

You are what you bet: eliciting risk attitudes from horse races

Still preliminary

Pierre-André Chiappori* Amit Gandhi†
Bernard Salanié‡ François Salanié§

August 29, 2008

Abstract

As a textbook model for contingent markets, horse races have provided an interesting way to study the attitude towards risk of bettors. We innovate on the literature by explicitly considering heterogeneous bettors, with different utility functions (possibly non-expected utility). Based on a simple single-crossing condition, we derive testable predictions; and we show that parimutuel data allow to uniquely identify the distribution of preferences among the population of bettors. We then estimate the model on data from US races. Preliminary empirical results illustrate how well families of preferences towards risk and/or uncertainty can explain the behavior of bettors.

*Columbia University.

†University of Wisconsin-Madison.

‡Columbia University. e-mail: bsalanie@columbia.edu

§Toulouse School of Economics (Lerna/Inra).

1 Introduction [to be completed]

Much empirical work in macroeconomics or finance assumes that attitudes toward risk are the same for all agents. This homogeneity assumption may receive several justifications, beyond the search for tractability. Since attitudes toward risk are not directly observable, financial exchanges may be more convincingly explained by differences in endowments or wealth. One may also refer to results showing the existence of a representative agent (but such an agent only exists under restrictive conditions, especially when uncertainty is introduced.)

Yet there is mounting experimental evidence that risk attitudes are massively heterogeneous. To quote but a few examples, Barsky et al. (1997) use survey questions and observations of actual behaviour to measure relative risk aversion. Results indicate that this parameter varies between 2 (for the first decile) and 25 (for the last decile), and that this heterogeneity is poorly explained by demographic variables. Guiso and Paiella (2003) report similar findings, and use the term “massive unexplained heterogeneity”. Finally Chiappori and Paiella (2007) observe the financial choices of a sample of households across time, and use these panel data to show that a model with constant relative risk aversion model well explains each household’s choices, although this coefficient is highly variable across households (its mean is equal to 4.2, for a median equal to 1.7).

These papers and others indicate that heterogeneous attitudes towards risk and uncertainty may be crucial to understand patterns on insurance and financial markets. Still the difficulty we alluded to above remains: to elicit different risk attitudes from the observation of choices, it seems that one needs rich data on individual behaviors. This paper adopts a different strategy. We focus on the case of horse races, which provide a textbook example for financial markets, and for which large samples are available. The races we focus on use parimutuel betting; that is, those lucky bettors that have bet on the winning horse share the total amount wagered in the race. Hence observing the odds of a horse is equivalent to observing its market share. How these market shares vary from race to race as a function of odds thus conveys some information on how bettors react to odds, and finally on their attitudes towards risk. Our main theoretical result establishes that under some conditions one can identify the entire distribution of preferences in the population of bettors.

2 Theoretical framework

Let us first explain how parimutuel betting is organized. Consider a race with $i = 1 \dots n$ horses. The simplest bets are “win bets”, bets on the winning horse: each dollar bet on horse i pays a net return R_i if horse i wins, and is lost if it loses. R_i is called the odds of horse i , and in parimutuel races it is determined by the following rule: all money wagered by bettors constitutes a pool that is redistributed to those that bet on the winning horse, apart from a share t corresponding to taxes and a house “take”. Accordingly, if s_i is the share of the pool corresponding to the sums wagered on horse i , then the payment to a winning bet of \$1 is

$$R_i + 1 = \frac{1 - t}{s_i} \quad (1)$$

Hence odds are not set by bookmakers; instead they are determined by the distribution (s_1, \dots, s_n) of bets among horses. The odds are low for those horses on which many bettors laid money (favorites), and they are high for outsiders (often called longshots)¹. Note that because $\sum s_i = 1$ these equations together imply

$$\frac{1}{1 - t} = \sum_i \frac{1}{R_i + 1} \quad (2)$$

Hence knowing the odds (R_1, \dots, R_n) allows to compute the take t , and thus the shares in the pool (s_1, \dots, s_n) .

We now define a n -horse *race* (\mathbf{p}, t) by a vector of positive probabilities $\mathbf{p} = (p_1, \dots, p_n)$ in the n -dimensional simplex, and a take $t \in]0, 1[$.

Here p_i is the objective probability that horse i wins the race; thus we start from the very strong assumption that all bettors agree on the probabilities, and these probabilities are correct. This is a useful starting point as it allows us to focus on heterogeneity in preferences; yet it clearly needs to be relaxed. This can be done in a number of ways, some of which in fact fit our framework very nicely. We will come back to this point as we go along.

Given homogeneous and correct beliefs, any bettor looks on a bet of one dollar on horse i as a lottery that returns R_i with probability p_i , and (-1) with probability $(1 - p_i)$. We denote this lottery by (R_i, p_i) , and call it a *gamble*. By convention, throughout the paper we index horses by decreasing probabilities ($p_1 > \dots > p_n > 0$), so that horse 1 is the favorite²

¹According to this formula odds can even be negative, if s_i is above $1 - t$. In our data this never happens.

²With homogeneous and correct beliefs, clearly R is ordered as $R_1 > \dots > R_n$, so that

2.1 Preferences and single-crossing

We consider a continuum of bettors, indexed by a parameter θ ; without loss of generality we assume that θ is uniformly distributed on $[0, 1]$. Because data on individual bets are not available, as in most of the literature we assume that in a given race, each bettor wagers a fixed amount (normalized to 1) on one of the horses in the race. Hence participation is not modeled, as bettors are not allowed to opt out. This implies in particular that the shares (s_i) in the pool defined above can be identified to market shares.

Each bettor θ is characterized by a utility function $V(R, p, \theta)$ over gambles, so that θ bets on the horse i that gives the highest value to $V(R_i, p_i, \theta)$. Recall from (1) that theoretically odds may be negative, so we define V on

$$]-1, +\infty[\times]0, 1] \times [0, 1]$$

Assume that V is continuously differentiable in (R, p, θ) , and is strictly increasing in R and p . Hence preferences satisfy a first order stochastic dominance property: if $p' < p$ and $R' \leq R$, then every bettor strictly prefers the gamble (R, p) to the gamble (R', p') . Since we have indexed horses by decreasing probabilities, from now on we only consider families of odds (R_1, \dots, R_n) that are ranked by increasing order.

Because V is differentiable³, we can safely define the marginal rate of substitution⁴

$$w(R, p, \theta) \equiv \frac{V_R}{V_p}(R, p, \theta) > 0$$

The properties of w fully determine the bettors' choices among gambles, and in fact this paper aims at identifying w from the observation of choices. We now introduce two additional requirements. The first one is quite mild, and essentially imposes that for each θ the rate of substitution w is neither zero nor infinity:

Assumption 1 (Inada) For any $\theta, R, p > 0$:

- for any $p' > 0$, there exists R' such that $V(R, p, \theta) < V(R', p', \theta)$;
- for any R' , there exists $p' > 0$ such that $V(R, p, \theta) > V(R', p', \theta)$.

notions of favourites and longshots can be defined equivalently on odds and on probabilities.

³Differentiability is not crucial; it just simplifies some of the equations.

⁴Throughout the paper subscripts to functions indicate partial derivatives.

Our main assumption imposes more structure to preferences:

Assumption 2 (*single-crossing*) Consider two gambles (R, p) and (R', p') , with $p' < p$. If for some θ we have

$$V(R, p, \theta) \leq V(R', p', \theta)$$

then for all $\theta' > \theta$

$$V(R, p, \theta) < V(R', p', \theta')$$

Under first-order stochastic dominance, the fact that θ prefers the gamble with the lowest winning probability ($p' < p$) implies that its odd is higher ($R' > R$): in that sense the gamble (R', p') is riskier. The assumption states that if an agent θ prefers the riskier gamble, then so does any agent θ' above θ . The single-crossing assumption thus imposes that agents can be sorted according to their “taste” for risk. Thus higher θ s prefer long-shots, while lower θ s prefer favorites.

The same point can also be made by applying a well-known sufficient condition for single-crossing, which expresses that the marginal rate of substitution $w = V_R/V_p$ is increasing with θ . Hence higher θ s value more an increase in the odd, compared to a reduction in the winning probability. We use this condition in the examples below, that illustrate the generality of the setting.

1. *Expected utility*: normalizing to zero the utility $u(-1, \theta)$ of losing the bet, we get

$$V(R, p, \theta) = pu(W_\theta + R, \theta)$$

where W_θ is the wealth of agent θ . Expected utility with reference point is a variant in which we assume, following an idea initially introduced by Kahneman and Tversky, that an agent’s utility depends of the gain she receives and not of her total wealth; so that the term W_θ disappears from the above definition. For our purpose these frameworks are equivalent, since the reference point of a bettor is fixed.

Normalizing W_θ to zero, single-crossing is shown to hold if the ratio u/u_R is decreasing with θ . This ratio is known as the fear-of-ruin index⁵. This index is a measure of risk-aversion for binomial risks; a lower fear-of-ruin index is equivalent to a lower risk-aversion for utility functions displaying constant (absolute or relative) risk aversion.

⁵See Foncel and Treich (2005) for a recent reference.

2. *Robust control*, as developed by Hansen and Sargent, models an expected utility agent that is unsure of whether the distribution p is the true one; a different distribution m may as well be the right one. In our simple setting, the utility from betting on horse i becomes

$$\min_m [m_i u(R_i) + a(\theta) e(m, p)]$$

where $e(m, p)$ is the relative entropy function that penalizes distortions from p :

$$e(m, p) = \sum_j m_j \log\left(\frac{m_j}{p_j}\right)$$

The positive coefficient $a(\theta)$ measures how better θ pays attention to probability distortions. It is now well-known⁶ that such a departure from the standard expected utility framework amounts to considering an expected utility maximizer with preferences

$$\bar{u}(R, \theta) = 1 - \exp(-u(R)/a(\theta))$$

so we are back to the preceding case. Because we have assumed that the “reference utility” u is the same for all agents, the model however imposes restrictions on the preferences of the agents, and so it is a strict subset of expected utility. Here it is easily shown that the fear-of-ruin index associated to \bar{u} is decreasing with θ if and only if $a(\theta)$ increases with θ . Hence single-crossing obtains under this simple condition. Intuitively, due to the minimum in the objective function agents tend to underestimate probabilities of favorable events. Winning a bet on an outsider is a very favorable event, so that agents with a low coefficient a tend to bet on favorites⁷, and conversely for agents with a high value for a . The case when a is infinite brings us back to (homogeneous) expected utility.

Note that this introduces a form of heterogeneity of beliefs in agents’ preferences; it is easy to see that probabilities are “exponentially tilted” so that when considering a bet on horse i , agent θ in fact behaves as if the probability that this horse wins is not p_i but

$$\frac{p_i \exp(-u(R_i)/a(\theta))}{1 - p_i + p_i \exp(-u(R_i)/a(\theta))};$$

since this tilting depends on θ , agents differ by their degree of pessimism.

⁶For a recent reference, see Hansen (2007, Section VI).

⁷One may in fact show the same properties for a wider class of penalty functions.

3. *Rank-Dependent Expected Utility Theory* also allows for a deformation of probabilities. That is, there exists an increasing function ρ such that the utility V writes

$$V(p, R, \theta) = \rho(p, \theta)u(R, \theta)$$

Note that, in general, both the utility and the probability deformation functions may vary with θ . Single-crossing then requires an additional condition, namely that ρ_p/ρ , which is positive, be decreasing in θ . Take the simplest case,

$$\rho(p, \theta) = p^{c(\theta)};$$

then we need c to be a decreasing function of θ , so that larger θ 's exaggerate small probabilities more (or underestimate them less) than smaller θ 's.

Again, this allows us to account for some heterogeneity in beliefs; here it seems more flexible than with the robust control approach, but it does not depend on odds. Taking into account more random heterogeneity in beliefs is also possible, but it would take us beyond nonparametric identification.

Many other families of preferences, such as cumulative prospect theory, may also be made compatible with our setting (though the single-crossing condition becomes more involved). The main limitation is that we require agents to only pay attention to consequences: the utility $V(R, p, \theta)$ derived from betting on one horse cannot depend on the characteristics of other horses. Therefore it seems difficult to make the setting encompass other features of risk attitudes, such as regret. Another interesting feature of preferences is how they deal with ambiguous probabilities, and whether bettors prefer or reject horses whose past performances are less well-known. One could specify preferences as in Klibanoff-Marinacci-Mukerji (2005), but single-crossing is not likely to hold: there is no reason why the winning probability of a favorite should be perceived as less (or more) ambiguous than the winning probability of a long-shot. Finally, the single-crossing assumption seems to exclude the case of a fully homogeneous population considered in Jullien and Salanié (2000): when V is independent of θ , the strict inequality in the assumption cannot hold. Still homogeneity can be dealt with as a limiting case, as we shall see.

2.2 Market Shares and Equilibrium

The winning probabilities are assumed exogenous and characterize the race. In contrast, the odds are endogenous: the bettors' behavior determines market shares, which in turn determine odds through the parimutuel rule (1).

In such a setting, it is natural to rely on the concept of rational expectations equilibria: agents determine their behavior given their anticipations on odds, and these anticipations are fulfilled in equilibrium. Gandhi (2006) characterizes these equilibria in a setting with general preferences, and proves existence and uniqueness under mild conditions. Moreover the correspondence between probabilities and odds is one-to-one, so that one can recover probabilities from observing odds. We provide below a version of these results adapted to our simpler framework (proofs are relegated to the Appendix).

Let us assume that the families \mathbf{p} and \mathbf{R} are given, and known to all agents. Each agent then optimizes on which horse to bet on, and a simple consequence of the single crossing condition is the following:

Proposition 1 *Suppose that \mathbf{p} and \mathbf{R} are such that all market shares are positive: $s_i > 0, i = 1, \dots, n$. Then there exists a family $(\theta_j)_{j=0, \dots, n}$, with $\theta_0 = 0 < \theta_1 < \dots < \theta_{n-1} < \theta_n = 1$, such that:*

- for all $i = 1, \dots, n$, if $\theta_{i-1} < \theta < \theta_i$ then bettor θ strictly prefers to bet on horse i than on any other horse;
- for all $i < n$ we have

$$V(p_i, R_i, \theta_i) = V(p_{i+1}, R_{i+1}, \theta_i) \quad (3)$$

In words: under single crossing, if we rank horses by increasing odds, we have a partition of the set of bettors into n intervals, each of them gathering bettors who bet for the same horse. The bounds of the intervals are defined by an indifference condition; namely, for $i = 1, \dots, n - 1$, there exists a *marginal bettor* θ_i who is indifferent between betting on horses i and $i + 1$.

As a simple corollary, because the distribution of θ is uniform the market shares s_i for horse $i = 1, \dots, n$ can be expressed as

$$s_i = \theta_i - \theta_{i-1}$$

or equivalently

$$\theta_i = \sum_{j \leq i} s_j$$

Recall that odds are determined from market shares as in (1) and (2), so that in equilibrium one must have $\theta_i = \theta_i(\mathbf{R})$, where

$$\theta_i(\mathbf{R}) \equiv \frac{\sum_{j \leq i} \frac{1}{R_{j+1}}}{\sum_j \frac{1}{R_{j+1}}} \quad i = 1, \dots, n \quad (4)$$

We can now define a market equilibrium. We want bettors to behave optimally given odds and probabilities, as expressed in (3); and we want odds to result from market shares, which is what the equalities $\theta_i = \theta_i(\mathbf{R})$ impose. This motivates the following definition:

Definition 1 Consider a race (\mathbf{p}, t) . $\mathbf{R} = (R_1, \dots, R_n)$ is a family of equilibrium odds if (2) holds and

$$\forall i < n \quad V(p_i, R_i, \theta_i(\mathbf{R})) = V(p_{i+1}, R_{i+1}, \theta_i(\mathbf{R})) \quad (5)$$

We then prove existence and uniqueness:

Proposition 2 Given a race (\mathbf{p}, t) , there exists a unique family \mathbf{R} of equilibrium odds.

Hence to each race one can associate a unique rational expectations equilibrium, with positive market shares. From an empirical viewpoint, however, the odds are directly observable, while probabilities have to be estimated. Fortunately, probabilities can be uniquely recovered from odds:

Proposition 3 For any \mathbf{R} ranked in increasing odds ($-1 < R_1 < \dots < R_n$), there exists a unique (\mathbf{p}, t) such that \mathbf{R} is a family of equilibrium odds for (\mathbf{p}, t) .

As already observed, the value of the take t is in fact given by (2), and results from the rules of parimutuel betting. On the other hand, probabilities result from preferences. The function $\mathbf{p}(\mathbf{R}) = (p_1(\mathbf{R}), \dots, p_n(\mathbf{R}))$ implicitly defined in Proposition 3 thus conveys some information on the underlying preferences of bettors. This function is continuously differentiable, from our assumptions on V . Since choices are fully determined by the marginal rates of substitution $w = V_R/V_p$, we shall say hereafter that $\mathbf{p}(\mathbf{R})$ characterizes market equilibria associated to the family V , or equivalently w .

3 Testable implications and identifiability

A natural question is whether our setting can generate testable predictions about observed behavior. In other words, does the theory impose further restrictions on the form of the function $\mathbf{p}(\mathbf{R})$?

3.1 Testable implications

Since V increases in p , we may define Γ as the inverse of V with respect to p :

$$\forall R, p, \theta \quad \Gamma(V(R, p, \theta), R, \theta) = p$$

One can then define a function G as

$$G(R', p', R, \theta) = \Gamma(V(R', p', \theta), R, \theta) \quad (6)$$

Now we can rewrite the equilibrium conditions in Definition 1 as

$$\forall i < n \quad p_i(\mathbf{R}) = G(R_{i+1}, p_{i+1}(\mathbf{R}), R_i, \theta_i(\mathbf{R})) \quad (7)$$

where $\theta_i(\mathbf{R})$ was defined in (4). We immediately obtain several properties of G :

Proposition 4 *If $\mathbf{p}(\mathbf{R})$ is the characterization of market equilibria associated to some family V , then there exists a function $G(R', p', R, \theta)$ such that*

- i) G is continuously differentiable, increasing with R' , p' and θ , and decreasing with R ;*
- ii) $G_{R'}/G_{p'}$ is independent of R ;*
- iii) $G(R, p, R, \theta) = p$;*
- iv) (7) holds for any family (R_1, \dots, R_n) .*

Properties i), ii), iii) derive from our assumptions on V , and the definition of Γ and G . As an illustration, recall that the single-crossing assumption states that for all $R < R'$ and $p > p'$

$$V(R, p, \theta) \leq V(R', p', \theta) \quad \Rightarrow \quad \forall \theta' > \theta \quad V(R, p, \theta') < V(R', p', \theta')$$

This is equivalent to

$$p \leq G(R', p', R, \theta) \quad \Rightarrow \quad \forall \theta' > \theta \quad p < G(R', p', R, \theta')$$

and thus G must be increasing with θ , as required in property i).

Property iv) means that the winning probability $p_i(\mathbf{R})$, which normally depends on the whole family of odds, can be computed from only four numbers: the pair of odds R_i and R_{i+1} , the index of the marginal consumer $\theta_i(\mathbf{R})$, and the probability $p_i(\mathbf{R})$ of the horse ranked by bettors just before $(i + 1)$. Hence $p_i(\mathbf{R})$ and $\theta_i(\mathbf{R})$ are sufficient statistics for the $n - 2$ odds that are missing from this list. Note moreover that G does not depend on the index i , on the number of horses n , nor on the take t . These and the other properties of G listed in Proposition 4 also provide directly testable predictions of our model.

3.2 Exhaustiveness and identification

Now suppose that we are given a very large sample of races, and that we know the family of odds and the identity of the winning horse for each race. We can then compute the empirical frequency of the event “horse R_i wins a race characterized by the family R ”, and collect these frequencies in a function $\mathbf{p}(\mathbf{R})$. We can then check whether there exists a function G verifying properties i)-iv) above.

The first step would be to check that the non-parametric regression of $p_i(\mathbf{R})$ on the four other variables $(R_{i+1}, p_{i+1}(\mathbf{R}), R_i, \theta_i(\mathbf{R}))$ has a perfect fit, which of course will not happen. Our theory does not allow at this stage for deviations from equilibrium, and so we will be content with checking that the fit is “good enough”. Suppose it to be the case. One may first ask whether properties i)-iv) are exhaustive; does the knowledge of G allow us to recover a family of utilities V , such that $\mathbf{p}(\mathbf{R})$ characterizes the market equilibria associated to V ? If moreover this family is unique, we would also have proven that preferences are non-parametrically identified.

We can easily prove that this model is nonparametrically identified. Use property ii) to define a function w by

$$w(R', p', \theta) = \frac{G_{R'}}{G_{p'}}(R', p', R, \theta)$$

Choose any V whose marginal rate of substitution V_R/V_p is equal to w . Then by definition

$$\frac{V_R}{V_p}(R', p', \theta) = \frac{G_{R'}}{G_{p'}}(R', p', R, \theta)$$

so that there exists a function \tilde{G} , increasing with its first argument, such that

$$G(R', p', R, \theta) = \tilde{G}(V(R', p', \theta), R, \theta)$$

Then i) implies that V is increasing with R' and p' . Notice that from iii), it must be the case that \tilde{G} is the inverse of V with respect to p .

Let us now prove that V verifies the single-crossing assumption. Assume that $V(R, p, \theta) \leq V(R', p', \theta)$, for $R < R'$ and $p > p'$. Since \tilde{G} is the inverse of V , we get

$$p \leq \tilde{G}(V(R', p', \theta), R, \theta) = G(R', p', R, \theta)$$

Since from property i) G is increasing with θ , we obtain that

$$p < G(R', p', R, \theta') = \tilde{G}(V(R', p', \theta'), R, \theta')$$

Finally, since \tilde{G} is the inverse of V we get

$$V(R, p, \theta') < V(R', p', \theta')$$

Therefore V verifies the single-crossing assumption, as announced.

Finally, since \tilde{G} is the inverse of V property iv) can be rewritten as

$$\forall i < n \quad V(R_i, p_i(\mathbf{R}), \theta_i(\mathbf{R})) = V(R_{i+1}, p_{i+1}(\mathbf{R}), \theta_i(\mathbf{R}))$$

and these are exactly the equilibrium conditions in Definition 1. Thus $\mathbf{p}(\mathbf{R})$ characterizes the market equilibria associated to V .

The previous paragraphs prove two results. Firstly, the properties i)-iv) stated in Proposition 4 are in fact **exhaustive**: since they are strong enough to ensure the existence of a family V satisfying our assumptions, no other testable implications can be found⁸.

Secondly, these properties are strong enough to recover a unique family of rates of substitution w . Indeed we want (6) to hold, but this implies

$$\frac{V_R}{V_p}(R', p', \theta) = \frac{G_{R'}}{G_{p'}}(R', p', R, \theta)$$

so that there is a unique candidate for w , and consequently the family V is identified up to an increasing function of θ .

⁸For simplicity we have omitted the predictions corresponding to the Inada assumption.

Remark: identification of course only holds on the support of the random variables that we defined. This has an important consequence in our setting. Assume that no race has more than n horses (or that not enough races have more than n horses). The favourite in each race by definition has the largest market share, and so we will always observe $\theta_1 > 1/n$. Since identification relies on boundary conditions in the θ_i 's, it follows that we cannot hope to recover the family of functions $V(.,.,\theta)$ for $\theta < 1/n$.

3.3 The case of expected utility

The analysis can be rephrased to deal with the case when V is a family of expected utility functionals:

$$V(R, p, \theta) = pu(R, \theta)$$

(we again normalized $u(-1, \theta)$ to zero). Then the indifference condition in Definition 1 becomes

$$p_{i+1}(\mathbf{R}) = p_i(\mathbf{R})u(R_i, \theta_i(\mathbf{R}))/u(R_{i+1}, \theta_i(\mathbf{R}))$$

We thus obtain a new testable implication:

v) G is linear with respect to p' .

Reciprocally, if i)-v) hold, then from v) we get

$$G(R', p', R, \theta) = p'H(R', R, \theta)$$

From ii) we obtain

$$\frac{\partial}{\partial R} \frac{H_{R'}}{H} = 0,$$

or equivalently

$$H(R', R, \theta) = A(R, \theta)B(R', \theta)$$

and from iii) we obtain $A(R, \theta)B(R, \theta) = 1$. Hence

$$G(R', p', R, \theta) = p' \frac{B(R', \theta)}{B(R, \theta)}$$

and B is the required von Neumann-Morgenstern utility function. Once more, this function is uniquely identified (up to a multiplicative constant), once G is estimated from the data.

In fact, say that we normalize $u(R_m, \theta) \equiv 1$ for some R_m . Then it is easy to see that the whole family of vNM functions can be recovered by the simple (but not very practical) formula

$$u(R, \theta) = E \left(\frac{p_{i+1}(\mathbf{R})}{p_i(\mathbf{R})} \mid R_i = R, R_{i+1} = R_m, \theta_i(\mathbf{R}) = \theta \right).$$

4 The Data

Our data consist of a large sample of thoroughbred races (the dominant form of organized horse racing worldwide) in the United States, spanning the years 2001 through 2004. The data were collected by professional handicappers from the racing portal paceadvantage.com, and a selection of the variables that they collect were shared with us. In particular, for each horse race in the data, the date, track, race number in the day, number of horses, the market odds for each horse, and finishing position for each horse are observed. Excluded from the data are variables that the handicappers use for competitive purposes, such as various measures of the historical speed of each horse.

For the present analysis, we focus on a single year of data, namely the year 2001. The 2001 data contain races from 77 tracks spread over 33 states. There are 448,314 horses in the sample. Of these horses, 111 were “purse only” horses, meaning that they ran a race, but were not bet upon and hence did not receive a value for betting odds. After exclusion of the corresponding races, there remains 447,387 horses and 54,199 races.

The average number of horses in each race is 8.26, with a min of 2 and a max of 17 (the standard deviation being 1.87). The betting odds over horses in the data range from extreme favorites (odds equaling .05, i.e., horses paying 5 cents on the dollar), to extreme longshots (odds equaling 999.9, i.e., horses paying close to 1000 dollars on the dollar). The mean and median odds on a horse are 15.1 and 8.3 respectively. Thus the distribution of odds is skewed to the right, as there are a sizeable fraction of extreme longshots, with 29,171 horses having odds over 50.

Our main covariate information at the level of a race, is the track, date, and the number of the race in a given day. The dates thus allow us to separate weekday races from weekend races, and by matching the zip code of the track with the 2000 Census, we can classify each track on an urban/rural scale. The urban/rural scale in the census captures a variety of zip code level information, including income, demographics, and commuting patterns.

For the present analysis, we abstract away from these covariates, and treat the data as homogenous sample of races. Thus we view the odds as

being generated through a stable population of preferences given the varying probabilities across races. Future work aims to incorporate the covariate information into the analysis, both as sources of variation in tastes and as driving selection of bettors. We also focus on races in which two different horses have different odds, so as to avoid a special numerical treatment for ties; and we exclude races in which the odds of a horse are above 50.

5 Estimation Strategy

The fundamental equation of our model indicates that each probability may be expressed as a function of four other variables:

$$\forall i < n \quad p_i(\mathbf{R}) = G(R_{i+1}, p_{i+1}(\mathbf{R}), R_i, \theta_i(\mathbf{R}))$$

Our estimation strategy aims at recovering both the probabilities and the function G . Since a winning probability $p_i(\mathbf{R})$ is the expectation of an indicator variable $1_i(\mathbf{R})$ that indicates whether horse i won race \mathbf{R} , the above equation can be written

$$p_i(\mathbf{R}) = E[1_i(\mathbf{R}) | R_{i+1}, p_{i+1}(\mathbf{R}), R_i, \theta_i(\mathbf{R})] \quad i < n$$

Such a recursive system of equations can be solved iteratively. Define a recursive process (p^m) through

$$p_i^{m+1}(\mathbf{R}) = E[1_i(\mathbf{R}) | R_{i+1}, p_{i+1}^m(\mathbf{R}), R_i, \theta_i(\mathbf{R})] \quad i < n$$

$$p_n^{m+1}(\mathbf{R}) = 1 - \sum_{j < n} p_j^{m+1}(\mathbf{R})$$

The expectation in the first equation can be computed using standard non-parametric techniques. If this process converges, it converges to probabilities that verify (7). Then both G and the probabilities are uniquely identified.

This seems an attractive strategy, and we indeed obtained estimates of probabilities and utility functionals this way. However, it turns out that this algorithm ends up fitting the market shares and neglecting the fit of the probabilities. At this stage, we ended up working with a different approach. We first estimate the $\mathbf{p}(\mathbf{R})$ function with a flexible specification. Without loss of generality, we can write

$$p_i(\mathbf{R}) = \frac{\exp(P_i(\mathbf{R}))}{\sum_{j=1}^n \exp(P_j(\mathbf{R}))}.$$

We now specify the P functions as

$$P_i(\mathbf{R}) = -\log(R_i + 1) + \sum_{k=1}^K a_k(R_i)T_k(\mathbf{R}).$$

If the sum on the right-hand side were all zero, then the probabilities $p_i(\mathbf{R})$ would all be equal to the “naive” (risk-neutral) probabilities

$$p_i^n(R) = \frac{1}{R_i + 1} \frac{1}{\sum_j \frac{1}{R_j + 1}};$$

the terms in the sum consist in cubic splines of 5 to 10 knots (the a_k 's) applied to four symmetric functions of \mathbf{R} (the T_k 's).

Once these probabilities are estimated, we can easily recover the expected utility functionals that best fit them. A little algebra suggests the generalized additive model:

$$\frac{p_{i+1}(\mathbf{R})}{p_i(\mathbf{R})} = \frac{p_{i+1}^n(\mathbf{R})}{p_i^n(\mathbf{R})} + \log u(R_i, \theta_i(\mathbf{R})) - \log u(R_{i+1}, \theta_i(\mathbf{R})).$$

This can be estimated using seminonparametric routines (see e.g. Hastie, Tibshirani and Friedman (2001)⁹)

The resulting u functions can easily be projected on special cases, like the exponentially tilted utilities of robust control.

In principle, this approach can be adapted to non-expected utility, although estimation is not so straightforward and may need to be iterative. We plan to do this in further work.

6 Results

The results discussed below were obtained on a subsample of 5,184 races.

Let us first examine how estimated probabilities differ from naive probabilities. Figure 1 plots the ratio of estimated probability to naive probability; it shows that the former is much smaller than the latter for small probabilities (longshots). This is the well-known favourite-longshot bias, according to which expected returns are much lower for longshots than for favourites.

⁹We used the package `mgcv` in R. The nonparametric procedures used in the rest of the paper use the package `np`.

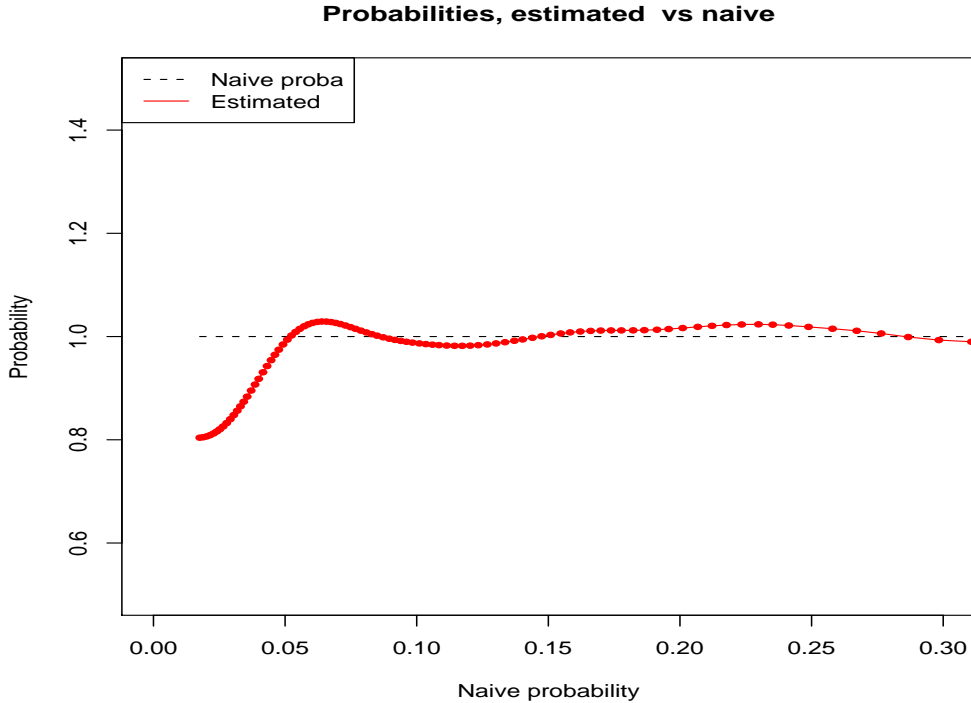


Figure 1: Ratio of p_i to p_i^n

Indeed, the expected return on a horse is $p_i R_i - (1 - p_i)$, which also equals

$$\frac{p_i}{(1 - t)p_i^n} - 1;$$

since t is close to 0.8 in most races in our sample, the lesson from the figure is clear: this market does not price risk in the most usual manner.

Once the probabilities are estimated, we fit a GAM model as explained above to recover the expected utility functionals. Let us start by imposing that all bettors have the same preferences. The results are plotted in Figure 2.

An obvious comparison point is risk-neutrality, the 45 degree line on the graph. A representative bettor does not seem to differ much from a risk-neutral bettor, in contrast with the results in Jullien and Salanié (2000). Their paper focused on the homogeneous case. It explicitly solved the system of equations giving the probabilities, using parameterized families of preferences. Parameters were then estimated by maximizing the likelihood associated to the winning probability of the observed winner in the sample of races. In the EU case, the best estimate was a CARA function with

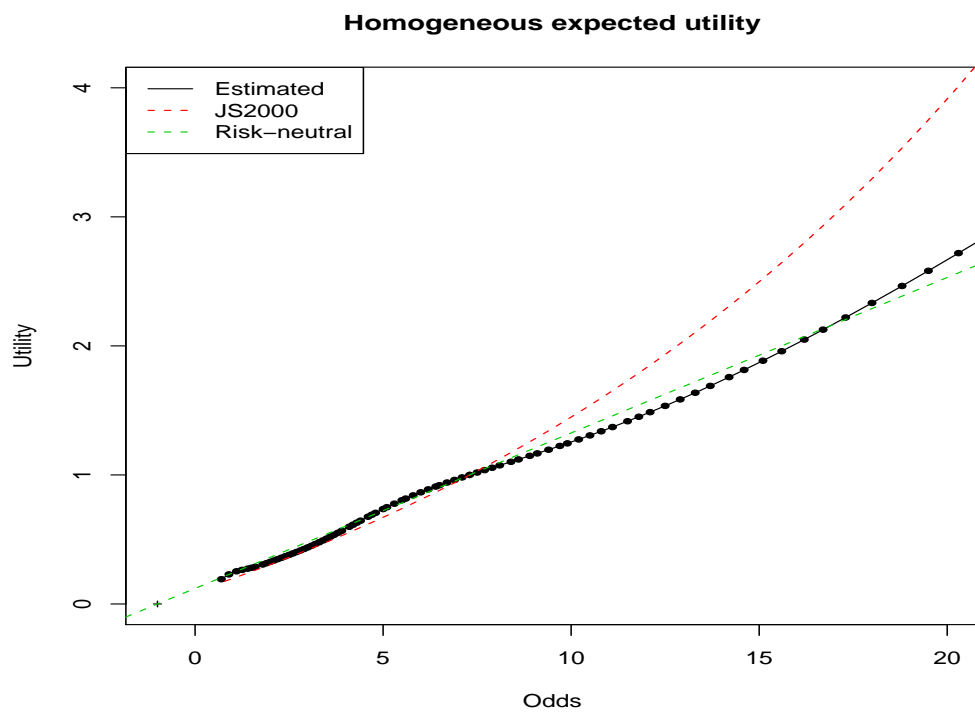


Figure 2: Estimated homogeneous expected utility functions

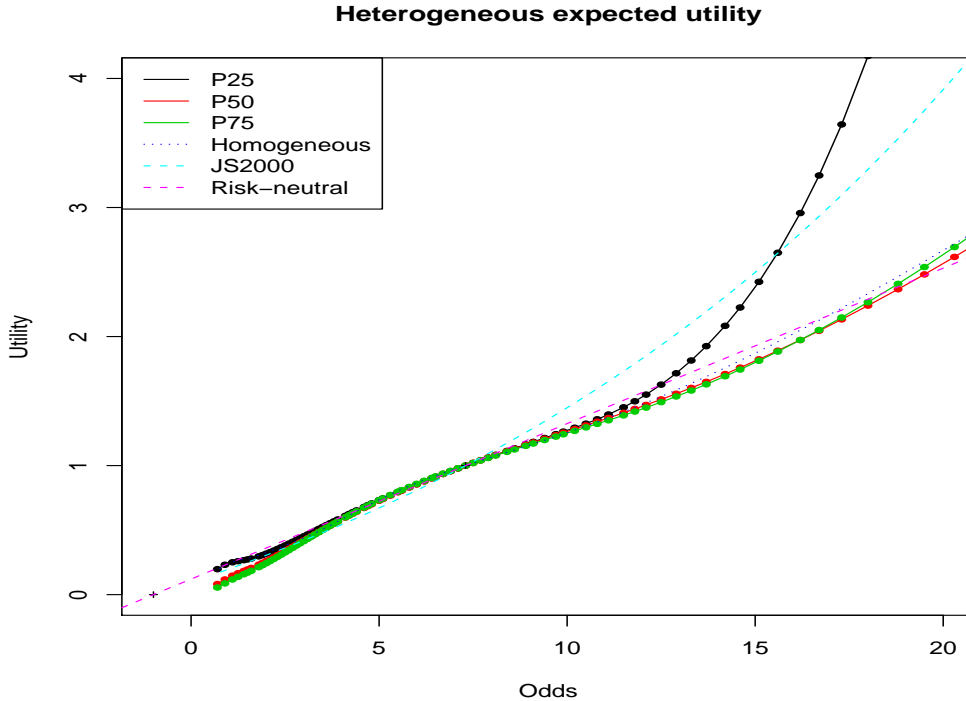


Figure 3: Estimated heterogeneous expected utility functions

risk-aversion equal to -0.06 , which is again plotted in figure 2. Our results, which are fully nonparametric, show that parametric forms may lead analysts astray in this field.

The case of heterogeneous preferences is of course of more interest to us. Figure 3 plots our results. The curves labeled P25, P50, P75 correspond to the three quartiles of the empirical distribution of $\theta_i(\mathbf{R})$ in our sample. The risk-neutrality line and the two homogeneous estimates (Jullien and Salanié’s and ours) are also plotted. Preferences again seem close to risk-neutrality, except for the very prudent bettors (with smaller θ ’s.)

When we allow for heterogeneous preferences, the generalized R^2 of the GAM regression above increases from 0.52 to 0.62. This is by no means spectacular, but it is of course hugely significant in statistical terms.

Recall that in the expected utility setting, our single-crossing assumption requires that the fear-of-ruin index u/u_R be a decreasing function of θ . The picture is rather mixed on our estimates, as shown on Figure 4: while single-crossing seems to hold for favourites, it is definitely violated for outsiders. This suggests that our model, while much more general than previous

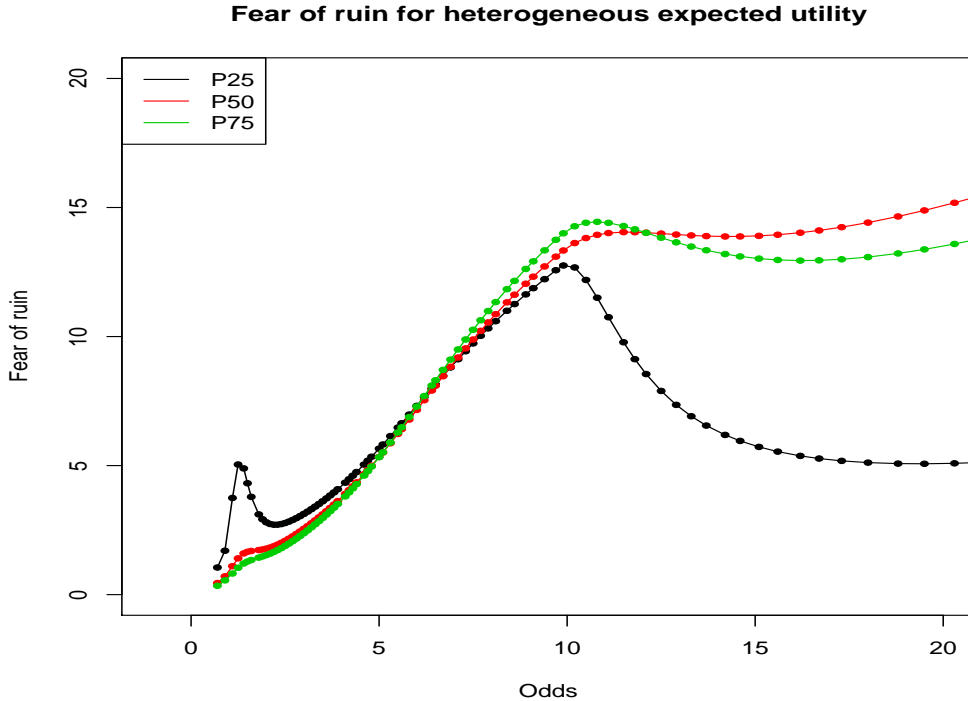


Figure 4: Estimated fear-of-ruin

attempts, is still not general enough.

Finally, we also estimated the robust control model, by running the (non-linear) regression

$$\frac{\frac{p_{i+1}(\mathbf{R})}{p_i(\mathbf{R})}}{\frac{p_{i+1}^n(\mathbf{R})}{p_i^n(\mathbf{R})}} = \frac{1 - \exp(-u(R_i)/a(\theta))}{1 - \exp(-u(R_{i+1})/a(\theta))},$$

which follows from elementary algebra. To estimate it, we use cubic splines again, with 30 knots for R and 20 knots for θ . The generalized R^2 of this regression is 0.55, whereas it is 0.63 for our homogeneous expected utility estimate.

Figure 5 plots the estimated $u(R)$ under robust control. It is not very different from the picture we had before, or again from risk-neutrality.

More interestingly, Figure 6 plots our estimated $1/a(\theta)$ as a function of θ . To interpret this graph, it is useful to recall that robust control tilts the

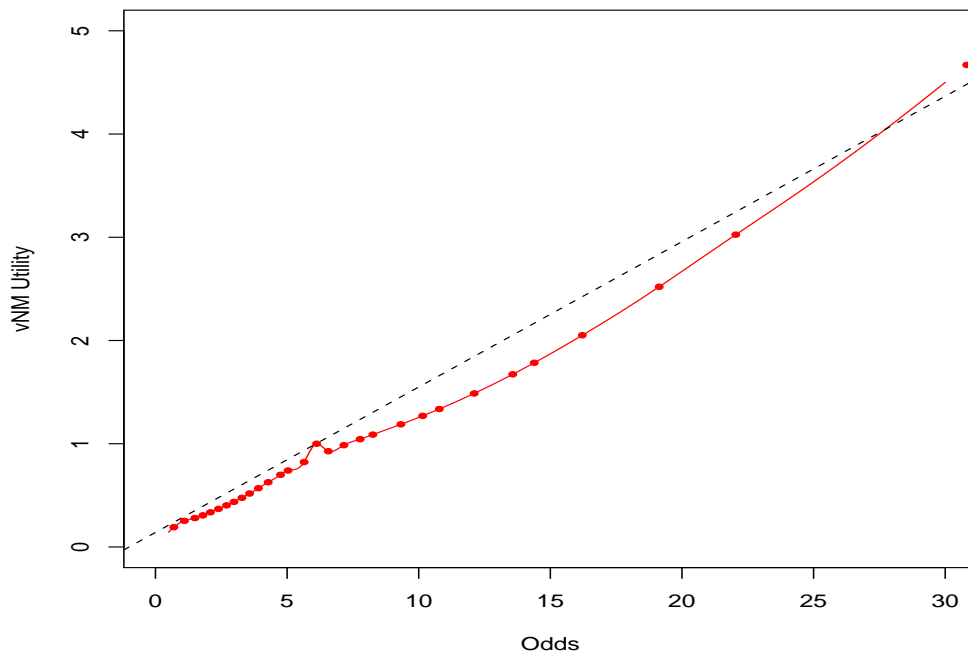


Figure 5: Estimated robust control expected utility function

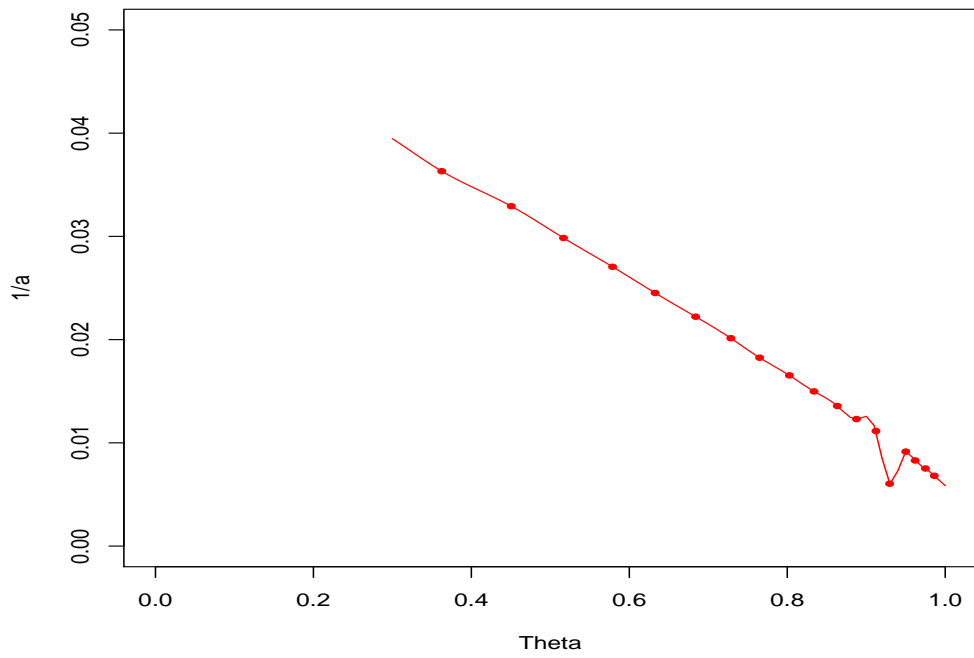


Figure 6: Estimated robust control $1/a(\theta)$ parameter

probabilities exponentially: the odds ratio of winning on a horse with odds R and probability p , which is normally

$$\frac{p}{1-p},$$

is multiplied by the factor $\exp(-u(R)/a(\theta))$. We normalized utility to be zero in -1 and 1 at the median odds, which are about 8 to 1 . Thus when $1/a$ is about 0.03 , the argument of the exponential is roughly $-3R/1000$. For a favourite with $R = 2$, the odds ratio would decrease by 0.6% ; for the median horse it would decrease by 2.4% , and for a longshot with $R = 50$ it would decrease by 15% . This is for a given agent θ of course; but according to the graph, agents which bet on longshots (the larger θ 's) have a very low $a(\theta)$ and do not care for robustness anyway.

Note that single-crossing easily holds in the robust control model, as it requires that the coefficient $a(\theta)$ be increasing.

Appendix

Proof of Proposition 1: from the single-crossing assumption, the set of agents that strictly prefer horse i to horse $j > i$ is an interval containing 0 . Similarly, the set of agents that strictly prefer horse i to horse $j < i$ is an interval containing 1 . Therefore the set of agents that strictly prefer horse i to all other horses is an interval. The single-crossing assumption also implies that these intervals are ranked by increasing i ; and that the set of agents indifferent between horse i and horse $(i + 1)$ is a singleton. Q.E.D.

Proof of Proposition 2: from definition 1, given a race (\mathbf{p}, t) we have to find a family \mathbf{R} such that

$$\forall i < n \quad V(p_i, R_i, (1-t) \sum_{j \leq i} \frac{1}{R_j + 1}) = V(p_{i+1}, R_{i+1}, (1-t) \sum_{j \leq i} \frac{1}{R_j + 1})$$

The right-hand-side is increasing with R_{i+1} , is strictly below the left-hand-side at $R_{i+1} = R_i$, and strictly above it for R_{i+1} high enough, from the Inada assumption. Thus this equality defines a unique R_{i+1} , such that $R_{i+1} > R_i$. Moreover the single-crossing assumption ensures that

$$V_\theta(p_i, R_i, (1-t) \sum_{j \leq i} \frac{1}{R_j + 1}) \leq V_\theta(p_{i+1}, R_{i+1}, (1-t) \sum_{j \leq i} \frac{1}{R_j + 1})$$

Since in addition $V_R > 0$, this proves that R_{i+1} is an increasing function of R_i , and a non-decreasing function of each R_j , $j < i$. Iterating this remark, we get that each R_{i+1} is an increasing function of R_1 . Replacing in (2), we get an equation in R_1 which has at most one solution. Existence follows from the fact that (R_1, \dots, R_n) forms an increasing sequence, so that by setting R_1 high enough we get $1/(1-t) > \sum_j 1/(1+R_j)$; and from the fact that when R_1 goes to -1 we get $1/(1-t) < \sum_j 1/(1+R_j)$. Q.E.D.

Proof of Proposition 3: If we know the odds, then we know the take and the market shares, from (1) and (2); and we also know the indexes $(\theta_j(\mathbf{R}))$ of marginal bettors, from (4). There only remains to find a family \mathbf{p} solution to the system

$$\forall i < n \quad V(R_i, p_i, \theta_i) = V(R_{i+1}, p_{i+1}, \theta_i)$$

Let us focus on positive probabilities. The right-hand-side is increasing with p_{i+1} , is strictly above the left-hand-side at $p_{i+1} = p_i$, and strictly below when p_{i+1} goes to zero from our Inada assumption: therefore p_{i+1} is uniquely defined, and $p_{i+1} < p_i$. Moreover p_{i+1} is an increasing function of p_i , and thus of p_1 . Finally p_1 is uniquely determined by $p_1 + \sum_{i < n} p_{i+1} = 1$ (existence follows from checking the cases $p_1 \rightarrow 0$ and $p_1 = 1$). Q.E.D.

References

- R. B. Barsky, F. T. Juster, M. S. Kimball and M. D. Shapiro (1997)**, "Preference Parameters and Behavioral Heterogeneity: An Experimental Approach in the Health and Retirement Study", *Quarterly Journal of Economics*, Vol. 112, No. 2, May, 537-579.
- P.A. Chiappori and M. Paiella (2007)**, "Relative Risk-Aversion is constant: evidence from panel data", Columbia University discussion paper.
- J. Foncel and N. Treich (2005)**, "Fear of ruin", *Journal of Risk and Uncertainty* 31, 289-300.
- L. Guiso and M. Paiella (2003)**, "Risk-aversion, wealth and background risk".
- A. Gandhi (2006)**, "Rational Expectations at the Racetrack: Testing Expected Utility Using Prediction Market Prices", Chicago University working paper.
- L.P. Hansen (2007)**, "Beliefs, Doubts and Learning: Valuing Macroeconomic Risk", Richard T. Ely Lecture, AEA Papers and Proceedings, May, 1-30.

T. Hastie, R. Tibshirani and J. Friedman (2001), *The Elements of Statistical Learning*, Springer.

B. Jullien and B. Salanié (2000), “Estimating Preferences under Risk: The Case of Racetrack Bettors”, *Journal of Political Economy*, vol. 108, p. 503-530.

P. Klibanoff, M. Marinacci and S. Mukerji (2005), “A smooth model of decision-making under ambiguity”, *Econometrica* 73-6, Nov., 1849-92.