Foundations and Trends® in
Microeconomics
4:1–2 (2008)

The Microeconomics of Insurance

Ray Rees and Achim Wambach

now
the essence of knowledge

The Microeconomics of Insurance

The Microeconomics of Insurance

Ray Rees

*Institut für Volkswirtschaftslehre
University of Munich
Ludwigstrasse 28/III VG
80539 Munich
Germany*

Ray.Rees@lrz.uni-muenchen.de

Achim Wambach

*Department of Economics
University of Cologne
50931 Cologne
Germany*

wambach@wiso.uni-koeln.de

now

the essence of knowledge

Boston – Delft

Foundations and Trends® in Microeconomics

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is R. Rees and A. Wambach, The Microeconomics of Insurance, Foundations and Trends® in Microeconomics, vol 4, no 1–2, pp 1–163, 2008

ISBN: 978-1-60198-108-0

Editorial Scope

Foundations and Trends® in Microeconomics will publish survey and tutorial articles in the following topics:

- Environmental Economics
- Contingent Valuation
- Environmental Health Risks
- Climate Change
- Endangered Species
- Market-based Policy Instruments
- Health Economics
- Moral Hazard
- Medical Care Markets
- Medical Malpractice
- Insurance economics
- Industrial Organization
- Theory of the Firm
- Regulatory Economics
- Market Structure
- Auctions
- Monopolies and Antitrust
- Transaction Cost Economics
- Labor Economics
- Labor Supply
- Labor Demand
- Labor Market Institutions
- Search Theory
- Wage Structure
- Income Distribution
- Race and Gender
- Law and Economics
- Models of Litigation
- Crime
- Torts, Contracts and Property
- Constitutional Law
- Public Economics
- Public Goods
- Environmental Taxation
- Social Insurance
- Public Finance
- International Taxation

Information for Librarians

Foundations and Trends® in Microeconomics, 2008, Volume 4, 6 issues. ISSN paper version 1547-9846. ISSN online version 1547-9854. Also available as a combined paper and online subscription.

Foundations and Trends® in
Microeconomics
Vol. 4, Nos. 1–2 (2008) 1–163

DOI: 10.1561/0700000023

The Microeconomics of Insurance

Ray Rees[1] and Achim Wambach[2]

[1] *Institut für Volkswirtschaftslehre, University of Munich, Ludwigstrasse 28/III VG, 80539 Munich, Germany, Ray.Rees@lrz.uni-muenchen.de*

[2] *Department of Economics, University of Cologne, 50931 Cologne, Germany, wambach@wiso.uni-koeln.de*

Abstract

In this relatively short survey, we present the core elements of the microeconomic analysis of insurance markets at a level suitable for senior undergraduate and graduate economics students. The aim of this analysis is to understand how insurance markets work, what their fundamental economic functions are, and how efficiently they may be expected to carry these out.

Contents

1

Introduction

When we consider some of the possible ways of dealing with the risks that inevitably impinge on human activities — lucky charms, prayers and incantations, sacrifices to the Gods, consulting astrologists — it is clear that insurance is by far the most rational. Entering into a contract under which one pays an insurance premium (a sum that may be small relative to the possible loss), in exchange for a promise of compensation if a claim is filed on occurrence of a loss, creates economic value even though nothing tangible is being produced. It is clearly also a very sophisticated transaction, which requires a well-developed economic infrastructure. The events which may give rise to insurable losses have to be carefully specified, the probabilities of the losses have to be assessed, so that premiums can be set that do not exceed the buyer's willingness to pay and make it possible for the insurer to meet the costs of claims and stay in business, while, given the fiduciary nature of the contract, buyers must be confident that they will actually receive compensation in the event of a claim. Insurance in its many and varied forms is a central aspect of economic activity in a modern society.

In this relatively short survey, we present the core elements of the microeconomic analysis of insurance markets, at a level suitable for senior undergraduate and graduate economics students. The aim of this analysis is to understand how insurance markets work, what their fundamental economic functions are, and how efficiently they may be expected to carry these out. We can give a brief outline of the coverage of the survey with the help of one simple model.

Consider an individual who has to decide how much insurance cover to buy. Formally, she maximizes her expected utility by choosing the optimal cover or indemnity C:

$$E[U] = (1 - \pi)u(W - P(C)) + \pi u(W - P(C) - L + C).$$

Here we assume that the individual has a von Neumann–Morgenstern utility function[1] $u(\cdot)$ which is increasing and strictly concave. Strict concavity implies that the individual is risk averse. π is the probability that a loss of size L occurs. W is her wealth in the event of no loss. P is the insurance premium paid, which can in general be thought of as a function of C, the cover.

As a simple example, assume you have a van Gogh with a market value of \$10 million hanging in your living room, and the probability of having the painting stolen is say $\pi = 0.001$, or one in a thousand. In this case $L = \$10$ million. You can buy insurance cover C by paying the premium $P(C) = pC$. For every \$1 you want to get paid in case of a loss, you have to pay $p \times \$1$. p is called the premium rate. Thus, if the premium rate is 0.002, or \$2 per \$1000 of cover, and you want to get all of your \$10 million back, you have to pay \$20,000 up front as a premium to the insurance company. Note that if the van Gogh is stolen, you have paid the premium already, so net you receive \$9,980,000 or $C - P(C)$.

This simple model is the starting point for all the discussion in the following chapters. In the first part of Section 2, we deal exclusively with this model and we investigate how the demand for insurance depends on the premium rate p, wealth W, the size and probability of the loss L and π, respectively, and the degree of risk aversion as reflected in the concavity of $u(\cdot)$.

[1] We assume throughout this survey that the reader is familiar with the basic elements of the economics of uncertainty. For treatments of this see Gravelle and Rees (2004, Chap. 17), (Gollier, 2001, Chap. 1–3), and Eeckhoudt et al. (2005).

However, there are limitations to the applicability of this model. Many real world features of insurance contracts such as deductibles, contracts with experience rating and coinsurance require more elaborate models. We will now discuss these limitations and indicate where the sections in this survey deal with the features which this simple model does not adequately take into account. (We use arrows to show where the modified models differ from the basic model above.)

1. *State dependent utility function*

$$E[U] = (1-\pi)\overset{\Downarrow}{\overbrace{\;u\;}}(W - P(C)) + \pi\overset{\Downarrow}{\overbrace{\;v\;}}(W - P(C) - L + C).$$

For some applications it is not sensible to assume that people have the same utility whether a loss has incurred or not, even if they are fully financially compensated for the loss. Assume that you own a gold bar that is stolen, but fully covered by an insurance policy. In this case you probably will not mind the loss. You just go out and buy yourself another gold bar. In the case of your van Gogh being stolen this might be different. If you are very attached to this painting, you will feel worse off even if the insurance company pays out the full price you have paid for it. The reason is that a particular van Gogh is not a tradable good which can be rebought in the market. Another example is health insurance — if you break your leg skiing, even with full insurance to cover the medical expenses you will feel worse than when you are healthy. These aspects are discussed in detail in Section 2, where we consider the demand for insurance in the presence of state dependent utility functions.

2. *Is there only one risk?*

$$E[U] = E_{\tilde{W}}[(1-\pi)u(\overset{\Downarrow}{\overbrace{\;\tilde{W}\;}} - P) + \pi u(\overset{\Downarrow}{\overbrace{\;\tilde{W}\;}} - P - L + C)].$$

In the simple model above,[2] the van Gogh is either stolen or not. However, in general individuals face more than one risk. Standard additional

[2] The symbol $E_{\tilde{W}}$ denotes the expected value taken with respect to the distribution of the random variable $\tilde{W}$.

risks like car accidents, illness, fire, etc. can be covered by separate insurance contracts. But there are also uninsurable risks around — for example income risk, as the return on shares and bonds you own is uncertain, or because your job is not secure. You might not know for sure how much money you are going to inherit from a benevolent grandmother, whether you will marry into money or not .... This feature is known as *background risk.* Also in Section 2 we analyze the situation where individuals face additional uninsurable risks (like the $\tilde{W}$ in the equation above). Now the demand for insurance will depend on whether those risks reinforce each other or whether they tend to offset each other so that they can be used as a hedging mechanism.

3. *Where does $P(C)$ come from?*

$$E[U] = (1-\pi)u(W - \overbrace{P(C)}^{\Downarrow}) + \pi u(W - \overbrace{P(C)}^{\Downarrow} - L + C).$$

In the simple model we have assumed that the individual faces some exogenous given premium function $P(C) = pC$. But who determines the premium? On what factors does it depend? In Section 3 where we discuss the supply of insurance, this will become clear. We consider premium setting on a competitive market. We will also discuss how insurance shareholders react to risks by *diversifying* their risks (risk spreading) and how insurance enables the insured to *pool* their risks. Finally, we discuss some aspects of the important subject of the regulation of insurance markets.

4. *Is there only one loss level possible?*

$$E[U] = (1-\pi)u(W - P) + \pi \overbrace{\sum_i \pi_i}^{\Downarrow} u(W - P - \overbrace{L_i + C_i}^{\Downarrow}).$$

In many situations a single loss level does not seem appropriate. Certainly, your van Gogh is either stolen or not, but in the case of a fire, for example, it could be partly or completely damaged. If you have a car accident, the damage can vary between some hundred dollars and many hundreds of thousands. Similarly in aviation insurance: A claim could have the size of a few hundred dollars for a damaged suitcase, but can increase to many millions of dollars for loss of a plane. As a matter of fact, one of the largest liability claims in the history of flight insurance resulted from the blowing up of the PanAm Boeing 747 over Lockerbie, Scotland. So far more than $510 million has been paid. More than one loss level is discussed with the help of the model of Raviv, which we present in Section 3. This model provides a synthesis of the demand for and supply of insurance in the case of many loss levels. In this model we will see *deductibles* and *coinsurance* emerging. By deductibles it is meant that the first D dollars of the loss have to be paid by the insured. Coinsurance applies if an additional dollar of loss is only partially covered. This might be the case if for example the insurance covers a fixed percentage of the loss.

5. *Is π known?*

$$E[U_i] = (1 - \overbrace{\pi_i}^{\Downarrow})u(W - P) + \overbrace{\pi_i}^{\Downarrow}u(W - P - L + C).$$

When determining the premium rate from the point of view of the insurer it is usually assumed that the probability of loss π is known. However, this may not necessarily be the case. You probably know much better than your insurer whether you are a cautious or a crazy driver, whether you have a healthy lifestyle or not, and so on. This is modeled by assuming that the insured knows her own π_i and the insurance company only has some information about the overall distribution of the π_i in the population. In those cases high risk types with a large π_i try to mimic low risk types and buy insurance which is not designed for them, causing losses for the insurer. This is known as *adverse selection.* In Section 4, we discuss the seminal paper by Rothschild and Stiglitz and other models which deal with this topic. The phenomenon of adverse selection allows us to understand why in some cases insurers offer several different contracts for the same risk. For your car insurance, for example, you might buy a contract with no deductible and a high premium rate or with a deductible and a lower premium rate. Offering a choice of contracts with different premium rates is a discriminating mechanism, which only makes sense if people differ in some unobservable characteristic. This analysis also allows us to discuss another feature which is commonly observed: *Categorical discrimination.* What are the pros and cons of conditioning a particular contract on gender or age, for example? Is it efficient to sell different contracts to males and females or to young and old drivers?

6. *Is the loss probability exogenous or endogenous?*

$$E[U] = (1 - \overbrace{\pi(e)}^{\Downarrow})u(W - P) + \overbrace{\pi(e)}^{\Downarrow}u(W - P - L + C) - \overbrace{c(e)}^{\Downarrow}.$$

In many situations the loss probability can be influenced *ex ante* by the insured. The degree of attentiveness you devote to the road is something you have control of. By increasing your concentration the loss probability is reduced: the derivative of the probability $\pi'(e) < 0$. However, the more you concentrate the less time you have for phone calls with your mobile phone, listening to the radio, etc., so there are costs of concentrating $(c(e))$ which increase if one employs more effort: marginal cost of effort $c'(e) > 0$. If a person is completely insured, she might not employ any effort as she is not liable for any damage. This problem is known as *ex ante moral hazard* and is discussed in detail in Section 5. Here we will find another reason why insurance companies may offer contracts with *partial insurance cover.* We also discuss there *ex-post* moral hazard, the situation, held to be prevalent in health insurance markets, in which the fact that health costs are covered by insurance may lead to demand for them being greater than the efficient level.

7. *Is the size of the loss observable?*

$$E[U] = (1 - \pi)u(W - P) + \pi u(W - P - \overbrace{L}^{\Downarrow} + C).$$

In some situations neither the occurrence of a loss nor the size of the loss is easily observable by the insurance firm. In those situations the insured might be tempted to overstate the size of a loss or to claim a loss which has not occurred. *Insurance fraud* is discussed at the end of Section 5. For obvious reasons the actual size of insurance fraud is difficult to measure. However estimates based on questionnaires suggest that for personal liability insurance around 20% of all claims are fraudulent.

We will discuss how contractual and institutional arrangements might cope with this problem.

8. *Why only one period?*

$$\begin{aligned} E[U] = {} & (1-\pi)u(W-P_0) + \pi u(W-P_0-L+C_0) \\ & + \overbrace{(1-\pi)[(1-\pi)u(W-P_N) + \pi u(W-P_N-L+C_N)]}^{\Downarrow} \\ & + \overbrace{\pi[(1-\pi)u(W-P_L) + \pi u(W-P_L-L+C_L)]}^{\Downarrow} . \end{aligned}$$

If insurance is sold under perfect information, it does not make any difference whether many single-period contracts or one many-period contract are sold. In reality however, we observe many contracts which have a dynamic component, such as *experience rating* contracts in the car or health insurance industry. In those cases, individuals pay a different premium in the future period depending on whether a loss has occurred or not (P_L or P_N, respectively). This phenomenon can be explained by the existence of asymmetric information, as in the adverse selection or moral hazard models mentioned above. In Section 4, we consider this issue in the context of adverse selection and show how experience rating may appear endogenously. Also in Section 5, as part of the discussion on moral hazard, *dynamic contracts* are considered. Another topic which is relevant when one discusses multi-period contracts is the issue of *renegotiation and commitment.* The crucial point here is that even if *ex ante* both the insurer and the insured agree to a longer lasting contract, *ex-post* it might be of advantage for both parties to change the terms of the contract in some circumstances.

2

The Demand for Insurance

2.1 Introduction

Insurance is bought by means of a contract specifying a set of events, the occurrence of which will create a financial loss for the buyer. The insurer undertakes to pay compensation, which we will call the cover, (or equivalently the indemnity) in the event of these losses. In exchange the buyer pays a premium for sure, usually at the time of entering into the contract. The basic characteristics of all insurance contracts are therefore: specified loss events, losses, cover, and premium. The "demand for insurance" can in the first place be interpreted as the *demand for cover*.

The details and complexity of specific insurance contracts will vary greatly with the particular kinds of risks being dealt with. Though for theoretical purposes we model insurance as completely defined by the above four elements, we should recognize that in applications to specific markets, for example health, life, property and liability insurance, it may often be necessary to adapt this general framework to the particular characteristics of the market concerned.

We can go beyond this descriptive account of the insurance contract to obtain a deeper interpretation of the demand for insurance, and of the economic role that insurance markets play. The effect of this interpretation is to place insurance squarely within the standard framework of microeconomics, and this has powerful analytical advantages, since it allows familiar and well-worked out methods and results to be applied.

The basis of the approach is the concept of the *state of the world.* For our purposes, it is sufficient to think of a state of the world as corresponding to an amount of the loss incurred by the insurance buyer. The situation in which she incurs no loss is one possible state, and there is then an additional state for each possible loss. The simplest case is that in which there is only one possible loss, so we have two states of the world. At the other extreme, losses may take any value in an interval $[0, L_m]$, in which case there is a continuum of possible states of the world, each defined by a point in the interval. We shall consider in this survey models of both these cases, as well as intermediate ones, but we begin here with the simplest case, already encountered in the Introduction.

We define the buyer's wealth[1] in each state of the world, W, as her *state contingent wealth.* Before entering into an insurance contract, the consumer has given endowments of state contingent wealth, W_0 if no loss occurs, and $W_0 - L$ given the occurrence of loss $L > 0$. If she buys insurance, she will receive under the contract an amount of compensation C that will generally depend on L, and will pay *for sure*, i.e., in every state of the world, a premium P. Thus with insurance her state contingent wealth becomes $W_0 - P$ in the no loss state, and $W_0 - L - P + C$ in the loss state. Then, by allowing the buyer to vary P and C, the insurance market is providing the consumer the means to vary her state contingent wealth away from the values she is initially endowed with. *Insurance permits trade in state contingent wealth* and in doing so allows the buyer to transfer wealth to the loss state from the no-loss state. Moreover, these state contingent wealth holdings can be interpreted as the "goods" in the standard microeconomic model of the consumer, and then the "demand for insurance"

[1] In fact, since most of what we do concerns only one time period, "wealth" and "income" can usually be used interchangeably.

becomes, under this interpretation, the *demand for state contingent wealth*.

In the rest of this section, we shall find it useful to consider both concepts of the demand for insurance — the demand for cover, and the demand for state contingent wealth — side by side, since each gives its own insights and interpretations. Common to both is the basic microeconomic framework of optimal choice. The demand for insurance is viewed as the solution to the problem of maximizing a utility function subject to a budget constraint. This utility function is taken from the theory of preferences under uncertainty usually referred to as the *Expected Utility Theory*. The theory of insurance demand can be regarded as an application, indeed one of the most successful applications, of this theory. Under it, the consumer is modeled as having a von Neumann–Morgenstern utility function $u(W)$, which is unique up to a positive linear transformation and is at least three times continuously differentiable. We assume the first derivative $u'(W) > 0$, more wealth is always preferred to less. Moreover, we assume that the insurance buyer is risk averse, and so $u''(W) < 0$, the utility function is strictly concave.[2] The sign of $u'''(W)$, which defines the curvature of the marginal utility function $u'(W)$, we leave open for the moment.

Note that the utility function is the same regardless of whether we are in the no loss or loss state. That is, the utility function is *state independent*. This is not always an appropriate assumption for insurance, and we consider the effects of changing it below.

An important characteristic of any utility function is its *Arrow–Pratt index of risk aversion*

$$A(W) \equiv -\frac{u''(W)}{u'(W)}, \tag{2.1}$$

which increases at any wealth level with the risk aversion of the consumer. The lower the index at a given wealth level, the greater the willingness to bear risk. As we shall see, this index arises frequently in comparative statics analysis of insurance demand, and the results

[2] Strict concavity implies that for any risk $\tilde{W}$, $E_{\tilde{W}}[u(\tilde{W})] < u(E_{\tilde{W}}[\tilde{W}])$, i.e., the consumer is risk averse in the sense of always preferring the expected value of a risk to owning the risk itself.

typically depend on whether it is increasing, constant or decreasing in W. We will usually consider all three cases.

Under this theory, given a set of alternative probability distributions of wealth, each of which gives a corresponding probability distribution of utilities, the decision taker chooses that distribution with the highest expected value of utility, hence the name. We now consider the insights into the demand for insurance this theory gives us.

2.2 Two Models of the Demand for Insurance

The first step is to define the buyer's budget constraint appropriately. Then, formulating the problem as the maximization of expected utility subject to this constraint, we can go on to generate the implications of the model. The simplest models have just two possible states of the world, a no loss state and a single state with loss L. The probability of this loss is π. Thus the expected value of wealth without insurance is

$$\bar{W} = (1 - \pi)W_0 + \pi(W_0 - L) = W_0 - \pi L \tag{2.2}$$

with πL the expected value of income loss. We always assume $L < W_0$.

Expected utility in the absence of insurance is

$$\bar{u}^0 = (1 - \pi)u(W_0) + \pi u(W_0 - L). \tag{2.3}$$

In the absence of insurance, the buyer has an uncertain wealth endowment with an expected utility of $\bar{u}^0$.

The insurer offers cover C at a *premium rate* p, where p is a pure number between zero and one. The *premium amount* is $P = pC$. We assume that the buyer can choose any value of $C \geq 0$. The nonnegativity restriction says simply that the buyer cannot offer a bet on the occurrence of the loss event, and is a realistic restriction on insurance markets. The constraint will in general not be binding at the equilibrium of a competitive insurance market. Nevertheless there are cases in which it should be considered explicitly, as we shall see. The assumption that cover is fully variable may well not hold in a real insurance market (one may for example only be able to choose full cover, $C = L$, as in health insurance, or there may be an upper limit on cover $C_{\max} < L$, as in auto insurance) but an important goal of the analysis is to understand why such restrictions exist, and so it is useful to begin by assuming the most general case of no restrictions (beyond nonnegativity) on cover. Other possibilities are considered below. Finally, it is convenient to express the premium as the product of cover and a premium *rate*. This is a common, but not universal, way of expressing insurance premia in reality, but of course a premium rate, the price of one monetary unit of cover, can always be inferred from values of P and C. The key point is the assumption that $p = P/C$ is constant and independent of C, so that the average and marginal cost of cover to the consumer are the same.

We obtain two alternative model formulations by defining demand in terms of cover, on the one hand, and state contingent wealth, on the other.

2.2.1 The Model of the Demand for Cover

We assume the buyer solves the problem

$$\max_{C \geq 0} \bar{u} = (1-\pi)u(W_0 - P) + \pi u(W_0 - L - P + C) \tag{2.4}$$

subject to the constraint

$$P = pC. \tag{2.5}$$

Clearly, the simplest way to solve this is to substitute from the constraint into the utility function and maximize

$$\bar{u}(C) = (1-\pi)u(W_0 - pC) + \pi u(W_0 - L + (1-p)C) \tag{2.6}$$

giving the Kuhn–Tucker condition

$$\begin{aligned}\bar{u}_C(C^*) = & -p(1-\pi)u'(W_0 - pC^*) + (1-p)\pi u' \\ & \times (W_0 - L + (1-p)C^*) \leq 0 \quad C^* \geq 0 \;\; \bar{u}_C(C^*)C^* = 0.\end{aligned} \tag{2.7}$$

Taking the second derivative of $\bar{u}(C)$, we have

$$\begin{aligned}\bar{u}_{CC}(C) = & \, p^2(1-\pi)u''(W_0 - pC) \\ & + (1-p)^2\pi u''(W_0 - L + (1-p)C) < 0,\end{aligned} \tag{2.8}$$

where the sign follows because of the strict concavity of the utility function at all $C \geq 0$. Thus expected utility is strictly concave in C, and the first-order condition $\bar{u}_C(C^*) = 0$ is both necessary and sufficient for optimal cover $C^* > 0$.

The condition implies two cases:

Optimal cover is positive:

$$C^* > 0 \Rightarrow \frac{p}{1-p} = \frac{\pi}{(1-\pi)} \frac{u'(W_0 - L + (1-p)C^*)}{u'(W_0 - pC^*)}. \tag{2.9}$$

Optimal cover is zero:

$$C^* = 0 \Rightarrow \frac{p}{1-p} \geq \frac{\pi}{1-\pi} \frac{u'(W_0 - L)}{u'(W_0)}. \tag{2.10}$$

Taking the case $C^* > 0$ and rearranging (2.9) we have

$$u'(W_0 - pC^*) = \frac{\pi}{p} \frac{(1-p)}{(1-\pi)} u'(W_0 - L + (1-p)C^*). \tag{2.11}$$

We call the case in which $p = \pi$ the case of a *fair premium*,[3] that where $p > \pi$ the case of a *positive loading*, and that where $p < \pi$ the case of a *negative loading*. We can then state the first results,[4] using (2.11), as:

$$p = \pi \Leftrightarrow u'(W_0 - pC^*) = u'(W_0 - L + (1-p)C^*) \Leftrightarrow C^* = L \tag{2.12}$$

$$p > \pi \Leftrightarrow u'(W_0 - pC^*) < u'(W_0 - L + (1-p)C^*) \Leftrightarrow C^* < L \tag{2.13}$$

$$p < \pi \Leftrightarrow u'(W_0 - pC^*) > u'(W_0 - L + (1-p)C^*) \Leftrightarrow C^* > L. \tag{2.14}$$

In words:

> *with a fair premium the buyer chooses full cover*;
> *with a positive loading the buyer chooses partial cover*;
> *with a negative loading the buyer chooses more than full cover.*

where the last two results follow from the fact that $u'(\cdot)$ is decreasing in wealth, i.e., from risk aversion.

Taking the case of zero cover, since risk aversion implies $u'(W_0 - L) > u'(W_0)$, p must be sufficiently greater than π for this case to be possible.

[3] So called because it equals the expected value of loss, and an insurer selling to a large number of buyers with identical, independent risks of this type would exactly break even in expected value. See Section 3 for further discussion.

[4] These were first derived in Mossin (1968), and are the most basic in the theory of the demand for insurance. They are often collectively referred to as the Mossin Theorem.

We can obtain a useful diagrammatic representation of the equilibrium as follows. Let $U(C,P)$ denote the objective function for the maximization problem in (2.4). Given (2.7), assume $C^* > 0$ and rewrite the condition as

$$\frac{\pi u'(W_0 - L - P^* + C^*)}{(1-\pi)u'(W_0 - P^*) + \pi u'(W_0 - L - P^* + C^*)} = p, \tag{2.15}$$

where $P^* = pC^*$ is the premium payment at the optimum. This is the condition that would be obtained by solving the problem of maximizing expected utility $U(C,P)$ with respect to P and C, and subject to the constraint in (2.5). We can interpret the ratio on the LHS of (2.15) as a marginal rate of substitution between P and C, i.e., as the slope of an indifference curve of $U(C,P)$ in (C,P)-space, and then this condition has the usual interpretation as the equality of marginal rate of substitution and price, or tangency of an indifference curve with a budget line.

This is illustrated in Figure 2.1. The lines show the constraint $P = pC$ for varying values of p. The indifference curves show (C,P)-pairs

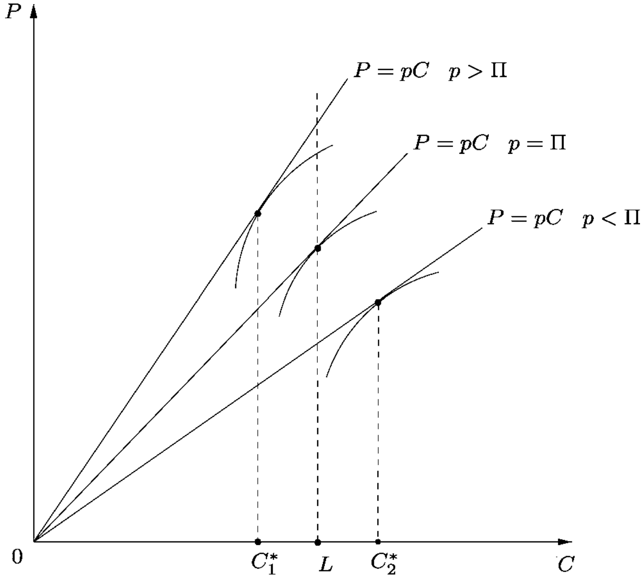

Fig. 2.1 Optimal choice of cover.

that yield given levels of expected utility. We shall justify the shape shown in a moment. Optimal C in each case is given by a point of tangency. For $p = \pi$, this point corresponds to L, as we have already established.

It remains to justify the shapes of the indifference curves shown in Figure 2.1. Along any indifference curve in the (C, P)-space, we must have

$$U(C,P) = (1-\pi)u(W_0 - P) + \pi u(W_0 - L - P + C) = k \tag{2.16}$$

for some constant k. Using subscripts to denote partial derivatives, we have

$$U_C = \pi u'(W_0 - L - P + C) \tag{2.17}$$

$$U_P = -[(1-\pi)u'(W_0 - P) + \pi u'(W_0 - L - P + C)] \tag{2.18}$$

$$U_{CC} = \pi u''(W_0 - L - P + C) \tag{2.19}$$

$$U_{CP} = \bar{u}_{PC} = -\pi u''(W_0 - L - P + C) \tag{2.20}$$

$$U_{PP} = (1-\pi)u''(W_0 - P) + \pi u''(W_0 - L - P + C). \tag{2.21}$$

Then, from the Implicit Function Theorem, we have that the slope of an indifference curve is

$$\frac{dP}{dC} = -\frac{U_C}{U_P} > 0 \tag{2.22}$$

so this justifies the positive slopes of the indifference curves in Figure 2.1. Moreover, setting $C^* = L$ gives

$$\frac{dP}{dC} = \pi \tag{2.23}$$

so that at $C^* = L$ on the C-axis, all indifference curves have the same slope, π.

To justify the curvature, consider first Figure 2.2. The characteristic of this curvature is that all points in the interior of the convex set formed by the indifference curve yield a higher level of expected utility than any point on the indifference curve. For example point A in the figure must yield a higher expected utility than point B because it offers higher cover for the same premium. Since B and C yield the same

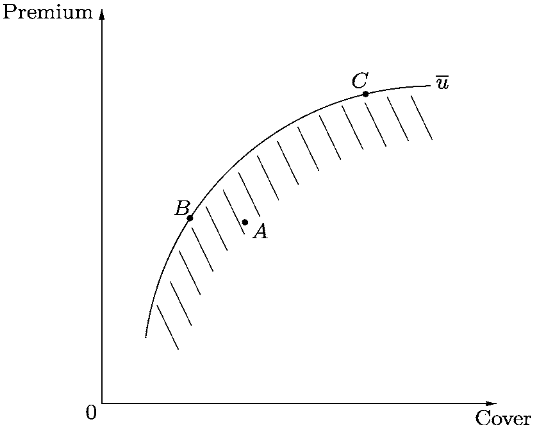

Fig. 2.2 Quasiconcavity.

expected utility, A must be better than C also. A function having this property is called *strictly quasiconcave.* Thus, we have to prove that the function $U(C, P)$ is strictly quasiconcave. The easiest way to do this is to show that $U(C, P)$ is strictly concave, because every strictly concave function is also strictly quasiconcave. $U(C, P)$ is strictly concave if the following conditions are satisfied:

$$U_{CC} < 0 \tag{2.24}$$

$$\begin{vmatrix} U_{CC} & U_{CP} \\ U_{PC} & U_{PP} \end{vmatrix} = U_{CC}U_{PP} - U_{PC}U_{CP} > 0. \tag{2.25}$$

The first condition is satisfied, because of risk aversion. By inserting the above expressions for the second-order partials and canceling terms we obtain that the determinant is equal to

$$\pi(1-\pi)u''(W_0 - P)u''(W_0 - L - P + C) > 0 \tag{2.26}$$

as required. Intuitively, since the utility function $u(W)$ is strictly concave in wealth, and wealth is linear in P and C, $U(C, P)$ is strictly concave in these variables.

From the first-order condition $\bar{u}_C(C^*) = 0$ we can in principle solve for optimal cover as a function of the exogenous variables of the problem: wealth, the premium rate (price), the amount of loss, and the loss probability

$$C^* = C(W_0, p, L, \pi). \tag{2.27}$$

We call this function the buyer's *cover demand function.* We consider its main properties below.

2.2.2 The Model of the Demand for State-Contingent Wealth

We now let the choice variables in the problem be the state contingent wealth values W_1 and W_2, respectively, where

$$W_1 = W_0 - pC \tag{2.28}$$

$$W_2 = W_0 - L + (1-p)C. \tag{2.29}$$

The buyer's expected utility is now written as

$$\bar{u}(W_1, W_2) = (1-\pi)u(W_1) + \pi u(W_2). \tag{2.30}$$

An indifference curve corresponding to this expected utility function is shown in Figure 2.3. Since $u(W)$ is strictly concave and the cross-derivative $\bar{u}_{12}$ is identically zero, the function $\bar{u}(W_1, W_2)$ is strictly concave and therefore strictly quasiconcave, and so the indifference curve has the curvature familiar from the standard model of the consumer. Its slope at a point (in absolute value) is given by

$$-\frac{dW_2}{dW_1} = \frac{(1-\pi)}{\pi}\frac{u'(W_1)}{u'(W_2)}. \tag{2.31}$$

Note therefore that at a point on the 45° line, generally called the certainty line, because along it $W_1 = W_2$, this becomes equal to the probability ratio or "odds ratio" $(1-\pi)/\pi$.

Now, solving for C in (2.28), substituting into (2.29) and rearranging gives

$$(1-p)[W_0 - W_1] + p[(W_0 - L) - W_2] = 0 \tag{2.32}$$

or

$$(1-p)W_1 + pW_2 = W_0 - pL. \tag{2.33}$$

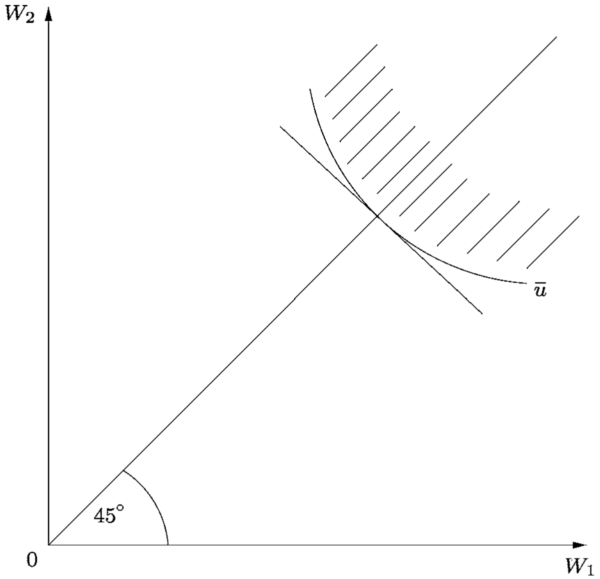

Fig. 2.3 Indifference curve.

We can interpret this as a standard budget constraint, with $(1-p)$ the price of W_1, p the price of W_2, and $W_0 - pL$, "endowed wealth," a constant, given p. The point where $W_1 = W_0$, $W_2 = W_0 - L$ clearly satisfies this constraint. Thus, we can draw the constraint as a line with slope $-(1-p)/p$, passing through the point $(W_0, W_0 - L)$, as shown in Figure 2.4. The interpretation is that by choosing $C > 0$, the buyer moves leftward from the initial endowment point $(W_0, W_0 - L)$, and, if there are no constraints on how much cover can be bought, all points on the line, including the fully insured wealth, W_F, are attainable. The price ratio or rate of exchange of the state contingent incomes is $(1-p)/p$. The demand for insurance can now be interpreted as the demand for W_2, wealth in the loss state. Note that the budget line is flatter, the higher is p.

The elimination of cover C to obtain this budget constraint in (W_1, W_2)-space is more than just a simple bit of algebra. It can be interpreted to mean that what an insurance market essentially does is to make available a budget constraint that allows the exchange of state

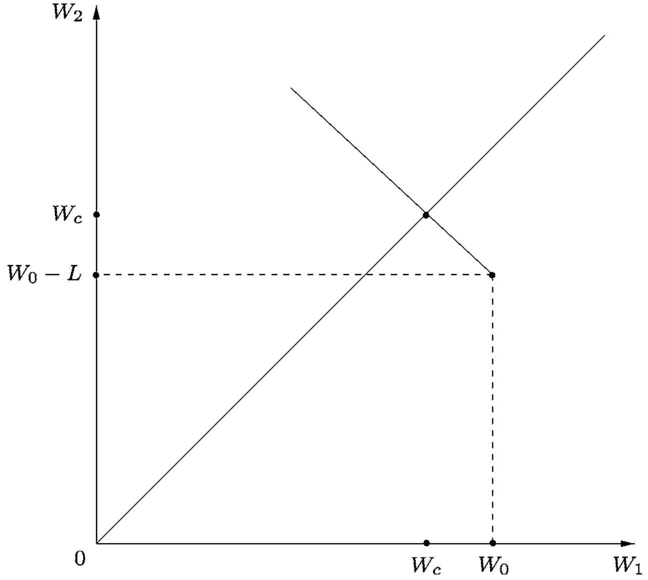

Fig. 2.4 Budget constraint.

contingent wealth: buying insurance means giving up wealth contingent on the no-loss state in exchange for wealth in the loss state, at a rate determined by the premium rate in the insurance contract.

Solving the problem of maximizing expected utility in (2.30) subject to the budget constraint (2.33) yields first-order conditions on the optimal state contingent incomes

$$(1 - \pi)u'(W_1^*) - \lambda(1 - p) = 0 \tag{2.34}$$

$$\pi u'(W_2^*) - \lambda p = 0 \tag{2.35}$$

$$(1 - p)W_1^* + pW_2^* = W_0 - pL. \tag{2.36}$$

The first two can be expressed as

$$\frac{(1 - \pi)}{\pi}\frac{u'(W_1^*)}{u'(W_2^*)} = \frac{1 - p}{p}, \tag{2.37}$$

which has the interpretation of equality of the marginal rate of substitution with the price ratio, or tangency of an indifference curve with budget line. Writing this condition as

$$u'(W_1^*) = \frac{\pi}{p}\frac{(1-p)}{(1-\pi)}u'(W_2^*) \tag{2.38}$$

allows us to derive the results

$$p = \pi \Leftrightarrow u'(W_1^*) = u'(W_2^*) \Leftrightarrow W_1^* = W_2^* \tag{2.39}$$

$$p > \pi \Leftrightarrow u'(W_1^*) < u'(W_2^*) \Leftrightarrow W_1^* > W_2^* \tag{2.40}$$

$$p < \pi \Leftrightarrow u'(W_1^*) > u'(W_2^*) \Leftrightarrow W_1^* < W_2^*. \tag{2.41}$$

Referring back to (2.28) and (2.29), equal state contingent wealth must imply full cover, a higher wealth in the no loss state must imply partial cover, and a higher wealth in the loss state must imply more than full cover. Thus we have the same results as before.

This solution is illustrated in Figure 2.5. Define the *expected value* or *fair odds line* by

$$(1-\pi)W_1 + \pi W_2 = \bar{W} = W_0 - \pi L. \tag{2.42}$$

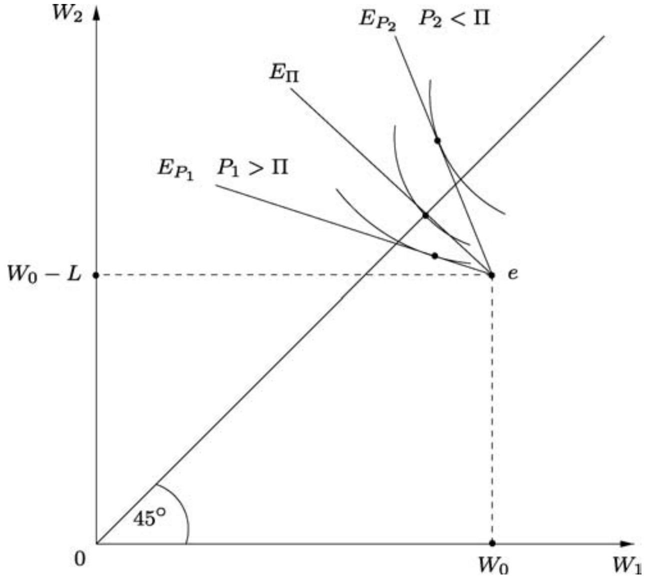

Fig. 2.5 Equilibrium.

This is clearly also a line passing through the initial endowment point. Note that any indifference curve in (W_1, W_2)-space has a slope of $(1-\pi)/\pi$ at the point at which it cuts the certainty line. Then clearly the cases of full, partial and more than full cover correspond to the cases in which the budget constraint defined by p is respectively, coincident with, flatter than, or steeper than the expected value line (see the figure), since the coverage chosen, as long as it is positive, is always at a point of tangency between an indifference curve and a budget line. Note that if the budget line is so flat that it is tangent to or intersects from below the indifference curve passing through the initial endowment point e, then we have the case where $C^* = 0$, the buyer stays at the initial endowment point and no cover is bought.

It is useful to be able to read off from the figure in state contingent wealth space the amount of cover bought. Figure 2.6 shows how to do this. Given the optimal point a, draw a line parallel to the certainty line. This therefore has a slope of 1, and cuts the line ce at b. Then the

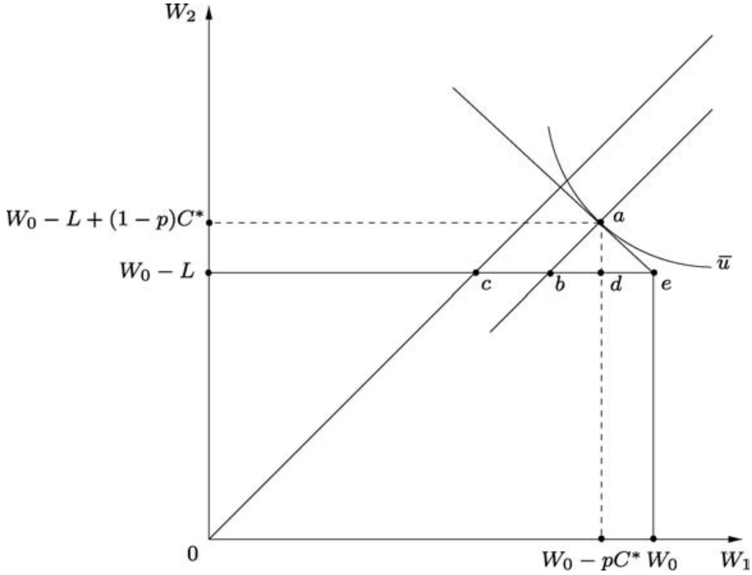

Fig. 2.6 Reading off cover.

length be represents the cover bought. To see this note that $ed = pC^*$, while $bd = ad = (1-p)C^*$. So $be = bd + de = pC^* + (1-p)C^* = C^*$.

This model allows us in principle to solve for the optimal state contingent wealth values as functions of the exogenous variables of the problem

$$W_s^* = W_s(W_0, p, L, \pi) \quad s = 1, 2. \tag{2.43}$$

Thus we have demand functions for state contingent wealth as a way of expressing the demand for insurance, alternative to that given by the demand-for-cover function. The two models are of course fully equivalent, and both are used frequently in the literature. The cover-demand model is more direct and often easier to handle mathematically. The advantage of the wealth-demand model on the other hand is that it allows the obvious similarities with standard consumer theory to be exploited, especially in the diagrammatic version. In the remainder of this survey, we will use whichever model seems more suitable for the purpose in hand.

2.3 Comparative Statics: The Properties of the Demand Functions

We want to explore the relationships between the optimal value of the endogenous variable, the demand for insurance, i.e., the demand for cover, and the exogenous variables that determine it, W_0, p, L, and π. For an algebraic treatment, the cover-demand model is the more suitable, but we also exploit its relationship with the wealth-demand model to obtain additional insights.

Recall that the first-order condition of the cover-demand model is

$$\begin{aligned}\bar{u}_C = &-p(1-\pi)u'(W_0 - pC^*)\\ &+ (1-p)\pi u'(W_0 - L + (1-p)C^*) = 0.\end{aligned} \tag{2.44}$$

Applying the Implicit Function Theorem we have that

$$\frac{\partial C^*}{\partial W_0} = -\frac{\bar{u}_{CW_0}}{\bar{u}_{CC}} \tag{2.45}$$

$$\frac{\partial C^*}{\partial p} = -\frac{\bar{u}_{Cp}}{\bar{u}_{CC}} \tag{2.46}$$

$$\frac{\partial C^*}{\partial L} = -\frac{\bar{u}_{CL}}{\bar{u}_{CC}} \tag{2.47}$$

$$\frac{\partial C^*}{\partial \pi} = -\frac{\bar{u}_{C\pi}}{\bar{u}_{CC}}. \tag{2.48}$$

We have already shown that, because of risk aversion, $\bar{u}_{CC} < 0$. Thus the sign of these derivatives is in each case the same as that of the numerator.

2.3.1 The Effect of a Change in Wealth

We have that

$$\begin{aligned}\bar{u}_{CW_0} = &-p(1-\pi)u''(W_0 - pC^*)\\ &+ (1-p)\pi u''(W_0 - L + (1-p)C^*).\end{aligned} \tag{2.49}$$

Consider first the case in which $p = \pi$ and so $C^* = L$. Inserting these values gives

$$\frac{\partial C^*}{\partial W_0} = -\frac{\bar{u}_{CW_0}}{\bar{u}_{CC}} = 0. \tag{2.50}$$

The reason is intuitively obvious. Since full cover is bought, and L stays unchanged, a change in wealth has no effect on insurance demand.

More interesting is the case in which $p > \pi$ and so $C^* < L$. In that case, from (2.49) we have $\bar{u}_{CW_0} \gtreqless 0$, i.e., the effect cannot be signed, insurance demand could increase or decrease with wealth.

This indeterminacy should not come as a surprise to anyone who knows standard consumer theory: wealth effects can typically go either way. Thus insurance cover can be an inferior or a normal good. It is however of interest to say a little more than this, by relating this term to the buyer's attitude to risk bearing. To do this we make use of the wealth-demand model. Given the optimal wealth in the two states, we have $W_1^* > W_2^*$ because of partial cover. From the first-order condition in the wealth-demand model we have

$$p(1-\pi) = \frac{(1-p)\pi u'(W_2^*)}{u'(W_1^*)}. \tag{2.51}$$

Substituting this into (2.49) gives

$$\bar{u}_{CW_0} = -u''(W_1^*)\frac{(1-p)\pi u'(W_2^*)}{u'(W_1^*)} + (1-p)\pi u'(W_2^*) \tag{2.52}$$

$$= (1-p)\pi u'(W_2^*)\left[\frac{u''(W_2^*)}{u'(W_2^*)} - \frac{u''(W_1^*)}{u'(W_1^*)}\right]. \tag{2.53}$$

Recall now the definition of the Arrow–Pratt measure of (absolute) risk aversion

$$A(W) \equiv -\frac{u''(W)}{u'(W)}. \tag{2.54}$$

We can then write

$$\bar{u}_{CW_0} = (1-p)\pi u'(y_2^*)[A(W_1^*) - A(W_2^*)]. \tag{2.55}$$

Thus

$$\bar{u}_{CW_0} \gtreqless 0 \tag{2.56}$$

according as

$$A(W_1^*) \gtreqless A(W_2^*). \tag{2.57}$$

Since $W_1^* > W_2^*$, insurance cover is a normal good if risk aversion increases or is constant with wealth ($A(W_1^*) \geq A(W_2^*)$), and an inferior good if risk aversion decreases with wealth ($A(W_1^*) < A(W_2^*)$). Since the latter is what we commonly expect, the conclusion is that we expect that insurance is an inferior good. The intuition is straightforward: if an increase in wealth increases the buyer's willingness to bear risk, then her demand for insurance falls, other things equal.

This could be bad news for insurance companies: the demand for insurance could well be predicted to fall as incomes rise. It could also be bad news for the theory, since a cursory glance at insurance market statistics shows that insurance demand has been growing with income over time. However, a resolution might well be tucked away in the "other things equal" clause. In reality, we would expect that as incomes rise, so does the value of the losses insured against. This is almost certainly true in health, life, property and liability insurance. As we now see, a *ceteris paribus* increase in the loss L increases the demand for insurance.

2.3.2 The Effect of a Change in Loss

We have that

$$\bar{u}_{CL} = -(1-p)\pi u''(W_0 - L + (1-p)C^*) > 0. \qquad (2.58)$$

Thus, as we would intuitively expect, given risk aversion, an increase in loss increases the demand for cover, *other things being equal.*

2.3.3 The Effect of a Change in Premium Rate

The effects of a price change on demand are always of central interest and importance. We have

$$\begin{aligned}\bar{u}_{Cp} = &-[(1-\pi)u'(W_1^*) + \pi u'(W_2^*)] \\ &+ [p(1-\pi)u''(W_1^*) - (1-p)\pi u''(W_2^*)]C^*. \qquad (2.59)\end{aligned}$$

This too cannot be unambiguously signed, since the first term is negative while the second could have either sign. But notice that the second term is just $-u_{CW_0}C^*$. In fact we have a standard Slutsky equation, which we can write as

$$\frac{\partial C^*}{\partial p} = -\frac{\bar{u}_{Cp}}{\bar{u}_{CC}} = \frac{(1-\pi)u'(W_1^*) + \pi u'(W_2^*)}{\bar{u}_{CC}} + C^*\frac{u_{CW_0}}{\bar{u}_{CC}}. \qquad (2.60)$$

The first term is the substitution effect, and is certainly negative ($\bar{u}_{CC} < 0$), while the second is the wealth effect and, as we have seen, could be positive or negative (or zero). If $u_{CW_0} \geq 0$, this wealth effect is negative or zero, and so the demand for cover certainly falls as the premium rate (price) rises. That is, there is no ambiguity if absolute risk aversion increases or is constant with wealth. On the other hand, if insurance is an inferior good, the wealth effect is positive and so works against the substitution effect. That is, insurance may be a Giffen good[5] if risk aversion decreases sufficiently with wealth.

The intuition is also easy to see. An increase in the premium rate increases the price of wealth in state 2 relative to that in state 1, and so, with utility held constant, W_1 will be substituted for W_2, implying

[5] Hoy and Robson (1981) were the first to show that insurance could be a Giffen good. Brys et al. (1989) generalize their analysis.

a reduced demand for cover. However, the increase in premium also reduces real wealth, to an extent dependent on the amount of cover already bought, C^*, and this will tend to increase the demand for insurance if risk aversion falls with wealth, and reduce it if risk aversion increases with wealth.

2.3.4 The Effect of a Change in Loss Probability

$$\bar{u}_{C\pi} = pu'(W_1^*) + (1-p)u'(W_2^*) > 0 \tag{2.61}$$

Thus, as we would expect, an increase in the risk of loss increases demand for cover.

Note, however, that in general we would not expect the premium rate to remain constant when the loss probability changes, though we need some theory of the supply side of the market before we can predict exactly how it would change. We would expect it to change in the same direction as the loss probability, however, which would mean combining the unambiguous effect of the change in probability with the ambiguous effect of a change in price. We should not therefore be surprised to find the overall result of a change in loss probability on the equilibrium amount of cover purchases, taking account of supply as well as demand effects, to be ambiguous.[6]

2.4 Multiple Loss States and Deductibles

The simple two-state model considered so far in this section is useful, but of course limited. One aspect of this limitation is that the idea of "partial cover" is very simple: in the single loss-state, $C < L$. In reality, when there are multiple loss states, there can be different types of partial cover. One example is the special case of *coinsurance* in which a fixed proportion of the loss is paid in each state. Another is the case of a *deductible:* nothing is paid for losses below a specified value, called the deductible, while, when losses exceed this value, the insured receives an amount equal to the loss *minus* the deductible.

[6] Note also that the idea of a "change in risk" becomes more complex and subtle when there is more than one loss state. See Eeckhoudt and Gollier (2000) for a good recent survey of the literature on the effects of a change in risk when there is a continuum of loss states.

In practice, a deductible is a much more commonly observed form of partial cover than coinsurance. We now examine one possible reason for this. It can be shown that, when offered a choice between a contract with a deductible and any other contract *with the same premium*, assumed to depend only on the expected cost to the insurer of the cover offered, a risk averse buyer will always choose the deductible.[7] This offers an explanation of the prevalence of deductibles and at the same time a confirmation of the predictive power of the theory.

We generalize the two-state model by assuming now that the possible loss lies in some given interval: $L \in [0, L_m]$, $L_m < W_0$, and has a given probability function $F(L)$ with density $f(L) = F'(L)$. Under proportional coinsurance we have cover

$$C = \alpha L \quad \alpha \in [0,1] \tag{2.62}$$

with $\alpha = 0$ implying no insurance and $\alpha = 1$ implying full cover. Under a deductible we have

$$C = 0 \qquad \text{for } L \le D \tag{2.63}$$

$$C = L - D \quad \text{for } L > D, \tag{2.64}$$

where D denotes the deductible, with $D = L_m$ implying no insurance and $D = 0$ full cover. The difference between the two contracts is illustrated in Figure 2.7, which shows cover as a function of loss.

Given the premium amount P, and an endowed wealth W_0 in the absence of loss, the buyer's state-contingent wealth in the case of proportional coinsurance is

$$W_\alpha = W_0 - L - P + C = W_0 - (1 - \alpha)L - P \tag{2.65}$$

and in the case of a deductible is

$$W_D = W_0 - L - P + C = W_0 - L - P + \max(0, L - D). \tag{2.66}$$

Figure 2.8 shows these wealth values. The important thing to note about a deductible is that for $L \ge D$, the insurance buyer is fully

[7] This was first shown in Arrow (1974). The elegant proof given here is due to Gollier and Schlesinger (1996).

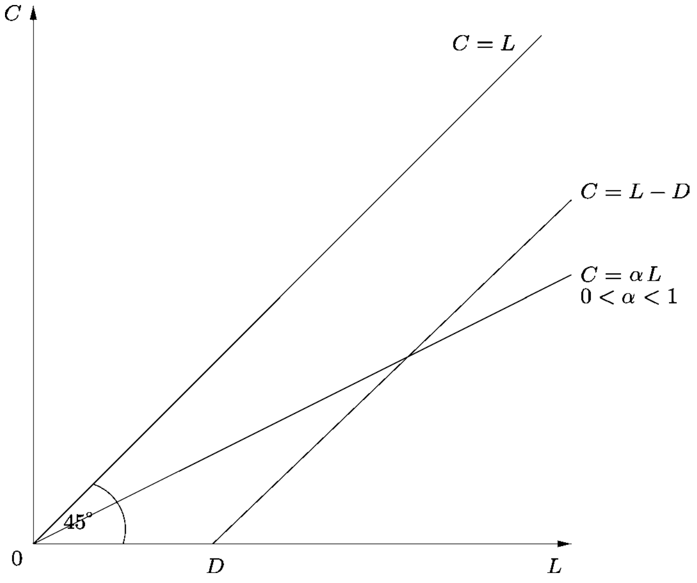

Fig. 2.7 Cover as a function of the loss.

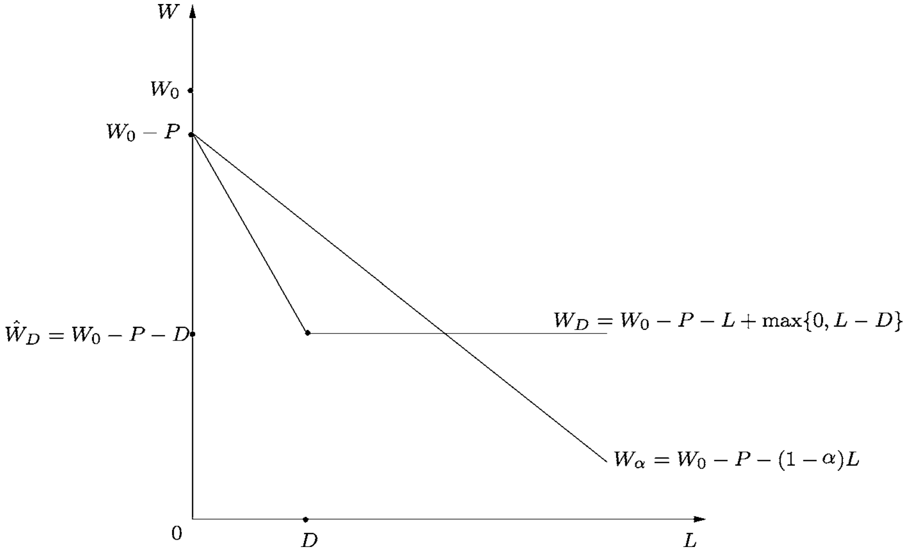
W
W_0
$W_0 - P$
$\hat{W}_D = W_0 - P - D$
$W_D = W_0 - P - L + \max\{0, L - D\}$
$W_\alpha = W_0 - P - (1 - \alpha)L$
0
D
L

Fig. 2.8 Wealth under coinsurance and deductible.

insured at the margin. For losses above the deductible, her wealth becomes certain, and equal to

$$\hat{W}_D = W_0 - L - P + (L - D) = W_0 - P - D. \tag{2.67}$$

It is this fact that accounts for the superiority, to a risk-averse buyer, of the deductible contract over other forms of contract with the same premium (and expected cost to the insurer). Under a deductible, wealth cannot fall below $\hat{W}_D$, however high the loss.

Consider now the probability distribution function for the buyer's wealth under a given deductible contract. We have

$$\Pr(W_D \leq W') \quad \text{for } W' \in [\hat{W}_D, W_0 - P] \tag{2.68}$$

$$= \Pr(L \geq L') \quad \text{for } L' = W_0 - P - W' \in [0, D] \tag{2.69}$$

$$= 1 - F(L') \tag{2.70}$$

$$\Pr(W_D < \hat{W}_D) = 0. \tag{2.71}$$

The function $H(W)$, showing $\Pr(W_D < W)$, $W \in [W_0 - P - L_m, W_0 - P]$, is illustrated in Figure 2.9. To the left of $\hat{W}_D$ it is just the horizontal axis, to the right it is given by $1 - F(L)$.

Now consider another type of contract without a deductible, but with the same expected value of cover as the deductible contract in question, and therefore the same premium. This alternative could be the proportional contract defined above, or indeed any other kind of coinsurance contract. We can show that it must have the kind of distribution function shown in the figure as $G(W)$, with the area *abc* equal to the area *cde*. But this then means that $H(W)$ is better than $G(W)$ in the sense of *second-order stochastic dominance*.[8] That is, $H(W)$ would be preferred to $G(W)$ by any risk averse buyer. $G(W)$ is riskier than $H(W)$, but has the same expected value.

[8] This is a specific way of defining an increase in risk. Take two continuous probability distribution functions $A(x)$ and $B(x)$, for x defined on an interval $[a, b]$, which need not be finite. $A(x)$ *first-order stochastically dominates* $B(x)$ if $A(x) \leq B(x)$ with strict inequality over some subinterval, and $A(x)$ *second-order stochastically dominates* $B(x)$ if $\int_a^{\hat{x}} [A(x) - B(x)]dx \leq 0$, for all $\hat{x} \in [a, b]$, with strict inequality over some subinterval. $B(x)$ is riskier than $A(x)$ in the sense that, as can be shown formally, $A(x)$ would be strictly preferred by every risk averse individual.

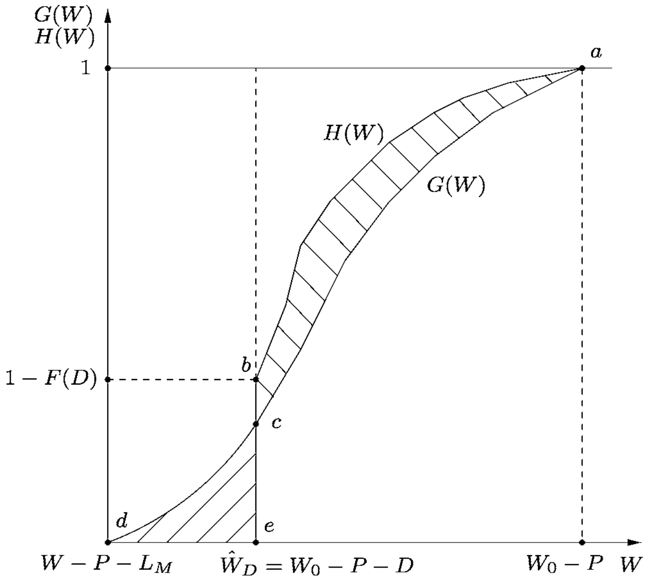

Fig. 2.9 The superiority of a deductible.

To see that any alternative to the deductible contract must have the general properties of $G(W)$, note the following points:

- If the deductible contract and the alternative contract have the same expected cost, i.e., expected value of cover, to the insurer, they imply the same expected wealth to the buyer. The expected value of wealth under each of the contracts is $E[W_0 - L - P + C]$, and so, given that $W_0, E[L]$ and P are the same in each case, $E_G[C] = E_H[C]$ implies that expected wealth is equal.
- It is a standard result that if two distributions with the same support have the same expected value, then the areas under the distribution functions are equal. Thus the areas under $H(W)$ and $G(W)$ in the figure must be the same.[9]

[9] Thus if two distributions $F(x)$ and $G(x)$, $x \in [a,b]$ have the same mean then $\int_a^b x[f(x) - g(x)]dx = 0$. Integrating by parts: $\int_a^b [F(x) - G(x)]dx = [x(F(x) - G(x))]_a^b - \int_a^b x[f(x) - g(x)]dx = 0$.

- This implies that if a contract has a lower distribution function than $H(W)$ to the right of $\hat{W}_D$ in the figure, it must have a higher distribution function to the left of $\hat{W}_D$, and the corresponding areas must be the same. Thus second order stochastic dominance applies.

The impressive aspect of this result is its generality and simplicity.

2.5 Insurance Demand with State Dependent Utility

It seems reasonable to believe that for at least some types of losses for which insurance can be bought, the utility of wealth will depend on whether or not a particular event takes place. Sickness is an obvious example. The utility of a given wealth if one is sick may well differ from that if one is healthy, and more importantly the marginal utility of wealth may differ as well. A formal extension of the model of insurance demand to this case is quite straightforward, particularly if we revert to the case of a single loss state.

We again take state 1 as the no-loss state and state 2 as the loss-state, but now denote the utility function in state $s = 1, 2$ as $u_s(W)$, with $u_1(W) > u_2(W)$, for all $(W) > 0$. Otherwise these are standard von Neumann–Morgenstern utility functions. For simplicity we assume in this subsection that insurance is offered at a fair premium, since this brings out clearly the implications of state dependent utilities. Let p therefore denote both the probability of loss and the premium rate. Formulating the insurance problem as one of choosing state contingent wealth values, the consumer solves

$$\max_{W_1 W_2} (1-p)u_1(W_1) + pu_2(W_2) \quad \text{s.t. } (1-p)W_1 + pW_2 = \bar{W}, \tag{2.72}$$

where $\bar{W}$ is the expected value of wealth. Assuming an interior solution, it is easy to see that the optimum requires

$$u_1'(W_1^*) = u_2'(W_2^*). \tag{2.73}$$

At a fair premium, the insurance buyer will always want to equalize marginal utilities of wealth across states. However, this necessarily implies equality of *wealth* across states if and only if the *marginal* *utility* of wealth is not state dependent. More generally, we want to see what this condition of equality of marginal utilities implies for the choice of wealth, and therefore of insurance cover, across states, when the marginal utility of wealth is state dependent.

We can distinguish three senses in which we could talk of "full insurance":

- choice of cover that equalizes marginal utilities of wealth across states.
- choice of cover that equalizes total utilities of wealth across states.
- choice of cover that equalizes wealth across states.

When utility is state independent and the premium is fair, these three coincide: choice of cover equalizes wealth, marginal and total utilities. Under state dependent utility, as we have just seen, marginal utilities will be equalized, but it remains an open question whether incomes and total utilities are equalized. To explore this further, we have to find an economically meaningful way of relating the state dependent utility functions to each other.

A nice way of doing this was developed by Cook and Graham (1977).[10] At every wealth level W, assume there is an amount of income $\omega(W)$ that satisfies

$$u_1(W - \omega(W)) = u_2(W). \tag{2.74}$$

We could define $\omega(W)$ as the consumer's maximal willingness to pay to be in the "good" state 1 rather than the "bad" state 2. The notation emphasizes that this willingness to pay may depend on the wealth level. Figure 2.10 illustrates this in the utility-wealth space. In the figure, for any given level of W in state 2, $\omega(W)$ gives the reduction in this income level required to yield an equal level of utility in state 1. It is just the horizontal difference between the two curves. This is a useful way to describe the relationship between the curves as W changes.

[10] See also Schlesinger (1984).

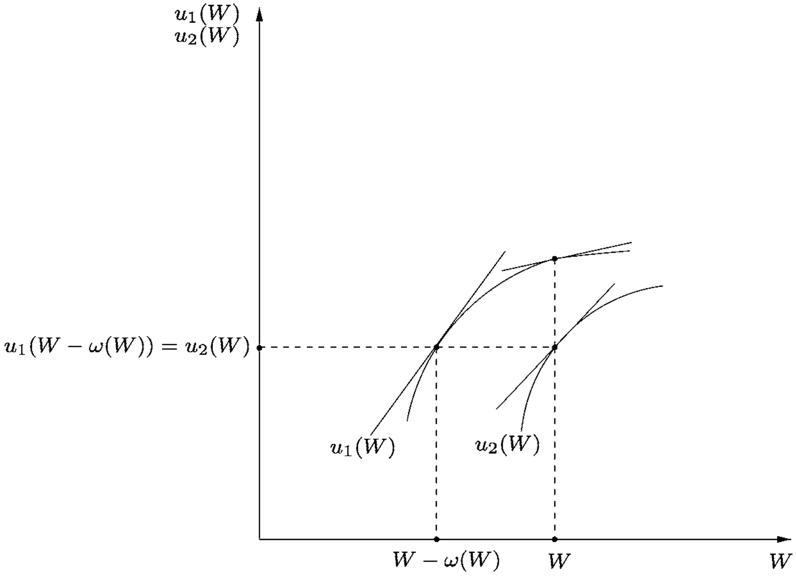

Fig. 2.10 State dependent utility.

To develop this further, since (2.74) is an identity, differentiating through with respect to W gives

$$u_1'(W-\omega(W))[1-\omega'(W)] = u_2'(W) \tag{2.75}$$

or

$$\omega'(W) = 1 - \frac{u_2'(W)}{u_1'(W-\omega(W))}. \tag{2.76}$$

Thus the way in which the willingness to pay changes as wealth varies is determined by the slopes of the utility functions at equal utility values. It seems reasonable to assume $\omega'(W) \geq 0$. For example, we would expect the willingness to pay to be healthy rather than sick at least not to fall with wealth. Thus we have

$$\omega'(W)=0 \Rightarrow u_2'(W) = u_1'(W-\omega(W)) \Rightarrow u_2'(W) > u_1'(W) \tag{2.77}$$

$$\omega'(W)>0 \Rightarrow u_2'(W) < u_1'(W-\omega(W)) \Rightarrow u_2'(W) \gtreqless u_1'(W) \tag{2.78}$$

for all W.

To see the effects of state dependent utility on the insurance decision, given the optimality condition in (2.73), we move to the state contingent wealth space in Figures 2.11 and 2.12. In the case where utility is not state dependent, we regard the 45° line as the certainty line, because equality of wealth implies equality of utilities. In the state dependent utility case, the 45° line still corresponds to certainty of *wealth*, but it no longer implies certainty of *utility*. A point on this line implies that utility in state 2 is below that in state 1 (refer back to Figure 2.10). In order to determine a locus of points at which utility across states is equal, i.e., certain, we know from (2.74) that we have to subtract $\omega(W)$ from each wealth level in state 2, the bad state, to obtain the wealth level in state 1, the good state, that yields the same utility level. Where $\omega'(W) = 0$, this implies the line shown as WW in Figure 2.11, whereas when $\omega'(W) > 0$ we have the curve WW in Figure 2.12.

To analyze the insurance decision, take first the case shown in Figure 2.11. The initial wealth values are as shown at point A, and the line passing through this point has slope $-(1-p)/p$. The optimality

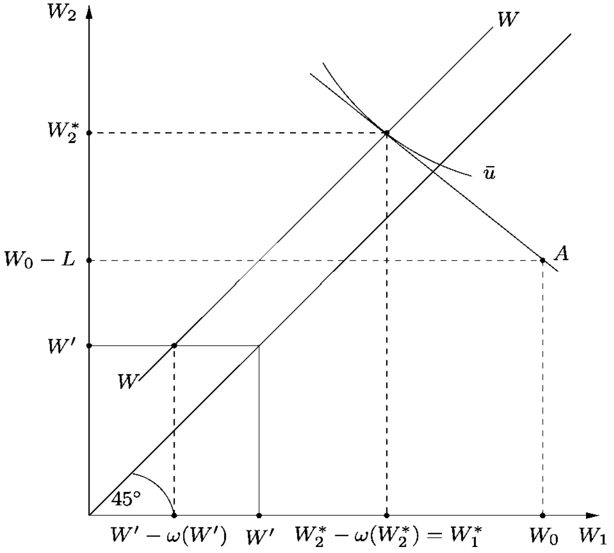

Fig. 2.11 Optimal insurance.

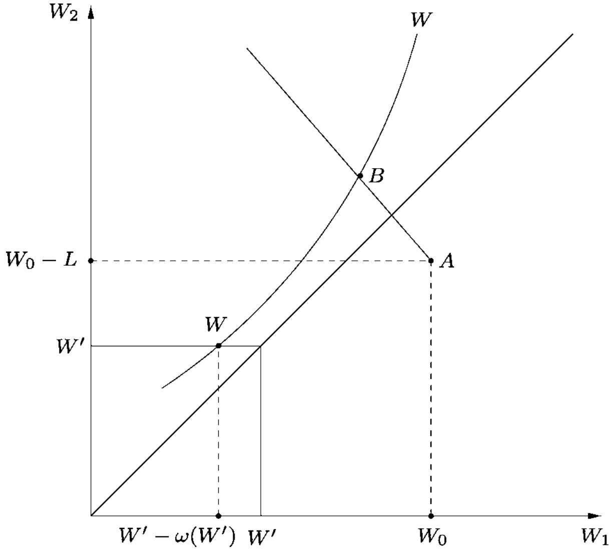

Fig. 2.12 Equal utility.

condition, given the fair premium, is that state contingent wealth after insurance cover is chosen must satisfy $u_1'(W_1^*) = u_2'(W_2^*)$. But if $\omega'(W) = 0$, (2.77) shows that we must have

$$u_2'(W_2^*) = u_1'(W_2^* - \omega(W_2^*)) \tag{2.79}$$

implying

$$W_1^* = W_2^* - \omega(W_2^*). \tag{2.80}$$

The tangency between budget constraint and indifference curve must take place on the line WW in Figure 2.11, because marginal utilities of wealth are equal along this line. Then there is full insurance of utilities, in the sense that $u_1(W_1^*) = u_2(W_2^*)$, i.e., utilities are equalized across states. But there is more than full insurance of wealth, since $W_2^* > W_1^*$.

If $\omega'(W_1^*) > 0$, the optimum cannot lie on WW in Figure 2.12, because, from (2.76), along that curve the marginal utility of wealth in state 2 is less than that in state 1. An optimal point must lie on the budget line to the right of where it intersects WW'. Thus utility remains less than fully insured, in the sense that $u_1(W_1^*) > u_2(W_2^*)$.

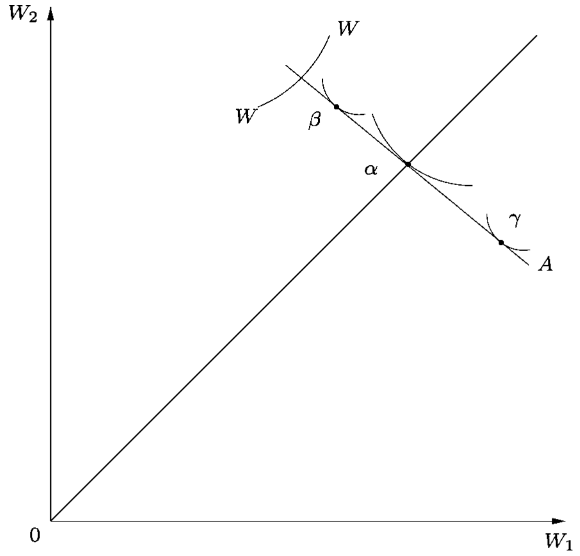

Fig. 2.13 Equilibrium possibilities.

That is all that can be said without making further assumptions about the relation between marginal utilities of income in the two states. Three cases are possible, as Figure 2.13 illustrates:

(a) $u_2'(W) = u_1'(W)$ at each W, so *marginal* utilities are state independent. Then the optimum is at α, where wealth is fully insured;

(b) $u_2'(W) > u_1'(W)$, so that at a given wealth, increasing wealth increases utility more in the bad state than in the good. Then the corresponding indifference curve through α must be flatter than the budget line, and the optimum must be at a point such as β, where more than full wealth insurance is bought;

(c) $u_2'(W) < u_1'(W)$, so that at a given wealth, increasing wealth increases utility more in the good state than in the bad. Then the corresponding indifference curve through α must be steeper than the budget line, and the optimum is at a point like γ, where less than full wealth insurance is bought.

An interesting implication of this analysis is that an insurance contract that restricts cover to the loss actually incurred — actual loss of income from employment, actual medical costs, in the case of health insurance — may not be optimal if marginal utility of wealth is state dependent. For example, if the marginal utility of wealth when ill is lower than that when healthy, then full cover would be suboptimal, because the individual would prefer to have more wealth when healthy and less when sick.

2.6 Insurance Demand and Uninsurable Risk

Up until now, it has been assumed that the insurance buyer faces only one type of loss, and insurance against this can always be bought. In reality insurance markets are typically *incomplete*, in the sense that not all risks an individual faces can be insured against. Thus one can buy insurance against wealth loss arising from ill health, but not against a loss due to fluctuations in business conditions leading to loss of overtime, short-time working, and loss of bonuses. In other words, part of one's wealth may be subject to "background risk" as it is called in the literature, which cannot be insured against. We now want to examine, in the simplest possible model, the effect the existence of an uninsurable risk can have on the purchase of insurance against an insurable risk, as well as the question of whether a welfare loss arises from the absence of a market for insurance against one of the risks. We know that the absence of a market cannot make the insurance buyer better off — one can always choose not to use a market if it is not optimal to do so. The question is whether the consumer is thereby made strictly worse off. Throughout this section, we assume that it is in principle possible to buy more than full cover against the risk of loss L. This could be perhaps because one can buy insurance from more than one insurer and sellers are unable to prevent this. The consequences of imposing an upper bound L on the amount of cover that can be bought should be obvious from the analysis.

Suppose an individual has a wealth W_0, and faces the loss L with probability π and a loss K with probability θ. There are then four possible states of the world, with associated wealth W_s, $s = 1,\ldots,4$ as set out in the following table. It is assumed that insurance cover C can be bought against risk of loss L at premium rate $p \geq \pi$. We are interested in the effect of the non-insurability of loss K on the buyer's choice of C.

Loss	0	L
0	$W_1 = W_0 - pC$	$W_3 = W_0 - L + (1 - p)C$
K	$W_2 = W_0 - pC - K$	$W_4 = W_0 - L + (1 - p)C - K$

The important point to note is that since only L can be insured against, it is possible to use the insurance market to transfer income only between *sets of states,* but not between all individual states. Insurance allows income to be exchanged between states 1 and 2, on the one hand, and 3 and 4 on the other, but not between 1 and 2, or between 3 and 4.

Denote the probability of state $s = 1,\ldots,4$ by ϕ_s. Clearly, since these four states are mutually exclusive and exhaustive, $\sum_s \phi_s = 1$. The exact values of these probabilities ϕ_s will depend on the nature of the stochastic relationship between the two losses. We consider here the three extreme cases:

(i) the two losses are statistically independent. In that case:
$\phi_1 = (1 - \pi)(1 - \theta)$ — neither loss occurs
$\phi_2 = (1 - \pi)\theta$ — only K occurs
$\phi_3 = \pi(1 - \theta)$ — only L occurs
$\phi_4 = \pi\theta$ — both losses occur

(ii) the two losses are perfectly positively correlated — either both occur or both do not occur. In effect then, there is only one loss, $L + K$, which for some reason can only be partially insured against. Then
$\phi_1 = (1 - \pi) = (1 - \theta)$ — neither loss occurs
$\phi_2 = \phi_3 = 0$ — we cannot have only one of the losses occurring
$\phi_4 = \pi = \theta$ — both losses occur

(iii) the losses are perfectly negatively correlated — if one occurs the other does not, and conversely. Then

$\pi = (1 - \theta)$, $\theta = (1 - \pi)$,
$\phi_1 = \phi_4 = 0$ — there is neither no loss nor both losses occurring together
$\phi_2 = \theta$ — only K occurs
$\phi_3 = \pi$ — only L occurs

The buyer will choose cover to solve

$$\max_C \bar{u}(C) = \sum_{s=1}^{4} \phi_s u(W_s) \quad \text{s.t.} \quad C \geq 0 \tag{2.81}$$

given the specific expressions for W_s in the Table. The general form of the first-order condition will be the same for cases (i)–(iii), but the interpretation will of course depend on the precise interpretation of the probabilities ϕ_s, which varies across the three cases. The first-order (Kuhn–Tucker) condition is

$$-p[\phi_1 u'(W_1^*) + \phi_2 u'(W_2^*)] + (1 - p)[\phi_3 u'(W_3^*) + \phi_4 u'(W_4^*)] \leq 0 \tag{2.82}$$

$$C^* \geq 0 \quad \bar{u}_C C^* = 0. \tag{2.83}$$

It is straightforward to show that the second-order condition is satisfied.

The condition shows that if $C^* > 0$,

$$\frac{\phi_1 u'(W_1^*) + \phi_2 u'(W_2^*)}{\phi_3 u'(W_3^*) + \phi_4 u'(W_4^*)} = \frac{(1 - p)}{p}. \tag{2.84}$$

Thus the marginal rate of substitution on the left-hand side has to be defined with reference to marginal utilities of wealth averaged over each subset of states within which state contingent wealth *cannot* be exchanged. This is simply because an increase in C reduces wealth in *both* states 1 and 2 and increases wealth in *both* states 3 and 4. In order to exchange incomes between states within a subset we would require an insurance market for the loss K.

We now want to see what effect the presence of the non-insurable risk has on the purchase of cover against the insurable risk.

Case (i): *Independence.*

Writing in the explicit expressions for the wealth W_s^* and probabilities ϕ_s, we obtain from the first-order condition

$$\frac{\pi(1-p)}{p(1-\pi)} \leq \frac{(1-\theta)u'(W_0 - pC^*) + \theta u'(W_0 - pC^* - K)}{(1-\theta)u'(W_0 - L + (1-p)C^*) + \theta u'(W_0 - L + (1-p)C^* - K)} \tag{2.85}$$

$$C^* \geq 0 \quad \bar{u}_C C^* = 0. \tag{2.86}$$

We now have to distinguish between two subcases:

(a) Fair premium, $p = \pi$. Then it is easy to see that $C^* = L$, we have full cover. Thus the background risk makes no difference to the optimal cover against L. To see this, note that the left-hand side of the condition becomes 1 in this case. If $C^* < L$, the denominator in the right-hand ratio must (because $u'' < 0$) be greater than the numerator, thus the ratio must be < 1 and the condition cannot be satisfied. If however $C^* = L > 0$ the ratio on the right-hand side is 1 and equals the left-hand side. If $C^* > L$, the numerator on the right-hand side is larger than the denominator and the condition is not satisfied.

Intuitively, one might think that, when insurance against L is available at a fair premium, one might overinsure (assuming this is feasible), to compensate for not being able to insure against K. In the independence case this intuition is false, because it simply results in expected marginal utility across the states in which L does occur becoming smaller than that across the states in which L does not occur.

(b) Positive loading, $p > \pi$. Then in that case $C^* = L$ cannot be optimal, because we just saw that the right-hand ratio would then equal 1. Assume that $L > C^* > 0$, i.e., the loading is not so large that no cover is bought. We want to know what effect on choice of cover introduction of the risk

K makes. In general, the answer depends on the precise form of the buyer's utility function. In fact we can show the following:

- Introducing K, suitably small, increases cover, if and only if absolute risk aversion decreases with wealth;
- Introducing K, suitably small, reduces cover, if and only if absolute risk aversion increases with wealth;
- Introducing K, suitably small, leaves cover unchanged, if and only if absolute risk aversion is constant.

Proof. We prove only the first, the others follow similarly. Note first that if we want to *increase* the ratio on the right-hand side of (2.85), we have to *increase* C^*, since this *reduces* both wealth values and *increases* both marginal utilities in the numerator, and *increases* both wealth values and *reduces* both marginal utilities in the denominator.

Now consider the equilibrium in the absence of the risk K. This has to satisfy the condition

$$\frac{\pi(1-p)}{p(1-\pi)} = \frac{u'(W_0 - pC^*)}{u'(W_0 - L + (1-p)C^*)}. \tag{2.87}$$

We know then that when we introduce K, since this leaves $\frac{\pi(1-p)}{p(1-\pi)}$ unchanged, if this reduces the value of the ratio on the right-hand side, we will have to increase C^* to restore equality. It is easy to show that the value of the ratio will be reduced (and cover therefore increased) if

$$\frac{u'(W_0 - pC^*)}{u'(W_0 - L + (1-p)C^*)} > \frac{u'(W_0 - pC^* - K)}{u'(W_0 - L + (1-p)C^* - K)} \tag{2.88}$$

that is, if

$$\frac{u'(W_0 - L + (1-p)C^* - K)}{u'(W_0 - L + (1-p)C^*)} > \frac{u'(W_0 - pC^* - K)}{u'(W_0 - pC^*)}. \tag{2.89}$$

For short, write this as

$$\frac{u'(W_3^* - K)}{u'(W_3^*)} > \frac{u'(W_1^* - K)}{u'(W_1^*)}. \tag{2.90}$$

Now assume that K is sufficiently small that we can use the simple Taylor series approximation

$$u'(W_s^* - K) \approx u'(W_s^*) - u''(W_s^*)K \quad s = 1,3. \tag{2.91}$$

Inserting these into (2.90) and canceling terms then gives

$$A(W_3^*) \equiv -\frac{u''(W_3^*)}{u'(W_3^*)} > -\frac{u''(W_1^*)}{u'(W_1^*)} \equiv A(W_1^*). \tag{2.92}$$

Since $W_3^* < W_1^*$ (partial cover), this gives the result.

Case (ii): *Perfect positive correlation.*
In this case we can show that ideally, if there is fair insurance the buyer would like to set $C^* = L + K$, i.e., over-insure on the L-market to compensate for not being able to insure against K. If $p > \pi$, the buyer would like to set $C^* < L + K$, for reasons with which we are already familiar, and so we just consider the case of a fair premium. Introducing the appropriate probabilities and incomes for this case into the first-order condition with $p = \pi$ gives

$$\frac{u'(W_0 - pC^*)}{u'(W_0 - L + (1-p)C^* - K)} = 1. \tag{2.93}$$

(Note, we can rule out the case in which $C^* = 0$ because then the ratio on the left-hand side is strictly less than one, which does not satisfy the Kuhn–Tucker condition). Clearly then this condition is satisfied if and only if $C^* = L + K$. This is then a case in which the noninsurability of K does not reduce welfare, though it does change behavior. However if, for some reason, cover is restricted in the L -market, for example by $C \leq L$, then the buyer chooses $C^* = L$ and is made strictly worse off by the non existence of the K-market.

Case (iii): *Perfect negative correlation.*
Inserting the appropriate probabilities and incomes into the first-order condition gives

$$\frac{(1-\pi)u'(W_0 - pC^* - K)}{\pi u'(W_0 - L + (1-p)C^*)} \geq \frac{1-p}{p}. \tag{2.94}$$

We take the fair premium case, in which the condition becomes

$$u'(W_0 - pC^* - K) \geq u'(W_0 - L + (1-p)C^*). \tag{2.95}$$

Suppose first that $C^* > 0$, so the condition must hold with equality. This then implies

$$pC^* + K = L - (1-p)C^* \tag{2.96}$$

or

$$C^* = L - K. \tag{2.97}$$

Now L and K are exogenous, with $L \gtreqless K$. Thus we have three possibilities:

(a) $L = K$. This implies $C^* = 0$, which is a contradiction. In fact in this case no cover is bought. The reason is that, because of the perfect negative correlation and the equality of K and L, income is certain with zero insurance cover.

(b) $L > K$. Then $C^* = L - K > 0$. In order to equalize incomes across the states, cover has to be bought which just makes up the difference between L and K.

Note a feature of these two cases: the introduction of the second risk K certainly makes a difference to the insurance decision on the purchase of cover on the market for insurance against L, but, because of the perfect negative correlation, there is no welfare loss arising from the absence of a market for insurance against K.

(c) $K > L$. Then we would have $C^* < 0$, which is assumed not to be possible, and again contradicts the assumption that $C^* > 0$. In fact in this case we have $C^* = 0$: buying positive cover would worsen the income inequality between the two states, since it reduces income in the state in which K occurs and L does not. The buyer would actually like to have negative cover, i.e., offer a bet on the occurrence of the loss L, since this would transfer income from the state in which L occurs to that in which K occurs. In this case also, the insurance decision on the L-market is certainly affected by the existence of the non insurable risk K. The buyer would be made better off if the K-market existed and the L-market did not.

2.7 Conclusions

In this section we have carried out a quite thorough, but by no means exhaustive,[11] analysis of the determinants of the demand for insurance, treating it as the outcome of expected utility maximization. The basic theorem, due to Mossin, is that a risk averse individual offered insurance at a fair premium will always choose full cover. On the other hand, there is a wide range of cases, ranging from a positive loading on the insurance premium, through state dependent utility to negatively correlated background risk, in which the buyer would prefer to have partial cover. An important further result was that, other things being equal, the insured would prefer to have this partial cover in the form of a deductible. These results are however confined to the demand side of the market. In the rest of this survey, we will encounter explanations for insurance contracts with partial cover arising out of the supply side of the market, first as a result of positive marginal costs of supplying cover, which provide a rationale for a positive loading, and then as a consequence of asymmetric information between insurance buyers and sellers. We will continue to be interested in the conditions under which this partial cover optimally takes the form of a deductible.

[11] One very special characteristic of the model was the assumption of only one loss state. In Section 3.4, we consider a model with a continuum of loss states, but do not carry out the comparative statics or consider the problem of an uninsurable background risk. Both these analyzes become considerably more complex with a continuum of loss states. For good recent surveys see the papers by Eeckhoudt and Gollier, Gollier and Schlesinger in Dionne (2000).

3

The Supply of Insurance

3.1 Introduction

In modeling the market supply of goods in general, we proceed by first developing a theory of the firm, and then analyzing its supply behavior. The key underlying relationship is the production function, showing how the output quantities that can feasibly be produced vary with the input quantities used. The general properties of this function are important because they determine the nature of the firm's costs, in particular how they vary with output. The production function and its properties are treated in a very general way. As economists we are not interested in the details of the engineering or technological relationships involved in producing some specific good, but only in their broad characteristics — the behavior of marginal productivity as input quantities vary, the nature of the returns to scale — that allow us to put restrictions on the form of the firm's cost function.

In most of the economics literature on insurance markets, a much simpler approach is taken. It is just assumed that the market is "competitive," the "production costs" of insurance are zero, and as a result there is a perfectly elastic supply of insurance cover at a fair premium.

This approach can be justified when the purpose of the model is to analyze specific issues that would only be unnecessarily complicated by a more complete specification of the supply side of the market. Later in this survey, for example, we shall see this in the analysis of the implications of information asymmetries for the existence and optimality of insurance market equilibrium. It will not suffice however for all aspects of the analysis of insurance markets. Therefore in this section, we consider more explicit models of the supply side of the insurance market.

Our first concern will be with the "technology" of insurance. This has two aspects. On the one hand, there are the activities involved in physically "producing" insurance: drawing up and selling new insurance contracts, administering the stock of existing contracts, processing claims, estimating loss probabilities, calculating premiums, and administering the overall business. The costs involved in these activities are often referred to as "transactions costs," but since they clearly extend beyond what in the economics literature are normally referred to as transactions costs, we will call them *insurance costs*. In Section 3.4, therefore we examine the model of Raviv (1979), which analyzes the implications of the existence of insurance costs for the design of insurance contracts on a competitive market.

The other aspect of insurance technology is conceptual rather than physical, and concerns the *pooling* and *spreading* of risk. When an insurer enters into insurance contracts with a number of individuals, or a group of individuals agrees mutually to provide insurance to each other, the probability distribution of the aggregate losses they may suffer differs from the loss distribution facing any one individual. We are interested in the nature of this aggregate loss distribution. In particular, we want to establish its properties as the number of individuals insured becomes large. This is the topic of risk pooling. In addition, the insurer will typically not be a single individual, but rather a group of individuals. Each member of this group may face *unlimited liability*, in the sense that he will be liable to meet insurance losses to the full extent of his wealth; or *limited liability*, where his possible loss is limited to the extent of his shareholding. An example of the former is the Lloyd's syndicate, where a group of individuals known as "names" finances the sale of insurance to non-members of the group. In the latter case we have the standard insurance *company* or *firm*. Also important is the mutual insurance company, in which the insureds are also the shareholders in the company, with limited liability. In each case the insurance losses are being spread over a number of individuals, and we are interested in the question of how this affects the premium that would be set, given that the individuals may themselves be risk averse.

Finally, a very important aspect of an insurer's operations are its investment activities. These arise in two ways. As we shall see, the insurer will have to hold reserves against the possibility that the aggregate value of loss claims will exceed its premium income. These will be invested in assets that yield a return. Second, since under every insurance contract premium revenue is collected in advance of the payout of any corresponding claim, this provides a flow of investible funds. For both these reasons large insurers are also major financial institutions. It is therefore of interest to examine how these two sides of the business, insurance and investment, interact. In this context, it is also interesting to consider the issue of the solvency regulation of insurance companies.

3.2 Risk Pooling

We assume that the insurer enters into insurance contracts with n individuals, and we make the further assumption that the distribution of claims costs under each contract is identical, and independent across contracts. This assumption of *identically and independently distributed* (*i.i.d.*) risks is not essential for determining the aggregate claims distribution, but is very helpful in greatly simplifying the technicalities involved, while losing little of interest to the economist. Thus each contract is assumed to have the same probability distribution of cover, and therefore of loss claims, $\tilde{C}_i$, with mean μ and variance σ^2, both finite, and with zero covariance between any pair of values C_i, C_j, $i, j = 1, \ldots, n$, $i \neq j$. It follows from the standard properties of the sum of *i.i.d.* random variables that $\tilde{C}^n = \sum_{i=1}^{n} \tilde{C}_i$, is also a random variable with mean $n\mu$. One immediate implication of this is that if the insurer sets the premium on each contract equal to the expected value of cover or claims cost μ, and insurance costs are zero, it will just break even *in expected value*, since its total premium revenue $n\mu$ will equal the expected value of claims costs. This is the reason for calling μ the fair premium.

We find the variance of $\tilde{C}^n$ as

$$E\left[\left(\sum_{i=1}^{n}\tilde{C}_i - n\mu\right)^2\right] = E\left[\left\{\sum_{i=1}^{n}(\tilde{C}_i - \mu)\right\}^2\right] = \sum_{i=1}^{n}E[(\tilde{C}_i - \mu)^2] = n\sigma^2. \tag{3.1}$$

Note that the variance of the aggregate claims cost increases linearly with n. It must be emphasized that any one realization of $\tilde{C}^n$, that is, actual aggregate claims costs in any one period, may be larger or smaller than $n\mu$, no matter how large the number of contracts sold, since the variance $n\sigma^2$ is always positive and increases with n. If the insurer is to avoid *insolvency*, i.e., the situation in which claims costs exceed the funds available to meet them, it will have to carry what are called technical or insurance reserves.

Now, it is reasonable to assume that each contract has a maximum cover $C_{\max}$, and so there is a maximum possible aggregate claims cost $nC_{\max}$. Thus, in principle, if the insurer sets a premium amount P per contract and also carries reserves (ignoring for the moment investment income) $R_{\max} = n(C_{\max} - P)$, it will have a zero probability of insolvency. In practice, however, the probability that actual claims costs will be in the region of $nC_{\max}$ is typically vanishingly small, while attempting to raise a capital of $R_{\max}$ could be extremely costly. Consequently, insurers proceed by choosing a so-called *ruin probability*, which we denote by ρ, and, given the distribution of aggregate claims costs, they then choose a level of reserves $R(\rho) = C_\rho - nP$, where C_ρ satisfies

$$\Pr[\tilde{C}^n > C_\rho] = \rho.$$

That is, reserves are set at a level such that the probability is ρ that actual claims costs will exceed premium revenue plus reserves (again ignoring investment income) and the insurer will be insolvent.

Figure 3.1 illustrates, for the case in which the insurer sets the fair premium, $P = \mu$. The aggregate loss claims distribution is bounded

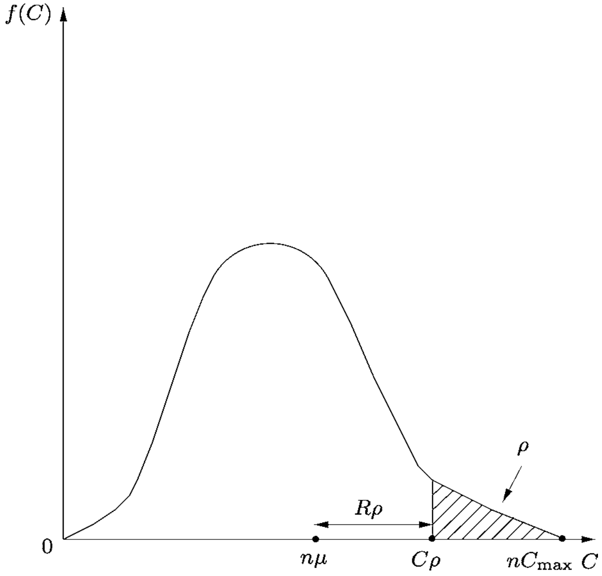

Fig. 3.1 Insurance reserves.

below by zero and above by $nC_{\max}$, and C_ρ is the value of aggregate claims such that with probability ρ the insurer will be insolvent. For a given value of ρ, the value C_ρ will typically increase with the number of contracts n, as will the value of the required reserves $R(\rho)$.

It is clearly of interest to ask how the ruin probability ρ is determined. It will result from a solution to the problem of the optimal trade-off between the costs associated with the risk of insolvency, which depends in part on insurance buyers' perceptions of this risk, and the cost of holding reserves. We shall explore this problem in more detail in Section 3.5. First, we consider the implications of the *Law of Large Numbers* for the value of the loss and insurance reserves *per contract.* Consider a particular realization $C_1, C_2, \ldots, C_n$ of the claims under the n individual contracts. We can regard this as a random sample from a distribution with mean μ and variance σ^2, both finite. Let $\bar{C}_n$ denote the sample mean, or average loss per contract, i.e., $\bar{C}_n = \frac{1}{n}\sum_{i=1}^{n} C_i$. Then the version of the Law of Large Numbers relevant for present purposes[1] says that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left[\left|\bar{C}_n - \mu\right| < \varepsilon\right] = 1. \tag{3.3}$$

In words, as n becomes increasingly large, this sample mean, the average loss claim per contract, will be arbitrarily close to the value μ with probability approaching 1. Put loosely, this says that for a sufficiently large number of insurance contracts, it is virtually certain that the loss per contract is just about equal to μ, the mean of the individual loss distribution. As the number of contracts increases, so the probability that the loss per contract lies outside an arbitrarily small interval around μ goes to zero.

Consider the variance of $\bar{C}_n$. This is given by

$$\begin{aligned} E\left[\left(\frac{1}{n}\sum_{i=1}^{n} C_i - \mu\right)^2\right] &= E\left[\frac{1}{n^2}\left(\sum_{i=1}^{n} C_i - n\mu\right)^2\right] \\ &= \frac{1}{n^2} E\left[\left(\sum_{i=1}^{n} C_i - n\mu\right)^2\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \end{aligned} \tag{3.4}$$

Thus the variance of the realized loss per contract about the mean of the individual loss distribution goes to zero as n goes to infinity.

This suggests that as the number of individual insurance contracts sold by an insurer becomes very large, the risk that the realized loss *per contract* will exceed the fair premium becomes vanishingly small. We can interpret this as a type of *economy of scale*: although the variance of aggregate claims increases with n, so the insurance reserves will typically have to increase in *absolute* amount, the required reserve *per contract* tends toward zero: required reserves increase less than propor-

[1] This can be derived from Chebyshev's Inequality: Let X be a random variable with finite mean μ and finite variance σ^2. Then for every $k > 0$

$$\Pr\left[|X - \mu| < k\sigma\right] \geq 1 - \frac{1}{k^2}. \tag{3.2}$$

Note that normality of the distribution of X, in this case the realized average loss per contract, is not required. See Cummins (1991) for a very thorough discussion of Laws of Large Numbers and the Central Limit Theorem in the insurance context.

tionately with size of the insurer, measured in terms of the number of individual insurance contracts.[2]

3.3 Risk Spreading

Suppose now that the insurer is either a syndicate with N members or a company with N shareholders. It will simplify the analysis, without losing much of economic interest, if we assume that these individuals are all identical and share the net income from the insurance business equally, so that each receives a share $s = 1/N$ of this net income. The main difference between these two types of insurer in the present context is that if it is a syndicate, the total wealth of the members will have to be at least equal to the insurance reserve implied by the chosen ruin probability, while for a company, the equity capital would have to be at least this amount. That said, we will ignore the distinction for the time being, by assuming it is costless to hold reserves.[3] We also assume that the individuals are risk averse. The question of interest is: what, if any, are the implications of increasing the number of individuals N in the syndicate or company, i.e., of spreading the risky income over a larger number of individuals? The intuition would be that this should in some sense reduce the riskiness of the individual incomes and therefore reduce the risk premium that they would demand as a condition of taking a share in the insurance, thus reducing the insurance premium. Again we would have a type of economy of scale. Support for this intuition, and understanding of the precise conditions under which it holds true, is given by the *Arrow–Lind Theorem.*

This theorem has many applications over and beyond insurance markets, but is also of central importance here. It confirms the intuitive idea that the larger the number of syndicate members who share in a given distribution of wealth from a risky insurance business, the smaller the cost of the risk associated with that business, even though

[2] See Cummins (1991) for numerical examples.

[3] In other words the insurance syndicate or company can costlessly raise capital K on which it has to pay an interest rate i, and then invest it at precisely that interest rate i. This is obviously a strongly simplifying assumption. We consider the implications of relaxing it below.

the individual syndicate members are risk averse. More importantly, it makes clear a necessary condition for this result, namely that *the covariance between the member's wealth from the insurance business and his marginal utility of wealth, if he does not share in this business, must be zero.*

Thus let $\tilde{Z}$ be the aggregate wealth from the insurance business, and $E[\tilde{Z}]$ its expected value. There are N members of the syndicate or shareholders of the company, and each receives a share $s = 1/N$ of the random wealth $\tilde{Z}$. Assume each member has an identical risk averse utility function $u(\cdot)$ and non-insurance wealth $\tilde{W}$, which may be a random variable. The key condition is that $\text{Cov}[\tilde{Z}, u'(\tilde{W})] = 0$ (which of course certainly holds if $\tilde{W}$ is certain).

Now, define the variable r as the certain amount of wealth which satisfies

$$E[u(\tilde{W} + s\tilde{Z} + r) = E[u(\tilde{W})]. \tag{3.5}$$

Note that this is an identity in r, and implicitly defines r as a function of s. We could think of r as the amount the individual would require to be paid to induce her to participate in the insurance business. If this is negative, it is the amount she would pay for a share in the business. It is obvious that as $N \to \infty$, i.e., as $s \to 0$, we have $r \to 0$. For example, a risk averse decision taker with a certain income would always be indifferent about accepting a fair bet if it is small enough — to the first order expected utility would be unchanged. What the theorem shows, however, is the somewhat less obvious fact that, on the given assumptions, the *sum* $Nr(s) = r(s)/s$ goes to $-E[\tilde{Z}]$ as N goes to infinity. We can interpret this as saying that for sufficiently large N, the aggregate market value of the insurance business can be taken as the expected value of its net income — we can treat the insurer as risk neutral. We now show this.

First, we can apply the Implicit Function Theorem to obtain, since the right-hand side of (3.6) is independent of s,

$$\frac{dr}{ds} = -\frac{E[u'(\tilde{W} + s\tilde{Z} + r)\tilde{Z}]}{E[u'(\tilde{W} + s\tilde{Z} + r)]}. \tag{3.6}$$

Now consider

$$\lim_{s \to 0} \frac{r(s)}{s}. \tag{3.7}$$

Since both numerator and denominator go to zero, we apply l'Hôpital's Rule

$$\begin{aligned}
\lim_{s \to 0} \frac{r(s)}{s} &= \lim_{s \to 0} \frac{dr(s)/ds}{ds/ds} && (3.8)\\
&= \lim_{s \to 0} -\frac{E[u'(\tilde{W} + s\tilde{Z} + r)\tilde{Z}]}{E[u'(\tilde{W} + s\tilde{Z} + r)]} && (3.9)\\
&= -\frac{E[u'(\tilde{W})\tilde{Z}]}{E[u'(\tilde{W})]} && (3.10)\\
&= \frac{-E[u'(\tilde{W})]E[\tilde{Z}] - \mathrm{Cov}(u'(\tilde{W}), \tilde{Z})}{E[u'(\tilde{W})]}. && (3.11)
\end{aligned}$$

Given that $\mathrm{Cov}(u'(\tilde{W}), \tilde{Z}) = 0$ we have

$$\lim_{s \to 0} \frac{r(s)}{s} = -E[\tilde{Z}]. \tag{3.12}$$

Thus, the aggregate value of the insurance business to the participants is equal to its expected value, with no adjustment for risk, if and only if the uncertain net income has a zero covariance with the individuals' marginal utility of income from outside the business. Note that if this covariance were positive, implying, since $u'' < 0$, a negative covariance between $\tilde{W}$ and $\tilde{Z}$, the aggregate value of the insurance business to its shareholders would exceed its expected value, and conversely if the covariance were negative. In the former case, the insurance business offers the shareholders a way of diversifying their asset portfolios.

The Arrow–Lind Theorem provides a basis for the assumption, often made in the microeconomics of insurance markets, that the insurer is risk neutral. $\tilde{Z}$ is the net wealth from the insurance business, $\tilde{W}$ would be the value of an individual's total human capital and net financial wealth, and so the assumption of zero covariance between these may or may not be regarded as plausible. One important reason why this may not be so is that the net wealth from the insurance business may itself contain the returns to the insurance company's asset portfolio, which is likely to be correlated with the return to an individual's portfolio. If the assumption of independence is rejected, recourse must be made to theories of the financial market, such as the CAPM, to find a way of valuing the returns of the insurance firm, and an objective function for it. This is not entirely without problems. For example, if the conditions under which the CAPM is valid are satisfied, it is hard to see why insurance markets would be necessary, since use of the capital market would allow individuals to construct portfolios that allow them to achieve an optimal allocation of state contingent income. Insurance markets would only be required when the capital market is incomplete,[4] but in that case it does not provide a clear way of valuing an insurance company, or indeed any other type of company. This is a paradox which seems to have received little attention in the insurance literature, but further consideration of it in this survey would take us too far afield.

3.4 Insurance Costs: The Raviv Model

We now analyze the implications of introducing insurance costs for a risk neutral insurer supplying insurance in a perfectly competitive market. In doing this we adapt a model first formulated by Raviv (1979), extending previous work of Arrow (1974) and Borch (1962). The importance of the Raviv model is that it shows how the existence of deductibles and coinsurance in the (equilibrium) insurance contract is related to the nature of insurance costs. Raviv used the methods of dynamic optimization to derive rigorously the results of the model. However, the main results can be established, albeit less rigorously, with a much simpler approach. We set this out here.

We take a single representative insurance buyer and a single representative insurance seller. The buyer faces a loss $L \in [0, L_m]$, with distribution function $F(L)$ and associated density $f(L)$ everywhere positive on this interval. Her wealth in state L is

$$W(P, C(L), L) = W_0 - P - L + C(L), \tag{3.13}$$

[4] In the sense that the the linear space spanned by the returns vectors of the available assets has smaller dimension than the set of states of the world. For further discussion of this at increasing levels of abstraction see Gravelle and Rees (2004, Chap. 21), Mas-Colell et al. (1996, Chap. 19), and Magill and Quinzii (1996).

where P is the premium amount under the insurance contract and $C(\cdot)$ is cover as a continuous function of loss, and we have the restriction

$$C(L) \geq 0 \tag{3.14}$$

at every L. The buyer's utility function is $u(W)$, with $u' > 0$, $u'' < 0$, she is strictly risk averse. The seller's cost function is $F + \gamma(C(L))$, where $F \geq 0$ is a fixed administrative and transactions cost per insurance contract, independent of the total number of contracts sold; $\gamma(C(L))$ is the variable insurance cost function, giving the cost of the insurance seller as a function of the amount paid by the insurer in the event of a loss claim L.[5] We assume

$$\gamma(0) = 0, \quad \gamma'(\cdot) \geq 0, \quad \gamma''(\cdot) \geq 0 \tag{3.15}$$

That is, the marginal cost of cover may be zero or positive, and, if positive, may be constant or increasing. If marginal cost is zero for all values of cover, then total variable costs must also be zero and only the fixed cost per contract F is relevant. The seller's income in state L is

$$z(P, C(L)) = P - C(L) - F - \gamma(C(L)). \tag{3.16}$$

We assume that the insurer is risk neutral in order to focus on the effects of insurance costs.[6] Note also that at any L,

$$\frac{\partial z}{\partial C} = -[1 + \gamma'(C(L))]. \tag{3.17}$$

In a competitive market, an equilibrium contract must maximize the expected utility of the representative buyer, subject to a zero expected profit constraint of the representative seller. If the former were not the case, any firm could profit from offering buyers a better contract than that currently on offer, while if the latter were not the case, entry or exit

[5] It may be hard to see why the cost incurred for any one buyer should increase with the amount of cover sold just to her. However, if all buyers are identical and costs increase with the total amount of cover sold for a fixed number of buyers, then we could regard this as being captured by the function $\gamma(\cdot)$. The implications of this function being identically zero are pointed out below.

[6] If the insurer is risk averse then we know from the general theory of risk sharing that the equilibrium will involve less than full cover even in the absence of insurance costs. See for example Gravelle and Rees (2004, Chap. 21).

of firms would take place. We therefore find the equilibrium contract by solving the problem

$$\max_{P,C(L)} \bar{u} = \int_0^{L_m} u(W(P,C(L),L))dF \tag{3.18}$$

$$\text{s.t.} \quad \int_0^{L_m} z(P,C(L))dF = 0 \tag{3.19}$$

$$C(L) \geq 0. \tag{3.20}$$

The Lagrange function for this problem is

$$\Lambda = \int_0^{L_m} u(W(P,C(L),L))dF + \lambda \int_0^{L_m} z(P,C(L))dF, \tag{3.21}$$

where it should be noted that the multiplier λ is independent of the state variable L. The first order (Kuhn–Tucker) conditions are

$$f(L)\{u'(W(P^*,C^*(L),L)) - \lambda^*[1 + \gamma'(C^*(L))]\} \leq 0$$

$$C^*(L) \geq 0 \quad C^* \frac{\partial \Lambda}{\partial C} = 0 \tag{3.22}$$

$$-\int_0^{L_m} u'(W(P^*,C^*(L),L))dF + \lambda^* = 0 \tag{3.23}$$

$$\int_0^{L_m} z(P^*,C^*(L))dF = 0 \tag{3.24}$$

We assume in everything that follows that cover is positive for at least some L, so that $P^* > 0$, otherwise there is nothing to talk about. Note that we treat the problem as one in *pointwise* maximization with respect to C at each $L \in [0, L_m]$.

We consider the possibility of the following types of contract (see Figure 3.2):

- *Deductible contract.* Over some interval of losses $[0, D^*]$ there is zero cover, while over the interval $(D^*, L_m]$ cover is positive.
- *Non-deductible contract.* Cover is positive over the entire interval $[0, L_m]$, $D^* = 0$.

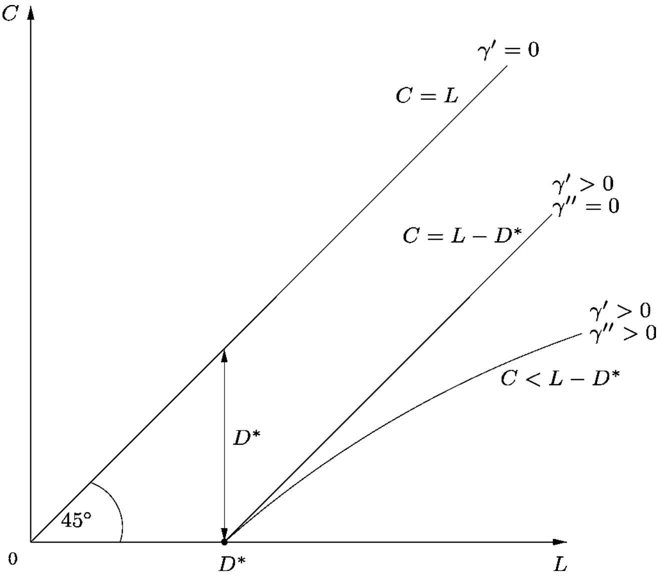

Fig. 3.2 Types of insurance contracts.

In each case, it is also of interest to ask about the relationship between loss and cover when cover is positive: is there

- *full cover above a deductible* ($C^* = L - D^*$),
- *full cover* ($C^* = L$), or
- *coinsurance above a deductible* ($0 < C^*(L) < L - D^*$)?

In fact we shall show the following:

- There is a deductible contract if and only if $\gamma'(\cdot) > 0$, marginal cost is positive.
- There is full cover if and only if $\gamma'(\cdot) = 0$, marginal cost is zero.
- There is coinsurance above a deductible if and only if $\gamma'(\cdot) > 0$ and $\gamma''(\cdot) > 0$, marginal cost is positive and increasing.
- If marginal cost is positive and constant there is full cover above a deductible.

First, we show briefly why, when the optimal contract involves partial cover, this must take the form of a deductible, i.e., zero cover over some *initial interval of loss.*

The Kuhn–Tucker conditions imply that if at some $L' > 0$ we have $C^*(L') = 0$, then $C^*(L) = 0$ for all $L < L'$, while if at some L'' we have $C^*(L'') > 0$ then we have $C^*(L) > 0$ for all $L > L''$. Thus, if there is not full cover at all loss levels, we have a deductible-type of contract. The key point is that $u'(W(P^*, C(L), L))$ is *increasing* in L for given C, and *decreasing* in C for given L, while $\lambda^*[1 + \gamma'(C(L))]$ is constant or increasing in C. Thus using the Kuhn–Tucker condition (3.23):

$$\begin{aligned} u'(W(P^*, C^*(L), L)) < u'(W(P^*, 0, L')) &\leq \lambda^*[1 + \gamma'(0)] \\ &\leq \lambda^*[1 + \gamma'(C^*(L))] \end{aligned} \tag{3.25}$$

for all $L < L'$, and this rules out the possibility of $C^*(L) > 0$.

Likewise we must have

$$\begin{aligned} u'(W(P^*, 0, L)) &> u'(W(P^*, C^*(L''), L'')) \\ &= \lambda^*[1 + \gamma'(C^*(L''))] \geq \lambda^*[1 + \gamma'(0)] \end{aligned} \tag{3.26}$$

for all $L > L''$ which rules out the possibility of $C^*(L) = 0$.

We now need to see more specifically what determines the structure of the optimal contract.

Proposition 3.1. The contract has a deductible $D > 0$ if and only if $\gamma'(\cdot) > 0$.

Proof. This is equivalent to saying that the contract has no deductible, i.e., $C^*(L) > 0$ for *all* L, if and only if $\gamma'(\cdot) = 0$.

(a) $\gamma'(\cdot) = 0 \Rightarrow C^*(L) > 0$ for *all* L

We prove this by contradiction. Suppose $\gamma'(\cdot) = 0$ but $C^*(L) = 0$ on some interval, say $[0, D)$. Then we have from the first-order conditions

$$u'(W(P^*, C^*(L), L)) < \lambda^* \qquad L \in [0, D) \tag{3.27}$$

$$u'(W(P^*, C^*(L), L)) = \lambda^* \qquad L \in [D, L_m]. \tag{3.28}$$

Integrating gives

$$\int_0^D u'(W(P^*, C^*(L), L))dF < \lambda^* \int_0^D dF \tag{3.29}$$

$$\int_D^{L_m} u'(W(P^*, C^*(L), L))dF = \lambda^* \int_D^{L_m} dF. \tag{3.30}$$

Adding, and recalling that $\int_0^{L_m} dF = 1$ gives

$$\int_0^{L_m} u'(W(P^*, C^*(L), L))dF < \lambda^*, \tag{3.31}$$

which contradicts condition (3.24). Thus $\gamma'(\cdot) = 0$ is a sufficient condition for $C^*(L) > 0$ for all L.

(b) $C^*(L) > 0$ for all $L \Rightarrow \gamma'(\cdot) = 0$.

Again we prove this by contradiction. Suppose $C^*(L) > 0$ for all L but $\gamma'(\hat{L}) > 0$ for some $L = \hat{L}$. Then from the first-order conditions we have

$$u'(W^*(P^*, C^*(\hat{L}), \hat{L})) = \lambda^*[1 + \gamma'(C^*(\hat{L}))]. \tag{3.32}$$

By continuity, $\gamma' > 0$ in some neighbourhood of $\hat{L}$. Integrating then gives

$$\int_0^{L_m} u'(W^*(P^*, C^*(L), L))dF = \lambda^* \int_0^{L_m} [1 + \gamma'(C^*(L))]dF > \lambda^*, \tag{3.33}$$

which again contradicts condition (3.24). Thus $\gamma'(\cdot) = 0$ is a necessary condition for there to be no deductible, $C^*(L) > 0$ for all L. □

We now want to examine the form of the relationship between optimal cover and loss when $\gamma'(\cdot) > 0$, i.e. there is a deductible. Thus we have

$$u'(W_0 - P^* - L + C^*(L)) = \lambda^*[1 + \gamma'(C^*(L))] \quad L \in (D^*, L_m]. \tag{3.34}$$

This is an identity in L, so differentiating through with respect to L we get

$$-u'' + u''\frac{dC^*}{dL} = \lambda^*\gamma''\frac{dC^*}{dL}. \tag{3.35}$$

The first-order condition gives $\lambda^* = u'/(1 + \gamma')$, and so substituting and rearranging gives

$$\frac{dC^*}{dL} = \frac{A}{A + \frac{\gamma''}{1+\gamma'}}, \tag{3.36}$$

where $A = -u''/u'$ is the Arrow–Pratt measure of risk aversion for the insurance buyer. This immediately gives the following results:

(a) If $\gamma'' = 0$, $dC^*/dL = 1$, and so, given $\gamma' > 0$, we have $C^* = L - D^*$, full insurance above a deductible.

(b) If $\gamma'' > 0$, $dC^*/dL < 1$, and so, given $\gamma' > 0$, we have $C^* < L - D^*$, coinsurance above a deductible.

Note finally that if $\gamma' = 0$, this implies $\gamma'' = 0$, in which case we have both no deductible, $D^* = 0$, and no coinsurance, $dC^*/dL = 1$, i.e., full cover given zero marginal insurance costs and a risk neutral insurer.

Consider now the nature of the premium in the optimal contract. The zero profit constraint implies that

$$P^* = F + \int_0^{L_m} [C^*(L) + \gamma(C^*(L))]dF. \tag{3.37}$$

This tells us that the optimal premium takes the form of a type of *two-part tariff*: there is a fixed charge per contract to meet the cost F, which is independent of the amount of cover, and then a charge $\int_0^{L_m} C^*(L)dF$ which is the expected cost of cover and represents a fair premium, and finally a loading $\int_0^{L_m} \gamma(C^*(L))dF$ to cover the variable insurance costs, which is also a part of the premium that varies with the amount of cover. This loading is zero if variable insurance costs are zero, in which case, since the premium offered as a function of cover is fair, the buyer will want full cover.

Note that it cannot be the case in equilibrium that an insurer includes F as a loading on the premium *per unit of cover*, since this would distort the choice of cover downward: For example the buyer would no longer choose full cover if there were zero variable insurance costs. Such a contract would always be displaced by one that required a fixed charge, independent of cover, to meet the fixed cost per contract F.[7] A contract would not contain a fixed charge only if $F = 0$. This model therefore gives very clear predictions about the form of the equilibrium insurance contract that will be offered on a competitive insurance market. Note that we must assume that F does not exceed the buyer's willingness to pay for the contract with premium P^* and cover $C^*(L)$.

Finally, we can consider the intuition for the structure of an optimal contract in the case where $\gamma'(\cdot) > 0$. Why does positive marginal insurance cost imply a deductible? The answer is easy to see from the first-order conditions. From (3.24), λ^* measures the marginal expected utility loss resulting from a marginal increase in the premium, while $[1 + \gamma'(C^*(L))]$ is the marginal increase in the premium resulting from a marginal increase in cover in state L. Thus over the interval $[0, D)$ we have, from the condition

$$u'(W^*(P^*,0,L)) < \lambda^*[1 + \gamma'(0)]. \tag{3.38}$$

that the marginal utility gain from a small increase in cover, from a value of zero, is less than the marginal utility loss resulting from the corresponding premium increase. For sufficiently small losses, it pays the buyer to accept the loss rather than pay the marginal insurance cost of meeting the claim. As the loss increases and marginal utility of income rises, however, the point is reached at which it becomes just worthwhile to pay the marginal insurance cost associated with meeting the claim. This happens at $L = D^*$.

3.5 Capital, Solvency, and Regulation

The Raviv model of the previous section solves for the optimal insurance contract on a competitive market without any consideration being given to the kinds of issues discussed in Section 3.2 earlier. It is implicitly assumed that the number of insurance contracts in aggregate is sufficiently large that the cost of reserves per contract is small enough to be ignored, and that the insurer will always provide enough reserves

[7] This is essentially the proposition that a lump sum tax is always Pareto superior to a tax per unit of output of a good, when both yield the same tax revenue.

to meet any level of claims due under the contracts. In other words, there is no risk of insolvency of the insurer.

In reality, in virtually all countries of the world that have insurance markets, there is some form of regulation of insurance companies, the main *raison d'être* of which is to protect insurance buyers against the risk of insolvency of their insurers. Why should this be necessary? Obviously, given the fiduciary nature of the insurance contract there is a risk of fraud, since a purported insurer who is paid a premium and then absconds makes a profit for sure, at least if he does not get caught and penalized. But large, well-established insurance companies that wish to remain in business for the long term would surely not need detailed regulatory intervention, over and above the standard laws of contract and corporate governance, to ensure that they carry enough reserves to meet their contractual obligations, or so it may be thought.[8]

However, we shall now present an analysis, based on contributions by Borch (1981), Finsinger and Pauly (1984), and Munch and Smallwood (1981), which shows that under limited liability, where a shareholder is liable for the debts of a company only up to the value of his shareholding, an unregulated insurer may find it optimal to put up *no* reserves against insolvency, but simply invest the premium income and declare bankruptcy if claims should exceed the proceeds of this investment. Moreover, this can be the case even when there is no opportunity cost to holding insurance reserves, though introducing such costs does raise the risk of insolvency. The reason for this result is the existence of a fundamental non-concavity in the insurer's objective function, stemming from the underlying technology of insurance, in the form of what is known as the *increasing failure rate property* that is characteristic of virtually all insurance loss-claims distributions. We will also argue however that this result depends on a further implicit assumption which, if relaxed, significantly modifies the conclusions of this analysis.

We consider an insurance company in business for the long term, and so taking decisions over an infinite time horizon, with a sequence of discrete time periods (say years). At the beginning of each year, the

[8] Indeed, one critical view of the detailed systems of regulatory control, such as that which existed in Germany up until the mid-1990's, was that their role was essentially to support an insurance market cartel.

insurer must decide on a level of reserve capital K for the insurance business, in the light of a given distribution of claims costs C, described by the distribution function $F(C)$ with (differentiable) density $f(C)$, defined over the interval $[0, C_{\max}]$. To focus on the case in which reserve capital is effectively costless to the insurer, we assume that it already owns an amount of capital sufficiently large as to be able to cover any level of claims, and which is invested on the capital market at a certain gross return $r > 1$. If some of this capital is transferred and committed to the insurance business, it can also be invested on the market at this same rate of return until the end of the period, when claims become payable.

The assumption that was left implicit in the models of Borch, Finsinger and Pauly and Munch and Smallwood is that premium income P is exogenous, and in particular independent of the level of capital chosen, and therefore of the insolvency risk of the insurer. This assumes that insurance buyers do not perceive any relationship between the insurer's capital and the likelihood that their claim will be met, but simply assume there is no solvency risk.[9] The premium income P is collected at the beginning of the period and invested, along with the insurance capital, in the riskless market asset. If at the end of the period assets $A \equiv (P + K)r$ are at least enough to meet claims costs C, then the insurer remains in business and receives a continuation value V, that is the expected present value of returns from the insurance business over all future periods. If claims costs turn out to be greater than assets, the insurance assets A are paid out and the insurer defaults on the remaining claims, losing the right to the continuation value V. Because of limited liability it does not have to pay out to claimants more than A.

The insurer can always choose to guarantee solvency by putting in enough capital, since we have assumed that the upper limit $C_{\max}$ on possible claims is finite. The question of interest is: under what circumstances would the insurer *choose* to stay solvent?

[9] In related work, which does not however consider the question of regulation, Doherty and Schlesinger (1990) and Schlesinger and Graf v. d. Schulenburg (1987) analyzed an individual's demand for insurance in the presence of insurer default risk.

It is assumed to want to maximize the expected present value of net wealth from the insurance business

$$\max_K V_0(K) = \int_0^A \left(\frac{V}{r} + K + P - \frac{C}{r}\right) dF - K \quad \text{s.t. } K \in [0, K_{\max}], \tag{3.39}$$

where $K_{\max} = (C_{\max}/r) - P$ is the capital required to ensure no default. The integrand is given by the present value of end of period wealth if solvent, consisting of the present value of its continuation value plus the present value of assets less claims. Note that the upper limit A on the integral reflects the existence of limited liability. If $C > A \equiv (P + K)r$, the insurer is insolvent, pays out the available assets, declares bankruptcy and loses the continuation value, and so the value of the integrand is in that case zero.

Now since at the beginning of each period the future is identical, we have $V = V_0(K)$, and so using this in (3.40) and rearranging gives

$$V_0(K) = \left[\int_0^A \left(K + P - \frac{C}{r}\right) dF - K\right] \Big/ \left(1 - \frac{F(A)}{r}\right). \tag{3.40}$$

So far nothing beyond differentiability has been assumed for the claims distribution F. It is an important fact of insurance technology however that actual insurance claims distributions typically belong to the class of “increasing failure rate” distributions, with the property that

$$\frac{d}{dC}\left[\frac{1 - F(C)}{f(C)}\right] = -(f'(1 - F) + f^2)f^2 < 0. \tag{3.41}$$

We now show that an important implication of this property is that *only corner solutions to the insurer’s wealth maximization problem are possible*: either it chooses $K = 0$, or $K = K_{\max}$.

Proposition 3.2. *Given the property of the claims distribution in (3.42), any solution to the insurer’s wealth maximization problem (3.40) is a corner solution.*

Proof. Suppose not, i.e., there exists a value $K^* \in (0, K_{\max})$ such that $V(K^*)$ is a maximum. Then $V_0'(K^*) = 0, V_0''(K^*) \leq 0$. Using (3.41) to evaluate these derivatives and recalling the definition of A gives

$$V_0'(K^*) = [V_0(K^*)f - (1 - F)]/\left(1 - \frac{F}{r}\right) = 0 \tag{3.42}$$

$$V_0''(K^*) = [V_0(K^*)f' + f]/\left(1 - \frac{F}{r}\right) \leq 0. \tag{3.43}$$

Then (3.43) implies

$$V_0(K^*) = (1 - F)/f \tag{3.44}$$

and substituting this into (3.44) gives

$$f^2 + f'(1 - F) \leq 0, \tag{3.45}$$

which cannot hold if the increasing failure rate property in (3.42) holds. Thus there cannot be an interior solution. By Weierstrass' Theorem, a solution to the problem must exist, since the objective function is continuous on the compact interval $[0, K_{\max}]$. Thus the optimum must be at an endpoint.

Which endpoint is optimal is given by the straightforward comparison of the values

$$V_0(0) = F(rP)(rP - \bar{C}_0)/[r - F(rP)] \tag{3.46}$$

$$V_0(K_{\max}) = (rP - \bar{C})/[r - 1], \tag{3.47}$$

where $\bar{C}$ is the mean of the claims distribution and $\bar{C}_0 = [F(rP)]^{-1} \int_0^{rP} CdF < \bar{C}$ is the mean of the claims distribution truncated at $C = rP$. This truncation represents the effect of limited liability. As these expressions clearly show, for a given premium income the advantage to not putting up any capital is that the expected present value of claims cost falls. The disadvantage is that there is a risk $1 - F(rP) > 0$ of going out of business. It is not possible in general to say that one of these endpoints is always better than the other, this will depend on the parameters of the problem. Figure 3.3 illustrates the possibilities. It shows possible shapes of the function $V_0(K)$ given the property in (3.42), and for different assumptions about where the optimum lies.

We can summarize these results as follows: Because of a particular characteristic of the technology of insurance, the property in (3.42),

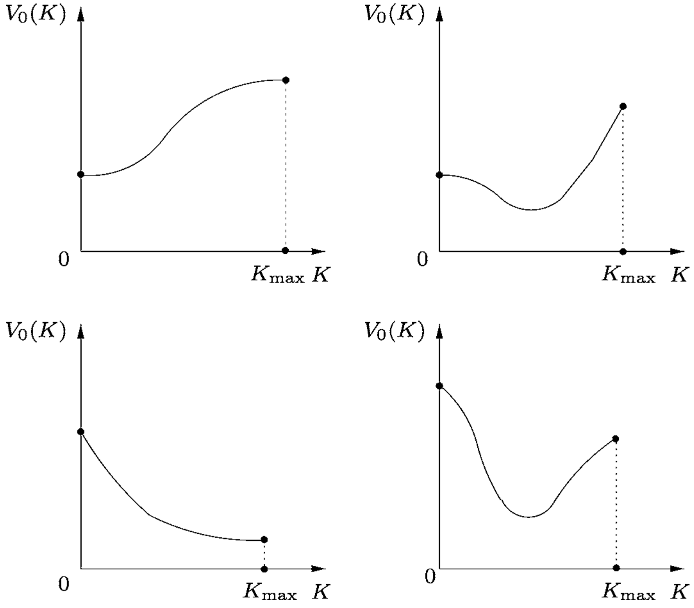

Fig. 3.3 Possible shapes of the value function.

there is a fundamental non-concavity in the problem of finding the optimal level of insurance reserves, which has the result that it will be optimal either to hold no reserves, or to hold such high reserves that the probability of insolvency is zero.

A limitation of the model is that it assumed that the interest rate was independent of the amount of capital raised, and that there were no other costs associated with raising capital for the insurance business. Clearly, the introduction of an increasing, *sufficiently convex* function relating capital costs to the amount of capital raised would create the required concavity in the objective function of the insurer and allow an interior solution for capital $K^* < K_{\max}$. A ruin probability could be thought of as a rule of thumb approximation to such an optimal interior solution. On the other hand, where the optimal corner solution is at $K = 0$, introducing such a capital cost function would not change this outcome, but rather is likely to widen the set of cases in which it would result. In that case the ruin probability would be that corresponding to no reserve capital.

The possibility of a solution in which the insurer would want to hold zero or low reserves could be thought of as providing a rationale for the commonly observed insolvency regulation. However, there is an obvious limitation of the model which undermines its persuasiveness in this regard. As we pointed out, the assumption of the exogeneity of the premium income implies that *the willingness to pay for insurance is independent of the insolvency risk of the insurer*. As Rees et al. (1999) show however, if we go to the opposite extreme, and assume that the insurance buyer is fully informed about the insolvency risk, then, under the assumptions of the above model, it will always pay the insurer to put up enough reserves to ensure a zero probability of insolvency.[10]

Why should regulation then be necessary? Clearly, the rationale for regulation must be based on an assumption about the *bounded rationality* of the insurance buyer, in the sense of an inability to calculate the implications for the insurer's solvency of its decision on capital reserves. This suggests that rather than detailed regulation of an insurer's capital reserves, it may be more effective and consistent with market competition and efficiency for regulatory agencies to concern themselves with collecting and disseminating to insurance buyers information on the default risks of insurance companies, as rating agencies do for potential investors. This will allow the market to bring about the efficient array of combinations of price and product quality, the latter defined by default probability.

[10] This is shown to hold both for monopoly and oligopoly insurance markets, where the latter are modeled as a Bertrand duopoly.

4

Adverse Selection

4.1 Introduction

Suppose that you have won an offer from the "endurance society" to join a professional group on a climb of Mount Everest. Statistically, 3.3% of all climbers leaving the basis camp have fatal accidents, while the ratio of people who reached the summit to those who died is 4:1. Still, considering the risk worth taking, you decide to join the tour. To take care of your family in case something untoward happens, you consider buying insurance. You find out that the life insurance company asks for a premium of \$100 per year for an insurance cover of \$100,000. Fortunately they do not ask whether you intend to climb Mount Everest. For your part, you would be prepared to go on the expedition even if you did have to pay the fair premium for that, so there is no way your actions are influenced by buying insurance. The insurance offer is very attractive. With a chance of death of 3.3% and a premium rate of 0.1% it makes sense even to overinsure (recalling the results from Section 2). There is no reason to stop with life insurance: accident insurance, disability insurance are next on the shopping list.

This is a situation the theory of adverse selection addresses. Individuals who know their own risk better than the insurance company use this knowledge when they buy insurance contracts. In the following we will investigate how insurance companies might react to the phenomenon of adverse selection, and what role — if any — the state can play to improve on the market's performance.

Adverse selection arises not only in life insurance markets, but pertains to all areas of the insurance industry: people know better whether they are reckless or careful drivers, whether they have healthy lifestyles or not, whether their property is well equipped against earthquakes or not. Although these examples contain some element of self-control, for example even a reckless driver might try to drive carefully, here we only consider that individuals innately differ *a priori*, i.e., before they acquire insurance. Influence over the risk probabilities will be discussed in the next section, where we turn to moral hazard.

As with all of the literature on asymmetric information, insurance markets are only one example of their applicability. Banks granting credits do not know the profitability of the projects, which however the borrower knows. Employers looking for new workers do not know the productivity of the potential employee, who might know it better. A government procuring defense equipment from the private sector has limited information on the costs of production which the companies themselves know in more detail. So although we concentrate on the insurance sector, many of the ideas and approaches can be generalized to other areas in economics.

We will start with a perfectly competitive market, which is the basis for most of the discussion of adverse selection in the insurance literature. We then consider further issues: categorical discrimination, endogenous information acquisition, long term contracts, and renegotiation.

4.2 Adverse Selection in Competitive Insurance Markets

4.2.1 The Basic Model

Adverse selection is defined as the situation where the insured has better information about her risk type than the insurer. We then say that the individual risk is her private information. For simplicity, we concentrate on two types only: High risks and low risks, with risk probability

π_h and $\pi_l < \pi_h$ of losing a sum L. Otherwise the individuals are identical. The insurer knows only the ratio of high risks to low risks in society. This is given by $\gamma_h/(1-\gamma_h)$ so that the average risk in society is: $\gamma_h\pi_h + (1-\gamma_h)\pi_l$. This is the probability that a randomly drawn insurance buyer will incur a loss.

Contracts specify premium rates only

The expected utility of an individual of type i if she buys an insurance contract (pC, C) is given by

$$Eu_i(pC, C) = (1-\pi_i)u(W - pC) + \pi_i u(W - L + (1-p)C). \quad (4.1)$$

As before, p is the premium rate, C is the amount of cover, W is the initial wealth of the individual.

As we know from the section on insurance demand under symmetric information, if the insured chooses her optimal insurance cover C^* then C^* will be larger (smaller) than L if π is larger (smaller) than p. In Figure 4.1 this is shown.

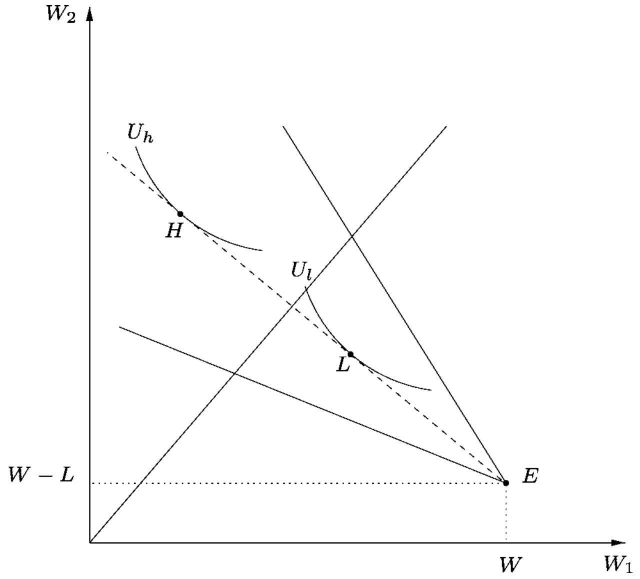

Fig. 4.1 Insurance cover for linear contracts.

On the two axes is wealth in the two states (no-accident, accident), point $E = (W, W - L)$ describes the initial endowment without insurance. The two solid lines are the zero profit lines for each type, which have the slope $-(1 - \pi_i)/\pi_i$ for $i = h, l$. If the contract line lies somewhere in between the two lines (the dotted line), then the high risks will overinsure (point H) while the low risks will optimally underinsure (point L). Both types still face an income risk after the insurance purchase. U_h and U_l denote the indifference curves of the two types.[1] Note that due to this over- and underinsurance, a fair pooled premium rate $p = \gamma_h \pi_h + (1 - \gamma_h)\pi_l$ will lead to a loss for the insurance companies. For profits made with L-types to offset losses made with H-types, they must buy the same contract on the pooled fair-odds line.

Thus there are two inefficiencies arising: First, individuals do not buy the efficient amount of insurance, i.e., they either over- or underinsure. Second, prices have to be larger than the average risk probability to avoid losses for the insurer. If there is a continuum of types, the insurance market may actually break down, as for any premium rate those lower risks who still buy a contract are not sufficient to subsidize the losses inflicted by the high risks. This argument works exactly as in Akerlof (1970) famous "lemons market" where the quality of a good is unknown. In such a case, governmental intervention might be useful, by, for example, obliging everyone to buy full insurance at the fair average premium.

Nonlinear insurance contracts

The analysis so far was restricted since it only concentrated on premium rates as the instrument available to the insurance companies. This may be a sensible approximation for life insurance, where individuals can buy several contracts from different companies, but in other areas of insurance this assumption does not necessarily hold. For example in the case of car insurance, concepts like deductible or partial insurance cover are meaningful, and, as we will see, are very useful in dealing with the problems of asymmetric information. Rothschild and Stiglitz (1976)

[1] In Section 2, it was shown that along the certainty line, the 45 degree line, the slope of the indifference curves is $-(1 - \pi_i)/\pi_i$. Therefore, the high (low) risk indifference curve is tangential to the dotted line to the left (right) of the certainty line.

and Wilson (1977) were the first to model insurers who offer contracts specifying both premium and amount of indemnity as a reaction to adverse selection.

In the following we allow insurance companies to set menus of price/ indemnity contracts. An implicit assumption of the analysis is that individuals buy only one contract with only one insurance company. This may be achieved through a clause in the contract or through legal requirement, which e.g., forbids overinsurance. This point is discussed later on in more detail. The presentation is based on the work by Rothschild and Stiglitz, but differs from theirs in two respects: First, the modern terminology of a game, rather than a specific notion of equilibrium, is used. Second, insurers are allowed to set a menu of contracts, while in Rothschild and Stiglitz' original paper insurers could only offer a single contract each.

Consider the following game: The players are as described above: There are two risk types in a society with N individuals where each individual is a h type with probability γ_h.[2] There are $M \geq 2$ risk neutral insurers in the market. The game proceeds as follows: At Stage 1, each insurer i offers a menu of contracts $\{\omega_i^k = (P_i^k, C_i^k), k = 1, 2, \ldots\}$ which specify premium and cover.[3] At Stage 2, each individual chooses one of the contracts which is optimal for her, if any. For the equilibrium later on we assume that if more than one insurer offers such a contract, individuals will split between the insurers equally. Then nature decides for each individual whether an accident occurs or not. Payments are made accordingly.

Before describing the equilibrium, one property has to be introduced:

Single Crossing Property: For every contract ω, the slope of the indifference curve of the low risks in a two-states-of-world diagram is steeper than the slope of the high risks.

[2] In the present model, the insured do not act strategically, they just choose the best contract available. Therefore, it suffices to assume that out of a population of N, $\gamma_h N$ types are high risks.

[3] It can easily be seen that contracts with random payments are never optimal, as they just confer some expected utility to the insured in case of an accident. This expected utility is the same for both types, so both types would be willing to pay the same amount of money in the accident state to avoid this uncertainty.

The single crossing property implies that indifference curves cross once only. This is usually assumed in all principal agent models, and it naturally holds in the present case, as $\frac{(1-\pi_h)u'(W_1)}{\pi_h u'(W_2)} < \frac{(1-\pi_l)u'(W_1)}{\pi_l u'(W_2)}$, where W_1 (W_2) is the income in the no-accident (accident) state. This assumption is crucial for the following analysis. However, if individuals differ in their wealth or some other characteristics in addition to their risks, then the single crossing property may be violated.[4]

For the Nash equilibrium we concentrate on a symmetric equilibrium in pure strategies for the insurers, where all customers of the same type choose the same contract. Mixed strategies, asymmetric equilibria and customers mixing between contracts will be discussed later.

Define by $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ the set of contracts offered in equilibrium. If all individuals of a certain type choose the same contract, then at most two active contracts are offered in equilibrium. Without loss of generality, we do not consider that insurers offer idle contracts as well.

Assuming that an equilibrium exists, five steps are required to derive the equilibrium contracts:

1. *Non-existence of a pooling equilibrium*: Suppose $\Omega = \{\omega_p\}$, i.e., all insurers offer the same contract in equilibrium. The corresponding final wealth of the insured is shown in Figure 4.2 (point P).

   For that to be a feasible outcome it has to lie below or on the pooling-zero-profit line (the dotted line), which is given by $W_2 = W_1 - L + I$ with $W_1 = W - P$ and $P = [\gamma_h \pi_h + (1 - \gamma_h)\pi_l]I$, as otherwise the insurers would make a loss. Assume first, that this outcome lies strictly below the pooling-zero-profit line. Then one insurer might offer another contract with a slightly lower premium which leads to outcome a (see Figure 4.2) with which it attracts all customers. If a is close to P then this is surely better than also offering ω_p and only taking $1/M$ of the customers.[5] Therefore

[4] See Smart (2000); Villeneuve (2003); Wambach (2000).

[5] This argument is the same as that used for the proof that in the standard Bertrand oligopoly two firms are enough to restore perfect competition.

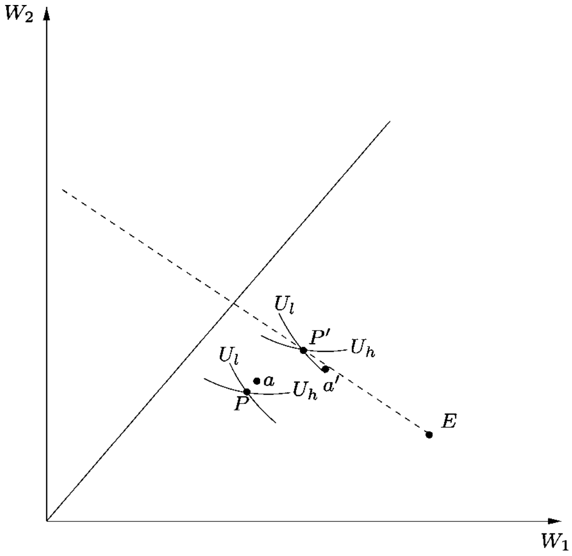

Fig. 4.2 Non-existence of a pooling equilibrium.

a point like P cannot be an equilibrium. Next assume that the outcome lies on the pooling-zero-profit line (P' in Figure 4.2). Due to the single crossing property, the indifference curves of the low and the high risk types cross each other at this point. From this it follows that there exist contracts which lead to outcomes like point a'. If a single insurer offers such a contract while all the others still offer P', this insurer will only attract the low risk types. As this contract would give approximately zero profit if taken by both risk types, it makes a profit if only the low risks buy it. Thus such a deviation — called a "cream-skimming" contract — is profitable. Hence, we note as a first result that no pooling contract can be an equilibrium outcome.

Therefore suppose that two contracts are offered in equilibrium: $\Omega = \{\omega_l, \omega_h\}$, where the first is taken by the low risks, while the latter is taken by the high risks.

2. *No contract makes a loss in equilibrium*: This is easy to see: if one of the two contracts makes a loss, then for every insurer

it is strictly better not to offer this contract, as long as the others still offer it. Note that this argument does not work in an asymmetric equilibrium: If only one insurer offers a loss-making contract, then by withdrawing this contract, individuals would choose a different contract at Stage 2, which might change the profitability of the other contracts this insurer offers. We return to this point later. As no contract can make a loss in a symmetric equilibrium, cross-subsidizing contracts, where for example the contract for the high risks makes a loss, while that for the low risks makes a profit, are ruled out. This is an important point to keep in mind, because as is shown later on, cross-subsidizing contracts might be (second best) efficient.

3. *No contract makes a profit in equilibrium*: Suppose that the contract for the low risks makes positive profits. There are two possible cases: Either the high risk types are indifferent between their contract and that for the low risks, or they are not. In the latter case, offering the low risks a slightly better deal will attract all low risks and make approximately the same profit per insured but more profit overall. This is again the standard Bertrand argument that price competition in a market without informational problems leads to a dissipation of profits. Therefore assume that the high risks are indifferent between the two contracts. Due to the single crossing property, there still exists a contract in the vicinity of ω_l such that only the low risks prefer this new contract. An insurer offering this new contract would then again attract all low risks instead of only $1/M$ if it were to stick with the old contract and make a profit with those. Note that with a similar argument, profit making contracts for the high risks can be excluded.
4. *The high risks obtain full insurance at the high risk fair premium*: As a result of steps 2 and 3 we know that the outcome for the high risks must lie on the high risks zero profit line. Suppose this point were at H' as drawn in Figure 4.3.

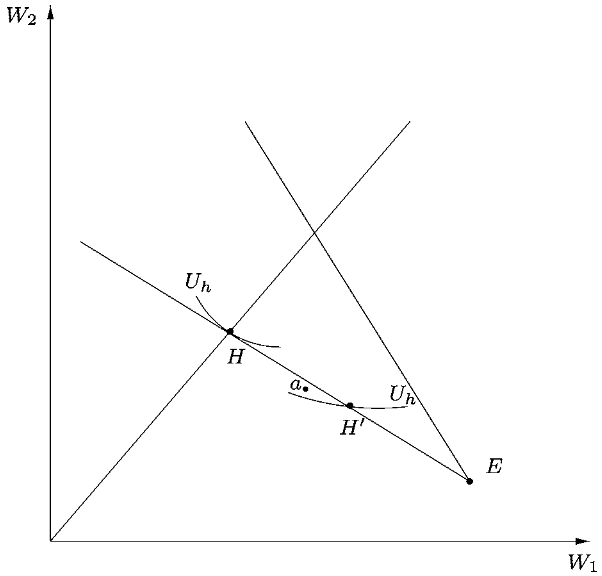

Fig. 4.3 High risks obtain full insurance.

As the indifference curve of the high risks cuts the zero profit line (remember that it is tangential only at the point of full insurance) offering a contract which leads to a would attract all high risks and lead to a profit for the company. Why are the low risks not a problem? We know that the contract for the low risks makes zero profit in equilibrium. Thus, even if the low risks now prefer a to their contract, so that they all switch to a once it is offered, that is good news for the insurer, as he will make a profit with the low risks at this contract as well. The same argument works at a point of overinsurance, because also there, the high risks indifference curves cut the zero-profit line. There is still scope for profitable deviations. The only possible outcome is at the point of full insurance where the indifference curve is tangential to the zero-profit line (point H).

5. *The low risks obtain partial insurance at their fair premium. The contract is such that the high risks are just indifferent between their contract and that for the low risks.*

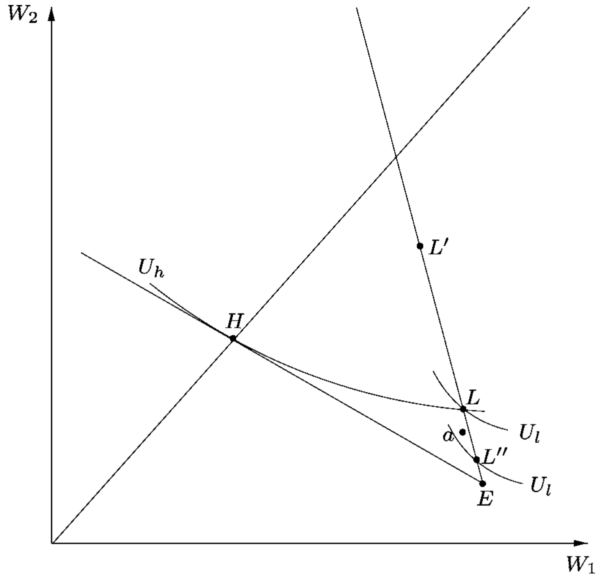

Fig. 4.4 Low risks obtain partial insurance.

This can be seen in Figure 4.4.

If the high risks are fully insured (outcome H), and the low risks receive a contract on their zero-profit line, then outcome L is the best one can do for them. Any contract with less partial insurance (those outcomes which lie above point L, e.g., L') are also preferred by the high risks. Therefore an insurer offering such a contract would attract the high risks with whom it would make a loss. For any contract on the zero-profit line below L (e.g., L''), there exists other contracts which all low risks and no high risk prefer and with which an insurer can make a profit (e.g., a).

This finishes the proof: contracts $\{\omega_l^{RS}, \omega_h^{RS}\}$ which lead to outcomes $\{L, H\}$ are the famous Rothschild–Stiglitz (RS) contracts: The high risks receive full insurance at their fair premium, while the low risks obtain partial insurance at their fair premium, and the high risks are indifferent between the two contracts.

This analysis is the basis for most of the work which has been done in the context of adverse selection in the insurance market. It displays three features which are testable with real world contracts: First, for a specific risk more than one contract is offered. In automobile insurance, for example, one can choose between different levels of deductible. This is usually also the case for health insurance contracts. Second, partial insurance is offered at a lower premium rate than the full insurance contract. The evidence for this is weaker, see e.g., Chiappori and Salanié (2000). Third, there is no cross-subsidization between contracts.

The structure of the RS contracts is very intuitive and gives a clear indication of possible market reactions to different risk types. Unfortunately, the discussion has one very serious limitation: The RS contracts do only constitute an equilibrium if the number of high risks is sufficiently large. As we next show, if γ_h is small, the RS contracts do not constitute a pure strategy Nash equilibrium.

The easiest way to see this is to show that in some cases a pooling contract might be such that it will be preferred by both risk types to their RS contracts, while still making a profit. Such an outcome is drawn in Figure 4.5 (point a).

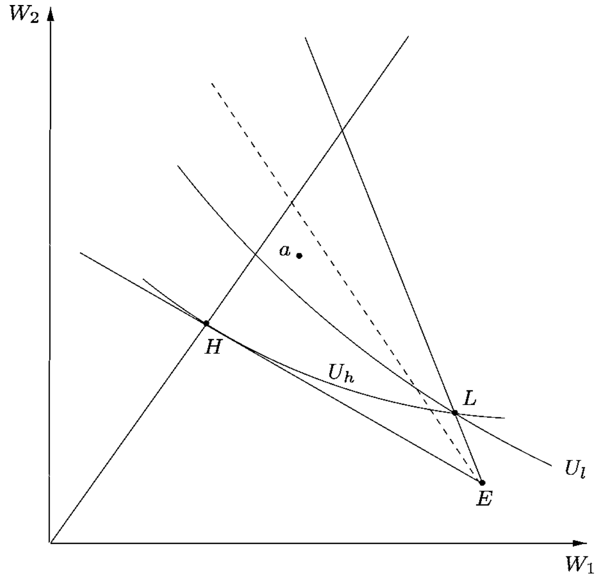

Fig. 4.5 Pooling contracts dominate the RS contracts.

If there are only a few high risks, the fair pooling line (the dotted line) will lie close to the low risk zero-profit line. Then the low risks prefer a contract on the pooling line to their RS-contract. An insurer offering a contract slightly below the pooling line (point a) can attract all customers and makes a strictly positive profit. But, as shown in step 1 above, a pooling contract cannot be an equilibrium. Therefore in that case we have to conclude that there does not exist an equilibrium under the assumptions we made on the strategies.

This non existence problem is not just a technicality but a serious problem, as for example in the health insurance market, one would expect that the ratio of high risks to low risks is relatively small.

Even if there is no profit making pooling contract which is better for the low risks than ω_l^{RS}, a pair of cross-subsidizing contracts (H', L') might be a profitable deviation, as shown in Figure 4.6.

If an insurer offers such a pair of contracts, he would make a loss with the contract for the high risks, but a profit with the contract for the low risks. Then if there are (again) sufficiently many low risks, both

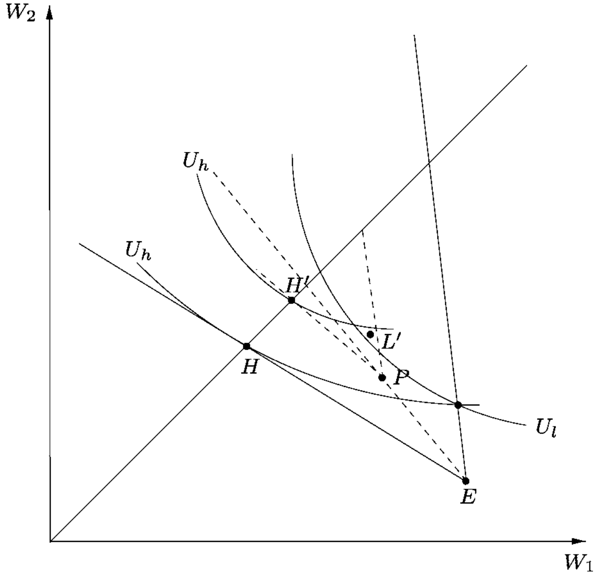

Fig. 4.6 Cross-subsidizing contracts dominate the RS contracts.

risk types can be made better off, while the insurer makes a positive profit overall. This can be seen as follows: In Figure 4.6, P lies on the pooling zero-profit line. Therefore any contract on the dotted line between P and H', which is the iso-profit line for the high risks, gives the insurer the same loss with a high risk type. Similarly, any contract on the iso-profit line for the low risks going through P would make the same profit with a low risk as contract P does. Now with outcome L' slightly below the iso-profit line for the low risks and below the high risk indifference curve which goes through H', the combination (H', L') leads to a profit overall. But, as shown in steps 2 and 3 above, profit making or loss making contracts cannot be equilibrium contracts. Thus also here we have to conclude that no equilibrium exists.

In the original work by Rothschild and Stiglitz, this cross-subsidization was not considered to be a problem, as only single-contract deviations were possible. But note that the pooling contract can also be seen as a degenerate pair of cross-subsidizing contracts. To summarize the results: If γ_h is sufficiently large, the RS contracts constitute a Nash equilibrium. If γ_h is smaller than some γ', a pair of cross-subsidizing contracts overturns the RS contracts, so a Nash-equilibrium in pure strategies does not exist. For even lower $\gamma_h < \gamma'' < \gamma'$, there exists a single pooling contract which will be preferred by both types to the RS contracts.

Before turning to the discussion on the equilibrium existence problem let us first check whether the assumption we made on the equilibrium strategies (insurers have symmetric strategies, all consumers of one type choose the same contract), restricted our analysis.

The only point in our proof of the RS equilibrium where the assumption that all insurers offer the same contracts in equilibrium played a role, was, when we showed that loss-making contracts cannot exist in equilibrium. Now, if only one insurer offers a loss-making contract which, say, is taken by the high risks, then by not offering such a contract, all high risks may choose another contract and could inflict a loss upon those insurers offering this contract. Therefore the insurer may perhaps rationally include the loss making contract in his menu of contracts. However, for that insurer not to make a loss altogether, he must offer a contract to the other type, in this case the low risks, with which he makes a profit. But here, the same “Bertrand-dynamic” as in step 3 above comes in: Due to the single crossing property, there always exists another contract for the low risks with slightly better terms which the high risks do not prefer. Another insurer offering this contract will attract all low risks and make a profit with those. Therefore such a constellation breaks down. This point is worth stressing: Usually the argument brought forward against cross-subsidizing contracts in a competitive market is that insurers will withdraw the loss making contract (as we did in step 2 in the previous section). In the case of asymmetric contract offers, however, cross-subsidization does not work because the other insurers do not allow the insurer with the loss-making contract to recover its losses. It is the profit making contract which is not sustainable in a competitive environment.

Now we relax the assumption that all customers of one type choose the same contract. Suppose one type of customer mixes between contracts. This implies that all the contracts have to lie on the same indifference curve for this type. Then in general the insurers make different profits per insured with each of these contracts. But, with a similar proof as in steps 2 and 3, neither loss making nor profit making contracts are possible in equilibrium. Therefore at most two contracts on the zero-profit line, one with underinsurance, the other with overinsurance are conceivable. However, if the other type does not even weakly prefer any of these contracts, then offering full insurance is strictly better (as was shown in step 4). If, on the other hand, the other type is indifferent between her contract and one of the two contracts, such that moving to full insurance is not feasible, then either the overinsurance contract (in case of the low risks) or the underinsurance contract (in case of the high risks) must be strictly preferred by the other type. Thus customers of one type choosing different contracts cannot be an equilibrium.

4.2.2 The “Equilibrium-Non-Existence” Debate

To be precise, although the literature talks about the equilibrium-non-existence problem, so far we have only shown that an equilibrium where insurers use pure strategies does not exist.

So what about mixed strategies? This would be the formal game theoretic solution to the non-existence problem. In the case of two insurers, who offer two contracts each, Dasgupta and Maskin (1986) (see also Rosenthal and Weiss, 1984) have indeed shown that an equilibrium in mixed strategies exists if the RS contracts do not constitute an equilibrium. In this context, mixed strategies mean that each insurer offers different sets of two contracts, each with some probability. The exact equilibrium is not known, however some details of the equilibrium can be obtained: First, insurers make zero expected profit; second, with any contract pair offered, the high risks obtain full insurance at a fair or better premium and the low risks obtain partial insurance at an unfair premium.

However, the economic interpretation of an equilibrium in mixed strategies is unclear: Are insurers supposed to be randomizing over contracts each year or perhaps each day? More worrisome is the fact that once an insurer has seen which particular set of contracts his competitors have offered, he will presumably prefer to respond by modifying his own contracts on offer. Thus it plays a crucial role how long contracts stay on the market and how often they can be modified. This is different in the RS equilibrium (in case it exists). Once contracts are offered, no insurer has an incentive to withdraw contracts or offer new ones, as the RS contracts are mutual best responses. In many contexts, mixed strategies are a sensible concept to use. As a description of the strategic interaction in an insurance market, however, equilibria in mixed strategies are more an indication of the limitations of our model. Perhaps it is too simplistic to assume that insurers only offer contracts out of which customers choose the best one available. Presumably there is something going on in an insurance market which we have not captured so far. We have to look for more sophisticated games.

In the late 1970's this was not done by extending the game structure, but rather by assuming different equilibrium concepts. The Rothschild and Stiglitz equilibrium definition is that *there is no contract outside the equilibrium set that, if offered, makes a profit.* (We already extended this to a menu of contracts.) In Wilson's equilibrium concept (Wilson, 1977), *every additional contract should stay profitable even if those contracts which make a loss after the introduction of the new contract are* *withdrawn.* It is easy to see that in this case a pooling contract might survive in equilibrium: Consider again step 1 above: Pooling was unstable because someone could offer a contract only to the low risks, i.e., to "skim off" the good risks. However, in the Wilson concept, if someone tries to attract the low risks only, all others will withdraw their loss-making pooling contract, because that contract would be bought by high risks only. Therefore the high risks also choose this newly offered contract, which makes it much less attractive to offer it in the first place. Wilson has shown that if the number of high risks is sufficiently large, his equilibrium coincides with the RS contracts. If the proportion of high risks is low, however then the equilibrium contract will be a pooling contract on the zero-profit line. To be precise, this contract is the best zero-profit pooling contract from the point of view of the low risks, i.e., it is a partial insurance contract where the low risks indifference curve is tangential to that line.

Extending Wilson, Miyazaki (1977) and Spence (1978) allow in addition that insurers offer more than one contract. Therefore cross-subsidization between contracts becomes possible. This leads to the so-called WMS equilibrium, which is the solution to the following maximization problem:

$$\begin{array}{ll} \max\limits_{P_l,I_l,P_h,I_h} & (1-\pi_l)u(w-P_l)+\pi_l u(w-P_l-L+I_l) \\ & \text{s.t.} \\ \text{IC} & (1-\pi_h)u(w-P_h)+\pi_h u(w-P_h-L+I_h) \\ & \quad \geq (1-\pi_h)u(w-P_l)+\pi_h u(w-P_l-L+I_l) \\ \text{PC} & \gamma_h(P_h-\pi_h L)+(1-\gamma_h)(P_l-\pi_l I_l)\geq 0 \end{array} \tag{4.2}$$

The utility of the low risk type is maximized under two constraints: The first constraint, the so-called incentive constraint, shows that implementable contracts must be incentive feasible, i.e., the high risk must weakly prefer their intended contract over the alternative. Since only high risks present an adverse selection problem (they have an incentive to mimic the low risks), it is their incentive constraint which needs to be considered. The second constraint, called the participation constraint, ensures that the insurers make nonnegative profit overall. The WMS equilibrium concept is quite often used in the insurance

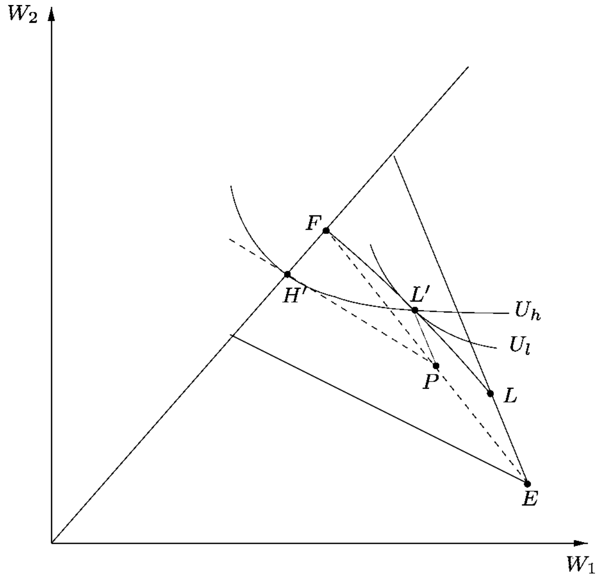

Fig. 4.7 WMS contracts.

literature. So it is worthwhile to spend some time to construct the WMS contracts. The construction is shown in Figure 4.7.

The line $F - L$ denotes all feasible outcomes for the low risks. That is, these are contracts which together with a specific contract for the high risks can be offered to the low risks such that the high risks would buy the contract designed for them and the insurer makes nonnegative profits. This line is constructed in the following way: First, select a point on the certainty line and interpret this as a full insurance contract for the high risks (H'). Draw from this point the iso-profit line of the high risks, the dotted line, until it cuts the pooling-zero profit line (point P). Recall that for the insurer it does not matter whether a high risk were to buy the contract at H' or at P. Beginning from P draw the iso-profit line of the low risks, also a dotted line. Now, where the indifference curve of the high risks which goes through (H') cuts this line at (L'), we have a feasible contract for the low risks. This contract pair is incentive compatible and makes zero profit for the insurer. Now by shifting (H') along the certainty line one can construct all possible contracts for the low risks, which leads to the line $F - L$. One endpoint of this curve, L is the RS contract of the low risks, while the other endpoint, F, must be the full insurance contract which lies on the pooling line. The WMS outcome is now given by the best contract possible for the low risks along this line, i.e., the point where the indifference curve of the low risks is just tangential to the line $F - L$. This point is shown as L'. Again, the high risks obtain full insurance, while the low risks obtain partial insurance.

Note that all contracts above and including L', together with the corresponding contract for the high risks, denote the Pareto frontier of the adverse selection problem. While F is the best contract (pair) from the point of view of the high risks, the WMS contracts are best for the low risks. We have drawn the diagram such that the RS contracts are not equilibrium contracts, which implies that they are not efficient. If γ_h however is sufficiently large, then point F will move downwards, L' will shift to the right and the WMS equilibrium corresponds to the Rothschild–Stiglitz outcome.

A different equilibrium concept was introduced by Riley (1979). While in the Wilson concept insurers anticipate that other insurers will withdraw contracts as a result of their deviation, here the deviating insurers anticipate that at least one other insurer will react by offering an additional contract. In this *reactive equilibrium*, if the RS contracts are offered, insurers shy away from offering deviating contracts as they anticipate that it will make a loss once other insurers have reacted to the deviation with a new offer. This is straightforward to see. If an insurer were to offer a deviating pooling contract, then due to the single crossing property another insurer can offer a contract which just attracts the low risks. Thus the deviating insurer is stuck with the high risks and makes a loss. Thus the Rothschild–Stiglitz outcome is an equilibrium even if there are only a few high risks.

The WMS equilibrium is attractive from an economic point of view, as the contracts are second best efficient, i.e., it is not possible given the informational asymmetry to offer an alternative pair of incentive compatible contracts that jointly break even and yield strictly higher expected utility for at least one of the risk types and no lower utility for either type (Crocker and Snow, 1985). That is what a competitive market is expected to lead to: Pareto efficient outcomes. It is this feature of the WMS equilibrium which makes it quite popular in the insurance literature. On the other hand, the Riley concept rationalizes the Rothschild–Stiglitz outcome even if it does not constitute a Nash equilibrium. In that case, the outcome would be inefficient. In both cases, however, equilibria are justified by introducing new and to some degree arbitrary new conditions. In the spirit of game theory, one would feel more confident if one could replicate these conditions in a fully specified game: Why should an insurer withdraw its contract in response to other insurers' entry? How long does it take to withdraw? Why do only deviators fear responses, why do not those insurers offering the equilibrium contracts fear deviators? What are other possible ways to interact strategically on the insurance market?

Both in the concepts of WMS and Riley, some form of dynamics, namely the possible reaction of insurers after the contracts have been offered, are considered. In the remainder of this section we will discuss selected models where this dynamic aspect is explicitly modeled as part of more elaborate games.

Hellwig (1987), based on the work by Grossman (1979), introduced a third stage in the model described above. Again, insurers offer a contract at Stage 1, then customers choose at Stage 2, but now insurers can withdraw their contract at Stage 3. This comes close to the Wilson concept, as insurers now have the ability to withdraw some contracts, depending on what the other insurers offered, and what the customers choose. The Nash equilibrium of this game is more elaborate to determine, as now customers by choosing contracts at Stage 2 reveal information which might be used in Stage 3. This is therefore a combination of a screening (by the insurers) and signaling (by the customers) model. Those readers who are familiar with signaling models know that usually many equilibria are possible, depending on the out-of-equilibrium beliefs. In this case, however, under specific belief refinements, only the Wilson pooling contract is robust, whenever the RS equilibrium does not exist. This approach is very useful as it explicitly models the equilibrium concept of Wilson. It shows that an equilibrium in pure strategies always exists. However, whether the possibility of withdrawing an accepted contract is really a descriptive characteristic of an insurance market remains an open question.[6]

It is not correct to conclude that if insurers were to offer a menu of contracts instead of a single contract at Stage 1 in Hellwig's model, the WMS contract pair would be the outcome. The reason is that as insurers can withdraw contracts at Stage 3, any insurer offering a WMS contract pair would, if high risks were indeed to choose their full insurance contract, withdraw this contract at Stage 3 as it makes a loss. This however will be anticipated by the high risks, thus they do not choose that contract in the first place. However, modifying the game structure somewhat will lead to the WMS outcome. As before, in Stage 1 insurers offer a menu of contracts. Then, in Stage 2 after observing all the offers on the market, insurers can withdraw their offers or part of their offer. At Stage 3, customers choose out of the remaining contracts on offer. Now offering the WMS contract pair can be stable — if anyone tries to cream skim the low risks, insurers can withdraw their contracts at Stage 2. If some or all insurers offer the WMS contract pair at Stage 2, then the insureds can choose their respective contract, as insurers are not allowed not to serve the contract anymore.[7]

In Asheim and Nilssen (1996) there are again three stages. In stage one, insurers make offers, in Stage 2 the insureds choose a contract. Now, in Stage 3 instead of withdrawing their contract(s), insurers can offer new contracts to their existing customers, who then can choose to either stick with their contract or take one of the new ones on offer. Although this sounds somewhat like the Riley concept, it differs in so far as insurers can only offer contracts to their own customers, and not to the whole market. The motivation for this model is the idea that an insurer can renegotiate the contracts with its own customers. At

[6] An alternative model would be to let customers send some form of signal first, after which the insurers make their contract offers, after which the customers choose the preferred contract (see Cho and Kreps, 1987). In that case, under appropriate belief refinements, the separating contracts of the Rothschild–Stiglitz type are the equilibrium contracts. In the insurance market it is however not clear, what kind of signal the customers might give at Stage 1. In other models, like for example the credit market, this signal could be a collateral.

[7] We have not seen this argument modeled. We were told (Hellwig, private communication), that in the late 1980's this reasoning appeared in some working papers, which however have not been published.

Stage 3, any insurer does not compete with the others for its customers anymore. Therefore it can offer cross-subsidizing contracts to its own customers as long as these contracts give them larger utility than the contract which made them sign in the first place. There is no danger of "cream-skimming" by the other insurers. It can be shown that, overall, the WMS outcome as the final contracts is the unique equilibrium of this game. One might criticise, however, that once renegotiation is explicitly introduced, it is not clear why the new contracts have to be offered to all customers. If customers have signed different contracts at Stage 2, then an insurer might offer different contracts depending on the contract the insured has already signed.

So far it is always assumed that some exclusivity condition can be enforced: Individuals only buy one insurance contract. This could be enforced by a clause in the contract stating that the insurance company will not pay if the insured receives a payment from another insurer. But it is not obvious that the information on the number of contracts bought is readily available to the insurance company: Who tells them whether there exists an additional policy or not? As suggested by Jaynes (1978) and later developed in game form by Hellwig (1988), the incentive insurance firms have to share or conceal information about their customers might be another strategic variable. This is modeled in a four stage game: In the first stage, insurers offer contracts and decide whether or not an exclusivity requirement is attached to this contract. In the second stage, consumers choose a combination of contracts. Then insurers at the third stage decide what information if any they want to divulge to which insurer and at the fourth stage, they choose, depending on the information they received, whether or not to enforce the exclusivity condition. In equilibrium, customers buy two types of contract: The Wilson pooling contract is sold by insurers who exchange information with each other. This policy is bought by everyone. The high risks amend this contract with a partial insurance contract at the high risk fair premium, such that they obtain full insurance. This latter additional contract is bought from insurers who do not reveal information about their customers. One interesting aspect of this result is that insurers do not screen the market. The contract sold to the low risks is a pooling contract which all the high risks buy as well. Then the high risks obtain additional coverage from different insurers. An example of this could be seen in many health insurance markets: Everyone buys the same standard insurance package, while only some acquire additional coverage through an "add-on insurance."

Instead of modifying the behavior of insurers, Inderst and Wambach (2001) modify the attributes of insurance companies. They assume that insurers face capacity constraints, which might result from limited capital available to the insurer. In that case, by offering deviating contracts an insurer cannot be sure that he obtains the mix of risk types he desires, as not everyone will turn up at this insurer. Under some assumptions on the severity of the capacity constraint and on the costs the customers face when they are rationed, it is shown that indeed only the high risks will turn up if someone offers a deviating pooling contract or a pair of cross-subsidizing contracts. The reason is that due to the single crossing property the high risks gain much more from a deviating contract, so they are more willing to endure the rationing which will occur at the deviating insurer. Therefore no insurer has an incentive to deviate, which implies that the RS contracts are always an equilibrium outcome of the game.

Sometimes it is argued that, due to the existence problem, there is some instability in an insurance market with adverse selection. Insurers might offer the RS contracts, then someone will find a better pooling offer, this one will then be overturned by someone who cream-skims the low risks, etc. An attempt to model these dynamics of an insurance market explicitly is given in Ania et al. (2002), where methods from evolutionary game theory are used. In their work the assumptions that insurers have perfect knowledge of the utility functions of the customers, their risk types, the number of different risk types, etc. are relaxed. Instead it is assumed that insurers offer contract menus and imitate successful behavior, i.e., in every period they observe the most profitable contracts on the market and copy those. In addition, once in a while they experiment with their own contracts (which is called "mutation" in the literature). Experimentation and mutation both stand for different explanations of this dynamical feature: Either insurers are supposed to experiment, trying to find ways to increase their profits or market shares by offering new contracts, or they mutate, which means that they make mistakes in pricing their contracts, thus offering new ones by accident. Two results are shown: First, if no profit making pooling contract is better for the low risks than the RS contract, then the RS contracts are the long run outcome of this evolutionary game. If a pooling contract is preferred, then the RS contracts are still the long run outcome if experimentation takes place only locally, i.e., insurers only add contracts close to the existing ones. The first result is interesting as it shows that even without detailed information insurers can learn to offer screening contracts. Furthermore, in an evolutionary context, possible deviations via cross-subsidizing contracts are not a problem. Insurers will quickly copy the profit making part of the cross-subsidizing pair of contracts while the first insurer to offer this set of policies withdraws the loss making one. Then the system works itself back to the RS outcome. The second result points to the destabilizing force of pooling contracts which the RS contracts do not share: Pooling contracts can be destabilized by small changes in the contract structure while to destabilize the RS contracts a substantial change in the contract conditions is required. This evolutionary model is explicitly dynamic as it discusses the very long run outcome. It is limited as neither strategic contract settings from side of the insurers, nor strategic choice of contracts from side of the customers is considered.

To summarize this section: The non-existence of an equilibrium in pure strategies of the Rothschild–Stiglitz model is still, after nearly three decades, a problem in the insurance literature. To remedy it the simple two-stage game has to be extended. So far, candidates for an equilibrium, if in the original model no equilibrium, exists are still the RS contracts, the Wilson pooling contract, the WMS cross-subsidizing contracts, and even a combination of Wilson and full insurance contracts.

This equilibrium debate is relevant, as adverse selection has quite often been brought forward as a reason for governmental intervention in insurance markets, in particular the health, pension and unemployment insurance markets. If the outcome were inefficient or if the market would not find an equilibrium, then it might make sense for the government to step in. However, note first that if the RS contracts do constitute an equilibrium, then they are efficient, so governmental intervention is not needed, except possibly for distributional reasons.[8] If they are not equilibrium contracts, then they are inefficient. However, as they are not sold in equilibrium, they also do not matter. It is true that in that case no equilibrium exists in the standard RS model. But rather than using this as a reason to call for governmental intervention one might interpret this as a sign that the model is not fully capturing what is going on in the insurance market when there are just a few high risks. And if one extends the model along the lines of WMS or Hellwig, then an efficient outcome, the WMS contracts, are obtained. So again the government would not improve the situation. The best case to make for regulation of insurance market is either to go back to the model with linear premium rates only as discussed in the beginning of the previous section, or to follow the reasoning along the lines of the models by Riley, Inderst and Wambach or Ania et al., where the RS contracts are always equilibrium contracts even if they are inefficient. In that case a governmental intervention can be welfare improving. To see this, recall that if there is a large number of low risks, then there exists a pooling contract on the fair pooling line which would be preferred by both the low and the high risks. Now, if the RS contracts are still equilibrium contracts in this case, (as they are in the models of Riley, etc.), then a regulation requiring insurers to offer this particular pooling insurance contract would lead to a welfare improvement, as both risk types are made better off and the insurers still make zero profits.

4.3 Categorical Discrimination

A common feature in the insurance markets is price discrimination between groups of customers: Young drivers pay a higher premium for automobile insurance. The premium depends on whether the car owner has a garage or not, what type of car she is driving, where she lives, etc. Health insurance premia are higher for older persons. They are also higher for women than men. On the other hand, one does

[8] An equilibrium in which high risk types are having to pay high premiums while low risk types receive little cover, albeit at a fair premium, may be considered socially inferior to one in which there is compulsory pooling, even if it involves redistribution between types.

not observe that the price of a good, say a TV set, depends on who buys it. In some cases, as for example railroad tickets, the price might depend on the age of the customer, but in general price discrimination is regarded as a sign of market power. In a competitive market one would expect that a price equal to marginal costs is charged. So what is different in the insurance sector, if any? Does discrimination, if it is not legally forbidden, develop endogenously in a market? Is there a need for governmental intervention? In this section we will tackle these questions.

First, we should distinguish between *discrimination* and *differentiation* in prices. When the marginal costs of supplying different groups of consumers differ, then in competitive markets the prices they face would also differ, and it is indeed Pareto efficient that they do so. This is price differentiation. Price discrimination, on the other hand, is a term usually reserved for situations in which different groups of buyers are charged different prices even when the marginal costs of supplying them are the same, and requires non-competitive market power, as well as the ability to prevent arbitrage, for it to be feasible. Now on an insurance market, if two groups of individuals have different loss probabilities, then we would expect premium differentiation on a competitive market, since the expected values of loss, the equivalent of marginal cost on an insurance market, are different. Unfortunately, in the insurance literature, this has come to be called discrimination, giving negative overtones to something that is in fact perfectly consistent with economic efficiency. In the following discussion of categorical discrimination, we will stick with the common term discrimination, but at certain points clarify whether the observed behavior indeed corresponds to premium discrimination or premium differentiation.

For simplicity we consider only two groups in society. For concreteness, call them males (with a proportion of λ_m) and females ($\lambda_f = 1 - \lambda_m$). As before, each individual faces a risk of losing L, and we assume the loss probabilities are the same for all members of a given group. If both groups also have the same risk-probability, then in a competitive market without transaction costs the premium rate for both would be equal to this common risk probability. So discrimination only makes sense if males and females differ in some relevant characteristic.

Assume that both groups have different risks, with the male loss probability $\pi_m > \pi_f$, the female loss probability. Again, in a competitive market, if discrimination is possible, both parties would receive full insurance at their fair premium, that is the males have to pay more than the females. However, in this case we should use the term "differentiation" rather than "discrimination," because this is effectively a case in which different groups of consumers face different marginal costs, and so the prices they pay should also (on efficiency grounds) differ.

If price discrimination is forbidden, then the result depends on the equilibrium concept used. Here we concentrate only on the Rothschild–Stiglitz (RS) outcome and the Wilson–Miyazaki–Spence (WMS) equilibrium, as these are the concepts which are predominantly used in the literature. In both cases insurers will try to attract the females (the good risks) not by offering a "women-only" contract (which is forbidden by assumption) but by offering contracts with more or less partial insurance. In an RS outcome, as we have seen in the previous section (Figure 4.4), the high risks (the males) would receive full insurance at their fair premium, while the low risks (the females) would obtain partial insurance at their fair premium. In comparison to that, females are better off if price discrimination were allowed: they would receive full insurance at their fair premium. The males are indifferent between discrimination and the Rothschild–Stiglitz equilibrium, because in both cases they obtain the same contract. Discrimination would be Pareto improving, because if discrimination were not possible, firms will screen the market by other means. In this case by offering partial insurance contracts. As in general this leads to inefficiencies, it might indeed be better to allow discrimination in the first place. This effect is particularly strong in an RS outcome.

If one considers a WMS equilibrium instead, this result does not hold. In that case the males are subsidized by the females if discrimination (or better: differentiation) is forbidden. In contrast, if discrimination were allowed, women would fare better while men would be worse off. This is probably the standard result one would expect from a switch in regime from no to full discrimination: High risks are worse off while low risks are better off. But note that due to the inefficiencies which arise in an insurance market under adverse selection, a social planner would always prefer to discriminate: She could still offer the males the same policy even after differentiation while the females can be made strictly better off. This is shown in Figure 4.8. Starting from a WMS outcome (H, L), if discrimination is allowed, under perfect competition the new outcomes are (H', L'). However, a social planner could for example offer the policies (H, L'') which would be a Pareto improvement compared to (H, L). This latter feature distinguishes discrimination in the insurance market from price discrimination in other markets: It might help to overcome some of the existing inefficiencies.

So far we have considered the case where the gender of the two types is a perfect signal of the riskiness. But in general one would expect that any category like gender or age does not reveal the risk type completely. There are probably still some cautious male and some risky female drivers around. Denote by γ_m the proportion of high risks in the male population, and by $\gamma_f < \gamma_m$ the proportion of high risks in the female

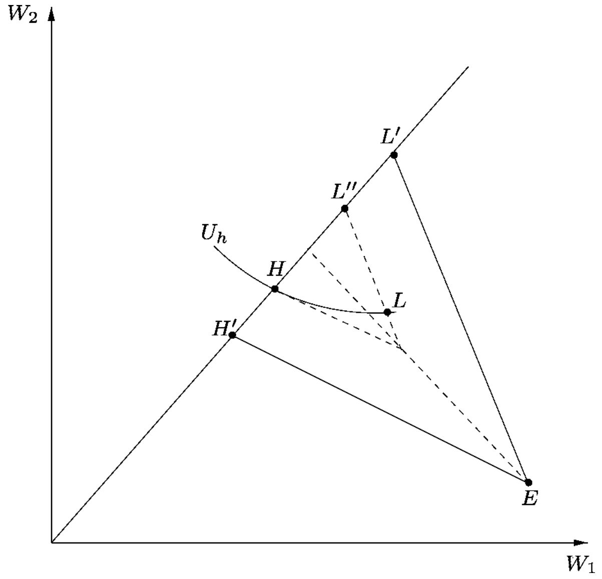

Fig. 4.8 Discrimination can be welfare improving.

population. That is, the average male risk, which is $\gamma_m \pi_h + (1 - \gamma_m)\pi_l$ is larger than the average female risk. Given that it is perfectly costless to distinguish between males and females, and that they have different average loss probabilities, even though within the group loss probabilities also differ, the question arises of whether premiums should be differentiated on the basis of the group to which an insured belongs. This is the issue of whether categorical discrimination is welfare improving or not, and of whether it should be prohibited by regulators. Clearly, if women are offered a lower premium than men, then low risk men are paying a price above their true marginal cost while high risk women are paying a price below theirs.

To discuss the outcome in a competitive market, consider first the RS equilibrium, as shown in Figure 4.4. A crucial feature of the RS equilibrium is that the policies for the two risk types are independent of their actual proportion in society. High risks obtain full insurance at their fair premium while low risks obtain the best possible contract at fair terms, which the high risks do not prefer. In that case, discriminating with respect to gender would not change any contract offered.

This result does not hold in a WMS equilibrium. The cross-subsidizing pair of WMS contracts does depend on the ratio of the two types. To see this, take again a look at Figure 4.7 from the previous section.

Recall that the line $F - L$ denotes all feasible outcomes for the low risks, with (H', L') as the WMS contracts. Now, by moving from non discrimination to discrimination, the curve $F - L$ will shift for the two groups. As male drivers are on average more risky, contract F will move down along the certainty line. Furthermore, L' moves closer to contract L, whose position is independent of γ_h. Indeed, if the average risk of males is sufficiently high, then the RS contracts become Pareto efficient, which implies $L' = L$. On the other hand, if in the female group just a few high risks are around, then their F will move upwards along the certainty line and their L' moves closer to F. We can thus make the following observations:

1. If it is possible to discriminate costlessly, then the insurance companies will do so. There will be winners and losers in

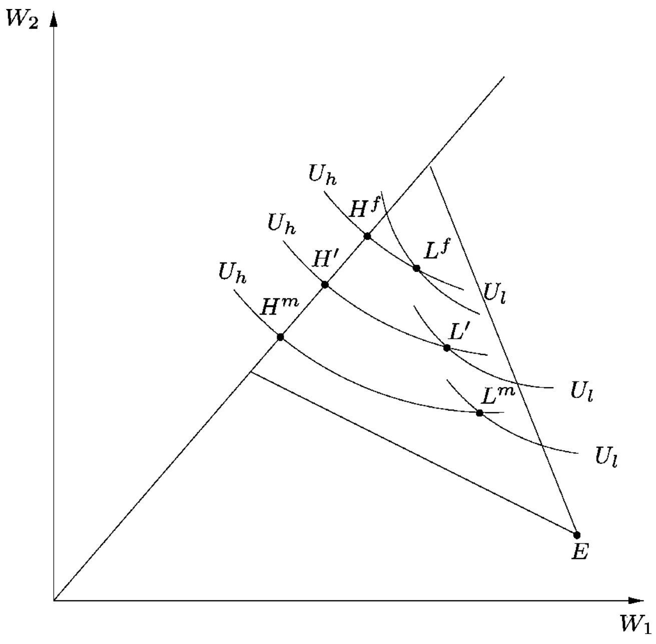

Fig. 4.9 Unregulated discrimination is not Pareto improving.

the market. This can be seen in Figure 4.9. While initially at outcome pair (H', L'), the females receive (H^f, L^f) while the males obtain (H^m, L^m). Note that in the case which we discussed in the beginning, if gender is a perfect indicator of the risk, $L^f = H^f$ at the low risk fair premium, while H^m is at the fair premium of the high risks.

2. A social planner selling insurance policies could not do worse, and sometimes better if she discriminates. The reason is that by discriminating, the Pareto frontier shifts outwards. As we have seen above, if the signal is fully revealing, then this holds naturally (Figure 4.8). But also if the signal is only partially revealing, the efficiency gains could be used to make everyone better off.

Crocker and Snow (1986) have shown that with an appropriate tax system, where the different contracts are taxed differently, the state could implement their desired outcome.

In that case everyone could be made better off by categorical discrimination. Without additional intervention by the government, however, there is usually someone who loses and someone who wins.

Endogenous criteria

An interesting extension of the analysis here is the case of discriminating with respect to endogenous quantities, like the type of car or whether the car is parked in a garage or not. In that case, discrimination might influence the behavior of the insured. Consider the following situation: Half of the population has a garage. The other half does not have a garage, but they would be willing to get one if the rent of garages were reduced by $20 per year. Owning a garage or not does not change the riskiness of a driver, but it happens to be the case that non garage owners are more dangerous drivers on average. So far, due to the legislation, everyone was fully insured at the same premium. Now discrimination is allowed. The insurance companies find out that garage owners have on average fewer accidents than non-owners, and they decide to offer garage owners a discount of $25 per year, while the non-owners have to pay $25 more. What will happen? First, all non-owners will rent a garage, as this reduces their car insurance premium by $50, while only $20 was required to rent a garage. However, if they do so the insurance companies make losses with the garage-owners policy, as everyone now has this policy so the average risk in that group of policyholders is as before. Next year, the insurer will thus raise the premium for the garage owners policy again by $25 to the old level while the contract of non owners stays as it is. So after all, everyone pays the same as before, but the non owners are made worse off. This is certainly a simplified example, but the general point should be clear: If one can decide to which group one belongs, different policies sold to different groups influence the composition of groups.

To conclude this section: In contrast to most other markets, where the costs of production do not depend on the customer who buys the good, discrimination with respect to some characteristics may be welfare improving in the insurance sector. However, if discrimination is not accompanied by lump sum redistribution between contract types, there are usually some types who win and others who lose.

4.4 Endogenous Information Acquisition

So far it was assumed that individuals know which risk type they are, i.e., the information structure was exogenous to the model. In this section, we investigate the situation where the individuals have no information *ex ante*, but they have a possibility to learn about their risk type. This topic has recently attracted a lot of attention, as genetic tests are a case at hand: By undergoing such a test, the individual learns more about her risk type. And now the crucial question, which is still being debated, is: Should the insurer have this information as well or not? Let us go through the model to analyze this case in more detail.

We now need three types of individuals, called H, U, and L. H and L consist of those individuals who already know that they are high and low risk, respectively, while U consists of those who do not know whether they are high or low risk.

Let γ_h be the probability for an U type to be a high risk. So the loss probability for an untested person is given by: $\pi_u = \gamma_h \pi_h + (1 - \gamma_h)\pi_l$.

Symmetric information
As a benchmark consider first the case where the insurer can observe whether the individual has taken a test or not. And, if she has taken a test, what result this test gave. As there is full information, it is clear that the equilibrium contracts will provide full cover at the fair premium for the respective types. The contracts, denoted by H^*, U^*, L^*, are shown in Figure 4.10.

Consider in this case the social value of a genetic test that identifies with complete accuracy whether someone has a loss probability of π_h or π_L. It is a long-established result[9] that, from a positive point of view, no insurance buyer will want to take the test, and from a normative

[9] For general analysis of the value of information in market economies under uncertainty see for example Arrow (1970), Hirshleifer (1971), Harris and Townsend (1981), and Milgrom and Stokey (1982).

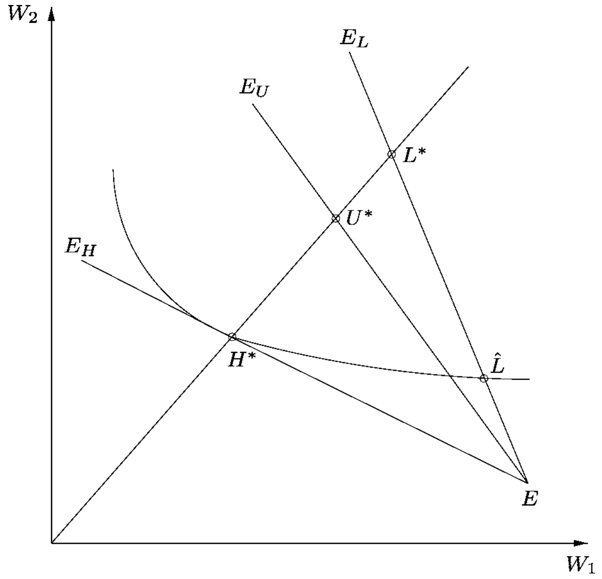

Fig. 4.10 Contracts under symmetric information and under asymmetric information if informational status is common knowledge.

point of view, the *ex ante* social value of the test is negative. This can be seen as follows:

Before testing, someone in U buys full cover at the fair premium $\pi_u L$ while after testing, since insurers can observe everyone's type, the premium will be either $\pi_L L$ or $\pi_h L$, respectively, depending on the outcome of the test. Thus she has a utility of $u(W - \pi_u L)$ if not tested, and $u(W - \pi_h L)$ with probability γ_h and $u(W - \pi_l L)$ with probability $(1 - \gamma_h)$, if tested. Then strict concavity of utility implies

$$u(W - \pi_u L) > \gamma_h u(W - \pi_h L) + (1 - \gamma_h)u(W - \pi_l L) \tag{4.3}$$

since

$$W - \pi_u L = \gamma_h(W - \pi_h L) + (1 - \gamma_h)(W - \pi_l L). \tag{4.4}$$

Intuitively, each individual in U prefers the certainty of premium $\pi_u L$ to the gamble on premia, with the same expected value, that the test represents. This phenomenon is known as *premium risk.* Then, since individuals in H and L gain nothing from taking the test, while those in U lose in expected utility, nobody takes the test.

Next consider the case where the insurer does not know the test result, but does know whether a test was undertaken or not.

Asymmetric information with respect to risk type only

Insurers can observe whether or not someone is in U, but not which of H or L she may be in. In the terminology of Doherty and Thistle (1996), they can observe *informational status*, the fact of knowing or not knowing one's risk type, but not the risk type itself. Now we have an adverse selection problem with respect to the H and L type, but not for the U type. By concentrating on the Rothschild–Stiglitz equilibrium, it is quite clear that the new equilibrium contracts will be H^*, U^*, and $\hat{L}$, where H^* and $\hat{L}$, are the RS contracts (see Figure 4.10).

It follows immediately that the premium risk individuals in U would be facing if they were to undertake a genetic test is now increased, as compared to the previous case. The expected utility resulting from testing falls, because with probability $1 - \gamma_h$ the individual post-testing will buy the contract $\hat{L}$, which yields a lower expected utility than L^*. Thus the U types will not become tested.

However, the fact that the test is costless and verifiable must imply that in this case the adverse selection problem disappears completely, if individuals can show their test results to the insurer. By showing that they are low risks individuals in L can obtain contract L^* instead of $\hat{L}$. So they have an incentive to take the test and all three types again get full insurance at their fair premium. This is of course a genuine welfare gain resulting from existence of the test, which is a costless and perfect signal of type.

Now we turn to the case where the insurer does not even know whether a test was undertaken or not. This case seems to become more and more relevant, as genetic tests are already being offered via internet, so they could be taken without knowledge of the insurer (even if one's credit card company may know about it!).

Completely asymmetric information

Assume now that insurers cannot observe which of the three subsets an individual is in. In the terminology of Doherty–Thistle, they cannot observe informational status. In that case, they cannot offer the contract U^*, since they would not be able to prevent individuals in H from taking this contract. In effect, they have an adverse selection problem with respect to the subsets H and U. As shown by Doherty and Thistle, this has a dramatic effect on the existence of premium risk. The competitive market equilibrium contracts that solve this adverse selection problem are H^* and $\hat{U}$ as shown in Figure 4.11. On the other hand, on the same argument as we just made, insurers will offer the contract L^* to anyone producing test results establishing that they are low risk. This will be done by everyone in L. Consider then an individual in U. Because the no-test contract for this individual is now $\hat{U}$ and not U^*, premium risk disappears and she will now have a positive expected gain from taking the test.

This can be seen as follows. If she takes the test, with probability γ_h she will be high risk and will receive the contract H^*, with a utility for certain that we denote as $u(H^*)$. With probability $1 - \gamma_h$ she will be low risk, provide the insurer the test result and receive the contract L^* with a certain utility of $u(L^*)$. Finally if she does not take the test

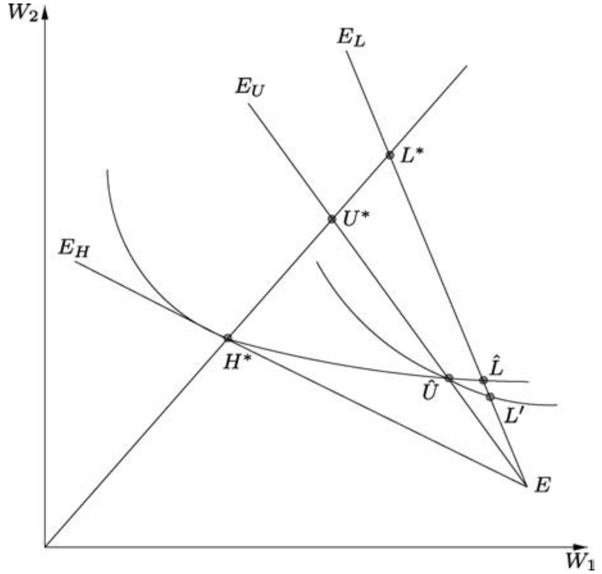

Fig. 4.11 Contracts under completely asymmetric information with and without voluntary provision of information.

she receives the contract $\hat{U}$ with expected utility

$$\bar{u}(\hat{U}) = (1-\pi_u)u(\hat{W}_1) + \pi_u u(\hat{W}_2), \tag{4.5}$$

where $\hat{W}_1$ and $\hat{W}_2$ are the wealth levels associated with $\hat{U}$ in the figure. Thus the value of taking the test is

$$\gamma_h u(H^*) + (1-\gamma_h)u(L^*) - \bar{u}(\hat{U}). \tag{4.6}$$

From the incentive compatibility constraint with regard to H^* and $\hat{U}$, we know that

$$u(H^*) = (1-\pi_h)u(\hat{W}_1) + \pi_h u(\hat{W}_2). \tag{4.7}$$

Substituting for $u(H^*)$ and $\bar{u}(\hat{U})$ in (4.6), and recalling that $\pi_u = \gamma_h\pi_h + (1-\gamma_h)\pi_l$, we obtain after simplifying that the value of taking the test is

$$(1-\gamma_h)\{u(L^*) - [(1-\pi_l)u(\hat{W}_1) + \pi_l u(\hat{W}_2)]\} > 0. \tag{4.8}$$

So untested individuals have an incentive to take a test. But this means that the situation shown in Figure 4.11 cannot be an equilibrium. There will be no takers for the contract $\hat{U}$. Instead, insurers will offer the two contracts H^* and L^*, the latter available only to those who present negative test results. All those in U will take the test, since the gamble with probability γ_h of H^* and $1-\gamma_h$ of L^* is strictly better than the certainty of H^*. Likewise all those in L will take the test since it is costless and brings them the contract L^*. Thus the only equilibrium of the insurance market in this case, is that everyone in U and in L takes the test, test results can be provided to insurers, those proving they are low risks receive the contract L^*, the remainder receive H^*. The end result is therefore the separating equilibrium $[H^*, L^*]$.[10]

This works as long as individuals are allowed to show the insurer their test results if they want to do so. However, in some countries information relating to genetic testing may neither be required by insurers nor voluntarily provided by buyers. With a similar argumentation as before it follows that in this case the equilibrium will be the

[10] If testing is costly, then this result need no longer hold. See Doherty and Thistle (1996) for details.

Rothschild–Stiglitz contracts $[H^*, \hat{L}]$ as shown in Figure 4.11. Insurers are initially faced with the three subsets, among which they cannot distinguish. So we have a double adverse selection problem. The candidates for a separating equilibrium would then be $[H^*, \hat{U}, L^{'}]$ in Figure 4.11. However, in that case it pays each buyer in U to be tested, because a lottery involving H^* and $L^{'}$ yields higher expected utility than $\hat{U}$. If she turns out to be high risk she is no worse off than at $\hat{U}$, and if low risk she will be better off (an L-type indifference curve drawn through point L' lies above $\hat{U}$). Thus insurers can conclude that everyone will either be in H or L, but low risks are no longer able to signal their type by the test. Hence we have the more costly signaling of the Rothschild–Stiglitz equilibrium.

So should we conclude that allowing insurers to see at least the test results is optimal? Or, even better, to get a full information scenario where insurers also can ask for the test results? The main argument why this is not necessarily the optimal regulation is that in both cases there will be a premium risk for the uninformed in case they take the test. So uninformed individuals will be deterred from taking a test if they know that in case the test result is positive, thus indicating an illness, they will be forced to pay a higher premium. Individuals will then even avoid useful tests, which help as they allow for early treatment. The optimal regulation is quite subtle and a lot of work has been done on this (e.g., Crocker and Snow, 1992; Doherty and Thistle, 1996; Brockett et al., 2000; Hoy and Polborn, 2000; Strohmenger and Wambach, 2000; Hoy and Ruse, 2005). In the model as it is set up here, making information symmetric and e.g., subsidizing high risks to avoid the risk premium would be optimal. Tabarrok (1994) has indeed proposed a genetic insurance, which is to be bought before the genetic test is undertaken and which insures against all negative financial consequences like, e.g., an increase in the health insurance premium which follows from a positive test result. However, if tests can be taken secretly, such an insurance market will break down. As one result it should have become clear that different insurance markets should be regulated differently, as the premium risk in the health insurance market is presumably much more serious than in the life insurance market, as people can more easily live without the latter than without the former. The regulation in the United States, which specifies no information disclosure for health insurance but disclosure for life insurance might go in the right direction. However, the resulting problems with adverse selection in the health insurance market must then be taken care of, e.g., with group insurance contracts or with a nationwide basic insurance contract (which would help to avoid cream skimming by the insurers).

4.5 Dynamic Adverse Selection

Looking around, we can observe many different forms of long term insurance contracts: In automobile insurance this year's premium depends on whether or not the insured had an accident last year. Health insurance contracts usually last for several years, and to exit from an existing contract might be costly. In unemployment insurance, the payment decreases with the time one is unemployed, while the premium usually does not depend on the duration of the employment. However, there also exist policies, as in the legal or liability insurance market, which only last for one year. Renewal is possible, but the terms of the contract are independent of whether a policy was acquired last year. What are the reasons for these differences?

The models under symmetric information are not useful in understanding long term contracts. Suppose a person faces a risk of having an accident in any one year of π. In a competitive market this person would receive insurance at the fair premium, and that in every year. No long term contract could do any better. As a matter of fact, any long term contract which makes zero profit in expectation is equivalent to a series of short term (one-period) insurance and saving contracts.[11] But if saving aspects are ignored, long term contracts do not improve efficiency under symmetric information.

[11] In some models, it is the saving motive which is the reason for long term contracts. If customers are credit constrained, for example, then a long term contract which specifies a very low premium in the beginning and a larger premium later on serves as a means of redistributing money across periods. Although in some cases this might be a relevant issue to discuss, we exclude this aspect in this section.

We have to turn to asymmetric information to understand the economics of long term contracts. Asymmetric information in the form of adverse selection will be discussed in this section, while moral hazard will be dealt with in Section 5.

4.5.1 Finitely Many Periods

As before we simplify by considering two risk types only. One, the high risk type, has accident probability π_h, and the other has $\pi_l < \pi_h$. The proportion of the high risks in the population is γ_h. Now consider a two-period model, where the agents face the same risky environment in each period. In such a situation, a long term insurance contract for type i consists of six parameters: P_i is the premium paid in period one. I_i is the net payment in case of an accident in period 1, paid out in period 1. $P_i(n)$ is the premium in period 2, if no accident has happened in period 1, and $P_i(a)$ is the corresponding premium if an $a$ccident has occurred. $I_i(n)$ and $I_i(a)$ are the corresponding net indemnities.

We speak of a long term contract if the optimal contract is such that it differs from two short term contracts. The problem with this definition is that it is not clear what two short term contracts would look like. This is not a problem if one considers a Rothschild–Stiglitz equilibrium only: As long as the risk probabilities stay constant over time, and there are at least some high risks who buy on the spot market in the second period, the spot market contracts in the two periods are the same. It is more difficult under the WMS equilibrium concept, where the contracts depend on the ratio of high risks to low risks. The problem is that in equilibrium everyone buys a long term contract, so that no-one is active on the spot market in period 2. So therefore also the ratio of high to low risks is undetermined, which makes it difficult to tell which WMS contracts are the relevant spot market contracts.

For our purpose we define a long term contract as a contract where the premium or indemnity in period 2 differs from that in period 1. As we will see, this interpretation is sufficient for the economics we want to bring across: The usefulness of *experience rating*.

Facing a long term insurance contract, the expected utility of an individual of type i is:

$$\begin{aligned}&(1-\pi_i)u(W-P_i)+\pi_i u(W-L+I_i)\\&\quad+(1-\pi_i)[(1-\pi_i)u(W-P_i(n))+\pi_i u(W-L+I_i(n))]\\&\quad+\pi_i[(1-\pi_i)u(W-P_i(a))+\pi_i u(W-L+I_i(a))],\end{aligned}\tag{4.9}$$

which we abbreviate to

$$Eu_i(P_i,I_i)+(1-\pi_i)Eu_i(P_i(n),I_i(n))+\pi_i Eu_i(P_i(a),I_i(a)).\tag{4.10}$$

By writing down this expression, we made a series of simplifying assumptions. First, it is assumed that the income of the agent in both periods is the same. Different wealth in the different periods would make the formulation more messy. Furthermore, the spot market contracts might also differ between the two periods, as the insured are in general more or less risk averse if they are poorer or richer. But overall the economics stays roughly the same. Second, the agent does not save but consumes all the income she has. Allowing for savings would create different problems, some of which will be discussed below. But note that if the contract specifies full insurance at the same price in both periods, then there will be no incentive to save anyway. A third assumption is that there is no discounting. This is not critical and just made for simplicity.

We discuss the results in a model of perfect competition by using the WMS equilibrium concept. The qualitative features of a long term contract do not differ when one uses the Rothschild–Stiglitz equilibrium instead, but formally the results are easier to see in the WMS framework.

To derive the WMS equilibrium, the following optimization problem has to be solved:

$$\begin{aligned}\max_{P_i,I_i,P_i(n),I_i(n),P_i(a),I_i(a)} \; & Eu_l(P_l,I_l)+(1-\pi_l)Eu_l(P_l(n),I_l(n))\\ & +\pi_l Eu_l(P_l(a),I_l(a))\end{aligned}$$

s.t.

$$\begin{aligned}&Eu_h(P_h, I_h) + (1-\pi_h)Eu_h(P_h(n), I_h(n)) + \pi_h Eu_h(P_h(a), I_h(a))\\&\quad \geq Eu_h(P_l, I_l) + (1-\pi_h)Eu_h(P_l(n), I_l(n)) + \pi_h Eu_h(P_l(a), I_l(a))\end{aligned} \tag{IC}$$

$$\begin{aligned}&(1-\gamma_h)\{(1-\pi_l)P_l - \pi_l I_l + (1-\pi_l)[(1-\pi_l)P_l(n) - \pi_l I_l(n)]\\&\qquad +\pi_l[(1-\pi_l)P_l(a) - \pi_l I_l(a)]\}\\&\qquad +\gamma_h\{(1-\pi_h)P_h - \pi_h I_h + (1-\pi_h)[(1-\pi_h)P_h(n)\\&\qquad -\pi_h I_h(n)] + \pi_h[(1-\pi_h)P_h(a) - \pi_h I_h(a)]\} \geq 0. \qquad \text{(PC)}\end{aligned} \tag{4.11}$$

This is the generalization of the maximization problem (4.2): Maximize the utility of the low risks such that the high risks prefer their contract to that of the low risks (the incentive constraint) and the insurance companies make no loss (the participation constraint).

First, let us work through the first-order conditions with respect to the high risk contract parameters, where λ (respectively, μ) is the Lagrange parameter of the incentive (respectively participation) constraint:

$$\begin{array}{lr} P_h & -\lambda(1-\pi_h)u'(W-P_h) + \mu\gamma_h(1-\pi_h) = 0\\ I_h & \lambda\pi_h u'(W-L+I_h) - \mu\gamma_h\pi_h = 0\\ P_h(n) & -\lambda(1-\pi_h)(1-\pi_h)u'(W-P_h(n)) + \mu(1-\pi_h)\gamma_h(1-\pi_h) = 0\\ I_h(n) & \lambda(1-\pi_h)\pi_h u'(W-L+I_h(n)) - \mu(1-\pi_h)\gamma_h\pi_h = 0\\ P_h(a) & -\lambda\pi_h(1-\pi_h)u'(W-P_h(a)) + \mu\pi_h\gamma_h(1-\pi_h) = 0\\ I_h(a) & \lambda\pi_h\pi_h u'(W-L+I_h(a)) - \mu\pi_h\gamma_h\pi_h = 0. \end{array} \tag{4.12}$$

Note that the "no-accident" (respectively "accident") first-order conditions are very much the same as those for period 1, the only difference is that they are multiplied by $(1-\pi_h)$ (or π_h, respectively) which cancels out. Thus it follows that the high risks obtain full insurance, and their wealth in all states of the world is the same: $P_h = P_h(n) = P_h(a)$ and $I_h = I_h(n) = I_h(a) = L - P_h$. High risks obtain full insurance and no income variations across periods. The exact premium charged depends on the degree of cross-subsidization between low and high risks, which in turn depends on the ratio of high to low risks in the population. If there are many high risks around, so that the Rothschild–Stiglitz equilibrium becomes relevant, then $P_h = \pi_h L$. Note that with our definition the optimal contract of the high risks is not a long term contract, as the terms of the contract do not differ between periods.

This result depends on the assumptions that wealth in both states is the same and utility functions do not differ. If these were relaxed, then the contract in period 2 would still not depend on whether an accident did occur or not in period one, but it would differ from the contract in period 1. Observe that marginal utility across states is equalized for the high risks, so this implies that wealth is equalized only if the utility functions are the same across states and time. And this implies that the contract is the same in the two periods, only if initial wealth in both periods is the same. Note also, that with such a contract the high risks have no incentive to save or dissave money.

The first-order conditions for the low risks are slightly more cumbersome:

$$\begin{array}{ll}
P_l & -(1-\pi_l)u'(W-P_l) + \lambda(1-\pi_h)u'(W-P_l) \\
 & \qquad +\mu(1-\gamma_h)(1-\pi_l) = 0 \\
I_h & \pi_l u'(W-L+I_l) - \lambda\pi_h u'(W-L+I_l) - \mu(1-\gamma_h)\pi_l = 0 \\
P_l(n) & -(1-\pi_l)^2 u'(W-P_l(n)) + \lambda(1-\pi_h)^2 u'(W-P_l(n)) \\
 & \qquad +\mu(1-\pi_l)^2(1-\gamma_h) = 0 \\
I_h(n) & (1-\pi_l)\pi_l u'(W-L+I_l(n)) - \lambda(1-\pi_h)\pi_h u'(W-L+I_l(n)) \\
 & \qquad -\mu(1-\pi_l)\pi_l(1-\gamma_h) = 0 \\
P_l(a) & -\pi_l(1-\pi_l)u'(W-P_l(a)) + \lambda\pi_h(1-\pi_h)u'(W-P_l(a)) \\
 & \qquad +\mu\pi_l(1-\pi_l)(1-\gamma_h) = 0 \\
I_h(a) & \pi_l^2 u'(W-L+I_l(a)) - \lambda\pi_h^2 u'(W-L+I_l(a)) - \mu\pi_l^2(1-\gamma_h) = 0.
\end{array} \tag{4.13}$$

Reformulating the expressions for the premia gives:

$$\begin{array}{ll}
P_l & u'(W-P_l) = [1-\lambda(1-\pi_h)/(1-\pi_l)]^{-1}\mu(1-\gamma_h) \\
P_l(n) & u'(W-P_l(n)) = [1-\lambda(1-\pi_h)^2/(1-\pi_l)^2]^{-1}\mu(1-\gamma_h) \\
P_l(a) & u'(W-P_l(a)) = [1-\lambda\pi_h(1-\pi_h)/\pi_l(1-\pi_l)]^{-1}\mu(1-\gamma_h).
\end{array} \tag{4.14}$$

As

$$\frac{\pi_h}{\pi_l}\frac{(1-\pi_h)}{(1-\pi_l)} > \frac{(1-\pi_h)}{(1-\pi_l)} > \frac{(1-\pi_h)^2}{(1-\pi_l)^2} \tag{4.15}$$

it follows that

$$P_l(a) > P_l > P_l(n). \tag{4.16}$$

The agent is "penalized" in period 2 if a loss occurred in period 1, but "rewarded" for no loss, i.e., the terms of her insurance contract are better if she does not have an accident. Note that there is nothing the agent can do about the accident, so that penalizing and rewarding are not meant in the sense that they give the insured an incentive to avoid losses. The contract structure is such that the high risks have no incentive to choose the contract designed for the low risks. And as they have a larger probability of having an accident, they are more afraid of the "penalty" which might occur.

Reformulating the expressions for the indemnity, and comparing all equations it is easy to see that the low risks receive partial insurance in both periods, and that the indemnity can be ranked as well:

$$I_l(a) < I_l < I_l(n) < L - P_l(n). \tag{4.17}$$

In this case, the assumption we made before that the insured cannot save may becomes binding, as the agent anticipates that she has different outcomes in period 2.

If there are more than two periods, the high risks still obtain full insurance in every period, independently of whether an accident did occur or not, while the low risks obtain a long term policy with partial insurance in every period. The premium indemnity schedule is given by $(P_l(t,j), I_l(t,j))$ where t denotes time and j the number of accidents which occurred already. It can be shown that $P_l(t,j)$ is increasing in j while $I_l(t,j)$ decreases in j for constant t (Cooper and Hayes, 1987). Note that it does not play a role when exactly the accident happened, only how often an accident occurred. The reason is that the exact timing does not give more information on the risk type: Any risk type has the same probability of, say, having accidents in periods 1, 2, 5, or in periods 2, 3, 6, respectively.

This structure of the low risk contract is a feature we observe in many insurance markets, most notably automobile insurance. It is known as *experience rating* or *bonus-malus* system. To make the argument again: These contracts provide the low risks the optimal protection under the constraint that the high risks do not choose this contract. In the model presented here, experience rating is *not* a means to provide incentives to take more care about accident prevention.

Before finishing this section, one remark on general adverse selection models is in order. As we have stressed several times, the problem we face in the insurance market is formally quite similar to that faced by a monopolist selling a good, a government procuring weapons, and many other principal agent models. However, with respect to long term contracts, one difference between the standard applications and the insurance market exists: In the insurance market, over time information about the type of agent is revealed, independently of which contract is signed. If someone has an accident, she is more likely to be a high risk than a low risk. In standard principal agent models this is different. If for example a firm does not sell any output to its customers, then it will obtain no new information concerning their willingness to pay. It actually turns out that in those standard principal agent models, the optimal long term contract is just the repetition of the one-period contract, which certainly differs from the result obtained for the insurance market.

4.5.2 Infinitely Many Periods

Although insurance contracts do not in general last longer than a lifetime, considering infinitely many periods is useful as it clearly brings out the advantage long term contracts provide. These can most easily be illustrated with an example. Suppose you have a coin, which might be manipulated such that heads appears twice as often as tails. But you do not know for sure. What would you do to find out?

You would probably throw the coin 1000 times or so, and write down how often heads appeared. One would expect that the manipulated coin comes up with heads much more often. Once we have this number, we can calculate the probability of this happening under the two scenarios.[12] For example, if head appeared 550 times, and the coin was expected to be manipulated with probability 1/2, then the revised belief that the coin is still manipulated is given by $2 \times 0.67^{550} 0.33^{450}/(0.5^{550} 0.5^{450} + 0.67^{550} 0.33^{450})$, which is approximately equal to 10^{-12}, quite a small number.

This effect can also be used in the insurance market. Actually, we have already done this before, just with 2 periods. If very many periods are possible, then one should obtain quite precise information about the riskiness of the type. In the limit, the information should be so good that the first best can be closely approximated. And it will, as we now show.

To obtain efficiency, both types of agents should obtain a contract which specifies full insurance in every period at their fair premium rate. However, to avoid that the high risk type buys the contract designed for the low risk type, the contract for the low risk type must provide some form of partial (or no) insurance if the number of accidents is too large as would be expected from a low risk person. On the other hand, the low risks should not suffer from this partial insurance. The way to achieve this is to set contracts in the following way: Let $(P_h, I_h) = (\pi_h L, (1 - \pi_h)L)$ in each period, and

$$(P_l, I_l) = \begin{cases} (\pi_l L, (1 - \pi_l)L) & \text{if } \frac{N}{T} < \pi_l + \delta(T) \\ (0,0) & \text{otherwise.} \end{cases} \tag{4.18}$$

Here T is the period and N is the number of accidents which occurred so far.

The whole trick lies in finding the appropriate $\delta(T)$ function, which should be large enough that the low risks have only an infinitesimal risk of obtaining zero insurance, and small enough that the high risks have a significant risk of losing cover if they choose this contract. As is quite obvious from the remarks above, $\delta(T)$ will be a decreasing function in T, more periods allow one to get much better information about the true risk type of the agent.

[12] This probability is given by $\binom{1000}{N} p^N (1 - p)^{1000-N}$, where N is the number of heads.

One function which satisfies the above-mentioned requirements is

$$\delta(T) = \sqrt{2\gamma\pi_l(1-\pi_l)\log[\log[T]]/T}, \tag{4.19}$$

where γ is some parameter larger than one.[13]

To summarize the results we obtained: Long term contracts allow for a weakening of the incentive compatibility constraint. Involuntarily the high risks reveal information about their type, because they have more accidents on average. So making the terms of the contract improve if no accident has occurred, but worsen if an accident did happen, the policy becomes more acceptable to the low risks than to the high risks. This is the structure known as bonus-malus systems or experience rating. In the limit of infinitely many periods, the first best can be achieved.

4.5.3 Renegotiation

So far we have assumed that long term contracts are enforceable, which is to say that once the contract is signed, both parties will stick to it. However, there are at least two reasons why the enforceability of long term contracts is limited. One is legal, the other economic.

In some situations, laws prevent long term contracts. Most famous is the prohibition of slavery: a worker is not allowed to commit herself to work with a company for 20 years, say. On the other hand, the company might well offer the worker a long term contract which it cannot breach, while she can. Similarly in insurance markets: In some sectors firms offer long term contracts, but the insured are allowed by law to opt out of the contract each year.

The other reason for renegotiation lies in economics, in particular the tendency for wanting to capture efficiency gains: If the long term

[13] The "Law of the Iterated Logarithm" states that for any sequence of independent identically distributed random variables $\{x^t\}$, with finite mean $\bar{x}$ and finite variance σ^2, and for any $\gamma > 1$, almost surely

$$\lim_T \sup \frac{|\bar{x} - T^{-1}\sum_{i=1}^{T} x^i|}{\sqrt{2\gamma\sigma^2 \log[\log[T]]/T}} < 1.$$

With $\sigma^2 = \pi_l(1-\pi_l)$, the low risks' average number of losses will almost surely for all but finitely many T be smaller than $\pi_l + \delta(T)$. It is slightly more demanding to show that the high risks indeed prefer their contract to that of the low risks. We refer the reader to Dionne (1983), who discusses the issue with a continuum of types.

contract is signed, the insurer knows the type of the insured. But we have seen that the contract for the low risk type is inefficient in the second period as it specifies partial insurance. But then profitable renegotiation between the insurer and the insured could take place. And if both parties agree to renegotiate, then no court will forbid them to change the conditions of the contract. However, surely, if renegotiation could take place, this will be anticipated by the high risks, who then might choose the low risk contract in expectation of profitable renegotiations. So a priori it is not clear what will happen, and we have to go into the model in more detail. We will discuss the two issues in turn.

One-party commitment First, consider the possibility of renegotiation due to laws, which, although stated as if they give the insured the flexibility to change firms every year, in effect prohibit the insured to commit to a binding long term contract.

We work again in the two period model used before. If only the insurer, but not the agent, can commit to long term contracts, the low risk might quit her contract if an accident occurred. Furthermore, a high risk type might perhaps choose the low risk type contract and, in case of an accident, change the insurer.

This introduces at least one further constraint in the optimization problem:

$$EU_l(P_l(a), I_l(a)) \geq EU_l(P_l(s), I_l(s)), \tag{4.20}$$

where $(P_l(s), I_l(s))$ is the contract offered to the low risks on the spot market in period 2. As already discussed above, it is not quite clear what this contract is, as it is only offered out of equilibrium. In equilibrium the low risks do not opt out of their long term contract. However, as high risks could also buy single period contracts out of equilibrium, a good starting point for the analysis would be to assume that the spot contracts are the Rothschild Stiglitz contracts. But note that these policies only play a role in so far as they give the outside option of a low risk type on the spot market, they do not change the qualitative structure of the optimal contract. In addition there might be a further constraint for the high risks, as mentioned above.

Instead of going through all the equations again, we discuss the results[14]: High risks obtain full insurance as before. It also still holds that in case of no accident the low risks are rewarded, that is the premium is lower and the net indemnity is larger in the second period. The policy in case of an accident however is modified. It can be shown that:

$$EU_l(P_l(n), I_l(n)) > EU_l(P_l(a), I_l(a)) > EU_l(P_l, I_l). \qquad (4.21)$$

The low risk type is better off in period 2 than in period 1, independent of whether she has had an accident or not, but she is even better off in case of no accident. This ranking of utilities is necessary such that the low risk type does not leave the insurer after period 1.

An interesting result follows from this in the case where there are many high risks. Then, as in the RS model, the high risks obtain full insurance at their fair premium and there is no cross-subsidization. The insurance company will then make an expected profit from the low risk type in period 1 and an expected loss in period 2. Even if the overall expected profit is not zero, insurers make lower profits per insured in later periods.

There is an interesting debate on whether insurers make a profit with their clients first and losses later on or whether it is the other way around. This model seems to suggest that it is the former case: insurers make losses later on so that customers do not prefer to change the company. As an example consider private health insurance markets, where in some cases the insured explicitly pays more in earlier periods to obtain lower premia in later periods.[15] While Dionne and Doherty (1994) find some evidence for "highballing," others have found evidence which suggests that insurer make losses first and profits later on (also known as "lowballing," D'Arcy and Doherty, 1990). One possible explanation for this could be that the insurers first learn about the type of the agent and exploit this knowledge at later stages. Com-

[14] A detailed analysis can be found in Cooper and Hayes (1987).

[15] There are other reasons for this phenomena: First, such a contract has a saving element in it: save now for the higher future premia. Second, individuals learn more about their risks when they grow older. Locking the customer in might be a means to prevent her from going to another insurer if she turns out to be a low risk, and to stay with the insurer only if she is a high risk.

petition then drives the market to zero expected profits, which implies losses first and profits later on (Kunreuther and Pauly, 1985).

Renegotiation in period 2 In contrast to the last subsection, now both parties can sign a long term contract. However, as discussed earlier, they cannot prevent themselves from renegotiating to a better contract. This is actually what the Coase Theorem says: *Bargaining will achieve an efficient allocation of resources whatever the allocation of property rights, if transaction costs are zero.* So if the insurer knows the type of insured for sure, such that there are no transactions costs due to asymmetric information, then an efficient outcome will be obtained.

If we assume that the insured are perfectly separated in period 1, this implies that both contracts for the low risk in period 2 have to specify full insurance, whether an accident has occurred or not. The only contracts which are fully efficient are those with full insurance. This then introduces two further constraints, the so-called renegotiation proofness constraints:

$$I_l(n) = L - P_l(n); \quad I_l(a) = L - P_l(a) \tag{4.22}$$

The contract can still be different in case of an accident and if no accident occurred, but it cannot specify partial insurance as before.

Note that due to renegotiation the low risks lose. The optimization problem in a WMS equilibrium is the same as before, but now with two additional constraints, and so the low risks cannot be made better off. We can say even more in a Rothschild–Stiglitz equilibrium: If renegotiation is possible, then the high risks still obtain the same full insurance contract at their fair premium, the firms still make zero profit, and the low risks are *worse* off, a clear Pareto worsening. The inefficiencies in the contract were used to prevent the high risks from taking the low risk contract. If for some reason these inefficiencies cannot be used, then the separation of types is much harder to achieve. Separation is still possible, but the contract in case of an accident has to be sufficiently bad, while that in case of no accident has to be very good. So in addition to the partial insurance she obtains in period one, the low risk type faces a larger future income risk than she would have had in the case of no renegotiation.

Dionne and Doherty (1994) combine both approaches and analyze a semicommitment contract (i.e., only the insurer can commit to a long term contract) with renegotiation (which occurs if the stated contract is inefficient). They obtain partial pooling in the first period where it is not optimal to separate the high risks from the low risks fully.

With this we end the section on adverse selection. We have discussed one possible reason for many aspects which can be observed in the real world: Partial insurance contracts, categorical discrimination, experience rating or bonus-malus systems. Although the phenomenon of adverse selection has been known for a long time, and the formalism was established more than than 30 years ago, markets under adverse selection are still a very active research area both in the insurance as well as in the general economics literature (see Chiappori et al., 2006).

We now turn to another problem of asymmetric information, which similar to adverse selection has turned out to be one of the most widely discussed topics over the last 20 years: Moral hazard.

5

Moral Hazard

5.1 Introduction

The introduction of the safety belt was celebrated as a major step forward to reduce the number of fatal automobile accidents. This seems obvious, as the risk of having major injuries is significantly reduced. In many countries, safety belts are now legally required to be worn.

However, although the number of injuries per accident decreased in the first years, the number of accidents actually increased. This latter effect was so dominant that the overall number of fatalities stayed roughly constant. And the incidence of injuries changed: Major injuries shifted away from the drivers of cars to pedestrians and cyclists. So altogether probably only the repair industry initially profited from the introduction of safety belts.[1] Safety belts are an insurance device. In case of an accident, less damage will occur to the driver. This is quite similar to an insurance contract, which pays out in case of an accident, making the damage less severe. Now people, who installed the "safety belt insurance" felt less inclined to drive safely, as they would have if

[1] The phenomena, that people adjust their behavior to regulation in a way which counteracts the intended effects of regulation, is called the "Peltzman-Effect" (Peltzman, 1975).

no safety belt existed.[2] This is a common phenomenon in the insurance market, known as *moral hazard*. Other examples are: Health insurance may induce people to be less careful when playing dangerous sports; having a property insurance will make one think whether it is really necessary to take care of your premises; with crop insurance, a farmer may work less hard to cultivate her fields.

We have been quite careful only to cite examples where individuals provide less effort in case of full insurance. In the literature, it is often assumed that individuals invest less financially in case of full insurance. For example, a person acquires fire insurance and does not install smoke detectors or fire sprinklers. Someone has earthquake insurance which makes her more willing to build a less earthquake-proof house. Installing burglar alarms, locks, or ferocious dogs might be considered less necessary if burglary insurance is available. The problem with these latter examples is that if the moral hazard problem consists of underinvestment, it should be possible to write the efficient investment into the contract. Surely, a clause "indemnity is only paid if sprinkler system A is installed" would lead the homeowner to install that system. Or flood damage insurance may specify that a house must not be built on a flood plain. Even if monitoring of the investment is costly, it is still possible *a priori*, and that should be included in the model. Therefore, in the following we concentrate on effort costs and not financial costs.

Unfortunately, the term "moral hazard" has a second meaning in the insurance literature. It is also used if people who hold health insurance consume more health services than would be optimal. This effect arises because insurance companies pay for treatment rather than indemnifying the patient. To distinguish this effect from the underprovision of effort as discussed here, we will denote it as *ex-post* moral hazard, because the behavior occurs *after* the loss has occurred.[3]

Apart from the insurance context, models with (*ex ante*) moral hazard cover a wide range of economic phenomena. Some examples: A manager, who is "insured" by receiving a fixed wage, has no incentive to

[2] People were asked whether they drove more dangerously with their safety belts on — they said no. However, when they were asked whether they would drive more carefully if their car had no safety belts, they agreed.

[3] *Ex-post* moral hazard is discussed in Zweifel et al. (2007, chap 6).

work hard. Banks, who invest into foreign countries, and which are "insured" through bailouts by the IMF if that country collapses, have less incentives to screen the projects they finance carefully. Students, who have passed their midterm exam with a good grade, and who are thus "insured" against failing the whole term, may be less inclined to work hard for the final exam.

The last example is quite instructive as it gives a hint of how to overcome some of the problems moral hazard creates. In many cases students do not just receive a pass/fail mark, they can also acquire different degrees like distinction, a prize for the best exam, etc. Thus, even if they are "insured" against failing, there is still the incentive to work hard to obtain a good grade. In the models we present, a similar effect will hold: Individuals must somehow profit from the effort they put in. If they do not, they do not incur the costs of effort.

In the following we will cover the literature step by step: We start with a simple example, with only two outcomes and two effort levels ("lazy" and "hard working"). This model serves two purposes: First, it discusses how the costs of effort can be modeled, and second, we see partial insurance coverage appearing as a second best contract. Then the model is extended to more than two effort levels. In this context, the famous problem of the "First-Order Approach" will be discussed. In a third step, continuous outcomes are considered. The most general case is instructive as it teaches that not many robust results can be obtained. However, one result which emerges is that moral hazard, although it creates inefficiencies, does not lead to a breakdown of the market.

Having derived the most general form in the static model, in the following section we turn to dynamic moral hazard problems. This is done in two steps: First, two periods are considered and second, infinitely many periods. The focus in this part lies on the question whether long-term contracts like for example experience rating can be useful to deal with moral hazard. Allowing for more than one period, renegotiation may become an issue again. This is discussed in Section 5.3.

The term moral hazard is used to describe the situation where at the contracting stage, both parties have the same information, while after the contract is signed, one party to the contract obtains private information which she exploits to her advantage, e.g., by providing too little effort (in the models described so far) or by consuming too much (in the case of *ex-post* moral hazard). Insurance fraud is another example of one party having an informational advantage, as only the insured knows whether the accident has occurred or not and how large the size of the loss is exactly. Depending on the technology available, the insurer might find out whether a claim is fraudulent or not. These cases are analyzed in Section 5.4.

5.2 Single Period Contracts Under Moral Hazard

5.2.1 The Basic Model

We use the by now well known two states of the world model: One state with no loss, the other where the loss occurs. In contrast to the standard model, in this section the probability of the damage is not exogenous, but can be influenced by the insured. Formally:

$$E[U] = (1 - \pi(e))u(W - P) + \pi(e)u(W - L + I) - c(e), \qquad (5.1)$$

where we have introduced $I = C - P$ as the net compensation (indemnity) in case of a loss. Here, $\pi(e)$ is the probability that a loss occurs, which satisfies $\pi'(e) < 0$, i.e., more effort (e) leads to a lower probability of an accident.

We have written $c(e)$, with the assumption that $c'(e) > 0$, as the "cost of effort" in utility units ("utils"). It is obvious that if someone puts in effort to prevent an accident from happening, then this must be costly in some form. However, how to model this simple insight is far less trivial. Unfortunately, there is no axiomatic approach which can tell us how to do it optimally (like e.g., the axioms of expected utility lead to von Neumann–Morgenstern utility functions). Several possibilities exist, we discuss only two: First, costs could be monetary, i.e., $u(W, e) = u(W - c^m(e))$. The advantage of this way of modeling is that it is easily interpreted. However, as we have argued above, if costs are indeed monetary, in many cases contracts could condition on these costs. Then, the moral hazard problem would disappear.[4] An alternative, which we use in the following, is: $u(W, e) = u(W) - c(e)$,

[4] For an analysis with monetary effort costs, see Arnott and Stiglitz (1988a).

i.e., the utility function is additively separable in income and effort. In this case the reduction in utility is independent of the state of the world, i.e., whether a loss has occurred or not, "costs" of effort are $c(e)$. Furthermore, preferences over lotteries do not depend on the amount of effort taken.[5]

One might wonder whether $u(W,e)$, the most general formulation, would not be appropriate. The main reason for this and for the choice we make is practicability: By going through the section it becomes clear that it is not always easy to find a solution to the moral hazard problem. For many specifications, the mathematical problem is not well defined.

Let us now go into the model. As mentioned above, there are only two states of the world. In addition, the agent has the choice between two effort levels: e_1 ("lazy") and $e_2 > e_1$ ("hard working").

First consider the "first best," where effort is observable and contractible. We know from Section 2 that without informational asymmetries full insurance is optimal. The premium will be $P = \pi(e)L$, depending on the effort level. Therefore either effort level e_1 or effort level e_2 is optimal, depending on which of the two expressions is larger:

$$u(w - \pi_1 L) - c(e_1) \gtreqless u(w - \pi_2 L) - c(e_2),$$

where $\pi_i = \pi(e_i)$, $i = 1,2$.

To make the problem interesting, let us assume that effort level e_2 is the first best effort level. If e_1 were preferred, then even in the case of non-observability of effort the moral hazard problem would cease to exist, as the agent could receive her full insurance contract as before and just stay lazy.

Now turn to the "second best," where effort is not observable. Suppose that also under asymmetric information the higher effort should be implemented. The contract (premium/indemnity) must be designed such that the agent will indeed work hard. The optimization problem

[5] Another advantage of an additively separable utility function is that random contracts, where the indemnity is paid out with some probability smaller than one, are never optimal. In case of monetary effort costs, this result does not hold (see Arnott and Stiglitz, 1988b).

is then given by

$$\begin{aligned}
&\max_{P,I}\ (1-\pi_2)u(W-P)+\pi_2 u(W-L+I)-c(e_2)\\
&\qquad \text{s.t.}\\
&\text{PC:}\ \ (1-\pi_2)P-\pi_2 I \geq \Pi \qquad\qquad (5.2)\\
&\text{IC:}\ \ (1-\pi_2)u(W-P)+\pi_2 u(W-L+I)-c(e_2)\\
&\qquad \geq (1-\pi_1)u(W-P)+\pi_1 u(W-L+I)-c(e_1).
\end{aligned}$$

As in the previous section, PC stands for participation constraint, i.e., the insurance company must obtain at least profit Π to agree to trade with the agent. Note, that by varying Π the whole efficiency boundary of this problem can be reached. Thus if $\Pi = 0$, we are in the competitive market situation. If Π is large enough, the solution to the monopoly problem will be obtained. This holds for moral hazard problems, because *ex ante* both parties have the same information. The asymmetric information issue arises after the contract is signed, when the insured decides on which effort to choose. This is in contrast to adverse selection models, where one party has an informational advantage at the time the contract is signed. There the structure of the result depends on how the bargaining power is distributed.[6]

IC is the incentive compatibility constraint. As effort is not observable by the insurer, the contract must be such that it is in the interest of the insured to put in effort e_2 rather than e_1. This looks quite similar to the adverse selection problem discussed previously. There, however, the incentive compatibility constraint was such that it prevented one type of agent choosing the contract of the other type. Here, there is only one type of insured. But this person must have an incentive to put in the desired effort level, which makes the IC necessary.

[6] In the insurance market under adverse selection, both in the monopoly problem (Stiglitz, 1977) as well as under perfect competition (as discussed in Section 4), the high risks obtain full insurance while the low risks are underinsured. However, for other forms of principal agent models the party whose contract is distorted may well depend on which party has the bargaining power.

This is a Kuhn–Tucker problem, but fortunately we know that both constraints have to be binding. PC binds as otherwise through a decrease of P by $\epsilon/U'(W-P)$ and an increase of I by $\epsilon/U'(W-L+I)$ with $\epsilon>0$ and small, the incentive constraint does not change, the participation constraint only changes marginally, which is all right if it was slack before, while the utility of the insured increases. Note that this follows from the assumption of additively separable utility functions. In the case of monetary costs, for example, it might happen that the PC does not bind at the optimum. The IC must be binding, because without it, we know that full insurance would be optimal. But that would lead the agent to put in effort e_1, which is a contradiction.

The Lagrange function is then given by

$$\begin{aligned} L=&(1-\pi_2)u(W-P)+\pi_2u(W-L+I)-c(e_2)\\ &+\lambda[(1-\pi_2)P-\pi_2I-\Pi]+\mu[(1-\pi_2)u(W-P)\\ &+\pi_2u(W-L+I)-c(e_2)-(1-\pi_1)u(W-P)\\ &-\pi_1u(W-L+I)+c(e_1)]. \end{aligned} \tag{5.3}$$

The first-order conditions with respect to P and I are:

$$\begin{aligned} &-(1-\pi_2)u'(W-P)+\lambda(1-\pi_2)-\mu[(1-\pi_2)u'(W-P)\\ &\qquad-(1-\pi_1)u'(W-P)]=0\\ &\pi_2u'(W-L+I)-\lambda\pi_2+\mu[\pi_2u'(W-L+I)\\ &\qquad-\pi_1u'(W-L+I)]=0. \end{aligned} \tag{5.4}$$

If we denote $u'(W-P)=u_1'$ and $u'(W-L+I)=u_2'$, then the first-order conditions can be written as

$$\begin{aligned} \frac{1}{u_1'}&=\lambda^{-1}+\frac{\mu}{\lambda}\frac{(1-\pi_2)-(1-\pi_1)}{(1-\pi_2)}\\ \frac{1}{u_2'}&=\lambda^{-1}+\frac{\mu}{\lambda}\frac{\pi_2-\pi_1}{\pi_2}. \end{aligned} \tag{5.5}$$

An expression of this form reappears over and over in the literature on moral hazard. One over the marginal utility is equal to a constant plus another constant times an expression, which depends positively on the change in probability for different effort levels for that state and negatively on the probability of that state. Note that the last factor can be written as $1 - (1 - \pi_1)/(1 - \pi_2)$ (and $1 - \pi_1/\pi_2$, respectively). The ratio of the probabilities is also known as the likelihood ratio. If the likelihood ratio becomes smaller, then u' decreases which implies that the wealth in that state increases. For a very small likelihood ratio this state is very unlikely to occur under the low effort level, i.e., if $(1 - \pi_1)/(1 - \pi_2) = 0.1$, then it is ten times more likely to obtain the state no-accident if effort level e_2 is used instead of e_1. Consumption is optimally chosen to be large in those states to give an incentive to work hard. In this example, if $\pi_2 = 0.1$, π_1/π_2 will be equal to 9.1. So, in the accident state consumption will be much lower to give a strong incentive to prevent the accident from happening, which is much less likely under the larger effort level.

In the present case note that as $\pi_2 < \pi_1$, we have $u_1' < u_2'$ which implies that $P < L - I$, i.e., there is less than full insurance. Although that was to be expected, it is useful to denote this as the first general insight:

> *To implement higher effort levels, the agent must not obtain full insurance.*

This is one example of what was mentioned in the introduction; the insured has to profit somehow from putting in more effort. Here, this is achieved by giving her a larger utility if no accident occurs than in case of an accident. A common feature observed in insurance contracts, namely deductibles and/or partial insurance, can be explained by this.

The problem is not solved yet. We have calculated how the contract would look if the higher effort level is implemented. It is not clear, however, whether this is optimal. Although in a first best world a higher effort level may be preferred, in a second best world it might be better to implement the lower effort level, as implementing high effort leads to inefficiencies due to partial insurance. So to find the overall solution, one has to check whether the high effort level with partial insurance or the low effort level with full insurance will give the insured a larger utility.

5.2.2 Many Effort Levels

In this section, we extend the model in one direction, namely to allow for more than two effort levels. The formal implementation of many effort levels gave rise to a lengthy debate in the literature which goes under the heading of the "First-Order Approach." We will say more on this later.

Suppose the possible effort levels are $e \in E$, where E is some discrete or continuous set. The problem, which has to be solved, is the following:

$$\begin{aligned}
&\max_{e\in E,P,I} && (1-\pi(e))u(W-P)+\pi(e)u(W-L+I)-c(e)\\
& && \text{s.t.}\\
&\text{PC:} && (1-\pi(e))P-\pi(e)I \geq \Pi\\
&\text{IC:} && e = \arg\max_{\tilde{e}\in E}\ [(1-\pi(\tilde{e}))u(W-P)\\
& && \qquad\qquad +\pi(\tilde{e})u(W-L+I)-c(\tilde{e})]\,.
\end{aligned}\tag{5.6}$$

The optimal contract has to be such that the agent prefers to choose effort e, i.e., e must maximize her utility given the contract. The way the IC is written captures this problem, but is unfortunately not very helpful for finding a solution. How do we deal with an *argmax* function? There are two possibilities on how to treat the IC in such a way that standard techniques from optimization theory can be used. One is to have discrete effort levels, i.e., $E = \{e_1, e_2, \ldots\}$, in which case the incentive compatibility constraints can be written for each effort level separately: $\forall e_i \in E$ with $e_i \neq e$

$$\begin{aligned}
\text{IC:}_i\quad &(1-\pi(e))u(W-P)+\pi(e)u(W-L+I^n)-c(e)\\
&\geq (1-\pi(e_i))u(W-P)+\pi(e_i)u(W-L+I)-c(e_i).
\end{aligned}\tag{5.7}$$

This direction was pursued by Grossman and Hart (1983). The big advantage of this approach is that, together with a finite number of outcomes, it is possible to show that the maximization problem given in (5.6) is well-defined, i.e., after reformulation the Kuhn–Tucker problem satisfies the conditions of a concave programming problem, for which we know that a solution exists.

The other possibility, and that is how we will proceed, is to use continuous effort levels and replace the incentive compatibility constraint by the first-order condition for the agent. This is the so-called *First-Order Approach*, used by Mirrlees (1971), Holmström (1979), among others. Mirrlees showed a potential flaw in this approach — it might not be well defined!

The problem is that it is not clear whether the first-order condition for the agent does describe the unique maximum: it could well describe a minimum, a saddle point, or a local, but not global maximum. If one wants to use the first order approach, one therefore always has to check whether the problem is well defined. Fortunately, in the present case it is, if we make a further assumption on the second derivative of the cost and probability function. Let us check this. For an interior solution, the IC can be replaced by

$$\text{IC:} \quad -\pi'(e)[u(W-P)-u(W-L+I)]-c'(e)=0. \tag{5.8}$$

As $\pi'(e)<0$ and $c'(e)>0$, from the first-order condition it already follows that $u(W-P)>u(W-L+I)$, i.e., partial insurance. Now check for the second-order condition:

$$-\pi''(e)[u(W-P)-u(W-L+I)]-c''(e)<0.$$

This holds for any partial insurance contract if $c''(e)>0$ and $\pi''(e)>0$, i.e., costs of effort are convex, and probability is a convex function of effort. Larger effort becomes more and more costly, and less and less productive.

The IC above already shows that to implement any effort level larger than $e_{\min}$, partial insurance is necessary. This is a very neat way of proving this result. It again shows the general trade-off in a moral hazard problem: More extensive insurance is desired as the agent is risk averse, while less insurance gives the agent more incentives to avoid the accident.

Allowing for more than two effort levels, another issue arises. Does the agent work harder or less hard in a situation under moral hazard compared to the first best, i.e., where effort is observable. As a first guess one would expect that due to the unobservability of effort the agent will always work less hard. However, as shown by Grossman and Hart (1983), this is not true in general. Examples can be constructed in which the optimal effort level is actually higher under asymmetric than under symmetric information.

> *In a second best world, the agent may either work less hard or harder than in the first best world.*

So far only two outcomes were possible. A partial insurance contract could therefore be either a contract with a deductible, or with coinsurance, or with a combination of these two. We now turn to continuous outcomes, to shed more light on the optimal contract structure.

5.2.3 Continuous Losses

This section discusses the most general case, with a continuous loss distribution. Here we distinguish two cases: Loss-prevention and loss-reduction. The former is easy to define and refers to the case where the agent can influence the probability of a loss. Loss reduction refers to the case where, by exercising effort, the insured influences the size of the loss. The formalization is however not trivial, and we will defer the discussion until we come to the section on loss reduction.[7]

Loss-prevention Loss prevention refers to the case where the insured controls the probability that a loss occurs. However, the distribution of losses cannot be influenced by the insured. So suppose losses are random with a distribution function $F(L)$ and density $f(L)$, defined on $L \in [\underline{L}, \bar{L}]$.

Denote by $I(L)$ the (net) indemnity as a function of the size of the loss. With the help of the "First-Order Approach," the optimization

[7] In the literature, loss-prevention is also known as self-protection, while loss-reduction is referred to as self-insurance, (see, e.g., Ehrlich and Becker, 1972). These wordings are ambiguous. It is not clear, for example, whether the common notion of protecting oneself does not also refer to loss reduction. For example, a bullet proof vest as a means of self protection does not prevent an attempt of murder, but it lowers the severity of the attack.

problem becomes the following:

$$\begin{aligned}
&\max_{e,P,I(L)} && (1-\pi(e))u(W-P)+\pi(e)\int_{\underline{L}}^{\bar{L}} u(W-L+I(L))f(L)\mathrm{dL}-c(e)\\
& && \text{s.t.}\\
&\text{PC:} && (1-\pi(e))P-\pi(e)\int_{\underline{L}}^{\bar{L}} I(L)f(L)\mathrm{dL}\geq \Pi\\
&\text{IC:} && \pi'(e)[-u(W-P)+\int_{\underline{L}}^{\bar{L}} u(W-L+I(L))f(L)\mathrm{dL}]-c'(e)=0.
\end{aligned}\tag{5.9}$$

Let us check whether the second-order condition for the agent has the correct sign:

$$\pi''(e)\left[-u(W-P)+\int_{\underline{L}}^{\bar{L}} u(W-L+I(L))f(L)\mathrm{dL}\right]-c''(e)<0,$$

which is satisfied if, as before, $\pi''(e)>0$ and $c''(e)>0$. Note that due to the IC the term in brackets is positive. In Section 2, we made the assumption that the cover should not be negative, as otherwise the insured would not report a loss. This introduces an additional constraint: $C(L)\geq 0$ (or $I(L)\geq -P$). Then, at the optimum:

$$\begin{aligned}
&\pi(e)u'(W-L+I(L))f(L)-\lambda\pi(e)f(L)\\
&\quad+\mu\pi'(e)u'(W-L+I(L))f(L)\leq 0
\end{aligned}\tag{5.10}$$

and $I(L)=-P$ if expression (5.10) is strictly lower than zero.

If the equality sign holds, reformulating (5.10) yields:

$$\frac{1}{u'(W-L+I(L))}=\lambda^{-1}+\mu/\lambda\frac{\pi'(e)}{\pi(e)}.\tag{5.11}$$

This equation looks quite similar to the equations in (5.5): one over the marginal utility is equal to some constant plus a term which depends on the change in probability divided by the probability. This latter expression is also known as the differential form of the likelihood ratio. Note that the right-hand side does not depend on L. This implies that $I(L)-L$ has to be constant. By inspection of the incentive compatibility constraint it is clear that this constant has to be lower than P.

With the additional constraint that $I(L) + P = C(L)$ cannot be negative, this leads to the following result:

> *If the agent can only influence the probability of an accident, then the optimal insurance contract has a deductible:* $C(L) = \max[L - D, 0]$.

This result is very instructive as it shows how incentive contracts operate. In this case, the insured can only influence the probability of loss, not the loss distribution. So to make her work hard, she needs to be punished in case a loss occurs, but rewarded for no loss. This is achieved by giving the insured an income of $W - P$ in the no-loss state, but $W - P - D$ in the loss state, if loss exceeds D. Apart from small losses, where we assumed that the insured cannot pay money back to the insurance company, the insured has the same income independent of the size of the loss. That is, she is fully insured against variations in the size of the loss. There is no reason to distort the payment, e.g., to provide full insurance for low loss levels, and partial insurance for high loss levels, as the insured cannot influence the distribution of losses. However, the insured is only partially insured against the occurrence of a loss, as she has to pay the deductible D herself. This is done to give her an incentive to reduce the probability of a loss.

Such an intuition also holds in a more general context: When designing an incentive contract, one has to be aware which quantities are influenced by the agent, and which are not, and condition the contract only on the former. Turning the argument around, another result can be shown to hold: If something observable and contractible is influenced by the agent's effort, then the contract should condition on this (Holmström, 1982). This so called "Sufficient Statistic Result" is quite strong, as it implies that optimal contracts should condition on possibly very many quantities. For example, in the case of a car accident, the indemnity should condition on the speed of the car, whether the radio was turned on or not, whether the driver was in a phone conversation, etc. as long as these quantities give indications on the preventive effort the agent has taken, and if they can be observed *ex-post.* Partially, this is achieved by the negligence clause, i.e., the insurance company pays less if someone behaved negligently.

The preventive effort a car driver exerts does not only influence the probability but also the severity of an accident. We now turn to this.

Loss-reduction Loss-reduction refers to a situation where by putting in effort the insured can influence the size of the loss. The straightforward way to formalize loss reduction would be to let loss be a function of effort, i.e., $L = L(e)$ with $L'(e) < 0$. However, if this is a deterministic function, even if e is not observable, the first best can be achieved. How?

Determine the first best effort level e^{FB} and premium P^{FB} in a contract with full insurance, i.e., the cover is equal to the size of the loss $C^{FB} = L(e^{FB})$. In the second best world, where effort is not observable, consider the following contract: In case of a loss the insurance company pays the insured the size of the loss if the realized loss is smaller than $L(e^{FB})$. If the realized loss is larger than $L(e^{FB})$ the insured receives nothing. For this service, the insured has to pay P^{FB}. What will the agent do? She will surely not work harder than e^{FB}, as this leads to the same full insurance outcome, but with larger effort costs. If she works less than e^{FB}, in case of a loss she will receive nothing. But this cannot be better than working e^{FB} and being fully insured. So the first best is obtained.[8] We do not know of any real world policy which has such a feature: Full insurance for small losses and zero insurance for large losses. However, insurance policies which are truncated at some point (with a maximum cover) are observed in many cases, most prominently in car insurance.

For modeling purposes this result is unsatisfying as it implies that moral hazard with respect to loss reduction is not a problem. But recall that we assumed that by putting in more effort the insured reduces the size of the loss in a deterministic manner. This picture changes if the decrease in losses has a stochastic element.

To model this, assume that the size of the loss is uncertain and that each effort level determines another distribution function $F(L, e)$ of the loss (with density $f(L, e)$). Larger effort could then be interpreted

[8] The particular contract is an example of a *forcing contract*, which effectively leaves the insured no choice over the effort she chooses.

as leading to a reduction in the expected loss, i.e., $\int_L LdF(L,e_1) < \int_L LdF(L,e_2)$ for $e_1 > e_2$. Larger effort shifts the loss curve such that the mean moves to the left. Also here we have to be very careful. Larger effort must shift the curve in such a way that the support of the loss function stays the same for all effort levels (see Figure 5.1).

By using the "First-Order Approach", the optimization problem becomes:

$$\begin{aligned}
&\max_{e,P,I(L)} \quad (1-\pi)u(W-P) + \pi \int_{\underline{L}}^{\bar{L}} u(W-L+I(L))f(L,e)\mathrm{dL} - c(e) \\
&\qquad\quad \text{s.t.} \\
&\text{PC:} \quad (1-\pi)P - \pi \int_{\underline{L}}^{\bar{L}} I(L)f(L,e)\mathrm{dL} \geq \Pi \qquad (5.12) \\
&\text{IC:} \quad \pi \int_{\underline{L}}^{\bar{L}} u(W-L+I(L))f_e(L,e)\ \mathrm{dL} - c'(e) = 0.
\end{aligned}$$

If for different loss levels the supports were different, then to prevent the agent from engaging in a particular effort one could penalize him

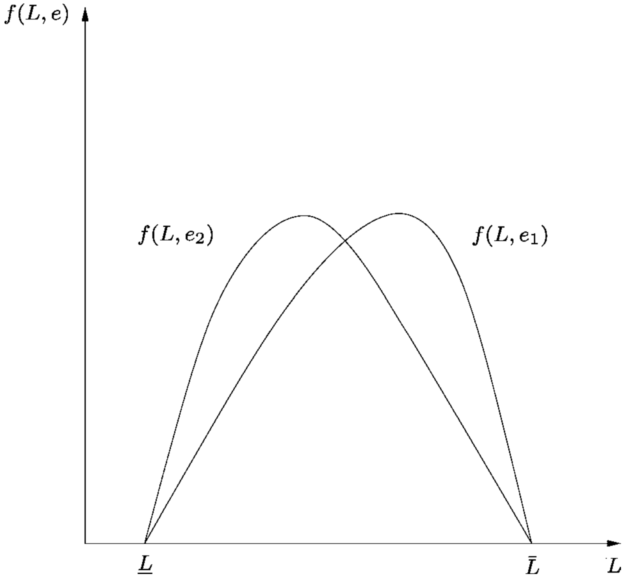

Fig. 5.1 Loss probability density for two different effort levels.

very hard if losses occur which are impossible under the desired effort level. But then, the first best would again be possible.

Check the second-order condition for the agent:

$$\pi \int_{\underline{L}}^{\bar{L}} u(W - L + I(L)) f_{ee}(L, e) \mathrm{dL} - c''(e) < 0. \tag{5.13}$$

Here is where the problem alluded to earlier arises. It is not clear at all, that this expression is smaller than zero for *any* net indemnity function $I(L)$. $f_{ee}(L, e)$ is the second-order derivative of the distribution function with respect to effort, which can well be positive for some effort levels. As mentioned above, there is a large literature around the first order approach, and only under restrictive assumptions can this approach be justified in the present context.[9]

We do, as many have done in the literature before us, just assume that the first-order condition describes the global maximum. But note, that if one wants to use this approach for a given specific problem, once one has found a candidate for an optimal premium-indemnity schedule, one has to check whether the second-order condition indeed holds.

Given the above maximization problem, the first-order condition with respect to $I(L)$ is given by

$$\begin{aligned} &\pi f(L, e) u'(W - L + I(L)) - \lambda \pi f(L, e) \\ &\quad + \mu \pi f_e(L, e) u'(W - L + I(L)) = 0 \end{aligned} \tag{5.14}$$

or, after reformulating:

$$\frac{1}{u'(W - L + I(L))} = \lambda^{-1} + \mu / \lambda \frac{f_e(L, e)}{f(L, e)}. \tag{5.15}$$

Again, an expression quite similar to Eqs. (5.5) and (5.11) above.

What can be said about the form of the optimal contract? The right-hand side of (5.15) depends on the distribution function and on how this distribution changes when effort changes. As effort is presumably hard to measure, the effects of a change in effort on the distribution are

[9] See, e.g., Rogerson (1985a,b) and Jewitt (1988).

even more elusive. With more assumptions on the distribution function some results can be obtained. If the distribution function satisfies the monotone likelihood ratio property (MLRP), which requires that the function $f(L,e_2)/f(L,e_1)$ is decreasing in L if e_2 is larger than e_1, then $-L + I(L)$ will be a decreasing function, i.e., the agent has lower utility for larger losses. MLRP can be interpreted as implying that a larger loss is more probable to have occurred under the lower effort level. In general, distribution functions might or might not satisfy MLRP. But even if MLRP holds, it can still be true that $I(L)$ decreases for larger losses, as MLRP only implies that $-L + I(L)$ increases.

This is a discouraging result. The optimal indemnity schedule can be quite arbitrary, and may not even look like an insurance contract at all. So the predictive power for real world insurance contracts is quite weak. Insurance economists help themselves to some degree by putting more structure on the contract, on the basis of external factors. As argued above, to get the insured to claim a loss, a negative cover can be excluded. In Section 2 it is also argued that overinsurance, i.e., a cover larger than the loss, could be excluded as this would give an incentive to cause the accident. Furthermore, if the insured can manipulate the size of the claim downwards or the size of the loss upwards, then $0 \leq C'(L) \leq 1$ has to hold as well, i.e., it should neither be possible by reducing the claim of the damage to obtain a larger payment, nor by increasing the size of the damage to obtain more money than the increase in loss. If one makes these assumptions, which might be considered sensible but which have nothing to do with the moral hazard problem, then, in addition to MLRP, one obtains an optimal insurance contract with possibly full insurance for low loss levels, partial insurance thereafter and a non-decreasing indemnity schedule.

One might argue that the generality we use for both the utility function and the distribution function are really not necessary. Firms do not know much about $f(L,e)$. They might know $f(L)$ from past experience, however how this distribution function changes for different effort levels is presumably not well-known. Also, the insured has very little knowledge about how her effort influences the loss distribution exactly. It might perhaps be reasonable to assume a simple form for the distribution function. A first attempt would be to model the loss distribution function as a normal distribution, where the agent controls the expected value. That is,

$$f(L,e) = \sqrt{2\pi\sigma^2}^{-1} \exp\left[-\frac{(L-(\bar{L}-e))^2}{2\pi\sigma^2}\right]. \tag{5.16}$$

In that case $f_e(L,e)/f(L,e) = a + bL$, where a and b are constant parameters. This is a very simple expression. If in addition the individual is assumed to have a logarithmic utility function $u(w) = \ln(w)$, and using the result from Eq. (5.15), linear contracts would seem to be the outcome, i.e., insurance contracts specify a percentage rate of losses they cover.

Unfortunately this conclusion is wrong. This can be seen as follows: If, as we assumed, losses are normally distributed, L can take on very large values. With a linear indemnity function, wealth can become negative. However, the logarithm of a negative number is not defined. Somehow it seems that we must have done something wrong with the mathematics. A well-defined model, a correct calculation, but now this caveat. As a matter of fact, this is an example of a problem where the first order approach fails. Mirrlees showed that if the agent can be punished sufficiently hard, the first best can be closely approximated with the help of a trigger contract similar to the one we have discussed above for deterministic loss functions: full insurance for losses below some value L^* and punishment (no insurance) for loss levels above this value. So we should keep in mind that the results presented above only hold if the first order approach is valid.

On the positive side, one result which is worth mentioning is that the insurance market will not break down completely, i.e., it is never optimal to sell no insurance contract at all. This can be seen in the following way: $(P, I(L)) = (0,0)$ would imply a corner solution to the maximization problem (5.12), which yields for the first-order condition with respect to $I(L)$:

$$\pi f(L,e)u'(W-L) - \lambda\pi f(L,e) + \mu\pi f_e(L,e)u'(W-L) \leq 0. \tag{5.17}$$

On the other hand, taking the derivative of the Lagrange function with respect to P at $P = 0$ gives:

$$-(1-\pi)u'(W) + \lambda(1-\pi) \leq 0. \tag{5.18}$$

As $\int_{\underline{L}}^{\bar{L}} f(L,e)dL = 1$ for all effort levels, it follows that $\int_{\underline{L}}^{\bar{L}} f_e(L,e)dL = 0$. So there exist values of L at which $f_e(L,e) > 0$. For those loss levels the inequality (5.17) implies that $u'(W-L) \leq \lambda$, while from the inequality (5.18) it follows that $u'(W) \geq \lambda$ which is a contradiction. So moral hazard creates inefficiencies, but does not lead to a complete market breakdown.

To summarize the results we have obtained in the static model:

- To give the insured an incentive to provide preventive effort, partial insurance is necessary.
- The second best effort level can be lower or larger than the first best effort level.
- Moral hazard leads to inefficiencies, but the market does not break down completely.
- "Sufficient Statistic Result": The optimal contract should not condition on quantities which do not reveal any additional information on the choice of effort by the insured. On the other hand, every (costless) signal whose occurrence does provide information about the effort chosen by the insured, should be part of the optimal incentive scheme.
- In case of loss-prevention, a simple deductible contract is optimal.
- In case of loss-reduction, the optimal contract depends on the exact characteristics of the environment. Under some assumptions on the distribution function and on the feasibility of contracts, the optimal contract is closer to a coinsurance contract than to an insurance policy with a deductible.

We now ask the question whether, as in the case of adverse selection, a dynamic contract which lasts over several periods can give better incentives to the agent and will avoid some of the inefficiencies.

5.3 Dynamic Moral Hazard

5.3.1 Many Periods

The new issues which arise in a multi-period model can best be seen in a simplified two period model, where the moral hazard problem only occurs in the first period. In both periods, the individual can either work hard or be lazy. However, while in period 1 effort is unobservable, we assume that in period 2 the insured will work hard for sure. To keep it simple there are again only two states of the world in each period.

Define the following quantities:

e_i: effort of the insured: $i = 1$: lazy, $i = 2$: hard working.

c_i: utility costs of effort if the insured exercises e_i.

π_i: probability of an accident, given effort e_i.

W_k^1: wealth (or consumption) of the agent in period 1, if no accident ($k = n$) or if an accident occurred ($k = a$).

W_{kj}^2: wealth (or consumption) of the insured in period 2, if in period 1 no (a) accident occurred ($k = n$ or $k = a$, respectively) and if an accident did not or did occur in period 2 ($j = n$ or $j = a$, respectively).

If the individual works hard in period 2, and exercises effort e_i in period 1, her expected utility is given by

$$\begin{aligned}&(1-\pi_i)u(W_n^1) + \pi_i u(W_a^1) - c_i + (1-\pi_i)[(1-\pi_2)u(W_{nn}^2)\\&\quad + \pi_2 u(W_{na}^2) - c_2] + \pi_i[(1-\pi_2)u(W_{an}^2) + \pi_2 u(W_{aa}^2) - c_2]. \qquad (5.19)\end{aligned}$$

It is assumed that the insured does not discount. Discounting will not change the result.

Assume that the insurer would like the agent to work hard in period 1. Otherwise, full insurance in period 1 would be optimal. Under the assumption of a competitive market, where the insurer makes zero profit, the optimal contract is then given by the solution to the following maximization problem:

Maximize (5.19) with $i = 2$, subject to

$$\begin{aligned}
\text{IC:}\quad & (1-\pi_2)[u(W_n^1) + (1-\pi_2)u(W_{nn}^2) + \pi_2 u(W_{na}^2)] \\
& \quad + \pi_2[u(W_a^1) + (1-\pi_2)u(W_{an}^2) + \pi_2 u(W_{aa}^2)] - 2c_2 \\
& \geq (1-\pi_1)[u(W_n^1) + (1-\pi_2)u(W_{nn}^2) + \pi_2 u(W_{na}^2)] \\
& \quad + \pi_1[u(W_a^1) + (1-\pi_2)u(W_{an}^2) + \pi_2 u(W_{aa}^2)] - c_1 - c_2. \\
\text{PC:}\quad & (1-\pi_2)[W_n^1 + (1-\pi_2)W_{nn}^2 + \pi_2 W_{na}^2] \\
& \quad + \pi_2[W_a^1 + (1-\pi_2)W_{an}^2 + \pi_2 W_{aa}^2] \leq \tilde{W}^1 + \tilde{W}^2.
\end{aligned} \tag{5.20}$$

The first constraint is the incentive compatibility constraint: The insured must be better off by working hard in period 1 than by staying lazy. The second constraint is the participation constraint of the insurer. The consumption of the agent must not be larger than her overall wealth in the two periods, which is given by $\tilde{W}^1 = W^1 - \pi_2 L$ plus $\tilde{W}^2 = W^2 - \pi_2 L$. W^i denotes the wealth of the insured in period i, if no accident occurs. $\pi_2 L$ is the fair insurance premium for each period.

Denoting the two Lagrange multipliers by λ and μ, and taking the first-order conditions with respect to $W_n^1, W_{nn}^2, W_{na}^2$ yields:

$$\begin{array}{lr}
W_n^1 & u'(C_n^1)[(1-\pi_2) - \lambda((1-\pi_2) - (1-\pi_1))] - \mu(1-\pi_2) = 0 \\
W_{nn}^2 & u'(C_{nn}^2)[(1-\pi_2)^2 - \lambda((1-\pi_2)^2 - (1-\pi_1)(1-\pi_2))] - \mu(1-\pi_2)^2 = 0 \\
W_{na}^2 & u'(C_{na}^2)[(1-\pi_2)\pi_2 - \lambda((1-\pi_2)\pi_2 - (1-\pi_1)\pi_2)] - \mu(1-\pi_2)\pi_2 = 0.
\end{array} \tag{5.21}$$

And therefore:

$$u'(W_n^1) = u'(W_{nn}^2) = u'(W_{na}^2). \tag{5.22}$$

The agent is fully insured in period 2 if no accident occurred in period 1 ($W_{nn}^2 = W_{na}^2$), and in addition, she consumes as much in period 2 as she does in period 1. By taking the other derivatives it follows that also $W_a^1 = W_{an}^2 = W_{aa}^2$, but with $W_a^1 < W_n^1$.

The intuition for this result is the following: The insured receives less in case of an accident than in case of no accident. This gives her an incentive to work hard in the first period. In the second period, the insured has no influence on the accident probability, so she is fully insured. Her consumption level in the second period is the same as in the first period. This is due to the *income smoothing* effect. The insured prefers a constant income stream to a variable one. With this construction, the incentive is distributed over the two periods: If an accident occurs, the insured is not only worse off in this period, but also in the next period. In other words, the insured still feels the consequences of today's accident tomorrow. On the other hand, this also implies that the deductible the insured has to pay in period 1 is less severe than the corresponding deductible would be in a one-period model. Actually, if the number of periods is increased, and the incentive problem is still only in the first period, the first best will be closer and closer approximated. The difference between the consumption levels following an accident or no accident becomes smaller and smaller. This is the first result:

Many periods allow that incentives for putting in effort today are distributed over time.

By the way, a similar result holds for example for a manager: She might work hard today not just because of this year's bonus payment, but also to increase her chances to become CEO in 10 years time.

Observe that the insured obtains full insurance in period 2. This is also what she would get if she were to buy a short term insurance contract in period 2, as there is no incentive problem at this stage. So one might wonder whether this consumption pattern could not also be achieved with a series of short term contracts. Actually, it can. Write $W_n^1 = W^1 - P^1 - S_n$, $W_{nn}^2 = W_{na}^2 = W^2 - \pi_2 L + S_n$ where P^1 is the premium the agent has to pay in period 1, S_n is the amount of money she saves in period 1, and $\pi_2 L$ is the fair premium in period 2. (Similarly, $W_a^1 = W^1 - P^1 - L + I^1 - S_a$ and so on.) Working backwards we see that two short term contracts will achieve the optimal outcome. In period 2, the insured buys a full insurance contract at the fair premium $\pi_2 L$. If she has no accident in period 1, she has wealth of $W^1 - P^1$. Depending on the size of W^2, the insured will either save or lend money, in order to smooth her income across the two periods. In any case, she will choose S_n such that $W^1 - P^1 - S_n = W^2 - \pi_2 L + S_n$. In the first period this behavior will be anticipated by the insurer, so he sets (P^1, I^1) such that the insured obtains the same final consumption stream as with the optimal long term contract, which also induces her to work hard in period 1. This is the next result:

Although incentives are distributed across periods, this can be achieved with single period contracts.

This result generalizes to the case where the insured can influence her effort also in period 2 and also in further periods. However, in that case one has to assume that the insurer can control the saving of the agent, which was not necessary so far. But as before in period 2 the insured is better off if she had no accident in period 1, and this can again be achieved with short term contracts.[10]

The assumption that the insurer can control the saving of the insured is problematic. If the insurer cannot do so, new problems arise. If savings are not observable, and the insured employs a mixed strategy for her saving behavior, then in the second period there will be a combination of a moral hazard and an adverse selection problem. Adverse selection arises because the insurer does not know the wealth of the insured and with that the risk aversion of the insured. Unfortunately, as shown by Chiappori et al. (1994), the optimal contract which implements any other than the minimum effort level, indeed involves randomized saving. The structure of the optimal contract under these circumstances has not yet been derived. There is however one exception where the savings of the insured do not create a problem. That is the case if the agent's utility function has constant absolute risk aversion, and her effort costs are monetary, i.e., $u(W,e) = -\exp[-r(W - c(e))]$. In that case the amount of savings is irrelevant for the incentive structure and does not create a problem. So a series of short term contracts can be optimal.

To summarize the results: Many periods allow a shifting of the incentives across periods. An individual is worse off in later periods if she has an accident today, i.e., her consumption next year differs according to whether she had an accident this year or not. This is desirable due to the income smoothing effect. With an appropriate combination of savings and insurance contracts, the optimal outcome can be achieved via short term contracts. In light of these results, bonus-malus systems

[10] See Fudenberg et al. (1990) and Malcomson and Spinnewyn (1988).

can be seen as an approximation to the optimal contract structure: Incentives are shifted across periods, but as savings are not controllable by the insurer, long term contracts are used.

5.3.2 Infinitely Many Periods

As in the case of adverse selection, with infinitely many periods the first best can be obtained. There are several ways to see this. One is to use a contract similar to the one we used in the adverse selection case: Pay full insurance as long as the average risk probability is close enough to the one expected under the first best effort level. If not, penalize the agent. This is the procedure used in Rubinstein and Yaari (1983) and Radner (1981). As the argumentation is quite similar to the one outlined in Section 4.5, we will not go into detail here.

Instead we discuss the solution along the lines of Fudenberg et al. (1990). They make the assumption that the agent is not allowed to borrow any money, but she can save. In addition it is assumed that the technology is the same in every period, as well as the wealth of the agent, i.e., without insurance the agent obtains W in every period if no accident occurs, and $W - L$ in case of an accident. The first best utility of the agent would be to consume $W - \pi_2 L$ in every period, which gives her a utility of: $u^* = u(W - \pi_2 L) - c_2$. As before, e_2 is the larger effort level, which would be implemented in the first best.

Fudenberg et al. show the following: There exists a series of short term contracts such that for every $\epsilon > 0$, there exists a discount rate $\delta(\epsilon) < 1$, such that the agent can ensure herself an average utility level $u^* - \epsilon$ for all $\delta > \delta(\epsilon)$.

The interesting thing about this result is the specific series of short term contracts used: Namely no insurance at all. The agent is fully responsible for what she does. The crux of the proof is to find a consumption strategy so that the agent can ensure herself a utility level close to the first best. This is achieved if the agent consumes slightly less than $W - \pi_2 L$ if her savings are large enough, say larger than some $\bar{W}$. If savings fall below this critical level, she consumes $W - L$ in every period and restarts saving. With such a strategy, wealth follows a stochastic process (a submartingale) with a positive drift rate.

From the theory of martingales it is known that eventually wealth will not fall below the critical level with a probability arbitrarily close to one. And if the discount rate is sufficiently small, the "bad years" in the beginning, where the agent accumulates savings, do not count for the overall average utility.

So infinitely many periods are a nice thing to consider theoretically, but the results are not very relevant for actual insurance markets.

5.3.3 Renegotiation

Renegotiation of contracts is a relevant constraint in adverse selection problems. Fortunately, there is much less to worry about if we have a moral hazard problem instead. In the last section, it was argued that the optimal long term contract can be implemented by repetitions of short term contracts. But each short term contract is immune against renegotiation. Renegotiation is only relevant if based on the information the two parties have, there exists a contract which is better for both. But the short term contract is already optimal. Otherwise the insurer and the insured would have chosen a different contract.

There is however an interesting renegotiation problem in a moral hazard situation if renegotiation can occur earlier. Consider the following scenario: A ship owner wants to insure her ship against accidents and drowning. The ship still needs to be built, and a crew has to be found. The insurer gives the owner an insurance contract with a deductible, so that the owner has an incentive to build the ship solidly, and find a good crew to sail with it. So they agree on a 10 year contract, specifying a premium for each year, and a deductible. The owner builds and launches her ship, recruits a crew and then comes back to the insurer arguing: "The insurance contract has a deductible to give me an incentive to work hard on the crew and the ship. So I built my ship properly and found a really good crew. But now, as the ship is on sea, there is nothing I can do about the quality anymore. So why don't we change the contract to a full insurance contract. I do not need the incentives anymore." At this point the logic is compelling and renegotiation of the contract would take place. However, anticipating this renegotiation, it is not clear whether the deductible contract was optimal in the first place.

This problem is discussed by several authors. The basic model used is always the same: The insurer offers a contract to the insured, which the insured can sign. At Stage 2 the insured exercises her effort, and then the contract is renegotiated. Finally, at Stage 3, the outcome occurs and payments are made.

Fudenberg and Tirole (1990) assume that when it comes to renegotiating, the insurer makes a new take-it-or-leave-it offer to the insured. Their main result is that under these assumptions the insured will either put in the lowest effort level or use a mixed strategy over effort levels. Why is that? Assume that the insured exercises any other effort level with probability one in equilibrium. Then at the renegotiation stage the insurer knows the effort and therefore the risk probability. There is no further incentive problem at the renegotiation stage, so indeed the optimal thing to do is to give the agent full insurance. However, this will be anticipated by the insured, so she would optimally shirk by putting in the lowest effort possible. Anticipating full insurance the agent does not work hard. Therefore no equilibrium exists where the insured works hard with probability one. Assume instead that the insured chooses a mixed strategy. Then at the renegotiation stage, the insurer does not know which effort precisely the insured has exercised, while the insured has this information. This is therefore an adverse selection problem, so the insurer will offer a menu of contracts with partial insurance for the low risk type (i.e., high effort type) and full insurance for the high risk type (i.e., low effort type).

Ma (1994) changes a "slight" detail in the setup of the game. He considers the case where the agent makes the renegotiation proposal instead of the principal. Therefore at this stage we have no longer a screening model, but a signaling one, where the agent can with her contract proposal signal her type. In this setup, renegotiation is not a problem, the standard second best contract can be obtained (under specific belief refinements). The reason for this is that when the insured comes to the insurer asking for a full insurance contract, the insurer believes that the insured has put in the lowest effort level, so he only accepts such a renegotiation if the premium is quite high. This makes it unattractive for the insured to even propose a full insurance contract. This in turn implies that the insurer will indeed only observe demands for a full insurance contract out of equilibrium which justifies his beliefs.

Finally, Hermalin and Katz (1991) change another detail of the setup. They allow the principal to observe the effort put in by the agent. Effort is still not verifiable in the sense that it cannot be written into the contract, but the principal can see how much effort was put in. Take the earlier ship owner example: Although it might be difficult to write into a contract that the crew has to be excellent, get on together well, not consume too much drink, etc., it might well be possible that the insurer can judge for himself how good the crew is. In that case, the first best can be achieved: Renegotiation is good news. This works in the following way: The insurer offers an incentive contract which implements the first best effort level. That is, the contract is such that the insured receives maximum utility if she exercises this first best effort level assuming that no renegotiation takes place. This is in general *not* the second best contract. Then, when it comes to renegotiation, the insurer makes a take-it-or-leave-it offer to the insured. For any effort level he observes, he will make a full insurance offer which leaves no additional rent to the insured and this offer will be accepted by the insured. Anticipating this behavior, the insured will put in the optimal effort. Any other effort level gives the insured a lower expected utility level under the original contract. When it comes to renegotiation the insured would then also obtain lower utility, because renegotiation does not increase the utility of the insured.

The difference in this model to the models above is that information is no longer asymmetrically distributed. Both the insured and the insurer can observe the effort, although it can still not be written into a contract. If effort were verifiable, i.e., could be written into a contract, then the first best is easily achieved. If it is only observable but not verifiable, then with an appropriate mechanism a first best can also be obtained (see also Moore and Repullo (1988)). In this case, a simple renegotiation procedure where the insurer makes a take-it-or-leave-it offer is just such a mechanism.

This completes the first part on moral hazard, where the incentive of the insured to prevent losses were analyzed. In the next section, we turn to another moral hazard problem, which is an even stronger headache for the insurance industry: Fraud.

5.4 Insurance Fraud

It should not come as a surprise that precise data on insurance fraud is hard to obtain. It is estimated that fraud in private liability insurance accounts for more than 25% of all claims. The Insurance Information Institute estimates that in the US, property and casualty insurance fraud cost insurers an estimated $30 billion in both 2004 and 2005.

Insurance fraud occurs when the insured makes a false claim which cannot easily be verified. Consider the extreme case: Suppose it is costless for the insured to state false claims while at the same time it is impossible (or infinitely costly) for the insurer to verify that claim. In that case, rational insureds who are not morally constrained have no incentive to report the true size of their loss. Anticipating this, the insurer will not offer a contract with different payments for different loss sizes, as the insured will always claim to have that particular loss with the largest cover. So, if it is at least verifiable that a loss occurred or not (even if the size is non verifiable), in equilibrium the insurer can only offer an insurance contract with a fixed payment independent of the size of the loss. If the occurrence of the loss is not even observable, then the insurance market will break down completely.

To make the problem economically more interesting (and more realistic) the literature has concentrated on two less extreme scenarios. In the first, the so-called "Costly State Verification" approach, the insurer has finite costs to verify a claim. For example the insurer has to send an expert to determine the true size of the claim. In the second, the insured finds it costly to manipulate the claim size, e.g., because false documents are required which are costly to obtain. This is the so-called "Costly State Falsification" approach. In the following sections, we analyze both cases in turn, before we discuss extensions of the models with respect to ethical behavior by the insured, commitment by the insurer, and the role of third parties.

5.4.1 Costly State Verification

The insurance model is similar to the ones analyzed so far. The individual faces a random loss L with distribution $F(L)$ on $[0, \bar{L}]$. The insurer charges a premium P. Now comes the difference. When a loss occurs, the insured reports loss $\tilde{L}$ to the insurer, where $\tilde{L}$ might or might not be equal to the true loss L. By incurring costs of c, the insurer can verify the report. We assume that with probability $\gamma(\tilde{L})$ he will do so.[11] Consequently, there will be two different cover functions: $C(\tilde{L})$ is the cover if the reported loss is $\tilde{L}$ and the insurer does not verify the claim. $C_a(L, \tilde{L})$ is the cover if the reported loss is $\tilde{L}$, the insurer verifies the claim and the loss is of size L. It is obvious that if the insurer detects a false report, he would like to punish the insured for this. In line with the literature we assume that the punishment takes the form of an exogenous payment B, which might be the result of the legal system in place. So the cover function $C_a(L, \tilde{L})$ is either $-B$ in case $\tilde{L} \neq L$ or $C_a(L, L)$ in case $\tilde{L} = L$.

When the loss has occurred the insured decides on whether to make a false statement or not. At this point, the size of the loss is her private information.

Consider first the (Townsend-) case of deterministic auditing. If the size of the report is such that the insurer audits with probability one, then it will be in the insured's interest to file truthfully. If the size of the report is such that the insurer will not audit, then the insured will report that loss which yields the highest cover among all losses in the no-audit range. This cover must be smaller than the cover for audited losses, otherwise only losses in the no-audit range will be reported. A contract which has the same cover for all non-audited losses then leads to the result that the insured has no incentive to lie about her loss.

[11] The initial literature on costly state verification assumed that the verification strategy by the insurer is deterministic. It then turned out that if the loss is below some threshold, the insurer will not examine the loss further, while losses above that threshold are verified. The insurance contract was then a deductible contract, where small losses were not covered and thus did not need verifying, while large losses have marginally full insurance (Townsend, 1979). Mookherjee and Png (1989), however, show that random auditing is welfare improving.

Concentrating on contracts without fraud also in the case of random auditing, the expected utility of the insured is given by

$$EU = \int_0^{\bar{L}} [(1-\gamma(L))u(W-P-L+C(L)) \\ +\gamma(L)u(W-P-L+C_a(L,L))]dF(L). \quad (5.23)$$

Assuming a competitive insurance market, the insurer has at least to break even with the contract, taking into account the costs for auditing the claims. Thus we have as the participation constraint:

$$\text{PC}: \; P - \int_0^{\bar{L}} [(1-\gamma(L))C(L) + \gamma(L)(C_a(L,L)+c)]dF(L) \geq 0. \quad (5.24)$$

The important new constraint, the incentive compatibility constraint, assures that it is not in the interest of the insured to make fraudulent claims. This must be true for each loss size L:

$$\text{IC}: \quad (1-\gamma(L))u(W-P-L+C(L)) + \gamma(L)u(W-P-L \\ +C_a(L,L)) \geq (1-\gamma(\tilde{L}))u(W-P-L+C(\tilde{L})) \\ +\gamma(\tilde{L})u(W-P-L-B) \;\; \forall \tilde{L} \in [0,\bar{L}]. \quad (5.25)$$

From inequality (5.25) it immediately follows that no loss needs to be audited with probability one to ensure truth telling.

While in the case of deterministic auditing the overall contract is relatively easy to determine, random auditing is more complicated. We follow Fagart and Picard (1999) and assume that the insured have constant absolute risk aversion (CARA): $u(W) = -e^{-\alpha W}$. CARA makes life easier as it implies that the incentive to make fraudulent claims does not depend on the size of the loss, because the factor $-e^{-\alpha(W-P-L)}$ cancels out in inequality (5.25). Also from (5.25) it follows that in order to establish truth telling, it is sufficient to show that the insured has no incentive to lie in case her true cover is minimal. This is usually the case at zero loss, where $C(0) = C(0,0) = 0$. By replacing $\tilde{L}$ by L the incentive compatibility constraint then reads:

$$u(W-P) = (1-\gamma(L))u(W-P+C(L)) + \gamma(L)u(W-P-B) \quad (5.26)$$

and therefore

$$\gamma(L) = \frac{u(W - P + C(L)) - u(W - P)}{u(W - P + C(L)) - u(W - P - B)} \in [0,1). \tag{5.27}$$

From Eq. (5.27) it follows that with an increase in cover $C(L)$ there will be an increase in auditing probability $\gamma(L)$.

We do not derive the full contract here (see Fagart and Picard, 1999, for details), but display the result in Figure 5.2. The top part shows the optimal auditing probability which for very small losses is zero, and is then increasing in the size of the loss. The optimal cover, which is shown in the lower part of Figure 5.2, has several interesting features which can be understood from our analysis so far. First, the optimal contract has a deductible. For very small losses it is not optimal to pay cover as this implies incurring the cost of auditing with some probability. For losses larger than the deductible there will be two cover functions: One when the loss report is not audited and a second, larger one, when the report is audited. The latter one displays marginal full insurance. This follows as the cover function $C_a(L, L)$ only enters expected utility (5.23)

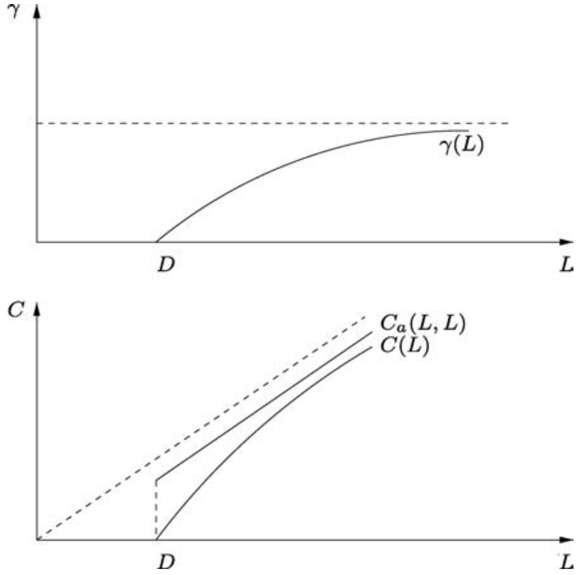

Fig. 5.2 Insurance contract under costly state verification.

and expected profit (5.24) but not the incentive constraint (5.27). Thus optimally this function is chosen such that the risk remains with the insurer. If there is no audit, the cover $C(L)$ is smaller than $C_a(L, L)$. This follows as an increase in $C(L)$ does not only imply more insurance, but also leads to an increase in $\gamma(L)$ and thus in higher expected costs for the insurer. As in the Raviv model of Section 3, these transaction costs have to be borne by the insured in the form of a lower coverage.

Let us add one remark on the assumption of a CARA utility function. If the insured has for example decreasing absolute risk aversion (DARA), then she is less risk averse if the loss is small. This implies that her incentive to make fraudulent claims increases further for small loss sizes. In that case it might be optimal to also have a positive cover for small losses, which in turn reduces the incentive to make a false claim. For further details, see Picard (2000).

There are several features in this model which are worth analyzing a bit further: Are people really just rational optimizers or do they abstain from fraudulent behavior because of moral concerns? Why should the insurer in such a framework audit the insured, if, after all, there will only be correct loss reports? We discuss these and further topics in the final subsection of this section.

5.4.2 Costly State Falsification

Under costly state falsification, the insurer has no means to audit the claim, while the insured finds it costly to manipulate the claim. As before, the insured faces a random loss L with distribution $F(L)$ on $[0, \bar{L}]$. Now, if a loss L occurs, and the insured makes a (possibly) false claim $\tilde{L} \geq L$, then she faces costs of $g(\tilde{L} - L)$.[12] These costs satisfy $g(0) = 0$, $g' > 0$, and $g'' > 0$. While Lacker and Weinberg (1989) assume that $g'(0) = \delta > 0$, Crocker and Morgan (1998) work with $g'(0) = 0$, such that a small manipulation of the claim is nearly costless.

As before, the insured pays a premium P and obtains cover, which again depends on the reported loss $C(\tilde{L})$. If the insured with loss L

[12] The more general analysis assumes $\tilde{L} \neq L$ and costs $g(|\tilde{L} - L|)$. See Crocker and Morgan (1998) for details.

chooses to report $\tilde{L}(L)$, then her expected utility can be written as

$$EU = \int_0^{\bar{L}} u(W - P - L + C(\tilde{L}(L)) - g(\tilde{L} - L))dF(L). \qquad (5.28)$$

The expected profit by the insurer has to be positive:

$$\text{PC:} \;\; P - \int_0^{\bar{L}} C(\tilde{L}(L))dF(L) \geq 0. \qquad (5.29)$$

Finally, it must be optimal for the insured to report $\tilde{L}$ when her loss is L. Assuming that the optimal $C(\tilde{L}(L))$ is differentiable, it thus follows that:

$$\begin{aligned} \text{IC:} \;\; & u'(W - P - L + C(\tilde{L}(L)) - g(\tilde{L} - L))) \\ & \times [C'(\tilde{L}(L)) - g'(\tilde{L} - L)] = 0. \end{aligned} \qquad (5.30)$$

The term in square brackets states that the marginal cover (as a function of the reported loss) is equal to the marginal costs of misstating the claim. Or in other words, the insured will inflate her claim until the marginal costs of doing so are equal to the marginal increase in cover. It is quite realistic to assume that $g' < 1$, i.e., increasing the claim by one unit costs less than this unit. In this case, from (5.30) it immediately follows that the optimal contract features partial insurance (as a function of the reported loss).

From Eq. (5.30) it also becomes clear that it makes a difference whether $g'(0) = 0$ or $g'(0) = \delta > 0$. In the latter case (as analysed by Lacker and Weinberg (1989)), the optimal contract has the form $C(\tilde{L}) = C_0 + \delta\tilde{L}$ if δ is sufficiently large. Inflating the claim by one unit will cost the insurer the amount δ, and will lead to an increase in cover of the amount δ. Thus it is optimal for the insured to report the correct loss size. If however $g'(0) = 0$, fraudulent claims can only be avoided by paying a fixed cover independent of the size of the claim. If one would like to have an increase in cover for larger claims, then some fraud cannot be avoided.

Coming back to the case with $g'(0) = \delta$ where the optimal contract has the form $C(L) = C_0 + \delta L$. Interestingly, the constant payment C_0 is larger than zero, which implies overinsurance for small losses. The reason for this is that with partial marginal insurance (i.e., $C'(L) = \delta < 1$), high losses tend to be underinsured. The degree of underinsurance would become larger if small losses were not covered. So by adding a constant payment, this underinsurance is mitigated (at the cost of having overinsurance for small losses).

To solve for the optimal contract in the general case, the revelation principle (Myerson, 1979) is used. This principle is a technical insight which shows that it suffices to concentrate on allocations which are functions of the types.[13] In the present context, an allocation consists of two parts: A report and a cover. Thus, for any loss level L we have to find a report $\tilde{L}(L)$ and a cover $C(\tilde{L}(L)) = C(L)$. The optimal contract is then chosen such that it maximizes the utility of the insured by considering the zero profit constraint and a "truthtelling" constraint. The latter constraint implies that in case the true loss is L, the utility of the insured is indeed maximized by the cover $C(L)$ and misreporting costs $g(\tilde{L}(L) - L))$, where the alternative would be to mimic a person with loss $L' \neq L$, which would result in cover $C(L') = C(\tilde{L}(L'))$ and misreporting costs $g(\tilde{L}(L') - L)$.

While we do not go through the full analysis (see Crocker and Morgan (1998) for details), we display the final contract in Figure 5.3. It has all the elements which we discussed before: Overinsurance for small loss reports, underinsurance for large loss reports. And finally partial marginal insurance, the exact form of which depends on the manipulating cost function g.

5.4.3 Extensions

The costly state verification and falsification models serve as the basis for the literature on insurance fraud. Several extensions to these models exist.

So far it is assumed that the rational insured will take every possibility to misstate her claim as long as this is profitable for her. However,

[13] Formally, message games are considered where the agent first reports her type (in this case the loss level), and then an allocation based on the report is implemented. The revelation principle states that it is sufficient to consider these kind of "direct mechanisms" where the agent truthfully reports her type.

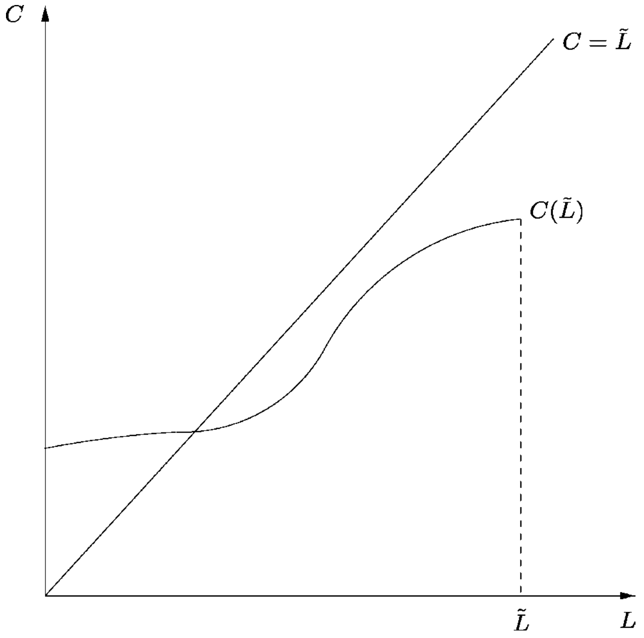

Fig. 5.3 Insurance contract under costly state falsification.

in reality, many people might abstain from this behavior for *moral reasons* (Tennyson, 1997). If the insurer cannot distinguish between honest and “opportunistic” individuals, there are two possible outcomes. The contract may be such that the dishonest individuals are deterred from misstating their claims. This however implies incurring auditing costs and having inefficient partial insurance. The alternative is, and this is the equilibrium outcome if there are sufficiently many honest people around, to give full insurance contracts without auditing, where the few dishonest people defraud the system (Picard, 1996). It is quite obvious that in this case the premium for the full insurance contract cannot be fair, as the honest insured have to pay for the inflated claims made by the opportunists.

In the case of costly state verification the optimal contract prevents insurance fraud completely. However, this implies that the insurer should not find it in his interest to audit the claims. If the insurer can commit to this auditing strategy, e.g., by establishing a national insurance fraud detection office or something similar, then such a contract can work. If *commitment is not possible*, however, one has to analyze a sequential setup where the insurer has to decide after the contract is written and the agent has made her report whether to audit or not. It turns out that in this case, insurance fraud cannot be completely avoided and auditing takes place with positive probability. That is clear: If there were no fraud, there would be no incentive to audit. If there were no audit, the insured would have a strong incentive to make fraudulent claims. Interestingly the optimal contract might even have overinsurance for large losses. This overinsurance will give the insurer a larger incentive to audit high loss claims, as there is more at stake (see Picard (1996) and Khalil (1997) for details).

While in the basic model either the insured or the insurer have costs to manipulate or audit a claim, in Bond and Crocker (1997) *both parties have to incur costs*: The insurer incurs costs in auditing the claim, while the insured can by exerting effort manipulate the costs of the insurer, i.e., she can make it more or less easy to audit the claim. The optimal insurance contract is then such that the insured has no or little incentive to influence the auditing costs. This is achieved by overcompensating losses which are easy to audit, while those which are hard to audit are underinsured.

In many situations of insurance fraud a *third party* is involved. This might for example be the person forging the documents. There are two directions the literature on third party fraud has pursued. First, starting with the work by Tirole (1986) and applied to the insurance framework by Brundin and Salanié (1997) collusion between the insured and a third party is analyzed. For example, the insured has to collude with a mechanic to write a large invoice. The second branch of the literature analyzes the incentives of the third party to commit fraud without the knowledge of the insured, i.e., when the damaged car is given to a garage the mechanic might try to inflate the bill. For these models it is necessary that the insured is not aware of the true size of the loss. In that case we would call the repair work a credence good (Wolinsky, 1993).

When the case where the insured colludes with the third party, say a mechanic, the model is quite similar to the costly state falsification model. The insured has to bribe the mechanic to make false statements, which implies that it is costly for the insured to manipulate the claim. In the simplest model, the insured and the mechanic bargain over the bribe the insured has to pay to the mechanic. In a typical Nash bargaining situation this bribe will then depend on the amount of money to be gained through misreporting, which itself depends on the slope of the cover function. If the insured and the mechanic split the gain by half, then in the formulation of the costly state falsification model the cost function g can be written as: $g(\tilde{L}, L) = \frac{1}{2}(C(\tilde{L}) - C(L))$, i.e., half of the gains of misstating the claim go to the mechanic and are therefore costs for the insured. Note that in contrast to the costly state falsification model this cost function depends on the cover, which itself is endogenously chosen by the insurer.

In the case where the third party, e.g., a mechanic, commits the fraud without the knowledge of the insured, the resulting outcome depends on the specifics of the market (for an overview see Dulleck and Kerschbamer (2006)). As the insured does not know the exact damage she has and will not find out after the damage is repaired, she can only detect wrong behavior by obtaining a second opinion. Wolinsky (1993) shows that as a response to this problem a market for only minor repairs might emerge. If a mechanic (or similarly a physician in case of health insurance) who has been given the car to be repaired, claims that he cannot do the job and refers the insured to a more specialized garage, then the insured can have some confidence in assuming that this diagnosis is correct, as the mechanic himself does not profit from it (at least as long as he is not linked to the specialized garage).[14] Intuitively one would expect that in such a framework a partial insurance contract would make the insured behave more sensitively toward controlling the third party. This is indeed the case, because a person with full insurance has no incentive to control costs. However, as Sülzle and Wambach (2005) show, a marginal increase in partial insurance might

[14] In Emons (1997) the mechanic might face capacity constraints. If the prices in the market are set correctly (which they will in a competitive market), then the mechanic just does not find it profitable to do too much repair work, as his order book is filled already.

lead to a situation where the mechanic behaves even more fraudulently. Less insurance cover makes the insured more cost sensitive so she asks more often for second opinions. On the other hand, the mechanics can expect more customers who are already on their second visit — so giving them a false diagnosis will just confirm their first (wrong) diagnosis and will then be accepted by the insured.

References

Akerlof, G. A. (1970), 'The market for lemons: Quality uncertainty in the market mechanism'. *Quarterly Journal of Economics* **84**, 488–500.

Ania, A., T. Tröger, and A. Wambach (2002), 'An evolutionary analysis of insurance markets with adverse selection'. *Games and Economic Behavior* **40**, 153–184.

Arnott, R. and J. E. Stiglitz (1988a), 'The basic analytics of moral hazard'. *Scandinavian Journal of Economics* **90**, 383–413.

Arnott, R. and J. E. Stiglitz (1988b), 'Randomization with asymmetric information'. *RAND Journal of Economics* **19**, 344–362.

Arrow, K. J. (1970), *Essays in the Theory of Risk-Bearing.* Amsterdam-London: North-Holland.

Arrow, K. J. (1974), 'Optimal insurance and generalized deductibles'. *Scandinavian Actuarial Journal* **1**, 1–42.

Asheim, G. B. and T. Nilssen (1996), 'Non-discriminating renegotiation in a competitive insurance market'. *European Economic Review* **10**, 1717–1736.

Bond, E. W. and K. J. Crocker (1997), 'Hardball and the soft touch: The economics of optimal insurance contracts with costly state verification and endogenous monitoring costs'. *Journal of Public Economics* **63**, 239–264.

Borch, K. (1962), 'Equilibrium in a reinsurance market'. *Econometrica* **30**, 424–444.

Borch, K. (1981), 'Is regulation of insurance companies necessary?'. In: H. Göppl and R. Henn (eds.): *Geld, Banken und Versicherungen,* Vol. 2. Königstein, pp. 717–731.

Brockett, P. L., R. MacMinn, and M. Carter (2000), 'Genetic testing, insurance economics and societal responsibility'. *North American Actuarial Journal* **3**(1), 1–20.

Brundin, I. and F. Salanié (1997), *Fraud in the Insurance Industry: An Organizational Approach.* Department of Economics, University of Toulouse, Mimeo.

Brys, E., G. Dionne, and L. Eeckhoudt (1989), 'More on insurance as a Giffen good'. *Journal of Risk and Uncertainty* **2**, 415–420.

Chiappori, P. A., B. Jullien, B. Salanié, and F. Salanié (2006), 'Asymmetric information in insurance: General testable implications'. *RAND Journal of Economics* **37**(4), 783–798.

Chiappori, P. A., I. Macho, P. Rey, and B. Salanié (1994), 'Repeated moral hazard: The role of memory, commitment and the access to credit markets'. *European Economic Review* **38**, 1527–1553.

Chiappori, P. A. and B. Salanié (2000), 'Testing for asymmetric information in insurance markets'. *Journal of Political Economy* **108**, 56–78.

Cho, I. and D. Kreps (1987), 'Signalling games and stable equilibria'. *Quarterly Journal of Economics* **52**, 179–222.

Cook, P. J. and D. A. Graham (1977), 'The Demand for insurance and protection: The case of irreplacable commodities'. *Quarterly Journal of Economics* **91**, 143–156.

Cooper, R. and B. Hayes (1987), 'Multi-period insurance contracts'. *International Journal of Industrial Organization* **5**, 211–231.

Crocker, K. J. and J. Morgan (1998), 'Is honesty the best policy? Curtailing insurance fraud through optimal incentive contracts'. *Journal of Political Economy* **106**(2), 355–375.

Crocker, K. J. and A. Snow (1985), 'The efficiency of competitive equilibria in insurance markets with asymmetric information'. *Journal of Public Economics* **26**, 207–219.

Crocker, K. J. and A. Snow (1986), 'The efficiency effects of categorical discrimination in the insurance industry'. *Journal of Political Economy* **94**, 321–344.

Crocker, K. J. and A. Snow (1992), 'The social value of hidden information in adverse selection economies'. *Journal of Public Economics* **48**, 317–347.

Cummins, D. (1991), 'Statistical and financial models of insurance pricing and the insurance firm'. *Journal of Risk and Insurance* **58**, 261–302.

D'Arcy, S. P. and N. Doherty (1990), 'Adverse selection, private information and lowballing in insurance markets'. *Journal of Business* **63**, 145–164.

Dasgupta, P. and E. Maskin (1986), 'The existence of equilibrium in discontinuous economic games, II: Applications'. *Review of Economic Studies* **53**, 27–41.

Dionne, G. (1983), 'Adverse selection and repeated insurance contracts'. *Geneva Papers on Risk and Insurance* **8**, 316–332.

Dionne, G. (2000), *Handbook of Insurance.* Kluwer.

Dionne, G. and N. A. Doherty (1994), 'Adverse selection, commitment, and renegotiation: Extension to and evidence from insurance markets'. *Journal of Political Economy* **102**, 209–235.

Doherty, N. A. and H. Schlesinger (1990), 'Rational insurance purchasing: Consideration of contract nonperformance'. *Quarterly Journal of Economics* **105**, 143–153.

Doherty, N. A. and P. D. Thistle (1996), 'Adverse selection with endogenous information in insurance markets'. *Journal of Public Economics* **63**, 83–102.

Dulleck, U. and R. Kerschbamer (2006), 'On doctors, mechanics and computer specialists — The economics of credence goods'. *Journal of Economic Literature* **44**, 5–42.

Eeckhoudt, L. and C. Gollier (2000), 'The effects of changes in risk on risk taking: A survey'. In: G. Dionne (ed.): *Handbook of Insurance.* Kluwer, pp. 117–130.

Eeckhoudt, L., C. Gollier, and H. Schlesinger (2005), *Economic and Financial Decisions under Risk*. Princeton University Press.

Ehrlich, I. and G. Becker (1972), 'Market insurance, self insurance and self protection'. *Journal of Political Economy* **80**, 623–648.

Emons, W. (1997), 'Credence goods and fraudulent experts'. *RAND Journal of Economics* **28**, 107–119.

Fagart, M.-C. and P. Picard (1999), 'Optimal insurance under random auditing'. *Geneva Papers on Risk and Insurance Theory* **24**, 29–54.

Finsinger, J. and M. Pauly (1984), 'Reserve levels and reserve requirements for profit maximising insurance firms'. In: G. Bamberg and K. Spremann (eds.): *Risk and Capital*. Berlin: Springer.

Fudenberg, D., B. Holmström, and P. Milgrom (1990), 'Short term contracts and long term agency relationships'. *Journal of Economic Theory* **51**, 1–31.

Fudenberg, D. and J. Tirole (1990), 'Moral hazard and renegotiation in agency contracts'. *Econometrica* **58**, 1279–1319.

Gollier, C. (2001), *The Economics of Risk and Time*. The MIT Press.

Gollier, C. and H. Schlesinger (1996), 'Arrow's theorem on the optimality of deductibles: A stochastic dominance approach'. *Economic Theory* **7**, 359–363.

Gravelle, H. and R. Rees (2004), *Microeconomics*. Prentice Hall, Third edition.

Grossman, H. (1979), 'Adverse selection, dissembling and competitive equilibrium'. *Bell Journal of Economics* **10**, 330–343.

Grossman, S. and O. Hart (1983), 'An analysis of the principal-agent problem'. *Econometrica* **51**, 7–45.

Harris, M. and R. M. Townsend (1981), 'Resource allocation under asymmetric information'. *Econometrica* **49**(1), 33–64.

Hellwig, M. F. (1987), 'Some recent developments in the theory of competition in markets with adverse selection'. *European Economic Review* **31**, 319–325.

Hellwig, M. F. (1988), 'A note on the specification of interfirm communication in insurance markets with adverse selection'. *Journal of Economic Theory* **46**, 154–163.

Hermalin, B. and M. Katz (1991), 'Moral hazard and verifiability'. *Econometrica* **59**, 1735–1754.

Hirshleifer, J. (1971), ‘The private and social value of information and the reward to inventive activity’. *American Economic Review* **61**(4), 561–574.

Holmström, B. (1979), ‘Moral hazard and observability’. *Bell Journal of Economics* **10**, 74–91.

Holmström, B. (1982), ‘Moral hazard in teams’. *Bell Journal of Economics* **13**, 324–340.

Hoy, M. and M. Polborn (2000), ‘The value of genetic information in the life insurance market’. *Journal of Public Economics* **78**(3), 235–252.

Hoy, M. and A. Robson (1981), ‘Insurance as a giffen good’. *Economics Letters* **8**, 47–51.

Hoy, M. and M. Ruse (2005), ‘Regulating genetic information in insurance markets’. *Risk Management & Insurance Review* **8**(2), 211–237.

Inderst, R. and A. Wambach (2001), ‘Competitive insurance markets under adverse selection and capacity constraints’. *European Economic Review* **45**, 1981–1992.

Jaynes, G. D. (1978), ‘Equilibria in monopolistically competitive insurance markets’. *Journal of Economic Theory* **19**, 394–422.

Jewitt, I. (1988), ‘Justifying the first-order approach to principal-agent problems’. *Econometrica* **57**, 1177–1190.

Khalil, F. C. (1997), ‘Auditing without commitment’. *RAND Journal of Economics* **28**, 629–640.

Kunreuther, H. and M. Pauly (1985), ‘Market equilibrium with private knowledge: An insurance example’. *Journal of Public Economics* **26**, 269–288.

Lacker, J. M. and J. A. Weinberg (1989), ‘Optimal contracts under costly state falsification’. *Journal of Political Economy* **97**(6), 1345–1363.

Ma, C. T. A. (1994), ‘Renegotiation and optimality in agency contracts’. *Review of Economic Studies* **61**, 109–129.

Magill, M. and M. Quinzii (1996), *Theory of Incomplete Markets,* Vol. 1. MIT Press.

Malcomson, J. and F. Spinnewyn (1988), ‘The multiperiod principal-agent problem’. *Review of Economic Studies* **55**, 391–408.

Mas-Colell, A., M. D. Whinston, and J. R. Green (1996), *Microeconomic Theory.* Oxford University Press.

Milgrom, P. and N. Stokey (1982), ‘Information, trade and common knowledge’. *Journal of Economic Theory* **26**, 17–27.

Mirrlees, J. A. (1971), ‘An exploration in the theory of optimal income taxation’. *Review of Economic Studies* **38**(2), 175–208.

Miyazaki, H. (1977), ‘The rat race and internal labour markets’. *Bell Journal of Economics* **8**, 394–418.

Mookherjee, D. and I. Png (1989), ‘Optimal auditing, insurance and redistribution’. *Quarterly Journal of Economics* **104**, 399–415.

Moore, J. and R. Repullo (1988), ‘Subgame perfect implementation’. *Econometrica* **56**, 1191–1220.

Mossin, J. (1968), ‘Aspects of rational insurance purchasing’. *Journal of Political Economy* **79**, 553–568.

Munch, P. and D. Smallwood (1981), ‘Theory of solvency regulation in the property and casualty industry’. In: G. Fromm (ed.): *Studies in Public Regulation.* Cambridge: MIT Press, pp. 119–180.

Myerson, R. B. (1979), ‘Incentive compatibility and the bargaining problem’. *Econometrica* **47**, 61–73.

Peltzman, S. (1975), ‘The effects of automobile saftety regulation’. *Journal of Political Economy* **83**, 677–726.

Picard, P. (1996), ‘Auditing claims in the insurance market with fraud: The credibility issue’. *Journal of Public Economics* **63**, 27–56.

Picard, P. (2000), ‘Economic analysis of insurance fraud’. In: G. Dionne (ed.): *Handbook of Insurance Economics.* Boston, pp. 315–362.

Radner, R. (1981), ‘Monitoring cooperative agreements in a repeated principal-agent relationship’. *Econometrica* **49**(5), 1127–1148.

Raviv, A. (1979), ‘The design of an optimal insurance policy’. *American Economic Review* **69**, 84–96.

Rees, R., H. Gravelle, and A. Wambach (1999), ‘Regulation of insurance markets’. *Geneva Papers on Risk and Insurance Theory* **24**, 55–68.

Riley, J. G. (1979), ‘Informational equilibrium’. *Econometrica* **47**, 331–359.

Rogerson, W. (1985a), ‘The first-order approach to principal-agent problems’. *Econometrica* **53**, 1357–1368.

Rogerson, W. (1985b), 'Repeated moral hazard'. *Econometrica* **53**, 69–76.

Rosenthal, R. W. and A. Weiss (1984), 'Mixed strategy equilibrium in a market with adverse selection'. *Review of Economic Studies* **51**, 333–342.

Rothschild, M. and J. Stiglitz (1976), 'Equilibrium in competitive insurance markets: An essay on the economics of imperfect information'. *Quarterly Journal of Economics* **80**, 629–649.

Rubinstein, A. and M. E. Yaari (1983), 'Repeated insurance contracts and moral hazard'. *Journal of Economic Theory* **30**, 74–97.

Schlesinger, H. (1984), 'Optimal insurance for irreplacable commodities'. *Journal of Risk and Insurance* **51**, 131–137.

Schlesinger, H. and J. M. Graf v. d. Schulenburg (1987), 'Risk aversion and the purchase of risky insurance'. *Journal of Economics* **47**, 309–314.

Smart, M. (2000), 'Competitive insurance markets with two unobservables'. *International Economic Review* **41**(1), 153–169.

Spence, M. (1978), 'Product differentiation and performance in insurance markets'. *Journal of Public Economics* **10**, 427–447.

Stiglitz, J. (1977), 'Monopoly, non-linear pricing and imperfect information: The insurance market'. *Review of Economic Studies* **44**, 407–430.

Strohmenger, R. and A. Wambach (2000), 'Adverse selection and categorical discrimination in the health insurance markets: The effects of genetic tests'. *Journal of Health Economics* **19**(2), 197–218.

Sülzle, K. and A. Wambach (2005), 'Insurance in a market for credence goods'. *Journal of Risk and Insurance* **72**, 159–176.

Tabarrok, A. (1994), 'Genetic testing: An economic and contractarian analysis'. *Journal of Health Economics* **13**, 75–91.

Tennyson, S. (1997), 'Economic institutions and individual ethics: A study of consumer attitudes toward insurance fraud'. *Journal of Economic Behaviour and Organization* **32**, 247–265.

Tirole, J. (1986), 'Hierarchies and bureaucracies: On the role of collusion in organizations'. *Journal of Law, Economics and Organization* **2**, 181–214.

Townsend, R. M. (1979), 'Optimal contracts and competitive markets with costly state verification'. *Journal of Economic Theory* **21**, 265–293.

Villeneuve, B. (2003), 'Concurrence et antisélection multidimensionnelle en assurance'. *Annales d'économie et de statistique* **69**, 119–142.

Wambach, A. (2000), 'Introducing heterogeneity in the Rothschild–Stiglitz model'. *Journal of Risk and Insurance* **67**, 579–591.

Wilson, C. (1977), 'A model of insurance markets with incomplete information'. *Journal of Economic Theory* **16**, 167–207.

Wolinsky, A. (1993), 'Competition in the market for informed expert services'. *RAND Journal of Economics* **24**, 380–398.

Zweifel, P., F. Breyer, and M. Kifmann (2007), *Health Economics.* Springer.