Inter-regional insurance

Ben Lockwood*

Department of Economics, University of Warwick, and CEPR, Coventry CV4 7AL, UK

Received 1 February 1997; received in revised form 1 December 1997; accepted 1 July 1998

Abstract

This paper considers the problem facing a central government which can insure regional governments (by use of intergovernmental grants) against region-specific and privately observed shocks either to income, or demand for, or cost of, the public good. Notable results are: (i) depending on the source of the shock, the grant may induce over- or undersupply of the public good relative to the Samuelson rule; (ii) with public good spillovers between regions, there is two-way distortion of public good supply — that is, qualitatively different distortions (relative to the Samuelson rule) for different values of the shock; (iii) the solution to the central government’s problem may depend qualitatively on whether regional taxation is lump-sum or distortionary. © 1999 Elsevier Science S.A. All rights reserved.

Keywords: Intergovernmental grants; Asymmetric information; Fiscal federalism

JEL classification: D70; H41; H77

1. Introduction

In many federal countries, intergovernmental grants — that is, revenue transfers from higher to lower levels of government — are important sources of revenue for lower levels of government.1 There is now a large literature explaining why such grants may be used. For example, it is often suggested that specific matching

* E-mail: b.lockwood@warwick.ac.uk
1 For example, in 1990, intergovernmental transfers as a share of GDP in the following federal countries were: US 5.4%, Canada 7.4%, Australia 8.8%, Germany 4.6%, and Switzerland 5.7% (Costello, 1993).
grants (grants which are proportional to public expenditure of a given type by the lower-level government) help internalize inter-regional spillovers from public goods (Oates, 1972), and that lump-sum grants are intended to redistribute income between regions (Boadway and Wildasin, 1984). More recently, it has been suggested\(^2\) that in the US, grants from central to state governments help insure residents against shocks to gross state product (Sorensen et al., 1995).

In these circumstances, and with symmetric information between central and regional government, the optimal structure of such grants is clear. There should be a lump-sum grant element, depending only on the characteristics of the region, that is used to redistribute or insure, and a grant element that is matched to regional expenditure on the public good, and which is set at a level that internalizes the public good spillover (Oates, 1972).

However, it has long been recognized (for example, Oates, 1972, pp. 11–14), that regional governments are likely to be better informed about demand for, and cost of, public good provision in their region than national government. In spite of this informal recognition of the fact of asymmetric information between central and regional government, it is only recently that the role of asymmetric information in determining intergovernmental grants has been formally examined (Boadway et al., 1998; Bordignon et al., 1996; Bucovetsky et al., 1998; Cornes and Silva, 1996; Raff and Wilson, 1995). This paper is also concerned with this issue, but differs from the existing literature in a number of distinct ways.

Firstly, I model explicitly a feature of the economy that (as remarked above) is usually thought to imply a role for grants from central to regional government, namely \textit{inter-regional spillovers}, or externalities,\(^3\) from the provision of regional public good (Oates, 1972; Boadway and Wildasin, 1984).

Secondly, I take an ex ante view of the design of intergovernmental grants by assuming that regions are ex ante identical but are subject to stochastic shocks as described below. Thus, in this interpretation, intergovernmental grants are risk-sharing contracts, or to put it another way, central government provides insurance

\(^2\)The paper by Sorensen et al. estimates that 14\% of a shock to gross state product is smoothed by Federal taxes, transfers and grants.

\(^3\)Raff and Wilson (1995) allow for labour mobility between jurisdictions. This does not, however, generate an externality as workers do not share in the economic rents generated by the public input to production, and nor do they help finance the provision of this public input. Thus, the two factors that generate migration externalities (see e.g. Boadway and Flatters, 1982) are not present in this model. Bucovetsky et al. (1998), on the other hand, do use a model with inter-regional tax externalities arising from the mobility of the tax base, capital.

\(^4\)Because regions are ex ante identical (and agents within regions are also identical ex post, after the productivity shock) we can take an explicit ‘public choice’ approach to the modelling of government behaviour at both national and regional levels, assuming that government decision-making is determined by the preferences of the representative voter. This is in contrast to Boadway et al. (1998), where the regional government is modelled as a bureaucratic agency with different objectives to the citizens it governs, Raff and Wilson (1995), where central government has an (unexplained) preference for redistribution, modelled by the assumption of a concave social welfare function.
to regional governments. Some other contributions to this literature, by contrast, rely on ex ante differences between regions to get results.

The third distinctive feature of this paper is that the modelling of stochastic shocks is rather general. Unlike the other papers in this area, we allow regions to vary ex post i.e. after the stochastic shock, in three ways; with respect to income, the cost of producing the public good, and with respect to the demand for the public good. These possibilities cover all the kinds of variation between regions allowed for in the literature so far, and allow a systematic investigation of how the structure of the optimal grant varies with the type of shock, which is one of the main features of this paper (see Table 1 below for a summary of these results). Also, again unlike other papers in this area, we allow regions to be of many, rather than two ‘types'; that is, we allow the stochastic shock hitting any region to be a continuously distributed random variable. This allows us to say more about the structure of the optimal grant than the rest of the literature.

The fourth distinctive feature of the analysis of this paper is that it studies the sensitivity of the results to different assumptions about what regional government activities central government can observe. In particular, we allow central government to observe — and condition the grant on — either the physical output of the public good, or on the revenue raised by regional governments.

Finally, this paper also considers the sensitivity of the results to the nature of regional taxation. Initially, to bring out the issues as clearly as possible, we suppose that regional governments have access to lump-sum taxation. We then extend all analysis to the case of distortionary taxation, and we see that several key results change. Other papers in this literature without exception assume distortionary taxation.

So, in summary, this paper offers a very general framework within which the assumptions made by the other studies in this area may be located and compared.

The key results of this paper may be described as follows. Firstly, in this setting, the central government’s problem is to choose a grant, conditional on its available information (public goods supply or revenue raised in the region) to maximize the expected utility of a typical region subject to budget balance, participation, and incentive constraints. Call this the grant allocation problem, or GAP. Given the concavity of regional utility in the grant, central government faces in the GAP the problem of trading off insurance of the regions against offering correct incentives.

\[\text{Boadway et al. (1998) and Cornes and Silva (1996) allow regions to differ only in the cost of public good provision, Bucovetsky et al. (1998) have differences in willingness to pay for the public good, Raff and Wilson (1995) allow differences only in the marginal product of a productivity-enhancing public input, and finally, Bordignon et al. (1996) allow regions to differ only in labour productivity.} \]

\[\text{It is well-known in the principal-agent literature that with a finite number of types, the optimal payment schedule (the payment from principal to agent as a function of the agent’s action, in this case the grant as a function of public good provision) that induces the agent to carry out, or implement, the principal’s most preferred incentive-compatible action is not unique. Indeed, with just two types, just two points on the optimal payment schedule can be identified.} \]
for public good provision. With a continuum of types, public good spillovers, and risk-averse regions, this is a non-trivial problem to solve.

The first result of the paper emphasizes that an insurance motive is necessary for asymmetric information to be important in the GAP. If regional utility is linear in the grant then the government can perfectly internalize spillovers in public good provision by offering to each region a simple linear matching grant, as in the textbook treatment (Oates, 1972; Boadway and Wildasin, 1984).

The main results of the paper characterize the optimal grant and the implied rule for public good provision when regions are risk-averse. These features turn out to depend on the nature of the shock, whether public good spillovers are present or not, and whether regional taxation is lump-sum or distortionary — a summary is given in Table 1 in Section 8 of the paper. To put the main results in context, note that in the benchmark case of symmetric information, the first-best can always be decentralized with a simple linear matching grant for region $i$ of the form $g_i = T_i + mg$, where $g_i$ is public good supply in region $i$, $m > 0$ is the economy-wide value of the spillover effect from an increment in $g$, and $T_i$ is independent of $g_i$ but is chosen to fully insure the region and so will depend on the (publicly known) characteristics of the region. To put it another way, $m$ is the per-unit Pigouvian subsidy to the region that exactly internalizes the spillover. So, with symmetric information, if there are no spillovers, the grant should be lump-sum.

With asymmetric information, the main results are the following.

1. The qualitative distortion in public goods supply (relative to the Samuelson rule\(^8\)) induced by the optimal grant varies with the source of the shock hitting the regions, and oversupply, rather than undersupply, of the public good is quite possible.

2. If there are inter-regional spillovers, then it turns out that (unlike in the standard principal-agent problem), public good supply is distorted at both ends of support of the distribution of the shock, and moreover, the distortion is qualitatively different at the top and bottom — two-way distortion. For example, with stochastic demand, in the regions where demand is highest (lowest), then the public good is oversupplied (undersupplied). Qualitatively different distortions at the top and bottom of the distribution of agent types is, as far as I know, a new result in the principal-agent literature,\(^9\) and a result of externalities between multiple agents.

\(^7\)This kind of result can be found for example in Oates (1972).

\(^8\)Or, in the case that the regional governments use distortionary taxes, the modified Samuelson rule where the marginal rate of transformation between public and private goods is multiplied by the marginal cost of public funds.

\(^9\)Such a result was independently discovered by Bucovetsky et al. (1998). The relationship between that paper and this is discussed in more detail in Section 7.
3. The GAP when the cost of producing the public good is stochastic has a rather different structure to the other two cases. In particular, the qualitative features of the solution depend on the curvature of the subutility function over the public good, as measured by the elasticity of marginal utility with respect to the public good. In particular, if this elasticity is unity everywhere (i.e. the subutility function is logarithmic), then the first-best can be achieved. If there are no spillovers, and the elasticity is less than (resp. greater than) unity, expenditure on the public good is too high (resp. too low) relative to the first-best.

4. When taxes are lump-sum, the solution to the GAP is invariant to whether the central government can condition grants on tax revenue raised by the region, or on the expenditure on the public good.\(^\text{10}\)

5. Under some conditions, the qualitative features of the solution to the GAP depend on whether regional governments levy lump-sum or distortionary taxes.

The rest of the paper is arranged as follows. Section 2 sets up the model and the GAP. Sections 3, 4 and 5 analyze the GAP when regions differ ex post in demand for the public good, cost of the public good, and income, respectively. Section 6 shows that the results are robust to whether the central government observes regional tax revenue or expenditure on the public good. Section 7 extends the analysis to the case where regional governments use distortionary taxes. Section 8 concludes and discusses related literature.

### 2. The model

We consider a country with \(n\) regions, each of which is inhabited by a representative immobile resident. The resident of region \(i\) has preferences over a public good, a private good and leisure given by

\[
u(g^i, e^i, \theta^i) + \nu(c^i), e^i = \frac{1}{n - \sum_{j=1}^{n} g^j}
\]

(1)

where \(c^i\) is the level of consumption of the private good in region \(i\), \(g^i\) is the level of public good supply in region \(i\), and \(\theta^i\) is a parameter that measures the willingness to pay, or demand for the public good, which may vary across regions. The variable \(e^i\) models a spillover effect\(^{11}\) from public good provision in other

\(^{10}\)In the variable cost case, the equivalence is between conditioning on revenue raised by the region, and conditioning on the expenditure of the region on the public good. See Section 6 for details.

\(^{11}\)This specification of spillovers is based on Oates (1972, Chapter 3). He in fact considers a model with symmetric information, \(n = 2\), regional preferences of the form \(u(c, g) = u(c, g + \alpha e)\), \(\alpha < 1\).
regions. This is a very stylized way of capturing spillovers, but leads to clean and simple results.

We assume that \( u_{gg} > 0, u_{ss} < 0, \) and \( u_{s} \geq 0 \) i.e. the spillover effect, if it is present, is positive. We allow any resident to have risk-neutral \((u'(c) = 1, c \in \mathbb{R}_+)\) or (strictly) risk-averse \((u'(c) < 0, c \in \mathbb{R}_+)\) preferences over consumption of the private good.

We model income variation across regions as follows. The resident of region \( i \) is endowed with one unit of labour time which she supplies inelastically. One unit of labour produces \( \lambda' \) units of the private good and \( \lambda'/\rho' \) units of the public good. Note that the coefficients \( \lambda', \rho' \) fully describe any linear technology producing public and private goods, with labour as the only input. With this technology, the resident of \( i \) will have a wage (and income) of \( \lambda' \), measured in units of the private good. Consequently, in what follows, we refer to \( \lambda' \) as income. Also, we refer to \( \rho' \) as the cost of the public good (relative to the numeraire private good).

There are two alternative and equivalent\(^{12}\) interpretations of \( \lambda' \). In both, one unit of labour now only produces one (resp. \( \rho' \)) unit of the private (resp. public) good. The first alternative interpretation is that the resident of \( i \) is endowed with \( \lambda' - 1 \) units of the numeraire good, as well as one unit of leisure. In this case, the wage will be unity and the total value of the resident’s endowment is \( \lambda' \). The second is that the time endowment of agents differs across regions, so that \( \lambda' \) is the time endowment of region \( i \). This is the way that Bordignon et al. (1996) model income variation across regions.

Finally, the regional government imposes a lump-sum tax \( \tau' \) on the resident, measured in units of the private good. So, with any of the above interpretations of \( \lambda' \), the resident’s budget constraint is

\[
c' = \lambda' - \rho'
\]

(2)

2.1. Regional government

The regional government in region \( i \) raises revenue \( \tau' \) from its resident, and receives a grant of \( \tau' \) measured in units of the private good, from central government. So, the budget constraint of the regional government can be written

\[
\tau' = \rho' g'
\]

(3)

where \( \rho' g' \) is the cost (in units of the private good) of \( g' \) units of the public good.

Combining private and regional government budget constraints (2), (3), we see that the indirect utility of the resident, taking into account the regional budget constraint can be written

\(^{12}\)These other two interpretations are not equivalent to the interpretation of \( \lambda \) as a labour productivity parameter when the supply of labour is elastic, and a distortionary income tax is levied (see Section 7).
Assume for the moment that central government conditions \( \tau^i \) on \( g^i \); this assumption is discussed in Section 2.2. Then, regional government chooses \( g^i \) to maximize the utility of the resident (4), taking into account the dependence of \( \tau^i \) on \( g^i \). So, \( g^i \) solves;

\[
\frac{u_i(g^i,e^i,\theta^i)}{v_i(\tau^i - \rho^i g^i + \lambda^i)} = \frac{\phi_i' - d\tau^i}{dg^i}
\]

where the left-hand side of (5) is the marginal rate of substitution between public and private good within the region, and the right-hand side is the marginal rate of transformation faced by the regional government.

We have allowed regions to differ in three different dimensions: demand for the public good, cost of producing the public good, and income. This captures all possible differences between regions except differences in spillover effects. However, it is too difficult — and not very enlightening — to allow regions to differ in more than one dimension at once. So, we consider three different cases in what follows.

- **Regions differ only in demands for the public good:** the \((\theta^i)_{i=1}^n\) are random, and \(\lambda^i = \rho^i = 1, \; i = 1, \ldots, n\).

- **Regions only differ in cost of producing the public good:** the \((\rho^i)_{i=1}^n\) are random, and \(\lambda^i = \theta^i = 1, \; i = 1, \ldots, n\).

- **Regions only differ in income:** the \((\lambda^i)_{i=1}^n\) are random, and \(\theta^i = \rho^i = 1, \; i = 1, \ldots, n\).

The random variables in each case are assumed to be continuously distributed on the intervals \([\theta, \bar{\theta}] = \Theta, \; [\rho, \bar{\rho}] = \mathbb{R}, \; [\lambda, \bar{\lambda}] = \Lambda\). In each case, the random variables are identically and independently distributed across regions. We denote the density and distribution of any of \(\theta^i, \rho^i, \lambda^i\) by \(f\) and \(F\), respectively.

### 2.2. Central government

In this section, we discuss the behavior of central government under the assumption that demands for the public good, parameterized by \(\theta^i\), are variable, but the discussion applies equally to the other two cases.

The only activity of central government is making grants to regions. In doing so, it acts to maximise the expected utility of any region, subject to the optimal responses by regional government (5), budget balance, and participation constraints. The grant from central government to the government of region \(i\), \(\tau^i\), can
in principle, be conditioned on all observable variables. The demand parameters $\theta^i$ are private information by assumption, but it is natural to assume that central government can observe one or more of $r^i$, $g^i$ and so could condition $\tau^i$ on one or more of these.$^{13}$ It is shown in Section 6 that in the stochastic demand and income cases it does not matter whether the grant $\tau^i$ is conditioned on either (or both) of $g^i$ or $r^i$. So, for purposes of presentation, we focus on the case where $\tau^i$ is conditioned on the level of public good provision $g^i$.

At this point, it is convenient to move to a mechanism design perspective. The central government can be thought of as choosing for each region $i$ a pair $\tau^i, g^i$ conditional only on region $i$’s cost announcement $\hat{u}_i$ i.e.

$$g^i = g(\hat{\theta}^i), \tau^i = \pi(\hat{\theta}^i)$$ (6)

This mechanism must satisfy truth-telling and participation constraints, and budget balance. To formulate these constraints, it is helpful to consider the limiting case as $n$, the number of regions, becomes ‘large’ i.e. $n \to \infty$. In this case, by the weak law of large numbers, the empirical distributions of the $g^i$ and $\tau^i$ across regions approximate the theoretical distribution generated by (6) and the distribution of $\hat{\theta}$. Consequently,

$$e^i = \frac{1}{n} \sum_{j \neq i} g^i \simeq \frac{1}{\hat{\theta}} \int g(\theta) f(\theta) d\theta = e, \quad i = 1, \ldots n$$ (7)

So, from the point of view of any given region, $e^i$ is (approximately) fixed when $n$ is large.

As all regions are ex ante identical, it is natural to assume that the central government’s objective is the expected utility of the resident of any region $i$, taken with respect to the demand shock to his region, which may be written again by the law of large numbers as

$$EU = \int_{\hat{\theta}} \left[ u(g(\theta), e, \theta) + \nu(\tau(\theta) - g(\theta) + 1) \right] f(\theta) d\theta$$ (8)

where $e$ is defined in (7), and $g(\cdot), \tau(\cdot)$ in (6).

The government chooses $(\tau(\theta), g(\theta))$ to maximise (8) subject to truth-telling and

$^{13}$In the existing literature, a variety of assumptions are made about the conditioning of $\tau^i$ on observable variables. The papers by Bordignon et al. (1996) and Bucovetsky et al. (1998) focus on the case where the central government conditions $\tau^i$ on a distortionary labour tax $\tau^i$; Boadway et al. (1998) on the case where $\tau^i$ is conditioned on $g^i$.

$^{14}$In the case where $n$ is fixed at some finite value, $g^i, \tau^i$ must be conditioned on the entire vector of announcements $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$. This is because transfers across regions must sum to zero. For example, if $n=2$, then the transfer to region 1 must depend on whether region 2 announces a high or low value of $g^i$. 

feasibility constraints. The truth-telling constraints\textsuperscript{15} require that any region with demand parameter $\theta$ prefers to announce $\theta$ rather than some $\theta'$ for every possible value of $e$;

$$u(g(\theta),e,\theta) + \nu(\tau(\theta) - g(\theta) + 1) \geq u(g(\theta'),e,\theta) + \nu(\tau(\theta') - g(\theta') + 1), \text{ } \theta' \neq \theta \tag{9}$$

The first kind of feasibility constraint is the participation constraint. The participation constraint for any regional government specifies that it would rather participate in the federation, and receive some insurance via the grant system, rather than secede. As is well-known in the incentives literature, this constraint can be either ex ante (i.e. the regional government decides whether for not to participate before $\theta'$ is realized) or ex post (i.e. the regional government decides after $\theta'$ is realized). We impose the ex ante participation constraint for two reasons. Firstly, the decision whether to join or leave a federation is a costly and long-term one. Secondly, with an ex post participation constraint, any region would obtain informational rent from its knowledge of $\theta$; this will complicate the analysis considerably and obscure the issues we wish to analyze. Formally, the participation constraint is

$$EU \geq \int_{\theta} \max_{g(\theta)} u(g(\theta),e,\theta) + \nu(-g(\theta) + 1)f(\theta)d\theta \tag{10}$$

where $e$ is defined in (7) above, and $EU$ is defined above.

The second feasibility constraint is the budget balance condition which says that in aggregate, transfers must sum to at most zero. For large $n$, this can be written approximately as

$$\int_{\theta} \tau(\theta)f(\theta)d\theta \leq 0 \tag{11}$$

So, the grant allocation problem (GAP) for the government is to choose $(g(\theta),\tau(\theta))$ to maximize (8) subject to (9), (10) and (11). As is usual in the mechanism design literature, we assume that $g$, $\tau$ are piecewise continuously differentiable functions.\textsuperscript{16} We will also assume additionally that $g(\cdot)$ is everywhere continuous.

\textsuperscript{15}The truth-telling constraints in this problem are ‘approximately’ Bayes-Nash for large $n$: that is, if (9) holds, for any $e > 0$, there exists $n$ large enough so that any region can gain less than $e$ by deviating, given that all other regions tell the truth.

\textsuperscript{16}That is, only discontinuous at a finite number of points, and possessing left- and right-hand derivatives at these points.
Some comments are in order at this point. Firstly, by setting \( t(\cdot) = 0 \) in the GAP, central government can always replicate what any region could get by seceding. So, we can be sure that the participation constraint (10) will never bind, and consequently it is ignored in what follows.

Secondly, note that except in the exceptional case where \( \nu \) is linear, the government has two objectives in the GAP: internalization of spillovers, and insurance of the regions against adverse demand shocks. We refer to the special cases of a linear \( \nu \) as the pure spillover GAP and \( u_e = 0 \) as the pure insurance GAP, respectively, in what follows.

Finally, the solution to the GAP will be a set of pairs \((g^*(\theta), \tau^*(\theta))_{\theta \in \Theta}\) — called an allocation in what follows. Any non-linear grant \( \tau^*(g) \) that induces a ‘type-\( \theta \)’ region (i.e. a region with \( \theta' = \theta \)) to choose \( g^*(\theta) \) as the solution to (5) for all \( \theta \in \Theta \) is said to decentralize this allocation. With a continuous type space, the grant that decentralizes the optimal allocation is unique, and the slope of this grant at any \( \theta \in \Theta \) is \( d \tau / dg = (d \tau^* / d \theta) / (d g^* / d \theta) \).

3. Grant allocation when regions differ in demand for the public good

As is usual in this kind of analysis, we start by considering the allocation \((\hat{g}(\theta), \hat{\tau}(\theta))_{\theta \in \Theta}\) that the central government could achieve if it could observe the \( \theta' \) — we call this the full-information allocation. This allocation maximizes (8) subject to (11) only. The first-order conditions to this problem, after simple rearrangement, yield the following:

\[
\frac{u_e(\hat{g}(\theta), \hat{\tau}(\theta))}{v'(\hat{\tau}(\theta) - \hat{g}(\theta) + 1)} + \frac{E_{\theta} u_e(\hat{g}(\theta), \hat{\tau}(\theta))}{v'((\hat{\tau}(\theta) - \hat{g}(\theta) + 1)} = 1, \quad \theta \in \Theta \quad (12)
\]

\[
v'(\hat{\tau}(\theta) - \hat{g}(\theta) + 1) = \mu, \quad \theta \in \Theta \quad (13)
\]

where \( \mu \) is the multiplier on the budget constraint, \( E_{\theta} x = \int_{\theta}^{\hat{\theta}} x(\theta) f(\theta) d\theta \) for any function of \( \theta \), and \( \hat{e} = E_{\theta} \hat{g} \). The second equation, (13) is a standard insurance condition which says that consumption of the private good is the same, whatever \( \theta \). The first equation (12) is just the Samuelson condition: it says that the marginal rate of substitution of the public good for the private good in the region with demand parameter \( \theta \), plus the sum (or more precisely, integral) of marginal rates of substitution of the spillover \( \epsilon \) for the private good, must be equal to the marginal rate of transformation, unity.

It is well-known (see e.g. Oates (1972), who made the point in a similar model to this one) that the full-information allocation can be decentralized by a simple linear matching grant. It is useful to state this as a benchmark result to which we can refer.

**Proposition 1. (Oates) In the full-information case, the allocation \((\hat{g}(\theta), \hat{\tau}(\theta))_{\theta \in \Theta}\)**
can be decentralized by a grant system where a type-$u$ region is offered a grant 
\[
\tau_u(g) = T_u + \hat{m}g, \text{ with}
\]
\[
\hat{m} = \frac{E_u\mu_u(\hat{g}(\theta),\hat{c},\theta)}{v'(\hat{\tau}(\theta) - \hat{g}(\theta) + 1)} = \mu
\]
Note that $\hat{m}$ is independent of $\theta$, and that if there are no externalities ($u_e=0$), then the grant is lump-sum.

**Proof.** To see this, note from (5), using $\rho'=1$, that the first-order condition for $g$ implies a region sets
\[
\frac{u_g}{v'} = 1 - \frac{d\tau_u}{dg} \tag{14}
\]
Also, as $\tau_u(g) = T_u + \hat{m}g$, and the definition of $\hat{m}$, we have
\[
\frac{d\tau_u}{dg} = \hat{m} = \frac{E_u\mu_u}{v'} \tag{15}
\]
Combining (14) and (15), we get (12) as required. Finally, $T_u$ can always be set to achieve the full-insurance condition (10) i.e.
\[
v'(T_u + \hat{m}\hat{g}(\theta) - \hat{g}(\theta) + 1) = \mu
\]
all $\theta \in \Theta$. ■

The first question that arises is whether the full-information grant allocation is feasible in the grant allocation problem when the $\theta'$ are private information. Our first result indicates that the answer to this depends crucially on whether utility is linear or strictly concave in the grant.

**Proposition 2.** If residents are risk-neutral (i.e. $v'=1$), then the full-information allocation \((\hat{g}(\theta),\hat{\tau}(\theta))_{\theta \in \Theta}\) is feasible in the GAP. Furthermore, the full-information allocation can be decentralized by a matching grant as in Proposition 1 above, except that $T$ is independent of $\theta$. If residents are risk-averse, ($v'<0$) then the full-information allocation is not feasible in the GAP.

**Proof.** If $v'=1$, it follows from (13) that $\mu=1$, so the full-insurance condition (13) is automatically satisfied, *whatever* $T$. So, $T$ can be chosen independent of $\theta$ to satisfy the central government budget constraint i.e. $T = -\hat{m}E_u\hat{g}$. Moreover, as noted above, $\hat{m}$ is independent of $\theta$. So, the linear matching grant described in the previous proposition is feasible in GAP, even if the $\theta'$ are private information.

On the other hand, if $v$ is strictly concave, then (13) implies that $\hat{\tau} - \hat{g}$ is the same for all $\theta$. Also, from (12) and the properties of $u$, it is easy to check that $\hat{g}(\theta)$ is strictly increasing in $\theta$. So, if the allocation is \((\hat{g}(\theta),\hat{\tau}(\theta))_{\theta \in \Theta}\), any type-$\theta$ region
strictly prefers to announce $\theta' > \theta$, violating truth-telling. So, $(g(\theta), \hat{f}(\theta))_{\theta \in \Theta}$ is not feasible in GAP, as claimed.

This is a very useful benchmark result that tells us that asymmetric information only makes a difference to the grant allocation problem if and only if citizens are risk-averse i.e. if there is a role for the central government in providing insurance. To put it another way, if there is no role for insurance, the central government can use a simple linear matching grant to internalize spillovers between regions without distorting public goods supply, even if the $\theta'$ are private information of the regional governments.\footnote{Note that as asserted in the Proposition, the grant is exactly of the form that has been proposed e.g. by Oates (1972) and Boadway and Wildasin (1984), for the case where information is symmetric.}

We now turn to the more interesting case where $v$ is strictly concave. Note that the key assumption of $u_{g > 0}$ implies that the marginal rate of substitution between $g$ and $\tau$, namely
\[
\frac{\partial \tau}{\partial g} \bigg|_{\text{u+u const.}} = 1 - \frac{u_g(g, e, \theta)}{v'(g - \tau + 1)}
\]
is decreasing in $\theta$. This is known as the single-crossing condition in the mechanism design literature. This is shown in Fig. 1, where we plot two indifference curves (i.e. level sets of $u(g, \theta) + v(\tau - g + 1)$) for values $\theta$, $\theta'$ with $\theta' > \theta$. Note also that the point at which the slope of the indifference curve is horizontal, namely $1 = u_{g}/v'$, is the point at which the supply of the public good will be optimal from the regional point of view. This value of $g$, $\tilde{g}(\theta)$, is increasing in $\theta$, as shown in Fig. 1.

Now define the utility to a type-$\theta$ region when it is truth-telling:
\[
w(\theta) = u(g(\theta), e, \theta) + v(\tau(\theta) - g(\theta) + 1)
\]
It is well-known in this literature (e.g. Laffont and Tirole, 1993) that when the single crossing condition holds, the constraints (9) can be replaced by the following:
\[
\frac{dw}{d\theta} = u_{g}(g(\theta), e, \theta) \quad (16)
\]
\[
\frac{dg}{d\theta} = 0 \quad (17)
\]
for almost all $\theta \in \Theta$. Constraints (16) are called the local or first-order truth-telling constraints, and (17) are called the monotonicity constraints.

Also, we will assume that $u$ is additively separable in $g$ and $e$, i.e.
\[
u(g,e,\theta) = u(g, \theta) + h(e)
\]
with $h$ concave in $e$, in order to get sharp results. Note that in this case, $w_{e} = u_{e}(g, \theta)$ is independent of $e$. We can now write the GAP as follows:

$$
\max_{g(\theta), \tau(\theta)} \int [u(g(\theta), \theta) + h(e) + u(\tau(\theta) - g(\theta) + 1)] f(\theta) d\theta \text{ s.t.}
$$

$$
\int \tau(\theta) f(\theta) d\theta \leq 0 \quad (18)
$$

$$
\int g(\theta) f(\theta) d\theta = e \quad (20)
$$
\[
\frac{dw}{d\theta} = u_\phi(g(\theta), \theta) \tag{21}
\]

\[
\frac{dg}{d\theta} = 0 \tag{22}
\]

This can be reformulated as a standard optimal control problem with integral constraints, for which Hestenes’ theorem gives first order necessary conditions for an optimum.

A general analysis of this problem is complicated by the fact that unless \(u\) is linear in \(\tau - g\) (which is ruled out by assumption) it is impossible to give any conditions in terms of the data of the model which guarantee ‘no bunching’ i.e. that (22) does not bound.\(^{18}\)

However, even with the possibility of bunching, we can establish some general results. To put these results in context, note that if \(\hat{g}(\theta)\) maximises regional expected utility \(EU\) subject only to the budget constraint (11), then whatever \(\tau(\theta)\), \(\hat{g}(\theta)\) must satisfy

\[
u_\phi + h_\tau - \nu' = 0 \tag{23}
\]

Consequently, we say that \(g\) is efficiently supplied if (23) holds, or equivalently, \((u_\phi + h_\tau)/\nu' = 1\). This can be interpreted as a kind of Samuelson rule for public good provision, and motivates the following definitions. Say that the public good is oversupplied (resp. undersupplied) if \((u_\phi + h_\tau)/\nu'\) is less than (resp. greater than) unity. Then we have:

**Proposition 3.** The solution to GAP has the following properties. First, there is incomplete insurance (\(\nu'\) is increasing in \(\theta\)). Second, in the pure insurance GAP \((h_\tau = 0)\), the public good is oversupplied \((u_\phi/\nu' < 1)\) except when there is no bunching at an endpoint of the distribution of \(\theta\) (i.e. \(\theta = \theta\) or \(\bar{\theta}\)) in which case there is efficient supply at this endpoint. Third, in the general case \((h_\tau > 0)\), if there is no bunching at \(\theta\), there exists a \(\theta < \theta < \bar{\theta}\) such that the public good is oversupplied \(((u_\phi + h_\tau)/\nu' < 1)\) if \(\theta > \theta\), and undersupplied \(((u_\phi + h_\tau)/\nu' > 1)\) if \(\theta < \theta\).

**Proof.** See Appendix. ■

The results on public good provision are illustrated in Fig. 2 for the no-bunching case.

\(^{18}\)In principal-agent problems where the agent’s utility is linear in income, rather than concave as here, simple sufficient conditions for no bunching can be given for many problems. For example, in the problem of regulation of a monopolist with unknown cost \(\theta\), monotonicity of the hazard rate of \(F\), the distribution function of cost, is sufficient (see e.g. Laffont and Tirole, 1993, p. 66).
The intuition for these results is best understood by considering the two-type case i.e. where \( \theta \) can take on only two values \( \theta_h, \theta_l \) with \( \theta_l < \theta_h \). First, in this case, it is easily shown\(^{19}\) that in the solution to GAP, the binding incentive constraint is the ‘upward’ one. Formally, if \((g_h^*, \tau_h^*)\) is the public good/grant pair allocated to a type \( i = h,l \) region at the solution to GAP, then a type-\( l \) region is indifferent\(^{20}\) between \((g_h^*, \tau_h^*)\) and \((g_l^*, \tau_l^*)\). The reason for this is that (as established above) the full-information allocation gives all types of regions the same level of consumption of the private good, but more of the public good to type-\( h \) regions. So, a type-\( l \) region prefers the full-information allocation of a type-\( h \) region to its own.

With two types, the analog of Proposition 3 in the pure insurance case is that there is oversupply of the public good in type-\( h \) regions, but efficient supply in type-\( l \) regions. The intuition for this is the usual one: by increasing \( g_h \) from its efficient level, central government can relax the binding incentive constraint. Specifically, suppose to the contrary that supply of the public good were efficient in type-\( h \) regions as shown in Fig. 3, where the binding incentive constraint is also shown.

\(^{19}\)A formal statement and proof of all the following claims about the two-type case is available on request from the author.

\(^{20}\)It can be shown that in the two-type case, there is no bunching, so \((g_h^*, \tau_h^*) \neq (g_l^*, \tau_l^*)\).
There, \( C_h = (\tau_h^*, g_h^*) \) and \( C_l = (\tau_l^*, g_l^*) \) are shown, both lying on an indifference curve of a type-\( l \) region, with an indifference curve of a type-\( h \) region passing through \( C_h \).

Then, from the figure, it is clear that \( g_h \) could be raised slightly at no first-order cost to type-\( h \) regions, moving \((\tau_h^*, g_h^*)\) to point \( C_h' \) in Fig. 3; as shown, such an increase would however, make the payoff to a type-\( l \) region from \((\tau_h^*, g_h^*)\) lower (as it values \( g \) less at the margin).

Why is loosening the incentive constraint valuable? It is valuable because there is incomplete insurance of the regions; as long as \( g_h^* > g_l^* \), the private consumption of type-\( h \) regions must be lower than that of type-\( l \) regions — otherwise, type-\( l \) regions would prefer to pretend to be type-\( h \). So, a loosening of the incentive constraint allows a small (balanced budget) increase in \( \tau_h \) and fall in \( \tau_l \) (shown as a move from \( C_h' \) to \( C_h'' \) in Fig. 3). This in turn, must raise the expected utility of any region from private consumption, as consumption is being shifted from a state of the world in which the region’s marginal utility of private consumption is low \( (\theta = \theta_l) \) to one where it is high \( (\theta = \theta_h) \).

We now turn to offer some intuition for two-way distortion when externalities are present, which is one of the main results of the paper. The analog of the two-way distortion result in the two-type case is that there is undersupply at \( \theta = \theta_l \) occurring simultaneously with oversupply at \( \theta = \theta_h \). The oversupply at \( \theta = \theta_h \) has already been discussed. To see the intuition for undersupply, we can first write out...
the binding incentive constraint in the GAP. As already discussed above, type-1 regions are indifferent between announcing $\theta_1$ and $\theta_2$, so, noting that $h(e)$ cancels from both sides, the binding incentive constraint is;

$$u(g^*_1, \theta_1) + \nu(1 + \tau^*_h - g^*_h) = u(g^*_h, \theta_1) + \nu(1 + \tau^*_h - g^*_h)$$  \hspace{1cm} (24)

Then, a small change in $g^*_h$, $\Delta$, will leave the right-hand side of (24) unchanged, but cause the left-hand side to change by

$$\delta = [u_g(g^*_h, \theta_1) - \nu'(1 + \tau^*_h - g^*_h)]\Delta$$  \hspace{1cm} (25)

Now suppose that $g^*_h$ is initially efficient i.e. solves

$$u_g(g^*_h, \theta_1) + h_\nu - \nu'(1 + \tau^*_h - g^*_h) = 0$$  \hspace{1cm} (26)

Then, from (25), (26) the change in the left-hand side of the incentive constraint (24), $\delta$, reduces to

$$\delta = - h_\nu \Delta$$

So, as $h_\nu > 0$ by assumption, if $\Delta < 0$ then the term on the left-hand side of the incentive constraint (24) will increase. In other words, decreasing $g^*_h$ by $\Delta$ from its efficient level will relax the incentive constraint. In turn, this is valuable, because (as explained above) it allows reallocation of private consumption between the types. Proposition 3 above describes the allocation $(g^*(\theta), \tau^*(\theta))_{\theta \in \Theta}$ that solves GAP. It is also possible — and definitely of interest — to say something about the non-linear grant $\tau^*(g)$ that decentralizes this allocation. The benchmark is the full-information case, where, as already noted in Proposition 1, the Pigouvian grant that decentralizes efficient public goods supply is a simple linear matching grant $\tau_u(g) = T_q + n g$, where $n = h_{\nu'}/\nu'$ is simply the value (measured in units of the numeraire private good) of the spillover from $g$.

In the asymmetric information case, $\nu'$ varies across regions, and so the expected value across all regions, of the spillover from any region, measured in units of the numeraire, is $E_{\theta} [h_{\nu'}/\nu']$. The key feature of the grant $\tau^*(g)$ that decentralizes the solution to GAP is that the increase in the grant following an increment in $g$ exceeds this spillover $E_{\theta} [h_{\nu'}/\nu']$:

**Proposition 4.** The grant $\tau^*(g)$ that decentralizes the allocation $(g^*(\theta), \tau^*(\theta))_{\theta \in \Theta}$ is a non-linear increasing function of $g$, almost everywhere differentiable, with slope

$$\frac{d\tau^*}{dg} = E_{\theta} [h_{\nu'}/\nu'] + \Phi(g)$$

where $\Phi(g) > 0$ if $g \in (g^*(\underline{\theta}), g^*(\bar{\theta}))$, and $\Phi(g) = 0$ if $g = g^*(\theta), g^*(\bar{\theta})$. 
Proof. See Appendix. ■

So, following an increment in $g$, a region gets an additional payment greater than that needed to internalise the spillover. The intuition for this follows from the oversupply result. To loosen the incentive constraints, oversupply is optimal, so regions must be induced to supply more $g$ than that level that internalizes the spillover.

4. Grant allocation when regions differ in costs

Here, the costs of production of the public good are variable, and private information, in the different regions. In this case, the GAP has a rather complex structure. It turns out that it is convenient to analyse the GAP under the assumption that the central government can observe the expenditure on the public good, $G = \rho' g'$, rather than the physical output. Then, we may write regional utility as

$$u^i = u(G/\rho', e^i) + v(\tau^i - G^i + 1)$$

(27)

where $e^i = \sum_{j \neq i} G_j/\rho(n - 1)$. This has the same form as (4), the payoff to a region when demands are stochastic, with $1/\rho'$ replacing $\theta'$, and $G'$ replacing $g'$. This suggests that the results of the previous section will carry over to the stochastic cost case.

However, matters are not so straightforward, as the key single crossing assumption that $u$ is single-signed must be checked in this set-up. Let $-g u_g / u_g = e u_g$ be the elasticity of marginal utility from the public good $g$, which measures the rate at which marginal utility declines. It is then easy to show that

$$u_{Gp} \leq 0 \iff 1 \geq e u_g$$

(28)

To see this, note that $u' = u(G/\rho)/\rho$, so $u_{Gp} = -(u_g + g u_g)/\rho^2$.

The equation (28) has several striking implications. First, without spillovers, the first-best allocation satisfies

$$\frac{u_{Gp}(\hat{G}/\rho)}{v'((\hat{\tau} - \hat{G} + 1) = 1, \rho \in R}$$

(29)

$$v'((\hat{\tau} - \hat{G} + 1) = \mu, \rho \in R$$

(30)

21In general it makes a difference whether $g'$ or $G'$ is observable by the principal (central government), as shown by Maskin and Riley (1985). Investigation of this issue is a topic for further work.

22Note that $e u_g$ is also the coefficient of relative risk-aversion (CRRA) of utility over $g$. 
So, from (29), (30), the sign of $\frac{\partial G}{\partial \rho}$ is given by the sign of $u_{G\rho}$.
Consequently, from (28), $\hat{G}$ will fall with $\rho$ (resp. rise with $\rho$) if $\varepsilon u_g$ is less than 1 (resp. greater than 1). The intuition for this is as follows. The insurance objective of the central government is to reallocate income to states in which the marginal utility of private consumption, $u'$, is high. These are low-$\rho$ (resp. high-$\rho$) states if and only if the price-elasticity of demand for the public good is greater than (less than) one. Given the separability in private and public consumption, the price-elasticity of demand for the public good is greater than (less than) one as $\varepsilon u_g$ is less than (greater than) one.

Interestingly, in the borderline case where $\varepsilon u_g = 1$ for all $g$, then $\hat{G}$ is constant in $\rho$, and so the first-best involves from (30), $\hat{\tau}$ constant in $\rho$, which in turn implies $\hat{\tau}(\rho) = 0$, all $\rho$, from the budget constraint. It follows immediately that in this case, the full-information allocation must be incentive-compatible: no region can get a different allocation by misreporting $\rho$.

Summarising:

**Proposition 5.** In the pure insurance case ($u_s = 0$), assume that $\varepsilon u_g = 1$ for all $g$ (i.e. $u(g) = \kappa \ln g$, $\kappa > 0$). Then the full-information allocation is feasible in the GAP. This allocation gives every region the same level of expenditure on public goods, and a zero grant.

Now consider the case where $\varepsilon u_g \neq 1$. For simplicity, we again restrict attention to the pure insurance case. If $\varepsilon u_g < 1$ (i.e. weakly diminishing marginal utility of $g$), we can set $\theta = 1/\rho$, and then from (28), $u_{G\rho} > 0$ everywhere, and so, the results of the previous section carry over straightforwardly. That is, expenditure on public goods $G^*$ (weakly) decreases with cost $\rho$, and from Proposition 3, $G$ is oversupplied relative to the full-information allocation.

On the other hand, if $\varepsilon u_g > 1$ (i.e. strongly diminishing marginal utility of $g$), then we can set $\theta = -1/\rho$, and then from (28), $u_{G\rho} > 0$ everywhere, and so, the results of the previous section carry over in reverse. That is, expenditure on public goods $G^*$ (weakly) increases with cost $\rho$, and from Proposition 3, $G$ is undersupplied relative to the full-information allocation.

So, we may summarise:

**Proposition 6.** Assume the pure insurance case ($u_s = 0$). If $\varepsilon u_g < 1$, all $g$, then at the solution to GAP, expenditure on the public good $G^*$ (weakly) decreases with cost $\rho$, and $G^*$ is oversupplied relative to the full-information allocation. If $\varepsilon u_g > 1$, all $g$, then at the solution to GAP, expenditure on the public good $G^*$ (weakly) increases with cost $\rho$, and $G$ is undersupplied relative to the full-information allocation.

To conclude, a striking and non-standard feature of the cost GAP is that the qualitative features of the solution depend on the curvature of the sub-utility
function over the public good. The intuition for this has already been touched on above; the size of \( \epsilon u_d \) determines the price elasticity of demand by regions for the public good, which in turn determine whether the central government wishes to allocate income to states where \( \rho \) is high, or where \( \rho \) is low.

5. Grant allocation when regions differ in income

In this section, we consider the case where the income parameter differs between regions. In this case, regional utility is

\[
 u' = u(g',e) + \nu(\tau' - g' + \lambda')
\]

Also, note that with \( \nu \) strictly concave, the marginal rate of substitution between \( \tau \) and \( g \), \( 1 - (u,\nu') \) is monotonically decreasing in \( \lambda \), so regional preferences satisfy the single-crossing condition. The first-best allocation \( (\tilde{g}(\lambda),\tilde{\tau}(\lambda))_{\lambda \in \Lambda} \) is defined by equations similar to (12), (13), i.e. a Samuelson condition and a full insurance condition. Also, Propositions 1 and 2, suitably modified apply to this case as well. So, again the case of interest is when regions are strictly risk-averse with respect to private consumption.

In this case, the GAP with stochastic income is formulated as follows. Again, we assume that utility is separable in \( g \) and \( e \) i.e. \( u(g,e) = u(g) + h(e) \). Also, let \( w(\lambda) \) be the utility of a truth-telling region of type \( \lambda \in \Lambda \). Then the GAP can be written:

\[
g(\lambda),\tau(\lambda) \int_{\Lambda} [u(g(\lambda)) + h(e) + \nu(\tau(\lambda) - g(\lambda) + \lambda)] f(\lambda) d\lambda \text{ s.t.} \quad (31)
\]

\[
\int_{\Lambda} \tau(\lambda) f(\lambda) d\lambda \leq 0 \quad (32)
\]

\[
\int_{\Lambda} g(\lambda) f(\lambda) d\lambda = e \quad (33)
\]

\[
\frac{dw}{d\lambda} = \nu'(\tau(\lambda) - g(\lambda) + \lambda) \quad (34)
\]

\[
\frac{dg}{d\lambda} = 0 \quad (35)
\]

Note that (34) and (35) are the local truth-telling conditions and monotonicity.
Proposition 7. The solution to GAP has the following properties. In the pure insurance GAP ($h_s = 0$), the public good is undersupplied $(u_i'/v'>1)$ except when there is no bunching at an endpoint of the distribution of $\lambda$ (i.e. $\lambda = \lambda$ or $\lambda$) in which case there is efficient supply at this endpoint. Second, in the general case ($h_s > 0$), if there is no bunching at $\lambda$, there exists a $\lambda < \lambda < \lambda$ such that the public good is oversupplied $((u_i + h_s)/v' < 1)$ if $\lambda < \lambda$, and undersupplied $((u_i + h_s)/v' > 1)$ if $\lambda > \lambda$.

Proof. See Appendix.

Proposition 7 is illustrated in Fig. 4 for the no-bunching case. The results are qualitatively exactly the reverse of the stochastic demand case. With no spillovers, there is now undersupply of the public good, rather than oversupply. With spillovers, there is now oversupply at $\lambda < \lambda$, and undersupply at $\lambda > \lambda$.

These differences with the stochastic demand case can be explained intuitively by reference to the two-type GAP i.e. with only two types $\lambda < \lambda$. In the full-information solution to the two-type GAP, both types of regions are given the

![Diagram](image)

Fig. 4. Distortions in public goods supply when income parameter $\lambda$ is variable.
same amount of the public good, and of the private good, which implies a smaller grant for rich \((\lambda, l)\) regions. So, the full-information solution is not incentive compatible as rich regions will declare themselves to be poor. This fact implies that at the solution to GAP, it is now the downward incentive constraint that binds at the solution i.e. type-\(h\) regions are now indifferent between \((g^*, \tau^*)\) and \((g^h, \tau^h)\). In the pure insurance case, this incentive constraint can then be loosened by reducing \(g^h\) below its efficient value. Moreover, the intuition for the two-way distortion result is the same as that given above for the stochastic demand case.

As in the variable demand case, we can also say something about the grant \(\tau^h(g)\) that decentralizes the allocation described in Proposition 7. The key feature is that the slope \(d\tau^h/dg\) is now less than the value of the spillover from \(g\), and the intuition is the same as in the variable demand case.

6. Expenditure vs. revenue-related grants

So far we have assumed that the central government conditions the grant on the expenditure of regional governments on the public good (in the stochastic income and demand cases, the relative price of the public good is unity, so the physical quantity of the public good produced, \(g\), is also the expenditure). The obvious alternative is to condition the grant on revenue raised by regional governments. It is important to investigate whether our results are robust to this alternative.

In fact, our results are robust in this sense, and this can be shown as follows. Let the G-GAP and r-GAP be the problems where grants are conditioned on \(G\) and \(r\) respectively (recall that \(G = pg = g\) in the variable demand case). Say that G-GAP and r-GAP are equivalent if at the solution to G-GAP, every region with a given \(u\) has the same consumption of the public and private good as it has at the solution to r-GAP. Then we have;

**Proposition 8.** If demands \(\theta\) or incomes \(\lambda\) are variable, the G-GAP and r-GAP are equivalent.

**Proof.** For brevity, we focus only on the case with \(\theta\) variable. First (using the fact that \(G = g\) in the variable demand case), inspection of the budget and truth-telling constraints reveals that \((G(\theta), \tau(\theta))\) is feasible in G-GAP if \((\tau(\theta), \tau(\theta))\) is feasible in r-GAP, with \(\tau(\theta) = G(\theta) - \tau(\theta)\).

So, G-GAP and r-GAP can only not be equivalent if \((G^*(\theta), \tau^*(\theta))\) solves G-GAP, but there exists some \((\tau'(\theta), \tau'(\theta))\) with \(\tau'(\theta) \neq \tau^*(\theta)\) that solves r-GAP. However, if this were the case, \((G'(\theta), \tau'(\theta))\) would be feasible in G-GAP and yield higher utility than \((G^*(\theta), \tau^*(\theta))\), contrary to assumption. ■

\(^{23}\) The details are omitted to save space; a full discussion can be found in an earlier version of this paper, Lockwood (1997).
Two further points are worth making. First, it is clear from the above discussion that the central government could not do any better if it could condition on both of $r$, $G$, rather than one or the other, so the equivalence result extends to this case. Second, it immediately follows that in the variable income and demand cases, the $g$-GAP (where the grant is conditioned on the physical output of the public good, $g$) is also equivalent to the G-GAP and the $r$-GAP. However, this remark does not apply to the variable cost case, for the following reason. The proof of Proposition 8 makes use of the fact that the link between $r$ and $G$ does not depend on any variable that is the private information of the region. In the case of variable costs, the link between revenue and output, $r = p g - \tau$, now depends on $p$, which is private information of the regions, so the above equivalence result no longer holds.

7. Distortionary taxes

This paper has focused on the basic issues by assuming a simple set-up where taxes are lump-sum. This is, however, a strong assumption, and it is desirable to see whether the results obtained so far are robust to the introduction of distortionary taxes. One simple way of doing this is to suppose that labour supply by residents in a region is elastic, and so a wage or consumption tax becomes distortionary (Bordignon et al., 1996; Boadway et al., 1998).

In this section, we extend our model to the case of elastic labour supply. We then show that under assumptions that are satisfied (for example) when preferences over the private good and leisure are Cobb-Douglas, the results in the variable demand and cost cases are robust: all the results obtained above for the case where regions are risk-averse ($y^0, 0$) extend immediately to the case of distortionary taxes.

However, in the case of variable incomes, the picture is less clear. This section shows that if we model ‘income’ shocks as in Bordignon et al. (1996), the results of Section 5 may be qualitatively reversed.

Distortionary taxes are introduced by supposing that leisure enters the representative resident’s utility, so that labour supply becomes elastic. Specifically, suppose that the resident in the region $i$ now has utility

$$u^i = u(g^i, e^i, \theta^i) + \omega(c^i, l^i)$$

where $\omega(c^i, l^i)$ is defined over consumption and leisure, $l^i$.

The formulation of the budget constraint (and indeed, the results obtained below), depend crucially now on how we interpret $\lambda^i$. As we saw in Section 2 above, there are three possible interpretations of $\lambda^i$, all of which are only equivalent with inelastic labour supply. For purposes of comparison to Bordignon et al. (1996), we now interpret $\lambda^i$ as the time endowment of the representative resident of region $i$. With this interpretation, the budget constraint is
The resident maximizes \( w(c', l') \) subject to (37). This implies indirect utility over private good and leisure of \( \Omega(t', \lambda') \) and labour supply of \( L' = L(t', \lambda') \).

We can now express regional utility over public good and transfer pairs \((g', \tau')\) in a form similar to (4) by using the following trick. Let \( t(r, \lambda) \) be the smallest tax rate \( t \) that solves

\[
r = tL(t, \lambda)
\]

for a given value of \( r \) i.e. \( r(t, \lambda) \) is the smallest tax rate that yields revenue \( r \). Then noting that \( r = \rho g - \tau \), we can write indirect utility over \((\tau, g, \lambda)\) as

\[
V(\tau - \rho g, \lambda) = \Omega(\rho g - \tau, \lambda, \lambda)
\]

We may now write the utility for region \( i \) as

\[
u_i = u(g_i, e_i, \theta_i) + V(\tau - \rho g_i, \lambda_i)
\]

Note the similarity between (39) and (4). The analysis of the GAP is now exactly as before, with \( V \) replacing \( v \). So, to show that results extend to the distortionary tax case, we only need show that \( V \) has the same qualitative properties as \( v \). In the interesting case where \( v \) is strictly concave, we have by assumption that \( v_{\tau} > 0, v_{g} < 0, \) and \( v_{\lambda} = v_{\tau} < 0 \) i.e. regional utility is strictly concave in the grant, and the marginal utility with respect to the grant is decreasing in the region’s own income. Some, but not all, of the properties of \( v \) may carry over to \( V \), as the following Proposition shows. Suppose that

\[
\omega(c, l) = c^{1-\beta} - \alpha l^{\alpha}, \quad \alpha + \beta < 1, \quad \alpha, \beta > 0
\]

Then \( \beta \) parameterizes the degree of concavity of indirect utility in income \((\beta = 0 \text{ risk-neutrality, and } \beta > 0 \text{ risk-aversion})\) and \( \alpha \) parameterizes the (compensated) elasticity of labour supply \((\alpha = 0 \text{ inelastic labour supply, } \alpha > 0 \text{ elastic})\). We then have the following result.

**Proposition 9.** If \( \omega \) is Cobb-Douglas i.e. as in (40), then

\[
V(\tau - \rho g, \lambda) = \kappa \left( \lambda \left( 1 - \frac{\alpha}{1 - \beta} \right) + \tau - \rho g \right)^{1-\beta-a} \lambda^\alpha, \quad \kappa > 0
\]

and consequently \( V_\tau > 0, V_\tau < 0 \). Also,

\[
\alpha > (1 - \beta) \beta \psi \Rightarrow V_{\psi} > 0
\]

\[
\alpha < 1 - \beta + \beta \psi \Rightarrow V_{\lambda} < 0
\]

where \( \psi = \lambda/r > 1 \).

**Proof.** See Appendix.
A number of observations can be made at this point. First, Proposition 9 tells us that in the Cobb-Douglas case, $V$ is always an increasing, strictly concave function of $\tau$. By Proposition 2 above, this is sufficient for the GAP to be non-trivial i.e. for the full-information allocation not to be decentralisable. However, $V_{\lambda}$ may have the opposite sign to the sign of $V_{\lambda}$ when taxes are lump-sum, which (as we see below) will have implications for the properties of the GAP. In particular, if $\alpha=0$ and $\beta>0$, then we are back in the case of lump-sum taxation, and from (41), $V_{\lambda}<0$. On the other hand, if risk-aversion $\beta$ is low, and/or the compensated elasticity of labour supply $\alpha$ is high, then from (41), $V_{\lambda}>0$ i.e. an additional unit of income may actually increase the marginal utility from the grant.

The intuition for this is that with elastic labour supply, an increase in $\lambda$ has two effects on $V$. First, it increases the tax base, and hence tax revenue for the region, and so is like an increase in $\tau$. As $V_{\tau}<0$, this indicates that this first effect tends to make $V_{\lambda}$ negative. However, an increase in $\lambda$ increases spending on leisure, and hence increases both $V$ and $V_{\tau}$. The second effect tends to make $V_{\lambda}$ positive, and dominates if labour supply is highly elastic.

Having established some properties of $V$, at least in the Cobb-Douglas case, we now turn to analyse the GAP with distortionary taxation. It is convenient first to discuss briefly the full-information solution to the GAP with distortionary taxes. To lighten the notational load, and for ease of comparison to Section 3, we focus on the case where regions vary in $u$ only (i.e. $r=l=1$). The first, and obvious point is that the full-information allocation of public goods, private goods and leisure is now not the first-best allocation, as there are distortionary taxes.

In fact, the conditions that define the full-information allocation can be written exactly as in (12), (13), except that the marginal utility of private consumption, $V_{\tau}$, is replaced with the marginal utility to regional government of an additional unit of tax revenue, $V_{\tau}$. We can rewrite the equivalent of (9), omitting the arguments of the derivatives for clarity, as

$$
\frac{\mu_e}{\Omega_e} + \frac{E_\theta \mu_c}{\Omega_\lambda} = \frac{V_{\tau}}{\Omega_{\lambda}}
$$

(42)

where $\Omega_{\lambda}$ is the resident’s marginal utility of income (evaluated at $\lambda=1$). So,

---

24 The Proposition provides sufficient, not necessary conditions for $V$ concave in $\tau$; so concavity may hold under much more general conditions than identified here.

25 One intuition as to why $V$ is strictly concave in $\tau$ is as follows. The higher the grant $\tau$, the lower $t$ needs to be at a given $g$. So, as the deadweight loss of the tax increases with the square of $t$, a given increase in the grant when the grant is low (and $t$ is high) allows a bigger reduction in deadweight loss, and therefore a bigger increment in utility, than the same increase in the grant when the grant is high (and $t$ is low).
\(V_i/\Omega_\lambda\) is the marginal cost of public funds,\(^{26}\) and (42) is just the standard second-best Samuelson rule,\(^{27}\) which says that the sum of marginal rates of substitution between public and private good is equal to the marginal cost of public funds (see e.g. Atkinson and Stiglitz, 1980).

Given these differences in the interpretation of the full-information allocation, our previous results in the variable demand and cost cases extend straightforwardly.

**Proposition 10.** Suppose that \(\omega(c, l)\) is Cobb-Douglas (or any other conditions sufficient for \(V_i>0, V_{\tau l}<0\)). Then the results in Propositions 3–6 concerning the GAP when the resident is risk-averse \((\omega<0)\) extend unchanged to the case of distortionary taxes.

**Proof.** By inspection, the relevant parts of the proofs of Propositions 3–6 require \(v_i>0, v_{\tau l}<0\). By Proposition 9, these properties are also true of \(V\), so Propositions 3–6 must carry over directly to the distortionary case where regional utility is \(u(g, e, \bar{h}, \lambda)\).

However, the results for the case where ‘income’ \(\lambda\) is stochastic may not extend directly in this way. The reason is that as discussed above, with fixed labour supply, \(v=\omega(\tau-g+\lambda)\), so \(v_{\tau k}=v_{\tau l}<0\) i.e. the marginal utility of the grant is decreasing in an additional unit of time endowment due simply to concavity of utility. When preferences over leisure and consumption are Cobb-Douglas, however, as described in Proposition 9, it is possible that \(V_{\tau k}>0\) when \(\beta\) is small relative to \(\alpha\).

**Proposition 11.** Suppose that \(\omega(c, l)\) is Cobb-Douglas and that income shocks arise from shocks to time endowments. If the labour supply elasticity of the representative resident is small enough relative to risk-aversion \((\alpha<(1-\beta)\beta)\), then Proposition 8 continues to hold with distortionary taxes. If however, the resident is risk-neutral and supplies labour elastically \((\alpha>\beta=0)\), then in the solution to the pure insurance GAP (i.e. with \(h_\tau=0\)), the public good is now oversupplied \((u_g/V_i<1)\) for all \(\lambda \in (\bar{\lambda}, \bar{\Lambda})\).

\(^{26}\)Generally, \(V_i/\Omega_\lambda > 1\). In the Cobb-Douglas case, it is easy to check that if \(r>0\),

\[
V_i/\Omega_\lambda = \frac{(1 - \frac{\alpha}{1 - \beta}) \lambda}{1 - (1 - \frac{\alpha}{1 - \beta}) \lambda - r} > 1.
\]

\(^{27}\)Also, note that the full-insurance condition \(V_i = \mu\) does not now imply that residents’ marginal utility of income \(\Omega_i\) is equalized across regions, but that the marginal utility of the grant, \(V_i\), is equalized; with distortionary taxation, these are two different things.
Proof. Using the fact that $\psi > 1, \alpha < (1 - \beta)\beta$ implies, from (41), that $V_{\tau \lambda} < 0$, all $g, \tau, \lambda$. Consequently, Proposition 8 carries over. If, on the other hand, $\alpha > \beta = 0$, then again from (41), $V_{\tau \lambda} > 0$, all $g, \tau, \lambda$. Then the distortions described in Proposition 8 are reversed.

An important conclusion, therefore, from Proposition 11 is that with incomes varying across regions, the qualitative properties of the GAP may depend crucially on whether taxes are lump-sum or distortionary. This result also relates to Bordignon et al. (1996). There, agents are risk-neutral, supply labour elastically, and vary in income in precisely the same way as in this section. As discussed in more detail in Section 8, they find that where distortion occurs, it is in the direction of oversupply of the public good, consistently with Proposition 11. So, Proposition 11 indicates that their results must be sensitive to the degree of risk-aversion over private consumption of the residents of a region — if this is high enough, their results will be reversed.

8. Conclusions and related literature

The main results of this paper are summarized in Table 1. Some general themes emerge from the analysis. Perhaps the most novel is that externalities between regions (agents) induce two-way distortion i.e. qualitatively different distortions in public goods supply at the endpoints of the distribution of types, which cannot occur in the standard principal-agent problem.

Secondly, (ignoring spillovers), the source of the shock hitting the regions, matters for the qualitative properties of the optimal grant. For example, with

Table 1
Summary of results

<table>
<thead>
<tr>
<th></th>
<th>Stochastic demand</th>
<th>Stochastic cost</th>
<th>Stochastic income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure spillover case</td>
<td>Full information allocation</td>
<td>Full information allocation</td>
<td>Full information allocation</td>
</tr>
<tr>
<td></td>
<td>feasible in GAP, decentralisable by a linear matching grant</td>
<td>feasible in GAP</td>
<td>feasible in GAP, decentralisable by a linear matching grant</td>
</tr>
<tr>
<td>Pure insurance case</td>
<td>Incomplete insurance, oversupply of public good except at endpoints of distribution of $\theta$</td>
<td>As in stochastic demand case if elasticity of marginal utility with respect to the public good is less than 1</td>
<td>Incomplete insurance, undersupply of public good for $\lambda &lt; \hat{\lambda}$ and oversupply for $\lambda &gt; \hat{\lambda}$</td>
</tr>
<tr>
<td>General case</td>
<td>Incomplete insurance, undersupply of public good for $\theta &lt; \hat{\theta}$ and oversupply for $\theta &gt; \hat{\theta}$</td>
<td>As in stochastic demand case if elasticity of marginal utility with respect to the public good is less than 1</td>
<td>Incomplete insurance, oversupply of public good for $\lambda &lt; \hat{\lambda}$ and undersupply for $\lambda &gt; \hat{\lambda}$</td>
</tr>
</tbody>
</table>

*This result is for the case where taxes are lump-sum. With distortionary taxes, we may have oversupply.
lump-sum taxes, (absent spillovers), there is oversupply with demand shocks, and undersupply with income shocks. Overall, there can be no presumption that asymmetric information will lower public good supply.

Thirdly, at least when preferences over the private good and leisure are Cobb-Douglas, the structure of regional taxation (i.e. lump-sum or distortionary) may matter for the qualitative properties of the optimal grant. For example, if regions differ in incomes then (in the pure insurance case), with lump-sum taxes, there is undersupply of the public good, but with distortionary taxes, there may be oversupply.

Fourthly, in the case where regions vary in the cost of producing the public good, the structure of the optimal grant is quite subtle, and depends on the elasticity of demand for the regional public good by its residents, as measured by the curvature of the sub-utility function over the public good.

With a local public good spillover, the picture is more complicated; whether taxation is lump-sum or distortionary, the source of the shock will matter for the nature of the distortions in public good supply at the top and bottom of the type distribution.

These results can be compared to the existing literature as follows. The papers closest to this one are Bordignon et al. (1996) — BMT — and Bucovetsky et al. (1998) — BMP. The BMT model is a two-type, two-region version of our stochastic income model with elastic labour supply, where there are no externalities between regions, the resident’s utility is assumed linear in consumption, income is modelled in the same way as in Section 7 of this paper, income shocks are independent, and where the grant is conditioned on the (labour) tax rate. One of their main results is that with asymmetric information about incomes, the tax rate in a high-income region is undistorted (whether the other region has high or low income) but the tax rate in a poor region is too high (whatever the income of the other region) if the relevant incentive constraint binds. It follows from the fact that the transfer is increasing in the tax that the public good is oversupplied in the low-income region if the relevant incentive constraint is binding (Proposition 4 in their paper). These distortions are qualitatively the same as those in Proposition 8, and so Proposition 8 can be regarded as an extension, in the direction of a continuum of types, of their Proposition 4.

BMP consider a two-type, many-region version of our stochastic demand model, with the additional complication that the tax base, capital, is interregionally mobile. The most relevant comparison is with their Proposition 4, where it is assumed (effectively) that the central government maximizes the expected utility of a typical region prior to the demand shock and that regional utility is linear in income. In our model, linearity of income implies no distortion in public good

28They also extend their model to allow for endogenous tax enforcement by regions, allowing for an element of moral hazard, so the relevant comparison is to the basic version of their model, without moral hazard.
provision, whereas they do find distortions; low-demand regions oversupply the public good and high-demand regions undersupply. So, they find endpoint distortion even with quasi-linear preferences, something that cannot occur with public good spillovers, and so must be due to the nature of the tax externalities in their model.

Acknowledgements

Early versions of this paper were presented at the Mini-Workshop on International Taxation, EPRU, Copenhagen, March 14–16th, 1996, and the HCM Workshop, EPRU, May 3rd–4th, 1996. I would like to thank conference participants, especially Massimo Bordignon, Sam Bucovetsky, Maurice Marchand, Pierre Pestieau, and Wolfgang Peters, and an editor and a referee for helpful comments.

Appendix

Proof of Proposition 3

(i) First, we prove the incomplete insurance result. First, the fact that \( g \) is (weakly) increasing in \( \theta \) and truth-telling require that \( \tau - g \) is weakly decreasing in \( \theta \) (otherwise, \( \theta \)-regions would announce \( \theta' > \theta \)). Consequently, private consumption \( c = \tau - g + 1 \) is (weakly) decreasing in \( \theta \), so from strict concavity of \( \nu(c) \), \( \nu'(\tau - g + 1) \) is increasing in \( \theta \).

(ii) We now restate the GAP of Section 3 in a slightly different form as a standard optimal control problem. The problem is to choose piecewise continuous control variables \( \tau(\theta), z(\theta) \) and a number \( e \) to maximise

\[
\int_{\theta}^{\bar{\theta}} [w(\theta) + h(\theta)] f(\theta) d\theta
\]

subject to the following constraints

\[
w(\theta) = u(g(\theta), \theta) + \nu(\tau(\theta) - g(\theta) + 1)
\]

\[
z(\theta) \geq 0
\]

\[
- \int_{\theta}^{\bar{\theta}} \tau(\theta) f(\theta) d\theta \geq 0
\]
\[
\int \limits_\theta (g(\theta) - e) f(\theta) d\theta = 0
\]

and the state equations:

\[
\frac{dw}{d\theta} = u(\theta, g(\theta), \theta) \quad (A.1)
\]

\[
\frac{dg}{d\theta} = z(\theta) \quad (A.2)
\]

Note that we treat \( g \) as a state variable, and introduce the new control variable \( z \), in order to deal with the monotonicity constraint, following Fudenberg and Tirole (1993). Also, \( w(\theta) \) is now defined as utility net of the spillover \( h \).

(iii) Stated in this way, the GAP is an optimal control problem with integral and inequality constraints. Hestenes’ Theorem gives necessary conditions for a solution to this problem, which can be conveniently derived using the generalized Hamiltonian (Takayama, 1985). This Hamiltonian is

\[
H = (w(\theta) + h(e)) f(\theta) + \mu_1(\theta)(\alpha(u, g, \theta) + \psi(\theta) - g(\theta) + 1) - w(\theta)) f(\theta) + \psi_1(\theta) u(\theta, g, \theta) + \psi_2(\theta) z(\theta) \quad (A.3)
\]

where \( \mu_1(\theta), \mu_2(\theta) \) are Lagrange multipliers on the first two constraints, \( \psi_1, \psi_2 \) are the Lagrange multipliers on the integral constraints, and \( \xi_1(\theta), \xi_2(\theta) \) are co-state variables. Note that Hestenes’ theorem asserts that \( \psi_1, \psi_2 \) are independent of \( \theta \), and that multipliers \( \mu_1(\theta), \psi_1 \) on the inequality constraints are non-negative.

The first-order necessary conditions for a solution to GAP can now be stated as the state equations (A.1), (A.2), plus;

\[
H_e = (h e - \psi_1) = 0 \quad (A.4)
\]

\[
H_c = (\mu_1 v - \psi_1) f = 0 \quad (A.5)
\]

\[
H_\xi = \mu_2 + \xi_2 = 0 \quad (A.6)
\]

\[
\frac{d\xi_1}{d\theta} = -H_w = - (1 - \mu_1) f \quad (A.7)
\]

\[
\frac{d\xi_2}{d\theta} = -H_k = - (\psi_2 + \mu_1(u_k - v')) f - \xi_1 u_{\theta \theta} \quad (A.8)
\]

Finally, as the problem is one of variable end- and initial points, we have the transversality conditions

\[
\xi_i(\theta) = \xi_i(\bar{\theta}) = 0, \quad i = 1, 2 \quad (A.9)
\]
(iv) We now derive some properties of the co-state variable \( \xi_1(\theta) \) from the first-order and transversality conditions. First, from (A.5) and (A.7), we have:

\[
\frac{d\xi_1}{d\theta} = \left( \frac{\psi_1}{\nu'} - 1 \right) f
\]  

(A.10)

As we have established in (i) above, \( \tau - g \) is decreasing in \( \theta \) (otherwise a type-\( \theta \) region would prefer to announce \( \theta' > \theta \)), and therefore \( \nu'(\tau - g + 1) \) is increasing in \( \theta \). It follows that \( \psi_1/\nu'-1 \) is decreasing in \( \theta \). Moreover, using (A.4), (A.9), (A.10) we get

\[
\int_{\theta}^{\bar{\theta}} \left( \frac{\psi_1}{\nu'} - 1 \right) f d\theta = \xi_1(\bar{\theta}) - \xi_1(\theta) = 0
\]  

(A.11)

As \( \psi_1/\nu'-1 \) is decreasing in \( \theta \), it follows from (A.11) that \( \psi_1/\nu'-1 \) is first positive and then negative as \( \theta \) increases. This fact, plus (A.10), implies that \( \xi_1(\theta) \) is a quasi-concave function of \( \theta \), and strictly positive except at the endpoints; \( \xi_1(\theta) > 0, \theta \in (\bar{\theta}, \hat{\theta}) \).

(v) We now turn to look at distortions in the supply of the public good. We assume for the moment that there is no bunching \( (\mu_2(\theta) = 0, \theta \in \Theta) \). So, as \( \mu_2(\theta) = 0 \) everywhere, from (A.6), \( \xi_2(\theta) = 0 \) everywhere also, implying \( \xi_2 > 0 \). So, using (A.5), (A.8);

\[
\frac{u_x + h_x}{\nu'} = 1 + \left( 1 - \frac{\nu'}{\psi_1} \right) \int_{\theta}^{\bar{\theta}} \frac{h_x}{\psi_1} f d\theta - \frac{\xi_1 u_{g \theta}}{\psi_1 f \nu'}
\]  

(A.12)

Now consider the pure insurance case \( (h_x = 0) \). Then, (A.12) reduces to

\[
\frac{u_x}{\nu'} = 1 - \frac{\xi_1 u_{g \theta}}{\psi_1 f \nu'} \leq 1
\]  

(A.13)

where the inequality follows from the fact that \( \xi_1 \geq 0 \) as shown above, \( u_{g \theta} > 0 \) by assumption, and \( \psi_1 > 0 \) as asserted above. Moreover, from the properties of \( \xi_1 \), unless \( \theta = \bar{\theta} \) or \( \hat{\theta} \), the inequality in (A.13) is strict. So, the public good is oversupplied on all intervals where there is no bunching.

Now consider the general case with \( h_x > 0 \). Then, as \( \nu'/\psi_1 \) is increasing in \( \theta \), and from (A.11), \( \nu'/\psi_1 = 1 \) for some \( \theta = \bar{\theta} \). Consequently, from (A.12) \( g \) is undersupplied (oversupplied) for \( \theta < \bar{\theta} \) \( (\theta > \bar{\theta}) \). But, as \( \xi_1(\theta) > 0 \), from (A.12) \( g \) will be oversupplied when \( \nu'/\psi_1 = 1 \). So, there is a \( \hat{\theta} < \bar{\theta} \) such that \( g \) is undersupplied (oversupplied) for \( \theta < \hat{\theta} \) \( (\theta > \hat{\theta}) \). Finally, \( \hat{\theta} > \bar{\theta} \) as \( \xi_2(\theta) = 0 \) and \( \nu'/\psi_1 < 1 \) at \( \theta = \bar{\theta} \).

(v) Finally, we deal with distortions in public goods supply when there is bunching on one or more intervals. Initially, consider the pure insurance case \( (h_x = 0) \). Now suppose that there exists an interval with bunching i.e. where \( g \) is
constant. By the assumed continuity of \( g(\cdot) \), these intervals are closed, so we may write the interval as \([\theta_-, \theta_+]\).

Let \( \theta_+ < \hat{\theta} \). Now, suppose that (A.13) is violated i.e. there is undersupply or efficient supply at some \( \hat{\theta} \in [\theta_-, \theta_+] \). Then, there must be undersupply at \( \hat{\theta} = \theta_+ \) as \( u_{\theta} > 0 \). For any \( \epsilon > 0 \) small enough, there is no bunching at \( \theta_+ + \epsilon \), and so from the above argument, there is oversupply at \( \theta_+ + \epsilon \). Then, \( g(\hat{\theta}) \) must be right-discontinuous at \( \hat{\theta} = \theta_+ \), a contradiction.

It remains to show that there is undersupply (or efficient supply) at the endpoint \( \hat{\theta} \) when there is bunching at \( \hat{\theta} \) i.e. \( g \) is constant on the interval \([\theta_-, \hat{\theta}]\).

Integrating (A.8) over this interval and rearranging, we get

\[
\int_{\hat{\theta}}^{\hat{\theta}} \mu_1(z) \left[ \frac{u_g(g(\hat{\theta}), z)}{v'(1 + \tau(\hat{\theta}) - g(\hat{\theta}))} - 1 \right] f(z) \, dz = -\int_{\hat{\theta}}^{\hat{\theta}} \xi_1(z) u_{\theta g}(g(\hat{\theta}), z) \, dz < 0
\]

Also, from (A.5),

\[
\mu_1(z) = \frac{\psi_1}{v'(1 + \tau(\hat{\theta}) - g(\hat{\theta}))} > 0, \quad z \in [\theta_-, \hat{\theta}]
\]

So, (A.14) implies that

\[
\frac{u_g(g(\hat{\theta}), \hat{\theta})}{v'(1 + \tau(\hat{\theta}) - g(\hat{\theta}))} < 1
\]

i.e. oversupply at \( \hat{\theta} = \hat{\theta} \).

Now consider the general case \( (h_0 > 0) \). Even if there is bunching at the solution to GAP, the argument in (v) above establishing two-way distortion still applies if there is no bunching at the lower endpoint of the distribution. For then the critical \( \hat{\theta} \) must still be greater than \( \hat{\theta} \). □

Proof of Proposition 4

First, by incentive-compatibility, \( \tau^* \) is increasing in \( g \), and so is almost everywhere differentiable. By definition of \( w \), on intervals without bunching, we have

\[
\frac{dw}{d\theta} = u_\theta + (u_g - v') \frac{dg}{d\theta} + v' \frac{d\tau}{d\theta}
\]

So, from local incentive compatibility condition (A.1), on these intervals we have

\[
(u_g - v') \frac{dg}{d\theta} + v' \frac{d\tau}{d\theta} = 0
\]

(B.1)
Rearranging (B.1), and using (A.12), we get
\[
\frac{d\tau}{dg} = \frac{d\tau}{d\theta} \frac{d\theta}{dg} = 1 - \frac{u_g}{\psi_1} = \frac{h_e}{\psi_1} + \frac{\xi_1(\theta(g))u_{g'\theta}(g, \theta(g))}{\psi_1 f(\theta(g))}
\]  
(B.2)

where \(\theta(g)\) is the inverse of \(g(\theta)\) (as \(dg/d\theta > 0\), we can always invert to get \(\theta(g)\)).

Using the fact that from (A.11), \(1/\psi_1 = E_u[1/\nu']\), and setting
\[
\Phi(g) = \frac{\xi_1(\theta(g))u_{g'\theta}(g, \theta(g))}{\psi_1 f(\theta(g))}
\]
then (B.2) becomes
\[
\frac{d\tau}{dg} = E_u \frac{h_e}{\nu'} + \Phi(g)
\]
as claimed. Using the fact that \(\xi_1(\theta(g)) > 0, \ g \in (g(\theta), \ g(\tilde{\theta}))\) and \(\xi_1(\theta(g)) = 0, \ g = g(\theta), \ g(\tilde{\theta})\) we see that \(\Phi(g)\) has the required properties.

**Proof of Proposition 7**

To solve the GAP, it is convenient to transform variables. Let \(s = \tau - g\) be the surplus grant (i.e. the excess over that needed to finance the public good). So, we can write the utility of a typical region as \(v(s + \lambda) + h(e) + u(\tau - s)\). Moreover, the monotonicity condition on \(g\), (35) plus incentive-compatibility, indicates that \(\tau\) must be decreasing in \(\lambda\) (otherwise, type-\(\lambda\) regions would have an incentive to announce \(\lambda' > \lambda\)). So, \(s = \tau - g\) must be decreasing in \(\lambda\). We can therefore write the GAP as the problem of choosing continuous control variables \(\tau(\lambda), \ z(\lambda)\) and a number \(e\) to maximise
\[
\int \left[w(\lambda) + h(e)\right] f(\lambda) d\lambda
\]
subject to the following constraints
\[
w(\lambda) = v(s(\lambda) + \lambda) + u(\tau(\lambda) - s(\lambda))
\]
\[-z(\lambda) \geq 0\]
\[-\int \tau(\lambda)f(\lambda)d\lambda \geq 0\]
\[ \dot{\lambda} - \int_{\lambda}^{(s(\lambda) + e)f(\lambda)d\lambda} \geq 0 \]

and the state equations;
\[ \frac{dw}{d\lambda} = u'(s(\lambda) + \lambda) \]
\[ \frac{ds}{d\lambda} = z(\lambda) \]

Again we treat \( s \) as a state variable, and introduce the new control variable \( z \), in order to deal with the monotonicity constraint that \( s \) must be non-increasing in \( \lambda \). The fourth constraint above is obtained by combining the budget constraint and the identity defining \( e \).

Forming the Hamiltonian as above, we get
\[ H = (w(\lambda) + h(e))f(\lambda) + \mu_1(\lambda)(u(\tau(\lambda) - s(\lambda)) + v(s(\lambda) + \lambda) - w(\lambda))f(\lambda) - \mu_2(\lambda)z(\lambda) - \psi_1(\tau(\lambda)f(\lambda) - \psi_2(s(\lambda) + e)f(\lambda) + \xi_1(\lambda)u'(s(\lambda) + \lambda) + \xi_2(\lambda)z(\lambda) \]
(C.1)

From (C.1), the first-order and transversality conditions are
\[ H_x = (h_x - \psi_2)f = 0 \] (C.2)
\[ H_\xi = (\mu_1u_x - \psi_1)f = 0 \] (C.3)
\[ H_\xi = - \mu_2 + \xi_2 = 0 \] (C.4)
\[ \frac{d\xi_1}{d\lambda} = - H_u = -(1 - \mu_1)f \] (C.5)
\[ \frac{d\xi_2}{d\lambda} = - H_s = (\psi_2 - \mu_1(u' - u_x))f - \xi_1u' \] (C.6)

and
\[ \xi_i(\theta) = \xi_i(\bar{\theta}) = 0, \quad i = 1, 2 \] (C.7)

First, from (C.3), and (C.5), we have:
\[ \frac{d\xi_1}{d\lambda} = \left( \frac{\psi_1}{u_x} - 1 \right)f \] (C.8)

As we have established above, \( g \) is increasing in \( \lambda \), and therefore from \( u_x < 0 \), \( u_x \) is decreasing in \( \lambda \). It follows that \( \psi_1/u_x - 1 \) is increasing in \( \lambda \). Moreover, using (C.7), (C.8) we get
Monotonicity of \( \lambda \psi_1 / u_g - 1 \) in \( \lambda \), plus (C.9), implies that \( \psi_1 / u_g - 1 \) is first negative and then positive as \( \lambda \) increases. This fact, plus (C.8), implies that \( \xi_1(\lambda) \) is a quasi-convex function of \( \lambda \), and strictly negative except at the endpoints: \( \xi_1(\lambda) < 0, \lambda \epsilon (\lambda, \bar{\lambda}) \).

For simplicity, we focus on the case where there is no bunching. Manipulation of (C.2), (C.3), (C.6) gives

\[
\frac{u_x + h_x}{u'} = 1 + \left( 1 - \frac{u_g}{\psi_1} \right) \frac{h_x}{u'} + \frac{\xi_1 v'u_g}{\psi_1 v'}
\]

(C.10)

In the pure insurance case, \( (h_x=0) \), using the properties of \( \xi_1 \), and strict concavity of \( v \), we see from (C.10) that provision is efficient at the endpoints of \( \Lambda \), and for \( \lambda \epsilon (\lambda, \bar{\lambda}) \), we have \( (u_x + h_x)/v' > 1 \) i.e. undersupply.

In the general case with \( h_x \neq 0 \), assume for the moment that \( \xi_1(\lambda) = 0 \). Then, as \( \psi_1 / u_g \) is increasing in \( \lambda \), and from (C.9), it follows that \( \psi_1 / u_g = 1 \) for some \( \lambda = \lambda_1 \). Consequently, from (C.10), \( g \) is oversupplied (undersupplied) for \( \lambda < \lambda_1 \). But, as \( \xi_1(\lambda) < 0 \), so \( g \) will be undersupplied when \( u_g / \psi_1 = 1 \). So, there is a \( \lambda < \lambda_1 \) such that \( g \) is oversupplied (undersupplied) for \( \lambda < \lambda_1 \). Finally, \( \lambda > \lambda_1 \), as \( \xi_1(\lambda) = 0 \) and \( u_g / \psi_1 > 1 \) at \( \lambda = \lambda_1 \).  

**Proof of Proposition 9**

(i) Let \( \omega(c, l) = e^{\gamma-a} l^\alpha, \gamma = 1 - \beta \). Then, it is easy to compute that indirect utility and labour supply are

\[
\Omega(t, \lambda) = \hat{\kappa} \lambda^\gamma (1 - t)^{\gamma-a}
\]

(D.1)

\[
L(t, \lambda) = \lambda \left( 1 - \frac{\alpha}{\gamma} \right)
\]

(D.2)

where \( \hat{\kappa} = \alpha^\alpha (\gamma - \alpha)^{\gamma-a} \gamma^a \). So, from (D.2), the only solution to \( tL(t, \lambda) = r \) is

\[
t = \frac{r}{\lambda \left( 1 - \frac{\alpha}{\gamma} \right)}
\]

(D.3)

Combining (D.1), (D.3), we get

\[
V(t, \lambda) = \hat{\kappa} \left( 1 - \frac{\alpha}{\gamma} \right)^{a-x} \left( 1 - \frac{\alpha}{\gamma} \right) \left( \lambda \left( 1 - \frac{\alpha}{\gamma} \right) - \rho g + \tau \right)^{\gamma-a} \lambda^a
\]

\[
= \kappa \left( \lambda \left( 1 - \frac{\alpha}{1 - \beta} \right) - \rho g + \tau \right)^{1-\beta-a} \lambda^a
\]
as required. Finally, as \( \alpha < 1 - \beta \), it is then easy to check that

\[
V_t = (1 - \beta - \alpha)\kappa \left( \lambda \left( 1 - \frac{\alpha}{1 - \beta} \right) - \rho g + \tau \right)^{-\beta(1 + \alpha)} \lambda^\alpha > 0
\]

(D.4)

Differentiating (D.4) again, we see that \( V_{tt} < 0 \), and after some rearrangement, that

\[
V_{tt} = \frac{V}{\lambda(1 + r)} [t \alpha - \beta]
\]

So, using (D.3), \( V_{tt} > 0 \) iff

\[
\alpha > \beta \lambda \left( 1 - \frac{\alpha}{1 - \beta} \right)^r
\]

which reduces to

\[
\alpha > \frac{(1 - \beta)\beta^2}{1 - \beta + \beta^2} \quad \psi = \lambda^r
\]

as required. ■

References