h_x^i}(\cdot)\boldsymbol{G^{i*}}\right)^T\left(\boldsymbol{h_x^i}(\cdot)\boldsymbol{G^{i*}}\right)\right\rangle\right)^T, \tag{4.3c}$$

$$\dot{\boldsymbol{\lambda}}_{\boldsymbol{F}}^{\boldsymbol{*T}}(t) = \mathbf{0}, \tag{4.3d}$$

$$\dot{\lambda_{z^i}}^{*T}(t) = 0, \tag{4.3e}$$

$$y^*(0) = y_0, \tag{4.3f}$$

$$\lambda_{x^i}^{*T}(t_{\mathrm{f}}) = 0, \tag{4.3g}$$

$$\boldsymbol{\lambda}_{\boldsymbol{G^i}}^{\boldsymbol{*T}}(t_{\mathrm{f}}) = \mathbf{0}, \tag{4.3h}$$

$$\boldsymbol{\lambda}_{\boldsymbol{F}}^{\boldsymbol{*T}}(t_{\mathrm{f}}) = -\frac{\partial \phi(\boldsymbol{F}(t_{\mathrm{f}}))}{\partial \boldsymbol{F}}, \tag{4.3i}$$

$$\lambda_{z^i}^{*T}(t_{\mathrm{f}}) = -\frac{-\partial \mu_i^*(M^i - z^{i*}(t_{\mathrm{f}}))}{\partial z} = -\mu_i^*, \tag{4.3j}$$

$$(u^*, w^*)(t) \in \arg\min_{u \in \mathcal{U}^{n_{\exp}}, w \in \mathcal{W}^{n_{\exp}}} \mathcal{H}(y^*(t), u, w, \lambda^*(t), p), \tag{4.3k}$$

$$0 \leq M - z^*(t_{\mathrm{f}}), \tag{4.3l}$$

$$0 \leq \mu^*, \tag{4.3m}$$

$$0 = \mu^{*T}(M - z^*(t_{\mathrm{f}})) \tag{4.3n}$$

with $i = 1 \dots n_{\exp}$ *and* $y, \lambda, \tilde{f}$ *defined as above.*

Proof. The proof follows directly from applying the maximum principle (2.4) to the control problem (3.5) and taking the partial derivatives of the Hamiltonian (4.2) and the objective function $\phi(\cdot)$ of the OED control problem with respect to the state variables $x^i(\cdot), \boldsymbol{G^i}(\cdot), \boldsymbol{F}(\cdot)$, and $z^i(\cdot)$. ∎

This corollary serves as a basis for further analysis. A closer look at (4.3k) and the Hamiltonian reveals structure.

COROLLARY 4.2. *The Hamiltonian* $\mathcal{H}$ *decouples with respect to* $u^i(\cdot)$ *and* $w^i(\cdot)$ *for all experiments* $i = 1 \dots n_{\exp}$. *Hence the optimal controls* $u^{i*}(\cdot)$ *and* $w^{i*}(\cdot)$ *can be determined independently from one another for given states* $y^*(\cdot)$, *adjoints* $\lambda^*(\cdot)$, *and parameters* p.

Proof. The proof follows directly from (4.2) and the fact that $f^i(\cdot)$ and the partial derivatives $\boldsymbol{f_x^i}(\cdot)$ and $\boldsymbol{f_p^i}(\cdot)$ do not depend on the sampling functions $w^i(\cdot)$. Let $\tilde{w}^T = (w^{1,*^T}(t), \dots, w^{i-1,*^T}(t), w^{i^T}, w^{i+1,*^T}(t), \dots, w^{n_{\exp},*^T})(t)$, then

$$\begin{aligned} w^{i*}(t) &\in \arg\min_{w^i \in \mathcal{W}} \mathcal{H}(y^*(t), u^*(t), \tilde{w}, \lambda^*(t), p) \\ &= \arg\min_{w^i \in \mathcal{W}} \left\langle \boldsymbol{\lambda_F}^*, w^i \left(\boldsymbol{h_x^i}(\cdot)\boldsymbol{G^{i*}}\right)^T \left(\boldsymbol{h_x^i}(\cdot)\boldsymbol{G^{i*}}\right) \right\rangle + \lambda_{z^i}^* w^i. \end{aligned} \tag{4.4}$$

Likewise, the experimental controls $u^{i*}(\cdot)$ are given as

$$\begin{aligned} u^{i*}(t) &\in \arg\min_{u^i \in \mathcal{U}} \mathcal{H}(y^*(t), \tilde{u}, w^*(t), \lambda^*(t), p) \\ &= \arg\min_{u^i \in \mathcal{U}} \lambda_{x^i}^{*T} f^i(\cdot) + \left\langle \boldsymbol{\lambda_{G^i}}^*, \boldsymbol{f_x^i}(\cdot)\boldsymbol{G^{i*}} + \boldsymbol{f_p^i}(\cdot) \right\rangle \end{aligned} \tag{4.5}$$

because the measurement function $h(\cdot)$ and its partial derivative do not depend explicitly on $u(\cdot)$. ∎

We would like to stress that the decoupling of the control functions holds only in the sense of necessary conditions of optimality, and for given optimal states and adjoints. Clearly they may influence one another indirectly. We come back to this issue in section 4.1.

A closer look at (4.4) reveals that the sampling control function $w(\cdot)$ enters linearly into the Hamiltonian. This implies that the sign of the switching function determines whether $w(\cdot) \in [0, w^{\max}]$ is at its lower or upper bound, which corresponds in our case to integer feasibility, $w(\cdot) \in \{0, w^{\max}\}$.

DEFINITION 4.3 (local and global information gain). *The matrix* $\boldsymbol{P^i}(t) \in \mathbb{R}^{n_p \times n_p}$,

$$\boldsymbol{P^i}(t) := \boldsymbol{P}(x^i(t), \boldsymbol{G^i}(t)) := \left(\boldsymbol{h_x^i}(x^i(t))\boldsymbol{G^i}(t)\right)^T \left(\boldsymbol{h_x^i}(x^i(t))\boldsymbol{G^i}(t)\right),$$

is called the local information gain *matrix of experiment i. Note that $\boldsymbol{P^i}(t)$ is positive semidefinite, and positive definite if the matrix $\boldsymbol{h_x^i}(x^i(t))\boldsymbol{G^i}(t)$ has full rank n_p.*

If $\boldsymbol{F^*}^{-1}(t_{\mathrm{f}})$ exists, we call

$$\boldsymbol{\Pi^i}(t) := \boldsymbol{F^*}^{-1}(t_{\mathrm{f}})\boldsymbol{P^i}(t)\boldsymbol{F^*}^{-1}(t_{\mathrm{f}}) \in \mathbb{R}^{n_p \times n_p}$$

the global information gain *matrix.*

We use Corollary 4.2 as a justification to concentrate our analysis on the case of a single experiment. Hence we leave the superscript i away for notational convenience, assuming $n_{\exp} = 1$, and come back to the multiexperiment case in section 4.3.

DEFINITION 4.4 (switching function). *The derivative of the Hamiltonian* (4.2)

$$\mathcal{H}_w(t) := \frac{\partial \mathcal{H}(\cdot)}{\partial w} = \langle \boldsymbol{\lambda_F}(t), \boldsymbol{P}(t) \rangle + \lambda_z(t)$$

is called the switching function with respect to $w(\cdot)$. The derivative

$$\mathcal{H}_u(t) := \frac{\partial \mathcal{H}(\cdot)}{\partial u} = \frac{\partial}{\partial u}\left(\lambda_x^{*T} f(\cdot) + \left\langle \boldsymbol{\lambda_G}^*, \boldsymbol{f_x}(\cdot)\boldsymbol{G^{i*}} + \boldsymbol{f_p}(\cdot) \right\rangle\right)$$

is called the switching function with respect to $u(\cdot)$.

We are now set to investigate the conditions for either measuring or not at a time t for different objective functions. From now on we assume that $(y^*, u^*, w^*, \lambda^*, \mu^*)(\cdot)$ is an optimal trajectory of the relaxed optimal control problem (3.5) with $n_{\text{exp}} = 1$ and $\mathcal{W} = [0, w^{\max}]$, and hence a solution of the boundary value problem (4.3).

LEMMA 4.5 (maximize trace of Fisher matrix). *Let* $\phi(\boldsymbol{F}(t_\text{f})) = \phi_A^F(\boldsymbol{F}(t_\text{f})) = -\text{trace}(\boldsymbol{F}(t_\text{f}))$ *be the objective function of the OED problem* (3.5)*, and let* $w^*(\cdot)$ *be an optimal control function. If*

$$\text{trace}\,(\boldsymbol{P}(t)) > \mu^*$$

for $t \in (0, t_\text{f})$*, then there exists a* $\delta > 0$ *such that* $w^*(t) = w^{max}$ *almost everywhere on* $[t - \delta, t + \delta]$.

Proof. As $w^*(t)$ is the pointwise minimizer of the Hamiltonian and according to Corollary 4.2 it decouples from the other control functions, and as it enters linearly, it is at its upper bound of $w^{\max}$ whenever the sign of the switching function is positive. The switching function is given by

$$\mathcal{H}_w(t) = \left\langle \boldsymbol{\lambda}_{\boldsymbol{F}}^*(t), \boldsymbol{P}(t) \right\rangle + \lambda_z^*(t).$$

With Corollary 4.1 we have

$$\begin{aligned}\mathcal{H}_w(t) &= \left\langle -\frac{\partial \phi(\boldsymbol{F}(t_\text{f}))}{\partial \boldsymbol{F}}, \boldsymbol{P}(t) \right\rangle - \mu^* \\ &= \left\langle -\frac{\partial -\text{trace}(\boldsymbol{F}(t_\text{f}))}{\partial \boldsymbol{F}}, \boldsymbol{P}(t) \right\rangle - \mu^*.\end{aligned}$$

Applying Lemma A.2 from the appendix we obtain

$$\mathcal{H}_w(t) = \text{trace}\,(\boldsymbol{P}(t)) - \mu^*.$$

As $\text{trace}(\boldsymbol{P}(t))$ is differentiable with respect to time, there exists a time interval of positive measure around t where this expression is also positive, which concludes the proof. ∎

LEMMA 4.6 (minimize trace of covariance matrix). *For the assumptions of Lemma* 4.5*, but the objective function*

$$\phi(\boldsymbol{F}(t_\text{f})) = \phi_C^F(\boldsymbol{F}(t_\text{f})) = \text{trace}(\boldsymbol{C}(t_\text{f})),$$

the sufficient condition for $w^*(t) = w^{max}$ *in an optimal solution is that*

$$\text{trace}\,(\boldsymbol{\Pi}(t)) > \mu^*$$

holds.

Proof. The argument is similar to the one in Lemma 4.5. We have

$$\begin{aligned}\mathcal{H}_w(t) &= -\left\langle \frac{\partial \text{trace}(\boldsymbol{F}^{*-1}(t_\text{f}))}{\partial \boldsymbol{F}}, \boldsymbol{P}(t) \right\rangle - \mu^* \\ &= -\left\langle \frac{\partial \text{trace}(\boldsymbol{F}^{*-1}(t_\text{f}))}{\partial \boldsymbol{F}^{-1}}, \frac{\partial \boldsymbol{F}^{*-1}(t_\text{f})}{\partial \boldsymbol{F}} \boldsymbol{P}(t) \right\rangle - \mu^*.\end{aligned}$$

Note here that the expression $\frac{\partial \boldsymbol{F}^{*-1}(t_\mathrm{f})}{\partial \boldsymbol{F}}\boldsymbol{P}(t)$ is a matrix in $\mathbb{R}^{n_p \times n_p}$ by virtue of Definition 1.2. Applying Lemma A.2 from the appendix we obtain

$$\mathcal{H}_w(t) = -\mathrm{trace}\left(\frac{\partial \boldsymbol{F}^{*-1}(t_\mathrm{f})}{\partial \boldsymbol{F}}\boldsymbol{P}(t)\right) - \mu^*.$$

To evaluate the directional derivative of the inverse operation we apply Lemma A.3 and obtain

$$\mathcal{H}_w(t) = \mathrm{trace}\left(\boldsymbol{F}^{*-1}(t_\mathrm{f})\boldsymbol{P}(t)\boldsymbol{F}^{*-1}(t_\mathrm{f})\right) - \mu^*$$

which concludes the proof, as $\boldsymbol{\Pi}(t) = \boldsymbol{F}^{*-1}(t_\mathrm{f})\boldsymbol{P}(t)\boldsymbol{F}^{*-1}(t_\mathrm{f})$. □

LEMMA 4.7 (minimization of max eigenvalue of covariance matrix). *For the assumptions of Lemma* 4.5, *but the objective function*

$$\phi(\boldsymbol{F}(t_\mathrm{f})) = \phi_E^C(\boldsymbol{F}(t_\mathrm{f})) = \max\{\lambda : \lambda \textit{ is eigenvalue of } \boldsymbol{C}(t_\mathrm{f})\},$$

the sufficient condition for $w^*(t) = w^{max}$ *in an optimal solution is that, if* $\lambda_{\max}$ *is a single eigenvalue,*

$$v^T \boldsymbol{\Pi}(t) v > \mu^*$$

holds, where $v \in \mathbb{R}^{n_p}$ *is an eigenvector of* $\boldsymbol{C}(t_\mathrm{f})$ *to* $\lambda_{\max}$ *with norm* 1.

LEMMA 4.8 (minimization of determinant of covariance matrix). *For the assumptions of Lemma* 4.5, *but the objective function*

$$\phi(\boldsymbol{F}(t_\mathrm{f})) = \phi_D^C(\boldsymbol{F}(t_\mathrm{f})) = \det(\boldsymbol{C}(t_\mathrm{f})),$$

the sufficient condition for $w^*(t) = w^{max}$ *in an optimal solution is that*

$$\det(\boldsymbol{C}^*(t_\mathrm{f})) \sum_{i,j=1}^{n_p} (\boldsymbol{F}^*(t_\mathrm{f}))_{i,j} (\boldsymbol{\Pi}(t))_{i,j} > \mu^*$$

holds.

The proofs of Lemmas 4.7 and 4.8 and for other objective functions are similar to the one in Lemma 4.6, making use of the appendix Lemmas A.4 and A.5.

The local information gain matrix $\boldsymbol{P}(t)$ is positive definite, whenever the measurement function is sensitive with respect to the parameters. This attribute carries over to the matrix state $\boldsymbol{F}(\cdot)$ in which $\boldsymbol{P}(t)$ is integrated, to the covariance matrix function (as the inverse of a positive definite matrix is also positive definite), and to the product of positive definite matrices. The considered functions of $\boldsymbol{P}(t)$ and $\boldsymbol{\Pi}(t)$ are hence all positive values; compare, e.g., Lemma A.1.

This implies for nonexistent constraints on the number of measurements with $\mu^* = 0$ the trivial conclusion that measuring all the time with $w(t) \equiv w^{\max}$ is optimal.

In the more interesting case when the constraint $c(z^*(t_\mathrm{f})) = M - z^*(t_\mathrm{f}) \geq 0$ is active, the Lagrange multiplier μ^* indicates the threshold. The Lagrange multipliers are also called shadow prices, as they indicate how much one gains from increasing a resource. In this particular case relaxing the measurement bound M yields the information gain μ^* in the objective function $\phi(\cdot)$.

The main difference between using the Fisher information matrix $\boldsymbol{F}(\cdot)$ and the covariance matrix $\boldsymbol{C}(\cdot) = \boldsymbol{F}^{-1}(\cdot)$, e.g., in Lemmas 4.5 and 4.6, lies in the local $\boldsymbol{P}(t)$

and global $\boldsymbol{\Pi}(t) = \boldsymbol{F}^{-1}(t_\mathrm{f})\boldsymbol{P}(t)\boldsymbol{F}^{-1}(t_\mathrm{f})$ information gain matrices that yield a sufficient criterion, respectively. The fact that the sufficient criterion for a maximization of the Fisher information matrix does not depend on the value of $\boldsymbol{F}^{-1}(t_\mathrm{f})$ has an important consequence. Modifying the value of $w(t)$, e.g., by rounding, does not have any recoupling effect on the criterion itself. Therefore, whenever $w(t) \notin \{0, w^{\max}\}$ on different time intervals, one can round these values up and down (making sure that $\int_{\mathcal{T}} w(\tau)\,\mathrm{d}\tau$ keeps the value of M) to obtain a feasible integer solution with the same objective function value. This is *not* the case when we have a covariance objective function, as measurable modifications of $w(t)$ have an impact on $\boldsymbol{F}(t_\mathrm{f})$ and hence also on $\boldsymbol{F}^{-1}(t_\mathrm{f})$ and the sufficient criterion.

The procedure for the case with finitely many measurements that enter as non-continuous jumps in finite difference equations (3.3) is very similar to the one above, only some definitions need to be modified. The main results are identical and we have the same criteria to validate whether the control values w_j^i are on their upper bound of $w^{\max}$ or not. The main difference is that measurements in the continuous setting average the information gain on a time interval, whereas point measurements are on the exact location of the maxima of the global information gain function.

4.1. Singular arcs. As we saw above, the sampling controls $w(t)$ enter linearly into the control problem. If for control problems with linear controls the switching function is zero on an interval of positive measure, one usually proceeds by taking higher order time derivatives of the switching function to determine an explicit representation of this singular control, which may occur if at all in even degree time derivatives as shown by [10]. This approach is not successful for sampling functions in experimental design problems.

LEMMA 4.9 (infinite order of singular arcs). *Let $n_u = 0$. For all values $j \in \mathbb{N}$ the time derivatives $S^j := \frac{\mathrm{d}^j}{\mathrm{d}t^j}\mathcal{H}_w(t)$ never depend explicitly on $w(\cdot)$.*

Proof. The switching functions above are functions of either $\boldsymbol{P}(t)$ or in the case of a covariance objective function of $\boldsymbol{F}^{*-1}(t_\mathrm{f})\boldsymbol{P}(t)\boldsymbol{F}^{*-1}(t_\mathrm{f})$. Taking the time derivative only affects $\boldsymbol{P}(t)$. We see that in

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{P}(t)}{\mathrm{d}t} &= \frac{\mathrm{d}\left(\boldsymbol{h}_{\boldsymbol{x}}(x(t))\boldsymbol{G}(t)\right)^T\left(\boldsymbol{h}_{\boldsymbol{x}}(x(t))\boldsymbol{G}(t)\right)}{\mathrm{d}t} \\
&= 2\left(\boldsymbol{h}_{\boldsymbol{x}}(x(t))\boldsymbol{G}(t)\right)^T\frac{\mathrm{d}\left(\boldsymbol{h}_{\boldsymbol{x}}(x(t))\boldsymbol{G}(t)\right)}{\mathrm{d}t} \\
&= 2\left(\boldsymbol{h}_{\boldsymbol{x}}(x(t))\boldsymbol{G}(t)\right)^T\left(\boldsymbol{h}_{\boldsymbol{xx}}(x(t))\dot{x}(t)\boldsymbol{G}(t) + \boldsymbol{h}_{\boldsymbol{x}}(x(t))\dot{\boldsymbol{G}}(t)\right) \\
&= 2\left(\boldsymbol{h}_{\boldsymbol{x}}(x(t))\boldsymbol{G}(t)\right)^T\left(\boldsymbol{h}_{\boldsymbol{xx}}(x(t))f(x(t),u(t),p)\boldsymbol{G}(t)\right. \\
&\quad \left. + \boldsymbol{h}_{\boldsymbol{x}}(x(t))(\boldsymbol{f}_{\boldsymbol{x}}(x(t),u(t),p)\boldsymbol{G}(t) + \boldsymbol{f}_{\boldsymbol{p}}(x(t),u(t),p))\right)
\end{aligned}$$

only time derivatives of $x(\cdot)$ and $\boldsymbol{G}(\cdot)$ appear. Also in higher order derivatives $\boldsymbol{F}(\cdot)$ and $z(\cdot)$ never enter, and as $n_u = 0$ no expressions from a singular control $u^*(\cdot)$ may appear, hence also $w(\cdot)$ never enters in any derivative. ∎

In the special case that $n_u = 0$, hence all controls enter linearly, another type of argument can also be applied. In [7] and also in [15] it is shown that under certain conditions when linearly entering controls are relaxed[1] towards a probability distribution $P(u(t))$, then the support of P is contained in the set on which the

[1] Attention: here *relax* does not refer to a relaxation between integer values $\{0, 1\}$ towards their convex hull $[0, 1]$ as in the rest of this paper!

Hamiltonian attains its minimum value. The boundary of the feasible set, $\{0,1\}^{n_{\exp}}$, can be chosen as the support by means of a simple convex combination. Thus, a control $\alpha^*(t)$ with

$$0 < k_1 \leq \alpha^*(t) \leq k_2 < 1$$

on an interval $[t_1, t_2]$ can be replaced by either 0 or 1 in this interval, and the maximum principle is satisfied for the new control too.

This is the case for $w^i(\cdot)$ when $(\boldsymbol{h}^{\boldsymbol{i}}_{\boldsymbol{x}}(x^i(t))\boldsymbol{G}^{\boldsymbol{i}}(t))^T \boldsymbol{h}^{\boldsymbol{i}}_{\boldsymbol{x}}(x^i(t))\boldsymbol{G}^{\boldsymbol{i}}(t) \equiv 0$ on $[t_1, t_2]$ (compare (3.5)), corresponding to an interval in which parameters of the systems cannot be identified because of nonexistent sensitivities.

The assumption that $n_u = 0$ is rather strong though. It is an open and interesting question, whether one can construct nontrivial instances of OED control problems for which the joint control vector $(u, w)(\cdot)$ is a singular control. This implies that the interplay of the singular controls results in a constant value of the global information gain matrix $\boldsymbol{\Pi}(t)$ on a measurable time interval.

Numerically, such sensitivity-seeking cases (following the terminology of [26]) can be resolved efficiently making use of the sum up rounding strategy, which we shortly apply to the OED case for the convenience of the reader. Note, however, that the question remains open if this algorithm to generate integer feasible solutions from noninteger ones in linear time and with guaranteed error bounds needs to be applied to OED problems at all.

4.2. Applying the integer gap lemma to OED. A first immediate advantage of the formulation (3.5) as a continuous optimal control problem is that we can apply the integer gap lemma proposed in [20]. In the interest of an easier presentation let us assume $w^{\max} = 1$. We first recall the sum up rounding strategy. We consider given measurable functions $\alpha^i : [0, t_{\mathrm{f}}] \mapsto [0,1]$ with $i = 1 \dots n_{\exp}$ and a time grid $0 = t_0 < t_1 < \cdots < t_m = t_{\mathrm{f}}$ on which we approximate the controls $\alpha^i(\cdot)$. We write $\Delta t_j := t_{j+1} - t_j$ and Δt for the maximum distance,

$$\Delta t := \max_{j=0\dots m-1} \Delta t_j = \max_{j=0\dots m-1} \{t_{j+1} - t_j\}. \tag{4.6}$$

Let then a function $\omega(\cdot) : [0, t_{\mathrm{f}}] \mapsto \{0,1\}^{n_{\exp}}$ be defined by

$$\omega^i(t) = p_{i,j}, \quad t \in [t_j, t_{j+1}), \tag{4.7}$$

where for $i = 1 \dots n_{\exp}$ and $j = 0 \dots m-1$ the $p_{i,j}$ are binary values given by

$$p_{i,j} = \begin{cases} 1 & \text{if } \int_0^{t_{j+1}} \alpha^i(\tau)\mathrm{d}\tau - \sum_{k=0}^{j-1} p_{i,k}\Delta t_k \geq 0.5\Delta t_j, \\ 0 & \text{else.} \end{cases} \tag{4.8}$$

We can now formulate the following corollary.

COROLLARY 4.10 (integer gap). *Let $(x^{i*}, \boldsymbol{G}^{\boldsymbol{i}*}, \boldsymbol{F}^*, z^{i*}, u^{i*}, \alpha^{i*})(\cdot)$ be a feasible trajectory of the relaxed problem (3.5) with the measurable functions $\alpha^{i*} : [0, t_{\mathrm{f}}] \to [0,1]$ replacing $w^i(\cdot)$ in problem (3.5), with $i = 1 \dots n_{\exp}$.*

Consider the trajectory $(x^{i*}, \boldsymbol{G}^{\boldsymbol{i}*}, \boldsymbol{F}^{SUR}, z^{i,SUR}, u^{i*}, \omega^{i,SUR})(\cdot)$ which consists of controls $\omega^{i,SUR}(\cdot)$ determined via sum up rounding (4.7)–(4.8) on a given time grid from $\alpha^{i*}(\cdot)$ and differential states $(\boldsymbol{F}^{SUR}, z^{i,SUR})(\cdot)$ that are obtained by solving the initial value problems in (3.5) for the fixed differential states $(x^{i*}, \boldsymbol{G}^{\boldsymbol{i}*})(\cdot)$ and $\omega^{i,SUR}(\cdot)$.

Then there exists a constant $\bar{C}$ such that

$$|z^{i,SUR}(t_\mathrm{f}) - z^{i*}(t_\mathrm{f})| \leq \bar{C}\Delta t, \quad i = 1, \dots, n_\mathrm{exp}. \tag{4.9}$$

Assume in addition that constants $C, M \in \mathbb{R}^+$ exist such that the functions

$$\hat{f}^i(x^{i*}, \boldsymbol{G}^{i*}) := \left(\boldsymbol{h}^{\boldsymbol{i}}_{\boldsymbol{x}}(x^i(t))\boldsymbol{G}^{\boldsymbol{i}}(t)\right)^T \left(\boldsymbol{h}^{\boldsymbol{i}}_{\boldsymbol{x}}(x^i(t))\boldsymbol{G}^{\boldsymbol{i}}(t)\right)$$

are differentiable with respect to time and it holds

$$\left\| \frac{\mathrm{d}}{\mathrm{d}t}\hat{f}^i(x^{i*}, \boldsymbol{G}^{i*}) \right\| \leq C$$

for all $i = 1 \dots n_\mathrm{exp}$, $t \in [0, t_\mathrm{f}]$ almost everywhere, and $\hat{f}^i(x^{i*}, \boldsymbol{G}^{i*})$ are essentially bounded by M. Then there exists a constant $\hat{C}$ such that

$$|\phi(\boldsymbol{F}^{SUR}(t_\mathrm{f})) - \phi(\boldsymbol{F}^*(t_\mathrm{f}))| \leq \hat{C}\Delta t. \tag{4.10}$$

Proof. The proof follows from Corollary 8 in [20] and the fact that all assumptions on the right-hand-side function are fulfilled. Note that the condition on the Lipschitz constant is automatically fulfilled, because $z(\cdot)$ and $\boldsymbol{F}(\cdot)$ do not enter in the right-hand side of the differential equations. ∎

Corollary 4.10 implies that the exact lower bound of the OED problem (3.5) can be obtained by solving the relaxed problem in which $w^i(t) \in \operatorname{conv} \Omega$ instead of $w^i(t) \in \Omega$. In other words, anything that can be done with fractional sampling can also be done with an integer number of measurements. However, the price might be a so-called chattering behavior, i.e., frequent switching between yes and no. Note that the famous *bang-bang principle* and the mentioned references [7, 15] state similar results, however, without the linear grid dependence of the Hausdorff distance that can be exploited numerically by means of an adaptive error control.

4.3. L^1 penalization and sparse controls. We are interested in how changes in the formulation of the optimization problem influence the role of the global information gain functions. We first consider an L^1 penalty term in the objective function. We are going back to the multiexperiment case and use the upperscript $i = 1 \dots n_\mathrm{exp}$.

COROLLARY 4.11 (switching function for L^1 penalty). *Let $\mathcal{H}^{\mathrm{old}}_{w^i}(\cdot)$ denote the switching function for problem* (3.5) *and $\mathcal{H}^{\mathrm{penalty}}_{w^i}(\cdot)$ the switching function with respect to $w^i(\cdot)$ for problem* (3.5) *with an objective function that is augmented by a Lagrange term,*

$$\min_{x^i, \boldsymbol{G}^{\boldsymbol{i}}, \boldsymbol{F}, z^i, u^i, w^i} \phi(\boldsymbol{F}(t_\mathrm{f})) + \int_{\mathcal{T}} \sum_{i=1}^{n_\mathrm{exp}} \epsilon^i w^i(\tau)\, \mathrm{d}\tau.$$

Assume that the maximum principle holds in normal form, i.e., $\lambda_0 = 1$. Then it holds

$$\mathcal{H}^{\mathrm{penalty}}_{w^i}(t) = \mathcal{H}^{\mathrm{old}}_{w^i}(t) - \epsilon^i.$$

Proof. By definition (2.2) of the Hamiltonian we have

$$\mathcal{H}^{\mathrm{penalty}}(t) = \mathcal{H}^{\mathrm{old}}(t) - \sum_{i=1}^{n_\mathrm{exp}} \epsilon^i w^i(t)$$

which already concludes the proof. ∎

Corollary 4.11 allows a direct connection between the penalization parameter ϵ and the information gain function. For the minimization of the trace of the covariance matrix (compare Lemma 4.6), this implies that a sufficient condition for $w^{i*}(t) = w^{\max}$ is

$$\operatorname{trace}\left(\mathbf{\Pi^i}(t)\right) > \epsilon^i + \mu^{i*}.$$

As a consequence, an optimal sampling design never performs measurements when the value of the trace of the information gain function is below the penalization parameter ϵ^i.

The case is similar for the time discrete OED problem (3.3). Assume we extend the objective with a penalization term

$$\sum_{i=1}^{n_{\exp}} \sum_{j=1}^{n_t^i} L^{\mathrm{tr}}(w_j^i) = \sum_{i=1}^{n_{\exp}} \sum_{j=1}^{n_t^i} \epsilon^i \; w_j^i,$$

then the derivative of the discrete Hamiltonian (2.8) with respect to the control w_j^i is again augmented by $-\epsilon^i$.

4.4. L^2 penalization and singular arcs. An alternative penalization is a L^2 penalization of the objective function with a Lagrange term

$$\int_{\mathcal{T}} \sum_{i=1}^{n_{\exp}} \epsilon^i \; w^i(\tau)^2 \; \mathrm{d}\tau.$$

This formulation has direct consequences. As the controls $u^i(\cdot)$ and $w^i(\cdot)$ decouple (compare Corollary 4.2), the optimal sampling design may be on the boundary of its domain, or can be determined via the necessary condition that the derivative of the Hamiltonian with respect to $w^i(\cdot)$ is zero, i.e.,

$$w^i(t) = \frac{1}{2\epsilon^i} \left[\operatorname{trace}\left(\mathbf{\Pi^i}(t)\right) - \mu^{i*}\right]$$

for the case of the minimization of the trace of the covariance matrix. This implies that $w^i(\cdot)$ may be a singular control with fractional values $w(t) \in (0, w^{\max})$. Hence, we discourage use of this formulation.

5. Numerical examples. In this section we illustrate several effects with numerical examples. Our analysis so far has been based on the so-called *first optimize, then discretize* approach. Now we solve the numerical OED problems with direct or *first-discretize–then-optimize* methods. In particular, we use the code `MS MINTOC` that has been developed for generic mixed-integer optimal control problems by the author. It is based on Bock's direct multiple shooting method, adaptive control discretizations, and switching time optimization. A comprehensive survey of how this algorithm works can be found in [22]. Note however that there are many specific structures that can, should, or even have to be exploited to take into account the special structure of the OED control problems in an efficient implementation. It is beyond the scope of this paper to go into details, instead we refer to [9, 13] for a more detailed discussion.

Having obtained an optimal solution, it is possible to evaluate the functions $\mathbf{\Pi^i}(t)$ for an a posteriori analysis. This is what we do in the following. As we have derived

an explicit formula for the switching functions $\mathbf{\Pi}^{i}(t)$ in terms of primal state variables, we do not even have to use discrete approximations of the adjoint variables.

Although the algorithm has also been applied to higher-dimensional problems, such as the bimolecular catalysis benchmark problem of [11], we focus here on two small-scale academic benchmark problems, that allow us to illustrate many of the interesting features of optimal sampling designs.

5.1. One-dimensional academic example. We are interested in estimating the parameter $p \in \mathbb{R}$ of the initial value problem

$$\dot{x}(t) = p\, x(t),\ t \in [0, t_{\mathrm{f}}], \quad x(0) = x_0.$$

We assume x_0 and t_{f} to be fixed and are only interested in when to measure, with an upper bound M on the measuring time. We can measure the state directly, $h(x(t)) = x(t)$. The experimental design problem (3.5) then simplifies to

$$\begin{aligned}
&\min_{x,G,F,z,w} \ \tfrac{1}{F(t_{\mathrm{f}})} \\
&\text{subject to} \\
&\qquad \dot{x}(t) = p\, x(t), \\
&\qquad \dot{G}(t) = p\, G(t) + x(t), \\
&\qquad \dot{F}(t) = w(t)\, G(t)^2, \\
&\qquad \dot{z}(t) = w(t), \\
&\qquad x(0) = x_0,\ \ G(0) = F(0) = z(0) = 0, \\
&\qquad w(t) \in \mathcal{W}, \\
&\qquad\quad 0 \le M - z(t_{\mathrm{f}})
\end{aligned} \tag{5.1}$$

with $t_{\mathrm{f}} = 1$, $M = 0.2 w^{\max}$.

Although problem (5.1) is as easy as an optimum experimental design problem can be, it already allows us to investigate certain phenomena that may occur. First, assume that $x_0 = 0$. This implies $\dot{x}(t) = \dot{G}(t) = 0$ for all $t \in \mathcal{T}$, and hence the degenerated case in which $G(\cdot) \equiv 0$ and the inverse of the Fisher information matrix does not even exist. If we were to maximize a function of the Fisher information matrix, the sampling design would be a singular decision, as there is no sensitivity with respect to the parameter throughout.

If we choose an initial value of $x_0 \neq 0$, this degenerated case does not occur: obviously a $0 < \tau < t_{\mathrm{f}}$ exists such that $\int_0^\tau x(t)\, \mathrm{d}t \neq 0$ and hence also $G(\tau) \neq 0$ and therefore $F(t_{\mathrm{f}}) > 0$. The global information function for (5.1) is given by

$$\mathbf{\Pi}(t) = \frac{G(t)^2}{F(t_{\mathrm{f}})^2}.$$

As the matrix is one dimensional, all considered criteria carry directly over to this expression. The switching function for (5.1) is given by $\mathcal{H}_w = \frac{G^2(t)}{F^2(t_{\mathrm{f}})} - \mu$. Hence it is clear that a singular arc with $\mathcal{H}_w = 0$ can only occur on an interval $[\tau_{\mathrm{s}}, \tau_{\mathrm{e}}]$ when $\dot{G}(\tau) = 0$ for $\tau \in [\tau_{\mathrm{s}}, \tau_{\mathrm{e}}]$ almost everywhere. With $\dot{G}(\tau) = pG(\tau) + x(\tau)$ this implies that also $x(\cdot)$ is constant on $[\tau_{\mathrm{s}}, \tau_{\mathrm{e}}]$, which is impossible for $x_0 \neq 0$. Therefore problem (5.1) with $x_0 \neq 0$ always has a bang-bang solution with respect to $w(\cdot)$.

We choose $x_0 = 1$ in the following. If $G(\cdot)$ happens to be a positive, monotonically increasing function on $\mathcal{T}$, then we can deduce that the optimal sampling $w(\cdot)$ is given by a $0-1$ arc, where the switching point is determined by the value of M. Such a

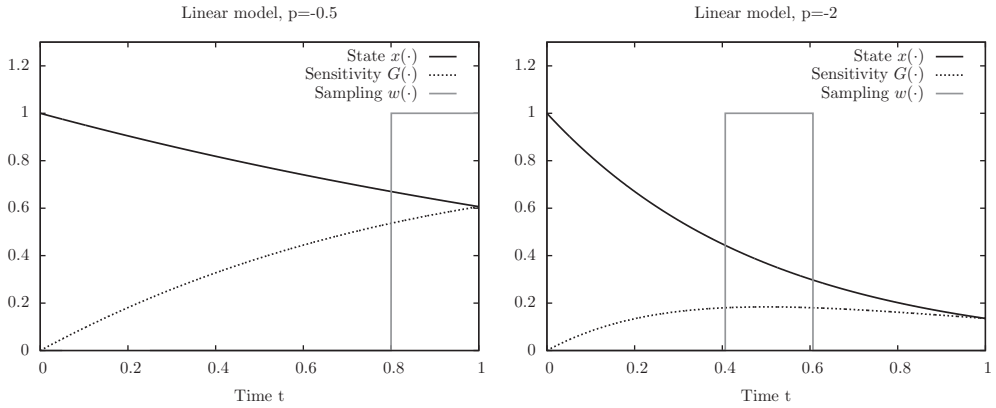

FIG. 5.1. *Linear optimum experimental design problem* (5.1) *with one state and one sampling function for different values of* p. *Left:* $p = -0.5$, *right:* $p = -2$.

scenario is obtained for the expected optimal parameter value of $p = -0.5$; compare Figure 5.1(left).

The switching structure depends not only on functions and initial values, but may also depend on the very value of p itself. An example with an optimal $0-1-0$ solution is depicted in Figure 5.1(right) for the value of $p = -2$. Here the optimal sampling is

$$w(t) = \begin{cases} 0, & t \in [0,\tau] \cup [\tau + 0.2, 1], \\ w^{\max}, & t \in [\tau, \tau + 0.2]. \end{cases} \tag{5.2}$$

Figure 5.1 also illustrates the connection between the discrete-time measurements in section 3.1 and the measurements on intervals as in section 3.2. If the interval width is reduced, the solutions eventually converge to a single point ($\arg\max_{t\in\mathcal{T}} \mathbf{\Pi}(t)$) and coincide with the optimal solution of (3.3).

One interesting feature of one-dimensional problems is that the effect of additional measurements is a pure scaling of $\mathbf{\Pi}(t)$, but not a qualitative change that results in measurements at different times. In other words: it is always optimal to measure as much as possible at the point/interval in time where $\mathbf{\Pi}(t)$ has its maximum value. The measurement reduces the value of $\mathbf{\Pi}(t)$, but its maximum remains in the same time point. This is visualized in Figure 5.6(left), where the optimal sampling (5.2) for different values of $w^{\max}$ results in differently scaled $\mathbf{\Pi}(t)$. We see in the next section that this is not necessarily the case for higher-dimensional OED problems.

5.2. Lotka–Volterra. We are interested in estimating the parameters $p_2, p_4 \in \mathbb{R}$ of the Lotka–Volterra-type predator-prey fish initial value problem

$$\begin{aligned} \dot{x_1}(t) &= p_1\, x_1(t) - p_2 x_1(t) x_2(t) - p_5 u(t) x_1(t),\ t \in [0, t_{\mathrm{f}}], \quad x_1(0) = 0.5, \\ \dot{x_2}(t) &= -p_3\, x_2(t) + p_4 x_1(t) x_2(t) - p_6 u(t) x_2(t),\ t \in [0, t_{\mathrm{f}}], \quad x_2(0) = 0.7, \end{aligned}$$

where $u(\cdot)$ is a fishing control that may or may not be fixed. The other parameters, the initial values and $t_{\mathrm{f}} = 12$, are fixed, consistent with a benchmark problem in mixed-integer optimal control, [21]. We are interested in how to fish and when to measure, again with an upper bound M on the measuring time. We can measure the states directly, $h^1(x(t)) = x_1(t)$ and $h^2(x(t)) = x_2(t)$. We use two different sampling

functions, $w^1(\cdot)$ and $w^2(\cdot)$, in the same experimental setting. This can be seen either as a two-dimensional measurement function $h(x(t))$, or as a special case of a multiple experiment, in which $u(\cdot)$, $x(\cdot)$, and $\boldsymbol{G}(\cdot)$ are identical. The experimental design problem (3.5) then reads

$$
\begin{aligned}
&\min_{x,\boldsymbol{G},\boldsymbol{F},z^1,z^2,u,w^1,w^2} \operatorname{trace}\left(\boldsymbol{F}^{-1}(t_\mathrm{f})\right)\\
&\text{subject to}\\
&\quad \dot{x_1}(t) = p_1\, x_1(t) - p_2 x_1(t)x_2(t) - p_5 u(t)x_1(t),\\
&\quad \dot{x_2}(t) = -p_3\, x_2(t) + p_4 x_1(t)x_2(t) - p_6 u(t)x_2(t),\\
&\quad \dot{G_{11}}(t) = f_{x11}(\cdot)\, G_{11}(t) + f_{x12}(\cdot)\, G_{21}(t) + f_{p12}(\cdot),\\
&\quad \dot{G_{12}}(t) = f_{x11}(\cdot)\, G_{12}(t) + f_{x12}(\cdot)\, G_{22}(t),\\
&\quad \dot{G_{21}}(t) = f_{x21}(\cdot)\, G_{11}(t) + f_{x22}(\cdot)\, G_{21}(t),\\
&\quad \dot{G_{22}}(t) = f_{x21}(\cdot)\, G_{12}(t) + f_{x22}(\cdot)\, G_{22}(t) + f_{p24}(\cdot),\\
&\quad \dot{F_{11}}(t) = w^1(t)G_{11}(t)^2 + w^2(t)G_{21}(t)^2,\\
&\quad \dot{F_{12}}(t) = w^1(t)G_{11}(t)G_{12}(t) + w^2(t)G_{21}(t)G_{22}(t),\\
&\quad \dot{F_{22}}(t) = w^1(t)G_{12}(t)^2 + w^2(t)G_{22}(t)^2,\\
&\quad \dot{z^1}(t) = w^1(t),\\
&\quad \dot{z^2}(t) = w^2(t),\\
&\quad x(0) = (0.5, 0.7),\\
&\quad \boldsymbol{G}(0) = \boldsymbol{F}(0) = \boldsymbol{0},\\
&\quad z^1(0) = z^2(0) = 0,\\
&\quad u(t) \in \mathcal{U},\ w^1(t) \in \mathcal{W},\ w^2(t) \in \mathcal{W},\\
&\quad 0 \le M - z(t_\mathrm{f})
\end{aligned}
\tag{5.3}
$$

with $t_\mathrm{f} = 12$, $p_1 = p_2 = p_3 = p_4 = 1$, and $p_5 = 0.4$, $p_6 = 0.2$, and $f_{x11}(\cdot) = \partial f_1(\cdot)/\partial x_1 = p_1 - p_2 x_2(t) - p_5 u(t)$, $f_{x12}(\cdot) = -p_2 x_1(t)$, $f_{x21}(\cdot) = p_4 x_2(t)$, $f_{x22}(\cdot) = -p_3 + p_4 x_1(t) - p_6 u(t)$, and $f_{p12}(\cdot) = \partial f_1(\cdot)/\partial p_2 = -x_1(t)x_2(t)$, $f_{p24}(\cdot) = \partial f_2(\cdot)/\partial p_4 = x_1(t)x_2(t)$.

Note that the state $F_{21}(\cdot) = F_{12}(\cdot)$ has been left out for reasons of symmetry. We start by looking at the case where the control function $u(\cdot)$ is fixed to zero. In this case the states and the sensitivities are given as the solution of the initial value problem, independent of the sampling functions $w^1(\cdot)$ and $w^2(\cdot)$. Figure 5.2 shows the trajectories of $x(\cdot)$ and $\boldsymbol{G}(\cdot)$.

We set $\mathcal{W} = [0, 1]$ and $M = (4, 4)$. The optimal solution for this control problem is plotted in Figure 5.3. It shows the sampling functions $w^1(\cdot)$ and $w^2(\cdot)$ and the trace of the global information gain matrices

$$
\boldsymbol{\Pi^1}(t) = \boldsymbol{F}^{-1}(t_\mathrm{f}) \begin{pmatrix} G_{11}(t)^2 & G_{11}(t)G_{12}(t) \\ G_{11}(t)G_{12}(t) & G_{12}(t)^2 \end{pmatrix} \boldsymbol{F}^{-1}(t_\mathrm{f}), \tag{5.4a}
$$

$$
\boldsymbol{\Pi^2}(t) = \boldsymbol{F}^{-1}(t_\mathrm{f}) \begin{pmatrix} G_{21}(t)^2 & G_{21}(t)G_{22}(t) \\ G_{21}(t)G_{22}(t) & G_{22}(t)^2 \end{pmatrix} \boldsymbol{F}^{-1}(t_\mathrm{f}) \tag{5.4b}
$$

with $\boldsymbol{F}^{-1}(t_\mathrm{f}) = \begin{pmatrix} F_{11}(t_\mathrm{f}) & F_{12}(t_\mathrm{f}) \\ F_{12}(t_\mathrm{f}) & F_{22}(t_\mathrm{f}) \end{pmatrix}^{-1}$.

Comparing this solution that measures at the time intervals when the integral over the trace of $\boldsymbol{\Pi}(t)$ is maximal to a simulated one with all measurements in the time intervals $[0, M_i]$ with $M_i = 4$, the main effect of the measurements seems to be a homogeneous downscaling over time, comparable to the one-dimensional case in

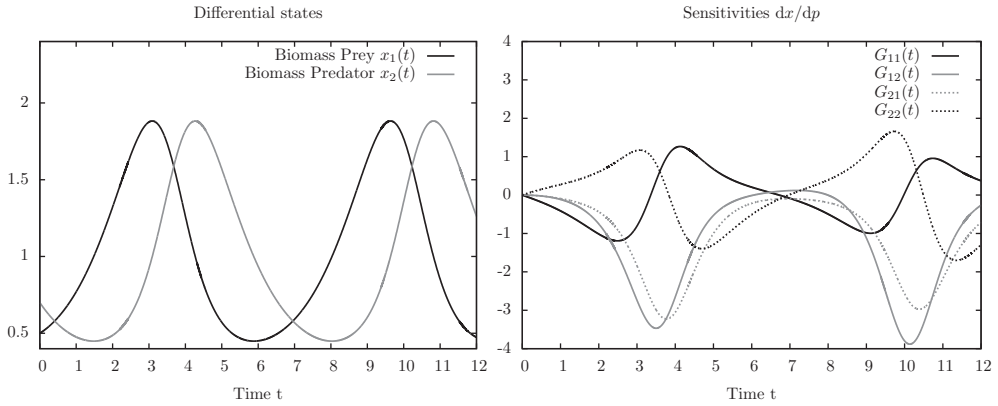

FIG. 5.2. *States and sensitivities of problem* (5.3) *for* $u(\cdot) \equiv 0$ *and* $p_2 = p_4 = 1$.

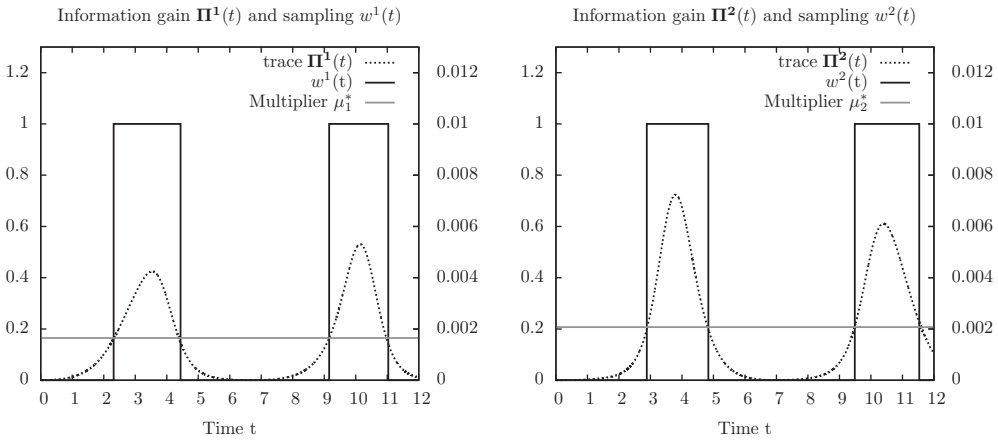

FIG. 5.3. *Optimal solution of problem* (5.3) *for* $u(\cdot) \equiv 0$ *and* $p_2 = p_4 = 1$. *Left: measurement of prey state* $h^1(x(t)) = x_1(t)$. *Right: measurement of predator state* $h^2(x(t)) = x_2(t)$. *The dotted lines show the traces of the functions* (5.4) *over time, their scale is given at the right borders of the plots. One clearly sees the connection between the timing of the optimal sampling, the evolution of the global information gain matrix, and the Lagrange multipliers of the total measurement constraint.*

the last example. The value of what could be gained by additional measurements is reduced by a factor of ≈ 10. These values for both measurement functions are, as we have seen in the last section, identical to the Lagrange multipliers μ_i^*. The numerical results for these Lagrange multipliers are also plotted as horizontal lines in Figure 5.3. As one expects they are identical to the maximal values of the trace of $\mathbf{\Pi}(t)$ outside of the time intervals in which measurements take place.

The same is true for the optimal solution for problem (5.3), again with $u(\cdot) \equiv 0$ and $M = (4, 4)$, but now $p_4 = 4$. The difference in parameters results in stronger oscillations and differences between the two differential states. The optimal sampling hence needs to take the heavy oscillations into account and do measurements on multiple intervals in time; see Figure 5.4. As one can observe, the optimal solution is a sampling design such that the values of the traces of $\mathbf{\Pi}(t)$ at the border points of the $w^i \equiv 1$ arcs are identical to the values of the corresponding Lagrange multipliers. Hence, performing a measurement does have an inhomogeneous (over time) effect on

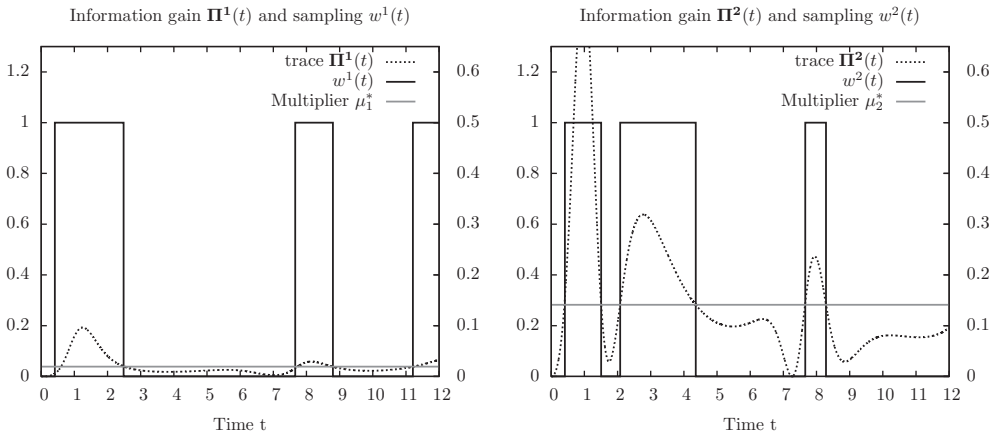

FIG. 5.4. *Optimal solution of problem* (5.3) *for* $u(\cdot) \equiv 0$ *and* $p_2 = 1, p_4 = 4$. *The traces of the information gain functions have more local maxima, hence the sampling is distributed in time. Note that the Lagrange multipliers indicate entry and exit of the functions into the intervals of measurement.*

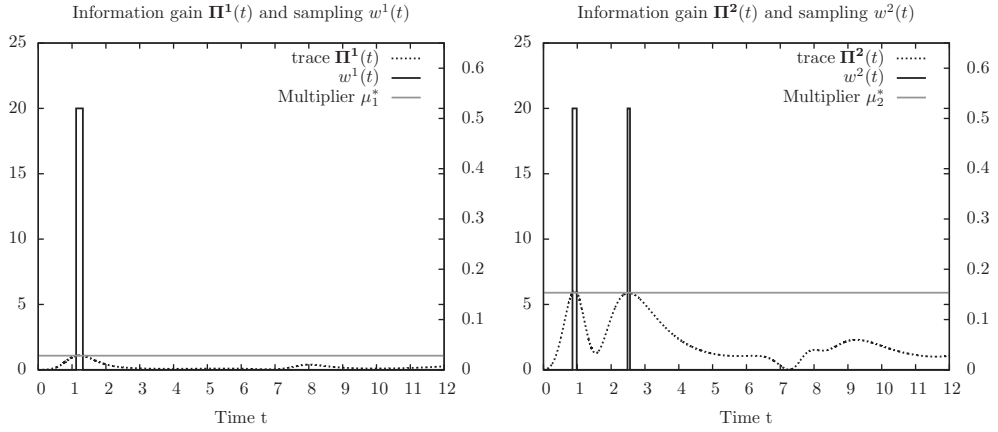

FIG. 5.5. *Optimal solution of problem* (5.3) *as in Figure* 5.4, *but now with* $w^{max} = 20$. *Comparing trace* $\mathbf{\Pi^1}(t)$ *to the one in Figure* 5.4, *one observes a modification and hence a change in the number of arcs with* $w^1(t) \equiv 1$. *The objective function value is reduced, which is reflected in the fact that the values of the optimal Lagrange multipliers* μ_i^* *are smaller than in Figure* 5.4.

the scaling of $\mathbf{\Pi}(t)$. The coupling between measurements at different points in time, and also between different experiments, takes place via the transversality conditions of the adjoint variables.

The inhomogeneous scaling can also be observed in Figure 5.5, where a sampling design for $w^{\max} = 20$ is plotted. One sees that fewer measurement intervals are chosen and that the shape of the local information gain function $\mathbf{\Pi^1}(t)$ is different from the one in Figure 5.4.

The same effect—an inhomogeneous scaling of the information gain function—is the reason why fractional values $w(\cdot) \notin \{0, 1\}$ may be obtained as optimal values when fixed time grids are used with piecewise constant controls. We use the same scenario as above, hence $u(\cdot) \equiv 0$, $M = (4, 4)$, and $p_4 = 4$. Additionally we fix $w^2(\cdot) \equiv 0$ and consider a piecewise constant control discretization on the grid $t_i = i$ with $i = 0 \ldots 12$.

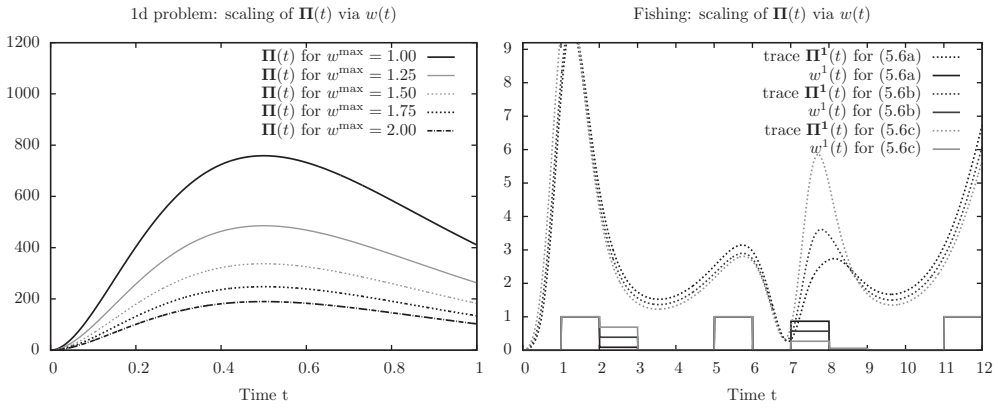

FIG. 5.6. *Left: Global information gain function for one-dimensional OED problem* (5.1) *and controls* $w(\cdot)$ *obtained from* (5.2) *for different values of* w^{max}*. Note that the information gain matrix is scaled uniformly over the whole time horizon. Right: Global information gain functions for OED problem* (5.3) *and controls* $w(\cdot)$ *obtained from* (5.5) *and either one from* (5.6a)–(5.6c)*. One sees that the information gain matrix* $\mathbf{\Pi^1}(t)$ *is scaled differently, depending on the values of* w_2 *and* w_7*. The optimal solution* (5.6b) *on this coarse grid is the solution which scales the information gain function in a way such that the integrated values on* $[2,3]$ *and* $[7,8]$ *are identical.*

We consider the trajectories for $w^1(t) = w_i$ when $t \in [t_i, t_{i+1}]$, $i = 0 \dots 11$ with

$$w_1 = w_5 = w_{11} = 1,\ w_0 = w_3 = w_4 = w_6 = w_9 = w_{10} = 0,\ w_8 = 0.0444, \tag{5.5}$$

and the three cases

$$w_2 = 0.0885, w_7 = 0.8671, \tag{5.6a}$$
$$w_2 = 0.3885, w_7 = 0.5671, \tag{5.6b}$$
$$w_2 = 0.6885, w_7 = 0.2671, \tag{5.6c}$$

where the trajectory corresponding to (5.6b) is the optimal one, and the two others have been slightly modified to visualize the effect of scaling the information gain matrix by modifying the sampling design. See Figure 5.6(right) for the corresponding information gain functions. One sees clearly the inhomogeneous scaling. The optimal solution (5.6b) on this coarse grid is the solution which scales the information gain function in a way such that the integrated values on $[2,3]$ and $[7,8]$ are identical. To get an integer feasible solution with $w(\cdot) \in \{0,1\}$ we therefore recommend refining the measurement grid rather than rounding.

Next, we shed some light on the case where we have additional degrees of freedom. We choose $\mathcal{U} = [0,1]$ and allow for additional fishing, again for the case $p_2 = p_4 = 1$. In Figure 5.7(left) one sees the optimal control $u^*(\cdot)$, which is also of bang-bang type. The effect of this control is an increase in amplitude of the states' oscillations, which leads to an increase in sensitivity information; see Figure 5.7(right). The corresponding optimal sampling design is plotted in Figure 5.8. The timing is comparable to the one in Figure 5.3. However, the combination of control function $u^*(\cdot)$ and the sampling design leads to a concentration of information in the time intervals in which measurements are being done. This is best seen by comparing the values of the Lagrange multipliers in Figure 5.3 of $\mu^* \approx (1.8, 2.6)10^{-3}$ versus the ones of Figure 5.8 with $\mu^* \approx (3, 3.6)10^{-4}$ which are one order of magnitude smaller.

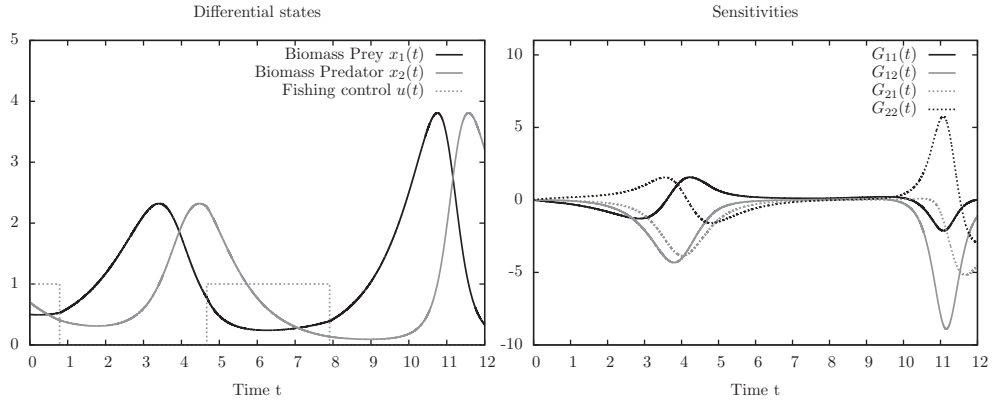

FIG. 5.7. *States and sensitivities of problem* (5.3) *for* $u(\cdot) \in \mathcal{U} = [0,1]$ *and* $p_2 = p_4 = 1$. *See the increased variation in amplitude compared to Figure* 5.2.

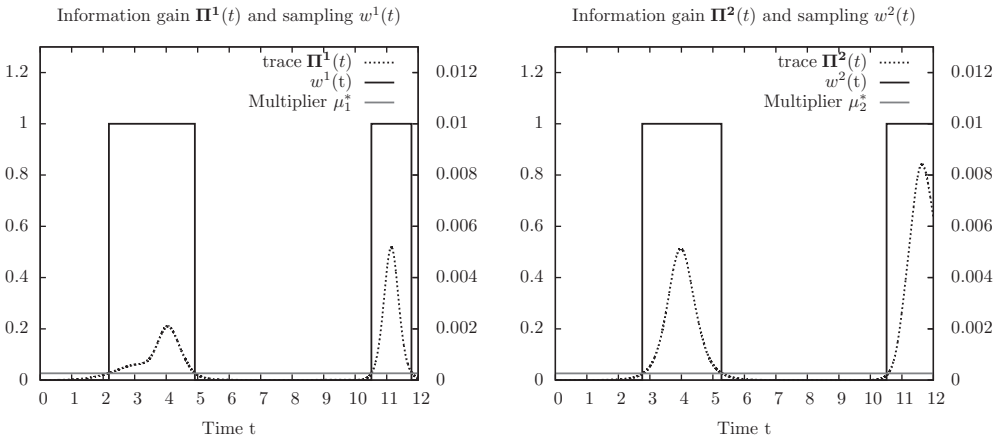

FIG. 5.8. *Optimal sampling corresponding to Figure* 5.7. *Note the reduction of the Lagrange multiplier by one order of magnitude compared to Figure* 5.3 *due to the amplification of states and sensitivities.*

As a last illustrating case study we consider an additional L^1 penalty of the sampling design in the objective function as discussed in section 4.3. We consider problem (5.3) for $u(\cdot) \equiv 0$ and $p_2 = p_4 = 1$ and $M = \infty$. The objective function now reads

$$\min_{x,\boldsymbol{G},\boldsymbol{F},z^1,z^2,u,w^1,w^2} \operatorname{trace}\left(\boldsymbol{F}^{-1}(t_\mathrm{f})\right) + \int_{\mathcal{T}} \epsilon(w^1(\tau) + w^2(\tau))\,\mathrm{d}\tau \tag{5.7}$$

with $\epsilon = 1$.

As can be seen in Figure 5.9, the L^1 penalization has the effect that the optimal sampling functions are given by

$$w^i(t) = \begin{cases} w^{\max}, & \operatorname{trace} \mathbf{\Pi^i}(t) \geq \epsilon, \\ 0, & \text{else.} \end{cases} \tag{5.8}$$

This implies that the value of ϵ in the problem formulation can be used to directly influence the optimal sampling design. Especially for ill-posed problems with small

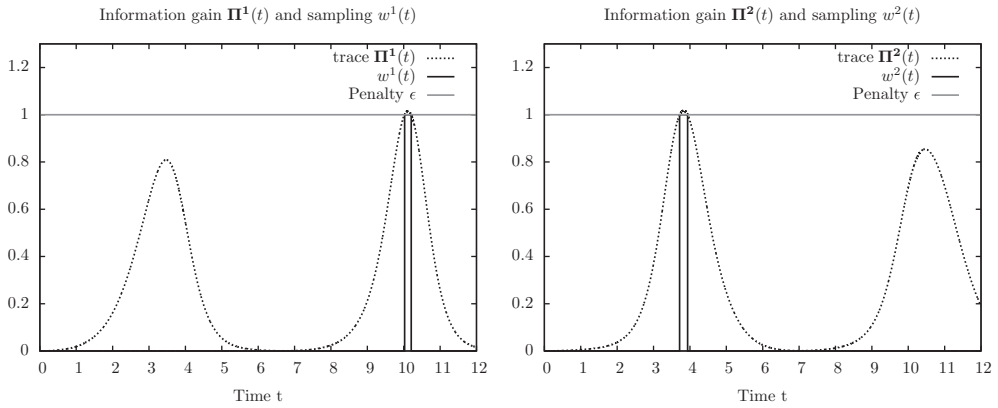

FIG. 5.9. *Optimal sampling for problem* (5.3) *with objective function augmented by linear penalty term* $\int_{\mathcal{T}} \epsilon(w^1(\tau) + w^2(\tau))\, \mathrm{d}\tau$. *The sampling functions* $w^i(t)$ *are at their upper bounds of* 1 *if and only if* trace $\mathbf{\Pi^i}(t) \geq \epsilon = 1$.

values in the information gain matrix $\mathbf{\Pi}(t)$ this penalization is beneficial from a numerical point of view, as it avoids flat regions in the objective function landscape that might lead to an increased number of iterations. Also it allows a direct economic interpretation by coupling the costs of a single measurement to the information gain. To give an idea of the impact on the number of iterations until convergence we consider an instance with both measurement functions, $u(\cdot) \in [0, 1]$ and $M = (4, 4)$. Dependent on the penalization value ϵ in (5.7) we get the following number of iterations (with default settings) with different NLP codes. Problem (5.3) has been discretized by means of a direct collocation (implicit Euler and piecewise constant controls) method. The resulting `AMPL` model can be found on `http://mintoc.de` in the benchmark library described in [19].

The following table shows iteration numbers which are *not* to be compared among one another as they indicate completely different things. Rather the impact of the penalization parameter ϵ within each row is of interest.

ϵ	0	10^{-3}	10^{-2}	10^{-1}	1	10
conopt 3.15C	1685	1707	1651	1666	1681	1677
ipopt 3.10.0	207	152	155	152	143	130
KNITRO 8.0.0	213	144	76	38	50	64
snopt 7.2-8	1670	2040	1473	2499	1530	1127

The significant reduction in iterations for higher values of ϵ for most solvers can also be observed with the optimal control package `MUSCOD-II`[2] that is based on Bock's direct multiple shooting method, hence an SQP-type method. However, the reduced gradient solver `conopt` seem to be less sensitive to the regularization. The optimal solutions corresponding to different ϵ are of course different, hence a comparison is somewhat arbitrary as the iteration number depends heavily on the starting point. Yet, it at least gives an indication of the potential of this regularization for certain numerical methods.

We discourage using an L^2 penalization as discussed in section 4.4. It often results in sensitivity-seeking arcs with values in the interior of $\mathcal{W}$, and there is no useful economic interpretation.

[2]Trunk version of September 1, 2012

6. Conclusions. We have applied the integer gap theorem and the maximum principle to an optimal control formulation of a generic optimum experimental design problem. Thus we were able to analyze the role of sampling functions that determine when measurements should be performed to maximize the information gain with respect to unknown model parameters. We showed the similarity between a continuous time formulation with measurements on intervals of time, and a formulation with measurements at single points in time. We defined the *information gain* functions that apply to both formulations as the result of a theoretical analysis of the necessary conditions of optimality. Based on information gain functions we were able to shed light on several aspects, both theoretical as by means of two numerical examples.

Differences between Fisher and covariance objective function. We showed that the information gain matrix for a Fisher objective function has a local character, whereas the one for a covariance objective function includes terms that depend on differential states at the end of the time horizon. This implies that measurements affect the information gain function in the covariance objective case, but not in the Fisher objective case. This noncorrelation for a maximization of a function of the Fisher information matrix has direct consequences: integral-neutral rounding of fractional solutions does not have any influence on the objective function. It also means that other experiments do not influence the choice of the measurements. Third, providing a feedback law in the context of first optimize then discretize methods is possible. All this is usually not true for covariance objective functions.

Scaling of global information gain function by measuring. Taking measurements changes the global information matrix $\mathbf{\Pi}(t)$. The impact may be in form of a uniform downscaling, but also as a nonhomogeneous over time modification. In the latter case it is not optimal to take as many measurements as possible in one single point of time, as is the case for a Fisher objective function or one-dimensional problems, if one allows more than one measurement per time point/interval. The coupling between the information function and the measurement functions takes place via the transversality conditions, thus the impact also carries over to other experiments and measurement functions.

Role of Lagrange multipliers. We showed that the Lagrange multipliers of constraints that limit the total number of measurements on the time horizon give a threshold for the information gain function. Whenever the function value is higher, measurements are performed, otherwise the value of w is 0.

Role of additional control functions. We used a numerical example to exemplarily demonstrate the effect of additional control functions on the shape of the information gain function.

Role of fixed grids and piecewise constant approximations. For the practically interesting case that optimizations are performed on a given measurement grid we showed that fractional solutions may be optimal. We recommend further refining the measurement grid instead of rounding.

Penalizations and ill-posed problems. By its very nature, optimal solutions result in small values of the global information gain function. This explains why OED problems are often ill-posed if the upper bounds on the total amount of measurements are chosen too high: additional measurements only yield small contributions to the objective function once the other measurements have been placed in an optimal way. As a remedy to overcome this intrinsic problem of OED we propose using L^1 penalizations of the measurement functions. We showed that the penalization parameter can be directly interpreted in terms of the information gain functions. Therefore such a formulation couples the costs of a measurement to a minimum amount of information

it has to yield, which makes sense from a practical point of view. Of course, the value of ϵ can also be decreased in a homotopy.

Appendix A. Useful lemmas. In this appendix we list several useful lemmas we use throughout the paper.

LEMMA A.1 (positive trace). *If $A \in \mathbb{R}^{n\times n}$ is positive definite, then* $\mathrm{trace}(A) > 0$.

Proof. As A is positive definite, it holds $x^T Ax > 0$ for all $x \in \mathbb{R}^n$, in particular for all unit vectors. Hence it follows $a_{ii} > 0$ for all $i = 1 \dots n$ and thus trivially $\mathrm{trace}(A) = \sum_{i=1}^n a_{ii} > 0$. ∎

LEMMA A.2 (derivative of trace function). *Let A be a quadratic $n \times n$ matrix. Then*

$$\left\langle \frac{\partial \mathrm{trace}(A)}{\partial A}, \Delta A \right\rangle = \mathrm{trace}(\Delta A). \tag{A.1}$$

Proof.

$$\begin{aligned}\left\langle \frac{\partial \mathrm{trace}(A)}{\partial A}, \Delta A \right\rangle &= \lim_{h\to 0} \frac{\mathrm{trace}(A + h\Delta A) - \mathrm{trace}(A)}{h} \\ &= \lim_{h\to 0} \frac{h\,\mathrm{trace}(\Delta A)}{h} = \mathrm{trace}(\Delta A). \qquad \blacksquare\end{aligned}$$

LEMMA A.3 (derivative of inverse operation). *Let $A \in \mathrm{GL}_n(\mathbb{R})$ be an invertible $n \times n$ matrix. Then*

$$\frac{\partial A^{-1}}{\partial A} \cdot \Delta A = -A^{-1}\Delta A A^{-1}. \tag{A.2}$$

LEMMA A.4 (derivative of eigenvalue operation). *Let $\lambda(A)$ be a single eigenvalue of the symmetric matrix $A \in \mathbb{R}^{n\times n}$. Let $z \in \mathbb{R}^n$ be an eigenvector of A to $\lambda(A)$ with norm* 1. *Then it holds*

$$\left\langle \frac{\partial \lambda(A)}{\partial A}, \Delta A \right\rangle = z^T \Delta A z. \tag{A.3}$$

LEMMA A.5 (derivative of determinant operation). *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric, positive definite matrix. Then it holds*

$$\left\langle \frac{\partial \det(A)}{\partial A}, \Delta A \right\rangle = \det(A) \sum_{i,j=1}^{n} A^{-1}_{i,j}\, \Delta A_{i,j}. \tag{A.4}$$

Proofs for the Lemmas A.3, A.4, and A.5 can be found in [11].

Acknowledgments. The author thanks the work groups of Georg Bock, Johannes Schlöder, and Stefan Körkel for helpful and stimulating discussions.

REFERENCES

[1] A.C. Atkinson, A.N. Donev, and R.D. Tobias, *Optimum Experimental Designs, with SAS*, Oxford University Press, Oxford, 2007.

[2] A. Bardow, W. Marquardt, V. Göke, H.J. Koss, and K. Lucas, *Model-based measurement of diffusion using raman spectroscopy*, AIChE J., 49 (2003), pp. 323–334.

[3] I. Bauer, H.G. Bock, S. Körkel, and J.P. Schlöder, *Numerical methods for optimum experimental design in DAE systems*, J. Comput. Appl. Math., 120 (2000), pp. 1–15.

[4] H.G. Bock and R.W. Longman, *Computation of optimal controls on disjoint control sets for minimum energy subway operation*, in Proceedings of the American Astronomical Society. Symposium on Engineering Science and Mechanics, Taiwan, 1982.
[5] A.E. Bryson and Y.-C. Ho, *Applied Optimal Control*, Wiley, New York, 1975.
[6] G. Franceschini and S. Macchietto, *Model-based design of experiments for parameter precision: State of the art*, Chem. Engrg. Sci., 63 (2008), pp. 4846–4872.
[7] R. Gamkrelidze, *Principles of Optimal Control Theory*, Plenum Press, New York, 1978.
[8] A.D. Ioffe and V.M. Tihomirov, *Theory of Extremal Problems*, Stud. Math. Appl. 6, North-Holland, Amsterdam, 1979.
[9] D. Janka, *Optimum Experimental Design and Multiple Shooting*, Master's thesis, Universität Heidelberg, Heidelberg, 2010.
[10] H.J. Kelley, R.E. Kopp, and H.G. Moyer, *Singular extremals*, in Topics in Optimization, G. Leitmann, ed., Academic Press, New York, 1967, pp. 63–101.
[11] S. Körkel, *Numerische Methoden für Optimale Versuchsplanungsprobleme bei Nichtlinearen DAE-Modellen*, Ph.D. thesis, Universität Heidelberg, Heidelberg, 2002.
[12] S. Körkel, E. Kostina, H.G. Bock, and J.P. Schlöder, *Numerical methods for optimal control problems in design of robust optimal experiments for nonlinear dynamic processes*, Optim. Methods Softw., 19 (2004), pp. 327–338.
[13] S. Körkel, A. Potschka, H.G. Bock, and S. Sager, *A multiple shooting formulation for optimum experimental design*, Math. Program., submitted.
[14] S. Körkel, H. Qu, G. Rücker, and S. Sager, *Derivative based vs. derivative free optimization methods for nonlinear optimum experimental design*, in Proceedings of HPCA2004 Conference, Shanghai, 2004, Springer, Berlin, 2005, pp. 339–345.
[15] E.J. McShane, *The calculus of variations from the beginning to optimal control theory*, SIAM J. Control Optim., 27 (1989), pp. 916–939.
[16] H.J. Pesch and R. Bulirsch, *The maximum principle, Bellman's equation and Caratheodory's work*, J. Optim. Theory Appl., 80 (1994), pp. 203–229.
[17] L.S. Pontryagin, V.G. Boltyanski, R.V. Gamkrelidze, and E.F. Miscenko, *The Mathematical Theory of Optimal Processes*, Wiley, Chichester, 1962.
[18] F. Pukelsheim, *Optimal Design of Experiments*, Classics Appl. Math. 50, SIAM, Philadelphia, 2006.
[19] S. Sager, *A benchmark library of mixed-integer optimal control problems*, in Mixed Integer Nonlinear Programming, J. Lee and S. Leyffer, eds., Springer, New York, 2012, pp. 631–670.
[20] S. Sager, H.G. Bock, and M. Diehl, *The integer approximation error in mixed-integer optimal control*, Math. Program. A, 133 (2012), pp. 1–23.
[21] S. Sager, H.G. Bock, M. Diehl, G. Reinelt, and J.P. Schlöder, *Numerical methods for optimal control with binary control functions applied to a Lotka-Volterra type fishing problem*, in Recent Advances in Optimization, A. Seeger, ed., Lectures Notes in Econom. and Math. Systems 563, Springer, Heidelberg, 2009, pp. 269–289.
[22] S. Sager, G. Reinelt, and H.G. Bock, *Direct methods with maximal lower bound for mixed-integer optimal control problems*, Math. Program., 118 (2009), pp. 109–149.
[23] K. Schittkowski, *Experimental design tools for ordinary and algebraic differential equations*, Math. Comput. Simulation, 79 (2007), pp. 521–538.
[24] J. Schöneberger, H. Arellano-Garcia, H. Thielert, S. Körkel, and G. Wozny, *Optimal experimental design of a catalytic fixed bed reactor*, in Proceedings of 18th European Symposium on Computer Aided Process Engineering, B. Braunschweig and X. Joulia, eds., Elsevier, Oxford, 2008.
[25] S.P. Sethi and G.L. Thompson, *Optimal Control Theory: Applications to Management Science and Economics*, 2nd ed., Springer, New York, 2005.
[26] B. Srinivasan, S. Palanki, and D. Bonvin, *Dynamic Optimization of Batch Processes:* I. *Characterization of the nominal solution*, Comput. Chem. Engrg., 27 (2003), pp. 1–26.