Lancaster University
Management School

Economics Working Paper Series

2020/010

# An augmented first-order approach for incentive problems

Philipp Renner

The Department of Economics
Lancaster University Management School
Lancaster LA1 4YX
UK

# An augmented first-order approach for incentive problems

Philipp Renner
Department of Economics
Lancaster University
phrenner@gmail.com

May 25, 2020

**Abstract**

Incentive constraints are constraints that are optimization problems themselves. If these problems are non convex then the first order approach fails. We propose an alternative solution method where we use the value function as an additional constraint. This ensures that all solutions are incentive compatible. To get the value function we use a function interpolator like sparse grids. We demonstrate our approach by solving two examples from the literature were it was shown that the first order approach fails.
**Keywords:** incentive constraints, first order approach, parametric optimization, value function approach.

**JEL codes:** C63, D80, D82.

## 1 Introduction

Incentive constraints are an important part of economic modelling. Whenever we encounter a situation where one party cannot observe the other's action, we need an incentive constraint to model the resulting interaction, encompassing optimal taxation, manager remuneration and mechanism design. We therefore need methods to deal with incentive constraints.

The general idea to handle these constraints, which are optimization problems themselves, is to replace them with their first order conditions. However it is well known that if the lower level is non-convex then the true solution might not even be amongst the stationary points of the first order approach (Mirrlees 1999). Even if the lower problem is convex the optimization problem resulting from the first order approach is difficult to deal with because the standard constraint qualifications fail at every feasible point (Ye and Zhu 1995).

Researchers have addressed this in multiple ways. First we can find conditions under which the lower level is a convex optimization problem, see Rogerson (1985) or recently Chade and Swinkels (2020). However, the trade-off is that either the assumptions are strict or that they do not work for a numerical treatment.

Second is using the first-order conditions and doing a post verification like in Abraham and Pavoni (2008). But this leaves us without an option if the verification fails. Furthermore since we know that the true solution might not even be a stationary point, we therefore only can certify feasibility and not optimality.

Lastly we can look for alternatives to the classical first-order approach. We showed in Renner and Schmedders (2015) that if we assume that the agent's problem is a rational function in his variables then we can use techniques from real algebraic geometry to solve the resulting bilevel optimization problem. There are two drawbacks to this approach. First, we need the rational function assumption which is not satisfied for many popular models. Second, the size of the problem goes up exponentially in the number of agent's decision variables.

In this paper we propose to augment the first order approach by adding the constraint that the agent's payoff has to be at least the value of his value function. This is called the value function approach in the mathematical programming literature and was first proposed by Outrata (1990). All optimal solutions of the resulting problem will therefore be incentive compatible.

An obvious problem is that we usually do not know the agent's value function and in general there is no closed form description of the value function. But if we regard the principal's choices as fixed then the agent's optimization problem is easy to solve. Therefore we will use an approximation of the agent's value function computed by solving the agent's problem at fixed set of points. The domain of the approximation will be the choices of the principal, e.g. if the principal has two variables to influence the agent then the agent's value function needs to be approximated over a two dimensional domain. Thus the complexity of the resulting problem is mostly independent of the number of choices for the agent but instead depends on the number of parameters.

The complexity of functional approximation depends on the dimension of the domain and suffers from the curse of dimensionality. Fortunately there are advances in function approximation that mitigate the curse and allow us to approximate several problems that were previously untracktable. If the domain of approximation can be mapped to a square then sparse grids (Smolyak 1963; Zenger 1991) can be used effectively to approximate even high dimensional functions, as in Brumm and Scheidegger (2017). If the domain of approximation is non square then we can use grid free methods from machine learning like Gaussian processes or neural networks (Murphy and Bach 2012).

We demonstrate the effectiveness of our approach on the two classical examples in the literature. We solve the original counter example from Mirrlees (1999) and the example from Araujo and Moreira (2001).

## 2 Value function approach for static principal agent problems

We want to solve the following problem

$$\begin{aligned}\max_{x,a}\ & F(x,a)\\ \text{s.t. }& G(x,a)\geq 0,\ H(x,a)=0\\ & a\in\arg\max_{\hat{a}\in\Gamma(x)} f(x,\hat{a})\end{aligned}\tag{1}$$

where $x\in\mathbb{R}^{n_p}$, $a\in\mathbb{R}^{n_a}$, $\Gamma(x)=\{\hat{a}\mid g(x,\hat{a})\geq 0, h(x,\hat{a})=0\}$ and $F,G,H,f,g,h$ are twice continuously differentiable functions.

The main issue is the incentive compatibility constraint

$$a\in\arg\max_{\hat{a}\in\Gamma(x)} f(x,\hat{a})\tag{2}$$

which tells us that if the principal wants to induce the agent to take action $a$, she has to take the agents preferences into account.

An immediate idea to solve (1) is to replace (2) with its first order conditions which yields

$$\begin{aligned}\max_{x,a}\ & F(x,a)\\ \text{s.t. }& G(x,a)\geq 0,\ H(x,a)=0\\ & \nabla_{\hat{a}}L(x,\hat{a},\lambda^g,\lambda^h)=0,\\ & g(x,\hat{a})\geq 0,\ h(x,\hat{a})=0,\\ & (\lambda^g)^T g(x,\hat{a})=0,\ \lambda^g\geq 0\end{aligned}\tag{3}$$

where $L(x,\hat{a},\lambda^g,\lambda^h)=f(x,\hat{a})+\sum_j\lambda^g_j g_j(x,\hat{a})+\sum_i\lambda^h_i h_i(x,\hat{a})$ is the Lagrangian of the agent's problem.

However, Dempe and Dutta (2012) show that even if the agent's problem is a convex optimization problem, there are two potential problems. If Slater's constraint qualification fails then global solutions to the reformulation do not necessarily correspond to global solutions of the principal agent problem. Further even if Slater holds then local solutions of the first order approach are not necessarily local solutions of the original problem.

More importantly if the agent's problem is none convex then Mirrlees (1999) showed that it is possible that the optimal solution to (1) is not amongst the stationary points of (3). The agent's problem becomes non-convex if we assume that he has a von Neumann Morgenstern utility where the probabilities depend on his own actions.

Therefore we desire alternative reformulations that allow us to solve problems where the reformulation (3) fails. If we assume that $f(x, a)$ attains its maximum on $\Gamma(x)$ for all $(x, a)$ satisfying $G(x, a) \geq 0,\ H(x, a) = 0$ then we can define the value function of the agent's problem

$$\varphi(x) = \max_{\hat{a} \in \Gamma(x)} f(x, \hat{a}) \tag{4}$$

Outrata (1990) proposed using the value function to rewrite problem (1) into

$$\begin{aligned} \max_{x,a}\ & F(x, a) \\ \text{s.t. }\ & f(x, a) \geq \varphi(x) \\ & G(x, a) \geq 0,\ H(x, a) = 0 \\ & g(x, a) \geq 0,\ f(x, a) = 0 \end{aligned} \tag{5}$$

It is immediate that (5) is equivalent to (1).

Ye and Zhu (2010) propose to combine the value function (5) with the first order approach (3) into the following problem

$$\begin{aligned} \max_{x,a}\ & F(x, a) \\ \text{s.t. }\ & f(x, a) \geq \varphi(x) \\ & G(x, a) \geq 0,\ H(x, a) = 0 \\ & \nabla_{\hat{a}} L(x, \hat{a}, \lambda^g, \lambda^h) = 0, \\ & g(x, \hat{a}) \geq 0,\ h(x, \hat{a}) = 0, \\ & (\lambda^g)^T g(x, \hat{a}) = 0,\ \lambda^g \geq 0 \end{aligned} \tag{6}$$

However in this form standard constraint qualifications will always fail due to the value function as a constraint. To remedy this Ye and Zhu (2010) use the value function constraint as a penalty and rewrite the problem to

$$\begin{aligned} \max_{x,a}\ & F(x, a) + \xi(f(x, a) - \varphi(x)) \\ \text{s.t. }\ & G(x, a) \geq 0,\ H(x, a) = 0 \\ & \nabla_{\hat{a}} L(x, \hat{a}, \lambda^g, \lambda^h) = 0, \\ & g(x, \hat{a}) \geq 0,\ h(x, \hat{a}) = 0, \\ & (\lambda^g)^T g(x, \hat{a}) = 0,\ \lambda^g \geq 0 \end{aligned} \tag{7}$$

Where $\xi > 0$ is a penalty parameter.

In general the value function $\varphi(x)$ is unknown to us. However for any given $x$ the agent's optimization problem is often easily solved. Therefore we propose to replace the true value function with a precomputed approximation where we use a function approximator. This means that we need to approximate a multidimensional function, a seemingly difficult task. However, there have been many advances in function approximation that make this originally impossible task tractable. There are two major options we can choose from.

First there are the grid based methods, the most efficient of them being adaptive sparse grids (Brumm and Scheidegger 2017). They have the advantage that they are well understood in both theory and practice. Sparse grids perform especially well when used to approximate Lipschitz continuous functions. The drawback is that we are restricted to a hypercube, i.e. a set of the form $\Pi_{i=1}^{n}[a_i, b_i]$. Fortunatley in many situations we can simply use a box that contains all possible values of $x$ that we want to plug in.

Second are the grid free methods that are used in Machine Learning, like Gaussian Processes (Scheidegger and Bilionis 2019) or Deep Neural Networks (Azinovic, Gaegauf, and Scheidegger 2019). Their advantage is that we can choose the points at which we approximate our function arbitrarily and therefore we can use them to approximate functions on any domain. However, they have worse theoretical properties in thus far that where we pick the sample points significantly influence the quality of our approximation and there are no theoretical save ways to efficiently pick these points.

Unfortunalty the value function $\varphi$ is non-differentiable in all relevant cases and so we need to look to non-smooth analysis to study its properties. For this we first need a new notion of a derivative.

**Definition 1.** Let $f : \mathbb{R}^n \to \bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ be an extended value function, $\bar{x} \in \mathbb{R}^n$ and $f(\bar{x})$ finite. Then the *regular subdifferential* of $f$ at $\bar{x}$ is defined as

$$\hat{\partial} f(\bar{x}) = \{v \in \mathbb{R}^n | f(x) \geq f(\bar{x}) + < v, x - \bar{x} > + o(\|x - \bar{x}\|)\}.$$

The *limiting subdifferential* of $f$ at $\bar{x}$ is

$$\partial f(\bar{x}) = \{v \in \mathbb{R}^n | \exists (v_k),\ v = \lim_k v_k, v_k \in \hat{\partial} f(x_k), x_k \to \bar{x}, f(x_k) \to f(\bar{x})\}.$$

Note when $f$ is convex then the above definitions are equivalent and equal to the subgradient from convex analysis, i.e.

$$\partial f(\bar{x}) = \{v \in \mathbb{R}^n | f(x) \geq f(\bar{x}) + < v, x - \bar{x} >\}.$$

The good news is that even though $\varphi$ is non-differentiable we can impose conditions under which the function is Lipschitz continuous. We use the results from Guo, Lin, Ye, and Zhang (2014) to ensure Lipschitz continuity. They require two conditions.

**Definition 2.** We say *restricted inf-compactness* holds around $\overline{x}$ if $\varphi(\overline{x})$ is finite and there is a compact set $A \subset \mathbb{R}^{n_a}$ and a $\varepsilon > 0$ such that, for all $x$ with $\|\overline{x} - x\| < \varepsilon$ for which $\varphi(x) > \varphi(\overline{x}) - \varepsilon$, the agent's problem (4) has a solution in $A$.

Note this is a very weak condition and it is trivially satisfied if for all $x$, $\Gamma(x) \subset A$ for some compact set $A$.

**Definition 3.** A point $(\overline{x}, \overline{a})$ with $\overline{a} \in \Gamma(\overline{x})$ is called *quasi-normal* if there is no nonzero vector $(\lambda^g, \lambda^h)$ such that

$$0 = \nabla g(\bar{x}, \bar{a})^T \lambda^g + \nabla h(\bar{x}, \bar{a})^T \lambda^h, \quad \lambda^g \geq 0$$

and there exists $(x^k, a^k) \to (\bar{x}, \bar{a})$ such that

$$\begin{aligned} &\lambda_i^g > 0 \Rightarrow \lambda_i^g g_i\left(x^k, a^k\right) > 0, \\ &\lambda_i^h \neq 0 \Rightarrow \lambda_i^h h_i\left(x^k, a^k\right) > 0. \end{aligned}$$

This is a weaker version of the Mangasarian-Fromovitz constraint qualification and therefore is easily satisfied.

**Theorem 1.** *(Guo, Lin, Ye, and Zhang 2014, Cor. 4.8) Let $S(\bar{x})$ denote the optimal solution mapping for the agent's problem* (4)*. If restricted inf-compactness holds around $\bar{x}$ and for each $\bar{a} \in S(\bar{x})$, $(\bar{x}, \bar{a})$ is quasi-normal, then the value function $\varphi(x)$ is Lipschitz continuous around $\bar{x}$. Furthermore the limiting subdifferential of $\varphi$ satisfies*

$$\partial \varphi(\bar{x}) \subseteq \widetilde{W}(\bar{x})$$

*where*

$$\widetilde{W}(\bar{x}) := \bigcup_{\bar{a} \in S(\bar{x})} \left\{ \nabla_x f(\bar{x}, \bar{a}) + \nabla_x g(\bar{x}, \bar{a})^T \lambda^g + \nabla_x h(\bar{x}, \bar{a})^T \lambda^h : \left(\lambda^g, \lambda^h\right) \in \mathcal{M}(\bar{x}, \bar{a}) \right\}$$

*where $\mathcal{M}(\bar{x}, \bar{a})$ is the set of quasi-normal multipliers defined as*

$$\mathcal{M}(\bar{x}, \bar{a}) := \left\{ \begin{array}{c|l} & 0 = \nabla_a f(\bar{x}, \bar{a}) + \nabla_a g(\bar{x}, \bar{a})^T \lambda^g + \nabla_a h(\bar{x}, \bar{a})^T \lambda^h, \lambda^g \geq 0 \\ \left(\lambda^g, \lambda^h\right) & \text{there exists } \left(x^k, a^k\right) \to (\bar{x}, \bar{a}) \text{ such that} \\ & \lambda_i^g > 0 \Rightarrow \lambda_i^g g_i\left(x^k, a^k\right) > 0 \\ & \lambda_i^h \neq 0 \Rightarrow \lambda_i^h h_i\left(x^k, a^k\right) > 0 \end{array} \right\}.$$

*Lastly if $\widetilde{W}(\bar{x}) = \{\zeta\}$ is a singleton then $\varphi(x)$ is strictly differentiable with $\nabla \varphi(x) = \zeta$.*

# 3 Classical static examples

To show the viability of our approach we compute the solution to two examples from the literature where it is known that the first-order approach fails.

First we look at the example from Mirrlees (1999), that is also discussed in Ye and Zhu (2010).

$$\begin{aligned} \max_{x,y} \; & -(x-2)^2-(y-1)^2 \\ \text{s.t. } & y \in \arg \max_{\tilde{y}\in[-2,2]} xe^{-(\tilde{y}+1)^2}+e^{-(\tilde{y}-1)^2} \end{aligned} \tag{8}$$

It is known that the optimal solution to this problem is at $(1, 0.95753)$. Using the first-order approach we obtain

$$\begin{aligned} \max_{x,y} \; & -(x-2)^2-(y-1)^2 \\ \text{s.t. } & -2e^{-(-1+y)^2}(-1+y)-2e^{-(1+y)^2}x(1+y)+\lambda_1^g-\lambda_2^g=0 \\ & \lambda_1^g(2+y)=0 \\ & \lambda_2^g(2-y)=0 \\ & -2\leq y\leq 2 \end{aligned} \tag{9}$$

We use SNOPT (Gill, Murray, and Saunders 2005) to solve the optimization problem and find the local solution $(1.991209, 0.8947100)$. Next we approximate the value function with an adaptive sparse grid with 315 points and resolve the problem with the value function approach (7). We find the solution $(0.9985342, 0.9575735)$ which is close to the true solution.

Next we look at the example from Araujo and Moreira (2001).

$$\begin{aligned} \max_{x_1,x_2,y} \; & (1-y^3)(1-x_1)+y^3(5-x_2) \\ \text{s.t. } & y \in \arg \max_{\tilde{y}\in[0,0.9]} (1-y^3)(\sqrt{x_1}-y^2)+y^3(\sqrt{x_2}-y^2) \end{aligned} \tag{10}$$

The optimal solution is found at $(0, 1.23457, 0.9)$. Using the first-order conditions we get the following problem

$$\begin{aligned} \max_{x_1,x_2,y} \; & (1-y^3)(1-x_1)+y^3(5-x_2) \\ \text{s.t. } & -3y^2\left(\sqrt{x_1}-y^2\right)+3y^2\left(\sqrt{x_2}-y^2\right)-2y^4-2\left(1-y^3\right)y+\lambda_1^g-\lambda_2^g=0 \\ & \lambda_1^g y=0 \\ & \lambda_2^g(0.9-y)=0 \\ & 0\leq y\leq 0.9 \end{aligned} \tag{11}$$

Without the value function we get the stationary point $(0, 0.5489, 0.9)$, which is not a feasible solution to the original problem. We approximate the agent's valuefunction with a grid with 258 points and get the solution $(0.0, 1.236787, 0.9)$.

# References

Abraham, A. and N. Pavoni (2008): "Efficient Allocations with Moral Hazard and Hidden Borrowing and Lending: A Recursive Formulation," *Review of Economic Dynamics*, 11, 781–803.

Araujo, A. and H. Moreira (2001): "A General Lagrangian Approach for Non-concave Moral Hazard Problems," *Journal of Mathematical Economics*, 35, 17–39.

Azinovic, M., L. Gaegauf, and S. Scheidegger (2019): "Deep Equilibrium Nets," SSRN Scholarly Paper ID 3393482, Social Science Research Network, Rochester, NY.

Brumm, J. and S. Scheidegger (2017): "Using Adaptive Sparse Grids to Solve High-Dimensional Dynamic Models," *Econometrica*, 85, 1575–1612.

Chade, H. and J. Swinkels (2020): "The no-upward-crossing condition, comparative statics, and the moral-hazard problem," *Theoretical Economics*, 15, 445–476, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/TE2937.

Dempe, S. and J. Dutta (2012): "Is bilevel programming a special case of a mathematical program with complementarity constraints?" *Mathematical Programming*, 131, 37–48.

Gill, P. E., W. Murray, and M. A. Saunders (2005): "SNOPT: An SQP Algorithm for Large-Scale Constrained Optimization," *SIAM Review*, 47, 99–131.

Guo, L., G.-H. Lin, J. J. Ye, and J. Zhang (2014): "Sensitivity Analysis of the Value Function for Parametric Mathematical Programs with Equilibrium Constraints," *SIAM Journal on Optimization*, 24, 1206–1237, publisher: Society for Industrial and Applied Mathematics.

Mirrlees, J. A. (1999): "The Theory of Moral Hazard and Unobservable Behaviour: Part I," *The Review of Economic Studies*, 66, 3–21, publisher: Oxford Academic.

Murphy, K. P. and F. Bach (2012): *Machine Learning: A Probabilistic Perspective*, Cambridge, MA: MIT Press.

Outrata, J. V. (1990): "On the numerical solution of a class of Stackelberg problems," *Zeitschrift fuer Operations Research*, 34, 255–277.

Rasmussen, C. E. and C. K. I. Williams (2006): *Gaussian Processes for Machine Learning*, University Press Group Limited.

Renner, P. and K. Schmedders (2015): "A Polynomial Optimization Approach to Principal-Agent Problems," *Econometrica*, 83, 729–769.

ROGERSON, W. P. (1985): "The First-Order Approach to Principal-Agent Problems," *Econometrica*, 53, 1357–1367.

SCHEIDEGGER, S. AND I. BILIONIS (2019): "Machine learning for high-dimensional dynamic stochastic economies," *Journal of Computational Science*, 33, 68–82.

SMOLYAK, S. (1963): "Quadrature and interpolation formulas for tensor products of certain classes of functions," *Dokl. Akad. Nauk SSSR*, 148, 1042–1045.

YE, J. J. AND D. ZHU (2010): "New Necessary Optimality Conditions for Bilevel Programs by Combining the MPEC and Value Function Approaches," *SIAM Journal on Optimization*, 20, 1885–1905, publisher: Society for Industrial and Applied Mathematics.

YE, J. J. AND D. L. ZHU (1995): "Optimality conditions for bilevel programming problems," *Optimization*, 33, 9–27, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/02331939508844060.

ZENGER, C. (1991): "Sparse Grids," in *Parallel Algorithms for Partial Differential Equations*, Vieweg Verlagsgesellschaft, vol. 31 of *Notes on Numerical Fluid Mechanics*, 241–251.