Using data to inform policy [*]

Maximilian Kasy [†]

December 19, 2013

Abstract

In this paper, a general framework is proposed for the use of (quasi-) experimental data when choosing policies such as tax rates or the level of inputs in a production process. The data are used to update expectations about social welfare as a function of policy, and the policy is chosen to maximize expected social welfare. We characterize the properties of the implied decision function. For settings where experimentation is feasible, we characterize experimental designs maximizing expected social welfare, assuming policies are chosen based on the data. We discuss several economic settings which are covered by our framework, and apply our methods to data from the RAND health insurance experiment. In this application, we obtain much smaller estimates of the optimal copay (18% vs. 50%) than those obtained using a conventional sufficient-statistic approach.

Our approach combines optimal policy theory, statistical decision theory, and nonparametric Bayesian methods. It explicitly takes into account economic theory, does not rely on restrictions of functional form or heterogeneity, and provides a transparent mapping from observations to policy choices. This mapping is optimal in finite samples.

1 Introduction

One of the main objectives of empirical research in economics is to inform policy choices. Examples include the choice of inputs into some production process (broadly defined), such as education or health, as well as the choice of policy parameters such as tax rates, co-payments in health insurance, replacement rates in social insurance, etc.

In this paper, a general framework is proposed for the use of (quasi-) experimental data when choosing such policies. Our framework assumes that the policy-maker wishes to maximize expected social welfare. The data are used to update expectations about social welfare under alternative choices of policy parameters. Our framework leads to a fairly simple and tractable mapping from observed data to optimal policy choices. In settings where experimentation is an option, it allows to calculate experimental designs which maximize ex-ante expected welfare, assuming the policy is chosen based on the

[*] I thank Gary Chamberlain, Raj Chetty, Nathan Hendren, and Michael Kremer as well as seminar participants at Stanford and UC Berkeley for helpful discussions and comments.

[†] Assistant professor, Department of Economics, Harvard University, and junior associate faculty, IHS Vienna. Address: Littauer Center 200, 1805 Cambridge Street, Cambridge, MA 02138. E-Mail: maximiliankasy@fas.harvard.edu.

experimental data. Our framework allows, finally, to assess the value of additional experimentation for increasing social welfare.

Our approach has several features often considered desirable. It explicitly incorporates the insights of economic theory, in particular optimal policy theory. It is nonparametric, and does not rely on any restrictions of functional form or the dimensionality of unobserved heterogeneity. It provides a transparent and easily visualized mapping from observed data to policy choices. It explicitly takes into account the uncertainty stemming from extrapolation outside the support of available data. The proposed procedures are, finally, optimal from the perspective of finite-sample statistical decision theory.

Our framework makes the following assumptions. (i) There is random variation of the policy choice variables in the available data.[1] (ii) The policy objective function can be written as a transformation of the relevant technological or behavioral relationships by some affine operator.[2] (iii) The prior distribution of these relationships is given by a Gaussian process. Under these assumptions, we provide explicit expressions, as well as characterizations, of optimal policy choices and of optimal experimental designs. We show how to calculate posterior expected social welfare, as a function of policy choices, and how to calculate the maximizer of posterior expected social welfare. We discuss the statistical properties of this maximizer, in particular its asymptotic normality, and the determinants of its asymptotic variance. We show, further, how ex-ante expected social welfare depends on experimental design, and how to choose an experimental design which maximizes ex-ante expected welfare, assuming the policy is chosen optimally given the experimental data.

Our framework aims to incorporate the advantages of various approaches to empirical work in economics. First, like "structural" approaches, it explicitly incorporates economic theory in order to evaluate policy counterfactuals. Second, like "causal" (or "reduced form") approaches, it does not rely on functional form restrictions or restrictions of heterogeneity for identification. Third, like "sufficient statistic" approaches, it is guided in its choice of objects of interest and of statistical loss by explicit models of optimal policy choice. Our framework has some limitations, as well. First, we essentially assume away the problem of external validity. Second, our experimental design is optimal for a particular, well defined policy problem – some experiments might end up being analyzed in a different way then originally intended. Third, policy choice is not the only purpose of economic research. The testing of theories, in particular, is beyond the scope of the framework discussed here. Finally, policy choice involves a number of normative decisions such as the welfare weights assigned to various groups or the willingness to pay for some outcome. These decisions should be a matter of public debate, and are beyond the scope of "expert knowledge."

This paper draws on a number of independent literatures including (i) optimal policy theory, (ii) statistical decision theory, (iii) machine learning and nonparametric Bayesian estimation, (iv) the methodological debates about the advantages of alternative approaches to empirical research in microeconomics, and (v) the empirical literature in applied microeconomics itself. We conclude this introduction by listing some references

[1] We later generalize to observational data where either conditional exogeneity given covariates holds, or there is a valid instrumental variable.

[2] This is not very restrictive and covers all standard models of optimal policy choice. The affine dependence allows for operations such as integration and differentiation, in particular.

for each of these literatures. Models of optimal policy in public finance have a long tradition going back at least to the discussion in Samuelson (1947) of social welfare functions, with classic contributions including Mirrlees (1971) and Baily (1978). The empirical implementation of such models using "sufficient statistics" is discussed in Chetty (2009) and Saez (2001). Textbook introductions to statistical decision theory are available, for instance, in Casella and Berger (2001) and Robert (2007). Gaussian process priors and nonparametric Bayesian function estimation are discussed extensively in Williams and Rasmussen (2006). A general and thorough treatment of nonparametric Bayesian statistics can be found in (Ghoshal and van der Vaart, 2013, forthcoming). The controversies surrounding "reduced form" and "structural" methods are well summarized in Deaton (2009), Imbens (2010), Angrist and Pischke (2010), and Nevo and Whinston (2010). The literature discussing the estimation of (educational) production functions using (quasi-) experimental data includes contributions such as Fryer (2011), Angrist and Lavy (1999), Krueger (1999), and Rivkin et al. (2005). A decision problem related to the problem discussed here is the problem of optimal assignment of a discrete treatment based on covariates and given experimental data. This problem has been studied by Manski (2004), Dehejia (2005), Bhattacharya and Dupas (2008), Hirano and Porter (2009), Chamberlain (2011) and Stoye (2011) among others.

The rest of this paper is structured as follows. Section 2 introduces the general setup studied in this paper. Section 3 discusses several optimal policy problems that fit into our framework, including optimal taxation and optimal insurance, the choice of inputs to a production process, and the allocation of treatments in the presence of peer effects. Section 4 discusses the optimal policy choice given experimental data, and the optimal experimental design. Our main theoretical results are discussed in sections 5 and 6. Section 5 develops the frequentist asymptotic theory of our estimators. Section 6 characterizes the optimal experimental design. Section 7 discusses extensions of our framework to settings with observational data where either (i) conditional exogeneity of treatment given available covariates is satisfied, or (ii) the data contain an exogenous instrumental variable. Section 8 applies our methods to data from the RAND health insurance experiment, using these data to estimate an optimal coinsurance rate, and compares the results to those obtained using the "sufficient statistic" approach. Section 9 concludes.

2 Setup

The setup we consider can be summarized as follows: We assume that the econometrician's role is to advise a policy maker, who wishes to maximize social welfare $u(t)$ through choice of t. The objective function u is unknown, but given by a known mapping from an average structural function m, where m corresponds to a structural function g. We can learn about m using experimental data (X_i, Y_i), where $E[Y_i|X_i = x] = m(x)$. Expected utility maximization by the policy maker requires a prior distribution for m; we assume that this prior distribution lies in the general (nonparametric) class of Gaussian process priors.

We next state our assumptions more formally, before discussing each of them in turn. Assumption 2 will be clarified by a series of examples in section 3.

Assumption 1 (Objective)

1. *Social welfare for policy choice* $t \in \mathbb{R}^{d_t}$ *is given by* $u(t)$*.*
2. u *is an unknown function.*
3. *The policy-maker aims to maximize expected social welfare.*

Assumption 2 (Average structural function and objective function)

1. *Social welfare* u *is equal to* $L \cdot m + u_0$*, where*
   - (a) m *is an element of*[3] $\mathscr{C}^1(\mathscr{X})$*,*
   - (b) L *is a linear operator from* $\mathscr{C}^1(\mathscr{X})$ *to* $\mathscr{C}^1(\mathscr{T})$*,*
   - (c) $\mathscr{X} \subset \mathbb{R}^{d_x}$*,* $\mathscr{T} \subset \mathbb{R}^{d_t}$*,*
   - (d) *and* L *and* u_0 *are known.*
2. $m(x) = E[g(x, \epsilon)]$*, where*
   - (a) $g(x, \epsilon)$ *is a structural function,*
   - (b) *and* $\epsilon \sim P_\epsilon$ *in the target population.*

Assumption 3 (Experimental data)

1. *The data* $(X_i, Y_i)_{i=1}^n$ *are i.i.d.*
2. $Y_i = g(X_i, \epsilon_i)$*.*
3. $X_i \perp \epsilon_i$*.*
4. $\epsilon_i \sim P_\epsilon$*, the distribution of* ϵ *in the target population.*

Assumption 4 (Prior)
The policy-maker has a prior which satisfies the following conditions.

1. $m \sim GP(\mu(.), C(., .))$*, where*
   - (a) $GP(\mu(.), C(., .))$ *is the law of a Gaussian process*
   - (b) *such that* $E[m(x)] = \mu(x)$*, and*
   - (c) $\mathrm{Cov}(m(x), m(x')) = C(x, x')$*.*
2. *The function* m *is independent of the probability distribution* P_X*.*

[3] $\mathscr{C}^k(\mathscr{X})$ denotes the space of k times continuously differentiable functions on $\mathscr{X}$, equipped with the norm $\|m\| = \sum_{j=1}^k \sup_{x \in \mathscr{X}} |m^{(j)}(x)|$.

Remarks

- **Assumption 1**

  We have assumed that u is an unknown function. An alternative notation for the same assumption would take u as a known function of an unknown set of parameters θ and the policy choice t.

  Assumption 1 states that the policy maker wishes to maximize expected social welfare, which puts us in the Bayesian paradigm.[4] Expectations are posterior expectations given the observed data. We equate the objective of statistical analysis and the policy-maker's objective.

- **Assumption 2**

  The definition of m as $m(x) = E[g(x, \epsilon)]$ equates it to the average structural function (cf. Blundell and Powell, 2003) corresponding to a structural function g and a distribution of unobserved heterogeneity ϵ. Note that we do *not* restrict the support or dimensionality of ϵ in any way, nor do we restrict the functional form of g.

  The objective function u is given by an affine transformation of m, that is $u - u_0$ is given by the image of the average structural function m under some general linear operator L. This assumption covers many cases of interest; the next section will discuss several examples. In these examples m corresponds for instance to demand functions, to the relationship between the tax base and tax parameters, or to production functions describing average output as a function of inputs.

  The functions u considered involve a population average. The function u might for instance be equal to a weighted average of private utility, or to an average potential outcome. Expected social welfare thus averages both over the distribution of heterogeneity in the population of interest, and over posterior uncertainty.

- **Assumption 3**

  This assumption immediately implies, that the average structural function m is identified from the observed data via $m(x) = E[Y_i | X_i = x]$.

  This assumption imposes several important restrictions. First, outcomes Y_i are given by the structural function g evaluated at X_i. This is known as the "stable unit treatment values" assumption (SUTVA, cf. Angrist et al. 1996). Note, however, that this assumption does not exclude consideration of equilibrium effects if we define the unit of observation appropriately.[5] Second, X_i is assumed to be (as if) randomly assigned. This guarantees the internal validity of a regression of Y on X. Third, we assume that the distribution of unobserved heterogeneity affecting outcomes is the same in the experimental population as in the target population. This guarantees the external validity of estimates based on the available data.

  When we consider experimental design, we will impose a minor variation of this assumption. We will consider the empirical distribution of the X_i to be chosen by

[4] In section 5 we will, however, analyze from a frequentist perspective the decision functions suggested by the Bayesian paradigm.

[5] Section 3.3 discusses a setting where we are explicitly interested in (equilibrium) peer effects, taking as our unit of observation the classroom.

the experimenter, while still maintaining that the distribution of ϵ_i is the same for all i and equal to the population distribution.

In section 7 we discuss how to relax assumption 3 to a quasi-experimental setting with instruments Z and controls W. We will assume that $P(\epsilon_i|Z_i, W_i) = P(\epsilon|W)$ where $P(\epsilon|W)$ is the conditional distribution of heterogeneity in the target population of the policy. This assumption relaxes the exogeneity of treatment X_i to exogeneity and exclusion of an instrument Z_i. It relaxes the external validity assumption to an assumption of external validity given controls.

- **Assumption 4**

  This assumption restricts the form of prior beliefs admitted for the policy-maker. It describes a general class of non-parametric priors for the function m. Gaussian process priors have become increasingly popular in the machine learning literature, as they lead to tractable Bayesian estimators without imposing parametric restrictions.[6] Note that this assumption does not restrict the first two moments of the prior, and therefore does not restrict the class of estimators based on posterior best linear predictors. In section 4 we discuss the implications of Gaussian process priors for posterior expectations. Appendix B provides some review of commonly used covariance kernels $C(.,.)$, and of the properties of m implied by these kernels.

2.1 Prior moments

In section 4 below, we derive posterior expectations for m and u given the data and given the Gaussian process prior of assumption 4. We then use the posterior expectation of u to characterize the optimal choice of t, which maximizes the posterior expectation of u.

In order to derive these posterior expectations and the optimal t, we need the prior covariance between observed outcomes and the value of the objective function at a given point, as well as the prior covariance between observed outcomes and the objective function's derivative. These covariances are given by the following definition. These covariances depend on the prior for m, as well as on the operator L mapping m into u. Lemma 1 characterizes these covariances. In section 3 we discuss a number of specific optimal policy problems, the corresponding operators L, and the corresponding covariances.

Definition 1 (Prior moments)
Define the function $\nu : \mathscr{T} \to \mathbb{R}$ as the prior expectation

$$\nu(t) := E[u(t)]. \tag{1}$$

Define the function $D : \mathscr{T} \times \mathscr{X} \to \mathbb{R}$ as the prior covariance

$$D(t, x) := \mathrm{Cov}(u(t), m(x)). \tag{2}$$

Define the function $K : \mathscr{T} \times \mathscr{T} \to \mathbb{R}$ as the prior covariance

$$K(t, t') := \mathrm{Cov}(u(t), u(t')). \tag{3}$$

[6] For an introduction to Gaussian process priors, see Williams and Rasmussen (2006); a thorough general treatment of nonparametric Bayesian statistics can be found in the forthcoming Ghoshal and van der Vaart (2013).

Define the function $B : \mathscr{T} \times \mathscr{X} \to \mathbb{R}$ as the prior covariance[7]

$$B(t, x) := \operatorname{Cov}\left(\frac{\partial}{\partial t} u(t), m(x)\right). \tag{4}$$

The following lemma characterizes these covariances under some high-level regularity conditions on the operator L and the covariance kernel C. In the applications considered in section 3 we directly check the implications of this lemma rather than the high level regularity conditions. In the statement of this lemma we write $L_{x'}C(x', x)$ to emphasize that this expression applies the linear operator L to $C(x', x)$ as a function of x' for fixed x.

Lemma 1 (Characterizing prior moments)
Under the regularity conditions stated below, the moments given by definition 1 are equal to

$$\nu(t) = (L \cdot \mu)(t) + u_0(t), \tag{5}$$

$$D(t, x) = L_{x'}C(x', x), \tag{6}$$

$$K(t, t') = L_x D(t, x) = L_x L_{x'} C(x', x), \text{ and} \tag{7}$$

$$B(t, x) = \frac{\partial}{\partial t} D(t, x) = \frac{\partial}{\partial t} L_{x'} C(x', x). \tag{8}$$

Sufficient conditions for these equalities are (i) m is a Gaussian random element with support in a separable Banach space with norm $\|.\|$ and finite expectation $E[\|m\|^2]$, and (ii) the following mappings are continuous linear functionals with respect to the norm $\|.\|$: (a) $m \to (L \cdot m)(t)$, (b) $m \to \frac{\partial}{\partial t}(L \cdot m)(t)$, and (c) the evaluation map $m \to m(x)$.

The proof of Lemma 1 and all further proofs are relegated to appendix A. The equalities of lemma 1 hold under more general conditions than those stated. In our applications, the linear operator L involves only integrals and pointwise linear transformations. For such operators, the exchangeability of linear operator and expectation follows immediately from the exchangeability of the order of integration (Fubini's theorem).

3 Optimal policy problems

The previous section has introduced the general optimal policy problem studied in this paper, where the policy objective is given by $u = L \cdot m + u_0$ for some average structural function $m(x) = E[g(x, \epsilon)]$ and some linear operator L. In this section we discuss a number of examples; optimal taxation and optimal insurance (section 3.1), optimal choice of inputs to a production function (section 3.2), and treatment assignment with peer effects (section 3.3). Examples such as these have been discussed extensively in the fields of public finance, development economics, and the economics of education. For each of these examples we are going to describe the operator L and the corresponding prior moments ν, D, B, and K.

[7] It is easy to show that u is (mean square) differentiable if the covariance kernel $K(x, x')$ is differentiable, see appendix B.

3.1 Optimal insurance and optimal taxation

Many problems of optimal policy in public finance share a similar structure. A typical example of such a public finance problem is optimal (unemployment) insurance, as first discussed by Baily (1978). Chetty (2006), building on insights of Feldstein (1999), has argued that a very general class of models lead to the same formulas characterizing optimal benefits. In terms of our setup, this general class of models can be summarized as follows.

Let $Y \in \{0, 1\}$ denote the state of an individual, where $Y = 1$ is the "bad" state (unemployment, sickness / seeing a doctor, ...). Let $t = x$ be the size of the transfer to individuals in the bad state. Let $Y^x = g(x, \epsilon)$ describe the (counterfactual) state of an individual given transfer level x. g is a behavioral relationship, reflecting all behavioral margins affecting the likelihood of being in state $Y = 1$. Let $m(x) = E[g(x, \epsilon)]$ be the share of the population that is in the bad state, given x. Let λ be the marginal value of income to individuals in the bad state, relative to the marginal value of government funds. We assume λ to be known.[8] [9]

With this notation, let us consider the effect of a marginal change dt of t on social welfare. There are potentially four components to the effect of such a change (cf. Chetty, 2009). This change has (i) a mechanical effect on private welfare of magnitude $\lambda m\ dt$ and (ii) a mechanical effect on government funds of $-m\ dt$. It has (iii) a behavioral effect on government revenues of magnitude $-tm'(t)\ dt$. Private utility maximization and the absence of externalities imply there is (iv) no behavioral effect on private welfare. This follows from envelope condition arguments which hold both in the continuous Y and in the discrete Y case.

Adding up these effects, we get the marginal effect of a change in t on social welfare,

$$u'(t) = (\lambda - 1) \cdot m(t) - t \cdot m'(t) = \lambda m(t) - \frac{\partial}{\partial t}(t \cdot m(t)). \tag{9}$$

The first order condition for optimal transfers is given by $u'(t^*) = 0$. Integrating and normalizing $u(0) = 0$ yields social welfare,[10]

$$u(t) = \lambda \int_0^t m(x)dx - t \cdot m(t). \tag{10}$$

We thus get $u_0 = 0$ and

$$(Lm)(t) = \lambda \int_0^t m(x)dx - t \cdot m(t). \tag{11}$$

Given the linear operator L, and given a Gaussian process prior for m as in assumption 4, we can characterize the prior moments ν, D, B, and K as in lemma 1; see box 1.

[8] In general, λ reflects social preferences for redistribution as well as private risk aversion. Emphasizing the latter interpretation, the public finance literature has attempted to estimate λ. I am inclined to share the perspective of Saez and Stantcheva (2012), who consider λ to ultimately reflect a political choice rather than an empirical parameter.

[9] We assume here that λ is known and constant. This is not necessary – all our results immediately generalize to non-constant and estimated λ. The resulting expressions are particularly simple if the posterior for λ is uncorrelated with the posterior for m.

[10] Notice the relationship between expression (10) and the "Harberger triangle." The latter corresponds to the special case where $\lambda = 1$.

For the optimal insurance model of section 3.1, the prior mean of u equals

$$\nu(t) = (L \cdot \mu)(t) = \lambda \int_0^t \mu(x)dx - t \cdot \mu(t). \tag{12}$$

The prior covariances D, B and K are given by

$$\begin{aligned}
D(t,x) = \mathrm{Cov}(u(t), m(x))) &= L_{x'} C(x', x) \\
&= \lambda \cdot \mathrm{Cov}\left(\int_0^t m(x')dx', m(x)\right) - t \cdot \mathrm{Cov}(m(t), m(x)) \\
&= \lambda \cdot \int_0^t C(x', x)dx' - t \cdot C(t,x), \qquad (13) \\
B(t,x) = \mathrm{Cov}\left(\frac{\partial}{\partial t} u(t), m(x)\right) &= \frac{\partial}{\partial t} D(t,x) = \frac{\partial}{\partial t} L_{x'} C(x', x) \\
&= (\lambda - 1) \cdot \mathrm{Cov}\left(m(t), m(x)\right) - t \cdot \mathrm{Cov}\left(\frac{\partial}{\partial t} m(t), m(x)\right) \\
&= (\lambda - 1) \cdot C(t,x) - t \cdot \frac{\partial}{\partial t} C(t,x), \text{ and} \qquad (14) \\
K(t,t') = \mathrm{Cov}(u(t), u(t')) &= L_x D(t,x) = L_x L_{x'} C(x', x) \\
&= \lambda^2 \cdot \int_0^t \int_0^{t'} C(x, x')dx'dx + t \cdot t' \cdot C(t, t') \\
&- \lambda \cdot \left(t' \cdot \int_0^t C(x, t')dx + t \cdot \int_0^{t'} C(x', x)dx'\right). \qquad (15)
\end{aligned}$$

Covariance Kernels 1: Optimal insurance

This setup is formally equivalent to a number of other problems considered in the literature. An example is the choice of the tax rate for the top tax bracket, as in Saez (2001). The above model applies immediately to this problem, once we change the interpretation of the parameters and some signs. To see this, let λ be the marginal value we assign to additional income for rich people relative to additional government revenues, let $Y^x = g(x, \epsilon)$ be the taxable income declared by an individual if she faces a top tax rate of x, and let $m(x)$ be the size of the tax base in the top bracket. The only difference to the insurance model is that Y is real-valued rather than binary, and that t is a tax rather than a transfer.

One difference between our treatment and the public finance literature lies in the absence of functional form restrictions. Chetty (2009) has surveyed the optimal policy literature in public finance, and emphasized the point that the first order conditions for optimal policy in general only involve some key behavioral elasticities *at the optimal policy.* The corresponding empirical implementations of such models estimate these behavioral elasticities *at some policy level.* Plugging these elasticities into the formulas for optimal policy implicitly assumes a parametric log-linear model for extrapolation of behavior outside the observed range. The non-parametric Bayesian approach which we propose might also need to extrapolate outside the support of the data using prior information, but it allows to quantify the uncertainty related to such an extrapolation. In contrast to the structural literature, our setting neither imposes functional form assumptions, nor does it restrict the dimensionality of unobserved heterogeneity ϵ.

A representation of the policy objective similar to the one above is more generally possible in settings which satisfy the following assumptions: The policy-maker's objective is to maximize a weighted sum of private utilities. Policy choices (such as tax rates or replacement rates) affect private choices. The government is subject to a budget constraint, or equivalently has alternative expenditures and revenues which pin down the marginal value of government revenues. If there are no externalities, these assumptions imply that the behavioral effects of policy choices on private welfare are zero at the margin, due to envelope conditions. This implies that welfare under a given policy choice only depends on some key behavioral relationship, for instance the tax base as a function of tax rates.

3.2 Optimal choice of inputs to a production process

Let us now consider settings where the policy-maker has to choose inputs $x \in \mathscr{X} \subset \mathbb{R}^{d_x}$ into some production process. The outcome of interest (output) is given by $Y = g(X, \epsilon)$, where g is the production function mapping observed inputs X and unobserved inputs ϵ into output Y. The policy-maker's objective is to maximize average (expected) output $E[Y]$, net of the costs of inputs. The unit-price of input x_j is given by p_j. The policy-maker's willingness to pay for a unit-increase in Y is given by λ. Equating $t = x$, this yields the objective function

$$u(t) = \lambda \cdot E[Y^t] - p' \cdot t. \tag{16}$$

Letting $m(t) = E[Y^t]$, $(Lm)(t) = \lambda \cdot m(t)$, and $u_0(t) = -p' \cdot t$, we get

$$u(t) = (Lm)(t) + u_0(t) = \lambda \cdot m(t) - p' \cdot t. \tag{17}$$

Given the linear operator L and the function u_0, and given a Gaussian process prior for m as in assumption 4, we can characterize the prior moments ν, D, B, and K as in lemma 1; see box 2.

For the production function model of section 3.2, the prior mean of u equals

$$\nu(t) = (L \cdot \mu)(t) + u_0(t) = \lambda \cdot \mu(t) - p' \cdot t. \tag{18}$$

The prior covariances D, B and K are given by

$$D(t,x) = \mathrm{Cov}(u(t), m(x))) = L_{x'} C(x', x) = \lambda \cdot C(t,x), \tag{19}$$

$$B(t,x) = \mathrm{Cov}(u'(t), m(x)) = \frac{\partial}{\partial t} D(t,x) = \frac{\partial}{\partial t} L_{x'} C(x', x) = \lambda \cdot \frac{\partial}{\partial t} C(t,x), \text{ and} \tag{20}$$

$$K(t,t') = \mathrm{Cov}(u(t), u(t')) = L_x D(t,x) = L_x L_{x'} C(x', x) = \lambda^2 \cdot C(t,t'). \tag{21}$$

Covariance Kernels 2: Choice of inputs

An important example of the kind of problem considered in this section is discussed in the literature on the economics of education. In this context, Y measures long-run student outcomes of interest, or (as a second best) proxies for these long-run outcomes such as test scores. The vector x is equal to educational inputs such as teachers per student (class size), teacher salaries (affecting self-selection into teaching), school facilities, extra tutoring, length of the school year, etc. λ is equal to the social willingness to pay for improvements in student long-run outcomes. A recent example of experimental evidence on the role of educational inputs is Fryer (2011), other references include Angrist and Lavy (1999), Krueger (1999), and Rivkin et al. (2005). Many further examples for such choice-of-inputs problems can be found in the experimental development economics literature, which aims to assess the (cost) effectiveness of various policies in achieving observable improvements; a survey can be found in Banerjee and Duflo (2008). Another problem fitting into the framework discussed here is the profit maximization problem of the firm, as treated in standard microeconomic theory (cf. Mas-Colell et al., 1995, chapter 5).

The policy problem discussed in this section assigns the same vector of inputs t to all units of observation. Section **??** considers a more general form of this problem, where the policy-maker has the option of conditioning the allocation of inputs t on some observed covariates w.

3.3 Peer effects

Two types of policies are discussed in the literature on social externalities or peer effects. The first assigns treatment to individuals who are members of fixed (predetermined) groups. The second assigns individuals with predetermined characteristics to groups. The objective function of both policy choice problems can be recast in terms of a reduced form relationship that fits into the framework of section 3.2. Here, as in other settings with externalities or equilibrium effects, we need to define our units of treatment in such a way that externalities only take place within these units. We next discuss a framework which directly addresses the problem of assigning treatment to individuals in fixed groups.

We then discuss how this framework can be re-interpreted to provide a solution to the problem of assigning individuals to groups.

Suppose individuals i are members of groups c ("classes") of size n_c, where c_i is such that $i \in c_i$. Let Y_i denote individual outcomes and define $Y_c = \frac{1}{n_c} \sum_{i \in c} Y_i$ as the average outcome within group c. Assume that the policy-maker's objective is to maximize the average outcome $E[Y_i]$ net of the cost of treatments. Assume that the outcome Y_i might depend not only on individual i's treatment X_i, but also on the treatments and outcomes of other members of group c_i,

$$Y_i = h(X_i, \{X_j, Y_j\}_{j \in c_i, j \neq i}, \epsilon_i). \tag{22}$$

The set of equations (22) for $i \in c$ forms a system of simultaneous equations. Solving for the equilibrium of this system we get a reduced form mapping from $\{X_i\}_{i \in c}$ and $\{\epsilon_i\}_{i \in c}$ to the vector of equilibrium outcomes $\{Y_i\}_{i \in c}$.[11] From this, we can in particular deduce the reduced form relationship

$$Y_c = g(\{X_j\}_{j \in c}, \epsilon_c), \tag{23}$$

where the group-level heterogeneity ϵ_c is a composite of all the individual-level ϵ_i, $\epsilon_c = \{\epsilon_i\}_{i \in c}$.

We assume (following Graham et al. 2008) that the individuals in a group are observationally equivalent ("exchangeable"), and in particular that any group-internal network structure is unknown to the researcher. Under this assumption, it is without loss of generality to assume that c only depends on the distribution of $(x_j)_{j \in c}$, but is invariant to permutations within the group.[12] If X_i is binary we can in particular define $X_c = \frac{1}{n_c} \sum_{i \in c} X_i$ and simplify the expression for c to

$$Y_c = g(X_c, n_c, \epsilon_c). \tag{24}$$

We assume that X_i is binary and that n_c is constant (or does not enter g) for the rest of this section. Notation gets more complicated without these restrictions; our arguments do, however, generalize.

Equation (24) describes the reduced form relationship that matters for the types of policy we consider. For our purposes it is not necessary to find a solution to the identification problem posed by simultaneity (the "reflection problem") which was discussed in Manski (1993). We do not need to distinguish between "exogenous" and "endogenous" peer effects.

Define the group-level average structural function m corresponding to the group-level reduced form relationship g as

$$m(x_c) := E[g(x_c, \epsilon_c)]. \tag{25}$$

As we assumed that the policy-maker's objective is to maximize average outcomes net of the cost of treatments, we get the objective function

$$u(t) = \lambda \cdot m(t) - p \cdot t. \tag{26}$$

[11] If there are multiple equilibria, the reduced form mapping additionally depends on an equilibrium selection mechanism, which we subsume into unobserved group heterogeneity ϵ_c.

[12] Note that this is a statement about the researcher's ignorance rather than about the underlying structure. For a more detailed discussion of this subtle point, see Graham et al. (2008).

Here p is the price of assigning treatment 1 to one individual. As before, we equate $t = x_c$, and λ is the policy-maker's willingness to pay for a unit-increase of individual outcomes Y. This is exactly the framework of section 3.2, so all the previous results apply.

This framework allows us to directly evaluate policies that choose the share t of individuals treated within each group. A typical example are the deworming programs studied in Miguel and Kremer (2003), where c indexes schools, i indexes students, $x_i = 1$ if a student was treated, and x_c is the share of students treated in a school. To the extent that there are positive externalities of treatment, as found in Miguel and Kremer (2003), the slope of m with respect to x is larger than individual treatment effects. To the extent that there are decreasing returns on a group level, less than complete treatment might be optimal.

The second of the two types of policies mentioned at the beginning of this section are policies that affect group structure. Such policies fit into a slightly modified version of our general setup.[13] Under a "double randomization" condition as in Graham et al. (2008), which guarantees that $X_c \perp \epsilon_c$, we can describe the problem as follows. In this case X_i is a predetermined individual characteristic, and $m(x_c)$ is the average (expected) outcome for groups with composition x_c. The policy-maker can choose a distribution of group compositions x_c which satisfies the constraints of the population distribution of individual characteristics x_i. She can choose any distribution of X_c which satisfies the constraint $E[X_c] = \overline{x}$. If the counterfactual policy also guarantees double randomization, we can describe the policy problem as

$$\max_{P(X_c)} E[m(x_c)] \quad \text{s.t.} \quad E[X_c] = \overline{x}.$$

We are now maximizing over a distribution for X_c, rather than requiring that all units c get the same treatment level. Denoting the Lagrange multiplier of the budget constraint by p/λ, the problem again takes the form

$$\max_{P(X_c)} E[\lambda \cdot m(X_c) - p \cdot X_c].$$

In general, there might be two values of x_c that maximize $\lambda \cdot m(x_c) - p \cdot x_c$ if the function m is not concave. To satisfy the budget constraint, the optimal policy might need to assign weight to both points.

4 Posterior, optimal policy choice, and optimal experimental design

In this section, we study the implications of the setup introduced in section 2 for optimal policy and experimental design; throughout we maintain assumptions 1 through 4. Section 4.1 characterizes the posterior mean of the regression function m, of the objective function $u = L \cdot m + u_0$, and of its derivative u'. The posterior is calculated conditional on experimental data (Y, X) that consist of i.i.d. draws (Y_i, X_i). Section 4.2 derives the optimal policy choice t^* conditional on the data X, Y. t^* maximizes the posterior expectation

[13] For a detailed discussion of the issues involved, see Graham et al. (2008).

of the objective function u (social welfare). Section 4.2 also discusses the optimal experimental design X^*, which maximizes ex ante expected welfare, assuming that t is chosen optimally ex post. Section 4.2, finally, provides expressions for the value of additional experimental observations, assuming the data are used to choose policy optimally.

Throughout the rest of the paper, we use θ as a shorthand for the data generating process so that expectations conditional on θ are "frequentist" expectations (such as $m(x) = E[Y_i|X_i = x, \theta]$), while unconditional expectations are "Bayesian" (such as $\widehat{m}(x) = E[m(x)|X, Y]$).

4.1 Posterior

In this subsection we derive and characterize the posterior mean of $m(x)$, of $u(t)$, and of $u'(t)$ given the matrix X and the vector of outcomes Y. Derivation and characterization of the posterior expectation of $m(x)$ is standard in the literature on Gaussian process regression, while the posterior mean of $u(t)$ and $u'(t)$ are specific to the setup considered in this paper.

In discussing the posterior expectation of $m(x)$, sections 4.1.1 through 4.1.3 provide a review of the literature on Gaussian process regression. This literature has early roots in geostatistics (Matheron, 1973) and is closely related to spline regression (Wahba, 1990). More recently, it has become central in the literature on machine learning; Williams and Rasmussen (2006) provide a good introduction. A general and thorough treatment of nonparametric Bayesian statistics can be found in (Ghoshal and van der Vaart, 2013, forthcoming).

4.1.1 The posterior expectation $\widehat{m}$ of m

We start by deriving the posterior expectation of the regression function $m(x) = E[Y_i|X_i = x, \theta]$. First, we give a representation of $\widehat{m}(x) := E[m(x)|X, Y]$ as the posterior best linear predictor of $m(x)$ in Y given X. We then provide an equivalent representation of $\widehat{m}(x)$ as the solution to a penalized regression problem. The combination of these two representations then motivates an asymptotic approximation in terms of the so called equivalent kernel.

Maintaining assumption 4, we use the following notation for the prior moments of m given X,[14]

$$\begin{aligned}\mu_i &= E[m(X_i)|X] = \mu(X_i),\\ C_{i,j} &= \mathrm{Cov}(m(X_i), m(X_j)|X) = C(X_i, X_j), \text{ and}\\ C_i(x) &= \mathrm{Cov}(m(x), m(X_i)|X) = C(x, X_i).\end{aligned} \tag{27}$$

Let furthermore μ, C, and $C(x)$ denote the vectors and matrix collecting these terms for $i, j = 1, \ldots, n$. In order to derive the posterior expectation of $m(x)$, we need to impose one more assumption.

[14] These prior moments do not depend on X by assumption 4.

Assumption 5 (Subjective expectations based on best linear predictor) *The subjective posterior expectations of the policy-maker are formed as posterior best linear predictors in Y assuming homoskedasticity of η_i, so that in particular*

$$\widehat{m}(x) = \underset{\check{m}(x)}{\operatorname{argmin}} \; E[(m(x) - \check{m}(x))^2|X]$$
$$s.t.\ \check{m}(x) = w_0(x) + \frac{1}{n}\sum_i w(x, X_i) \cdot Y_i \text{ for all } x. \tag{28}$$

Under this assumption,[15] the posterior mean of $m(x)$ is given by

$$\begin{aligned}\widehat{m}(x) = E[m(x)|X, Y] &= E[m(x)|X] + \operatorname{Cov}(m(x), Y|X) \cdot \operatorname{Var}(Y|X)^{-1} \cdot (Y - E[Y|X]) \\ &= \mu(x) + C(x) \cdot \left[C + \sigma^2 I\right]^{-1} \cdot (Y - \mu).\end{aligned} \tag{29}$$

Instead of directly imposing assumption 5, we could also assume that conditional on θ and X, the residuals $\eta_i := Y_i - m(X_i)$ are i.i.d. normally distributed with mean 0 and variance σ^2. This would yield a full specification of the prior.

4.1.2 Penalized regression representation of $\widehat{m}$

There are several alternative representations of the posterior expectation $\widehat{m}(x) = E[Y|X = x]$. It can in particular be written as the solution to a penalized regression, or as a "maximum a posteriori";[16]

$$\widehat{m} = \underset{\check{m}(.)}{\operatorname{argmin}} \left[\frac{1}{\sigma^2} \cdot \sum_i (Y_i - \check{m}(X_i))^2 + \|\check{m} - \mu\|_C^2\right], \tag{30}$$

where $\|m - \mu\|_C^2$ is a penalty term. The norm $\|m\|_C^2$ is the so called reproducing kernel Hilbert space norm corresponding to the covariance kernel C. It is defined as the norm corresponding to an inner product on the space of all linear combinations of functions of the form $C(x, .)$, and their limits, where $\langle C(x_1, .), C(x_2, .)\rangle = C(x_1, x_2)$. The representation (30) of $\widehat{m}$ shows the connection between Gaussian Process regression and spline regression, where $\|m\|_C^2$ is a smoothness penalty. For an introduction to the theory of reproducing kernel Hilbert spaces see Wahba (1990) and van der Vaart and van Zanten (2008b).

4.1.3 Equivalent kernel

By equation (29), the posterior expectation can be written in the form $\widehat{m}(x) = w_0(x) + \frac{1}{n}\sum_i w(x, X_i) \cdot Y_i$ for some weight function w. From this representation for $\widehat{m}$, it is straightforward to see that the weight function $w(., X_i)$ corresponds to the estimate of

[15] Note that almost all standard nonparametric regression methods (kernel, local polynomial, series, spline) are linear in Y, so that our setting is not more restrictive than any of these.

[16] To gain an intuition for this representation, consider the case where X_i has finite support, $\mu = 0$, and the residuals $\eta_i := Y_i - m(X_i)$ are i.i.d. normally distributed. In this case the right hand side of equation (30) corresponds to the joint log likelihood of m and Y. The squared norm $\|m\|_C^2$ is then equal to $m^t \cdot \operatorname{Var}(m)^{-1} \cdot m$. Since m and Y are jointly normally distributed, the posterior mean is furthermore equal to the posterior mode.

$\widehat{m}$ we would obtain if we had $Y_i = n$ and $Y_j = 0$ for $j \neq i$, and if we replace μ by 0. Combining this insight with representation (30), we get

$$\begin{aligned} w(., X_i) &= \underset{\check{w}(.)}{\operatorname{argmin}} \left[\sum_{j \neq i} \check{w}(X_i)^2 + (n - \check{w}(X_i))^2 + \sigma^2 \cdot \|\check{w}\|_C^2 \right] \\ &= \underset{\check{w}(.)}{\operatorname{argmin}} \left[\frac{1}{2} \int \check{w}(x)^2 dF_n(x) + \frac{\sigma^2}{2n} \cdot \|\check{w}\|_C^2 - \check{w}(X_i) \right] \qquad (31) \end{aligned}$$

where F_n is the empirical distribution function of X. This representation suggests to approximate w with the solution to the minimization problem

$$\overline{w}(., x') = \underset{\check{w}(.)}{\operatorname{argmin}} \left[\frac{1}{2} \int \check{w}(x)^2 dF(x) + \frac{\sigma^2}{2n} \cdot \|\check{w}\|_C^2 - \check{w}(x') \right]. \qquad (32)$$

The solution to this latter minimization problem is called the equivalent kernel, (cf. Silverman, 1984; Sollich and Williams, 2005; Williams and Rasmussen, 2006, chapter 7). The equivalent kernel does not depend on the data, except through the sample size n which scales the penalty term $\|m\|_C^2$. The validity of this approximation hinges on the uniform closeness of $\int m(x)^2 dF_n(x)$ and $\int m(x)^2 dF(x)$. We use implications of this approximation in section 5, where we derive the frequentist asymptotic behavior of our proposed policy choice function.

4.1.4 The posterior expectation $\widehat{u}$ of u

Our main object of interest is not m but u. The posterior expectation of m, however, maps easily into the posterior expectation of u, given the linearity of $u = L \cdot m + u_0$. Under assumption 5, the posterior expectation $\widehat{u}(t) := E[u(t)|X, Y]$ is again given by the posterior best linear predictor of u in Y given X. Recall that we defined $D(t, x) := \operatorname{Cov}(u(t), m(x))$, and that under the conditions of lemma 1 we had $D(t, x) = L_{x'} C(x', x)$. We get

$$\begin{aligned} \widehat{u}(t) = E[u(t)|X, Y] &= E[u(t)|X] + \operatorname{Cov}(u(t), Y|X) \cdot \operatorname{Var}(Y|X)^{-1} \cdot (Y - E[Y|X]) \\ &= \nu(t) + D(t) \cdot \left[C + \sigma^2 I\right]^{-1} \cdot (Y - \mu). \qquad (33) \end{aligned}$$

The linearity of conditional expectations furthermore implies

$$\widehat{u} = L \cdot \widehat{m} + u_0 = (L \cdot w_0 + u_0) + \sum_i (L \cdot w(., X_i)) \cdot Y_i. \qquad (34)$$

In order to approximate this representation, we can substitute $\overline{w}$ for w. This yields an equivalent kernel for the posterior mean of u, where the kernel is again independent of the data.

4.1.5 The posterior variance of m, u and u'

Since our estimator $\widehat{m}(x)$ is a posterior best linear predictor, we get that the residual $m(x) - \widehat{m}(x)$ is uncorrelated with $\widehat{m}(x)$ under the prior. This implies that the residual variance is given by

$$\operatorname{Var}(m(x) - \widehat{m}(x)) = \operatorname{Var}(m(x)) - \operatorname{Var}(\widehat{m}(x)),$$

the difference between the prior variance of $m(x)$, and the prior variance of the estimator $\widehat{m}(x)$. Under the assumption of joint normality this variance is equal to the posterior variance of $m(x)$. The same considerations hold for $\widehat{u}$ and $\widehat{u}'$, and their respective best linear predictors.

The prior variance of $m(x)$ is given by $\operatorname{Var}(m(x)) = C(x,x)$ by assumption, and similarly $\operatorname{Var}(u(t)) = K(t,t)$ (cf. Lemma 1) and $\operatorname{Var}(u'(t)) = \frac{\partial^2}{\partial t \partial t'} K(t,t')$. The prior variance of their best linear predictors equals

$$\begin{aligned}
\operatorname{Var}(\widehat{m}(x)) &= C(x) \cdot \left[C + \sigma^2 I\right]^{-1} \cdot C(x)', \\
\operatorname{Var}(\widehat{u}(t)) &= D(t) \cdot \left[C + \sigma^2 I\right]^{-1} \cdot D(t)', \text{ and} \\
\operatorname{Var}(\widehat{u}'(t)) &= B(t) \cdot \left[C + \sigma^2 I\right]^{-1} \cdot B(t)'. \qquad (35)
\end{aligned}$$

4.2 Optimal policy and optimal experimental design

In section 4.1, we derived and discussed the posterior expectation $\widehat{u}$ of the policy-maker's objective function u given experimental data X, Y. In this section we use this posterior expectation to characterize the optimal policy choice t^* given the data, and the optimal experimental design X^*. Both are chosen to maximize the expected value of u. We further discuss the return to increasing sample size n, and the value of conducting an experiment, in terms of increasing expected social welfare.

4.2.1 Optimal policy

Our analysis is motivated by the problem of choosing t given experimental data X, Y. Under assumptions 1, 2, and 3, the optimal t satisfies

$$\widehat{t^*} = \widehat{t^*}(X,Y) \in \underset{t}{\operatorname{argmax}}\ \widehat{u}(t), \qquad (36)$$

where as before $\widehat{u}(t) = E[u(t)|X,Y]$. Under assumptions 4 and 5, the objective function of this maximization problem is given by

$$\widehat{u}(t) = \nu(t) + D(t) \cdot \left[C + \sigma^2 I\right]^{-1} \cdot (Y - \mu),$$

as shown in section 4.1. Assuming differentiability of $\widehat{u}$, which immediately follows from differentiability of the covariance kernel D, the first order condition for $\widehat{t^*}$ is $\frac{\partial}{\partial t}\widehat{u}(\widehat{t^*}) = 0$. By similar arguments as before we get

$$\begin{aligned}
\frac{\partial}{\partial t}\widehat{u}(t) = E[u'(t)|X,Y] &= E[u'(t)|X] + \operatorname{Cov}(u'(t), Y|X) \cdot \operatorname{Var}(Y|X)^{-1} \cdot (Y - E[Y|X]) \\
&= \nu'(t) + B(t) \cdot \left[C + \sigma^2 I\right]^{-1} \cdot (Y - \mu). \qquad (37)
\end{aligned}$$

The problem of finding the optimal policy $\widehat{t^*}$ reduces, at least locally, to the problem of finding a solution to the equation $\widehat{u}'(t) = 0$. This is reflected in the asymptotic distribution of $\widehat{t^*}$ derived in section 5, which is driven by variation in $\widehat{u}'(t^*)$. As far as numerical implementation is concerned, any standard optimization algorithm, such as the Newton-Raphson algorithm, can be used to find the maximum $\widehat{t^*}$ of $\widehat{u}$.

4.2.2 Optimal experimental design

Consider next the choice of an optimal experimental design X^*. The optimal design maximizes ex-ante expected welfare. Ex ante welfare, as a function of X, is defined assuming that the policy t is chosen as $t^*(X, Y)$ once the experiment is completed, and X and Y are known. Define

$$\widehat{\upsilon}(X) := E[\max_t \widehat{u}(t)|X] = E[\widehat{u}(\widehat{t^*})|X]. \tag{38}$$

Then the optimal experimental design satisfies

$$X^* \in \underset{X}{\operatorname{argmax}}\ \widehat{\upsilon}(X). \tag{39}$$

Under the same assumptions as before, we can plug in our expressions for $\widehat{u}$ to get

$$\widehat{\upsilon}(X) := E[\max_t \nu(t) + D(t) \cdot \left[C + \sigma^2 I\right]^{-1} \cdot (Y - \mu)],$$

where $\operatorname{Var}(Y) = C + \sigma^2 I$. The dependence on X of this expression is implicit, through the dependence of D, C, and the distribution of Y on the design points X_i.

In section 6 below we discuss characterizations of the optimal design X^*. These characterizations are based on a continuous approximation of our model, where we replace the problem of choosing a discrete design distribution $F_n(x)$ (i.e., choosing X) by the problem of choosing a continuous design density $f(x)$.

4.2.3 Optimal sample size

Consider, finally, the value of adding observations to our sample, and the value of the whole experiment. Both are characterized by the following value function, based on the value of the optimal experimental design for experiments of size n.[17]

$$\widehat{\upsilon}(n) := \max_{X \in \mathscr{X}^n} \widehat{\upsilon}(X) = \max_{X \in \mathscr{X}^n} E[\max_t \widehat{u}(t)|X] = E[\widehat{u}(\widehat{t^*})|X = \mathbf{X}^*]. \tag{40}$$

The value of adding an observation to the sample is given by $\widehat{\upsilon}(n+1) - \widehat{\upsilon}(n)$. The value of the whole experiment is given by $\widehat{\upsilon}(n) - \widehat{\upsilon}(0)$, where

$$\widehat{\upsilon}(0) = \max_t E[u(t)]$$

is the prior expected maximum of u. The optimal sample size satisfies

$$n^* = \underset{n}{\operatorname{argmax}} \left(\widehat{\upsilon}(n) - \sum_{i=1}^{n} c(i) \right). \tag{41}$$

Here $c(i)$ is the cost of an additional unit of observation at sample size i, in an appropriately chosen scale.

[17] Alternatively, if for some reason a non-optimal design is chosen, we could define a similar value function without the $\max_X$ step.

5 Frequentist asymptotics

In this section, we characterize the frequentist asymptotic behavior of $\widehat{t}^*$ and $u(\widehat{t}^*)$. Under certain regularity conditions, $\widehat{t}^*$ is asymptotically normal with a rate of convergence slower than $\sqrt{n}$. This limiting distribution and all other limits, expectations, and variances in this section are conditional on θ, that is conditional on the population distribution of (X, Y).

Theorem 1 (Asymptotic normality of the optimal policy choice) *Under assumptions 1 through 6,*

$$V_n^{-1/2} \cdot \left(\widehat{t}^* - t^* - B_n\right) \to^d N\left(0, I\right), \tag{42}$$

for

$$\begin{aligned} V_n &= l^2 \cdot \left(u''(t^*)^{-1} \cdot V_{n,m'} \cdot u''(t^*)^{-1}\right), \\ B_n &= -l \cdot u''(t^*)^{-1} \cdot B_{n,m'}, \end{aligned}$$

where $V_n \to 0$, $B_n \to 0$, and $l = -\widehat{t}^*$ for the optimal insurance problem; $l = \lambda$ for the optimal choice of inputs problem. $V_{n,m'}$ and $B_{n,m'}$, the asymptotic variance and bias of $\widehat{m}'(t^*)$, are defined in assumption 6.

The proof of this theorem can again be found in appendix A. Theorem 1 holds under the following regularity conditions. These conditions restrict the function u to be smooth and have a negative definite second derivative at the optimum. These conditions also place high-level restrictions on the asymptotic behavior of the regression $\widehat{m}$. More primitive conditions for the asymptotic behavior of $\widehat{m}$ estimated by Gaussian process regression are discussed in section 5.1.

Assumption 6 (Regularity conditions)

1. *u is twice continuously differentiable and $u''(t^*)$ is negative definite.*
2. *t has compact support, and t^* is at an interior point of this support.*
3. *u corresponds to the objective of one of the policy problems considered in section 3.*
4. *$\widehat{m}$ converges to m in $\mathscr{C}^2$.*
5. *$(\widehat{m}(t^*) - m(t^*)) = o_p(\widehat{m}'(t^*) - m'(t^*))$ and*

   $$V_{n,m'}^{-1/2} \cdot (\widehat{m}'(t^*) - m'(t^*) - B_{n,m'}) \to_d N(0, I)$$

   for sequences $V_{n,m'} \to 0$ and $B_{n,m'} \to 0$.

Remarks:

- The framework analyzed in sections 1 to 4 is based on a decision theoretic, Bayesian paradigm. The policy choice $\widehat{t}^*$ is by construction optimal in finite samples for the objective of maximizing u under the given prior. This optimality is to be understood in an *average* sense across possible functions m.

  The arguments in this section are asymptotic (approximate) and frequentist, that is, *conditional* on m. This analysis of conditional performance is a useful complement to the analysis of average performance.

- To gain intuition for theorem 1, note that we can think of $\widehat{t}^* = \text{argmax}_t\, \widehat{u}(t)$ as an M-estimator. Asymptotic normality of $\widehat{t}^*$ follows if we can show (i) consistency of $\widehat{u}$ as an element of $\mathscr{C}^2$, and (ii) asymptotic normality of $\widehat{u}'(t^*)$, (cf. van der Vaart, 2000, chapter 5). Asymptotic normality of $\widehat{u}'(t^*)$ follows from asymptotic normality of $\widehat{m}'$ for the applications discussed in section 3. Similarly, $\mathscr{C}^2$ consistency of $\widehat{u}$ follows from $\mathscr{C}^2$ consistency of $\widehat{m}$. The asymptotic behavior of $\widehat{m}$ is discussed in subsection 5.1.

- This line of reasoning does not require that $\widehat{m}$ be estimated using Gaussian process regression. Indeed the same results would follow for any of a number of nonparametric regression methods that imply asymptotic normality of $\widehat{m}'$. This is true in particular for kernel regression and for sieve- (series-) regression.

- Theorem 1 does not impose any assumptions of homoskedasticity or normality of residuals. A prior belief of homoskedasticity was used in section 4.2 to derive the posterior expectation of $\widehat{u}$, the optimal policy choice $\widehat{t}^*$, and the objective function of experimental design $\widehat{v}$. Homoskedasticity is not needed for the asymptotic theory developed here.

- Theorem 1 shows asymptotic normality of $\widehat{t}^*$. This allows to construct asymptotically valid frequentist confidence sets for t^*, given an estimator of the variance of $\widehat{t}^*$. Proposition 1 below provides such an estimator.

Theorem 1 characterizes the asymptotics of the maximizer $\widehat{t}^*$ of posterior expected social welfare. Based on this characterization, we can derive the asymptotic behavior of $u(\widehat{t}^*)$, the maximum value of posterior expected social welfare.

Corollary 1 (Asymptotic distribution of regret) *Under the assumptions of theorem 1, if $t \in \mathbb{R}$, and*

- *if $V_n^{-1/2} \cdot B_n \to 0$, then*

$$\frac{2}{V_n \cdot u''(t^*)} \cdot \big(u(t^*) - u(\widehat{t}^*)\big) \to^d \chi_1^2, \tag{43}$$

  where

$$\frac{2}{V_n \cdot u''(t^*)} = \frac{2 \cdot u''(t^*)}{V_{n,m'} \cdot l^2}.$$

- *If $V_n^{-1/2} \cdot B_n \to \infty$, then*

$$V_{n,u^*}^{-1/2} \cdot \big(u(\widehat{t}^*) - u(t^*) - B_{n,u^*}\big) \to N(0,1), \tag{44}$$

  where

$$\begin{aligned} V_{n,u^*} &= u''(t^*)^2 \cdot V_n \cdot B_n^2 \\ &= l^4 \cdot V_{n,m'}(t^*) \cdot B_{n,m'}(t^*)^2 \\ B_{n,u^*} &= u''(t^*) \cdot B_n^2/2 \\ &= l^2 \cdot u''(t^*) \cdot B_{n,m'}(t^*)^2/2. \end{aligned}$$

Proposition 1 (Estimating the variance of $\widehat{t^*}$) *Under the assumptions of theorem 1, the variance of $\widehat{t}^*$ can be estimated by*

$$\widehat{V} = \widehat{l}^2 \cdot \left(\widehat{u}''(\widehat{t}^*)^{-1} \cdot \widehat{V}_{m'} \cdot \widehat{u}''(\widehat{t}^*)^{-1}\right), \tag{45}$$

where

$$\widehat{V}_{m'} = \sum_i w_t(\widehat{t}^*, X_i)^2 \cdot (Y_i - \widehat{m}(X_i))^2, \tag{46}$$

and w_t is the derivative of the coefficients of the best linear predictor defined in assumption 5. This estimator satisfies $\widehat{V}/V_n \to^p 1$ if the following additional conditions hold:

1. *$1/M < E[\epsilon_i^2|X_i = x] < M$ and $E[\epsilon_i^4|X_i = x] \leq M$ where $\epsilon_i = Y_i - m(X_i)$, for all x and some constant $M < \infty$.*
2. *$W_n := \frac{\max_i w_{t,i}^2}{\sum_i w_{t,i}^2} = o_p(n^{-1/2})$.*
3. *$\|\widehat{m} - m\|_n = o_p\left(n^{-1} \cdot W_n^{-1}\right)$, where $\|.\|_n$ is the L^2 norm w.r.t. $P_n(x)$, $\|g\|_n^2 = \int g^2(x) dP_n(x)$.*

5.1 The asymptotic behavior of $\widehat{m'}$

The asymptotic results of theorem 1, corollary 1, and proposition 1 are based on the assumption of $\mathscr{C}^2$ consistency of $\widehat{m}$ and asymptotic normality of $\widehat{m}'(t^*)$. These are standard properties of nonparametric regression estimates of $m(x) = E[Y|X = x]$ when m is estimated using estimation methods such as kernel regression or series regression; the development of corresponding results for Gaussian process regression is an area of active research. Here we provide a proof of asymptotic normality based on an approximation of Gaussian process regression by a kernel regression estimator.

General discussions of the current state of research on the convergence properties of Gaussian process regression can be found in van der Vaart and van Zanten (2008b), van der Vaart and van Zanten (2008a), and (Ghoshal and van der Vaart, 2013, forthcoming). Conditions for consistency of $\widehat{m}$ are discussed, in particular, in Ghoshal and van der Vaart (2013). They require m to be in the appropriately defined closure of the Reproducing Kernel Hilbert Space corresponding to the Gaussian process prior for m.

We prove asymptotic normality of $\widehat{m}'$ for settings where $\widehat{m}$ can be approximated by a kernel regression estimator using the so called equivalent kernel. Silverman (1984) and Sollich and Williams (2005) have shown in the context of special cases (smoothing spline regression kernel, squared exponential kernel), that such an approximation is asymptotically valid. From this approximation by kernel regression we get asymptotic normality of $\widehat{m}$ and of $\widehat{m}'$ using standard results on kernel regressions, with rates of convergence which are slower than $\sqrt{n}$. For a review of the asymptotic behavior of kernel regressions see for instance (Li and Racine, 2011, chapter 2).

How can the equivalent kernel be characterized? In section 4.1 we saw that the weight function of Gaussian process regression can be approximated by weights which are independent of the empirical distribution of X, $F_n(x)$. We shall now apply two further steps of approximation, (i) approximating the design density $f(x)$ by a locally constant density $f(x) = \rho/n$, and (ii) approximating the shape of the kernel by its large n limit, after suitable rescaling. These approximations yield

$$\widehat{m}(x) = m_0(x) + \frac{1}{\rho}\sum_i \frac{1}{h(\rho)^{d_x}} K\left(\frac{X_i - x}{h(\rho)}\right) \cdot Y_i + \Delta_m(x), \tag{47}$$

where we define

$$K_\rho^m = \underset{K(.)}{\operatorname{argmin}} \left[\frac{1}{2} \cdot \int K(x)^2 dx + \frac{\sigma^2}{2\rho} \cdot \|K\|_C^2 - K(0)\right], \tag{48}$$

and

$$K(x) = \lim_{n\to\infty} h(\rho)^{d_x} K_\rho^m(x \cdot h(\rho)). \tag{49}$$

In this equation $h(\rho)$ is a suitably chosen bandwidth parameter that yields a non-degenerate limit kernel K, and $\rho = f(x) \cdot n$. Lemma 2 assumes that we can asymptotically neglect the remainder Δ_m of this approximation.

Lemma 2 (Asymptotics of $\widehat{m}'$)
Assume that the remainder term $\Delta_m(x)$ in equation (47) converges to 0 in $\mathscr{C}^2$, given θ, with respect to the norm $\|\Delta\| = \sup_x |\Delta(x)| + \sup_x |\Delta'(x)| + \sup_x |\Delta''(x)|$.[18] Denote $\sigma^2(x) = \mathrm{Var}(Y_i|X_i = x, \theta)$, assume $\sigma^2(x)$ is continuous in x, and denote j the smallest integer such that

$$\int K(\tilde{x})\tilde{x}^j d\tilde{x} \neq 0.$$

Assume that $m \in \mathscr{C}^{j+1}$. Then

$$V_{n,m'}^{-1/2} \cdot (\widehat{m}'(x) - m'(x) - B_{n,m'}) \to_d N(0, I)$$

given θ for

$$V_{n,m'} = \frac{\sigma^2(x)}{\rho \cdot h(\rho)^{d_x+2}} \int K'(\tilde{x}) \cdot K'(\tilde{x})^t d\tilde{x}$$

and

$$B_{n,m'} = \frac{m^{(j+1)}(x) \cdot h(\rho)^j}{(j+1)!} \cdot \int K(\tilde{x})\tilde{x}^j d\tilde{x}.$$

[18] Whether this is a valid assumption depends on the covariance kernel C, the rate of convergence of F_n to F, and the smoothness of f. A necessary condition on the kernel $C(x, x')$ is that it depends only on the difference $x - x'$. In pioneering work, Silverman (1984) has shown that this assumption holds for the case of spline regression, where $\|K\|_C^2 = \int K''(x)^2 dx$ and $d_x = 1$. In this case $K(t) = \frac{1}{2}\exp(-|t|/\sqrt{2})\sin(|t|/\sqrt{2} + \pi/4)$ and $h(\rho) = (\sigma^2/\rho)^{1/4}$.

More recently Sollich and Williams (2005) have shown that a similar result holds for the squared exponential kernel, $C(x, x') = \frac{1}{\sqrt{2\pi}}\exp(-(x - x')^2/(2l^2))$. In this case the bandwidth is equal to $h(\rho) = \sqrt{2\pi^2 l^2/\log(\rho(2\pi l^2)^{d_x/2}/\sigma^2)}$.

6 Characterizing the optimal experimental design

In this section we characterize the optimal experimental design, that is the distribution of design points X_i which maximizes expected social welfare $\widehat{v}$, assuming that the policy t will be chosen optimally given the experimental data. We employ an approximation to the setup considered so far. This approximation assumes that we have to choose a design density $f(x)$ rather than a discrete set of design points x_i. Using this approximation allows for a more elegant and simple characterization of optimal designs. This approximation is exactly the same which motivates the equivalent kernel introduced in section 4.1.3 above.

This section is structured as follows. We first introduce the continuous model and provide a brief discussion. We then derive the posterior mean $\widehat{u}$ of u, adapting the results of section 4.1 to the continuous case. Using the resulting expressions for $\widehat{u}$, we get the prior distribution of $\widehat{u}$; in particular the prior covariance kernel for $\widehat{u}$. The prior expectation of social welfare is given by $\widehat{v} = E[\max_t \widehat{u}(t)]$. This implies that the prior distribution of $\widehat{u}$ determines $\widehat{v}$, which is the objective our experimental design aims to maximize. We can characterize the effect of perturbations to the design density f on the prior covariance kernel of $\widehat{u}$. This leads to our central result, giving the effect of perturbations to the design density on $\widehat{v}$. This result allows us to provide first order conditions for the optimal experimental design. An asymptotic approximation to these first order conditions yields easily implementable guidelines for design. The section concludes with a discussion of the value – in social welfare terms – of increasing sample size.

6.1 A continuous model

We continue to assume that the policy-maker aims to maximize expected social welfare (assumption 1), that the objective function u is given by $u = Lm + u_0$ for some linear operator L (assumption 2), and that the policy maker has a prior distribution for the average structural function m of $m \sim GP(\mu, C)$ (assumption 4). We also maintain the assumption that posterior expectations are formed as best linear predictors in Y (assumption 5). We drop assumption 3, which states that we observe n i.i.d. draws (X_i, Y_i) such that $E[Y_i|X_i] = m(X_i)$. We replace this assumption by assumption 7, which states that we observe a (generalized) stochastic process $Y(x)$, where Y is equal to m plus some appropriately scaled white noise process. The scaling depends on the design density, where a larger density translates into lower noise.

Assumption 7 (Continuous model) *The policy-maker observes $Y(.)$, where*

$$Y(x) = m(x) + \frac{\sigma}{\sqrt{nf(x)}} dW(x). \tag{50}$$

In this equation dW is a standard white noise process independent of m, and f is the design density.

Remarks:

- Like continuous white noise, the process $Y(.)$ in assumption 7 has to be understood as a *generalized* stochastic process. That is, $Y(x)$ is not a well defined random variable, but $\int Y(x)g(x)dx$ is for any g which is continuous and bounded. The properties of generalized stochastic processes are pinned down by the joint distribution of such integrals; we will characterize this joint distribution shortly.

A concise discussion of white noise and generalized stochastic processes can be found in Appendix I of Yaglom (1962). For a thorough treatment of stochastic processes such as the one in equation (50), see Karatzas and Shreve (1991). In this paper we take it as given that assumption 7 as well as the moments considered below are meaningful, and ignore measure theoretic issues.

- In the next subsection we consider the posterior distribution of m and u given Y, and in particular the posterior expectation $\widehat{u} = E[u|Y]$. Calculating this posterior expectation is in fact a variant of the classic "Wiener filtering" problem, introduced by Wiener (1949).
- Assumption 7 provides the natural continuous equivalent of the discrete setting we considered so far. To illustrate this, suppose for a moment that m and f are constant on the intervals $(j, j+1]$ for integers j. Then observing Y is equivalent to observing $Y^j := \int_j^{j+1} Y(x)dx$ for every j, where (conditional on m) $Y^j \sim N(m(j+1), \sigma^2/(nf(j+1)))$ and $Y^j \perp Y^k$ for all $j \neq k$. But this is the same model as before, where we observe $nf(j+1)$ design points in the interval $(j, j+1]$, and outcomes have variance σ^2. Put differently, we can equivalently think of a larger design density as increasing the number of observations or as decreasing the variance of noise.
- Assumption 7 also rationalizes the equivalent kernel of equation (32) (see equation (57) below), which we similarly motivated by an approximation of the discrete design X by a continuous design density $f(x)$.

In order to proceed, we need to introduce some additional notation. It is common to identify matrices A with the linear mapping $x \to Ax$. We can similarly identify functions such as the covariance kernel C with integral operators $g \to Cg$. For the kernel C corresponding to the Gaussian process m, we write in linear operator notation

$$\begin{aligned}(Cg)(x) &:= \operatorname{Cov}\left(m(x), \int m(x')g(x')dx'\right)\\ &= \int C(x,x')g(x')dx' \end{aligned} \tag{51}$$

For the noise process $\frac{\sigma}{\sqrt{nf(x)}}dW(x)$ of assumption 7, we have

$$(Eg)(x) := \frac{\sigma^2}{nf(x)} \cdot g(x) = \frac{\sigma^2}{n} \cdot F^{-1}, \tag{52}$$

where F is the operator

$$(Fg)(x) = f(x) \cdot g(x). \tag{53}$$

These equations can in fact be taken as providing a definition of the noise process $\frac{\sigma}{\sqrt{nf(x)}}dW(x)$ of assumption 7. Heuristically, this definition of the covariance operator E can be justified by the formal calculation

$$(Eg)(x) = \operatorname{Cov}\left(\frac{\sigma}{\sqrt{nf(x)}}dW(x), \int \frac{\sigma}{\sqrt{nf(x')}}dW(x')g(x')dx'\right) = \frac{\sigma^2}{nf(x)} \cdot g(x).$$

6.2 Posterior means

Under assumption 5, $\widehat{m} = E[m|Y]$ is given by the best linear predictor of m in Y,[19] so that in particular

$$\widehat{m} = \mu + W(Y - \mu) \tag{54}$$

for a linear operator W. This is, in fact, the operator corresponding to the equivalent kernel $\overline{w}$ of equation (32). Using the operator notation just introduced, we have $\mathrm{Cov}(\widehat{m}, Y) = W\,\mathrm{Var}(Y) = W(C + E)$, and $\mathrm{Cov}(m, Y) = \mathrm{Var}(m) = C$. The usual conditions for the best linear predictor of $m(x)$ given Y imply that

$$\mathrm{Cov}\left(\widehat{m}(x) - m(x), Y(.)\right) = 0 \tag{55}$$

for all x. Imposing this condition for all x implies the operator equation

$$W(C + E) - C = 0, \tag{56}$$

and thus

$$W = C(C + E)^{-1}. \tag{57}$$

Invertibility of $(C+E)$ holds because of the positive semidefiniteness of C and the positive definiteness of E. This equation gives an alternative representation for the equivalent kernel $\overline{w}$, as the solution to an integral equation. (Recall that in section 4.1, we derived the equivalent kernel as the solution to a functional minimization problem.) Using the linearity of expectations and the regularity conditions of lemma 1,

$$\begin{aligned}\widehat{u} = E[u|Y] &= E[Lm + u_0|Y] \\ &= L\widehat{m} + u_0 \\ &= \nu + LW(Y - \mu) \\ &= \nu + LC(C + E)^{-1}(Y - \mu).\end{aligned} \tag{58}$$

Note that $LC = D$ is the operator corresponding to the covariance kernel D given by definition 1.

6.3 The prior distribution of the posterior mean $\widehat{u}$

The prior variance (covariance kernel) of Y is given by $C + E$. The prior variance V of $\widehat{u}$ is therefore given by

$$V := LW(C + E)(LW)^t = LC(C + E)^{-1}(LC)^t. \tag{59}$$

Note the analogy to the algebra of finite dimensional covariance matrices and best linear predictors. We summarize what we have derived so far in the following lemma.

Lemma 3 (Prior distribution of $\widehat{u}$)
Suppose that assumptions 2, 4, 5 and 7 hold. Then the prior distribution of $\widehat{u}$ is given by

$$\widehat{u} \sim GP(\nu, LC(C + E)^{-1}(LC)^t), \tag{60}$$

where $E = \frac{\sigma^2}{n} \cdot F^{-1}$.

[19] This section replicates section 4.1.1 for the continuous case.

Note that the experimental design density f and the sample size n only affect the prior distribution of $\widehat{u}$ through the operator E. Our objective is to find conditions for the optimal f which maximizes $\upsilon(f) = E[\max \widehat{u}]$. The following reasoning leads to first order conditions for the optimal f.

Consider local perturbations of the density f by δ times a density $\tilde{f}$, or equivalently of the operator F by δ times the operator $\tilde{F}$. This results in the covariance kernel

$$V(\delta) = LC(C + E(\delta))^{-1}(LC)^t. \tag{61}$$

for $E(\delta) = \frac{\sigma^2}{n} \cdot (F + \delta\tilde{F})^{-1}$. Taking derivatives at $\delta = 0$ (denoted by subscripts), we get

$$\begin{aligned} V_\delta &= -LC(C+E)^{-1}E_\delta(C+E)^{-1}(LC)^t \\ &= -\overline{W}E_\delta\overline{W}^t, \end{aligned} \tag{62}$$

where we denote $\overline{W} := LW = LC(C + E)^{-1}$ the equivalent kernel for $\widehat{u}$, and where

$$E_\delta = -\frac{\sigma^2}{n} \cdot F^{-1}\tilde{F}F^{-1}. \tag{63}$$

Expanding equation (62), we get

$$(V_\delta g)(x) = (-\overline{W}E_\delta\overline{W}^t g)(x) = \frac{\sigma^2}{n}\int \overline{w}(x, x')\overline{w}(x'', x')\frac{\tilde{f}(x')}{f(x')^2}g(x'')dx'dx''. \tag{64}$$

Remarks:

- Equations (62) and (64) reflect an envelope condition for the optimal weights W of the best linear predictor. Changing the data generating process could affect the variance of $\widehat{u}$ both directly and through its effect on W, but by optimality of W, in terms of mean squared error, the latter effect drops out.

- Better estimates of u lead to better estimates of t^* and thus to higher expected social welfare. Better estimates $\widehat{u}$ of u are those with lower expected residual variance, or equivalently with larger prior variance. Equation (64) can thus be understood to imply "decreasing returns" of an increase in the design density at a point, where the "returns" are in terms of the variance of the predictor.

- We do not need to use the normality of the prior and the noise process implicit in assumption 7 to derive equation (64), and we will not need it for the conclusions of the following sections to hold. As in the discrete X case, normality is not essential for our argument as long as we restrict attention to best linear predictors.

6.4 The optimal design

With these preliminaries, we can now derive a general characterization of the optimal experimental design f^*. We consider the effect on expected social welfare $\widehat{\upsilon}(f^* + \delta \cdot \tilde{f})$ of local perturbations $\delta \cdot \tilde{f}$ to the optimal design density f^*. It is sufficient to consider perturbations that add a point mass at points x' to the design density f; knowing the effect of such perturbations allows to pin down the effect of any perturbation. Theorem

2 provides the central result of this section. This theorem characterizes the effect of such local perturbations on expected welfare $\widehat{v}$. Corollary 2 then uses the result of theorem 2 to characterize the optimal experimental design, which maximizes $\widehat{v}(f)$ subject to the constraint $\int f(x)dx = 0$. Proposition 2 uses an asymptotic approximation to the condition of corollary 2 that allows to easily calculate an approximately optimal design.

Theorem 2 (Perturbations to the design and expected social welfare)
Suppose that assumptions 2, 3, 4 and 7 hold. Let $\tilde{f}$ be a unit point mass at the point x'. The marginal effect on expected social welfare ν of a perturbation of the design density f by $\tilde{f}$ is then given by

$$\frac{\partial}{\partial \delta} \widehat{v}(f + \delta \tilde{f}) = \frac{1}{2} \frac{\sigma^2}{n f^2(x')} E \left[\operatorname{tr} \left[\widehat{u}''(\widehat{t}^*)^{-1} \cdot \left(\overline{w}_t(\widehat{t}^*, x') \overline{w}_t(\widehat{t}^*, x')^t \right) \right] \right] .$$

Remark: The proof of this theorem can be found in appendix A. The basic idea of the proof is as follows. By equation (64), the first order effect on the prior distribution of $\widehat{u}$ of a perturbation of f which adds a unit point mass at x' is the same as the effect of adding the random function

$$v(x) = \frac{\sigma}{\sqrt{n} f(x')} \cdot \overline{w}(x, x') \cdot \epsilon$$

to $\widehat{u}$, where ϵ is standard normal. Based on this equivalent representation, we can then use a Taylor expansion of $\widehat{u}(\widehat{t}^*)$, where the perturbation enters both through its effect on $\widehat{u}$ and through its effect on $\widehat{t}^*$.

With this result at hand, it is now straightforward to prove the following corollary.

Corollary 2 (The optimal experimental design)
Suppose that assumptions 2, 3, 4 and 7 hold. Then the optimal experimental design density $f^* = \operatorname{argmax}_f \widehat{v}(f)$ is characterized by the set of first order conditions

$$\frac{1}{f^{*2}(x)} E \left[\operatorname{tr} \left[\widehat{u}''(\widehat{t}^*)^{-1} \cdot \left(\overline{w}_t(\widehat{t}^*, x) \overline{w}_t(\widehat{t}^*, x)^t \right) \right] \right] = \kappa, \tag{65}$$

where equation (65) holds for all x. In this equation κ is a Lagrange multiplier on the constraint $\int f = 1$, and $\overline{w}_t$ is the derivative of the equivalent kernel $\overline{w}$ with respect to its first component.

Remark: Corollary 2 provides an exact finite sample characterization of the optimal design for the continuous-X setup. Any solution to these first order conditions has to account for the fact that the joint distribution of $\widehat{u}$ and $\widehat{t}^*$, as well as the equivalent kernel $\overline{w}$ all depend on the choice of f^*. It is possible to estimate these magnitudes using simulation methods, and solve the first order condition numerically. Alternatively, and that is what we will do next, we can consider the large-n case. Asymptotically the first order condition simplifies considerably. For large n, we can approximate $\overline{w}$ by the kernel $K_{h(nf^*)}$ discussed in section 5.1. We can furthermore approximate $\widehat{u}''$ by u'', and $\widehat{t}^*$ by t^*. This is done in the following proposition.

Proposition 2 (The asymptotically optimal experimental design)
Maintain the assumptions of theorem 2. Under the conditions stated below, the optimal design f^* satisfies the set of first order conditions

$$f^{*2}(t) \cdot h(nf^*(t))^{d_x+2} = l^2 \cdot \frac{f^{t^*}(t)}{\kappa} \operatorname{tr}\left[E[u''(t)^{-1}|t^* = t] \cdot \Lambda\right] \tag{66}$$

for all t, up to a remainder which converges to 0 uniformly on the support of f^{t^*}. In this expression $\Lambda = \int K'(x)K'(x)^t dx$, and $l = -t$ for the optimal insurance problem; $l = \lambda$ for the optimal choice of inputs problem. Sufficient conditions for this to hold are

1. *$\widehat{u}''(t) = u''(t) + o_p(1)$ uniformly in t,*
2. *u'' is positive definite and continuous,*
3. *$\widehat{f^{t^*}} = f^{t^*} + o(1)$ uniformly in t,*
4. *f^{t^*} is positive and continuous on the compact support of t,*
5. *$w_t(x,t) = \frac{1}{h^{d_x+1}} K'\left(\frac{x-t}{h}\right) + o(1)$ uniformly in x and t, where $h = h(nf^*(t)) = o(1)$,*
6. *u corresponds to the objective of one of the policy problems considered in section 3.*

The o_p notation here is to be understood as referring to an average both over the data generating process and over the prior.

Remarks:

- The conditions imposed by proposition 2 are "high-level." More primitive conditions can be derived by building on the asymptotic arguments of section 5 as well as on the recent literature on the asymptotics of Gaussian process regression.
- Instead of doing so, we propose the following approach:[20] (i) Construct an approximately optimal design using the condition of proposition 2. (ii) Plug this design into the exact optimality condition of corollary 2. Calculate the remainder term of the first order condition. (iii) If this remainder is large, numerically solve for the exact optimum. Use the approximately optimal design as a starting point for the optimization algorithm.
- The first order condition provided in proposition 2 is of relatively simple form. The expression
$$l^2 \cdot f^{t^*}(t) \cdot \operatorname{tr}\left[E[u''(t)^{-1}|t^* = t] \cdot \Lambda\right]$$
is independent of the design density and only needs to be calculated once, for instance using simulation methods. The expression $f^{*2}(t) \cdot h(nf^*)^{d_x+2}$ is a monotonic transformation of the design density f^*. We know that for the squared exponential covariance kernel $h(\rho) = \sqrt{2\pi^2 l^2 / \log(\rho(2\pi l^2)^{d_x/2}/\sigma^2)}$ (cf. Sollich and Williams, 2005); for spline regression with $\|K\|_C^2 = \int K''(x)^2 dx$ we have $h(\rho) = (\sigma^2/\rho)^{1/4}$, (cf. Silverman, 1984). It is thus fairly straightforward to solve for the approximately optimal design density. The following algorithm provides a simple way of doing so.

[20] This approach might be advisable whether or not suitable primitive conditions for the result of proposition 2 are fulfilled

1. Pick a number N of Monte Carlo replications, choosing N as large as practically feasible. Draw N times from the prior distribution for u, store t^*_j and $u''_j(t^*_j)$ for $j = 1, \ldots, N$.

2. Estimate $f^{t^*}(t) \cdot \operatorname{tr}\left[E[u''(t)^{-1}|t^* = t] \cdot \Lambda\right]$ by
$$\frac{1}{N}\sum_j L(t^*_j - t) \cdot \operatorname{tr}\left[(u''_j(t^*_j))^{-1} \cdot \Lambda\right]$$
for some positive kernel L integrating to 1, where L is of appropriate bandwidth. This yields an estimate of the right hand side of equation (66) up to a multiplicative constant. If K can be written as $K(x) = \prod_l \tilde{K}(x_l)$ for a kernel function $\tilde{K}$, then Λ can be replaced by the identity matrix.

3. Start with a guess for the multiplicative constant (Λ/κ in the one-dimensional case). Solve for f^*, by inverting the left hand side of equation (66), using this guess.

4. Check whether f^* integrates to 1. If it does not, adjust your guess of the constant, and iterate until you found a value that leads to a valid design density.

5. Calculate the cumulative distribution function F^* corresponding to f^*. Turn f^* into a discrete experimental design by choosing $X_i = F^{*-1}(i/n - 1/(2n))$ for $i = 1, \ldots, n$.

Algorithm 1: Approximately optimal experimental design

6.5 The optimal sample size

So far we have considered the problem of finding the optimal density f^*, given the sample size n. Assume now that the marginal cost of increasing the sample size is equal to $c(n)$, where c is measured in the appropriate units. Under this assumption, how should we choose n to maximize expected social welfare, net of experimental costs?

To answer this question we characterize the marginal return to an increase of n based on a variation of the arguments which lead to theorem 2. Under the assumption that either (i) the experimental design density f^* is chosen independently of sample size, or (ii) the experimental design f^* is chosen optimally given sample size, we can ignore the effect of sample size on the choice of design. In case (ii) this holds by an envelope condition argument. The value function for sample size n is defined as $\widehat{v}(n) := E[\max_t \widehat{u}(t)]$ in case (i) and as $\widehat{v}(n) := \max_f E[\max_t \widehat{u}(t)]$ in case (ii).

Theorem 3 (The optimal sample size) *Under the assumptions of theorem 2, and assuming that either (i) the experimental design density f^* is chosen independently of sample size, or (ii) the design density f^* is chosen optimally given sample size, the marginal effect of an increase in sample size n on social welfare is equal to*

$$\frac{\partial}{\partial n}\widehat{v}(n) = \frac{1}{2}\frac{\sigma^2}{n^2} E\left[\operatorname{tr}\left[\widehat{u}''(\widehat{t}^*)^{-1} \cdot \left(\int \frac{\overline{w}_t(\widehat{t}^*, x')^2}{f^*(x')} dx'\right)\right]\right]. \tag{67}$$

In case (ii), this marginal effect is also equal to

$$\frac{\partial}{\partial n}\widehat{\upsilon}(n) = \frac{1}{2}\frac{\sigma^2}{n} \cdot \kappa, \tag{68}$$

where κ is the Lagrange multiplier of corollary 2.

Under the assumption that the marginal cost of increasing sample size is equal to $c(n)$, the optimal sample size satisfies $\frac{\partial}{\partial n}\widehat{\upsilon}(n) = c(n)$. Theorem 3 thus allows to evaluate the first order condition for the optimal sample size.

7 Extensions

Up to this point, this paper has considered the conceptual baseline case of fully random treatment assignment in a sample drawn at random from the population of interest. In this section, we discuss two extensions of this baseline case. The first extension assumes that additional covariates are available, and that exogeneity of treatment only holds conditional on these covariates. The second extension assumes that an instrument is available which is exogeneous, while treatment itself is possibly endogeneous. For either extension, we characterize the posterior expectation of the average structural function m and of social welfare u; the rest of our analysis then applies immediately.

7.1 Conditional independence

Assumption 3 posited the availability of experimental data which are such that potential outcomes Y^x and experimental treatments X are independent. We now generalize to settings where we assume conditional independence instead of independence. We replace assumption 3 (experimental data) and assumption 4 (prior) by the following assumption.

Assumption 8 (Conditional independence and prior)
The available data satisfy

1. $(X_i, Y_i, W_i)_{i=1}^n$ *are i.i.d.*
2. $Y_i = g(X_i, \epsilon_i)$.
3. $X_i \perp \epsilon_i | W_i$.
4. $\epsilon_i \sim P_\epsilon$*, the distribution of ϵ in the target population.*

Let $k(x, w) = E[Y|X = x, W = w]$. The policy-maker has a prior which satisfies the following conditions.

1. $k \sim GP(\mu^k(.), C(., .))$*, where*
   - *(a)* $GP(\mu^k(.), C^k(., .))$ *is the law of a Gaussian process*
   - *(b) such that* $E[k(x, w)] = \mu^k(x, w)$*, and*
   - *(c)* $\mathrm{Cov}(k(x, w), k(x', w')) = C^k((x, w), (x', w'))$.
2. $P_W \sim DP(\alpha, P_W^0)$*, where*

   (a) *$DP(\alpha, P_W^0)$ is the law of a Dirichlet process*
   (b) *such that $E[P_W(.)] = P_W^0(.)$, and*
   (c) *α is the "precision" of the prior.*[21]

3. *The function m, the probability distribution P_W, and the probability distribution $P_{X|W}$ are mutually independent.*

Under this assumption, we can generalize our previous analysis to optimal policy choice controlling for covariates.

7.1.1 The posterior expectation $\widehat{m}$ of m

We first derive the posterior expectation of the average structural function m, which is defined as before, $m(x) = E[g(x, \epsilon)]$. Under the conditional independence assumption, we can write

$$m(x) = E[g(x, \epsilon)] = \int k(x, w) dP_W(w). \tag{69}$$

Under the assumption of independence of the prior components, we can take posterior expectations in this expression for m to get

$$\widehat{m}(x) = \int \widehat{k}(x, w) d\widehat{P}_W(w), \tag{70}$$

where $\widehat{k}$ and $\widehat{P}_W$ are the corresponding posterior expectations.

Maintaining assumption 8, we use the following notation for the prior moments of k given X and W,

$$\begin{aligned}
\mu_i^k &= E[k(X_i)|X, W] = \mu^k(X_i, W_i), \\
C_{i,j}^k &= \mathrm{Cov}(k(X_i, W_i, k(X_j, W_j)|X, W) = C((X_i, W_i), (X_j, W_j)), \text{ and} \\
C_i^k(x, w) &= \mathrm{Cov}(k(x, w), k(X_i, W_i)|X, W) = C^k((x, w), (X_i, W_i)).
\end{aligned}$$

Let furthermore μ^k, C^k, and $C^k(x, w)$ denote the vectors and matrix collecting these terms for $i, j = 1, \ldots, n$. Assuming again that the posterior expectation of k is the posterior best linear predictor of k, the posterior mean of $k(x, w)$ is given by

$$\begin{aligned}
\widehat{k}(x, w) &= E[k(x, w)|X, Y, W] \\
&= E[k(x, w)|X, W] + \mathrm{Cov}(k(x, w), Y|X, W) \cdot \mathrm{Var}(Y|X, W)^{-1} \cdot (Y - E[Y|X, W]) \\
&= \mu^k(x, w) + C^k(x, w) \cdot \left[C^k + \sigma^2 I\right]^{-1} \cdot (Y - \mu^k).
\end{aligned} \tag{71}$$

Invoking the assumption that the prior for P_W is Dirichlet implies

$$d\widehat{P}_W(w) = \frac{\alpha}{\alpha + n} dP_W^0 + \frac{n}{\alpha + n} dP_W^n, \tag{72}$$

where P_W^n is the empirical distribution of W in the sample.

[21] For a discussion of Dirichlet priors, see for instance Ghosh and Ramamoorthi (2003).

Combination of equations (70), (71), and (72) yields

$$\begin{aligned}\widehat{m}(x) &= \int \widehat{k}(x,w) d\widehat{P}_W(w) \\ &= \int \left[\mu^k(x,w) + C^k(x,w) \cdot \left[C^k + \sigma^2 I\right]^{-1} \cdot (Y - \mu^k)\right] d\left[\frac{\alpha}{\alpha+n} P_W^0 + \frac{n}{\alpha+n} P_W^n\right] \\ &= \widehat{\mu}(x) + \widehat{C}(x) \cdot \left[C^k + \sigma^2 I\right]^{-1} \cdot (Y - \mu^k), \end{aligned} \tag{73}$$

where

$$\begin{aligned}\widehat{\mu}(x) :=& \int \mu^k(x,w) d\widehat{P}_W(w) \\ =& \frac{\alpha}{\alpha+n} \int \mu^k(x,w) dP_W^0(w) + \frac{1}{\alpha+n} \sum_i \mu^k(x, W_i),\end{aligned}$$

and

$$\begin{aligned}\widehat{C}(x) :=& \int C^k(x,w) d\widehat{P}_W(w) \\ =& \frac{\alpha}{\alpha+n} \int C^k(x,w) dP_W^0(w) + \frac{1}{\alpha+n} \sum_i C^k(x, W_i).\end{aligned}$$

7.1.2 The posterior expectations of u and u'

The posterior expectation of m again maps linearly into the posterior expectation of u by the linearity of $u = L \cdot m + u_0$. We get, in particular,

$$\begin{aligned}\widehat{u}(t) &= (L \cdot \widehat{m})(t) + u_0(t) \\ &= \widehat{\nu}(t) + \widehat{D}(t) \cdot \left[C^k + \sigma^2 I\right]^{-1} \cdot (Y - \mu^k), \end{aligned} \tag{74}$$

where

$$\begin{aligned}\widehat{\nu} :=& (L \cdot \widehat{\mu})(t) + u_0 \\ =& \int L_x \mu^k(x,w) d\widehat{P}_W(w) + u_0,\end{aligned}$$

and

$$\widehat{D}(t) := \int L_x C^k(x,w) d\widehat{P}_W(w).$$

As before, the optimal t satisfies

$$\widehat{t^*} = \widehat{t^*}(X,Y) \in \underset{t}{\operatorname{argmax}}\ \widehat{u}(t).$$

Assuming again differentiability of $\widehat{u}$, the first order condition for $\widehat{t^*}$ is $\frac{\partial}{\partial t}\widehat{u}(\widehat{t^*}) = 0$. We can write this first order condition more explicitly as

$$\begin{aligned}\frac{\partial}{\partial t}\widehat{u}(t) &= \frac{\partial}{\partial t}(L \cdot \widehat{m})(t) + \frac{\partial}{\partial t} u_0(t) \\ &= \widehat{\nu'}(t) + \widehat{B}(t) \cdot \left[C^k + \sigma^2 I\right]^{-1} \cdot (Y - \mu^k), \end{aligned} \tag{75}$$

where

$$\begin{aligned}\widehat{\nu'} :=& \frac{\partial}{\partial t}\left(L \cdot \widehat{\mu}\right)(t) + \frac{\partial}{\partial t}u_0 \\ =& \int \frac{\partial}{\partial t} L_x \mu^k(x,w) d\widehat{P}_W(w) + \frac{\partial}{\partial t}u_0,\end{aligned}$$

and

$$\widehat{B}(t) := \int \frac{\partial}{\partial t} L_x C^k(x,w) d\widehat{P}_W(w).$$

7.1.3 Multiplicative covariance kernels

Many covariance kernels C^k are multiplicative in the sense that they can be written in the form

$$C^k((x,w),(x',w')) = C(x,x') \cdot C^w(w,w'). \tag{76}$$

This is the case in particular for the squared exponential covariance kernel. If the covariance kernel has such a structure, then we get the following simplifications:

$$\begin{aligned}\widehat{C}(x)_i &= c_i \cdot C(x, X_i) \\ \widehat{D}(t)_i &= c_i \cdot D(t, X_i) \\ \widehat{B}(t)_i &= c_i \cdot B(t, X_i),\end{aligned}$$

where

$$c_i := \int C^w(w, W_i) d\widehat{P}_W(w) \tag{77}$$

and D and B are as in definition 1. The analysis in the conditional independence setting is then completely analogous to the analysis in the experimental setting, except that "correction factors" c_i are applied to all covariances.

7.2 Instruments

The main discussion in this paper assumed that we have experimental data; the last section assumed we have data such that treatment X is exogeneous conditional on a set of covariates. In this section we assume instead that an exogenous instrument Z is available.

Nonparametric identification with instrumental variables is the subject of a large literature in econometrics. Two popular approaches for continuous treatments are (i) conditional moment restrictions, as in Newey and Powell (2003), and (ii) control functions, as in Imbens and Newey (2009). Both of these rely on restrictions on the dimensionality of unobserved heterogeneity. In this section we follow the identification approach of Kasy (2013a), which does not rely on such restrictions. In the latter paper, triangular systems of the following form are considered:

$$\begin{aligned}Y &= g(X, \epsilon) \\ X &= h(Z, \eta)\end{aligned} \tag{78}$$

where X, Y, Z are random variables taking their values in $\mathbb{R}$, the unobservables ϵ, η have their support in an arbitrary measurable space of unrestricted dimensionality, and

$$Z \perp (\epsilon, \eta). \tag{79}$$

It is shown in Kasy (2013a) that under these assumptions, if the first stage relationship $h(z, v)$ is strictly increasing in z for all v and if some continuity conditions are satisfied, then

$$m(x) = E[g(x, \epsilon)] = \int E[Y_i|X_i = x, Z_i = z] dF_{Z^x}(z), \tag{80}$$

where

$$F_{Z^x}(z) = P(X_i \geq x|Z_i = z). \tag{81}$$

Assumption 9 (Instrumental variables and prior)
The available data satisfy

1. $(X_i, Y_i, Z_i)_{i=1}^n$ *are i.i.d.*
2. $m(x) = \int E[Y|X = x, Z = z] dF_{Z^x}(z)$*, where* $F_{Z^x}(z) = P(X \geq x|Z = z)$*.*

Let $k(x, z) = E[Y|X = x, Z = z]$*. The policy-maker has a prior which satisfies the following conditions.*

1. $k \sim GP(\mu^k(.), C(., .))$*, where*
   (a) $GP(\mu^k(.), C^k(., .))$ *is the law of a Gaussian process*
   (b) *such that* $E[k(x, z)] = \mu^k(x, z)$*, and*
   (c) $\mathrm{Cov}(k(x, z), k(x', z')) = C^k((x, z), (x', z'))$*.*
2. $P_{X|Z} \sim \mathscr{L}$*, where* $\mathscr{L}$ *is some general prior for the conditional distribution of* X *given* Z*.*
3. *The function* m*, the probability distribution* P_Z*, and the probability distribution* $P_{X|Z}$ *are mutually independent.*

Under this assumption, we can proceed to a large extent as in section 7.1. As before

$$\widehat{k}(x, z) = E[k(x, z)|X, Y, Z] = \mu^k(x, z) + C^k(x, z) \cdot \left[C^k + \sigma^2 I\right]^{-1} \cdot (Y - \mu^k).$$

Furthermore,

$$\begin{aligned}\widehat{m}(z) &= \int \widehat{k}(x, z) d\widehat{F_{Z^x}}(z) \\ &= \widehat{\mu}(x) + \widehat{C}(x) \cdot \left[C^k + \sigma^2 I\right]^{-1} \cdot (Y - \mu^k),\end{aligned} \tag{82}$$

where

$$\widehat{\mu}(x) = \int \mu^k(x, w) d\widehat{F_{Z^x}}(z)$$

and

$$\widehat{C}(x) = \int C^k(x, w) d\widehat{F_{Z^x}}(z).$$

From here on, we can proceed as before, to find the posterior expectations of $\widehat{u}$ and $\widehat{u'}$, the optimal policy choice $\widehat{t}^*$, etc.

In this discussion we have not explicitly characterized the posterior expectation of $P(X_i \geq x|Z_i = z)$, that is, $\widehat{F_{Z^x}}$. Assumption 9 left the prior distribution for $P_{X|Z}$ unspecified. The reason is that there seems to be no nonparametric prior for conditional distributions which allows for an easy analytic characterization of the posterior mean of $P_{X|Z}$. This contrasts with the elegant and simple posterior means for conditional expectations such as k, and of unconditional distributions such as P_W as in section 7.1. That said, there are a number of nonparametric priors for conditional distributions which have been proposed in the literature. One class of approaches considers the conditional distributions corresponding to continuous joint distributions. For the latter, Dirichlet process mixtures are an appropriate class of nonparametric priors; see for instance (Ghosh and Ramamoorthi, 2003, section 3.6). Another class of approaches assumes that $P(X_i \geq x|Z_i = z)$ is given by the value of some link function (such as the logistic cdf), applied to a latent function l, where the prior for the latter is given by a Gaussian process; see for instance (Williams and Rasmussen, 2006, section 3.3). For any of these approaches, we can obtain draws from the distribution F_{Z^x} using Markov Chain Monte Carlo or related methods. Such draws then allow us to evaluate $\widehat{m}$, $\widehat{u}$, and $\widehat{u'}$ in a straightforward manner.

8 Application - The RAND health insurance experiment

As an empirical application for the results derived in this paper we consider is the choice of coinsurance rates in health insurance. A policy maker who is in charge of a public health insurance program has to choose how much to redistribute from healthy contributors to those in need of health care. Such redistribution is justified on grounds of risk aversion (insurance) and on grounds of equity (redistribution). Lowering coinsurance rates redistributes more. It also affects public expenditures mechanically and through the behavioral response of possibly increased health care spending. We use the data of the RAND health insurance experiment in order to estimate this behavioral response and use the estimated relationship to determine the optimal coinsurance rate under an assumption about the relative welfare weight assigned to those in need of health expenditures.

Background and data

The following discussion is based on the review of the RAND experiment provided by Aron-Dine et al. (2013). The RAND experiment, which took place between 1974 and 1981, provided health insurance to more than 5,800 individuals from about 2,000 households in six different locations across the United States. Families participating in the experiment were assigned to plans with one of six coinsurance rates, where the coinsurance rate is the share of medical expenditures paid by the enrollee. Four of the six plans simply set different overall coinsurance rates of 95, 50, 25, or 0 percent (free care). The other two plans were somewhat more complicated, with higher coinsurance rates for dental and outpatient mental health services, or for outpatient services in general. For the sake of simplicity of our discussion, we neglect data from the last two plans and focus our analysis on the first four plans. The probability of assignment to each of these was .32 for the free care plan, .07 for the 50% coinsurance plan, .11 for 25% coinsurance plan, and .19 for the 95% coinsurance plan.

Families were additionally randomly assigned, within each of the six plans, to different out-of-pocket maximums, referred to as the Maximum Dollar Expenditure. The possible Maximum Dollar Expenditure limits were 5, 10, or 15 percent of family income, up to a maximum of $750 or $1,000 (roughly $3,000 or $4,000 in 2011 dollars). We pool data across Maximum Dollar Expenditure, and only consider the effect of coinsurance rates on expenditures.

Results

As a first step, we replicate some of the results of Aron-Dine et al. (2013). We estimate predicted expenditures, using specifications corresponding to those used by Aron-Dine et al. (2013) for the estimation of the treatment effects reported in rows 2 and 3 in each of the panels of their Table 3. The chosen regression specification controls for month × site fixed effects and year fixed effects (this is necessary, since treatment was only conditionally random). This specification additionally corrects for under-reporting of spending, by proportionally scaling up spending for outpatient services based on estimated rates of under-reporting (as discussed in Aron-Dine et al. (2013), this adjustment has only a minor impact on results). Table 1 reports predicted values for the share of families with any spending, and for the average amount of spending, within each of the treatment categories. Column 3 and 4 of this table control additionally for a rich set of predetermined covariates, to correct for imbalance in the assignment (this correction again has only a minor effect). As can be seen from this table, spending is essentially unaffected by the coinsurance rate in the range from 95% coinsurance to 25% coinsurance. Only when approaching the free-care treatment does there appear to be an effect of the coinsurance rate on spending.

We next apply the methods proposed in this paper to these data. We consider the model of optimal insurance, as discussed in section 3.1. Consider first estimation of m, the response function which gives expected spending as a function of the subsidy rate t. The subsidy rate t equals 1 minus the coinsurance rate. We use a prior for m which is a Gaussian process prior with squared exponential covariance kernel (cf. appendix B), and which is made uninformative about the level and slope of m. We use the same controls as in column 4 of table 1, so that our estimate $\widehat{m}$ is effectively a smooth interpolation of the estimates in table 1, column 4. Figure 1 shows our estimate $\widehat{m}$, as well as the estimated slope of m, $\widehat{m}'$. As to be expected based on the predicted values of table 1, $\widehat{m}$ is flat over most of it's support and curves upward toward the right, as t approaches 1, corresponding to the free care plan.

Assuming (and this is a crucial parameter), that the marginal welfare weight on income for those needing medical services relative to healthy contributors is given by λ, we can write expected social welfare as a function of the subsidy rate, $\widehat{u}(t) = \lambda \int_0^t \widehat{m}(x)dx - t \cdot \widehat{m}(t)$, and the expected marginal social return to an increase of t as $\widehat{u}'(t) = (\lambda - 1) \cdot \widehat{m}(t) - t \cdot \widehat{m}'(t)$. Figure 2 plots our estimate of social welfare $\widehat{u}$ and its derivative $\widehat{u}'$, assuming that $\lambda = 1.5$. The optimal policy choice $\widehat{t^*}$ solves the first order condition $\widehat{u}'(\widehat{t^*}) = 0$. We find an optimal policy choice of $\widehat{t^*} = 0.82$, corresponding to a coinsurance rate of 18%. As the objective function is fairly flat around this point, the free care plan performs almost as well in terms

Table 1: Expected spending for different coinsurance rates

	(1) Share with any	(2) Spending in $	(3) Share with any	(4) Spending in $
Free Care	0.931	2166.1	0.932	2173.9
	(0.006)	(78.76)	(0.006)	(72.06)
25% Coinsurance	0.853	1535.9	0.852	1580.1
	(0.013)	(130.5)	(0.012)	(115.2)
50% Coinsurance	0.832	1590.7	0.826	1634.1
	(0.018)	(273.7)	(0.016)	(279.6)
95% Coinsurance	0.808	1691.6	0.810	1639.2
	(0.011)	(95.40)	(0.009)	(88.48)
N	14777	14777	14777	14777

Notes: This table shows OLS estimates of average health care expenditures for the different treatment arms. Columns (1) and (2) control for month × site fixed effects and year fixed effects, columns (3) and (4) control additionally for a large set of further predetermined covariates. All regressions are pooled across Maximum Dollar Expenditure values.

of social welfare.

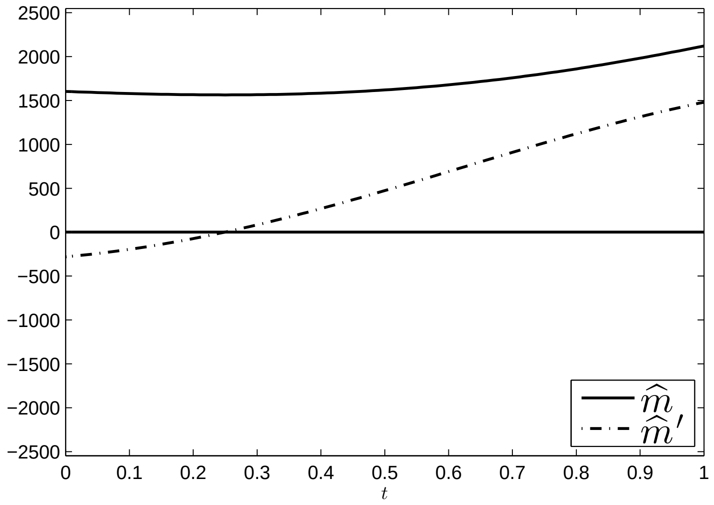
Based on the considerations of section 5, we can also perform frequentist inference for the optimal choice of t. Figure 3 plots a pointwise frequentist 95% confidence band for u'. The intersection of this confidence band with the horizontal axis yields an asymptotically valid 95% confidence interval for the optimal policy choice, which in this case ranges from a subsidy rate of about 70% to a subsidy rate of 100%, that is free care.

To check the robustness of our results, we replicate the analysis without controlling for predetermined covariates. This corresponds to the estimates in column 2 of table 1. Figure 4 replicates figures 1, 2, and 3 using this alternative specification. As can be seen, this results in a slightly higher estimated slope of m toward the right end of its support, a social welfare function which peaks slightly further left, and a downward shifted estimate of the slope of social welfare. The corresponding optimal policy choice equals $\widehat{t^*} = .75$, corresponding to a copay of 25%. The 95% confidence interval for the optimal t ranges from .64 to .98. Recall, however, that it is more plausible that the specification using predetermined controls yields the right estimates of m and u, and that this specification suggests higher insurance / lower copay rates.

Comparison to the conventional "sufficient statistic" approach

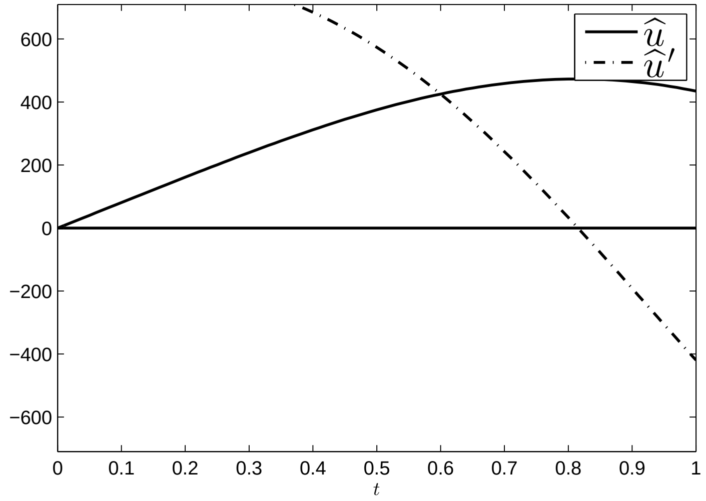
It is useful to compare these results – in particular the optimal choice $\widehat{t^*} = 0.82$ – to the recommended coinsurance rate obtained using a more conventional "sufficient statistic" approach as in Chetty (2009). Under the assumptions of section 3.1, we can rewrite the

Figure 1: Estimated dependence of health expenditures on copay using the RAND health insurance experiment



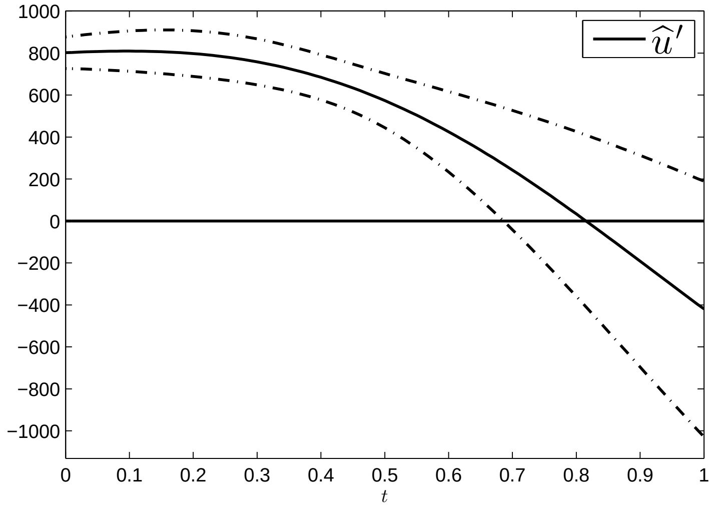
Notes: This graph shows our estimate of m, which describes expected health care expenditures as a function of the subsidy rate t and our estimate of m'.

Figure 2: Estimated social welfare function using the RAND health insurance experiment



Notes: This graph shows our estimate of u, which describes social welfare as a function of t assuming $\lambda = 1.5$, and our estimate of u'.

Figure 3: 95% confidence band for the slope of social welfare



Notes: This graph shows a pointwise frequentist .95 confidence band for u'.

marginal social return to an increase of t as

$$\begin{aligned} u'(t) &= (\lambda - 1) \cdot m(t) - t \cdot m'(t) \\ &= m(t) \cdot [(\lambda - 1) - t \cdot m'(t)/m(t)] \\ &= m(t) \cdot \left[(\lambda - 1) - \eta \cdot \frac{t}{1-t}\right], \end{aligned}$$

where η is the elasticity of health-care expenditures m with respect to copay $1-t$,

$$\eta := -\frac{\partial m(t)}{\partial(1-t)} \cdot \frac{1-t}{m(t)}. \tag{83}$$

Note that η is a function of t unless m is log-linear. Solving the first order condition $u'(t^*) = 0$ yields

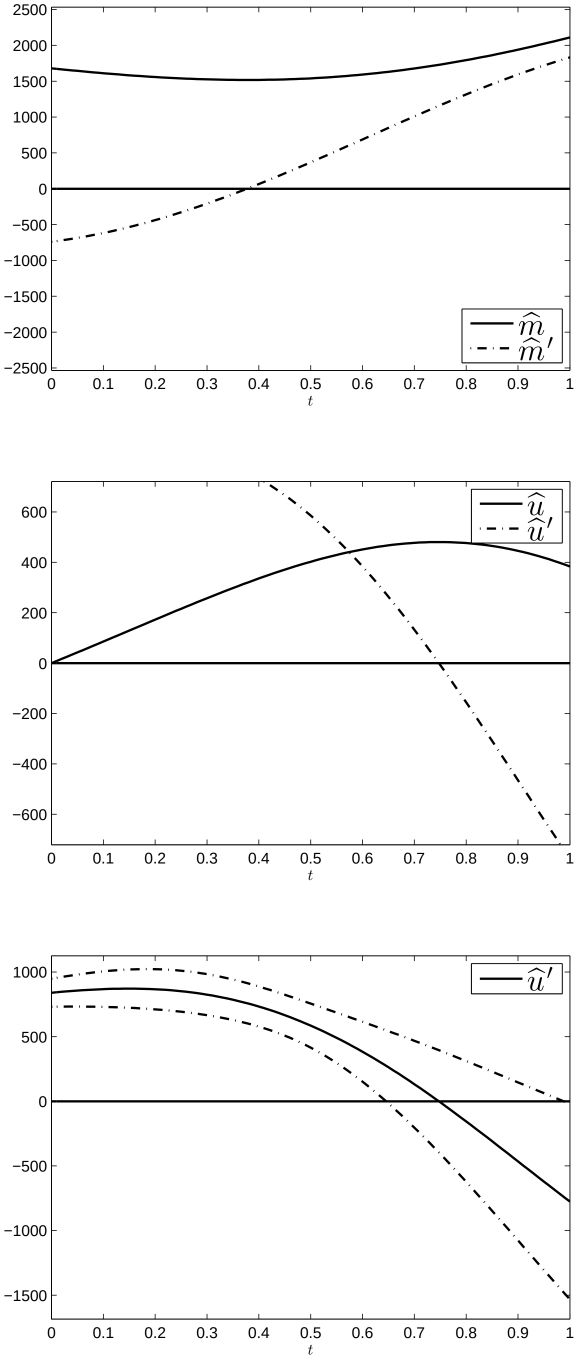
$$t^* = \frac{1}{1 + \eta/(\lambda - 1)}. \tag{84}$$

The approach suggested by Chetty (2009) (and many other contributions to the literature) is to obtain an estimate $\widehat{\eta}$ of the elasticity η, and to plug it into this formula to obtain a recommended copay of

$$\tilde{t}^* = \frac{1}{1 + \widehat{\eta}/(\lambda - 1)}.$$

Assuming the same relative marginal value λ of income for those requiring health-care expenditures as before, $\lambda = 1.5$, yields $\tilde{t}^* = 1/(1 + 2 \cdot \widehat{\eta})$.

Figure 4: Replicating estimates without use of controls



Notes: These graphs replicate figures 1 through 3, without using controls.

It remains to obtain an estimate of η. The most natural approach would fit a linear regression of $\log(Y)$ on $\log(1-t)$ as well as the appropriate controls, and take the negative of the coefficient on $\log(1-t)$ as the estimate of η. This turns out not to be feasible in the present context, given that $t = 1$ for an important part of the experimental sample. For these observations, the log-linear specification clearly makes no sense.

Various estimates for η based on the RAND experiment have been proposed in the literature, as discussed by Aron-Dine et al. (2013). The most famous estimate, constructed by the RAND investigators themselves, is given by $\widehat{\eta} = 0.2$. This estimate was constructed in a fairly complicated manner, based on so-called "arc-elasticities" for pairwise comparisons and averaging across these comparisons. Plugging this estimate into the sufficient-statistic formula yields $\tilde{t}^* = 1/1.4 \approx 0.7$, that is a suggested copay of approximately 30%. This is 12 percentage points or 50 percent higher than the optimal copay of 18% according to our calculations.

Table 4 of Aron-Dine et al. (2013) presents various alternative estimates $\widehat{\eta}$, based on the more standard definition of an elasticity underlying our derivation of the sufficient statistic formula. Their estimates, omitting the free-care plan from calculations, are slightly larger than 0.5. Plugging this into the formula for $\tilde{t}^*$ yields $\tilde{t}^* \approx 1/2 = .5$ – that is a suggested copay of approximately 50%. This is 32 percentage points or almost 180 percent higher than the optimal copay of 18% according to our calculations.

Where do these large differences come from? Note that we have (i) assumed the same model and the same social welfare function, (ii) have imposed the same value of $\lambda = 1.5$, and (iii) used the same data to estimate the behavioral relationship m. The difference stems from the way m, and thus u, are estimated. First, and most importantly, we have not imposed the log-linear functional form implicit in the constant-elasticity specification. In fact, η seems to vary significantly across t. Second, we estimate m in levels rather than in logs. This gives a larger weight to infrequent but high occurrences of large expenditures Y. Estimation in levels is appropriate according to the decision-theoretic setup. Third, we estimate m in a Bayesian way. This imposes some "shrinkage." Note however that we have used a prior which is uninformative about the level and slope of m, so that this seems only a minor driver of the difference in estimates.

9 Conclusion

In this paper, a general framework is proposed for the use of (quasi-) experimental data when choosing policies such as tax rates or inputs into a production process. This framework is based on the following assumptions: (i) There is random variation of the policy choice variables in the available data (possibly conditional on covariates), or there is a valid instrument for the policy choice. (ii) The policy objective function can be written as a transformation of the relevant technological or behavioral relationships by some affine operator. (iii) The prior distribution of these relationships is given by a Gaussian process. Under these assumptions we provide explicit expressions, as well as characterizations, of optimal policy choices and of optimal experimental designs. This setting leads to policy choice procedures which are easy to implement in practice. The resulting policy choices may differ strongly from those suggested by sufficient statistic approaches (which implicitly rely on stronger assumptions), as demonstrated by our empirical analysis of data from the RAND health insurance experiment.

Let us conclude by briefly discussing some of the larger context of the approach proposed in this paper. Arguably there are several conceptually distinct roles of econometrics. The first of these is the purely statistical task of forecasting or prediction, answering questions of the form "What will be?". This task needs no notion of causality. The second is the task of testing theories, answering questions of the form "Is statement A correct?". This task usually involves some notion of causality or structural invariance. The third is the task of informing policy choice, answering questions of the form "What should we do?". This task generally involves notions of causality or counterfactuals, as well as a normative evaluation of possible counterfactual outcomes. It is to this last task that the present paper aims to contribute.

In contrast to much of the literature in econometrics, this paper takes a Bayesian approach to solving a decision problem. The central motivation for taking such an approach is the complete class theorem, which loosely speaking says the following: Any decision procedure which is admissible (not dominated) is a Bayesian procedure. Put differently, we can not avoid trading off risk across different states of the world – Bayesian procedures just make this tradeoff explicit by assigning weights ("prior probabilities") to these states. One of the main objections to Bayesian procedures is that their results are not replicable because they depend on the choice of weights (prior). While a valid and important concern, the complete class theorem implies that non-Bayesian procedures are either (i) inadmissible, or (ii) in fact Bayesian procedures based on a conventional choice of weights, or (iii) Bayesian procedures based on a "subjective" but implicit choice of weights. While there is much to be said for procedures in category (ii), explicitly Bayesian solutions of decision problems seem preferable to those in categories (i) and (iii). Category (iii) includes any nonparametric procedure involving the choice of tuning or bandwidth parameters. Indeed, as the equivalence between Gaussian process regression and spline regression, and the asymptotic equivalence between Gaussian process regression and Kernel regression shows, there is a direct correspondence between these tuning parameters and prior assumptions about smoothness.

A Proofs

Proof of Lemma 1:
The claims are immediate once we can show the required exchangeability of expectation and linear operator L, i.e.,

$$E[(L\cdot m)(t)] = (L\cdot E[m])(t),$$

$$E[(L\cdot m)(t)\cdot m(x)] = L_{x'}E[m(x')\cdot m(x)](t).$$

etc. This exchangeability follows from Lemma 2.1 in van der Vaart and van Zanten (2008b), which implies the existence of the required Pettis integrals. □

Proof of theorem 1:

- For the policy problems of section 3, $u = \lambda \int_0^t m(x)dx - t\cdot m(t)$ or $u = \lambda m(t) - p't$.
- Consistency of $\widehat{m}$ in $\mathscr{C}^2$ then immediately implies consistency of $\widehat{u}$ in $\mathscr{C}^2$ on the compact support of t.
- This, combined with strict convexity of u at its maximum, in turn implies consistency of $\widehat{t^*}$.
- Given smoothness of u, and using the first order conditions for $\widehat{t^*}$, a Taylor expansion yields
$$0 = \widehat{u}'(\widehat{t^*}) = \widehat{u}'(t^*) + \widehat{u}''(\tilde{t}^*)\cdot\left(\widehat{t^*} - t^*\right),$$
where $\tilde{t}^* \in [\widehat{t^*}, t^*]$, and thus
$$\widehat{t^*} - t^* = -u''(t^*)^{-1}\cdot\widehat{u}'(t^*) + \Delta_t,$$
where
$$\Delta_t = \widehat{u}'(t^*)\cdot\left(\widehat{u}''(\tilde{t}^*)^{-1} - u''(t^*)^{-1}\right)$$
- $\Delta_t = o_p(\widehat{u}'(t^*))$ by consistency of $\widehat{t^*}$ and $\mathscr{C}^2$ consistency of $\widehat{u}$.
- For the policy problems of section 3, $\widehat{u}'(t) = (\lambda-1)\cdot\widehat{m}(t) - t\cdot\widehat{m}'(t)$ or $\widehat{u}'(t) = \lambda\cdot\widehat{m}'(t) - p$. By assumption 6, $(\widehat{m}(t^*) - m(t^*)) = o_p(\widehat{m}'(t^*) - m'(t^*))$. This implies $\widehat{u}'(t^*) - u'(t^*) = l\cdot(\widehat{m}'(t^*) - m'(t^*))\cdot(1+o_p(1))$.
- Asymptotic normality of $\widehat{m}(t^*)$ therefore implies
$$V_{n,u'}^{-1/2}\cdot(\widehat{u}'(t^*) - u'(t^*) - B_{n,u'}) \to_d N(0,I),$$
where
$$V_{n,u'} = l^2\cdot V_{n,m'}$$
$$B_{n,u'} = l\cdot B_{n,m'}.$$
for $l = t$ or $l = \lambda$, depending on the application.

- Collecting these results yields the asymptotic normality of

$$\widehat{t^*} - t^* = -l \cdot u''(t^*)^{-1} \cdot (\widehat{m}'(t^*) - m'(t^*)) \cdot (1 + o_p(1))$$

  and the expressions for V_n and B_n.

□

Proof of Corollary 1:
By the first order condition for t^* $u'(t^*) = 0$, and thus, using the fact that $u \in \mathscr{C}^2$ and the consistency of $\widehat{t^*}$ for t^*,

$$u(t^*) - u(\widehat{t^*}) = -\frac{1}{2}(\widehat{t^*} - t^*)^t \cdot u''(t^*) \cdot (\widehat{t^*} - t^*) + o_p(\|\widehat{t^*} - t^*\|^2).$$

The claims follow from this equation, the continuous mapping theorem and the δ-method.
□

Proof of proposition 1:

- Consistency of $\widehat{t^*}$ and $\mathscr{C}^2$ consistency of $\widehat{u}$ implies consistency of $\widehat{u}(\widehat{t^*})$ for $u(t^*)$.
- By the form of the estimator $\widehat{m}$ (assumption 5) and the independence of observations, we can choose

$$V_{m'} = \text{Var}(\widehat{m}'(t^*)|X, \theta) = \sum_i w_x(x, X_i)^2 \sigma^2(X_i).$$

  The subscript x denotes the partial derivative $w_x(x, X_i) := \partial/\partial x \ w(x, X_i)$. We need to show $\widehat{V}_{m'}/V_{m'} \to^p 1$.
- Let $\epsilon_i = Y_i - m(X_i)$, $\widehat{\epsilon}_i = Y_i - \widehat{m}(X_i)$, and $w_{t,i} = w_t(\widehat{t^*}, X_i)$. We can decompose

$$\frac{\widehat{V}_{m'}}{V_{m'}} = 1 + \frac{\sum w_{t,i}^2[\epsilon_i^2 - \sigma_i^2]}{\sum w_{t,i}^2 \sigma_i^2} + \frac{\sum w_{t,i}^2[\widehat{\epsilon}_i^2 - \epsilon_i^2]}{\sum w_{t,i}^2 \sigma_i^2} = 1 + R_1 + R_2.$$

  We have to show $R_1 \to^p 0$ and $R_2 \to^p 0$.
- Note that $E[R_1|X] = 0$ and, under the assumption of bounded 2nd and 4th moments,

$$\text{Var}(R_1|X) = \frac{\sum w_{t,i}^4 \, \text{Var}(\epsilon_i^2|X_i)}{\left(\sum w_{t,i}^2 \sigma_i^2\right)^2} \leq M^2 \cdot \sum \left(\frac{w_{t,i}^2}{\sum w_{t,i}^2} \right)^2 \leq M^2 \cdot n \cdot W_n^2$$

  and thus $\text{Var}(R_1|X) \to 0$ under the assumption on the large n behavior of W_n, which implies $R_1 \to^p 0$.

- For the second term, note that $\widehat{\epsilon}_i = \epsilon_i + m_i - \widehat{m}_i$, therefore $\widehat{\epsilon}_i^2 - \epsilon_i^2 = (m_i - \widehat{m}_i)^2 + 2\epsilon_i(m_i - \widehat{m}_i)$, and thus

$$
\begin{aligned}
R_2 &= \frac{\sum w_{t,i}^2 (f_i - \widehat{f}_i)^2}{\sum w_{t,i}^2 \sigma_i^2} + \frac{\sum w_{t,i}^2 2\epsilon_i (m_i - \widehat{m}_i)}{\sum w_{t,i}^2 \sigma_i^2} \\
&\leq M \cdot n \cdot W_n \cdot \left[\|\widehat{m} - f\|_n^2 + 2\|\epsilon\|_n \cdot \|\widehat{m} - m\|_n \right].
\end{aligned}
$$

  It is easy to see that $\|\epsilon\|_n = O_p(1)$. We get that $R_2 = O_p(n \cdot W_n \cdot \|\widehat{m} - m\|_n)$, and thus $R_2 \to 0$ under our assumption on the rate of convergence of $\widehat{m}$.

□

Proof of Lemma 2:
This follows from the equivalent kernel representation and the asymptotics of kernel regression estimators; (see for instance Li and Racine, 2011, chapter 2). By consistency, smoothness, and a change of variables,

$$
\begin{aligned}
\mathrm{Var}(\widehat{m}'(x)|\theta) &= \frac{n}{\rho^2} \int \left[\frac{1}{h(\rho)^{d_x+1}} K'\left(\frac{\tilde{x} - x}{h(\rho)} \right) \right]^2 \cdot \sigma^2(\tilde{x}) \cdot f(\tilde{x}) d\tilde{x} \\
&= (1 + o(1)) \cdot \frac{\sigma^2(x)}{\rho} \int \left[\frac{1}{h(\rho)^{d_x+1}} K'\left(\frac{\tilde{x} - x}{h(\rho)} \right) \right]^2 d\tilde{x} \\
&= (1 + o(1)) \cdot \frac{\sigma^2(x)}{\rho \cdot h(\rho)^{d_x+2}} \int K'(\tilde{x}) \cdot K'(\tilde{x})^t d\tilde{x}
\end{aligned}
$$

and, by partial integration and change of variables,

$$
\begin{aligned}
E[\widehat{m}'(x)|\theta] - m'(x) &= -\frac{n}{\rho} \int \frac{1}{h(\rho)^{d_x+1}} K'\left(\frac{\tilde{x} - x}{h(\rho)} \right) \cdot [m(\tilde{x}) - m(x)] \cdot f(\tilde{x}) d\tilde{x} \\
&= (1 + o(1)) \cdot \int \frac{1}{h(\rho)^{d_x}} K\left(\frac{\tilde{x} - x}{h(\rho)} \right) \cdot m'(\tilde{x}) d\tilde{x} \\
&= (1 + o(1)) \cdot \int K(\tilde{x}) \cdot m'(x + h \cdot \tilde{x}) d\tilde{x}.
\end{aligned}
$$

Asymptotic normality follows from a central limit theorem for triangular arrays. □

Proof of theorem 2: We are interested perturbation of the design density f by a point mass $\tilde{f}$ at the point x'. We can approximate $\tilde{f}$ by a sequence of continuous densities $\tilde{f}_j(x) = 1/j\tilde{K}(x'/j)$ for some continuous kernel function $\tilde{K}$ integrating to 1. Equation (64) gives the effect of perturbations by $\tilde{f}_j$ on V. Taking the limit with respect to j, we get that the marginal effect of perturbations by $\tilde{f}$ is equal to

$$
(V_\delta g)(x) = \frac{\sigma^2}{nf(x')^2} \int \overline{w}(x, x') \overline{w}(x'', x') g(x'') dx''. \tag{85}
$$

This expression for V_δ is equal to the covariance kernel of the random function

$$
v(x) = \frac{\sigma}{\sqrt{n} f(x')} \cdot \overline{w}(x, x') \cdot \epsilon, \tag{86}
$$

where $\epsilon \sim N(0,1)$. Modifying the experimental design density f by adding a point mass of measure δ at x' is thus equivalent, for small δ, to adding the random function $\sqrt{\delta} \cdot v$ to $\widehat{u}$, where $v \perp \widehat{u}$. Let us therefore consider the random function

$$\widehat{u}(\delta) := \widehat{u} + \sqrt{\delta} \cdot v, \tag{87}$$

the optimal policy $\widehat{t}^*(\delta) = \operatorname{argmax}_t \widehat{u}(\delta, t)$ and the corresponding expected social welfare $\widehat{v}(\delta) = E[\max \widehat{u}(\delta)]$. Taylor expansion of these objects, using that $\widehat{u}'(\widehat{t}^*) = 0$, yields

$$\begin{aligned}\widehat{u}(\delta, \widehat{t}^*(\delta)) = \widehat{u}(t^*) &+ \frac{1}{2}(\widehat{t}^*(\delta) - \widehat{t}^*)^t \cdot \widehat{u}''(\widehat{t}^*) \cdot (\widehat{t}^*(\delta) - \widehat{t}^*) \\ &+ \sqrt{\delta}[v(\widehat{t}^*) + v'(\widehat{t}^*) \cdot (\widehat{t}^*(\delta) - \widehat{t}^*)] + o_p(\delta\epsilon^2),\end{aligned} \tag{88}$$

and

$$\widehat{t}^*(\delta) = \widehat{t}^* - \sqrt{\delta} \cdot \widehat{u}''(\widehat{t}^*)^{-1} \cdot v'(\widehat{t}^*) + o_p(\sqrt{\delta}\epsilon). \tag{89}$$

Plugging the expansion for $\widehat{t}^*(\delta)$ into the expansion for $\widehat{u}(\delta, \widehat{t}^*(\delta))$, we get

$$\widehat{u}(\delta, \widehat{t}^*(\delta)) = \widehat{u}(t^*) + \sqrt{\delta} v(\widehat{t}^*) + \delta \cdot \frac{1}{2} v'(\widehat{t}^*)^t \cdot \widehat{u}''(\widehat{t}^*)^{-1} \cdot v'(\widehat{t}^*) + o_p(\delta\epsilon^2). \tag{90}$$

By construction of v, we have $E[v(t)] = 0$, and $v'(t) = \frac{\sigma}{\sqrt{nf(t)}} \cdot \overline{w}_t(t, x') \cdot \epsilon$. Taking expectations conditional on $\widehat{u}$ thus yields

$$E[\widehat{u}(\delta, \widehat{t}^*(\delta))|\widehat{u}] = \widehat{u}(t^*) + \delta \cdot \frac{1}{2} \frac{\sigma^2}{nf^2(x')} \operatorname{tr}\left[\widehat{u}''(\widehat{t}^*)^{-1} \cdot \left(\overline{w}_t(\widehat{t}^*, x')\overline{w}_t(\widehat{t}^*, x')^t\right)\right] + o(\delta). \tag{91}$$

Taking the derivative with respect to δ and the expectation over the distribution of $\widehat{u}$ proves the claim. □

Proof of Corollary 2:
The optimal design f^* solves the variational problem

$$\begin{aligned}&\max \widehat{v}(f) \\ &s.t. \int f(x)dx = 1.\end{aligned} \tag{92}$$

The first order conditions to this problem are given by

$$\frac{\partial}{\partial\delta}\widehat{v}(f + \delta\tilde{f}) = \tilde{\kappa} \cdot \int \tilde{f}(x)dx = \tilde{\kappa}, \tag{93}$$

where $\tilde{f}$ might be any probability density function. Considering perturbations $\tilde{f}$ that approximate point masses at x, and using the limiting expression for $\frac{\partial}{\partial\delta}\widehat{v}(f + \delta\tilde{f})$ given by theorem 2 yields the claim of the corollary. □

Proof of proposition 2:
By corollary 2 of theorem 2,

$$\frac{1}{f^{*2}(x)} E\left[\operatorname{tr}\left[\widehat{u}''(\widehat{t}^*)^{-1} \cdot \left(\overline{w}_t(\widehat{t}^*, x)\overline{w}_t(\widehat{t}^*, x)^t\right)\right]\right] = \kappa$$

for all x. Under the maintained conditions,

$$
\begin{aligned}
&E\left[\operatorname{tr}\left[\widehat{u}''(\widehat{t}^*)^{-1}\cdot\left(\overline{w}_t(\widehat{t}^*,x)\overline{w}_t(\widehat{t}^*,x)^t\right)\right]\right]\\
=&E\left[\operatorname{tr}\left[(u''(\widehat{t}^*)^{-1}+o_p(1))\cdot\frac{l^2}{h^{2d_x+2}}\left(K'\left(\frac{x-\widehat{t}^*}{h}\right)K'\left(\frac{x-\widehat{t}^*}{h}\right)^t+o(1)\right)\right]\right]\\
=&\int E[u''(t)^{-1}|t^*=t]\cdot\frac{l^2}{h^{2d_x+2}}K'\left(\frac{x-\widehat{t}^*}{h}\right)K'\left(\frac{x-\widehat{t}^*}{h}\right)^t f^{t^*}(t)dt+o(1).
\end{aligned}
$$

A change of the variable of integration, combined with the continuity and positivity of f^{t^*} and compactness of support yields the claim. □

Proof of Theorem 3:

Under either of the two cases, we can treat f^* as fixed as we vary sample size. In case (ii) this holds by an envelope condition argument. By similar arguments as those leading to theorem 2, the dependence of the covariance kernel of $\widehat{u}$ on n is given by

$$
V_n=-\overline{W}E_n\overline{W}^t=\frac{1}{n}\overline{W}E\overline{W}^t,
$$

since $E=\frac{\sigma^2}{n}F^{-1}$. Writing this in integral form, we get

$$
(V_ng)(x)=\frac{\sigma^2}{n^2}\int\overline{w}(x,x')\overline{w}(x'',x')\frac{1}{f^*(x')}g(x'')dx'dx''.
$$

Increasing sample size n is therefore equivalent, to first order, to adding the random function

$$
v(x)=\int\overline{w}(x,x')\frac{\sigma}{n\sqrt{f^*(x)}}d\tilde{W}(x)
$$

to $\widehat{u}$, where $d\tilde{W}$ is a standard white noise process independent of $\widehat{u}$. For this process v, we get $E[v(t)]=0$,

$$
v'(t)=\int\overline{w}_t(t,x')\frac{\sigma}{n\sqrt{f^*(x)}}d\tilde{W}(x),
$$

and

$$
\operatorname{Var}(v'(t)|\widehat{u})=\frac{\sigma^2}{n^2}\int\frac{\overline{w}_t(t,x')^2}{f^*(x')}dx'.
$$

The same Taylor expansions as in the proof of theorem 2 apply, so that

$$
\widehat{u}(\delta,\widehat{t}^*(\delta))=\widehat{u}(t^*)+\sqrt{\delta}v(\widehat{t}^*)+\delta\cdot\frac{1}{2}v'(\widehat{t}^*)^t\cdot\widehat{u}''(\widehat{t}^*)^{-1}\cdot v'(\widehat{t}^*)+o_p(\delta\cdot\|v\|).
$$

and

$$
E[\widehat{u}(\delta,\widehat{t}^*(\delta))|\widehat{u}]=\widehat{u}(t^*)+\delta\cdot\frac{1}{2}\operatorname{tr}\left[\widehat{u}''(\widehat{t}^*)^{-1}\cdot(\operatorname{Var}(v'(t)|\widehat{u}))\right]+o(\delta).
$$

Taking expectations over the distribution of $\widehat{u}$ and then the derivative with respect to δ proves the first claim. The second claim is an immediate consequence of theorem 2. □

B Common choices for the covariance kernel C.

This section provides a brief review of some popular covariance kernels for Gaussian process priors.

Linear models

Consider a set of regressors W which might include interactions, powers, and other transformations of observed variables X, $W = w(X)$. Assume that we have a multivariate normal prior for the coefficients of a linear regression model, where

$$\begin{aligned} Y &= w(X)\beta + \epsilon \\ E[\epsilon|\beta, X] &= 0 \\ E[\beta|X] &= 0 \\ \mathrm{Var}(\beta|X) &= \Sigma_\beta. \end{aligned}$$

This assumption implies a Gaussian process prior for $m(x) = E[Y|X = x] = w(x) \cdot \beta$, with covariance kernel

$$C(x_1, x_2) = w(x_1)' \cdot \Sigma_\beta \cdot w(x_2).$$

Squared exponential covariance function

A common choice of prior in the machine learning literature (cf. Williams and Rasmussen, 2006) is defined by the covariance kernel

$$C(x_1, x_2) = \exp\left(-\frac{1}{2l^2}\|x_1 - x_2\|^2\right), \tag{94}$$

where $\|.\|$ is some appropriately defined norm measuring the distance between covariate vectors. The parameter l determines the length scale of the process.

This prior does not restrict functional form and can accommodate any shape of m. In this sense it is a nonparametric prior. One attractive feature of the squared exponential covariance kernel is that is puts all its mass on smooth functions, in the sense that m is infinitely mean-square differentiable. A function is mean-square differentiable if the normalized differences of m converge in L^2 to some function $\partial m(x)/\partial x$,

$$\frac{m(x + \epsilon) - m(x)}{\|\epsilon\|} \to^{L^2} \frac{\partial m(x)}{\partial x}$$

as $\|\epsilon\| \to 0$, cf. Williams and Rasmussen (2006, p81). Infinite mean square differentiability holds for all processes that have a covariance kernel C which is infinitely differentiable around points where $x_1 = x_2$.

The length scale l, and more generally the norm $\|x_1 - x_2\|$, determines the smoothness of the process, where larger length scales correspond to smoother processes. One measure of smoothness are the expected number of "upcrossings" at 0, i.e., the expected number of times the process crosses 0 from below in the interval $[0, 1]$. For a one-dimensional process with squared exponential kernel, this number equals $1/(2\pi l)$, cf. again Williams and Rasmussen (2006, p81).

Noninformativeness

We might want to consider priors which are "non-informative" about some key parameters of the behavioral relationships under consideration, such as level and slope, while at the same time using our prior assumptions about smoothness of m.[22] One way to formalize such non-informativeness is to consider limit cases where the prior variance for some parameters goes to infinity, and to use the corresponding limit estimators of m and u as posterior expectations.

In particular, given a covariance kernel K for a stochastic process n as well as a subset of regressors X^*, consider the process

$$\begin{aligned}
Y &= n(X) + X^*\beta + \epsilon \\
E[n] &= 0 \\
E[\beta|X] &= 0 \\
E[\epsilon|X] &= 0 \\
\mathrm{Cov}(n(x_1), n(x_2)) &= K(x_1, x_2) \\
\mathrm{Var}(\beta|X) &= \lambda\Sigma_\beta \\
\mathrm{Var}(\epsilon|X, \beta) &= \sigma^2 I \\
\beta^d &\perp n.
\end{aligned}$$

For this process we get

$$C(x_1, x_2) = K(x_1, x_2) + \lambda x_1^* \Sigma_\beta x_2^*.$$

In the limit-case $\lambda \to \infty$ the prior over β becomes "non-informative." In the limit we get (cf. Kasy 2013b)

$$\widehat{m}(x) = x_1\widehat{\beta} + K(x)(K + \sigma^2 I)^{-1}(Y - X_1\widehat{\beta}), \tag{95}$$

where

$$\widehat{\beta} = \left(X'(K + \sigma^2 I)^{-1}X\right)^{-1} X'(K + \sigma^2 I)^{-1}Y. \tag{96}$$

C Explicit formulas and graphs for covariance kernels

In section 3 we introduced several policy problems that fit into our general framework, and derived expressions for the covariance kernels B, D and K for these problems from a general covariance kernel C. In this section we provide analytic expressions and plots of these kernels, assuming that the covariance kernel C is equal to a squared exponential kernel

$$C(x_1, x_2) = \exp\left(-\frac{(x_1 - x_2)^2}{2l^2}\right) = \frac{1}{\phi(0)} \cdot \phi\left(\frac{x_1 - x_2}{l}\right) \tag{97}$$

where ϕ is the standard normal pdf and l is a parameter determining the length scale of the kernel.

[22] And note that *any* nonparametric estimation method has to use assumptions about smoothness!

The form of the operator L for the optimal insurance and optimal taxation problems discussed in section 3.1 was

$$(Lm)(t) = \lambda \int_0^t m(x)dx - t \cdot m(t).$$

For this operator L and kernel C, the prior (co)variances D, K and B are given by

$$\begin{aligned} D(t,x) &= \lambda \cdot \int_0^t C(x',x)dx' - t \cdot C(t,x) \\ &= \frac{1}{\phi(0)} \cdot \left[\lambda l \cdot \left(\Phi\left(\frac{t-x}{l}\right) - \Phi\left(\frac{-x}{l}\right) \right) - t \cdot \phi\left(\frac{t-x}{l}\right) \right], \end{aligned}$$

where Φ denotes the standard normal cdf,

$$\begin{aligned} K(t,t') &= \lambda^2 \cdot \int_0^t \int_0^{t'} C(x,x')dx'dx + t \cdot t' \cdot C(t,t') \\ &- \lambda \cdot \left(t' \cdot \int_0^t C(x,t')dx + t \cdot \int_0^{t'} C(x',x)dx' \right) \\ &= \frac{1}{\phi(0)} \cdot \left[\lambda^2 \cdot \int_0^t l \cdot \left(\Phi\left(\frac{x-t}{l}\right) - \Phi\left(\frac{-t}{l}\right) \right) dx + t \cdot t' \cdot \phi\left(\frac{x_1 - x_2}{l}\right) \right. \\ &\left. -\lambda l \cdot \left(t' \cdot \left(\Phi\left(\frac{t-t'}{l}\right) - \Phi\left(\frac{-t'}{l}\right) \right) + t \cdot \left(\Phi\left(\frac{t'-t}{l}\right) - \Phi\left(\frac{-t}{l}\right) \right) \right) \right] \end{aligned}$$

and

$$\begin{aligned} B(t,x) &= (\lambda - 1) \cdot C(t,x) - t \cdot \frac{\partial}{\partial t} C(t,x) \\ &= \frac{1}{\phi(0)} \cdot \left[(\lambda - 1) \cdot \phi\left(\frac{t-x}{l}\right) - \frac{t}{l} \cdot \phi'\left(\frac{t-x}{l}\right) \right] \\ &= \frac{\phi\left(\frac{t-x}{l}\right)}{\phi(0)} \cdot \left[(\lambda - 1) - \frac{t \cdot (t-x)}{l^2} \right]. \end{aligned}$$

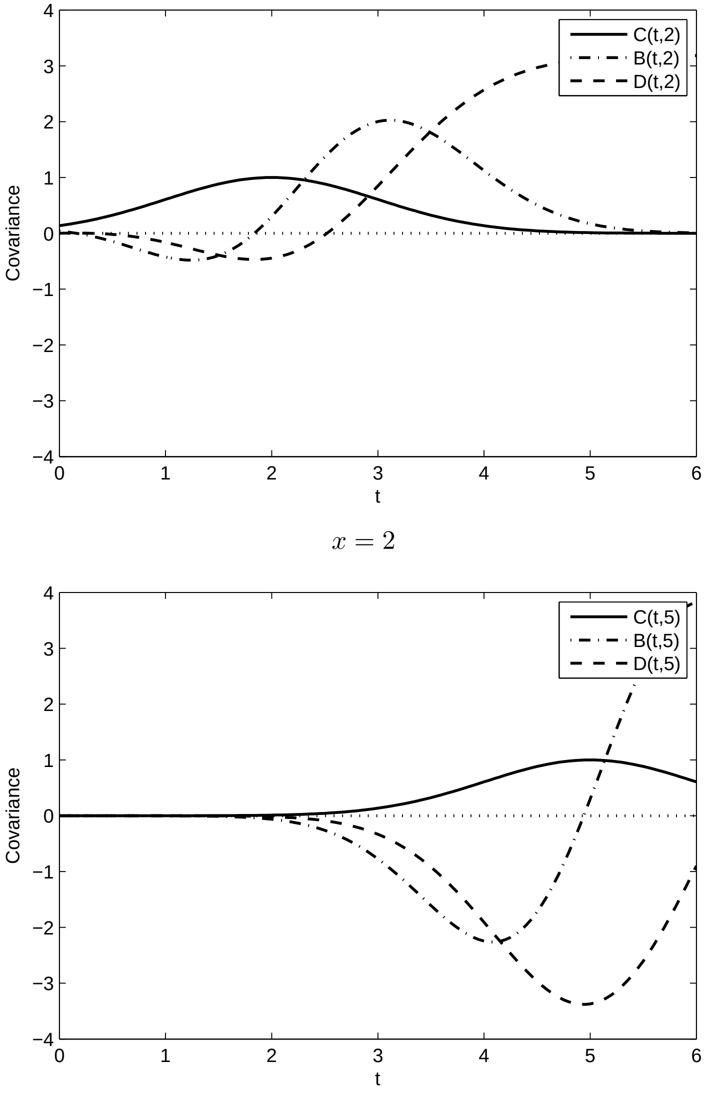
We finally get

$$\begin{aligned} \mathrm{Var}(u'(t)) &= \mathrm{Var}((\lambda - 1) \cdot m(t) - t \cdot m'(t)) = \\ &= (\lambda - 1)^2 \cdot C(t,t) - (\lambda - 1) \cdot t \cdot \frac{\partial}{\partial t'} C(t,t') + t^2 \cdot \frac{\partial^2}{\partial t' \partial t} C(t,t') \\ &= \frac{1}{\phi(0)} \cdot \phi\left(\frac{x_1 - x_2}{l}\right) \cdot \left[(\lambda - 1)^2 + \frac{t^2}{l^2} \right], \end{aligned}$$

where in the last expression several terms drop out since they are multiplied by $(t - t)/l$; in particular $\mathrm{Cov}(m(t), m'(t)) = 0$ for the squared exponential covariance kernel.

The following figures illustrate.

Figure 5: Covariance kernels for the optimal insurance problem



$x = 2$

$x = 5$

References

Angrist, J., G. Imbens, and D. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

Angrist, J. and V. Lavy (1999): "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *The Quarterly Journal of Economics*, 114, 533–575.

Angrist, J. D. and J.-S. Pischke (2010): "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics," *Journal of Economic Perspectives*, 24, 3–30.

Aron-Dine, A., L. Einav, and A. Finkelstein (2013): "The RAND Health Insurance Experiment, Three Decades Later," *Journal of Economic Perspectives*, 27, 197–222.

Baily, M. (1978): "Some aspects of optimal unemployment insurance," *Journal of Public Economics*, 10, 379–402.

Banerjee, A. and E. Duflo (2008): "The experimental approach to development economics," Tech. rep., National Bureau of Economic Research.

Bhattacharya, D. and P. Dupas (2008): "Inferring Welfare Maximizing Treatment Assignment under Budget Constraints," *NBER Working paper.*

Blundell, R. and J. Powell (2003): "Endogeneity in nonparametric and semiparametric regression models," in *Advances in economics and econometrics: Theory and applications, eighth world congress*, vol. 2, 655–679.

Casella, G. and R. Berger (2001): *Statistical inference*, Duxbury Press.

Chamberlain, G. (2011): "Bayesian Aspects of Treatment Choice," *Oxford Handbook of Bayesian Econometrics.*

Chetty, R. (2006): "A general formula for the optimal level of social insurance," *Journal of Public Economics*, 90, 1879–1901.

——— (2009): "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods," *Annual Review of Economics*, 1, 451–488.

Deaton, A. (2009): "Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development," *NBER working paper.*

Dehejia, R. (2005): "Program evaluation as a decision problem," *Journal of Econometrics*, 125, 141–173.

Feldstein, M. (1999): "Tax avoidance and the deadweight loss of the income tax," *Review of Economics and Statistics*, 81, 674–680.

Fryer, R. (2011): "Injecting successful charter school strategies into traditional public schools: Early results from an experiment in Houston," Tech. rep., National Bureau of Economic Research.

Ghosh, J. and R. Ramamoorthi (2003): *Bayesian nonparametrics*, Springer Verlag.

Ghoshal, S. and A. van der Vaart (2013): *Fundamentals of Nonparametric Bayesian Inference*, Cambridge Unversity Press.

Graham, B. S., G. Imbens, and G. Ridder (2008): "Measuring the average outcome and inequality effects of segregation in the presence of social spillovers," *working paper.*

Hirano, K. and J. Porter (2009): "Asymptotics for statistical treatment rules," *Econometrica*, 77, 1683–1701.

Imbens, G. W. (2010): "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, 48, 399–423.

Imbens, G. W. and W. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512.

Karatzas, I. and S. Shreve (1991): *Brownian motion and stochastic calculus*, vol. 113, Springer Verlag.

Kasy, M. (2013a): "Instrumental variables with unrestricted heterogeneity and continuous treatment," *working paper.*

——— (2013b): "Why experimenters should not randomize, and what they should do instead," *working paper.*

Krueger, A. (1999): "Experimental estimates of education production functions," *The Quarterly Journal of Economics*, 114, 497–532.

Li, Q. and J. Racine (2011): *Nonparametric econometrics: Theory and practice*, Princeton University Press.

Manski, C. F. (1993): "Identification of endogenous social effects: The reflection problem," *The Review of Economic Studies*, 60, 531–542.

——— (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, pp. 1221–1246.

Mas-Colell, A., M. Whinston, and J. Green (1995): *Microeconomic theory*, Oxford university press.

Matheron, G. (1973): "The intrinsic random functions and their applications," *Advances in applied probability*, 439–468.

Miguel, E. and M. Kremer (2003): "Worms: identifying impacts on education and health in the presence of treatment externalities," *Econometrica*, 72, 159–217.

Mirrlees, J. (1971): "An exploration in the theory of optimum income taxation," *The Review of Economic Studies*, 175–208.

Nevo, A. and M. Whinston (2010): "Taking the Dogma out of Econometrics: Structural Modeling and Credible Inference," *The Journal of Economic Perspectives*, 24, 69–81.

Newey, W. K. and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

Rivkin, S., E. Hanushek, and J. Kain (2005): "Teachers, schools, and academic achievement," *Econometrica*, 73, 417–458.

Robert, C. (2007): *The Bayesian choice: from decision-theoretic foundations to computational implementation*, Springer Verlag.

Saez, E. (2001): "Using elasticities to derive optimal income tax rates," *The Review of Economic Studies*, 68, 205–229.

Saez, E. and S. Stantcheva (2012): "Optimal Tax Theory with Endogenous Social Marginal Welfare Weights," .

Samuelson, P. (1947): "Foundations of Economic Analysis," .

Silverman, B. (1984): "Spline smoothing: the equivalent variable kernel method," *The Annals of Statistics*, 898–916.

Sollich, P. and C. Williams (2005): "Using the Equivalent Kernel to Understand Gaussian Process Regression," *working paper.*

Stoye, J. (2011): "Minimax regret treatment choice with covariates or with limited validity of experiments," *Journal of Econometrics.*

van der Vaart, A. (2000): *Asymptotic statistics*, Cambridge University Press.

van der Vaart, A. and J. van Zanten (2008a): "Rates of contraction of posterior distributions based on Gaussian process priors," *The Annals of Statistics*, 36, 1435–1463.

——— (2008b): "Reproducing kernel Hilbert spaces of Gaussian priors," *IMS Collections*, 3, 200–222.

Wahba, G. (1990): *Spline models for observational data*, vol. 59, Society for Industrial Mathematics.

Wiener, N. (1949): *The interpolation, extrapolation and smoothing of stationary time series*, J Wiley.

Williams, C. and C. Rasmussen (2006): *Gaussian processes for machine learning*, MIT Press.

Yaglom, A. M. (1962): *An introduction to the theory of stationary random functions*, Prentice-Hall.