The numeraire e-variable and reverse information projection

Johannes Ruf

Department of Mathematics, LSE

with Martin Larsson, Aaditya Ramdas

Bachelier Finance Society One World Seminar, April 2024
Martin Larsson

Aaditya Ramdas
A New Interpretation of Information Rate
reproduced with permission of AT&T

By J. L. Kelly, Jr.

(Manuscript received March 21, 1956)

If the input symbols to a communication channel represent the outcomes of a chance event on which bets are available at odds consistent with their probabilities (i.e., “fair” odds), a gambler can use the knowledge given him by the received symbols to cause his money to grow exponentially. The maximum exponential rate of growth of the gambler’s capital is equal to the rate of transmission of information over the channel. This result is generalized to include the case of arbitrary odds.
Kelly betting

► Assume iid coin flips $Z_i$ of bias $p > 1/2$ (for known $p$).
► We start with one dollar, and we are able to do “double or nothing” bets. (In each round, we bet some fraction $\lambda$ of our wealth on heads, if the coin is heads, we earn the amount we bet, and if it is tails, we lose the amount we bet.)
► What fraction $\lambda$ of our wealth should we bet at each step?
► Wealth after $t$ rounds:

$$W_t(\lambda) = \prod_{s=1}^{t}(1 + \lambda(2Z_i - 1)).$$
Kelly betting (cont’ed)

\[ W_t(\lambda) = \prod_{s=1}^{t} (1 + \lambda(2Z_i - 1)) \]

Kelly suggested: choose \( \lambda \) to maximise

\[ \lim_{t \to \infty} \frac{\mathbb{E}[\log W_t(\lambda)]}{t}. \]

Solution: bet \( \lambda^* = 2(p - 1/2) \) on heads.

Optimal wealth:

\[ W_t(\lambda^*) = \exp(t \times H(p|0.5) + o(t)), \]

where \( H \) is the relative entropy (KL divergence).
OPTIMAL GAMBLING SYSTEMS FOR FAVORABLE GAMES

L. BREIMAN
UNIVERSITY OF CALIFORNIA, LOS ANGELES

1. Introduction

Assume that we are hardened and unscrupulous types with an infinitely wealthy friend. We induce him to match any bet we wish to make on the event that a coin biased in our favor will turn up heads. That is, at every toss we have probability $p > 1/2$ of doubling the amount of our bet. If we are clever, as well as unscrupulous, we soon begin to worry about how much of our available fortune to bet at every toss. Betting everything we have on heads on every toss will lead to almost certain bankruptcy. On the other hand, if we bet a small, but fixed, fraction (we assume throughout that money is infinitely divisible) of our available fortune at every toss, then the law of large numbers informs us that our fortune converges almost surely to plus infinity. What to do?

- Generalises Kelly betting to other settings.
- Proves that the Kelly criterion also asymptotically optimises
  1. expected time to reach a threshold wealth;
  2. expected wealth at some threshold time.
In finance, e.g., in SPT or in the benchmark approach, lots of insights into the numéraire portfolio:

The numéraire portfolio in semimartingale financial models

Ioannis Karatzas · Constantinos Kardaras

Received: 30 May 2006 / Accepted: 2 April 2007 / Published online: 8 August 2007 © Springer-Verlag 2007

Abstract We study the existence of the numéraire portfolio under predictable convex constraints in a general semimartingale model of a financial market. The numéraire prices determined exogenously and can be traded without friction, i.e., transaction costs are non-existent or negligible. Our main concern will be a problem of dynamic stochastic optimization: to find a trading strategy whose wealth appears “better” when compared to the wealth generated by any other strategy, in the sense that the ratio of the two processes is a supermartingale. If such a strategy exists, it is essentially unique and is called numéraire portfolio. Necessary and sufficient conditions for the existence of such a portfolio are given in terms of the market data.

(Also Kallsen-Goll, Becherer, Fernholz, Platen, ...)

Portfolio Theory and Arbitrage
A Course in Mathematical Finance

Ioannis Karatzas
Constantinos Kardaras
Testing by betting

In order to test a hypothesis, one sets up a game such that: if the null is true, no strategy can systematically make (toy) money, but if the null is false, then a good betting strategy can make money.

Shafer & Vovk

- Wealth in the game is directly a measure of evidence against the null.
- Each strategy of the statistician = a different estimator or test statistic.
- So there are “good” and “bad” strategies for betting, just as there are good and bad estimators or test statistics.
- Kelly’s game corresponds to $H_0$: “fair coin” against $H_1$: “bias” $p$.

(For standard NP framework, see also Cvitanić & Karatzas.)
Testing by betting: point null and point alternative

Setup:
- Let $Z_i$ be observations.
- $H_0 : Z_i \sim^{iid} P$; $H_1 : Z_i \sim^{iid} Q$.
- Assumption: $P$ and $Q$ are equivalent.
Testing by betting: point null and point alternative

Setup:

- Let $Z_i$ be observations.
- $H_0 : Z_i \sim^{iid} P$; $H_1 : Z_i \sim^{iid} Q$.
- Assumption: $P$ and $Q$ are equivalent.

Game:

- Initial capital $W_0 = 1$.
- For each $t = 1, 2, \ldots$:
  Statistician declares “bet” $B_t : \mathcal{Z} \rightarrow [0, \infty)$ s.t.

\[
\mathbb{E}_P[B_t(Z_t)|Z_1, \ldots Z_{t-1}] \leq 1.
\]

Reality reveals $Z_t$.
Statistician’s wealth becomes $W_t = W_{t-1}B_t(Z_t)$. 
What is the log-optimal betting strategy?

- The log-optimal bet is the likelihood ratio $B_t(z) = q(z)/p(z)$.
- The corresponding wealth process is

$$W_t = \prod_{s=1}^{t} \frac{q(Z_s)}{p(Z_s)}.$$

- $(W_t)$ is a positive $P$-martingale.
- This choice of bets maximises $\mathbb{E}_Q[\log(W_t)]$, which equals (under the optimal choice) $tH(Q|P)$. 

Testing by betting: point null and point alternative (cont’ed)
To summarise what was known

- For testing a point null $P$ against point alternative $Q$, likelihood ratios are optimal per-round bets.
- The optimal wealth is the likelihood ratio process.
- The optimal rate of growth (exponent) of the wealth is exactly the KL divergence or relative entropy of $Q$ to $P$. 

What about composite nulls?

- Significant progress by Peter Grünwald and coauthors in two papers (“Safe testing” and “Universal reverse information projection and optimal e-statistics”).
To summarise what was known

- For testing a point null $P$ against point alternative $Q$, likelihood ratios are optimal per-round bets.
- The optimal wealth is the likelihood ratio process.
- The optimal rate of growth (exponent) of the wealth is exactly the KL divergence or relative entropy of $Q$ to $P$.

What about composite nulls?

- Significant progress by Peter Grünwald and coauthors in two papers (“Safe testing” and “Universal reverse information projection and optimal e-statistics”).
Our setting

- A composite null hypothesis $\mathcal{P}$ and a point alternative hypothesis $Q$.
- The data is either drawn from some $P \in \mathcal{P}$ (the null is true), or from $Q$ (the null is false).
- A valid bet is an “e-variable”, which is a $X \geq 0$ such that $\mathbb{E}_P[X] \leq 1$ for every $P \in \mathcal{P}$. Think of $X$ as being the multiplier of your wealth in each round of a multi-round game.
- Question: What is the optimal one-round bet $X^*$? Is it unique? Can one characterise/derive it?

Answer: It is the likelihood ratio of $Q$ to a special element $P^*$, which we call the Reverse Information Projection (RIPr).
Our setting

- A composite null hypothesis $\mathcal{P}$ and a point alternative hypothesis $Q$.
- The data is either drawn from some $P \in \mathcal{P}$ (the null is true), or from $Q$ (the null is false).
- A valid bet is an “e-variable”, which is a $X \geq 0$ such that $\mathbb{E}_P[X] \leq 1$ for every $P \in \mathcal{P}$. Think of $X$ as being the multiplier of your wealth in each round of a multi-round game.
- Question: What is the optimal one-round bet $X^*$? Is it unique? Can one characterise/derive it?
- Answer: It is the likelihood ratio of $Q$ to a special element $P^*$, which we call the Reverse Information Projection (RIPr).
Introducing $X^*$, the “numeraire” e-variable

Theorem
Under no assumptions on $\mathcal{P}$ and $Q$, there always exists a special e-variable (bet) $X^*$, which satisfies two properties:

1. $X^* \geq 0$ and $\mathbb{E}_P[X^*] \leq 1$, $P \in \mathcal{P}$ (e-variable or fair bet property).

2. $\mathbb{E}_Q[X/X^*] \leq 1$ for any e-variable $X$ (“numeraire property”).

Further, $X^*$ is unique up to Q-nullsets. In fact, $X^*$ is the numeraire if and only if it is log-optimal.
Introducing $X^*$, the “numeraire” e-variable

**Theorem**

Under no assumptions on $\mathcal{P}$ and $Q$, there always exists a special e-variable (bet) $X^*$, which satisfies two properties:

1. $X^* \geq 0$ and $\mathbb{E}_P[X^*] \leq 1$, $P \in \mathcal{P}$ (e-variable or fair bet property).
2. $\mathbb{E}_Q[X/X^*] \leq 1$ for any e-variable $X$ (“numeraire property”).

Further, $X^*$ is unique up to $Q$-nullsets. In fact, $X^*$ is the numeraire if and only if it is log-optimal.

**Proof technique:** Komlós-type lemma (Delbaen & Schachermayer) and observing that numéraire property corresponds to first-order condition of log-maximisation.
Finiteness of the numeraire

Theorem
The following conditions are equivalent:

1. The numeraire is $Q$-almost surely finite.
2. Every e-variable is $Q$-almost surely finite.
3. $Q$ is absolutely continuous with respect to $\mathcal{P}$.
4. The set of e-variables $\mathcal{E}$ is bounded in probability under $Q$. 

Proof technique: A decomposition result of Brannath and Schachermayer (1999). (See also Karatzas & Kardaras, 2007.)
Finiteness of the numeraire

Theorem
The following conditions are equivalent:

1. The numeraire is $Q$-almost surely finite.
2. Every $e$-variable is $Q$-almost surely finite.
3. $Q$ is absolutely continuous with respect to $P$.
4. The set of $e$-variables $\mathcal{E}$ is bounded in probability under $Q$.


(See also Karatzas & Kardaras, 2007.)
Introducing $P^*$, the reverse information projection

**Definition**

Define the measure $P^*$ by

$$\frac{dP^*}{dQ} = \frac{1}{X^*}.$$ 

- $P^*$ is not a probability measure in general, but a sub-probability measure, i.e., $P^*(\Omega) \leq 1$.
- $P^*$ lies in the bipolar of $\mathcal{P}$, which is defined as follows. The polar is

$$\mathcal{P}^\circ = \{X \geq 0 : \mathbb{E}_P[X] \leq 1, \text{ for all } P \in \mathcal{P}\} = \mathcal{E}. $$

The bipolar is

$$\mathcal{P}^{\circ \circ} = \{P \in M_+: \mathbb{E}_P[X] \leq 1 \text{ for all } X \in \mathcal{P}^\circ\},$$

which we also call “the effective null hypothesis”.
Relative entropy

**Simplifying assumption** in the rest of the talk: \( Q \) absolutely continuous w.r.t. \( P \), meaning that whenever \( P(A) = 0 \) for every \( P \in \mathcal{P} \), we also have \( Q(A) = 0 \). (Very weak assumption, just for presentation).

Recall that the KL divergence or relative entropy is defined as
\[
H(Q | P) = \mathbb{E}_Q \log \frac{dQ}{dP}
\]
if \( Q \) absolutely continuous wrt \( P \).

If \( P \) denotes the absolutely continuous part of \( P \) wrt \( Q \), we can rewrite
\[
H(Q | P) = \mathbb{E}_Q \left( -\log \frac{dP}{dQ} \right).
\]
**Relative entropy**

**Simplifying assumption** in the rest of the talk: \( Q \) absolutely continuous w.r.t. \( \mathcal{P} \), meaning that whenever \( P(A) = 0 \) for every \( P \in \mathcal{P} \), we also have \( Q(A) = 0 \). (Very weak assumption, just for presentation).

- Recall that the KL divergence or relative entropy is defined as

\[
H(Q|P) = \mathbb{E}_Q \left[ \log \frac{dQ}{dP} \right]
\]

if \( Q \) absolutely continuous wrt \( P \).

- If \( P^a \) denotes the absolutely continuous part of \( P \) wrt \( Q \), we can rewrite

\[
H(Q|P) = \mathbb{E}_Q \left[ -\log \frac{dP^a}{dQ} \right].
\]
Strong duality of \((X^*, P^*)\)

**Theorem**

Let \(P^*\) be an element of \(\mathcal{P}^\infty\) that is equivalent to \(Q\). TFAE:

1. \(P^*\) is the RIPr.
2. \(\mathbb{E}_Q \left[ \frac{dP^a}{dP^*} \right] \leq 1\) for all \(P \in \mathcal{P}^\infty\).
3. \(\mathbb{E}_Q \left[ \log \frac{dP^a}{dP^*} \right] \leq 0\) for all \(P \in \mathcal{P}^\infty\).
Strong duality of \((X^*, P^*)\)

**Theorem**

Let \(P^*\) be an element of \(\mathcal{P}^{\infty}\) that is equivalent to \(Q\). TFAE:

1. \(P^*\) is the RIPv.
2. \(E_Q \left[ \frac{dP^a}{dP^*} \right] \leq 1\) for all \(P \in \mathcal{P}^{\infty}\).
3. \(E_Q \left[ \log \frac{dP^a}{dP^*} \right] \leq 0\) for all \(P \in \mathcal{P}^{\infty}\).

If any of these hold and \(X^*\) is the numeraire, one has the strong duality relation

\[
E_Q[\log X^*] = \sup_{X \in \mathcal{E}} E_Q[\log X] = \inf_{P \in \mathcal{P}^{\infty}} H(Q \mid P) = H(Q \mid P^*),
\]

where some, and then all, of these quantities may be \(+\infty\).

\((E_Q[\log X] \text{ is understood as } -\infty \text{ whenever } E_Q[(\log X)^-] = \infty.)\)
Estimation of Mixture Models

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Qiang (Jonathan) Li

Dissertation Director: Andrew R. Barron

May 1999
4.2 A New Information Projection Theory

In our theory, we reverse the order of the arguments in the K-L divergence. An analogous information projection theory is obtained. Applications to maximum likelihood estimation require this reversal of the order in the K-L divergence. We build upon a theory of Bell and Cover [1980], who in a portfolio selection context developed the story under an assumption that a minimizer of $D(P\|Q), Q \in \mathcal{C}$ exists.

Again we consider a convex set $\mathcal{C}$ of probability measures. Let $P$ be a probability measure of our interest. Define

$$D(P\|\mathcal{C}) = \inf_{Q \in \mathcal{C}} D(P\|Q).$$

Similar to T-C theory, we also want to establish existence, uniqueness and characterizing Pythagorean Identity of a projection $P^*$ of $P$ onto $\mathcal{C}$.

**DEFINITION 4.2 (Reversed Information Projection)** Given a probability measure $P$ with a density $p$ and a convex set $\mathcal{C}$ of densities $q$, a function $q^*$ is called the (reversed) information projection if for every $q_n$ with $D(p\|q_n) \to D(p\|\mathcal{C})$, we have $\log q_n \to \log q^*$ in $L^1(P)$.

**THEOREM 4.3 (Properties of the Reversed I-Projection)** Let $\mathcal{C}$ be a convex set of probability measures $Q$ with densities $q$ and let $P$ be a target measure with density $p$. Then the reversed I-projection $q^*$ of $P$ exists and is unique. Moreover it satisfies the following properties:

1. $D(p\|q^*) = \inf_{Q \in \mathcal{C}} D(p\|Q),$

2. $c_q = \int p \frac{q}{q^*} \leq 1, \forall q \in \mathcal{C},$

3. $D(p\|q) \geq D(p\|q^*) + D(p\|\rho)$ where $\rho = \frac{q^*}{c_q}$ is a density depending on $q$. 

Our theory recovers these as a special case
Our theory avoids all these assumptions.
A bit more about the bipolar $\mathcal{P}^{\circ \circ}$.

Let $M_1$ denote the set of all probability measures.

Let $M_+$ denote the set of all nonnegative measures.

A set $C \subset M_+$ is called “solid” if, for every $P \in C$ we also have $P' \in C$ whenever $P' \leq P$.

**Special cases**

- If $\mathcal{P}$ is finite, then $\mathcal{P}^{\circ \circ} \cap M_1 = \text{conv}(\mathcal{P})$.
- If a reference measure $\mu$ exists for $\mathcal{P}$, then every $P \in \mathcal{P}^{\circ \circ}$ is also absolutely continuous wrt $\mu$, and $\mathcal{P}^{\circ \circ}$ is the smallest $\mu$-closed solid convex set that contains $\mathcal{P}$. 
A bit more about \((X^*, P^*)\)

Recall

\[
\frac{dP^*}{dQ} = \frac{1}{X^*}.
\]

**Theorem**

Let \(X\) be an e-variable that is \(Q\)-almost surely strictly positive and let \(P\) be given by \(dP/dQ = 1/X\). Then \(X\) is the numeraire if and only if \(P\) belongs to the effective null \(\mathcal{P}^{\circ\circ}\). Thus the numeraire is the only e-variable that is also a likelihood ratio between \(Q\) and some equivalent element of \(\mathcal{P}^{\circ\circ}\).
Example 1: symmetric distributions

- Let $Z$ denote the data, here $\mathbb{R}$-valued.
- Assume that

$$\mathcal{P} = \{ P \in M_1 : Z \text{ and } -Z \text{ have the same distribution under } P \}.$$  

- Note that $\mathcal{P}$ has no dominating reference measure.
- Suppose $Q$ has a Lebesgue density $q$ (just for presentation).
- Older theory does not apply in this case.
- RIPr is a sub-probability measure with density

$$p^*(z) = \frac{1}{2} (q(z) + q(-z)) 1_{\{q(z) > 0\}}.$$  

- It is a probability density iff $Q$ has symmetric support.
- Numeraire is

$$X^* = \frac{\frac{dQ}{dP^*}}{q(Z) + q(-Z)} = \frac{2q(Z)}{q(Z) + q(-Z)}.$$
Example 2: 1-subGaussian

Let $Z$ denote the data again, $\mathbb{R}$-valued.

Assume that

$$\mathcal{P} = \left\{ P \in M_1 : \mathbb{E}_P[e^{\lambda Z} - \lambda^2/2] \leq 1 \text{ for all } \lambda \in [0, \infty) \right\}.$$ 

Above condition implies that $\mathbb{E}_P[Z] \leq 0$ for all $P \in \mathcal{P}$.

Suppose $Q = \mathcal{N}(\mu, 1)$ for some known mean $\mu$.

Older theory does not apply in this case (no dominating measure exists).

RIPr is $\mathcal{N}(0, 1)$ and numeraire is

$$X^* = e^{\mu Z} - \mu^2/2.$$
Example 3: a parametric example from Lardy et al.

- $\mathcal{P}$ consists of two unit-variance Gaussians $P_1$ and $P_2$ with mean $+1$ and $-1$.
- $Q$ is the standard Cauchy distribution.
- The RIPr is $P^* = (P_1 + P_2)/2$. The numeraire is
  \[ X^* = \frac{2q(Z)}{p_1(Z) + p_2(Z)} \]
- Can be generalised to arbitrary symmetric alternative $Q$. 
Summary

- We have fully generalised Kelly betting to composite nulls and point alternatives, yielding a strong duality between \((X^*, P^*)\).
- We have defined the reverse information projection \(P^*\) (RIPr) and the the optimal e-variable \(X^*\) (numeraire) without any assumptions.
- We showed how to apply this theory to new nonparametric settings that were previously out of reach.

Lots of work to be done ...
Thank you!

arXiv:2402.18810