THE UNIVERSITY OF CHICAGO

LOG, STOCK AND TWO SIMPLE LOTTERIES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF ECONOMICS

BY
SERGIY VERSTYUK

CHICAGO, ILLINOIS
DECEMBER 2017

The limits of my language mean the limits of my world.

—Ludwig Wittgenstein

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Pietro Veronesi, George M. Constantinides, Doron Ravid, Matt Taddy; as well as Christian Julliard, Leonid Kogan, Nicholas Polson, Xiao Qiao, Philip J. Reny, Lawrence D. W. Schmidt, Harald Uhlig.

ABSTRACT

This work studies the problem of decision-making under risk by agents whose information processing abilities may be limited. The constructed theoretical framework grounds on findings from economic laboratory experiments, incorporates existing neuroscience knowledge, and is implemented using information-theoretic formalism. Activation of the above information-processing constraints distorts the subjective perception of the objective stochastic environment the agent operates in, and the constrained-optimal decision-making requires appropriate adjustments. In the selected application, a general equilibrium macro-finance model, such biases of subjective perspective as overconfidence, pessimism and categorization thus emerge endogenously. The theoretical implications receive empirical support in a mutually consistent way: according to (cross-checked) calibrations, they allow us to reverse-engineer and rationalize the phenomena known as the equity premium/risk-free rate puzzles; as well as contribute to the understanding of such regularities as the portfolio underdiversification puzzle, style investing and the non-monotone pricing kernel puzzle. On the other hand, these results also help rationalize, by formulating certain optimizing foundations behind, the experimental evidence that underlies the Allais paradox and that is systematized in, e.g., the (cumulative) prospect theory.

CHAPTER 1

INTRODUCTION

This work takes the issue of computational feasibility seriously, and studies the problem of investment under conditions of risk while explicitly accounting for the constraints imposed by the limited cognitive—or information processing—capabilities of decision-makers (but recognizing nevertheless rationality of the latter, and their endeavor to make the most of it). We ask how such limitations bias the subjective perspective on the objective stochastic reality, and how they are manifested in the ultimate investment decisions and observed market prices. For instance, how exactly the value of investment opportunities is assessed and acted upon.

The "complexity-reducing simplification" arising as a result of such constraints generally distorts the agent's perception of the stochastic environment he/she operates in. Applying these simplification mechanics to the problem of investment portfolio choice induces investors to substitute the "complex" original objective probability distribution defining the investment returns with its proxy: the subjective distribution characterized by relatively lower variance ("overconfidence") and decreased mean ("pessimism"). The degree of these biases in the investor's subjective perspective depends on the information processing capacity available to him. If the capacity is sufficiently large, the biases vanish and the subjective perspective coincides with the objective picture, bringing about the benchmark case presumed by conventional "rational expectations". However, even when the capacity limit takes its toll, but the perspective is biased in an optimal way, i.e., the simplification of the subjective distribution is implemented according to a rational adjustment procedure that we specify, then the effects of "overconfidence" and "pessimism" cancel each other out and resulting decisions do not systematically deviate from the optimal ones. That is, investment choices are approximately equal to their unconstrained counterparts and ultimate economic outcomes are close to those achieved under "rational expectations", yet less computational resources are utilized.

The presented framework receives empirical support, taking as a testing ground the so-called equity premium and risk-free rate "puzzles" at the intersection between macroeconomics and finance (Mehra and Prescott, 1985; Weil, 1989). The necessary prerequisite is to recognize that the subjective probability distribution may differ from the objective one, and that the observed sample probability distribution may fail to reflect the full complexity of the unobserved population distribution. Then, we can infer the unobserved objective distribution indirectly by calibrating the magnitude of the information processing capacity so that the implied putative population distribution would in turn imply that the model does fit all the relevant empirical criteria. As a result, the constructed consumption-based asset pricing model matches the equity premium and risk-free rate observed in the post-World-War-II United States relying on highly plausible levels of relative risk-aversion (between 2 and 4). Additional empirical yardsticks represented by various moments of consumption or dividend growth data are not far off either. This is consistent with the interpretation that investors indeed behave rationally and are sufficiently efficient in processing information, even accounting for potential cognitive limitations. While the premium investors seem to leave on the table is likely to be just an illusion of perspective.

For the above results to be convincing, the calibrated magnitude of information processing capacity should be disciplined somehow, for instance, cross-checked with the relevant measurements coming from alternative sources. Extensive laboratory experiments with human subjects on the tasks involving decision-making under risk crystallized, for instance, in (cumulative) prospect theory of Kahneman and Tversky (1979, 1992) allow, under the imperative of their subjects' rationality, to identify the demonstrated magnitudes of information processing capacity, providing such independent checkpoint for comparison. Reassuringly, feeding the model with measurements informed by economic experiments produces results similar to those obtained in the calibrations presented above, which furnishes the desired confirmation. Our suggested interpretation of the experimental data is that Allais' (1953) paradox as well as Kahneman and Tversky's theory, specifically the probability weighting

transformations it stipulates, embody rough heuristics that effectively implement an alternative adjustment procedure to go together with the complexity-reducing simplification required in a challenging stochastic environment (though other interpretations are conceivable, see the main text for elaboration).

Establishment of these interconnections also has interesting theoretical implications. The connection between the current dissertation's framework and the experimental results of Kahneman and Tversky is not just a matter of consistency with existing empirical evidence, but in fact suggests optimizing foundations for (some aspects of) the purely descriptive prospect theory that Kahneman and Tversky have formulated on the basis of their laboratory findings. Additionally, the distinction—as well as the fundamental connection—between complex objective and simplified subjective probability distributions reconciles the allegedly mutually inconsistent normative theoretical desiderata of standard von Neumann-Morgenstern axioms on the one hand (which, in our framework, apply to the former distribution) with positive empirical obervations of decision-making under risk on the other hand (which correspond to the latter distribution), thus killing two birds with one stone.

Continuing the list of useful theoretical contributions, within the presented framework endogenously emerge such phenomena as "overconfidence" (i.e., an agent's subjective variance is lower than the objective one), "pessimism" (subjective mean is lower than objective) and "categorization" (correlations are subjectively amplified, hence similar assets are clustered together into one asset class). All being popular mechanisms in theoretical work, they are usually imposed exogenously rather than explicitly derived from first principles.[1]

The empirical picture painted so far renders only the macro-level landscape, but we may well go beyond the broad brushstrokes and zoom deeper into micro-level features. For instance, the principal approach of the current work by means of the above categorization

1. Finishing the list of theoretical results, the take-aways of a relatively more auxiliary nature and technical flavor include a convenient method for description/encoding of probability distributions (see §G); as well as an introduction of the distinction between effective and physical information processing capacities, which explains some confusing empirical measurements (see §G, §I and §D).

and overconfidence mechanisms contributes to a unified understanding of such empirical regularities as the "portfolio underdiversification puzzle", (variations in) style investing, and the "non-monotone pricing kernel puzzle" (or the implied volatility "smile") by offering their rational structural explanations that are quantitatively consistent with the corresponding results of non-structural estimation exercises.

In the current work, information processing is formally modeled as information transmission between (or, equivalently, information storage to/retrieval from) different parts of the brain. Evaluating the merits of available decisions, e.g., by calculating the expected value of a lottery, requires sending a message that describes the corresponding probability distribution from one, perceptive part of agent's brain to another, calculating part, and computing the statistic of interest. However, if the probability distribution is "too complex" relative to the agent's information processing limitations (his individual "bandwidth"), the procedure becomes infeasible. In such a case the distribution has to be "simplified": the decision-maker chooses another probability distribution that roughly approximates the original one but possesses lower "complexity", and uses this subjective simplified distribution in computations of the relevant statistic.

Resorting to technical jargon, the complexity of the distribution is formally quantified by its "Shannon entropy" (a certain measure of the distribution's dispersion), the complexity of the message discribing it is proportional to the complexity of the distribution itself (due to the entropy's relation to the length of efficient communication/compression code); the message is transmitted via the communication channel, and the costs of information transmission are formalized by the capacity limit of this communication channel as measured by the "mutual information" between channel input and output (a specific constraint imposed on the relationship between entropies of ingoing and outgoing messages). Thus, a probability distribution is "too complex" if, loosely speaking, its entropy is larger than the channel capacity.

Before proceeding, it is worth clarifying that the theory of "rational inattention" proposed

by Sims (1998, 2003, 2006) also uses the mutual information constraint to formalize the costs of information transmission, but there are fundamental differences with the approach taken here. Conceptually, Sims's theory restricts the agent's external perceptions and focuses on uncertainty about the current state, while this work deals with constraints on the agent's internal cognition and centers on uncertainty about the future state. Operationally, in rational inattention models the state has already realized but is not yet observed, and the agent, before observing the state and making a decision, chooses an optimal information acquisition strategy that lets him learn the realized state as accurately as his limited capacity allows, which in turn implies that every period nature is sending to an agent the information about one realization of a random variable. While in the present work the state has not yet realized, and the agent, before making a decision and before realization of the state, chooses an optimal information processing strategy that lets him summarize/represent the space of possible states as accurately as his limited capacity allows, which in turn implies that every period one part of agent's brain is sending to another part of agent's brain the information about the whole probability distribution of a random variable. Thus, the differences are in the timing of events, the object of approximation, the task faced by the agent, the players and interactions involved, as well as in the dimensionality of transmitted messages. See §H for more details about the difference.

Then, the "architecture" of our framework builds from the bottom up: taking simple lotteries as primitives (arguably, lotteries to students of choice under uncertainty are what fruit flies are to geneticists) and completing the construction with the tree model in the style of Lucas (1978) (a structural microfounded general equilibrium workhorse model for stocks in macroeconomics and finance).

The "mortar" binding this construction contains several key ingredients of methodological nature. The process of choice under uncertainty is implemented using information-theoretic formalism (see Appendix §G for details, also see §H). Empirically, it is grounded on experimental evidence: for instance, on the work of Gabaix and Laibson (2000), who study how

humans make decisions under risk, and whose findings can be interpreted as being consistent with their subjects resorting to complexity/entropy-reducing approximations; together with the results of Kahneman and Tversky (1979, 1992), who uncover a number of systematic biases in human decision-making under uncertainty, and whose findings may be interpreted as counteracting the deceptive effects of the above entropy-reducing approximations. It is also guided by existing knowledge in neuroscience, for example, explicitly incorporating the concepts such as "working memory" or "bottleneck", as well as the mechanisms such as coordinate transformations and adjustments or recursive processing (see Appendix §I).

Lastly, we touch upon some technical characteristics of the basic framework underlying the analysis. The (macro-)economic setup is deliberately primitive and conventional: a workhorse Lucas tree model, which prevents contamination of the analysis with details not crucial to the main narrative and aids tractability. The fundamental assumptions are very frugal: the strongest one is the log-Normality of the stock returns. The optimization problem is framed in a modular format: there are two segregated sub-problems, which improves transparency as well as facilitates potential modifications and extensions. The formal model admits analytical closed-form solutions: in potential applications, including the one pursued here, the solution is as tractable as the standard canonical case, in spite of being more general.

The present research is connected to several very distinct literature clusters, and their exhaustive review is beyond the scope of this dissertation (a succinct overview is offered in Appendix §D). The most closely related works are the following. Gabaix (2014a, 2014b) proposes a similar approach to the simplification of the environment, which is based on a discrete and sparse representation of data. The rational inattention theory (Sims, 2003, 2006; Matějka and Sims, 2010; Matějka and McKay, 2014; Ravid, 2016) has successfully introduced information-theoretic methods into economics; it also uses the channel capacity (mutual information) constraint to model the transmission of information, although in a structurally different form and with different aims (see the text above, as well as Appendix

§H). Woodford (2012, 2014) departs from the rational inattention theory and uses the channel capacity mechanism while taking related neuroscientific and experimental evidence very seriously, which makes his papers quite closely connected to this work. Steiner and Stewart (2016) formulate a model that is also neurologically motivated, albeit one focusing on the effect of noise (rather than complexity) in processing information, and which, in special circumstances of prevailing negatively-skewed lotteries, similarly finds Kahneman and Tversky's prospect theory-style probability weighting as optimal second-bests. As to the empirical applications, this dissertation primarily focuses on macro-finance, it is particularly close to the asset pricing literature that deals with tail events and large-variance distributions such as Weitzman (2007) or Tsai and Wachter (2016) (also see Veronesi, 2004; and separately, Polkovnichenko, 2005; Polkovnichenko and Zhao, 2013). Other works in this area similarly concerned with the expectation formation process and likewise motivated by existing psychological evidence along with the agents' desire for simplification are Fuster et al. (2012) and Bordalo et al. (2016) (also see an older work by Carroll, 2003; as well as a recent survey by Manski, 2017). Authors here who also emphasize the distinction between the investor's and the econometrician's perspectives are Hansen and Richard (1987), Hansen (2007), Hansen and Sargent (2010), Weitzman (2007), as well as Gabaix and Laibson (2001). Our main differences lie along the dimensions of the area of application, conceptual and methodological approach, parsimony and tractability, experimental and neuroscientific inputs employed.

CHAPTER 2

THEORY

2.1 Simple lotteries, hard decisions

The aim here is to examine how decisions under risk[1] are made in a setting where the relevant probability distributions are "too complex" to be used in the agent's optimization process.

As a motivating example, consider the following decision problem. Figure 2.1 presents the payoffs and their corresponding probabilities for two simple lotteries. In principle, from the information given one can calculate all the necessary characteristics of the payoff distributions. For instance, a risk-neutral player cares only about the mean, which is just a probability-weighted sum of each lottery's payoffs — a very simple computation. A player with a mean-variance utility function cares about the first two moments of the payoff distributions — a slightly more involved computation. But what if these arithmetic operations, however elementary they are, can not be performed in full? Say, because of a time constraint: think of having only 1 second per contingency, i.e., 16 seconds to choose between the two lotteries (in case you tried it, see the footnote[2]).

Formally, the problem looks as follows:

$$\max_{\theta \in \{1,2\}} \mathrm{E}\left[\varphi(\varrho(\boldsymbol{x}|\theta))\right], \tag{2.1}$$

where the state vector $\boldsymbol{x}$ is idiosyncratically distributed as $\boldsymbol{x} \sim G(\boldsymbol{x})$, the lottery choice $\theta \in \{1, 2\}$, the payoff function $\varrho(\boldsymbol{x}|\theta)$ depends on the chosen lottery, and $\varphi(\varrho)$ is a player's decision criterion (i.e., some kind of felicity function). The player sees all the payoffs as well as the probabilities and has to estimate the expected payoff (or, strictly speaking, expected

1. In this work, we deal only with the notion of risk, which defines a situation when probabilities associated with a random variable are known, and only the actual random variable realizations are not known. We ignore the notion of Knightian uncertainty, in which even the probabilities themselves are unknown. The terms "risk" and "uncertainty" are thus used interchangeably throughout.

2. The lottery on the right has higher mean (5.0 vs. 4.9) and lower variance (14.5 vs. 14.6).

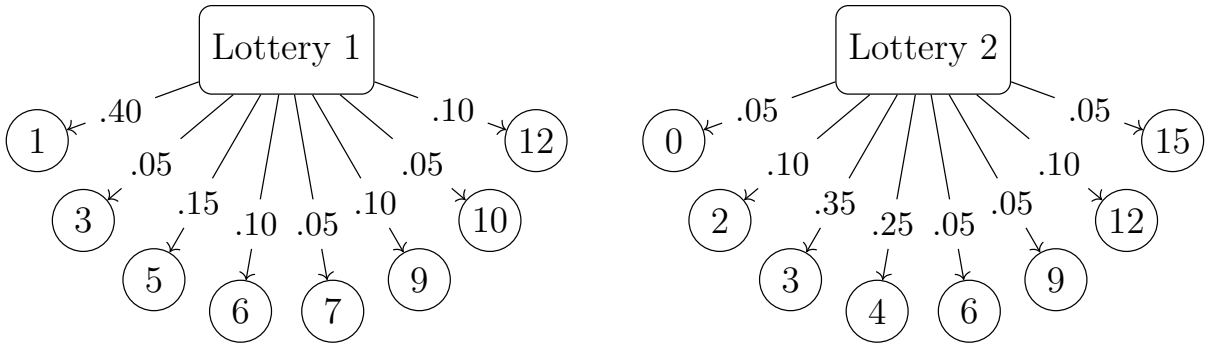


Figure 2.1: Two lotteries.

felicity) under the time constraint.

Gabaix and Laibson (2000) investigate a more complicated version of the same problem in laboratory conditions. They use multinomial recombining trees, each consisting of 10 root nodes and 4 to 9 levels of leaf nodes that are connected by probabilistic edges, with intermediate payoffs in every node; the goal is to select a root node with the highest expected value. Their paper proposes the following heuristic for evaluating the tree payoffs: consider only transit probabilities larger than a certain cutoff probability and calculate the expected payoff ignoring the less probable edges (it is required that the payoffs have a zero mean, and there are no extreme outlier payoffs). The authors interpret this decision rule as simulating the future by identifying typical or representative scenarios. They conduct an experiment where human subjects have to evaluate 12 such trees within 40 minutes, and the proposed algorithm most accurately matches the empirical distribution of choices, in particular, outperforming the fully rational model of behavior.

Effectively, the above method reduces the computational costs of evaluating the expectation in (2.1) by carefully changing the distribution of the random variable in focus to one with lower "uncertainty" (or entropy) at the cost of some "approximation error".

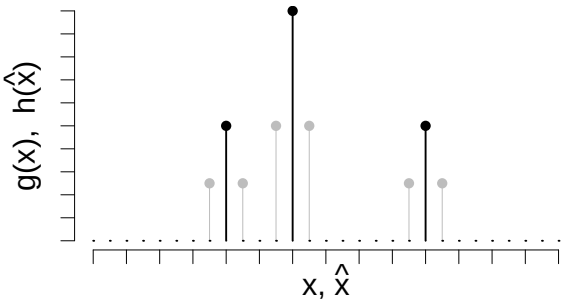


Figure 2.2: Entropy reduction primer
(higher-entropy PMF $g(\cdot)$ (in grey) vs. lower-entropy $h(\cdot)$ (in black)).

2.2 Algorithm for decision-making under risk

Our next task is to formalize the idea that the problem of decisions under risk can be simplified via entropy reduction. Appendix §G outlines the algorithm in detail. Here (as well as in §H) we provide an intuitive overview of its key features.

Figure 2.2 presents the probability mass function (PMF) $g(\cdot)$ for some random variable of interest x, that is a state variable such as next year's real GDP growth rate or a company's revenues. It is rather "complex", with "fine" partitions that allow for "many" (6 here) possible outcomes. In real life, decision-makers usually reduce the space of outcomes to a refined ensemble of just a few possible scenarios: say, a "baseline", "optimistic" and "pessimistic". Such a reduction makes the problem easier to analyze as well as to present and discuss it in business meetings, conference calls, or in written communication. In Figure 2.2 this is represented by PMF $h(\cdot)$, which is for convenience defined over a new synthetic random variable $\hat{x}$. This new distribution is relatively "simple", with "coarse" partitions that allow for just a "few" (3 here) possible outcomes.

The complexity (or uncertainty, dispersion) of a random variable can be measured by its entropy. Formally, (Shannon) entropy of a discretely distributed random variable x is defined as

$$\mathcal{E}(g(x)) := -\sum_{i=1}^{|\text{supp}(g)|} g(x_i) \log g(x_i), \tag{2.2}$$

where $|\text{supp}(g)|$ is the cardinality of the support set of $g(\cdot)$.[3,4] Units of measurement are either *bits* (when the logarithms of base 2 are used), or *nats* (for natural logarithms), with 1 nat $:= \log_2 e \approx 1.44$ bits.

For the discrete distributions primer presented in Figure 2.2, the "complex" distribution $g(\cdot)$ is characterized by higher entropy than the "simple" distribution $h(\cdot)$, so that replacing one with the other in the process of simplification leads to a reduction in the entropy measure. When the space of outcomes is continuous, the so-called differential entropy just invokes integration in place of summation in the above formula (2.2). The entropy reduction can be thought of as a reduction of dispersion, which in the case of continuous distributions is associated with a lower variance and the thinner tails of a distribution. (At first pass, it may seem counterintuitive that such a simplifying approximation unambiguously leads to a variance reduction, rather than the perturbations potentially going either way. The reasons are of the technical, information/measure-theoretic nature, and should be clear from the detailed exposition in Appendix §G. However, the unambiguous direction of the change in the variance is an important feature of the landscape indeed, as demonstrated in the rest of this work.)

In practice, such a simplification can be embodied in a subjective amalgamation of several low-probability events, like a serious epidemic and a major volcano eruption, into a single "disaster" event with the probability doubled and the effect averaged, which necessarily amounts to a reduction in entropy (as illustrated in Figure 2.2). But perhaps even more importantly, the fact that our statistical samples are finite means that the extreme tail events are likely to be "undersampled" with their (sample, and in the end those used in decision-making) probabilities equated to zero, which again mechanically entails entropy reduction. As is the case with regard to the chances of an expropriation of private property in the U.K. and the U.S., or maybe Germany and Japan, in the period after World War II,

3. We use $\log(\cdot)$ for $\log_2(\cdot)$ and $\ln(\cdot)$ for $\log_e(\cdot)$ throughout.

4. Shannon entropy is a concept originating in the information theory. For a textbook treatment of the information theory, see Cover and Thomas (2006) or MacKay (2003).

even though the latter two countries themselves experienced such an event right before the start of the indicated time period.[5] A crucial point in our argument is that it pertains to both favorable and unfavorable events, thus leading to the reduction in the dispersion/variance even if the means are not affected (so this can be thought as a "rare-events" narrative as opposed to "rare-disasters" or "peso-problem" hypotheses).

While from a statistical perspective the entropy of a random variable is some measure of its dispersion, there is another side to the coin. From the information-theoretic perspective, (Shannon) entropy of a random variable is the average length of code that can be used to efficiently carry information about the variable's outcomes. For example, the discrete distribution $h(\hat{x})$ presented in Figure 2.2 envisages three possible realizations with probabilities $\{0.25, 0.5, 0.25\}$. According to equation (2.2), this implies the entropy of 1.5 bits, which in turn means that the average codeword length is the same 1.5 bits. Indeed, such a binary alphabet would be $\{00, 1, 01\}$ for the corresponding realizations of the random variable $\hat{x}$ (with codeword "1" used on average twice more frequently than the other two). This would be a more compressed representation than what could be achieved for the higher-entropy distribution $g(x)$ that entails six different realizations with probabilities $\{0.125, 0.125, 0.25, 0.25, 0.125, 0.125\}$, which implies the entropy as well as the average codeword length of 2.5 bits on the basis of alphabet $\{000, 001, 10, 11, 010, 011\}$. Moreover, the above allows to deduce that the length of the code that summarizes the distribution itself is proportional to its entropy.

The instrumental value of this dispersion-code duality is that such a code permits us to (constructively) quantify the demands on the computational or, more broadly, information processing capacity. Consider a simple lottery represented again by distribution $g(\cdot)$ or $h(\cdot)$ from Figure 2.2, whose numerical payoff probabilities are listed in the preceding paragraph, and with some arbitrary payoff values, e.g., as those in Appendix §G (or §H).

5. In the Bayesian approach the "undersampling" issue is usually addressed by introducing additional prior information, but problems not less challenging emerge, see Weitzman (2007).

Intuitively, one can think of the task of evaluating this lottery as (a) learning and importing the complex probability distribution $g(\cdot)$ into the external-perceptive part of the brain; (b) drawing realizations from it in a Monte Carlo-type experiment; and (c) transmitting these values via a communication channel to the internal-cognitive part of the brain that calculates the statistics of interest using the transmitted random draws. A communication channel that has a limited capacity would require shorter codes, which are made possible by the simple distribution $h(\cdot)$. As the number of draws increases, the simulated statistics converge to their exact theoretical values, albeit those corresponding to distribution $h(\cdot)$ rather than $g(\cdot)$. Below we provide a slightly more technical account of the mechanisms involved, but skipping to part §2.3 should not preclude understanding the rest of the work.

Algorithm: Structurally speaking, in our framework the process of mental evaluation of a lottery comprises (i) the summarization of the given information about the lottery distribution (via a communication channel transmitting such a description), (ii) the loading and providing access to this information within a working memory (allocated in the memory storage), (iii) the calculation of lottery's value (via a communication channel transmitting intermediate iterations of such computations). Potential information processing "bottlenecks" may arise at any of the three milestones above, depending on which of the capacity constraints is binding: description channel capacity, storage memory capacity, or computation channel capacity. Such bottlenecks on the way of information flow can preclude the procedure's smooth completion that is necessary for making an optimal decision, with complex distribution $g(\cdot)$ being more susceptible to this predicament than simple distribution $h(\cdot)$. Figure 2.3 offers a sneak preview of the formal algorithm's mechanics at this particular juncture.

The leading illustration is probably memory and its capacity, which is relatively well studied in neuroscience (see Appendix §I) as well as recognized as an important concept in the economic literature (see Appendix §D for examples). Specifically, cognitive psychol-

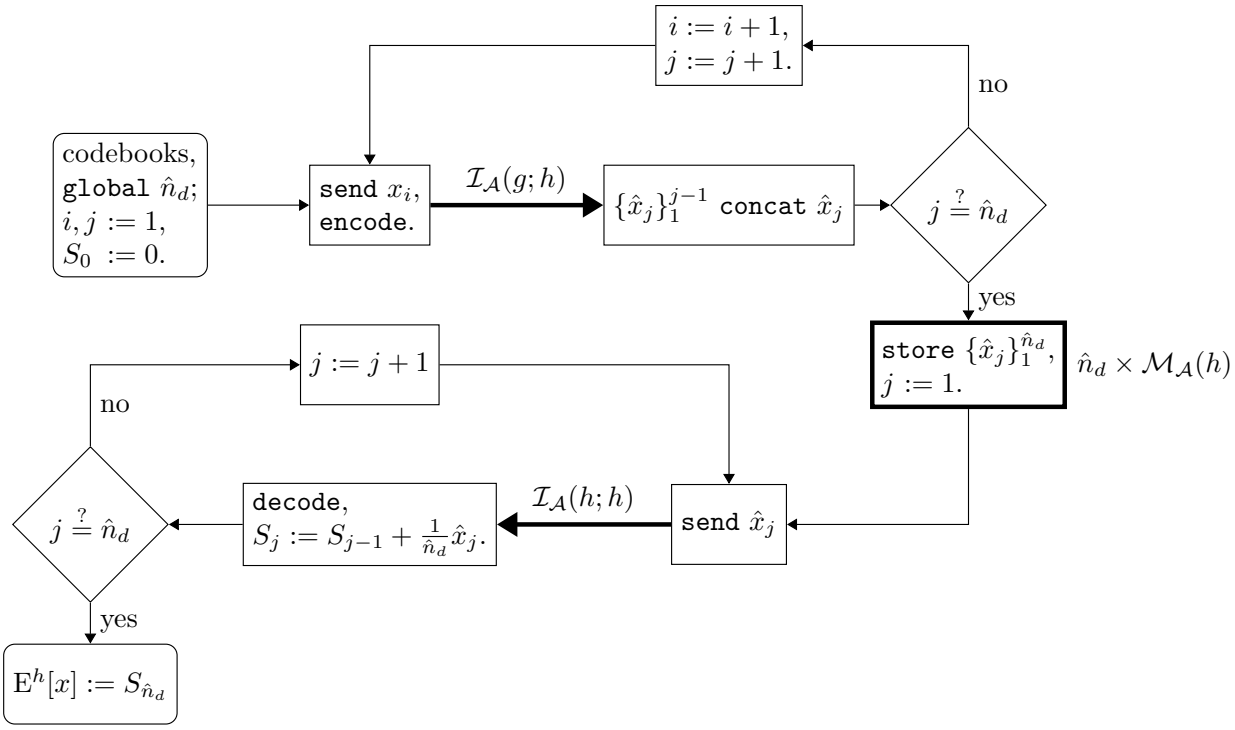


Figure 2.3: Information processing algorithm primer (potential bottlenecks shown in bold).

ogy and neuroscience define working memory as a limited capacity system that temporarily maintains and stores information to support human thought processes by providing an interface between perception, long-term memory and action (Baddeley, 2003). A probability distribution that is characterized by high entropy may require too complex a code to represent it, which puts unrealistic demands on the working memory's capacity and prohibits the utilization of a given distribution in mental information processing. Thus, appealing to neuroscience permits us to (structurally) motivate the above quantification of information-processing costs.

The exposition in Appendix §G makes the above statements formal and expands them, demonstrating how general the postulated framework is (not only providing for alternative information processing "bottlenecks", but also allowing for alternative deterministic or probabilistic Monte Carlo-type information processing schemes). In terms of functionality, the algorithm in Appendix §G starts with an evaluation of a simple lottery (formalized as ancillary procedure $\mathcal{P}_{\int}$) and at the end provides a recipe for the evaluation of expectations

of arbitrary functions of random variables that may be required for decision-making under risk.

2.3 Information constraint and informational problem

Now we are going to introduce an important object for subsequent exploration. Available effective capacity (from now on, information processing capacity) κ serves as a bound in the information processing capacity/mutual information constraint (in short, information constraint):

$$\kappa \geq \mathcal{I}(g(\boldsymbol{x}); h(\hat{\boldsymbol{x}})) = \mathcal{E}(g(\boldsymbol{x})) - \mathcal{E}(f(\boldsymbol{x}|\hat{\boldsymbol{x}})) = \mathcal{E}(h(\hat{\boldsymbol{x}})) - \mathcal{E}(f(\hat{\boldsymbol{x}}|\boldsymbol{x})) =$$
$$= \mathcal{E}(g(\boldsymbol{x})) + \mathcal{E}(h(\hat{\boldsymbol{x}})) - \mathcal{E}(f(\boldsymbol{x}, \hat{\boldsymbol{x}})) \tag{2.3}$$

(note the switch to random vectors, typed in boldface).

Information constraint is the instrumental take-away of the abstract algorithm presented in Appendix §G as well as in part §2.2 above. It quantifies and puts a bound on the costs (in bits of information processing capacity) to utilize the probability density $h(\hat{\boldsymbol{x}})$, for example, to calculate its expected value in the process of solving a maximization problem. The above density is a simplified proxy for a given original probability density $g(\boldsymbol{x})$, where the copy can be exact if the original density is already simple enough. To clarify the notation, $f(\boldsymbol{x}, \hat{\boldsymbol{x}})$ above is an ancillary function that captures the overall structure of stochastic interrelationships in the forthcoming optimization problem; it is a joint multivariate probability density function of $\boldsymbol{x}$, which is distributed according to its marginal density $g(\boldsymbol{x})$, and of $\hat{\boldsymbol{x}}$, which in turn is distributed according to marginal density $h(\hat{\boldsymbol{x}})$.

Intuitively, with the complexity of a random variable $\boldsymbol{x}$ fixed, one can reduce the information-processing costs by letting $\boldsymbol{x}$ being only imperfectly represented by a variable $\hat{\boldsymbol{x}}$ and carry some random "approximation" errors via raising the "uncertainty"/entropy of the conditional distribution of $\boldsymbol{x}$ given $\hat{\boldsymbol{x}}$ (refer to the RHS of the first equality in the top row of

information constraint equation 2.3); e.g., in Figure 2.2 knowing the realization of $\hat{\boldsymbol{x}}$ leaves a half-half chance of guessing the realization of $\boldsymbol{x}$. Equivalently, if a random variable $\hat{\boldsymbol{x}}$ is simple enough, then there is no need for further entropy reduction, and mutual information may be equated with entropy of $\hat{\boldsymbol{x}}$ by letting the distribution of $\hat{\boldsymbol{x}}$ conditionally on $\boldsymbol{x}$ being degenerate with zero entropy, in the sense that a variable $\boldsymbol{x}$ contains all the information about a variable $\hat{\boldsymbol{x}}$ (the RHS of the second equality in the top row of equation 2.3); e.g., in Figure 2.2 knowing $\boldsymbol{x}$ makes $\hat{\boldsymbol{x}}$ a certainty.

Alternatively, information constraint (2.3) trades-off how fine [coarse] probability measure $h(\hat{\boldsymbol{x}})$ is (middle term of equation in the last row) versus how [in]accurate approximation of $\boldsymbol{x}$ by $\hat{\boldsymbol{x}}$ is (rightmost term), taking $g(\boldsymbol{x})$ (leftmost term) as given. Put differently, with $g(\boldsymbol{x})$ fixed, to satisfy the information constraint one can either (i) reduce the entropy of $h(\hat{\boldsymbol{x}})$ by making it a coarser probability measure, or (ii) increase the entropy of $f(\boldsymbol{x}, \hat{\boldsymbol{x}})$ by making $\hat{\boldsymbol{x}}$ a less accurate approximation of $\boldsymbol{x}$. Note that this is exactly what the heuristic of Gabaix and Laibson (2000) amounts to in the end: a simplified distribution characterized by lower "uncertainty"/entropy at the cost of some "approximation error". In addition to the experimental findings of Gabaix and Laibson (2000), such behavior is also in agreement with neuroscientific evidence on humans' categorical perception (Goldstone and Hendrickson, 2010; Fleming et al., 2013). It is also worth mentioning that the information constraint has the same form as in Sims (2003, 2006), but here it has a different motivation and interpretation (more on this in §1 and §H).

Henceforth, we refer to the original random variable $\boldsymbol{x}$ and its probability distribution $g(\boldsymbol{x})$ as to "true", objective, unconstrained entities, while labeling the simplified random variable $\hat{\boldsymbol{x}}$ and its distribution $h(\hat{\boldsymbol{x}})$ as "approximating", subjective, constrained counterparts to and versions of the former.

Now, let us formulate what we call the informational problem, $\mathcal{P}_{\mathcal{I}}$:

$$\min_{f(\boldsymbol{x},\hat{\boldsymbol{x}})} \mathrm{E}^{f}\left[d(\boldsymbol{x},\hat{\boldsymbol{x}})\right] = \int_{\mathrm{supp}(h)} \int_{\mathrm{supp}(g)} d(\boldsymbol{x},\hat{\boldsymbol{x}})\, f(\boldsymbol{x},\hat{\boldsymbol{x}})\, d\boldsymbol{x}\, d\hat{\boldsymbol{x}} \qquad \{\mathcal{P}_{\mathcal{I}}\}$$

subject to the information constraint

$$\mathcal{I}(g(\boldsymbol{x}); h(\hat{\boldsymbol{x}})) \leq \kappa, \qquad [\lambda]$$

as well as the necessary technical restrictions (hereafter assumed implicitly)

$$\int_{\text{supp}(h)} f(\boldsymbol{x}, \hat{\boldsymbol{x}})\, d\hat{\boldsymbol{x}} = g(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in \text{supp}(g), \qquad [\mu(\boldsymbol{x})]$$

$$f(\boldsymbol{x}, \hat{\boldsymbol{x}}) \geq 0 \qquad \forall \boldsymbol{x} \in \text{supp}(g), \hat{\boldsymbol{x}} \in \text{supp}(h). \qquad [\nu(\boldsymbol{x}, \hat{\boldsymbol{x}})]$$

An agent seeks to minimize the expected "distortion" from using the approximating distribution $h(\hat{\boldsymbol{x}})$ instead of the true distribution $g(\boldsymbol{x})$ subject to information constraint, i.e., given a bound $\kappa > 0$ on the mutual information between these two distributions (Lagrange multipliers on each constraint are specified on the right in square brackets). The distortion function $d(\boldsymbol{x}, \hat{\boldsymbol{x}})$ (i.e., a distance metric for its two arguments) is taken as given. (The choice of appropriate distortion function is problem-specific, we will discuss it later in part §3.2.)[6]

Solution to problem $\mathcal{P}_{\mathcal{I}}$ is provided in Proposition 1.

Proposition 1 (General Solution to Informational Problem)**.** *Let $\boldsymbol{x}$ be a random vector distributed according to an absolutely continuous probability distribution function $G(\boldsymbol{x})$ with a probability density function $g(\boldsymbol{x})$, $d(\boldsymbol{x}, \hat{\boldsymbol{x}})$ be a distortion function for vectors $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ satisfying the condition*

$$\exists\, \hat{\boldsymbol{x}} : \quad \int_{\text{supp}(g)} d(\boldsymbol{x}, \hat{\boldsymbol{x}})\, g(\boldsymbol{x})\, d\boldsymbol{x} < \infty.$$

Then solution to the informational problem specified in $\mathcal{P}_{\mathcal{I}}$ is given by a conditional probability

6. In terms of Appendix §G's algorithm, our understanding is that in practice the informational problem is solved before Generating codebook at the Simplification step of the algorithm. Basically, this is the fundamental source of information-processing cost savings that we ultimately benefit from: by bearing the fixed costs at the Simplification step, the variable costs are saved in the following steps of the algorithm, which may lead to dramatic overall savings as the latter costs accumulate very quickly during the numerous iterations required to execute the remaining steps. This rules out a kind of infinite regress critique.

density

$$f(\boldsymbol{x}|\hat{\boldsymbol{x}}) = \exp\left(\frac{1}{\lambda}\nu(\boldsymbol{x},\hat{\boldsymbol{x}}) - \frac{1}{\lambda}\mu(\boldsymbol{x}) - \frac{1}{\lambda}d(\boldsymbol{x},\hat{\boldsymbol{x}})\right), \qquad \forall \hat{\boldsymbol{x}} \in \text{supp}(h).$$

Proof. See Appendix §F.1.

□

The form of the solution may seem counterintuitive at first. However, the key is to realize that the conditional distribution of $\boldsymbol{x}$ given $\hat{\boldsymbol{x}}$ that it provides is actually the distribution of "approximation error", i.e., the deviation of the original random variable $\boldsymbol{x}$ from its simplified counterpart $\hat{\boldsymbol{x}}$. Given the knowledge of $g(\boldsymbol{x})$, it is often easier to proceed by just guessing the density $h(\hat{\boldsymbol{x}})$, which we are chiefly interested in, and then verifying that, together with the deduced conditional distribution, the implied joint density

$$f(\boldsymbol{x},\hat{\boldsymbol{x}}) := f(\boldsymbol{x}|\hat{\boldsymbol{x}})h(\hat{\boldsymbol{x}})$$

satisfies the necessary requirements. Later in §3.3 we will demonstrate how this can be done with an example.

Whenever the information constraint does not bind, the expected distortion can be reduced to zero, because then $f(\boldsymbol{x}|\hat{\boldsymbol{x}})$ becomes the Dirac delta function centered at $\hat{\boldsymbol{x}}$, $\delta(\boldsymbol{x}-\hat{\boldsymbol{x}})$, $\forall \boldsymbol{x} \in \text{supp}(g), \hat{\boldsymbol{x}} \in \text{supp}(h)$; also $\text{supp}(h) := \text{supp}(g)$; and $\hat{\boldsymbol{x}} = \boldsymbol{x}$, $\forall \boldsymbol{x} \in \text{supp}(g)$, almost surely; as well as $h(\boldsymbol{x}) = g(\boldsymbol{x})$, $\forall \boldsymbol{x} \in \text{supp}(g)$ (thereby inducing "rational expectations"). (More on this later.)

Lastly, we claim that the following property holds.

Proposition 2 (Flexible Mean Property)**.** *Problem $\mathcal{P}_{\mathcal{I}}$ always admits the solution such that $\mathrm{E}^h[\hat{\boldsymbol{x}}] + \check{\boldsymbol{\mu}} = \mathrm{E}^g[\boldsymbol{x}]$ (provided the latter exists) for any bias $\check{\boldsymbol{\mu}} \in \mathbb{G}$, where $\mathbb{G}$ is some sufficient (extension) field for* $\text{domain}(g)$*.*

Proof. Trivially, the information constraint restricts only the mutual information between

the random variables $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$, which depends on the range but not the domain of probability density functions.

□

Thus, Proposition 2 allows to treat the mean of the simplified random variable, $\mathrm{E}^h[\hat{\boldsymbol{x}}]$, as a free parameter in the formulation of problem $\mathcal{P}_{\mathcal{I}}$ and as some exogenously defined control in the corresponding solution of Proposition 1.

In the following part, we apply the presented theoretical constructs to a suitable model economy.

CHAPTER 3

MODEL

Arguably, the limitations on the complexity of the computations that agents are able to undertake, as well as the distortions in perception and the deviations of decisions from the optimal ones that may be induced by such limitations are germane to many situations with uncertain outcomes, and to investment choices in particular.

3.1 Investment portfolio choice problem

Economic setup: Consider the following economic setting, which is essentially a variation of Lucas (1978) tree model (also refer to Breeden, 1979). For an easier exposition, first we formulate a two-period model, and an infinite-horizon extension follows later.

A representative agent with a lifespan of two periods lives in an exchange economy with opportunities to invest competitively in 1 risk-free and K risky assets. Risky assets are composed of one-period-lived "trees". The unit prices and quantities of shares in the risky trees purchased in period t are denoted by $\boldsymbol{P}_t$ and $\boldsymbol{q}_t$, respectively. The investments in them bring stochastic dividends $\boldsymbol{D}_{t+1}$, or "fruits", at the beginning of period $(t+1)$. A K-sized random vector $\boldsymbol{D}_{t+1}$ is distributed, given $\boldsymbol{D}_t$, according to probability density function $g_D(\boldsymbol{D}_{t+1}|\boldsymbol{D}_t)$. The risk-free asset is also composed of a one-period-lived tree. The unit price and quantity of shares in the risk-free tree purchased in period t are denoted by $P_{0,t}$ and $q_{0,t}$, respectively. The investments in it bring deterministic dividends $D_{0,t+1}$, the same type of fruits as above, at the beginning of period $(t+1)$. A constant scalar $D_{0,t+1}$ is normalized to 1 for all periods t. The fruits are perishable, output can not be stored between periods. We denote by C_t the agent's time-t consumption, and by $u(C_t)$ his per-period utility function, assumed to have a constant relative risk-aversion form, that is discounted at subjective rate β. This endowment economy comprises $\widehat{\boldsymbol{q}}$, a strictly positive K-sized constant vector, of risky trees, whose shares are initially owned by the representative agent. A risk-free tree is

fictitious, the economy comprises $\widehat{q}_0 = 0$ of them, i.e., it exists in zero net supply and can be thought of as a cash credit technology.

Problem: The consumer-investor is interested in solving the following consumption and portfolio choice problem, $\mathcal{P}_q$:

$$\max_{C_t,\{q_{0,t},\boldsymbol{q}_t\}} \{u(C_s) + \beta \mathrm{E}_t^g\left[u(C_{t+1})\right]\} = \{u(C_s) + \beta \int_{\mathbb{R}_+^K} u(C_{t+1})\, g_D(\boldsymbol{D}_{t+1}|\boldsymbol{D}_t)\, d\boldsymbol{D}_{t+1}\} \;\; \{\mathcal{P}_q\}$$

subject to budget constraints

$$C_t + P_{0,t}q_{0,t} + \boldsymbol{P}_t^\intercal \boldsymbol{q}_t = q_{0,t-1} + (\boldsymbol{P}_t + \boldsymbol{D}_t)^\intercal \boldsymbol{q}_{t-1},$$
$$C_{t+1} = q_{0,t} + \boldsymbol{D}_{t+1}^\intercal \boldsymbol{q}_t,$$

control variables' domain restriction $C_t, \{q_{0,t}, \boldsymbol{q}_t\} \in \mathbb{R}_+ \times \mathbb{R}^{K+1}$, with $\{q_{0,t-1}, \boldsymbol{q}_{t-1}\}$ and $\boldsymbol{D}_t$ given, with $u(C_t) = C_t^{1-\gamma}/(1-\gamma)$, and where

$$g_D(\boldsymbol{D}_{t+1}|\boldsymbol{D}_t) \text{ is given.}$$

In words, the representative agent chooses consumption and investment values that maximize his current and expected future utility that at the same time satisfy the budget constraints. The expectation is taken with respect to a given objective probability density function that defines the distribution of the stochastic fruit-dividends (which are produced by tree-assets the agents invests in, and which are then eaten as consumption goods).

Infeasibility in general: Using the notation introduced previously in §2, finding optimal solution to problem $\mathcal{P}_q$ requires directly maximizing

$$\mathrm{E}_t^g\left[\varphi^\sharp(\boldsymbol{x}|\boldsymbol{\theta})\right] := u(W_t - \{P_{0,t}, \boldsymbol{P}_t\}^\intercal \boldsymbol{\theta}) + \beta \mathrm{E}_t^g\left[u([1\ \boldsymbol{x}^\intercal]\boldsymbol{\theta})\right],$$

with W_t and $\{P_{0,t}, \boldsymbol{P}_t\}$ known, as well as with $\boldsymbol{x} := \boldsymbol{D}_{t+1}$, and $\boldsymbol{\theta} := \{q_{0,t}, \boldsymbol{q}_t\}$.

Our consumer-investor is assumed to know (say, to have learned by time t) the structure of the problem, i.e., the exact specification of the utility function $u(\cdot)$, the felicity function $\varphi^\sharp(\boldsymbol{\theta}, \boldsymbol{x})$, and the (stationary) distribution $g(\boldsymbol{x})$. Thus, the environment in terms of stochasticity is as primitive as possible. Nevertheless, in general the problem $\mathcal{P}_q$ is infeasible to solve: potentially it violates the information processing capacity constraint. Even though the agent obtains $g(\cdot)$ as an input into the algorithm of Appendix §G, he may be unable to execute the full procedure.

Because of his limited information processing capacities—required to deal with random variables and, in particular, to form expectations about the functions of random variables—such an investor will necessarily have to use subjective "distorted" probability distributions instead of objective ones. The former are less costly, but leave room for certain discrepancies in the computations, and hence in the perceived landscape of the stochastic environment and the resulting investment decisions.

Therefore, the agent focuses instead on maximizing

$$\mathrm{E}_t^h \left[\varphi(\hat{\boldsymbol{x}}|\boldsymbol{\theta})\right] := u(W_t - \{P_{0,t}, \boldsymbol{P}_t\}^\intercal \boldsymbol{\theta}) + \beta \mathrm{E}_t^h \left[u([1 \ \hat{\boldsymbol{x}}^\intercal]\boldsymbol{\theta})\right],$$

where

$$h(\hat{\boldsymbol{x}}) \text{ solves } \mathcal{P}_\mathcal{I} \text{ given } d(\boldsymbol{x}, \hat{\boldsymbol{x}}) \text{ and } \kappa.$$

To start with, note that the non-stochastic time-t objects are unaffected. However, the latest formulation makes it clear that in solving the stochastic optimization problem, we are using the subjective probability density $h(\hat{\boldsymbol{x}})$ in place of the objective density $g(\boldsymbol{x})$, with the discrepancy between the two densities depending on the available information processing capacity κ. While the distortion function $d(\boldsymbol{x}, \hat{\boldsymbol{x}})$ is chosen to be just some reasonable measure of distance between $\varphi^\sharp(\boldsymbol{x}|\boldsymbol{\theta})$ and $\varphi(\hat{\boldsymbol{x}}|\boldsymbol{\theta})$. More explicit formulations can be found

in Appendix §J.[1]

3.2 Feasible investment portfolio choice problem

Problem: A feasible version of our consumption and investment problem would recognize that the information constraint may be binding, in contrast to the previous formulation that effectively assumes the constraint is not binding, or is "slack". That is, the consumption and portfolio choice problem $\mathcal{P}_q$ has to be combined with the informational problem $\mathcal{P}_{\mathcal{I}}$. This is done without loss of generality, but allows to relax the straitjacket discipline of the standard problem's restrictions. The solution to the informational problem would provide an optimal (with respect to the distortion metric used) probability density function $h(\cdot)$, with respect to which the investment problem will in turn be solved.

Thus, a feasible version of the consumption and portfolio choice problem, $\mathcal{P}_{q\mathcal{I}}$, is formulated as follows:

$$\max_{C_t,\{q_{0,t},\boldsymbol{q}_t\}} \{u(C_s)+\beta \mathrm{E}_t^h\left[u(C_{t+1})\right]\} = \{u(C_s)+\beta \int_{\mathbb{R}_+^K} u(C_{t+1})\, h_D(\hat{\boldsymbol{D}}_{t+1}|\hat{\boldsymbol{D}}_t)\, d\hat{\boldsymbol{D}}_{t+1}\} \;\; \{\mathcal{P}_{q\mathcal{I}}\}$$

subject to budget constraints

$$C_t + P_{0,t}q_{0,t} + \boldsymbol{P}_t^{\intercal}\boldsymbol{q}_t = q_{0,t-1} + (\boldsymbol{P}_t + \hat{\boldsymbol{D}}_t)^{\intercal}\boldsymbol{q}_{t-1},$$
$$C_{t+1} = q_{0,t} + \hat{\boldsymbol{D}}_{t+1}^{\intercal}\boldsymbol{q}_t,$$

control variables' domain restriction $C_t, \{q_{0,t}, \boldsymbol{q}_t\} \in \mathbb{R}_+ \times \mathbb{R}^{K+1}$, with $\{q_{0,t-1}, \boldsymbol{q}_{t-1}\}$ and $\hat{\boldsymbol{D}}_t$

1. Our understanding is that the solution to the informational problem $\mathcal{P}_{\mathcal{I}}$ has already been learned by time t (e.g., calculated and made available in some effective analogue of abstract formal symbolic-like representation) or is just computationally easy relative to the solution of the unconstrained problem $\mathcal{P}_q$.

given, with $u(C_t) = C_t^{1-\gamma}/(1-\gamma)$, and where

$$h_D(\hat{\boldsymbol{D}}_{t+1}|\hat{\boldsymbol{D}}_t) \text{ solves } \mathcal{P}_{\mathcal{I}} \text{ given } d(\boldsymbol{D}_{t+1}, \hat{\boldsymbol{D}}_{t+1}) \text{ and } \kappa,$$
$$g_D(\boldsymbol{D}_{t+1}|\boldsymbol{D}_t) \text{ is given.}$$

The crucial difference from before is that in the feasible formulation of the consumption and portfolio choice problem the expectation is now taken with respect to the endogenous subjective probability density function for stochastic fruit-dividends, which itself has to be obtained as an optimal solution to the auxiliary informational problem. Also, note that in time t, the subjective random variables coincide with their objective counterparts (almost surely), but we signify them with hats nevertheless to preserve notational succession across the time periods.

Extension to infinite horizon: The economic setting is modified so that a representative agent has an infinite lifespan. Risky assets are composed of infinitely-lived trees. They bring stochastic dividends $\boldsymbol{D}_{t+1}$ at the beginning of period $(t+1)$. A K-sized random vector $\boldsymbol{D}_s$ follows a time-homogeneous (stationary) Markov chain defined by transition probability density function $g_D(\boldsymbol{D}_{s+1}|\boldsymbol{D}_s) = g_D(\boldsymbol{D}_{t+1}|\boldsymbol{D}_t)$ for all periods $s > t$. The risk-free asset is still composed of a one-period-lived tree. Its deterministic dividends $D_{0,t+1}$ are normalized to 1 for all periods t.

The formulations in terms of a sequence problem for both the infeasible as well as the feasible versions of this dynamic programming problem are available in Appendix §A. The Bellman equation corresponding to the infinite-horizon feasible problem, $\mathcal{P}_{\mathcal{QI}}$, is

$$\{\mathcal{P}_{\mathcal{QI}}\}$$
$$v(\{q_{0,t-1}, \boldsymbol{q}_{t-1}\}, \hat{\boldsymbol{D}}_t) = \max_{C_t, \{q_{0,t}, \boldsymbol{q}_t\}} \left\{ u(C_t) + \beta \mathrm{E}_t^h \left[v(\{q_{0,t}, \boldsymbol{q}_t\}, \hat{\boldsymbol{D}}_{t+1}) \right] \right\} \qquad \text{(PQI-1)}$$

subject to

$$C_t + P_{0,t} q_{0,t} + \boldsymbol{P}_t^{\intercal} \boldsymbol{q}_t = q_{0,t-1} + (\boldsymbol{P}_t + \hat{\boldsymbol{D}}_t)^{\intercal} \boldsymbol{q}_{t-1}, \tag{PQI-2}$$

domain restriction $C_t, \{q_{0,t}, \boldsymbol{q}_t\} \in \mathbb{R}_+ \times \mathbb{R}^{K+1}$, with the same utility function specification, and where (spelling out the relationship to $\mathcal{P}_{\mathcal{I}}$ explicitly)

$$h_D(\hat{\boldsymbol{D}}_{t+1} | \hat{\boldsymbol{D}}_t) := \int_{\text{supp}(g_D)} f(\boldsymbol{D}_{t+1}, \hat{\boldsymbol{D}}_{t+1} | \boldsymbol{D}_t, \hat{\boldsymbol{D}}_t) \, d\boldsymbol{D}_{t+1}, \tag{PQI-3}$$

$$f_D(\boldsymbol{D}_{t+1}, \hat{\boldsymbol{D}}_{t+1} | \boldsymbol{D}_t, \hat{\boldsymbol{D}}_t) := \arg \Big\{ \min_{f(\cdot,\cdot)} \mathrm{E}^f \left[d(v^\sharp(\{q_{0,t}, \boldsymbol{q}_t\}, \boldsymbol{D}_{t+1}), v(\{q_{0,t}, \boldsymbol{q}_t\}, \hat{\boldsymbol{D}}_{t+1})) \right] \text{ s.t. } \mathcal{I}(g_D(\boldsymbol{D}_{t+1} | \boldsymbol{D}_t); h_D(\hat{\boldsymbol{D}}_{t+1} | \hat{\boldsymbol{D}}_t)) \leq \kappa \Big\}, \tag{PQI-4}$$

$$g_D(\boldsymbol{D}_{t+1} | \boldsymbol{D}_t) \text{ is given.} \tag{PQI-5}$$

Above we denote the maximum value function as $v^\sharp(\{q_{0,t-1}, \boldsymbol{q}_{t-1}\}, \boldsymbol{D}_t)$ for the infeasible/unconstrained case, and as $v(\{q_{0,t-1}, \boldsymbol{q}_{t-1}\}, \hat{\boldsymbol{D}}_t)$ for the feasible/constrained case. Using the notation introduced earlier, now $\varphi^\sharp(\boldsymbol{x}|\boldsymbol{\theta}) := v^\sharp(\boldsymbol{\theta}, \boldsymbol{x})$ and $\varphi(\hat{\boldsymbol{x}}|\boldsymbol{\theta}) := v(\boldsymbol{\theta}, \hat{\boldsymbol{x}})$. Because the objective function incorporates an additional constraint, for a given $\boldsymbol{\theta}$-parameter, $\varphi(\hat{\boldsymbol{x}}|\boldsymbol{\theta})$ lies weakly below the unconstrained $\varphi^\sharp(\boldsymbol{x}|\boldsymbol{\theta})$, as does $v(\boldsymbol{\theta}, \hat{\boldsymbol{x}})$ relative to $v^\sharp(\boldsymbol{\theta}, \boldsymbol{x})$.

The Bellman equation's formulation is standard except that the probability density function $h_D(\cdot)$ with respect to which it is defined stems from (PQI-4), the solution to auxiliary sub-problem $\mathcal{P}_{\mathcal{I}}$.

Lastly, note that the environment in terms of dynamics and stochasticity is as primitive as possible, the agent solves the same problem again and again.[2]

Appropriate distortion function: The solution to $\mathcal{P}_{\mathcal{I}}$, referred to in (PQI-4), requires choosing some appropriate distortion function $d(\cdot, \cdot)$. Specifying one that is both reasonable

2. Implicitly, in terms of Bayesian inference with regard to $\boldsymbol{x}$ and $g(\cdot)$, we are in a degenerate situation with a "flat", never-updated prior. The task of embedding non-trivial Bayesian updating into our framework, particularly into algorithm of Appendix §G, is outside the scope of this dissertation.

in terms of the objective of larger problem $\mathcal{P}_{\mathcal{QI}}$, and convenient to work with analytically, is a laborous task that we attend to next. Besides somewhat technical arguments, an important intermediate result we show below is a pessimistic downward adjustment of the mean of the approximating distribution, which is required in order to compensate for the latter's lower entropy (and variance).

Define the new variables, W_{t+1} (for wealth, cum dividend), $R_{0,t+1}$ and $\boldsymbol{R}_{t+1}$ (gross returns), as well as, to avoid any confusion, $y_{0,t}$ and $\boldsymbol{y}_t$ (nominal investments):

$$W_{t+1} := q_{0,t} + (\boldsymbol{P}_{t+1} + \boldsymbol{D}_{t+1})^{\intercal} \boldsymbol{q}_t =: \tag{3.1}$$

$$=: P_{0,t} R_{0,t+1} q_{0,t} + (\mathrm{diag}(\boldsymbol{P}_t)\, \boldsymbol{R}_{t+1})^{\intercal}\, \boldsymbol{q}_t =: \tag{3.2}$$

$$=: R_{0,t+1} y_{0,t} + \boldsymbol{R}_{t+1}^{\intercal} \boldsymbol{y}_t; \tag{3.3}$$

and similarly $\hat{W}_{t+1}$, $\hat{\boldsymbol{R}}_{t+1}$.

Now, restrict the stochastic processes considered to independent identically distributed random variables (rather than Markovian), that is $g_R(\boldsymbol{R}_{t+1}|\boldsymbol{R}_t) := g_R(\boldsymbol{R}_{t+1})$. Furthermore, assume the log-Normality of the returns, i.e., $g_R(\boldsymbol{R}_{t+1})$ is $\log\mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$; for log-returns $\boldsymbol{r}_{t+1} := \ln \boldsymbol{R}_{t+1}$ (also reserving $r_{0,t+1} := \ln R_{0,t+1}$) this means

$$g_r(\boldsymbol{r}_{t+1}) \text{ is } \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r) \tag{3.4}$$

(we will "reverse-engineer" later what $g_D(\boldsymbol{D}_{t+1}|\boldsymbol{D}_t)$ this requires).[3]

Realize that the value function has the following form:

$$v^{\sharp}(\{q_{0,t}, \boldsymbol{q}_t\}, \boldsymbol{D}_{t+1}) = A\,(W_{t+1})^{1-\gamma}\,, \tag{3.5}$$

where $A := (1-\beta)^{-\gamma}/(1-\gamma)$; and analogously $v(\{q_{0,t}, \boldsymbol{q}_t\}, \hat{\boldsymbol{D}}_{t+1})$.[4]

3. A parametric log-Normal probability distribution is assumed here for analytical convenience, it is just a theoretical proxy for some non-parametric distribution dealt with in practically relevant problems.

4. We abstract away from the special fact that in the general equilibrium of this particular exchange

Thus, a reasonable choice for the distortion function in the sense of L^2 norm (squared) would be $d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}) := ||v^\sharp(\{q_{0,t}, \boldsymbol{q}_t\}, \boldsymbol{D}_{t+1}) - v(\{q_{0,t}, \boldsymbol{q}_t\}, \hat{\boldsymbol{D}}_{t+1})||_2^2$. However, given the CRRA functional form, we modify it to

$$d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}) := \frac{1}{(1-\gamma)^2} || \ln |v^\sharp(\{q_{0,t}, \boldsymbol{q}_t\}, \boldsymbol{D}_{t+1})| - \ln |v(\{q_{0,t}, \boldsymbol{q}_t\}, \hat{\boldsymbol{D}}_{t+1})| \, ||_2^2 =$$
$$= (\ln W_{t+1} - \ln \hat{W}_{t+1})^2. \tag{3.6}$$

It is worth saying explicitly that a distortion function so formulated is appropriate with regard to its economic grounding (a well-defined distance measure in felicity terms), but it also reflects some discretion with regard to the norm (L^2, squared) and the transformation (logarithmic) used.

Next, we produce an adaptation of the above formulation of the distortion function that is, for our purposes, more convenient to use.

Proposition 3 (Distortion Function)**.** *The above distortion function, in the context of problem $\mathcal{P}_{\mathcal{QI}}$ and given the distributional assumptions, can be reformulated as follows:*

$$d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}) = (\ln W_{t+1} - \ln \hat{W}_{t+1})^2 \approx$$
$$\approx \left(\boldsymbol{\omega}_t^\intercal (\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r)\right)^2, \tag{P3-1}$$

where

$$\boldsymbol{\omega}_t := \mathrm{diag}(\boldsymbol{P}_t)\boldsymbol{q}_t \,/\, W_t$$

is a K-vector of the shares of wealth invested in risky assets, while the mean $\hat{\boldsymbol{\mu}}_r$ of the simplified random variable $\hat{\boldsymbol{r}}_{t+1}$ equals

$$\hat{\boldsymbol{\mu}}_r := \boldsymbol{\mu}_r + \check{\boldsymbol{\mu}}_r, \tag{P3-2}$$

economy, $v(\cdot)$ actually attains and equals $v^\sharp(\cdot)$ identically.

and where

$$\check{\boldsymbol{\mu}}_r := \frac{1}{2}\text{diag}^{-1}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r) - \frac{1}{2}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r)\boldsymbol{\omega}_t, \tag{P3-3}$$

is a bias term, with $\hat{\boldsymbol{\Sigma}}_r$ *denoting the variance-covariance matrix for* $\hat{\boldsymbol{r}}_{t+1}$*.*

Under an additional assumption about the timing of an update of vector $\boldsymbol{\omega}_t$ *(the requirement to minimize maximum loss), the following refinement can be made:*

$$\begin{aligned}\left(\boldsymbol{\omega}_t^{\mathsf{T}}(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r)\right)^2 \propto (\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r(\widehat{\omega}_t))^{\mathsf{T}}(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r(\widehat{\omega}_t)) =: \\ =: d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}),\end{aligned} \tag{P3-4}$$

where

$$\begin{aligned}\check{\boldsymbol{\mu}}_r(\widehat{\omega}_t) &:= \frac{1}{2}\text{diag}^{-1}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r)(1 - \widehat{\omega}_t) = \\ &= \frac{1}{2}\text{diag}(\sigma_{r,1}^2 - \hat{\sigma}_{r,1}^2, \cdots, \sigma_{r,K}^2 - \hat{\sigma}_{r,K}^2)\mathbf{1}(1 - \widehat{\omega}_t)\end{aligned} \tag{P3-5}$$

is another bias term, with $\widehat{\omega}_t$ *denoting the total share of wealth invested in risky assets, i.e.,*

$$\widehat{\omega}_t := \mathbf{1}^{\mathsf{T}}\boldsymbol{\omega}_t.$$

Proof. See Appendix §F.2.

□

The function $\text{diag}^{-1}(\cdot)$ above is an inverse of the function $\text{diag}(\cdot)$, where the latter takes as an argument a vector and returns a diagonal matrix with a given vector's elements on the main diagonal, hence the former takes a diagonal (or just square) matrix and returns a column vector with the main diagonal's elements of a given matrix. Also, from now on the elements of a matrix are denoted with the same letter as the matrix but in lowercase.

An important intermediate result stated in the Proposition above is that in order to cancel out the effect on the wealth dynamics (our ultimate benchmark) of replacing the original

variance, $\boldsymbol{\Sigma}_r$, with the simplified variance, $\hat{\boldsymbol{\Sigma}}_r$, and thus to ensure the expected growth rate of the simplified log-wealth, $\ln \hat{W}_{t+1}$, equals that of the original log-wealth, $\ln W_{t+1}$, we also must adjust the mean of the simplified random variable, $\hat{\boldsymbol{\mu}}_r$, as shown in equation (P3-2). The corresponding bias term $\check{\boldsymbol{\mu}}_r$ consists of two sub-terms (equation P3-3).

The first (on the RHS of equation P3-3), rather technical one, is an artefact of the continuous-time approximation based on the (geometric) Brownian motion. It ensures that the expected return on the simplified and original wealth would coincide if the simplified wealth indeed followed the dynamics captured by the simplified variance; hence it can be viewed as a term setting the "origin" point.

The second one (on the RHS of equation P3-3) is more intuitive. It adjusts the mean downward from the origin (for positive risky investments) with the magnitude of the adjustment increasing in the share of wealth invested as well as in the size of the discrepancy between the original and simplified variances (a non-negative quantity, as shown later in §3.3); hence it can be viewed as a term imposing "pessimism". While also being a byproduct of the continuous-time approximation undertaken in the derivations above, the second adjustment term entails a downward shift in the expectation of log-return $\hat{\boldsymbol{r}}_{t+1}$ and effectively adopts a conservative view at the potential investment opportunities.

Remark 1 (Decision Rule- vs. Subjective Perception-Adjustment)**.** *A more direct way of correcting for the difference between the original and simplified vairances is to adjust the decision rules (i.e., policy functions dictating the choice of control variables $C_t, \{q_{0,t}, \boldsymbol{q}_t\}$) rather than the subjective perception (i.e., the mean of the simplified distribution $\hat{\boldsymbol{\mu}}_r$). Adjusting the decision rules shifts control variables $C_t, \{q_{0,t}, \boldsymbol{q}_t\}$ directly; which takes extra $(K+1)$ adjustment parameters. Adjusting the mean shifts control variables $C_t, \{q_{0,t}, \boldsymbol{q}_t\}$ indirectly, by affecting the state variables $\{q_{0,t-1}, \boldsymbol{q}_{t-1}\}, \hat{\boldsymbol{D}}_t$ and the chosen subjective probability density $h_D(\hat{\boldsymbol{D}}_{t+1}|\hat{\boldsymbol{D}}_t)$, via $h_r(\hat{\boldsymbol{r}}_{t+1})$, that control variables are functionally dependent on; which takes extra K adjustment parameters. Reasoning in terms of degrees of freedom, the latter*

option, with its lower number of adjustment parameters, is (weakly) more restrictive.[5]

Therefore, we are dealing with a simplified distribution $h_r(\hat{\boldsymbol{r}}_{t+1})$ that is biased, but biased in an optimal, expected distortion-minimizing way (see Appendix §F.2.1). For example, in the case of one risky asset that is the sole constituent of the investment portfolio (i.e., $\omega_t = 1$), the term setting the origin places the mean as if the risky asset is driven by a stochastic process that is determined by the simplified rather than the true variance (effectively matching the expected values of simplified and true returns, $\mathrm{E}_t^h[\hat{R}_{t+1}] = \mathrm{E}_t^g[R_{t+1}]$), while the pessimism term guarantees that the simplified mean does not depart from the true one (thus, matching the means $\hat{\mu}_r = \mu_r$). This results in the subjective expected return on the simplified portfolio being pessimistically biased and undershooting its objectively expected level (that is, $\mathrm{E}_t^h[\hat{W}_{t+1}/W_t] = \mathrm{E}_t^h[\hat{R}_{t+1}] = \exp(\hat{\mu}_r + 0.5\hat{\Sigma}_r) < \exp(\mu_r + 0.5\Sigma_r) = \mathrm{E}_t^g[R_{t+1}] = \mathrm{E}_t^g[W_{t+1}/W_t]$), but it actually ensures optimality in the end.

Remark 2 (Computational Benefits of Mean Adjustment)**.** *It may seem unreasonable to reduce information processing costs by using an approximating random variable with a simplified variance instead of the original one only to increase the burden down the line by necessitating the manipulations with the bias term (another kind of infinite regress critique). The reason the proposed approach works is because (conditionally on $\boldsymbol{\Sigma}_r$, $\hat{\boldsymbol{\Sigma}}_r$ and $\boldsymbol{\omega}_t$) the bias term $\breve{\boldsymbol{\mu}}_r$ is a non-stochastic object and possesses zero (in discrete case, $-\infty$ in continuous case) entropy, hence manipulating it is less computationally intensive than it is in the case of stochastic objects; e.g., consider the summation operations required to implement integration. (It can also be understood from a measure-theoretic standpoint as an issue of dimensionality: the stochastic objects (i.e., random variables as measurable functions from a space of outcomes to a measurable space) are characterized by a non-trivial profile on the corresponding measurable space, while zero-entropy objects (i.e., fixed constants) have a flat profile—if singletons, otherwise flat except a single atom—and the corresponding space is in some sense degenerate.)*

5. The author thanks Michael Woodford for raising this issue.

An additional assumption, required only for $K > 1$ cases, that disciplines the optimization-related information-processing and prohibits the dependence of $d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1})$ on $\boldsymbol{\omega}_t$ (by replacing the dependence on exact $\boldsymbol{\omega}_t$ with agnostic motive of minimization of maximum loss for any possible $\boldsymbol{\omega}_t$) allows to deduce a reasonable yet operationally applicable distortion function given in equation (P3-4).[6] It produces a convenient sum-of-squares formulation for suitably translated (bias-corrected) differences between the true and approximated log-returns. This finalizes the operational definition of the distortion function we are going to use for the informational sub-problem of problem $\mathcal{P}_{\mathcal{QI}}$.

Decorrelation via coordinate change: Part of our solution method relies on random variables being uncorrelated. This is implemented by transforming coordinates to the principal axes of the variance-covariance matrix, represented by its eigenvectors.

Thus, define the transformed random variables and parameters as

$$\boldsymbol{x} := \boldsymbol{\Xi}^{\intercal} \boldsymbol{r}_{t+1}, \quad (3.7) \qquad \hat{\boldsymbol{x}} := \boldsymbol{\Xi}^{\intercal} \hat{\boldsymbol{r}}_{t+1}, \quad (3.8) \qquad \check{\boldsymbol{\mu}}(\widehat{\omega}_t) := \boldsymbol{\Xi}^{\intercal} \check{\boldsymbol{\mu}}_r(\widehat{\omega}_t), \quad (3.9)$$

where $\boldsymbol{\Xi}$ is a square matrix with eigenvectors in its columns that is obtained from eigendecomposition (or diagonalization) of matrix $\boldsymbol{\Sigma}_r$.[7] The eigendecomposition procedure, $\mathcal{P}_{\boxslash}$, implies the following relationships:

$$\{\boldsymbol{\Sigma}, \boldsymbol{\Xi}\} := \text{eigendecompose}(\boldsymbol{\Sigma}_r); \qquad \{\mathcal{P}_{\boxslash}\}$$

6. Note that for any vector $\check{\boldsymbol{\mu}}_r(\widehat{\omega}_t)$, the refined distortion function does not depend on the scalar $\widehat{\omega}_t$; in other words, given the bias term $\check{\boldsymbol{\mu}}_r(\widehat{\omega}_t)$, the solution to the informational problem is invariant to the chosen value of the bound $\widehat{\omega}_t$. This fact will be put to work later.

7. In terms of Appendix §G's algorithm, our understanding is that this transformation is performed when solving the informational problem before the Generating codebook step of the algorithm presented in Appendix §G (the inverse transformation may be conducted either before or after maximizing the objective function in the process of solving the consumption and portfolio choice problem).

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_K^2 \end{bmatrix}, \qquad \boldsymbol{\Sigma}_r = \boldsymbol{\Xi}\boldsymbol{\Sigma}\boldsymbol{\Xi}^{-1}, \qquad \boldsymbol{\Xi}^{\intercal} = \boldsymbol{\Xi}^{-1}.$$

Proposition B.1 in Appendix §B states that the decorrelating transformation which facilitates our subsequent solution is innocuous.

Equilibrium: An equilibrium of the economy formed by the exogenously given economic setting described in §3.1 and the endogenously chosen optimal solutions to the feasible problem $\mathcal{P}_{\mathcal{QI}}$ is a collection of a continuous price function $\{P_0(\boldsymbol{D}_t), \boldsymbol{P}(\boldsymbol{D}_t)\} : \mathbb{R}_+^K \mapsto \mathbb{R}_+^{K+1}$, a continuous and bounded value function $v(\{q_{0,t-1}, \boldsymbol{q}_{t-1}\}, \boldsymbol{D}_t) : \mathbb{R}_+^{K+1} \times \mathbb{R}_+^K \mapsto \mathbb{R}$, and an absolutely continuous joint probability distribution function $F_D(\boldsymbol{D}_{t+1}, \hat{\boldsymbol{D}}_{t+1} | \boldsymbol{D}_t, \hat{\boldsymbol{D}}_t) : \mathbb{R}_+^K \times \mathbb{R}_+^K \mapsto [0, 1]$ such that:

(i) [consumption and investment optimality] Bellman equation (PQI-1) subject to the budget constraint (PQI-2), control variable's domain restriction, no-Ponzi-schemes constraint and with given utility function specification is satisfied;

(ii) [consumption and investment coherence] goods and asset markets clear, i.e.,

$$C_t = \widehat{\boldsymbol{q}}^{\intercal} \boldsymbol{D}_t, \qquad \boldsymbol{q}_t = \widehat{\boldsymbol{q}}, \qquad q_{0,t} = \widehat{q}_0;$$

(iii) [informational optimality] joint probability density function $f_r(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1})$ solves (utilizing the decorrelating transformation) the informational problem $\mathcal{P}_{\mathcal{I}}$ with $g_r(\boldsymbol{r}_{t+1})$ as the true density and κ as the information processing capacity;

(iv) [informational coherence] probability density functions for dividends $f_D(\boldsymbol{D}_{t+1}, \hat{\boldsymbol{D}}_{t+1} | \boldsymbol{D}_t, \hat{\boldsymbol{D}}_t)$ (referred to in equation PQI-4), $g_D(\boldsymbol{D}_{t+1} | \boldsymbol{D}_t)$ (referred to in PQI-5) and $h_D(\hat{\boldsymbol{D}}_{t+1} | \hat{\boldsymbol{D}}_t)$ (referred to in PQI-3) are consistent with the densities for returns $f_r(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1})$, $g_r(\boldsymbol{r}_{t+1})$ and $h_r(\hat{\boldsymbol{r}}_{t+1})$, also the correlated random variables' densities $f_r(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1})$,

$g_r(\boldsymbol{r}_{t+1})$ and $h_r(\hat{\boldsymbol{r}}_{t+1})$ are consistent with the decorrelated variables' densities $f(\boldsymbol{x}, \hat{\boldsymbol{x}})$, $g(\boldsymbol{x})$ and $h(\hat{\boldsymbol{x}})$, i.e., $\forall \boldsymbol{D}_{t+1}, \hat{\boldsymbol{D}}_{t+1} \in \mathbb{R}_+^K$:

$$\begin{aligned} f_D(\boldsymbol{D}_{t+1}, \hat{\boldsymbol{D}}_{t+1} | \boldsymbol{D}_t, \hat{\boldsymbol{D}}_t) = f_r(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}) = f_r(\boldsymbol{\Xi}\boldsymbol{x}, \boldsymbol{\Xi}\hat{\boldsymbol{x}}) = f(\boldsymbol{x}, \hat{\boldsymbol{x}}), \\ g_D(\boldsymbol{D}_{t+1} | \boldsymbol{D}_t) = \quad g_r(\boldsymbol{r}_{t+1}) \quad = \quad g_r(\boldsymbol{\Xi}\boldsymbol{x}) \quad = g(\boldsymbol{x}), \\ h_D(\hat{\boldsymbol{D}}_{t+1} | \hat{\boldsymbol{D}}_t) = \quad h_r(\hat{\boldsymbol{r}}_{t+1}) \quad = \quad h_r(\boldsymbol{\Xi}\hat{\boldsymbol{x}}) \quad = h(\hat{\boldsymbol{x}}). \end{aligned}$$

A policy function determining the optimal investment in tree shares $\boldsymbol{q}(\{q_{0,t-1}, \boldsymbol{q}_{t-1}\}, \hat{\boldsymbol{D}}_t)$ could be added to the list of equilibrium objects, but it is a constant function that is identically equal to $\widehat{\boldsymbol{q}}$ because the considered economy is an autarky. Although not made explicit, the informational coherence equations subsume the Jacobians of the transformations.

Note that the conditions (iii) and (iv) replace the traditional rational expectations assumption (Lucas, 1978). Otherwise, the notion of equilibrium is standard. The existence of an equilibrium is proven by constructing its instance and solving the model.

3.3 Solution

The consumption and investment segment of the larger problem is fairly standard, so here we only focus on the crucial elements of the informational part, benefiting from the clearly segregated formulations of these two sub-problems. (The full solution to the feasible consumption and portfolio choice problem $\mathcal{P}_{\mathcal{QI}}$ is available in Appendix §C.)

Taking the general solution to $\mathcal{P}_{\mathcal{I}}$ from Proposition 1, exploiting the flexible mean property due to Proposition 2, using the (refined) distortion function from Proposition 3, and applying decorrelating transformation allowed by Proposition B.1 yields:

$$f(\boldsymbol{x}|\hat{\boldsymbol{x}}) = \exp\left(\frac{1}{\lambda}\nu(\boldsymbol{x}, \hat{\boldsymbol{x}}) - \frac{1}{\lambda}\mu(\boldsymbol{x}) - \frac{1}{\lambda}(\boldsymbol{x} - \hat{\boldsymbol{x}} + \check{\boldsymbol{\mu}}(\widehat{\omega}_t))^{\intercal}(\boldsymbol{x} - \hat{\boldsymbol{x}} + \check{\boldsymbol{\mu}}(\widehat{\omega}_t))\right), \quad \forall \hat{\boldsymbol{x}} \in \operatorname{supp}(h). \tag{3.10}$$

Given our knowledge about the probability distribution of $\boldsymbol{x}$, we can solve for the whole stochastic structure of the relationship between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$, as shown in the following Proposition. But an attentive reader has already spotted the kernel of a Gaussian probability density

function in the last equation, which suggests the direction of our subsequent exploration and explains the labor we put into specification of the distance function.

Proposition 4 (Specific Solution to Informational Problem)**.** *Let the general solution to the informational problem, which is specialized to the chosen distortion function and accounts for the decorrelating transformation, be given by the conditional probability density function* $f(\boldsymbol{x}|\hat{\boldsymbol{x}})$ *from (3.10), where the random vector* $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, with the constant vector* $\boldsymbol{\mu}$ *implicitly defined via (3.7) and the matrix* $\boldsymbol{\Sigma}$ *described in the relationships* $\mathcal{P}_{\boxslash}$*.*

Then, the specific solution to the informational problem can take one of two forms, depending on the magnitude of information processing capacity κ *(i.e., tightness of shadow price/Lagrange multiplier on information constraint* λ*):*

(a) Interior solution ("large" κ*, "small"* λ*).*

$$f(\boldsymbol{x}|\hat{\boldsymbol{x}}) = (2\pi)^{-\frac{K}{2}} \left|\frac{\lambda}{2}\boldsymbol{I}_K\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}} + \check{\boldsymbol{\mu}}(\widehat{\omega}_t))^{\intercal} \left(\frac{\lambda}{2}\boldsymbol{I}_K\right)^{-1} (\boldsymbol{x} - \hat{\boldsymbol{x}} + \check{\boldsymbol{\mu}}(\widehat{\omega}_t))\right), \quad \forall \hat{\boldsymbol{x}} \in \mathbb{R}^K;$$

$$\boldsymbol{x} = \hat{\boldsymbol{x}} - \check{\boldsymbol{\mu}}(\widehat{\omega}_t) + \boldsymbol{\epsilon},$$

where

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}), \qquad \boldsymbol{\Psi} = \frac{\lambda}{2}\boldsymbol{I}_K,$$
$$\hat{\boldsymbol{x}} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}(\widehat{\omega}_t), \hat{\boldsymbol{\Sigma}}), \qquad \hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \boldsymbol{\Psi};$$

$$\lambda = 2\left(e^{-2\kappa}|\boldsymbol{\Sigma}|\right)^{\frac{1}{K}}.$$

Interior solution is valid if the following condition holds: $\sigma_k^2 > \frac{\lambda}{2},\ \forall k \in \{1, \ldots, K\}$.

(b) Boundary solution ("small" κ, "large" λ).

$$f(\boldsymbol{x}|\hat{\boldsymbol{x}}) = (2\pi)^{-\frac{K}{2}} \, |\boldsymbol{\Psi}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}} + \check{\boldsymbol{\mu}}(\widehat{\omega}_t))^{\intercal} \boldsymbol{\Psi}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{x}} + \check{\boldsymbol{\mu}}(\widehat{\omega}_t))\right),$$

$$\forall \hat{\boldsymbol{x}} \in \operatorname{supp}(h);$$

$$\boldsymbol{x} = \hat{\boldsymbol{x}} - \check{\boldsymbol{\mu}}(\widehat{\omega}_t) + \boldsymbol{\epsilon},$$

where

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}), \qquad \boldsymbol{\Psi} = \begin{bmatrix} \lambda/2 & & 0 & 0 & \cdots & 0 \\ & \ddots & & \vdots & \ddots & \vdots \\ 0 & & \lambda/2 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \sigma^2_{k^*+1} & & 0 \\ \vdots & \ddots & \vdots & & \ddots & \\ 0 & \cdots & 0 & 0 & & \sigma^2_K \end{bmatrix},$$

$$\hat{\boldsymbol{x}} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}(\widehat{\omega}_t), \hat{\boldsymbol{\Sigma}}), \qquad \hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \boldsymbol{\Psi};$$

$$\begin{aligned} \{\sigma^2_k\}_1^K &:= \operatorname{sortdescending}(\{\sigma^2_k\}_1^K), \\ k^* &:= \arg \min_{k \in \{1,\ldots,K\}} \{\sigma^2_k \mid \sigma^2_k > \frac{\lambda}{2}\}, \\ \lambda &= 2\left(e^{-2\kappa} \sigma^{-2}_{k^*+1} \cdots \sigma^{-2}_K |\boldsymbol{\Sigma}|\right)^{\frac{1}{k^*}}. \end{aligned}$$

(The last $(K - k^*)$ elements of vector $\hat{\boldsymbol{x}}$ are to be understood as deterministic scalars, or alternatively as Dirac delta functions centered at $\{\hat{\mu}_{k^*+1}(\widehat{\omega}_t), \cdots, \hat{\mu}_K(\widehat{\omega}_t)\}$ at a cost of abusing the notation when dealing with their Radon-Nikodym derivatives.)

Boundary solution is valid if the following condition holds: $\exists k \in \{1, \ldots, K\} : \ \sigma^2_k \leq \frac{\lambda}{2}$.

Proof. See Appendix §F.4.

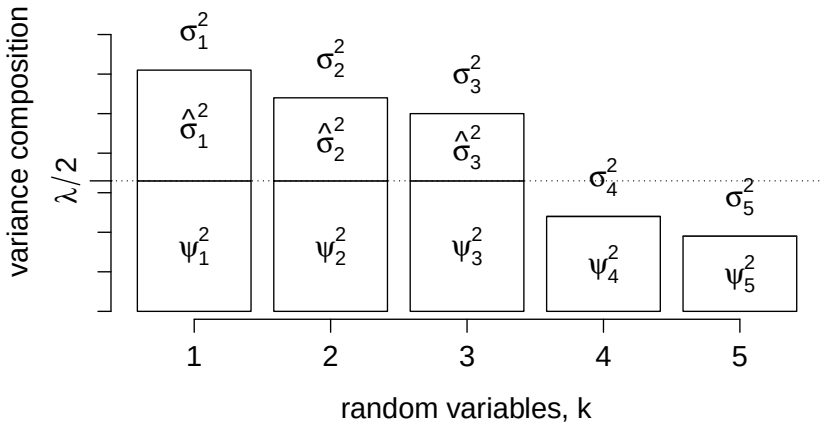


Figure 3.1: "Reverse water-filling".

□

The qualifiers 'interior' and 'boundary' in the formulation of the above Proposition should be understood in relation to the Cartesian product set $\bigtimes_1^K [0, \sigma_k^2]$. Figure 3.1 illustrates the (relatively more general) "reverse water-filling" logic of the boundary solution of Proposition 4. Consider a situation when information processing capacity κ is large enough so that information constraint does not bind: then the corresponding Lagrange multiplier λ is 0; elements on the diagonal of the variance-covariance matrix for approximation error $\mathbf{\Psi}$ are all 0, $\psi_k^2 = 0, \ \forall k \in \{1, \ldots, K\}$; while diagonal elements of the variance-covariance matrices $\hat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}$ are equalized, $\hat{\sigma}_k^2 = \sigma_k^2, \ \forall k \in \{1, \ldots, K\}$. If κ decreases and the information constraint starts to bind, initially the following happens: λ increases; ψ_k^2 rises slightly but remains equal for all k; while $\hat{\sigma}_k^2$ equal the difference between σ_k^2 and the level of ψ_k^2 for all k. If κ decreases even further, a qualitative change in the picture occurs, and we move from the interior to the boundary solution case: at some point λ increases enough for ψ_k^2 to catch up with the lowest σ_k^2 and the corresponding $\hat{\sigma}_k^2$ to become 0; then the remaining ψ_l^2 for $l \neq k$ depart from ψ_k^2 and continue to rise, so that the "reverse water-filling" process goes on for the rest of $\hat{\sigma}_l^2$. And so on.

From the information-processing perspective, σ_k^2 represent the total information available for processing, $\hat{\sigma}_k^2$ represent the information that is actually processed, while ψ_k^2 represent

the information that is omitted and consitutes the approximation errors. As information processing capacity decreases, the information omissions disproportionally affect the data dimensions with low informational content (small eigenvalues σ_k^2).[8]

The above configuration of the solution is not just a technical attribute, but also has deep economic implications.

Corollary 1 (Specific Solution to Informational Problem: Dispersion Folding)**.** *The specific solution to the informational problem, as given in the statement of Proposition 4, is characterized by the folded dispersions (or collapsed randomness/distributions) of the less volatile components of random vector $\hat{\boldsymbol{x}}$ in the boundary solution case. I.e., the corresponding subjective variances become 0:*

$$\hat{\sigma}_k^2 = 0, \quad \forall k > k^*.$$

Proof. Immediate from Proposition 4.

□

Considering (the more revealing case here) of boundary solution, the role of simplification is manifested in dropping some of the random variables' dimensions (or the random variables themselves, if they are uncorrelated) from the agent's approximation. Due to the effects of entropy/variance reduction culminating in its "folding" or "collapse"/"contraction", such random variables' dimensions are replaced with non-stochastic objects, that is by their—sufficiently biased—means (cf. the sparsity logic of Gabaix, 2014a).

Intuitively, in the case of two uncorrelated random variables (alternatively, two random variables' dimensions) x_1 and x_2, according to Corollary 1 as the "folded" random variable (dimension) $\hat{x}_2$ effectively becomes non-stochastic, a simple univariate (one-dimensional)

8. Lastly, note that in our case we benefit from the Normality of distribution $g(\cdot)$, among others assumptions: then the resulting distribution $h(\cdot)$ turns out to be Normal as well, but is characterized by lower variance. The result is not always as straight-forward: e.g., in the case of distribution $g(\cdot)$ having a bounded support, a discretely distributed solution for $h(\cdot)$ arises in the analysis of Matějka and Sims (2010).

approximating model based on $\hat{x}_1$ emerges subjectively.

For example, consider a vineyard: a garden planted with both mature and young grapevines (think of them as two different "trees", each representing an aggregate of identical vines). The formers' "payoff" is determined by their exposure to light and humidity conditions, while the latters' by their time to first harvest and healthiness of the development; hence, the two are uncorrelated. If the young vines are also few in number, a binding capacity constraint may lead to total disregard of the presence of these less important (and thus barely worth spending capacity on) vines.

Taking a slightly more involved case of correlated random variables, five different grapes may be perfectly characterized by such attributes as their acidity, body, flavor (say, spice), sugar and tannin levels. Here, a binding capacity constraint may result in focusing on the most crucial attributes (say, acidity, body and sugar levels) and ignoring the rest (flavor and tannin levels). (Corollary 2 develops this theme a little further.)

The results of Proposition 4 in economically interesting terms such as returns, i.e., after the inversion of the decorrelating transformation, are presented in the subsequent Proposition 5.

Proposition 5 (Specific Solution to Informational Problem: Representation in Economic Terms)**.** *The specific solution to the informational problem given in the statement of Proposition 4 can be equivalently represented in terms of returns. In particular, the following decomposition is valid:*

$$\boldsymbol{r}_{t+1} = \hat{\boldsymbol{r}}_{t+1} - \check{\boldsymbol{\mu}}_r(\widehat{\omega}_t) + \boldsymbol{\epsilon}_{r,t+1}, \qquad \forall \boldsymbol{r}_{t+1} \in \mathbb{R}^K,$$

also producing

$$\boldsymbol{\Sigma}_r = \hat{\boldsymbol{\Sigma}}_r + \boldsymbol{\Psi}_r, \qquad \forall \boldsymbol{\Sigma}_r \textit{ that is } K \times K \textit{ positive semi-definite},$$

where

$$\hat{\boldsymbol{r}}_{t+1} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_r(\widehat{\omega}_t), \hat{\boldsymbol{\Sigma}}_r),$$
$$\boldsymbol{\epsilon}_{r,t+1} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_r),$$

with $\hat{\boldsymbol{\mu}}_r(\widehat{\omega}_t)$ *and* $\check{\boldsymbol{\mu}}_r(\widehat{\omega}_t)$ *given by Proposition 3, as well as with* $\hat{\boldsymbol{\Sigma}}_r$ *and* $\boldsymbol{\Psi}_r$ *defined as*

$$\hat{\boldsymbol{\Sigma}}_r := \boldsymbol{\Xi}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Xi}^{-1}, \qquad \boldsymbol{\Psi}_r := \boldsymbol{\Xi}\boldsymbol{\Psi}\boldsymbol{\Xi}^{-1}$$

basing on the results of Proposition 4.

Proof. See Appendix §F.5.

□

Moreover, arguments in Appendix §C.3 allow to replace $\check{\boldsymbol{\mu}}_r(\widehat{\omega}_t)$ and $\hat{\boldsymbol{\mu}}_r(\widehat{\omega}_t)$ with, respectively, $\check{\boldsymbol{\mu}}_r$ and $\hat{\boldsymbol{\mu}}_r$ in the statement of Proposition 5 (as well as in Proposition 4).[9]

To sum up, Proposition 3 demonstrates the appropriateness of our approximation procedure and allows to derive the normality of $\hat{\boldsymbol{r}}_{t+1}$ with the subsequent decomposition equations of Propositions 4–5, providing us with great analytical convenience; while Appendix §C.3 ensures that approximation accuracy result from Proposition 3 has not been lost in the process.

There are two main economic results in Proposition 5: (i) due to simplification $\hat{\boldsymbol{\Sigma}}_r$, the variance-covariance matrix of the simplified log-returns $\hat{\boldsymbol{r}}_{t+1}$, is smaller than $\boldsymbol{\Sigma}_r$, its counterpart for the original log-returns $\boldsymbol{r}_{t+1}$ (i.e., the difference $\hat{\boldsymbol{\Sigma}}_r - \boldsymbol{\Sigma}_r$ is a negative semi-definite matrix); (ii) as compensation for the simplification above, $\hat{\boldsymbol{\mu}}_r$, the mean of $\hat{\boldsymbol{r}}_{t+1}$,

9. There are also a couple of minor technical details of note. To aid subsequent analysis, it may be worthwhile highlighting that in the interior solution case the variance-covariance matrix of approximation errors for returns is again diagonal and remains unchanged: $\boldsymbol{\Psi}_r = \boldsymbol{\Psi}$. Another revealing result in the interior solution case is that the optimal mean bias term defined in Proposition 3.1, $\check{\boldsymbol{\mu}}_r$, takes the form conformable with the bias term from Proposition 3.3, $\check{\boldsymbol{\mu}}_r(\widehat{\omega}_t)$: $\check{\boldsymbol{\mu}}_r = \frac{1}{2}\mathrm{diag}^{-1}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r) \odot (\mathbf{1} - \boldsymbol{\omega}_t)$, where $\odot$ denotes the Hadamard product.

is biased toward pessimism in comparison to $\boldsymbol{\mu}_r$, its counterpart for $\boldsymbol{r}_{t+1}$. Both results are obtained endogenously and follow from §3.2.

In addition to the claim that $\hat{\boldsymbol{\Sigma}}_r$ is smaller than $\boldsymbol{\Sigma}_r$, there are more facets to this attribute of the solution.

Corollary 2 (Specific Solution to Informational Problem: Correlation Inflation)**.** *The specific solution to the informational problem represented in economic terms, as given in the statement of Proposition 5, is characterized by:*

(a) The inflated correlations between the elements of $\hat{\boldsymbol{r}}_{t+1}$ *relative to those for the elements of* $\boldsymbol{r}_{t+1}$ *in the interior solution case. I.e., the generic correlation coefficient's subjective version moves away from its objective value towards* 1 *(or* -1*):*

$$|\hat{\rho}_{r,kl}| \geq |\rho_{r,kl}|, \qquad \forall k, l \in \{1, \ldots, K\};$$

which can be seen directly in the relationship

$$\hat{\rho}_{r,kl} = \rho_{r,kl} \times \frac{\left(\sum_{m=1}^{K} \xi_{km}^2 \sigma_m^2\right)^{1/2} \left(\sum_{m=1}^{K} \xi_{lm}^2 \sigma_m^2\right)^{1/2}}{\left(\sum_{m=1}^{K} \xi_{km}^2 \sigma_m^2 - \psi_1^2\right)^{1/2} \left(\sum_{m=1}^{K} \xi_{lm}^2 \sigma_m^2 - \psi_1^2\right)^{1/2}}, \quad \forall k, l \in \{1, \ldots, K\},$$

where

$$\psi_1^2 := \left(e^{-2\kappa} |\boldsymbol{\Sigma}|\right)^{\frac{1}{K}} < \min_{m \in \{1,\ldots,K\}} \sigma_m^2, \qquad \sum_{m=1}^{K} \xi_{km}^2 = 1, \quad \forall k \in \{1, \ldots, K\}.$$

(b) Either inflated or shrinking correlations between the elements of $\hat{\boldsymbol{r}}_{t+1}$ *relative to those for the elements of* $\boldsymbol{r}_{t+1}$ *in the boundary solution case. I.e., the generic correlation coefficient's subjective version may move away from its objective value either toward* 0 *or* 1 *(*-1*):*

$$|\hat{\rho}_{r,kl}| \gtreqless |\rho_{r,kl}|, \qquad \forall k, l \in \{1, \ldots, K\}.$$

Proof. See Appendix §F.7.

□

In spite of the ambiguous result for the specific correlation coefficients—and even that only for the boundary solution case—the region of the parameter space affected by such indeterminacy is (loosely speaking) "small". Because, together with the shrinking diagonal variance terms, inflation of the off-diagonal covariance terms contributes to achieving a variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_r$ smaller than $\boldsymbol{\Sigma}_r$—mechanics that apply both to the interior and boundary solution cases, thus making the "correlation inflation" outcome robust. Furthermore, note that the reduced variances and biased correlations are the reflections of the trade-off allowed by the information processing capacity constraint (2.3).

Intuitively, Corollary 2 shows how the inflation of the correlations between the elements of $\hat{\boldsymbol{r}}_{t+1}$ as compared to the correlations between the elements of $\boldsymbol{r}_{t+1}$ emerges subjectively; which effectively leads to further attraction of the positively correlated elements and the repulsion of the negatively correlated ones that ultimately results in the pooling of the random vector's components into relatively detached categories.[10]

For example, think of Cabernet Sauvignon, Pinot Noir and Shiraz grapevines being pulled together into one category, Pinot Grigio and Sauvignon Blanc grapevines into another category, and with two categories of plants being pushed apart as very distinct kinds of capital goods—say, "red" vs. "white"—that are characterized by different attributes.

10. In computational cognitive science, as neural nework models undergo supervised learning to perform categorization tasks, they demonstrate an emergent property of categorical perception: the latter is characterized by within-category compression and between-category separation, similarly to the "correlation inflation" effect above. For details, see Tijsseling and Harnad (1997), Damper and Harnad (2000).

CHAPTER 4

DISCUSSION

4.1 Theoretical results

The information processing capacity κ that is low enough to make the information constraint (2.3) binding induces a subjective probability measure $h(\cdot)$ that is different from the objective measure $g(\cdot)$. Using the former in place of the latter for making decisions under risk is less computationally burdensome, but at the same time biases the decision-making environment in a certain, predictable direction. Although this discrepancy may give rise to decision outcomes deviating from the unconstrained, "rational expectations" alternative, constrained optimality (offered by the solution to the feasible problem $\mathcal{P}_{\mathcal{QI}}$) is still within reach as long as optimal adjustments are made. Figure 4.1 illustrates the differences between the objective landscape of the stochastic environment and the subjective perspective on this stochastic landscape for a case with two risky assets (under condition that the necessary adjustments are indeed undertaken).

One key result can be viewed as the effective **"overconfidence"**. Because of the constraint on utilized information processing capacity, the subjective probability measure $h(\cdot)$ is characterized by lower entropy than the objective probability measure $g(\cdot)$, i.e., the former is a coarser version of the latter. In our case (with log-normal payoffs), the entropy reduction is achieved solely by decreasing the variance of the relevant random variables. Specifically, the variance decomposition equation from Proposition 5 implies that $\hat{\mathbf{\Sigma}}_r$, the variance-covariance matrix of $\hat{\boldsymbol{r}}_{t+1}$, is smaller than $\mathbf{\Sigma}_r$, its counterpart for $\boldsymbol{r}_{t+1}$ (the difference is a negative semi-definite matrix). This is reflected in Figure 4.1 by the relatively more peaked probability densities in the right panel.

An additional important result can be viewed as necessary **"pessimism"**. In order to compensate for this entropy-reducing (hence, variance-decreasing in our case) simplification, the means of approximating random variables have to be adjusted. Thereby, a biased second

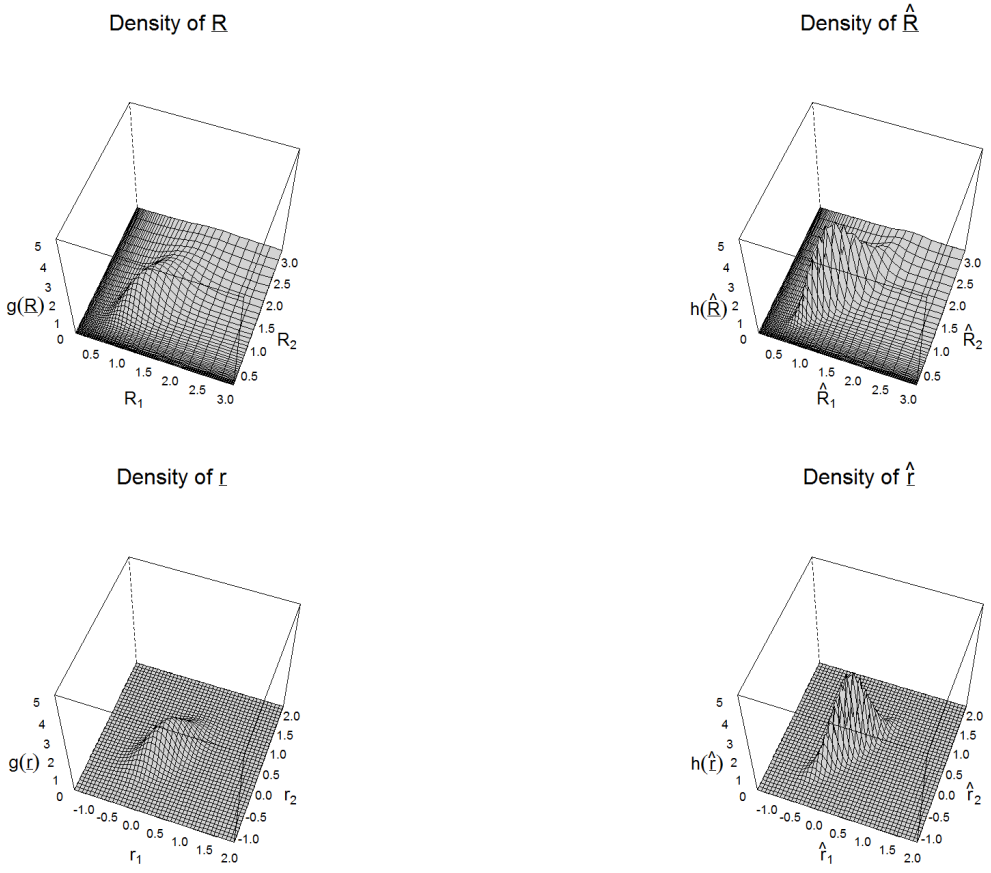


Figure 4.1: Objective and subjective probability densities, two risky assets (parameterizations used are $\boldsymbol{R}_{t+1} \sim \log \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$, $\boldsymbol{r}_{t+1} \sim \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$, $\boldsymbol{\mu}_r = [0.10; 0.20]$, $\boldsymbol{\Sigma}_r = [0.10, 0.08; 0.08, 0.16]$, $\mathrm{E}_t^g[\boldsymbol{R}_{t+1}] = [1.16; 1.32]$ (left panel); and, for $\kappa = 1$ nat ≈ 1.44 bits with $\boldsymbol{\omega}_t = [0.5; 0.5]$, $\hat{\boldsymbol{R}}_{t+1} \sim \log \mathcal{N}(\hat{\boldsymbol{\mu}}_r, \hat{\boldsymbol{\Sigma}}_r)$, $\hat{\boldsymbol{r}}_{t+1} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_r, \hat{\boldsymbol{\Sigma}}_r)$, $\hat{\boldsymbol{\mu}}_r = [0.11; 0.21]$, $\hat{\boldsymbol{\Sigma}}_r = [0.06, 0.08; 0.08, 0.12]$, $\mathrm{E}_t^h[\hat{\boldsymbol{R}}_{t+1}] = [1.15; 1.31]$ (right panel)).

moment necessitates a compensating bias in the first moment. Effectively, the required adjustment amounts to adopting a subjectively pessimistic view at the future state of the world (i.e., for a positive investment its expected value is biased downward). The exact bias term $\check{\boldsymbol{\mu}}_r$, as given by Proposition 3, depends on the shares of wealth invested in different risky assets, $\boldsymbol{\omega}_t$, as well as on the discrepancy between objective and subjective variances, $(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r)$. Failure to adopt such a compensating adjustment, or using an incorrect adjustment term leads to suboptimal decision outcomes and subsequent losses.[1] For instance, ignoring the pessimistic adjustment and setting the mean $\hat{\boldsymbol{\mu}}_r$ of log-returns $\hat{\boldsymbol{r}}_{t+1}$ so as to match the subjectively expected level of returns $\mathrm{E}_t^h[\hat{\boldsymbol{R}}_{t+1}]$ to the objectively expected one $\mathrm{E}_t^g[\boldsymbol{R}_{t+1}]$ for a risk-averse agent would result in relative "overinvestment" into risky assets in terms of the wealth share. Figure 4.1 demonstrates the optimal result with the expected value of the top right panel's probability distribution moved slightly to the west (although the bottom right panel's distribution is actually shifted to the east).

Another key result is emerging **"categorization"**.[2,3] In our case, the positively cor-

1. Note that this result is not necessarily at odds with our initial motivating experiment due to Gabaix and Laibson (2000), which implies "overconfidence" without "pessimism". If the differences of simplified variances from true variances are about the same for different root nodes in the choice set, the mean bias term is roughly equalized between available choices, and thus can be ignored in their setting.

2. The existing literature distinguishes at least two kinds of categorization: bundling of random variables' support set partitions (i.e., bundling of states; e.g., coarsening of the state space through merging of several states into one) and pooling of the random variables/different random vector's elements themselves (i.e., pooling of types; e.g., conditionally on one random variable the other converges to a non-stochastic (Dirac delta) function). (In principle, merging of support set partitions also happens in the process of quantization of a continuous random variable, but we focus on different issues at this point.)
Technically, entropy/variance reduction for uncorrelated random variables leads, at its extreme, to "dispersion folding" or "randomness collapse"/"distribution contraction" (folding and dropping out low-volatility categories, think of it as categorization in vs. categorization away; this formally corresponds to bundling of states). However, for correlated random variables the above pertains to random variables' dimensions, and leads to "correlation inflation" (clustering into similar categories, categorization together vs. categorization apart; this formally corresponds to pooling of types). The former is dealt with by Corrollary 1, the latter by Corrollary 2.
Importantly, now we examine onle one of the two categorization mechanisms: instead of focusing on categorization as "folding" and dropping out (less important) random variables, here we focus on categorization as "correlation inflation" between random variables. We discuss the economic role of the other mechanism in more detail elsewhere, as a part of a separate line of investigation.

3. In the words of Herbert Simon (1997), "The human being striving for rationality and restricted within the limits of his knowledge has developed some working procedures that partially overcome these difficulties. These procedures consist in assuming that he can isolate from the rest of the world a closed system containing

related elements of $\boldsymbol{r}_{t+1}$ become even more correlated in $\hat{\boldsymbol{r}}_{t+1}$, thus similarly behaved co-moving assets exhibit a sort of attraction. While negatively correlated elements become even more so, leading to the repulsion of the counter-moving assets. These dynamics engender subjective clustering of different assets into relatively distinct categories, "asset classes". The mechanics behind it are the same as before: the covariance terms in $\hat{\boldsymbol{\Sigma}}$, the variance-covariance matrix of decorrelated simplified random variables $\hat{\boldsymbol{x}}$, are unchanged, but the variance terms are reduced. Consequently, leading to the correlation coefficients exceeding those in $\boldsymbol{\Sigma}$, the variance-covariance matrix of decorrelated original random variables $\boldsymbol{x}$. Strictly speaking, the effect on the correlations in $\hat{\boldsymbol{\Sigma}}_r$ is ambiguous and depends on the tightness of the information constraint and the combination of the diagonal elements in $\boldsymbol{\Sigma}$, as well as on decorrelating eigenvectors $\boldsymbol{\Xi}$. However, overall the "correlation inflation" outcome dominates: the entropy of a Gaussian random vector is a one-to-one map with the determinant of its variance-covariance matrix, and so can be reduced only by decreasing the variances of and/or increasing the (absolute magnitudes of) covariances between the random vector's elements.

As a consequence, assets in the same category (i.e., positively correlated ones) would tend to be treated as more similar than they actually are, while assets in different categories (negatively correlated) would seem more different that they really are. For this reason, without proper adjustment of the mean, as say is happening in expected value-matching that results in overall "overinvesment" into risky assets, both "over-" and "underinvestment" is possible for specific risky assets depending on the relative size and direction of the correlation biases. It can be seen in Figure 4.1 that the random variables described by the probability densities of the right panel are more aligned along the west-east axis and hence exhibit a higher pairwise correlation ($\hat{\rho}_{r,12} = 0.90 > 0.63 = \rho_{r,12}$).[4]

---

a limited number of variables and a limited range of consequences." In our framework, the "limited number of variables" idea roughly corresponds to the "folding" effect, while the "limited range of consequences" to "correlation inflation" and entropy/variance reduction more generally.

4. One way to look at this categorization result is through the lens of principal component analysis. Propositions 4 and 5 reveal the effective amplification of the relative magnitude of the largest eigenvalues

Nevertheless, the **decision outcomes** (say, prices $P^*_{0,t}$ and $\boldsymbol{P}^*_t$) for agents implementing the approximation procedures correctly by construction, in accordance with Proposition 3, achieve constrained optimality. That is, are in the "neighborhood" of fully optimal unconstrained "rational expectations" outcomes, with the "radius" of the neighborhood inversely related to the magnitude of information processing capacity κ. Technically, decision errors appear due to the approximation to the wealth process, which leads to a positive expected distortion $\mathrm{E}^f_t\left[d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1})\right] = \boldsymbol{\Psi}_r \geq \mathbf{0}$. The approximation is very accurate in practice though (see the comments on Proposition 3.1 in §F.2). Also note that decision errors are non-systematic/symmetric around the fully optimal levels (see Proposition 5), hence they are likely to cancel out on aggregate; and are relatively smaller for contingencies that impact welfare the most (see §J). At the same time, non-negligible radius of the neighborhood of deviations from full optimality leaves room for positive **trading volumes** even between agents having access to exactly same information and identical in all other respects except different levels of κ (although agent multiplicity is not modeled explicitly here).[5]

4.2 Mapping theory to empirics

When presenting the schematic process of decision-making under risk that underlies our framework in part §2.2, we used lotteries as a basic example to illustrate it. The algorithm has some given "true" probability distribution $g(\cdot)$ as an input, which later has to be approximated by distribution $h(\cdot)$ in order to facilitate the processing of information and ultimately making the decisions. However, in contrast to the experiments with lotteries considered

of the subjective variance-covariance matrix that in turn leads to the amplification of the relative share of the subjective random variables' variance captured by their first principal components. As long as any two variables share the same leading principal components—in other words, the absolute magnitude of their pairwise correlation coefficient is high—then an increase in these leading principal components' importance increases (in absolute terms) the correlation between the two variables.

5. We again abstract away from the fact that in our particular exchange economy, consumption and investment outcomes coincide in constrained and unconstrained cases, and (as long as competitive equilibrium is unique/markets are complete) so do prices, hence there is actually no room for decision errors as well as trade.

there, when dealing with real-world applications we as econometricians only have access to sample data and do not know in full the "true" distribution that investors are working with. We interpret this "true" probability distribution $g(\cdot)$ classically as an objective, population distribution that describes (in reduced form) the data generating process, although it can also be understood in a Bayesian manner as a posterior distribution based at least partly on some prior beliefs (the latter distribution is sometimes termed subjective, but we reserve that name for other uses).

However it is understood, we assume that investors, who are inside actors in the model economy, know that true distribution and work with it, while we as econometricians, who are outside observers of the economic activities, have just a sample of realizations from the unobserved true distribution at our disposal. Since some of the contingencies, e.g., low-probability rare events, have not realized, they are missing from the observed data sample. Hence, the sample distribution is likely to have reduced entropy, and thus by definition manifests a form of simplification that we are concerned about. Accordingly, we treat it as such, as a simplified, approximate distribution $\hbar(\cdot)$.

One critically important caveat is in order though. Consider a univariate case, with the whole stock market as a single risky asset. The sample $\hbar(\cdot)$ that we have at our disposal may be a dislocated representation of the population distribution $g(\cdot)$; but, even if so, there is no solid reason to assert in which direction it is biased. Hence, by principle of insufficient reason, the mean of the sample distribution of returns is postulated to match the mean of the population distribution (instead of being adjusted in line with Proposition 3). That is, the expected values of the returns are equal, $\mathrm{E}_t^{\hbar}[\hat{R}_{t+1}] = \mathrm{E}_t^{g}[R_{t+1}]$; but equality does not hold for the means of the logarithms of returns, $\mathrm{E}_t^{\hbar}[\hat{r}_{t+1}] = \hat{\mu}_r \neq \mu_r = \mathrm{E}_t^{g}[r_{t+1}]$ (in fact, the left-hand-side quantity is larger, generally over the extent commanded by optimal mean adjustment that includes pessimistic bias, as evidenced by the equality of the levels of expected returns). In other words, this sample distribution $\hbar(\cdot)$ is simplified in the sense of possessing lower entropy, but it is not any investor's optimal simplified distribution $h(\cdot)$.

(Strictly speaking, since optimal bias term depends on the share of wealth invested in risky asset, ω_t, absence of the pessimistic correction stemming from equality of expected values of the returns corresponds to the case when $\omega_t = 0$ (see Proposition 3 and the explanations that follow); i.e., for an investor with zero wealth invested in risky asset, or really any agent that does not participate on the market, provided he does not simplify beyond sample distribution, this is actually the optimal bias term, and for him $h(\cdot) := \hbar(\cdot)$.) We will need to account and correct for this fact when taking our theoretical framework to data.

More formally, if we align in increasing order a sequence of information processing capacity parameters $\kappa^{(i)}$ for $i \in \{0, \ldots, m, \ldots, n\}$, we can associate each of them to some value of the variance parameter $\hat{\Sigma}_r^{(i)}$ within the universe of univariate Normal probability distributions, with, say, $\kappa^\sharp$ associated to what we have called true variance Σ_r (see Figure 4.2). Moreover, each of the variance values above induces a respective Normal probability density, $h_r(\hat{r}|\kappa^{(i)})$ or $g_r(r|\kappa^\sharp)$, with the former being either optimally biased or not (ones that are not optimally biased, or "centered", are reminiscent of the notion of "noncentral" distributions in the classical theory of probability distributions). Unfortunately, we do not know the $\kappa^{(i)}$ or $\hat{\Sigma}_r^{(i)}$ of any investor on the actual financial market. By necessity, our method of moving ahead is to rely on the fact that one of these variance values $\hat{\Sigma}_r^{(m)}$ must equal the sample variance $\hat{\Sigma}_r$ that we observe in the data (this is without loss of generality, as $\hat{\Sigma}_r^{(m)}$ may be arbitrarily close to true variance Σ_r). There is nothing special about investor indexed by m (for instance, we do not even know whether majority of the wealth is allocated by investors with κ above or below $\kappa^{(m)}$); rather because of him possessing information processing capacity just sufficient to work with the sample distribution, he happens to be a convenient working example. A valid empirical strategy must be flexible enough so as not to impose restrictions that contradict the logic of Figure 4.2.

This is the task of the following part, which will verify how consistent is the presented framework with empirical evidence.

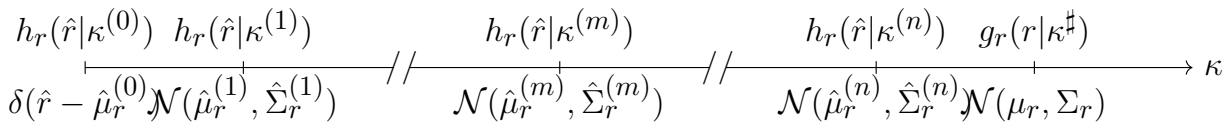


Figure 4.2: Family of probability distributions indexed by complexity level (distributions are univariate Normal; complexity measured by required information processing capacity κ, which is (ceteris paribus) in one-to-one correspondence to size of variance; variances rise from left to right, means may or may not be optimally biased).

CHAPTER 5

EMPIRICS

Now we are going to calibrate the (general equilibrium) consumption and portfolio choice model presented in §3 on U.S. consumption as well as government bill and stock returns data. We complement the model calibration results with a brief summary of further empirical support found in the diverse assortments of the existing literature.

5.1 Main calibrations

Key moments: The top panel of Table 5.1 provides key macro-financial facts on the United States after World War II. Following the discussion in §4.2, we postulate that these sample statistics characterize the approximating (joint) probability distribution via the moments of its marginals $\hbar_R(\hat{R}_{t+1})$, $\hbar_C(\hat{C}_{t+1})$ and $\hbar_D(\hat{D}_{t+1})$, that is the subjective picture revealed to and observed by econometricians. In terms of our framework, this is the solution to the feasible consumption and portfolio choice problem $\mathcal{P}_{\mathcal{QI}}$ of §3 for $K = 1$ with the simplified variance $\hat{\Sigma}_r < \Sigma_r$ and the matched expected return $\mathrm{E}^{\hbar}[\hat{R}_{t+1}] = \mathrm{E}^{g}[R_{t+1}]$.[1]

In the bottom panel of Table 5.1, we feed the empirically observed sample data $h_r(\hat{r}_{t+1})$ into the model solution equations of §3.3 under different assumptions about the information processing capacity κ and the coefficient of relative risk aversion γ. For the relevant probability density $\pi(\cdot)$, the risk-free rate there is calculated as

$$\mathrm{E}^{\pi}[R_0] := \frac{1}{\mathrm{E}^{\pi}[M_{t+1}]} \tag{5.1}$$

and the risk premium as

$$\mathrm{E}^{\pi}[R - R_0] := -\frac{\mathrm{Cov}^{\pi}[R_{t+1}, M_{t+1}]}{\mathrm{E}^{\pi}[M_{t+1}]}; \tag{5.2}$$

1. Our model abstracts away from labor income, although empirically it constitutes a large share of national income.

Table 5.1: Calibration Results

	$\mathrm{E}^{\hbar}[R_0]$ $\sqrt{\mathrm{V}^{\hbar}[R_0]}$	$\mathrm{E}^{\hbar}[R-R_0]$ $\sqrt{\mathrm{V}^{\hbar}[R]}$	$\mathrm{E}^{\hbar}[\Delta c]$ $\sqrt{\mathrm{V}^{\hbar}[\Delta c]}$	$\mathrm{E}^{\hbar}[\Delta d]$ $\sqrt{\mathrm{V}^{\hbar}[\Delta d]}$
Empirical, $\hbar$	0.88 1.26	8.23 17.05	1.94 0.99	2.43 4.23

	γ:	1			2			3			4			5		
	κ	$\mathrm{E}^{\cdot}[R_0]$ $\sqrt{\mathrm{V}^{\cdot}[R_0]}$	$\mathrm{E}^{\cdot}[R-R_0]$ $\sqrt{\mathrm{V}^{\cdot}[R]}$	$\mathrm{E}^{\cdot}[\Delta d]$ $\sqrt{\mathrm{V}^{\cdot}[\Delta d]}$	$\mathrm{E}^{\cdot}[R_0]$ $\sqrt{\mathrm{V}^{\cdot}[R_0]}$	$\mathrm{E}^{\cdot}[R-R_0]$ $\sqrt{\mathrm{V}^{\cdot}[R]}$	$\mathrm{E}^{\cdot}[\Delta d]$ $\sqrt{\mathrm{V}^{\cdot}[\Delta d]}$	$\mathrm{E}^{\cdot}[R_0]$ $\sqrt{\mathrm{V}^{\cdot}[R_0]}$	$\mathrm{E}^{\cdot}[R-R_0]$ $\sqrt{\mathrm{V}^{\cdot}[R]}$	$\mathrm{E}^{\cdot}[\Delta d]$ $\sqrt{\mathrm{V}^{\cdot}[\Delta d]}$	$\mathrm{E}^{\cdot}[R_0]$ $\sqrt{\mathrm{V}^{\cdot}[R_0]}$	$\mathrm{E}^{\cdot}[R-R_0]$ $\sqrt{\mathrm{V}^{\cdot}[R]}$	$\mathrm{E}^{\cdot}[\Delta d]$ $\sqrt{\mathrm{V}^{\cdot}[\Delta d]}$	$\mathrm{E}^{\cdot}[R_0]$ $\sqrt{\mathrm{V}^{\cdot}[R_0]}$	$\mathrm{E}^{\cdot}[R-R_0]$ $\sqrt{\mathrm{V}^{\cdot}[R]}$	$\mathrm{E}^{\cdot}[\Delta d]$ $\sqrt{\mathrm{V}^{\cdot}[\Delta d]}$
Model, $\hbar$	n/a	6.18 n/a	2.85 17.05	8.07 17.01	6.17 n/a	2.86 17.05	3.60 8.41	6.17 n/a	2.86 17.05	2.30 5.59	6.17 n/a	2.86 17.05	1.69 4.19	6.19 n/a	2.86 17.05	1.34 3.35
Model, g	0.1	−6.32 n/a	16.25 40.37	8.07 40.26	−6.35 n/a	16.27 40.36	1.99 19.22	−6.36 n/a	16.28 40.36	0.89 13.10	−6.36 n/a	16.28 40.36	0.50 9.81	−6.35 n/a	16.28 40.36	0.33 7.85
Model, g	0.2	0.36 n/a	8.78 29.80	8.07 29.73	0.34 n/a	8.79 29.14	2.87 14.64	0.33 n/a	8.80 29.80	1.66 9.73	0.33 n/a	8.80 29.80	1.16 7.29	0.35 n/a	8.80 29.80	0.89 5.83
Model, g	0.3	2.65 n/a	6.38 25.44	8.07 25.37	2.64 n/a	6.39 25.44	3.16 12.52	2.64 n/a	6.39 25.44	1.92 8.32	2.64 n/a	6.39 25.44	1.37 6.23	2.65 n/a	6.39 25.44	1.07 4.98
Model, g	0.4	3.80 n/a	5.21 23.01	8.07 22.95	3.79 n/a	5.22 23.01	3.30 11.34	3.79 n/a	5.22 23.01	2.05 7.53	3.79 n/a	5.22 23.01	1.48 5.64	3.80 n/a	5.22 23.01	1.16 4.51
Model, g	0.5	4.48 n/a	4.53 21.47	8.07 21.41	4.47 n/a	4.54 21.47	3.39 10.58	4.46 n/a	4.54 21.47	2.12 7.03	4.46 n/a	4.54 21.47	1.54 5.27	4.48 n/a	4.54 21.47	1.21 4.21

Notes: The top panel "Empirical" presents statistical moments for the real risk-free and market (excess) returns, the real per capita consumption growth as well as the real dividend growth. The data sample is U.S. 1948:Q1–2014:Q4, at quarterly frequency (see Appendix §E for a description of data sources). The bottom panel with "Model" rows presents the same statistical moments as above (the real per capita consumption growth is omitted being identically equal to the real dividend growth) for different combinations of the information processing capacity and the coefficient of relative risk-aversion. The CRRA utility, $\beta = 0.99$, and the probability density of returns ($\hbar$ or g) are the models' only inputs (see text for a detailed description). The measurement units are nats for information processing capacity, and percentage points converted into annualized terms for the economic variables.

where

$$M_{t+1} := \beta \frac{u'(C_{t+1})}{u'(C_t)} = \beta \frac{u'(D_{t+1})}{u'(D_t)} \tag{5.3}$$

is a (model-induced) stochastic discount factor (SDF), also called pricing kernel or marginal rate of substitution between consumption at time t and time $t+1$.

As a baseline, we start with the conventional approach that assumes the information constraint does not bind and $\hbar_r(\cdot)$ coincides with $g_r(\cdot)$; see the rows corresponding to "Model, $\hbar$" of Table 5.1. Specifically, we evaluate our model equations using as inputs the probability density of observed returns ($\hat{R} \sim \log \mathcal{N}(\hat{\mu}_r, \hat{\Sigma}_r)$ for the values of $\hat{\mu}_r$ and $\hat{\Sigma}_r$ estimated in sample) as well as the discount factor $\beta := 0.99$ for the plausible values of the RRA coefficient γ (implemented at quarterly frequency). It turns out, however, that the computed solution is not consistent with the rest of the observed data. Most striking are the low values obtained for the risk premium $\mathrm{E}^{\hbar}[R - R_0]$ and the high values for the risk-free rate R_0 for the realistic choices of parameter γ. These are the manifestations of the classic "equity premium puzzle" (Mehra and Prescott, 1985) and "risk-free rate puzzle" (Weil, 1989). From such a result we can infer that the probability distribution (and/or SDF specification) used in the actual investment decision-making differs from what we fed into the model.

Of course, if we take our information processing framework seriously, the above calibration results should not be surprising. Optimal, rational behavior should account for the usage of $\hat{\Sigma}_r$ in place of Σ_r. As discussed in §3.2 and §4.1, this can be addressed by adjusting the mean $\hat{\mu}_r$ and hence the expected return $\mathrm{E}^{\hbar}[\hat{R}_{t+1}]$ downward, which by the general-equilibirum logic shifts the demand from a risky toward a risk-free asset. The hurdle is that the mean bias term depends on unobserved true Σ_r, inferring which requires knowledge of the value of information processing capacity κ. The magnitude of the information processing capacity is not known, but by "reverse-engineering" logic we can calibrate it. Taking it as a free parameter, we search for the value of κ that produces Σ_r and thus $\hat{\mu}_r$ that lets the model fit empirical observations picked as testing criteria. Note, however, that in this calibration

exercise so as to facilitate comparisons with §5.2 later on, instead of adjusting downward the mean $\hat{\mu}_r$ of the distribution $\hbar_r(\cdot)$ (and thus obtaining the properly mean-adjusted optimal simplified distribution $h_r(\cdot)$), we follow a slightly different but equivalent in terms of outcomes route. We adjust upward the variance, i.e., back-out the true Σ_r (as well as true μ_r using the restriction $\mathrm{E}^g[R_{t+1}] = \mathrm{E}^{\hbar}[\hat{R}_{t+1}]$) and work directly with the putative distribution $g_r(\cdot)$.

The results obtained in the same manner as those for "Model, $\hbar$" earlier, but using as a probability density input the adjusted density $g_r(r)$, are placed in rows "Model, g" of Table 5.1. (For comparison, the unconstrained "Model, $\hbar$" effectively corresponds to "Model, g" with $\kappa \geq \kappa^\sharp := 3.4$.) In the vicinity of $\kappa = 0.2$ and $\gamma = 3$, the model results compare well with the data. The risk premium is slightly above 8% and the risk-free rate is slightly below 1% are close to empirically observed values. Of course, some words of skepticism are in order: the level of the risky asset's return was used as the model input. However, the ultimate split within this given level into the layer of the risk-free return and a risk premium on top is dictated by the model. Moreover, the mean and standard deviation of consumption growth (as well as dividend growth, since in our model featuring just one risky asset they coincide) are at low and high single-digit values, respectively, which looks plausible in comparison to empirical yardsticks; and this time no direct relation to model inputs can beget any skepticism.[2] The result of the adjustment that transforms the sample, or simplified probability density function $\hbar_r(\hat{r})$ into what we argue is an estimate of a population, or true probability density $g_r(r)$ is available on the left panel of Figure 5.1.[3]

2. The true distribution's mean consumption (or dividend) growth obtained in our calibration is 1.66%, which is not too far from empirical values of 1.94% (2.43%), adding plausibility. The corresponding standard deviation is 9.73%, which is naturally larger than 0.99% (4.23%) observed in sample, but is comparable to alternative results. Specifically, it is lower than the 17% standard deviation of welfare equivalent consumption growth in Weitzman (2007); it falls in between the 4.24% and 14.84% standard deviations of, respectively, consumption and dividend growth over 600,000 years of simulated data featuring rare booms and disasters in Tsai and Wachter (2016); it is also consistemt with the finding in Malloy et al. (2009) that consumption growth of stockholders exhibits a sensitivity of a factor of about 3 to 4 relatively to aggregate consumption growth.

3. Given that the true distribution $g_r(r)$ is not known and what we work with is just its calibration/estimate, one may find it more apt to denote variable r as $\hat{r}$.

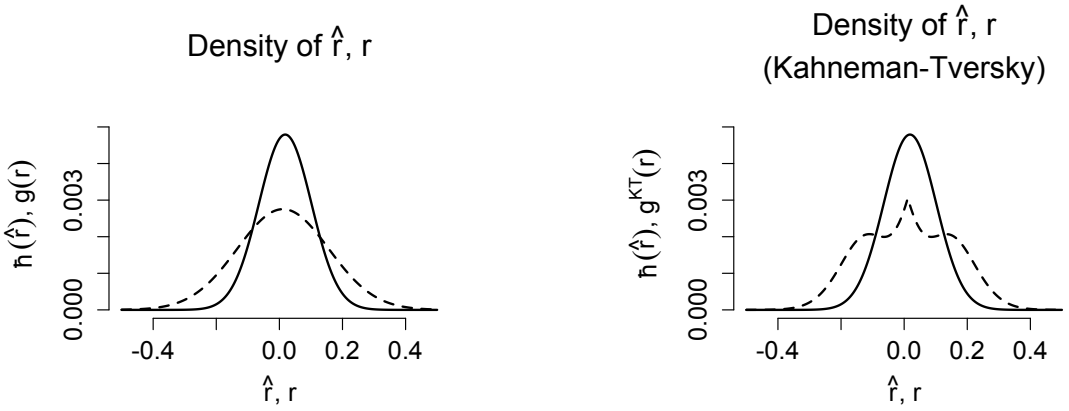


Figure 5.1: Empirical (solid) vs. adjusted (dashed) probability densities (parameterizations used are $\hat{r} \sim \mathcal{N}(\hat{\mu}_r, \hat{\Sigma}_r)$ with $\hat{\mu}_r := 0.0184$ and $\hat{\Sigma}_r := 0.0069$ vs. $r \sim \mathcal{N}(\mu_r, \Sigma_r)$ with $\mu_r := 0.0114$ and $\Sigma_r := 0.0210$, corresponding to "Model, g" from Table 5.1 with $\kappa = 0.2$ and $\gamma = 3$ (left panel); $\hat{r}$ as above vs. $r \sim g_r^{KT}(\mu_r, \Sigma_r)$ with $\mu_r := 0.0115$ and Σ_r determined by CPT-style adjustment, corresponding to "Model, g^{KT}" from Table 5.3 with any value of γ (right panel)).

Additional indicators: Table 5.2 provides some additional macroeconomic and financial data on the Unites States after World War II, together with the corresponding characteristics for the model specifications emphasized in the previous paragraph. One additional indicator is the correlation between the returns and consumption growth, which, being notably below 1, compares adequately to its empirical counterpart (in view of the fact that we are considering just a single risky asset).

The two other added indicators are the Sharpe ratio and the Hansen-Jagannathan bound. The first defined as

$$\mathrm{SR}^{\cdot} := \frac{\mathrm{E}^{\cdot}[R] - R_0}{\sqrt{\mathrm{V}^{\cdot}[R]}} \leq \frac{\sqrt{\mathrm{V}^{\cdot}[M]}}{\mathrm{E}^{\cdot}[M]} =: \mathrm{HJ}^{\cdot}, \tag{5.4}$$

where the bound on the right-hand side defines the second one. The former quantity reflects the expected risk-return trade-off. The latter was formulated by Hansen and Jagannathan (1991) basing on the statistical properties of a valid SDF; it serves an important role in the diagnostics of the asset-pricing models, often posing an impenetrable barrier for a theoretical Sharpe ratio to catch up with its empirical measurements.

Table 5.2: Calibration Results, with Additional Details

	κ	γ	$\mathrm{E}^{\cdot}[R_0]$	$\sqrt{\mathrm{V}^{\cdot}[R_0]}$	$\mathrm{E}^{\cdot}[R-R_0]$	$\sqrt{\mathrm{V}^{\cdot}[R]}$	$\mathrm{E}^{\cdot}[\Delta c]$	$\sqrt{\mathrm{V}^{\cdot}[\Delta c]}$	$\mathrm{E}^{\cdot}[\Delta d]$	$\sqrt{\mathrm{V}^{\cdot}[\Delta d]}$	$\rho^{\cdot}(r,\Delta c)$	$\mathrm{SR}^{\cdot}$	$\mathrm{HJ}^{\cdot}$
Empirical, $\hbar$	n/a	n/a	0.88	1.26	8.23	17.05	1.94	0.99	2.43	4.23	0.18	0.23	n/a
Model, g	0.2	1	0.36	n/a	8.78	29.80	8.07	29.73	8.07	29.73	1.00	0.14	0.15
Model, g	0.2	2	0.34	n/a	8.79	29.80	2.87	14.64	2.87	14.64	0.99	0.14	0.15
Model, g	0.2	3	0.33	n/a	8.80	29.80	1.66	9.73	1.66	9.73	0.92	0.14	0.15
Model, g	0.2	4	0.33	n/a	8.80	29.80	1.16	7.29	1.16	7.29	0.85	0.14	0.15
Model, g	0.2	5	0.35	n/a	8.80	29.80	0.89	5.83	0.89	5.83	0.78	0.14	0.15
Model, g	0.1	3	-6.36	n/a	16.28	40.36	0.89	13.10	0.89	13.10	0.92	0.19	0.20
Model, g	0.2	3	0.33	n/a	8.80	29.80	1.66	9.73	1.66	9.73	0.92	0.14	0.15
Model, g	0.3	3	2.64	n/a	6.39	25.44	1.92	8.32	1.92	8.32	0.92	0.12	0.12
Model, g	0.4	3	3.79	n/a	5.22	23.01	2.05	7.53	2.05	7.53	0.92	0.11	0.11
Model, g	0.5	3	4.46	n/a	4.54	21.47	2.12	7.03	2.12	7.03	0.92	0.10	0.11

Notes: The columns present the information processing capacity, the coefficient of relative risk-aversion; the real risk-free and market (excess) returns, the real per capita consumption growth as well as the real dividend growth with their statistical moments; the Sharpe Ratio denoted as "SR", and the Hansen-Jagannathan bound denoted as "HJ". The "Empirical" row: the data sample is U.S. 1948:Q1–2014:Q4, at quarterly frequency (see Appendix §E for a description of data sources). The "Model" rows: the CRRA utility, $\beta = 0.99$, and the probability density of returns ($\hbar$ or g) as the models' only inputs (see text for a detailed description). The measurement units are nats for the information processing capacity, and percentage points converted into annualized terms for the economic variables.

A Sharpe ratio around 0.14 is below the empirically observed value of 0.23, but is closer than what is the case for the baseline "Model, $\hbar$" of Table 5.1, where the Sharpe ratio and the Hansen-Jagannathan bound are both equal to 0.08 (not shown in the tables). Of course, the improved performance of the latest model is natural, looking at the adjustment through the lens of Hansen-Jagannathan bound. The process of entropy reduction has reduced the relevant variances, including that of the SDF, which directly leads to a decrease in the numerator of the HJ bound; while reversing the reduction in variance via appropriate adjustment increases the variance of the SDF and improves model's performance as far as the HJ bound is concerned.[4]

Interpretation: Our preferred interpretation of these calibration results is as follows. The model we use is deliberately primitive, even crude perhaps (although truly respecting the general equilibrium discipline; and with fully optimizing agents, including their information processing side which produces the probability distributions they use for making investment decisions). It takes as inputs only the probability distribution of the risky asset's returns, a commonly accepted level of the subjective discount factor β, and a RRA coefficient γ that is allowed to take one of the plausible values. The probability distributions we consider are the empirically observed distribution of returns (which we assume to take a parametric log-Normal form with parameters estimated from available sample data) and the inferred "true" (population) probability distribution of returns that is computed conditionally on the hypothesized information processing capacity κ inherent to market participants. Anyone who takes the (joint) empirical probability distribution at face value (as "naïvely" does an econometrician, who uses the sample variance $\hat{\Sigma}_r$ and the unbiased sample expected return $\mathrm{E}^{\hbar}[\hat{R}_{t+1}] = \mathrm{E}^{g}[R_{t+1}]$, though does not participate on the market) finds it inconsistent with the simple model outlined above, in particular he ends up being puzzled by "high" premium

4. Strictly speaking, matching Sharpe ratios is disputable as a relevant yardstick, because in contrast to traded assets, "disagreeing" with a Sharpe ratio does not provide investors with incentives to initiate the opposite investment/trade position.

on risky assets and "low" return on riskless assets.

On a closer look, however, this is just an illusion. More sophisticated agents ("professional" investors, who are themselves active market participants) recognize the fact that the stochastic environment represented by the population probability distribution is more complex than what an empirical probability distribution makes of it. Inferring their beliefs about the unobserved complex "true" distribution indirectly allows to rationalize the actual consumption and investment choices leading to observed (and erroneously perceived as "high") premium on the risky assets and (correspondingly, "low") return on the risk-free assets. Even though investors may disagree with the simplistic empirical probability distributions and consider them inappropriate for investment decisions, the observed market levels of risk-free and risky returns do not look off from their perspective (since the latter involves adoption of certain adjustments in actual decision-making).

Quantitatively, the key output of our calibrations is the estimate of the information processing capacity κ. The results in Table 5.1 suggest that the model fits the observed data best with the RRA parameter $\gamma = 3$ and the information processing capacity around

$$\kappa = 0.2 \text{ nats} \approx 0.3 \text{ bits}.$$

At first sight, the value we obtained for κ seems implausibly low. However, in support of this measurement, note that (i) it corresponds to a differential entropy whose magnitudes may look counter-intuitive, even though the induced probability adjustments are clearly substantial (cf. density plots on the left panel of Figure 5.1); (ii) κ is a measure of effective rather than available full physical capacity $\mathcal{K}^*$, which in principle may be orders of magnitude larger (see Appendix §G for explanation of the difference); (iii) experimental studies in psychology and neuroscience have been arriving at rather low measurements of information processing capacity as well (e.g., see the classical paper by Miller (1956), whose very title "The Magical Number Seven, Plus or Minus Two" reflects how small the estimates of such

capacity in bits—for discretely distributed information in that case—are). Lastly, to put it into context, the magnitude of the effective capacity that we obtained constitutes about 5% of the magnitude minimally required by the unconstrained benchmark "Model, $\hbar$".

The calibrated value of κ can be interpreted as the effective information-processing capacity of a representative investor who makes decisions using a sample distribution as his simplified probability distribution (and adopting optimal adjustments required to account for the discrepancy between the simplified/sample and the original/population distributions). (Of course, this is suggested just for convenience; individual market participants may have information-processing capacities larger or smaller than κ above, implying that their simplified distributions may be relatively more or less complex than the sample one, with corresponding modifications to their bias adjustments.)

Note that it may seem equally sensible to view the κ parameter we have calibrated not as the information-processing capacity of agents who use sample distribution as a simplified one, but just as a quantification of the distance between population and sample distributions (with the measure of mutual information used as a distance metric). However, such a "black-box" non-structural interpretation would rule out the very framework we use to digest and incorporate the entropy-reduction experiments from part §2 as well as to justify empirical regularities such as (variations in) style investing later in part §5.3.

Furthermore, these calibration results allow us to assess the welfare losses of a representative investor (holding the U.S. market portfolio as proxied by S&P500 Composite) who relies on the sample distribution together with the appropriate mean adjustment: using definition (3.6) as well as Propositions 3 and 5, annualized ex ante mean deviation in the level of (log) value function constitutes 1.02% of its current level.

5.2 Cross-checking with experimental evidence

The calibrations in the preceding part could be criticized as not necessarily following any particular theoretical framework, but rather reflecting an addition of an extra free parameter

that cancels out and stands in for some fundamental errors (e.g., errors in model specification or parametric assumptions). That the κ we have calibrated is just an atheoretical plug-in parameter that has nothing to do with the hypothesized information processing capacity (although then it would be quite surprising that the same one extra parameter helps fitting more than one additonal empirical facts, i.e., also the consumption/dividend moments).

Our theoretical case is strengthened if we can demonstrate that the magnitude of the information processing capacity that has arisen in the above calibrations is consistent with alternative measurements. Specifically, with those obtained independently in contexts that are similar to ours and are also related to the phenomena of human information processing. The data from laboratory experiments with human subjects gathered and summarized by Kahneman and Tversky as well as many other authors provide such an alternative independent source of relevant measurements.

5.2.1 Calibrations utilizing experimental measurements

We are going to use the stylized facts gathered in numerous laboratory experiments not directly, but through the lens of the prospect theory of Kahneman and Tversky that summarizes them very accurately and compactly.

Prospect theory and the connection: The prospect theory (Kahneman and Tversky, 1979) together with its refinement, the cumulative prospect theory (Tversky and Kahneman, 1992), applies to choice under risk and uncertainty and captures vast experimental evidence collected on such choices, offering robust empirical predictions. In particular, it is consistent with Allais' (1953) experimental results that challenge the conventional expected utility theory of choice. However, as the authors themselves emphasize, the (cumulative) prospect theory is not normative/prescriptive, but rather positive/descriptive, purely phenomenological. It's modern interpretation is that this is not a preference theory, but a theory of "default actions" that are "appropriate to maximize experienced utility only on average" (Bossaerts

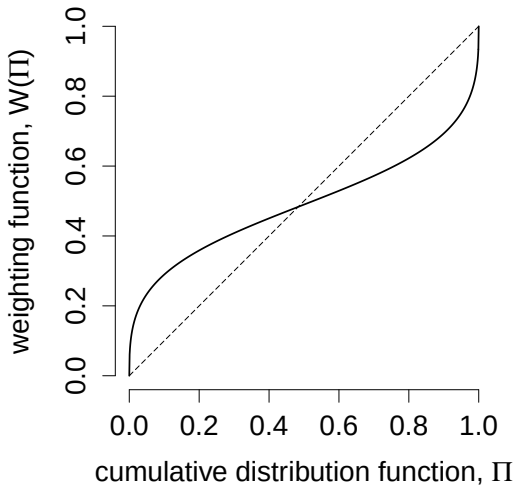


Figure 5.2: Cumulative prospect theory's probability weighting function primer.

et al., 2008), and which are actively used by less experienced agents (see Fox et al., 1996; List, 2004; also see List and Haigh, 2005).

The (cumulative) prospect theory stipulates that (after the initial framing stage) the overall value (or utility) of an uncertain "prospect" is a sum of values of the outcomes each multiplied by the decision weight. The outcome is defined with respect to a reference point, the value function potentially differs for positive ("concave for gains") and negative ("convex and steeper for losses") deviations from the reference point, and that the decision weights possibly do not coincide with the presented probabilities (they are not even necessarily being additive to 1). It is the decision weight aspect that we are chiefly interested in. In the leading formulation of the theory, the decision weight function is defined as a non-linear transformation of the cumulative probability distribution function of the outcome, such that it overweights the small probabilities and underweights the moderate and large ones (see Figure 5.2 for a primer).

The value function-related prescriptions of the prospect theory can be reconciled with the (fairly standard) felicity measurement side of our model by designating as a reference point $W_t = 0$ in the value function and equivalently $C_t = 0$ in the utility function (closing down the issue of asymmetric treatment of gains and losses due to the non-negativity of W_t

and C_t).

Since entropy reduction results in relative underweighting of the low probabilities, their relative overweighting prescribed by the prospect theory at least partially cancels out the distortion (we will provide a more detailed expositon of this argument later). Assuming that the degree of adjustment measured during the experiments of Kahneman and Tversky indeed reveals the optimal (in constrained, second-best, or Herbert Simon's "ecologically rational", sense) behavior and reflects the average κ that constrains humans in real-life investment decisions, the effect of probability weighting allows to estimate Σ_r and, in turn, indirectly calibrate κ.

We take the parametric measurements of the (cumulative) prospect theory's probability weighting function from Camerer and Ho (1994), who use data from nine different experimental studies to estimate the parameters we are interested in. The functional form they use is

$$W(\Pi) := \frac{\Pi^\eta}{(\Pi^\eta + (1-\Pi)^\eta)^{1/\eta}},$$

where $\Pi(\cdot)$ is some cumulative distribution function, $W(\cdot)$ is the cumulative probability weight for a given partition of the probability space, and $\eta \leq 1$ is a parameter, with $\eta = 1$ corresponding to the standard linear probability weighting. Camerer and Ho's estimate of η is 0.56.

Application: Now, we apply the prospect theory-style probability weighting adjustment using the functional form and parameter estimate of Camerer and Ho (1994). As a reference value, with respect to which the adjustment procedure determines what probability quantile a particular realization belongs to, we use the mean of the distribution we end up with, μ_r (which is obtained by finding such a value for μ_r that would ensure the expected value of return is preserved, $\mathrm{E}^g[R_{t+1}] = \mathrm{E}^{\hbar}[\hat{R}_{t+1}]$). The right panel of Figure 5.1 shows the end product of the probability weighting adjustments implied by the prospect theory of Kahneman and Tversky, depicted as $g_r^{KT}(r)$, along with the sample distribution, $\hbar_r(\hat{r})$.

Visually, the similarity between the two putative objective distributions $g_r(r)$ and $g_r^{KT}(r)$ is striking, their location and scale parameters seem well attuned. A quantitative comparison is provided by Table 5.3, which presents the results of an exercise analogous to the one conducted for "Model, g" with $\kappa = 0.2$ and $\gamma = 3$ in Tables 5.1 and 5.2, but this time using the distribution $g_r^{KT}(r)$ that is produced following the stipulations of the prospect theory. For $\gamma = 3$, the procedure of Kahneman and Tversky turns out the results that roughly match those we got in the calibrations of §5.1.

Taking Σ_r and $\hat{\Sigma}_r$ from "Model, g^{KT}" with RRA parameter $\gamma = 3$ as well as using Propositions 4 and 5 (including results in the proof of the former proposition),[5] or just relying on the evident from above proximity between two sets of calibration results, we arrive at the same estimate of the information processing capacity as earlier in §5.1, i.e., $\kappa = 0.2$ nats ≈ 0.3 bits.[6]

Next, we are going to interpret the calibrations based on the above experimental evidence and discuss how they fit into the broader framework.

5.2.2 Interpretation of paradoxical behavior observed in experiments

The consistency between the results in Tables 5.1–5.2 and Table 5.3 can be taken as a verification of the measurements in §5.1. Moreover, we argue that these two sets of calibrations

5. Strictly speaking, this method gives an upper bound: given Normal $\hbar_r(\hat{r})$, we calculate κ under the condition that $g_r^{KT}(r)$ is also Normal, but the probability adjustment produces a distribution that is clearly not Gaussian (as can be seen from the right panel of Figure 5.1), which means that for the given variance, it has lower entropy than the Gaussian (due to the latter's maximum-entropy property), and thus the required κ is in fact lower. (If we knew that the approximation error for the probability-adjusted distribution is also additive, we could use deconvolution procedure to deduce the error's distribution and then compute κ using the usual formula for mutual information; but this is not the case.) As it turns out, for our application this bound is pretty tight.

6. To finish with this part, it is worth noting that a large experimental study by Bruhin et al. (2010) provides additional estimates of the cumulative prospect theory's probability weight parameters. However, its focus is on individual heterogeneity, and it is hard to aggregate the results across different risk-taking types in a sensible way. Their measurements for the subjects that exhibit a significant deviation from the linear probability weighting differ from the aggregated results of Camerer and Ho (1994) and produce a larger adjustment of the variance upward, resulting in an even higher equity premium and lower risk-free rate. This potentially gives us some idea about the variability and upper bounds for the adjustments we are considering here.

Table 5.3: Calibration Results for Kahneman-Tversky Approach, with Additional Details

	κ	γ	$\mathrm{E}^{\cdot}[R_0]$	$\sqrt{\mathrm{V}^{\cdot}[R_0]}$	$\mathrm{E}^{\cdot}[R-R_0]$	$\sqrt{\mathrm{V}^{\cdot}[R]}$	$\mathrm{E}^{\cdot}[\Delta c]$	$\sqrt{\mathrm{V}^{\cdot}[\Delta c]}$	$\mathrm{E}^{\cdot}[\Delta d]$	$\sqrt{\mathrm{V}^{\cdot}[\Delta d]}$	$\rho^{\cdot}(r,\Delta c)$	$\mathrm{SR}^{\cdot}$	$\mathrm{HJ}^{\cdot}$
Empirical, $\hbar$	n/a	n/a	0.88	1.26	8.23	17.05	1.94	0.99	2.43	4.23	0.18	0.23	n/a
Model, g	0.2	3	0.33	n/a	8.80	29.80	1.66	9.73	1.66	9.73	0.92	0.14	0.15
Model, g^{KT}	n/a	1	0.63	n/a	8.47	29.19	8.05	29.11	8.05	29.11	1.00	0.14	0.14
Model, g^{KT}	n/a	2	0.70	n/a	8.43	29.10	2.92	14.34	2.92	14.34	1.00	0.14	0.14
Model, g^{KT}	n/a	3	0.71	n/a	8.42	29.09	1.71	9.53	1.71	9.53	0.90	0.14	0.14
Model, g^{KT}	n/a	4	0.70	n/a	8.41	29.06	1.19	7.13	1.19	7.13	0.77	0.14	4.31
Model, g^{KT}	n/a	5	0.75	n/a	8.39	29.03	0.92	5.69	0.92	5.69	0.63	0.14	3193.06

Notes: The columns present the information processing capacity, the coefficient of relative risk-aversion; the real risk-free and market (excess) returns, the real per capita consumption growth as well as the real dividend growth with their statistical moments; the Sharpe Ratio denoted as "SR", and the Hansen-Jagannathan bound denoted as "HJ". The "Empirical" row: the data sample is U.S. 1948:Q1–2014:Q4, at quarterly frequency (see Appendix §E for a description of data sources). The "Model" rows: the CRRA utility, $\beta = 0.99$, and the probability density of returns ($\hbar$, g or g^{KT}) as the models' only inputs (see text for a detailed description). The measurement units are nats for the information processing capacity, and percentage points converted into annualized terms for the economic variables.

are not just related, but represent different sides of the same coin.

Variance counter-adjustment: The reduction in information processing costs accomplished by the presented approach attains the constrained optimality only under accurate implementation of the procedure's components. For instance, the variance-decreasing simplification requires a corresponding adjustment in the mean. Failure to implement such an adjustment correctly results in suboptimal decision biases and potentially avoidable utility losses. That was the first, rational method of adjustment.

Alternatively, recall the variance decomposition equation from Proposition 5:

$$\boldsymbol{\Sigma}_r = \hat{\boldsymbol{\Sigma}}_r + \boldsymbol{\Psi}_r.$$

As can be clearly seen here, adding a (not too large) positive semi-definite matrix to the simplified variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_r$ would move it closer (in any reasonable matrix norm sense) to the original matrix $\boldsymbol{\Sigma}_r$. Thereby, the variance reduction can be compensated for by the variance counter-adjustment undoing the reduction, at least partially, rather than by an adjustment to the mean.

One way of implementing such a variance adjustment can be the mechanics uncovered by Kahneman and Tversky (1979, 1992). The weighting function transformation of the outcome probabilities entailed by the prospect theory effectively replaces the presented probabilities with adjusted ones, where the latter much resemble the former reshaped into a flatter form, i.e., move closer toward a uniform distribution that possesses comparatively higher entropy and a larger variance. Such a transformation of probabilities reflects the commonly exhibited in laboratory studies behavior known as the Allais paradox. From the perspective of our approach, the probability weighting-related stipulations of Kahneman and Tversky's theory may be interpreted as a sort of variance counter-adjustment, that is as an attempt to recover (and utilize in decision-making) the true, relatively higher-entropy variance after having simplified it (due to considerations of information processing costs) to its approximate lower-

entropy version, i.e., after moving it away from the uniform distribution.

In terms of information-processing costs, using a relatively large true variance is more expensive (as explained in Remark 2) than the simplified alternative of manipulating the mean adjustment term that was considered earlier. However, an accurate calculation of the former object is itself not computationally costless.[7] Thus, some agents would find the simplified decision-making procedure presented in §2 and §3 still too hard to implement. For this reason, alternatives that are computationally even cheaper, albeit at the expense of larger approximation errors, are of great practical interest.

Indeed, from our perspective the adjustment implied by the prospect theory can be viewed as a shortcut heuristic (cf. Gigerenzer et al., 2000; Gilovich et al., 2002), a suboptimal but computationally cheap technique that is deployed at the penultimate stage of decision making after solving the informational sub-problem and before proceeding to consumption and investment sub-problem. In this case, solving the consumption-investment problem still involves operations with high-entropy objects, but—with default prospect theory probability adjustments—those can be executed with "default", efficiently implemented methods. This is the second, heuristic method of adjustment.

Therefore, on the one hand the experimental evidence provides an alternative measurement of the information-processing capacity κ that exhibits the consistency we were seeking as a disciplining cross-check on (or cross-validation of) the calibrations in §5.1; as well as offers a practical example of the mechanism that implements one form of adjustment to the decision-making landscape that becomes necessary when the above capacity is actually binding.

7. The magnitude of the mean bias term depends on $\boldsymbol{\omega}_t$, the shares of wealth invested in different risky assets, which are not known in advance and whose accurate calculation requires a considerably long (when started far from the optimum) sequence of iterations that Appendix C.3 is devoted to. (Again, adopting inaccurate mean adjustment due to, say, only infrequent revision of the bias term or ignoring the term altogether, results in subsequent utility losses.)

Rationalization of paradoxical experimental evidence: On the other hand, the very consistency of the findings in experiments with the calibrated version of our framework also says something about the former, that is characterizes the paradoxical experimental behavior itself. For instance, since it effectively produces much the same decision-making landscape as when an investor uses the sample data and then follows our optimal adjustment procedure, the heuristic of Kahneman and Tversky's prospect theory can be interpreted as a procedure tailored to the practical needs of agents who rely on the sample data in their decision-making. Actually, an interpretation along these lines may explain, that is provide an optimizing basis and thus "rationalize" the experimental evidence on the behavior of human subjects uncovered by Allais and embodied in the prospect theory of Kahneman and Tversky. In such a case, the magnitude of κ we have measured reflects effective information processing capacity of an "average" agent who is sampling the environment and using the collected information (accumulated experience) to make optimal (or rather "ecologically rational") decisions; i.e., in our general approach of a representative investor using a representative sample of market data and trying to come up with (constrained) optimal decisions. The above logic agrees with the modern interpretation of the prospect theory as a theory of "default actions" that are optimal only on average (Bossaerts et al., 2008).

Direction of causality: To be fair, the direction of cause and effect can conceivably be reversed. In the classical, optimizing treatment above we take as given the unobserved large-entropy "true" (i.e., population) probability distribution, which is then approximated by its simplified counterpart (such as sample distribution). While the behavior leading to the Allais paradox as well as Kahneman and Tversky's probability adjustments play a purely instrumental role, they are just heuristics tailored exactly to those agents who use the sample distribution as the simplified distribution in their decision-making.

At the same time, as has already been mentioned earlier, we can not dismiss the Bayesian-fashioned treatment of the whole problem, where the unobserved high-entropy "true" distri-

bution is based on some (prior) beliefs (which do not have to be "right", just consistent with a Nash equilibrium, perhaps a self-fulfilling one). And we take as given the Allais paradox as well as the prospect theory of Kahneman and Tversky that reflects these intrinsic beliefs directly by summarizing and embodying their behavioral manifestations, thus in the big scheme of things being a cause rather than an effect, a source generating the high-entropy "true" distribution rather than a means of restoring it from the available approximating distribution. The latter treatment is closer to the spirit of Kahneman and Tversky themselves, originally at least.

Actually, in (unconditional) equilibrium these two treatments, the classical adjustment-motivated (which entails modern, optimizing interpretation of the experimental evidence and the prospect theory) versus Bayesian-flavored beliefs-based one (which implies more conventional, either-gross-mistakes-or-bizarre-preferences-cum-priors view on Allais' and Kahneman-Tversky's findings), are equivalent operationally for investors and observationally for econometricians. However, from a dynamic standpoint, the time-conditioned equilibria are potentially distinguishable, which allows us to formulate a testable prediction: the classical interpretation presumes that over time as more data are accumulated the sample probability distribution will converge toward the currently unobserved true, population distribution, while under the Bayesian-like interpretation relying on intrinsic beliefs the sample distribution will remain essentially unchanged and permanently differ from the "true" distribution (as long as the beliefs themselves are fixed). Testing and breaking this dichotomy is a topic for a separate paper, but seems within reach given a long enough observation history.

As a matter of fact, some existing works suggest the sample distribution is not representative and bound to evolve (e.g., see McGrattan and Prescott, 2003 and 2005; Dimson et al., 2003; Fama and French, 2002). Which favors the classical interpretation. On the other hand, it's prudent to take the classical story with a grain of salt, i.e., that such a close match of calibration results between Tables 5.1–5.2 and Table 5.3 is not entirely due to a representative investor adjusting the observed sample distribution to some exogenously

given unobserved population distribution (how likely it is that people's probability-weighting parameters have been accurately learned or evolutionarily optimized for post-World-War-II sample of financial returns?), hence at least some effect in the reverse direction is plausible. Which adds weight to the Bayesian-fashioned interpretation. (Implicitly, we are ruling out the third alternative, a pure coincidence.) Overall, some degree of self-reinforcing two-way interaction is probably at play here.[8]

As a final remark, widely accepted Stein-type "shrinkage" techniques (e.g., see Ledoit and Wolf, 2004a; Jagannathan and Ma, 2003) can be reinterpreted in terms of our framework as yet another way of implementing variance counter-adjustment that recovers the original variance-covariance matrix $\mathbf{\Sigma}_r$ from the simplified matrix $\hat{\mathbf{\Sigma}}_r$. This point is elaborated further in Appendix §L.

5.3 Some further empirical results

Next, we briefly talk about further empirical results consistent with the presented theoretical framework. These empirical results are based on the existing literature, this dissertation's contribution is to provide a unifying theoretical framework that may explain them, thus accumulating additional supporting evidence for our work. We particularly emphasize the empirical findings in the literature associated with probability weighting transformations as in the (cumulative) prospect theory of Kahneman and Tversky: the intimate interrelationship of our framework with Kahneman and Tversky's account has been demonstrated in §5.2 above, and now their theory can also provide a bridge to a lot of useful quantitative empirical results.

First, our result regarding the subjective ampification of the correlations between different

8. Historically, this is not the first work that refers to the prospect theory-style probability weighting in rationalizing the observed equity premium and related empirical phenomena. The others include an early attempt by Epstein and Zin (1990), as well as Routledge and Zin (2010), De Giorgi and Legg (2012), also notwithstanding the related contribution of Barberis and Huang (2008). However, all of the other works represent the latter, beliefs-based—as opposed to the former, adjustment-motivated—view on the direction of causality.

risky assets has a number of practical consequences.

Correlation inflation leads to apparent **"underdiversification"**, meaning that from a subjective perspective investment portfolios may look less diversified than they actually are (recall how correlation coefficient rises from the true value of 0.63 to the approximating level of 0.90 in the simple illustration of §4.1). This generates the familiar "portfolio concentration/underdiversification puzzle" (see Blume and Friend (1975), Statman (1987), Kelly (1995) for individual investors' domestic portfolios; French and Poterba (1991) for country-level international portfolios) in a way fundamentally related to the equity premium puzzle that we analyzed above; with the basic resolution being that using true, higher variance would reduce the assets' (true) cross-correlations and make further diversification less attractive than it erroneously seemed before.[9]

Quantitatively speaking, the work of Polkovnichenko (2005) shows that such empirical observations can be explained by probability weighting of the type stipulated by Kahneman and Tversky's theory, and are thus in agreement with current dissertation's framework (via the above-mentioned intimate connection between the two).

In the case of positively correlated assets, their subjective clustering into **asset classes** emerges endogenously, since correlation inflation entails categorization. For example, stocks in the Australian mining company BHP Billiton and Chicago Mercantile Exchange futures contracts on crude WTI oil may have a "true" correlation of returns well below 1, and yet be subjectively viewed by investors as correlated more tightly than that and treated as a single asset class "commodities"; while shares of U.S. companies with very different business fundamentals may be mechanically merged into an asset class "small value" or "technology" stocks. Such effects engender the (self-reinforcing) popularity of operating

9. In spite of being a subjective problem of perspective, it may have real consequences for an investor that relies on the simplified variance $\hat{\mathbf{\Sigma}}_r$ but the matched expected return $\mathrm{E}^h[\hat{\boldsymbol{R}}_{t+1}] = \mathrm{E}^g[\boldsymbol{R}_{t+1}]$, and maximizes portfolio return subject to a constraint on the accepted level of portoflio variance. In general, he will underappreciate the benefits of diversification: in a simple example, expanding a portfolio from one asset with variance $\hat{\sigma}_r$ to a portfolio split equally between two assets with equal variances $\hat{\sigma}_{r,1} = \hat{\sigma}_{r,2} := \hat{\sigma}_r$ and correlation coefficient $\hat{\rho}_{r,12}$ lowers the portfolio variance by $0.5\hat{\sigma}_r(1 - \hat{\rho}_{r,12})$, with the latter quantity contracting as $\hat{\sigma}_r$ falls and $\hat{\rho}_{r,12}$ rises further.

in terms of aggregated asset classes instead of disaggregated assets among investors and econometricians alike, fueling the interest in "asset allocation", "asset comovement" and "style investing" (see Sharpe, 1992; Brinson et al., 1986 and 1991; Doeswijk et al., 2014; Fama and French, 1993; Barberis and Shleifer, 2003; Barberis et al., 2005).

A straightforward theoretical prediction of our approach is that agents with lower information processing capacity will be relatively more predisposed to such clustering, manifesting stronger real effects. In turn, more sophisticated investors tend to be professional market participants who, due to competitive forces in the labor market, are supposed to be less capacity-constrained and less prone to suboptimal decisions, thus enabling empirical identification. The style investing phenomenon is very well studied in the literature, it provides enough evidence and variation to verify this claim. Indeed, there are ample confirmations of such clustering, it has a real effect on demand and outcomes, and that effect goes over and above the class member's true characteristics, its underlying fundamentals (e.g., Pindyck and Rotemberg, 1990; Chan et al., 2000; Teo and Woo, 2004; Froot and Teo, 2008; Choi and Sias, 2009). Moreover, unsystematic findings scattered in several style investing papers provide empirical evidence that correlation, or herding, of investment decisions is stronger within less sophisticated retail investors than it is among more sophisticated institutional investors. An explicit comparison confirming this result is conducted in Kumar and Lee (2006) as well as, more prominently, in Jame and Tong (2014), with the latter work reporting the values for a popular measure of such herding in two market constituencies we are concerned about at 4.01% and 2.09%, respectively. This is consistent with the above theoretical prediction.

In the case of negatively correlated assets, they may endogenously form subjective **hedging instruments**. For instance, portfolios of government bonds and portfolios of stocks may have a slightly negative "true" correlation of returns in some regimes/time periods, usually involving so-called "flight-to-quality" episodes (Li, 2002; Connolly et al., 2005; Guidolin and Timmermann, 2007; Andersson et al., 2008; Yang et al., 2009), but subjective amplification of such correlations could be responsible for an often-held view of bonds serving the role of a

hedge for stocks (for examples, see Canner et al., 1997). Such hedging motives are the main focus of the "strategic asset allocation" literature, e.g., see Brennan et al. (1997), Campbell and Viceira (2002b), also refer to Wachter (2010) for a more recent review paper.

Second, another result related to overconfidence may be responsible for **risk-aversion/ implied volaility "smile"** pattern that is observed empirically in the options segment of financial markets. Fundamentally, this "non-monotone pricing kernel puzzle" is a phenomenon that proves to be challenging for conventional theoretical models to convincingly address (e.g., an overview can be found in Ziegler, 2007). However, the fact that the simplified variance-covariance matrix of returns is smaller than its original counterpart, and one way to counteract this is to adjust it upwards—for instance, using probability weighting as stipulated by the prospect theory of Kahneman and Tversky—has direct relevance to the above phenomenon in finance.

Specifically, Kliger and Levy (2009) show that prospect theory-style probability weighting helps explain observed options prices. In contrast to the parametric approach taken in the previous paper, Polkovnichenko and Zhao (2013) use an agnostic non-parametric procedure to estimate an empirical SDF as well as a probability density function from options data, and then (conditionally on utility function specification assumptions) extract empirical probability weighting functions, which turn out to deviate from linearity and to be in line with the probability adjustments presented above. Thus (again utilizing the relationship between Kahneman and Tversky's prospect theory and this dissertation's framework), lending support to our theoretical argument.

CHAPTER 6

CONCLUSION

We conclude with a terse summary as well as make several closing remarks that discuss additional aspects of our work and put it into a broader perspective.

The ambition of the current dissertation is to improve our understanding of how in the stochastic, risky environment economic decisions are made by real people, whose information-processing abilities may be limited, as opposed to fictitious entities endowed with unbounded computational resources. The dissertation develops a positive (rather than normative) theoretical framework for decision-making under risk, which presumes rational optimizing behavior of the agents, builds from first principles (and yet is very tractable analytically), follows the discipline of information theory, is consistent with theoretical and empirical findings in neuroscience as well as with the results of economic laboratory experiments. In particular, neuroscientific and information-theoretic arguments allow us to structurally motivate and constructively quantify the costs of information processing. The selected application is an investment problem in the context of a general equilibrium Lucas tree model. The constructed model receives empirical support that is based on the calibrations against U.S. macro-financial data, with calibrated parameters later cross-checked and confirmed by the measurements obtained in laboratory experiments involving human subjects. This approach also produces a collection of accompanying theoretical results, which are quantitatively consistent with well-known empirical regularities and prove to be of separate interest.

One theoretical off-shoot that is worth highlighting is, effectively, a formulation of rational optimization-driven foundations behind certain paradoxical behavior observed in the experiments and the probability adjustments that are stipulated by the (cumulative) prospect theory of Kahneman and Tversky. Accordingly, these adjustments can be interpreted as a computationally cheap method of recovering the true high-entropy distribution from its approximate lower-entropy version, with the latter produced as a result of simplification that is necessitated by information-processing capacity restrictions.

In turn, introducing the distinction between true objective and approximating subjective probability distributions provides a rational unifying framework for a theoretical normative axiomatic approach to and empirical positive description of optimal choice under risk. Indeed, rational behavior of the decision-makers entails that their optimal decisions made effectively under the true distribution obviously satisfy the standard von Neumann-Morgenstern axioms, but viewed under (some implementations of) an approximating distribution these same decisions seem to deviate from rationality and exihibit phenomena such as the Allais paradox or the equity-premium puzzle.

This work develops an approach to evaluating expectations of stochastic objects that explicitly accounts for constraints imposed by the available information processing capacity. This is done without loss of generality, as traditional "rational expectations" are nested within and emerge as a special case when the information constraint is not binding. Such generalization effectively allows to explore the information processing demands of the rational expectation formation. It turns out that, from a technical standpoint, as a process of computing an integral with respect to some probability measure its requirements are not prohibitive, because various adjustments and rough heuristics can reduce the computational costs dramatically without substantial efficiency losses, thus producing the decision outcomes approaching the "rational expectations" benchmark quite closely (at least given the assumptions made here). However, from a more conceptual standpoint, as an equilibrium notion it imposes strong restrictions, and forcing the objective and subjective distributions to coincide is not innocuous and may lead to very misleading, even puzzling results. In short, "rational expectations" provide a sufficiently robust benchmark, but may be too inflexible to be applied blindly.

Within the universe of alternative structural consumption-based asset pricing models the current work belongs to the literature cluster that focuses on probability distributions involved as a source of missing explanatory power. This literature cluster hosts papers mainly dealing with consumption dynamics $(c_{t+1} - c_t)$ in the SDF, whose probability distribution

could be adjusted to account for heavy tails (say, basing on Bayesian updating of unknown parameters as in Weitzman, 2007), for jumps (rare disasters of Rietz, 1988; Veronesi, 2004; Barro, 2006; or rare booms of Tsai and Wachter, 2016), or to incorporate stochastic expected growth rate (long-run risk of Bansal and Yaron, 2004; Hansen et al., 2008).[1] But it also hosts works concerned about returns $\boldsymbol{r}_{t+1}$, whose distribution might have to be adjusted in light of capital income taxation (McGrattan and Prescott, 2003 and 2005) or upwardly biased selected sample (Dimson et al., 2003; Fama and French, 2002), and, going beyond first moment, heavier than Gaussian tails (as captured by Student's t or Lévy distribution, e.g., see classical reference Mandebrot, 1963, as well as Fama, 1963), possibly with jumps (Poisson processes). A different literature cluster is formed by papers focusing on the SDF M_{t+1} itself and on the underlying utility function, which may be amended with features such as recursive specification (Epstein and Zin, 1989 and 1991), ambiguity aversion (Hansen and Sargent, 2007b; Hansen, 2007), habit formation (Campbell and Cochrane, 1999; Menzly et al., 2004), or an extended consumption horizon (ultimate consumption risk of Parker and Julliard, 2005). Arguably, two clusters partitioning the literature above are fundamentally just two sides of the same coin—which can be rigorously shown by bringing into play a change of the probability measure, a well-known technique in financial economics and in stochastic calculus more broadly—thus constituting an encompassing non-contradictory perspective on alternative consumption-based asset-pricing models. Further elaboration on this topic deserves a separate paper, however.

Lastly, we should emphasize that our treatment is more general than it may seem at first: it also accounts for information processing performed with the aid of machines, which is relevant for any empirical parallels going beyond toy examples. Appendix §M fleshes out this point in more detail.

1. Zooming slightly deeper, the dichotomy of the two treatments discussed earlier in §5.2.2, classical adjustment-motivated versus Bayesian-like beliefs-based one, is in some sense akin to the difference between Bansal and Yaron (2004) or Tsai and Wachter (2016) as opposed to Weitzman (2007), respectively.

APPENDIX A

EXTENSION TO INFINITE HORIZON (SEQUENCE PROBLEMS)

A.1 Investment portfolio choice problem

The consumer-investor is interested in solving the following consumption and portfolio choice problem, $\mathcal{P}_{\mathcal{Q}}$:

$$\max_{\{C_s,\{q_{0,s},\boldsymbol{q}_s\}\}_t^\infty} \mathrm{E}_t^g\left[\sum_{s=t}^{\infty}\beta^{s-t}u(C_s)\right] = \int_{\mathbb{R}_+^K}\sum_{s=t}^{\infty}\beta^{s-t}u(C_s)\,g_D(\boldsymbol{D}_s|\boldsymbol{D}_{s-1})\,d\boldsymbol{D}_s \qquad \{\mathcal{P}_{\mathcal{Q}}\}$$

subject to a sequence of budget constraints

$$C_s + P_{0,s}q_{0,s} + \boldsymbol{P}_s^\intercal\boldsymbol{q}_s = q_{0,s-1} + (\boldsymbol{P}_s+\boldsymbol{D}_s)^\intercal\boldsymbol{q}_{s-1}, \qquad \forall s\geq t,$$

control variables' domain restrictions $C_s, \{q_{0,s},\boldsymbol{q}_s\}\in\mathbb{R}_+\times\mathbb{R}^{K+1}, \forall s\geq t$, as well as the no-Ponzi-schemes constraint, also listing here the usual transversality condition for optimality,

$$\lim_{T\to\infty}\left\{\left(\prod_{s=t}^{T}P_{0,s}\right)q_{0,T-1} + \mathbf{1}^\intercal\left(\prod_{s=t}^{T}\mathrm{diag}(\boldsymbol{P}_s+\boldsymbol{D}_s)^{-1}\mathrm{diag}(\boldsymbol{P}_s)\right)\boldsymbol{q}_{T-1}\right\}\geq 0 \quad \text{a.s. (under } g_D\text{)},$$
$$\lim_{s\to\infty}\mathrm{E}_t^g[\beta^{s-t}u'(C_s)(P_{0,s}q_{0,s}+\boldsymbol{P}_s^\intercal\boldsymbol{q}_s)] = 0;$$

with $u(C_s) = C_s^{1-\gamma}/(1-\gamma)$, and where

$$g_D(\boldsymbol{D}_{s+1}|\boldsymbol{D}_s) \text{ is given}, \qquad \forall s\geq t.$$

(Alternatively, the no-Ponzi-schemes constraint and the transversality condition can be replaced with a compact domain for admissible control variables that covers the borrowing/short-selling and the asset supply limits.)

In words, the representative agent would like to choose stochastic consumption and investment plans that maximize an expected discounted sum of per-period utilities and at the same time satisfy the sequence of budget constraints (as well as the technical conditions ruling out pathological and ensuring valid solutions). The expectation is taken with respect to a given objective probability density function that defines the distribution of the stochastic fruit-dividends.

This is a standard dynamic programming problem. The state variables are $\{q_{0,t-1}, \boldsymbol{q}_{t-1}\}$ and $\boldsymbol{D}_t$. Denote the maximum value function as $v^\sharp(\{q_{0,t-1}, \boldsymbol{q}_{t-1}\}, \boldsymbol{D}_t)$. The corresponding Bellman equation is then:

$$v^\sharp(\{q_{0,t-1}, \boldsymbol{q}_{t-1}\}, \boldsymbol{D}_t) = \max_{C_t, \{q_{0,t}, \boldsymbol{q}_t\}} \left\{ u(C_t) + \beta \mathrm{E}_t^g \left[v^\sharp(\{q_{0,t}, \boldsymbol{q}_t\}, \boldsymbol{D}_{t+1}) \right] \right\}$$

subject to

$$C_t + P_{0,t} q_{0,t} + \boldsymbol{P}_t^\intercal \boldsymbol{q}_t = q_{0,t-1} + (\boldsymbol{P}_t + \boldsymbol{D}_t)^\intercal \boldsymbol{q}_{t-1},$$

domain restriction $C_t, \{q_{0,t}, \boldsymbol{q}_t\} \in \mathbb{R}_+ \times \mathbb{R}^{K+1}$, as well as the same no-Ponzi-schemes condition; also with the same utlity function specification, and where

$$g_D(\boldsymbol{D}_{t+1} | \boldsymbol{D}_t) \text{ is given.}$$

A.2 Feasible investment portfolio choice problem

A feasible version of the consumption and portfolio choice problem, $\mathcal{P}_{\mathcal{QI}}$, is formulated as follows:

$$\max_{\{C_s, \{q_{0,s}, \boldsymbol{q}_s\}\}_t^\infty} \mathrm{E}_t^h \left[\sum_{s=t}^{\infty} \beta^{s-t} u(C_s) \right] = \int_{\mathbb{R}_+^K} \sum_{s=t}^{\infty} \beta^{s-t} u(C_s)\, h_D(\hat{\boldsymbol{D}}_s | \hat{\boldsymbol{D}}_{s-1})\, d\hat{\boldsymbol{D}}_s \qquad \{\mathcal{P}_{\mathcal{QI}}\}$$

subject to a sequence of budget constraints

$$C_s + P_{0,s}q_{0,s} + \boldsymbol{P}_s^\intercal \boldsymbol{q}_s = q_{0,s-1} + (\boldsymbol{P}_s + \hat{\boldsymbol{D}}_s)^\intercal \boldsymbol{q}_{s-1}, \qquad \forall s \geq t,$$

control variables' domain restrictions $C_s, \{q_{0,s}, \boldsymbol{q}_s\} \in \mathbb{R}_+ \times \mathbb{R}^{K+1}, \forall s \geq t$, as well as the no-Ponzi-schemes constraint, also listing here the usual transversality condition for optimality,

$$\lim_{T\to\infty} \left\{ \left(\prod_{s=t}^{T} P_{0,s} \right) q_{0,T-1} + \mathbf{1}^\intercal \left(\prod_{s=t}^{T} \operatorname{diag}(\boldsymbol{P}_s + \hat{\boldsymbol{D}}_s)^{-1} \operatorname{diag}(\boldsymbol{P}_s) \right) \boldsymbol{q}_{T-1} \right\} \geq 0 \quad \text{a.s. (under } h_D\text{)},$$

$$\lim_{s\to\infty} \mathrm{E}_t^h[\beta^{s-t} u'(C_s)(P_{0,s}q_{0,s} + \boldsymbol{P}_s^\intercal \boldsymbol{q}_s)] = 0;$$

with $u(C_s) = C_s^{1-\gamma}/(1-\gamma)$, and where

$$h_D(\hat{\boldsymbol{D}}_{s+1}|\hat{\boldsymbol{D}}_s) \text{ solves } \mathcal{P}_{\mathcal{I}} \text{ given } d(\boldsymbol{D}_s, \hat{\boldsymbol{D}}_s) \text{ and } \kappa, \qquad \forall s \geq t,$$

$$g_D(\boldsymbol{D}_{s+1}|\boldsymbol{D}_s) \text{ is given}, \qquad \forall s \geq t.$$

The crucial difference from before is that in the feasible formulation of the consumption and portfolio choice problem the expectation is now taken with respect to the endogenous subjective probability density function for stochastic fruit-dividends, which itself has to be obtained as an optimal solution to the auxiliary informational problem.

The corresponding Bellman equation becomes:

$$v(\{q_{0,t-1}, \boldsymbol{q}_{t-1}\}, \hat{\boldsymbol{D}}_t) = \max_{C_t, \{q_{0,t}, \boldsymbol{q}_t\}} \left\{ u(C_t) + \beta \mathrm{E}_t^h \left[v(\{q_{0,t}, \boldsymbol{q}_t\}, \hat{\boldsymbol{D}}_{t+1}) \right] \right\}$$

subject to

$$C_t + P_{0,t}q_{0,t} + \boldsymbol{P}_t^\intercal \boldsymbol{q}_t = q_{0,t-1} + (\boldsymbol{P}_t + \hat{\boldsymbol{D}}_t)^\intercal \boldsymbol{q}_{t-1},$$

domain restriction $C_t, \{q_{0,t}, \boldsymbol{q}_t\} \in \mathbb{R}_+ \times \mathbb{R}^{K+1}$, as well as the same no-Ponzi-schemes condi-

tion; also with the same utility function specification, and where

$$h_D(\hat{\boldsymbol{D}}_{t+1}|\hat{\boldsymbol{D}}_t) := \int_{\text{supp}(g_D)} f(\boldsymbol{D}_{t+1}, \hat{\boldsymbol{D}}_{t+1}|\boldsymbol{D}_t, \hat{\boldsymbol{D}}_t)\, d\boldsymbol{D}_{t+1},$$

$$f_D(\boldsymbol{D}_{t+1}, \hat{\boldsymbol{D}}_{t+1}|\boldsymbol{D}_t, \hat{\boldsymbol{D}}_t) := \arg\left\{ \min_{f(\cdot,\cdot)} \mathrm{E}^f \left[d(v^\sharp(\{q_{0,t}, \boldsymbol{q}_t\}, \boldsymbol{D}_{t+1}), v(\{q_{0,t}, \boldsymbol{q}_t\}, \hat{\boldsymbol{D}}_{t+1})) \right] \right.$$
$$\left. \text{s.t. } \mathcal{I}(g_D(\boldsymbol{D}_{t+1}|\boldsymbol{D}_t); h_D(\hat{\boldsymbol{D}}_{t+1}|\hat{\boldsymbol{D}}_t)) \leq \kappa \right\},$$

$g_D(\boldsymbol{D}_{t+1}|\boldsymbol{D}_t)$ is given.

The Bellman equation's formulation is standard except that the probability density function $h_D(\cdot)$ with respect to which it is defined stems from the solution to auxiliary sub-problem $\mathcal{P}_{\mathcal{I}}$.[1]

1. Note that the no-Ponzi-schemes constraint here holds also for the original probability distribution $g_D(\cdot)$ as long as original and simplified distributions are absolutely continuous with respect to each other.

APPENDIX B

INVARIANCE TO DECORRELATION

Proposition B.1 (Invariance to Decorrelation)**.** *Informational problem $\mathcal{P}_{\mathcal{I}}$ with distortion function defined in Proposition 3 is unaffected by decorrelation: both $d(\cdot\,,\cdot)$ and $\mathcal{I}(\cdot\,;\cdot)$ are invariant to linear transformations.*

Proof. See Appendix §F.3.

□

APPENDIX C

SOLUTION (TECHNICAL DETAILS)

This Apendix presents the solution to feasible consumption and portfolio choice problem $\mathcal{P}_{\mathcal{QI}}$. We start with the consumption and investment segment of the larger problem, dealing with the informational sub-problem afterwards. Clearly segregated formulations of these two sub-problems allow to formally solve each of them pretty much independently.

C.1 Solution to consumption and investment sub-problem

Here we solve problem $\mathcal{P}_{\mathcal{QI}}$ taking $h_D(\hat{\boldsymbol{D}}_{t+1})$ as given, i.e. focusing only on expressions (PQI-1)–(PQI-2) while respecting the domain, no-Ponzi-schemes and utility function restrictions. Essentially, this is a portfolio choice problem of Samuelson (1969), as well as Merton (1969), with price behavior related to underlying dividend dynamics as in Lucas (1978), and Breeden (1979).

First-order necessary conditions for the interior optimum is of the usual form:

$$P_{0,t} = \mathrm{E}_t^h \left[\beta \frac{u'(C_{t+1})}{u'(C_t)} \right] = \mathrm{E}_t^h \left[\beta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} \right], \tag{C.1}$$

$$\boldsymbol{P}_t = \mathrm{E}_t^h \left[\beta \frac{u'(C_{t+1})}{u'(C_t)} \left(\boldsymbol{P}_{t+1} + \hat{\boldsymbol{D}}_{t+1} \right) \right] = \mathrm{E}_t^h \left[\beta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} \left(\boldsymbol{P}_{t+1} + \hat{\boldsymbol{D}}_{t+1} \right) \right]. \tag{C.2}$$

We do not provide the full argument, and only mention the importance of realizing that due to i.i.d.-assumption, $\mathrm{E}_t^h \left[\beta \left(\widehat{\boldsymbol{q}}^{\intercal} \hat{\boldsymbol{D}}_{t+1} \right)^{-\gamma} \left(\boldsymbol{P}(\hat{\boldsymbol{D}}_{t+1}) + \hat{\boldsymbol{D}}_{t+1} \right) \right]$ ends up being just a vector of constants. Leaving verification to the reader, we simply state that the optimal solution to consumption and investment sub-parts of the full problem is characterized by the expressions below (unfortunately, completely closed-form analytical solutions are not available in general

even for the unconstrained problem $\mathcal{P}_{\mathcal{Q}}$):

$$C_t^* = (1-\beta)W_t = (\boldsymbol{q}_{t-1})^\intercal \boldsymbol{D}_t, \tag{C.3}$$

$$P_{0,t}^* q_{0,t}^* + (\boldsymbol{P}_t^*)^\intercal \boldsymbol{q}_t^* = \beta W_t, \tag{C.4}$$

$$\{q_{0,t}^*, \boldsymbol{q}_t^*\} = \{0, \widehat{\boldsymbol{q}}\}, \tag{C.5}$$

$$P_{0,t}^* = P_0(\boldsymbol{D}_t) = \beta(\widehat{\boldsymbol{q}}^\intercal \boldsymbol{D}_t)^\gamma \, \mathrm{E}_t^h \left[\frac{1}{(\widehat{\boldsymbol{q}}^\intercal \hat{\boldsymbol{D}}_{t+1})^\gamma} \right], \tag{C.6}$$

$$\boldsymbol{P}_t^* = \boldsymbol{P}(\boldsymbol{D}_t) = \frac{\beta}{1-\beta}(\widehat{\boldsymbol{q}}^\intercal \boldsymbol{D}_t)^\gamma \, \mathrm{E}_t^h \left[\frac{1}{(\widehat{\boldsymbol{q}}^\intercal \hat{\boldsymbol{D}}_{t+1})^\gamma} \hat{\boldsymbol{D}}_{t+1} \right], \tag{C.7}$$

$$v_t^* = v(\{q_{0,t-1}, \boldsymbol{q}_{t-1}\}, \boldsymbol{D}_t) = A W_t^{1-\gamma}, \tag{C.8}$$

where, according to the definition in (3.1),

$$W_t = q_{0,t-1} + (\boldsymbol{P}_t^* + \boldsymbol{D}_t)^\intercal \boldsymbol{q}_{t-1},$$

and, in line with (3.5),

$$A = \frac{(1-\beta)^{-\gamma}}{1-\gamma}.$$

In the optimum, consumption and total investments are each constant shares of current wealth; and the value function takes the same CRRA form as the utility function, only in terms of wealth.

C.2 Solution to informational sub-problem

Now we turn to the informational part of the larger problem $\mathcal{P}_{\mathcal{QI}}$. It is basically solved in the main text and crucial details of the solution are presented in §3.3 (with §C.3 furnishing some auxiliary results), the rest is available in §C.4.

C.3 Updating of the mean

Notice that

$$\begin{aligned}\hat{\boldsymbol{\mu}}_r(\widehat{\omega}_t) = \boldsymbol{\mu}_r + \check{\boldsymbol{\mu}}_r(\widehat{\omega}_t) = \boldsymbol{\mu}_r + \frac{1}{2}\mathrm{diag}^{-1}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r)\mathbf{1}(1-\widehat{\omega}_t) \neq \\ \neq \boldsymbol{\mu}_r + \frac{1}{2}\mathrm{diag}^{-1}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r) - \frac{1}{2}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r)\boldsymbol{\omega}_t = \boldsymbol{\mu}_r + \check{\boldsymbol{\mu}}_r = \\ = \hat{\boldsymbol{\mu}}_r\end{aligned}$$

in general. (For $\boldsymbol{\omega}_t$ with positive elements, in the interior solution case it is easy to see that $\hat{\boldsymbol{\mu}}_r(\widehat{\omega}_t) < \hat{\boldsymbol{\mu}}_r$; but this does not always hold in the boundary solution case, as can be shown by a simple counterexample.) However, Proposition 3.1 states that optimal accounting for the discrepancy between original, $\boldsymbol{\Sigma}_r$, and simplified, $\hat{\boldsymbol{\Sigma}}_r$, variance-covariance matrices requires using the latter value for the mean, $\hat{\boldsymbol{\mu}}_r$.

Less formally, we may posit that the mean $\hat{\boldsymbol{\mu}}_r(\widehat{\omega}_t)$ is trained over time off achieved decision outcomes, thus approaching $\hat{\boldsymbol{\mu}}_r$ in the course of "supervised learning".

Alternatively and more formally, we may postulate the following procedure for iterative updating of the mean. In each iteration ι of the optimization process, proposed choice of parameter value $\boldsymbol{\theta}_\iota := \{q_{0,t,\iota}, \boldsymbol{q}_{t,\iota}\}$ that has been accepted is immediately reflected in the corresponding value of $\boldsymbol{\omega}_{t,\iota}$ (which is possible since the latter is then just a function of known values of $\{P_{0,t}, \boldsymbol{P}_t\}$, W_t as well as $\{q_{0,t,\iota}, \boldsymbol{q}_{t,\iota}\}$; and with such auxiliary routine embedded into function $\varphi(\boldsymbol{x}|\boldsymbol{\theta}_\iota)$). In turn, this update allows to compute the values of $\check{\boldsymbol{\mu}}_{r,\iota}$ and $\hat{\boldsymbol{\mu}}_{r,\iota}$. Remember that conditional on the value of $\check{\boldsymbol{\mu}}_{r,\iota}$, distortion function from Proposition 3.3 is otherwise invariant, hence the rest of the solution to informational problem is unaffected, and results of Propositions 4–5 still hold except for updated values of $\check{\boldsymbol{\mu}}_{r,\iota}$ and $\hat{\boldsymbol{\mu}}_{r,\iota}$. Since our environment is sufficiently "well-behaved", both $\boldsymbol{\theta}_\iota$ and $\hat{\boldsymbol{\mu}}_{r,\iota}$ will converge to their optimal values $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\mu}}_r$ simultaneously. As a result, we have the following Proposition.

Proposition C.1 (Specific Solution to Informational Problem: Representation in Economic

Terms with Updating of the Mean)**.** *Assume the procedure for iterative updating of the mean described in the text. Then statement of Proposition 5 holds for* $\check{\boldsymbol{\mu}}_r(\widehat{\omega}_t)$ *and* $\hat{\boldsymbol{\mu}}_r(\widehat{\omega}_t)$ *replaced with, respectively,* $\check{\boldsymbol{\mu}}_r$ *and* $\hat{\boldsymbol{\mu}}_r$ *throughout.*

Proof. See Appendix §F.6.

□

Note that the result of Proposition C.1 is achieved for any admissible starting value of $\widehat{\omega}_t \in \mathbb{R}_+$. It is also noteworthy that postulated iterative updating procedure is akin to the (iterative or continuous) updating requirement discussed in Appendix §K, which we have ruled out appealing to robustness in Proposition 3.2 instead; but at this point the problem is much simpler and requirements needed for implementing the procedure seem more realistic.

C.4 Solution to informational sub-problem (continued)

Lastly, touching upon the informational coherence, optimal solution to the informational sub-problem amounts to the following joint probability density:

$$
\begin{aligned}
f(\boldsymbol{x}, \hat{\boldsymbol{x}}) &= f(\boldsymbol{x}|\hat{\boldsymbol{x}}) h(\hat{\boldsymbol{x}}) = \\
&= (2\pi)^{-\frac{K}{2}} \left|\boldsymbol{\Psi}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{x}}+\check{\boldsymbol{\mu}})^{\intercal}\boldsymbol{\Psi}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{x}}+\check{\boldsymbol{\mu}})} \times (2\pi)^{-\frac{K}{2}} \left|\hat{\boldsymbol{\Sigma}}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\hat{\boldsymbol{x}}-\hat{\boldsymbol{\mu}})^{\intercal}\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{x}}-\hat{\boldsymbol{\mu}})} = \\
&= (2\pi)^{-\frac{2K}{2}} \left|\begin{bmatrix} \boldsymbol{\Sigma} & \hat{\boldsymbol{\Sigma}} \\ \hat{\boldsymbol{\Sigma}} & \hat{\boldsymbol{\Sigma}} \end{bmatrix}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \begin{bmatrix} \boldsymbol{x}-\boldsymbol{\mu} \\ \hat{\boldsymbol{x}}-\hat{\boldsymbol{\mu}} \end{bmatrix}^{\intercal} \begin{bmatrix} \boldsymbol{\Sigma} & \hat{\boldsymbol{\Sigma}} \\ \hat{\boldsymbol{\Sigma}} & \hat{\boldsymbol{\Sigma}} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{x}-\boldsymbol{\mu} \\ \hat{\boldsymbol{x}}-\hat{\boldsymbol{\mu}} \end{bmatrix} \right),
\end{aligned}
$$

which, after substituting $\boldsymbol{\Xi}^{\intercal}\boldsymbol{r}_{t+1}$ for $\boldsymbol{x}$, $\boldsymbol{\Xi}^{\intercal}\hat{\boldsymbol{r}}_{t+1}$ for $\hat{\boldsymbol{x}}$, $\boldsymbol{\Xi}^{\intercal}\boldsymbol{\mu}_r$ for $\boldsymbol{\mu}$, $\boldsymbol{\Xi}^{\intercal}\hat{\boldsymbol{\mu}}_r$ for $\hat{\boldsymbol{\mu}}$, $\boldsymbol{\Xi}^{\intercal}\boldsymbol{\Sigma}_r\boldsymbol{\Xi}$ for $\boldsymbol{\Sigma}$, and $\boldsymbol{\Xi}^{\intercal}\hat{\boldsymbol{\Sigma}}_r\boldsymbol{\Xi}$ for $\hat{\boldsymbol{\Sigma}}$, produces $f_r(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1})$:

$$
\begin{aligned}
f(\boldsymbol{x}, \hat{\boldsymbol{x}}) &= (2\pi)^{-\frac{2K}{2}} \left|\begin{bmatrix} \boldsymbol{\Sigma}_r & \hat{\boldsymbol{\Sigma}}_r \\ \hat{\boldsymbol{\Sigma}}_r & \hat{\boldsymbol{\Sigma}}_r \end{bmatrix}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \begin{bmatrix} \boldsymbol{r}_{t+1}-\boldsymbol{\mu}_r \\ \hat{\boldsymbol{r}}_{t+1}-\hat{\boldsymbol{\mu}}_r \end{bmatrix}^{\intercal} \begin{bmatrix} \boldsymbol{\Sigma}_r & \hat{\boldsymbol{\Sigma}}_r \\ \hat{\boldsymbol{\Sigma}}_r & \hat{\boldsymbol{\Sigma}}_r \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{r}_{t+1}-\boldsymbol{\mu}_r \\ \hat{\boldsymbol{r}}_{t+1}-\hat{\boldsymbol{\mu}}_r \end{bmatrix} \right) = \\
&=: f_r(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}).
\end{aligned}
$$

That is, $f(\cdot,\cdot)$ and $f_r(\cdot,\cdot)$ have the same multivariate-$\mathcal{N}$ form, i.e. for parameters consisting of mean vector $\boldsymbol{\Theta}_1$ and variance-covariance matrix $\boldsymbol{\Theta}_2$,

$$f(\boldsymbol{\chi},\hat{\boldsymbol{\chi}}|\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2) = f_r(\boldsymbol{\chi},\hat{\boldsymbol{\chi}}|\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2) \text{ is } \mathcal{N}(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2) \quad \forall \boldsymbol{\chi},\hat{\boldsymbol{\chi}} \in \mathbb{R}^K.$$

Obviously, analogous relationship holds for $g(\cdot)$ and $g_r(\cdot)$, as well as $h(\cdot)$ and $h_r(\cdot)$.

C.5 Merging two sub-problems' solutions

Finally, we use the results from §C.1 and §C.2 to tie up the loose ends concerning the probability distributions of dividends.

Having got the solution for $\boldsymbol{P}_t^*$ in (C.7), we can combine definition (3.2) with the result of Proposition 5 to "reverse-engineer" approximating probability density function for dividends so that by construction it would be coherent with approximating density of returns deduced in the Proposition:

$$h_r(\hat{\boldsymbol{r}}_{t+1}) = (2\pi)^{-\frac{K}{2}}|\hat{\boldsymbol{\Sigma}}_r|^{-\frac{1}{2}}e^{-\frac{1}{2}(\hat{\boldsymbol{r}}(\hat{\boldsymbol{D}}_{t+1}|\hat{\boldsymbol{D}}_t)-\hat{\boldsymbol{\mu}}_r)^\intercal\hat{\boldsymbol{\Sigma}}_r^{-1}(\hat{\boldsymbol{r}}(\hat{\boldsymbol{D}}_{t+1}|\hat{\boldsymbol{D}}_t)-\hat{\boldsymbol{\mu}}_r)} =: h_D(\hat{\boldsymbol{D}}_{t+1}|\hat{\boldsymbol{D}}_t),$$

where

$$\begin{aligned}\hat{\boldsymbol{r}}(\hat{\boldsymbol{D}}_{t+1}|\hat{\boldsymbol{D}}_t) = \ln \hat{\boldsymbol{R}}(\hat{\boldsymbol{D}}_{t+1}|\hat{\boldsymbol{D}}_t) = \ln\left(\operatorname{diag}\left(\boldsymbol{P}(\boldsymbol{D}_t)\right)^{-1}\left(\boldsymbol{P}(\hat{\boldsymbol{D}}_{t+1})+\hat{\boldsymbol{D}}_{t+1}\right)\right) :=\\ := \ln\left(\operatorname{diag}\left(\frac{\beta}{1-\beta}(\widehat{\boldsymbol{q}}^\intercal\boldsymbol{D}_t)^\gamma\,\boldsymbol{E}\right)^{-1}\left(\frac{\beta}{1-\beta}(\widehat{\boldsymbol{q}}^\intercal\hat{\boldsymbol{D}}_{t+1})^\gamma\,\boldsymbol{E}+\hat{\boldsymbol{D}}_{t+1}\right)\right).\end{aligned}$$

Here, we have introduced for the following constant vector a shortcut notation

$$\boldsymbol{E} := \mathrm{E}_t^h\left[\frac{1}{(\widehat{\boldsymbol{q}}^\intercal\hat{\boldsymbol{D}}_{t+1})^\gamma}\hat{\boldsymbol{D}}_{t+1}\right].$$

Since in equilibirum prices are dictated by the solution to consumption and investment subproblem from §C.1, we use the same optimal price function (C.7), together with definition

(3.2) and assumption (3.4), also to back out the true density for dividends:

$$g_r(\boldsymbol{r}_{t+1}) = (2\pi)^{-\frac{K}{2}} |\boldsymbol{\Sigma}_r|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{r}(\boldsymbol{D}_{t+1}|\boldsymbol{D}_t)-\boldsymbol{\mu}_r)^\intercal \boldsymbol{\Sigma}_r^{-1}(\boldsymbol{r}(\boldsymbol{D}_{t+1}|\boldsymbol{D}_t)-\boldsymbol{\mu}_r)} =: g_D(\boldsymbol{D}_{t+1}|\boldsymbol{D}_t),$$

where

$$\begin{aligned}\boldsymbol{r}(\boldsymbol{D}_{t+1}|\boldsymbol{D}_t) = \ln \boldsymbol{R}(\boldsymbol{D}_{t+1}|\boldsymbol{D}_t) = \ln\left(\operatorname{diag}\left(\boldsymbol{P}(\boldsymbol{D}_t)\right)^{-1}\left(\boldsymbol{P}(\boldsymbol{D}_{t+1}) + \boldsymbol{D}_{t+1}\right)\right) := \\ := \ln\left(\operatorname{diag}\left(\frac{\beta}{1-\beta}(\widehat{\boldsymbol{q}}^\intercal \boldsymbol{D}_t)^\gamma \boldsymbol{E}\right)^{-1}\left(\frac{\beta}{1-\beta}(\widehat{\boldsymbol{q}}^\intercal \boldsymbol{D}_{t+1})^\gamma \boldsymbol{E} + \boldsymbol{D}_{t+1}\right)\right).\end{aligned}$$

Note that constant vector $\boldsymbol{E}$ used here in the determination of true $\boldsymbol{r}(\boldsymbol{D}_{t+1}|\boldsymbol{D}_t)$ is the same as in the case of approximating $\hat{\boldsymbol{r}}(\hat{\boldsymbol{D}}_{t+1}|\hat{\boldsymbol{D}}_t)$, and is still defined in terms of simplified probability density according to (C.5).

The corresponding result for $f_r(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1})$ and $f_D(\boldsymbol{D}_{t+1}, \hat{\boldsymbol{D}}_{t+1}|\boldsymbol{D}_t, \hat{\boldsymbol{D}}_t)$ follows in a similar manner, thus validating informational coherence.

APPENDIX D

RELATED LITERATURE

Related themes in the literature only briefly mentioned above include:

- information-theoretic methods: classical works (Shannon, 1948; Jaynes, 2003; Cover and Thomas, 2006; MacKay, 2003; in economics Marschak, 1959, 1968, 1971), rational inattention based on mutual information (Sims, 1998, 2003, 2006, 2010; Matějka and Sims, 2010; Matějka and McKay, 2014; Matějka, 2014a, 2014b; Caplin and Dean, forthcoming, 2013; Ravid, 2016; also refer to Woodford, 2012, 2014), model diagnostics and belief measurement based on relative entropy (Stutzer, 1995, 1996; Hansen and Sargent, 2007a, 2007b; Hansen, 2007; Backus et al., 2014; Ghosh et al., 2013; Chen et al., 2015; Alvarez and Jermann, 2005; overview available in Hansen, 2014);

- statistical learning and simplification: "coarsening" (Al-Najjar and Pai, 2014), regularization (Chen and Haykin, 2002; Bickel and Li, 2006; Hastie et al., 2009; Bishop, 2006; Murphy, 2012), shrinkage (Stein, 1956; James and Stein, 1961; in finance Ledoit and Wolf, 2003, 2004a, 2004b; Jagannathan and Ma, 2003; Won et al., 2012; Jorion, 1985; 1986);

- bounded rationality focusing on decision costs and ensuing simplification: classical works (Simon 1997, 1957; Gigerenzer and Selten, 2001; Gigerenzer et al., 2000; Gilovich et al., 2002), modern works (Gabaix and Laibson, 2000, 2005; Gabaix et al., 2006; Fuster et al., 2012; Bordalo et al., 2016; Gabaix, 2014a, 2014b; Mullainathan, 2002b; Jehiel, 2005; also see Carroll, 2003), with costs due to memory limitations (Gilboa and Schmeidler, 1995; Wilson, 2014; Mullainathan, 2002a);

- bounded rationality focusing on decision criterion and ensuing belief distortions, including "optimism"/"pessimism" and "overconfidence"/"doubt": endogenous distortions (Hansen, 2007; Hansen and Sargent 2007a, 2007b; Brunnermeier and Parker,

2005; Brunnermeier et al., 2007; Brunnermeier et al., 2013); exogenous distortions (Cecchetti et al., 2000; Abel, 2002; Scheinkman and Xiong, 2003; Peng and Xiong, 2006);

- experimental evidence: entropy reduction (Gabaix and Laibson, 2000; Goldstone and Hendrickson, 2010; Fleming et al., 2013), overconfidence (Camerer, 1995), prospect theory and probability weighting (Kahneman and Tversky, 1979, 1992; Hsu et al., 2009; Bossaerts et al., 2008; Fox et al., 1996; List, 2004; List and Haigh, 2005; as well as Tversky and Kahneman, 1971, 1972, 1974; reviews given in Fox and Poldrack, 2008; Barberis, 2013b), together with its measurements (based on laboratory experiments in Abdellaui, 2000; Abdellaui et al., 2005; Bleichrodt et al., 1999; Bleichrodt and Pinto, 2000; Bruhin et al., 2010; Fehr-Duda and Epper, 2012; Gonzalez and Wu, 1999; Noussair and Vogt, 2013; Prelec, 1998; Wu and Gonzalez, 1996; with meta-analysis given in Fox and Poldrack, 2008; Stott, 2006; Camerer and Ho, 1994; based on real financial markets in Gurevich et al., 2009; Kliger and Levy, 2009; Polkovnichenko, 2005; Polkovnichenko and Zhao, 2013; Dierkes, 2013; Hens and Reichlin, 2013), probability weighting within rank-dependent subjective expected utility theories (Quigging, 1982; Yaari, 1987; Luce, 1988; Schmeidler, 1989; Gul, 1991; reviews given in Fehr-Duda and Epper, 2012; Barberis, 2013a);

- neuroscience and psychology: textbooks (Dayan and Abbott, 2001; Squire et al., 2008), thematic volumes (Baddeley et al., 2000; Doya et al., 2007; Glimcher et al., 2008), journal review issues (Schultz, 2008; Bayer, 2008; Bunge, 2005; Pammi and Srinivasan, 2013), information-processing capacity (classical papers Miller, 1956; Barlow, 1961), working memory capacity (Shiffrin and Nosofsky, 1994; Kleinberg and Kaufman, 1971; Bor et al., 2003; Owen, 2004; Bor and Owen, 2006; Migliore et al., 2008; with review given in Cowan, 2000), selected models (Eliasmith et al., 2012; Rao, 2010; as well as Padoa-Schioppa and Rustichini, 2014, 2015);

- consumption-based asset pricing and "equity-premium puzzle": reviews (Campbell, 2003; Ludvigson, 2013), classical (Mehra and Prescott, 1985;, Weil, 1989; Hansen and Singleton, 1983), mainstream explanations (Rietz, 1988; Veronesi, 2004; Barro, 2006; Tsai and Wachter, 2015, 2016; Weitzman, 2007; Bansal and Yaron, 2004; Hansen et al., 2008; McGrattan and Prescott, 2003, 2005; Dimson et al., 2003; Fama and French, 2002 (also see Mandelbrot, 1963, with Fama, 1963); Epstein and Zin, 1989, 1991; Hansen and Sargent, 2007b; Hansen, 2007; Hansen et al., 2007; Campbell and Cochrane, 1999; Menzly et al., 2004; Parker and Julliard, 2005; as well as Cecchetti et al., 2000; Abel, 2003), probability-weighting based explanations (Epstein and Zin, 1990; Barberis and Huang, 2008; Routledge and Zin, 2010; De Giorgi and Legg, 2012; Post and Levy, 2005; He and Zhou, 2011; Xia and Zhou, 2012; Dierkes, 2013);

- consumption-based asset pricing and other puzzles: "non-monotone pricing kernel puzzle" (Aït-Sahalia and Lo, 2000; Jackwerth, 2000; Rosenberg and Engle, 2002; with reviews given in Ziegler, 2007; Hens and Reichlin, 2013; and with explanations based on probability weighting offered in Kliger and Levy, 2009; Polkovnichenko and Zhao, 2013; Dierkes, 2013; Gurevich et al., 2009), "asset classes" ("asset allocation" in Brinson et al., 1986, 1991; Sharpe, 1992; Fama and French, 1993; Doeswijk et al., 2014; with "strategic" undertone and stocks versus bonds in Brennan et al., 1997; Campbell and Viceira, 2002b; Wachter, 2010; Shiller and Beltratti, 1992; Canner et al., 1997; Campbell et al., 2003; Campbell and Ammer, 1993; Fleming et al., 1998; Connolly et al., 2005; Andersson et al., 2008; Guidolin and Timmermann, 2007; Li, 2002, 2003; Yang et al., 2009; "comovement" in Shiller, 1989; Pindyck and Rotemberg, 1990; Kumar and Lee, 2006; Barber et al., 2009; Barberis et al., 2005; Veldkamp, 2006; "style investing" in Barberis and Schleifer, 2003; Peng and Xiong, 2006; Brown and Goetzmann, 1997; Froot and Teo, 2008; Kumar, 2009; Teo and Woo, 2004; Chan et al., 2000; Wahal and Yavuz, 2013; Cooper et al., 2005; Boyer, 2011; Froot and Teo, 2008; Choi and Sias, 2009; Jame and Tong, 2014), "underdiversification puzzle" (Blume and

Friend, 1975; Statman, 1987; Kelly, 1995; French and Poterba, 1991; with analysis and explanations offered in Polkovnichenko, 2005; Goetzmann and Kumar, 2008; Anderson, 2013; Goetzmann et al., 2005; Li, 2003; Van Nieuwerburgh and Veldkamp, 2009, 2010);

- "free energy" and decision-making (Ortega and Braun, 2013; Friston, 2009, 2010; Still et al., 2012).

APPENDIX E

DATA

Real per capita consumption series is constructed using quarterly data on nominal seasonally adjusted at annual rates personal consumption expenditures for nondurable goods and services, corresponding seasonally adjusted price indexes, as well as population size from U.S. Bureau for Economic Analysis's NIPA tables.

Real risk-free and market returns are constructed using nominal risk-free and market returns from Fama-French online data library, converted from monthly into quarterly frequency, as well as the personal consumption expenditures for nondurable goods and services deflator applied to nominal consumption series as described above.

Real dividend series is constructed using nominal dividends for S&P500 Composite at quarterly frequency from Robert Shiller's online data collection as well as the personal consumption expenditures for nondurable goods and services deflator applied to nominal consumption series as described above.

Our data cover 1947:Q4–2014:Q4 time period.

APPENDIX F

PROOFS OF PROPOSITIONS

F.1 Proof of Proposition 1

Proof. In information theory $\mathcal{P}_{\mathcal{I}}$ is known as the problem of finding the distortion rate function (e.g., see Cover and Thomas 2006). This is a standard calculus of variations problem that can be solved using the method of Lagrange multipliers.

Form the Lagrangian functional:

$$\begin{aligned}
\mathcal{L} := & \int_{\text{supp}(h)} \int_{\text{supp}(g)} d(\boldsymbol{x}, \hat{\boldsymbol{x}})\, f(\boldsymbol{x}, \hat{\boldsymbol{x}})\, d\boldsymbol{x}\, d\hat{\boldsymbol{x}} + \\
& + \lambda \left[\int_{\text{supp}(h)} \int_{\text{supp}(g)} f(\boldsymbol{x}, \hat{\boldsymbol{x}})\, \ln f(\boldsymbol{x}, \hat{\boldsymbol{x}})\, d\boldsymbol{x}\, d\hat{\boldsymbol{x}} - \right. \\
& \quad - \int_{\text{supp}(h)} \underbrace{\left(\int_{\text{supp}(g)} f(\boldsymbol{x}, \hat{\boldsymbol{x}})\, d\boldsymbol{x} \right)}_{h(\hat{\boldsymbol{x}})} \ln \underbrace{\left(\int_{\text{supp}(g)} f(\boldsymbol{x}, \hat{\boldsymbol{x}})\, d\boldsymbol{x} \right)}_{h(\hat{\boldsymbol{x}})} d\hat{\boldsymbol{x}} - \\
& \quad \left. - \int_{\text{supp}(g)} g(\boldsymbol{x})\, \ln g(\boldsymbol{x})\, d\boldsymbol{x} - \kappa \right] + \\
& + \mu(\boldsymbol{x}) \left[\int_{\text{supp}(h)} f(\boldsymbol{x}, \hat{\boldsymbol{x}})\, d\hat{\boldsymbol{x}} - g(\boldsymbol{x}) \right] + \\
& + \nu(\boldsymbol{x}, \hat{\boldsymbol{x}}) \left[-f(\boldsymbol{x}, \hat{\boldsymbol{x}}) \right].
\end{aligned}$$

The integrability condition ensures the optimum exists. Absolute continuity rules out boundary solutions. The objective is convex in $f(\hat{\boldsymbol{x}}|\boldsymbol{x})$ for fixed $g(\boldsymbol{x})$, so the first order condition with respect to $f(\boldsymbol{x}, \hat{\boldsymbol{x}})$ is sufficient for interior minimum. Equalize the corresponding functional derivative to 0 to obtain

$$\begin{aligned}
0 =: \frac{\delta \mathcal{L}}{\delta f(\boldsymbol{x}, \hat{\boldsymbol{x}})} = d(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \lambda \left\{ \ln f(\boldsymbol{x}, \hat{\boldsymbol{x}}) + 1 - \ln \left(\int_{\text{supp}(g)} f(\boldsymbol{x}, \hat{\boldsymbol{x}})\, d\boldsymbol{x} \right) - 1 \right\} + \\
+ \mu(\boldsymbol{x}) - \nu(\boldsymbol{x}, \hat{\boldsymbol{x}}).
\end{aligned}$$

Rearranging yields the following solution to minimization problem:

$$f(\boldsymbol{x}|\hat{\boldsymbol{x}}) = e^{\frac{1}{\lambda}\nu(\boldsymbol{x},\hat{\boldsymbol{x}})-\frac{1}{\lambda}\mu(\boldsymbol{x})-\frac{1}{\lambda}d(\boldsymbol{x},\hat{\boldsymbol{x}})}.$$

□

F.2 Proof of Proposition 3, with additional comments

The statement and its proof is split into three auxiliary propositions.

Proposition 3.1 (Approximate Distortion Function)**.** *Let distortion function*

$$d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}) = (\ln W_{t+1} - \ln \hat{W}_{t+1})^2$$

be a random scalar defined in the context of problem $\mathcal{P}_{\mathcal{QI}}$.

Then, under distributional assumptions given, it approximately equals

$$d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}) \approx \left(\boldsymbol{\omega}_t^{\intercal}(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r)\right)^2,$$

where

$$\boldsymbol{\omega}_t := \mathrm{diag}(\boldsymbol{P}_t)\boldsymbol{q}_t \,/\, W_t$$

is a K-vector of the shares of wealth invested in risky assets; $\hat{\boldsymbol{\Sigma}}_r$ is a variance-covariance matrix for $\hat{\boldsymbol{r}}_{t+1}$; while the mean of simplified random variable $\hat{\boldsymbol{\mu}}_r$ equals

$$\hat{\boldsymbol{\mu}}_r := \boldsymbol{\mu}_r + \check{\boldsymbol{\mu}}_r,$$

with bias $\check{\boldsymbol{\mu}}_r$ defined as

$$\check{\boldsymbol{\mu}}_r := \frac{1}{2}\mathrm{diag}^{-1}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r) - \frac{1}{2}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r)\boldsymbol{\omega}_t.$$

As interval between time periods t and $(t+1)$ shrinks, in the limit the approximation turns into exact expression.

Proof. See Appendix §F.2.1.

□

The approximation is based on continuous-time representation of the stochastic processes involved ($\ln W_{t+1}$ and $\boldsymbol{\omega}_t^\intercal \boldsymbol{r}_{t+1}$), it is relatively innocuous. The general approach is related to that of Campbell and Viceira (2002a). On the other hand, similar approximation result can also be obtained directly using second-order Taylor expansion.

Approximate distortion function from Proposition 3.1 is purpose-specific: it adjusts to and depends on $\boldsymbol{\omega}_t$, the shares of wealth considered for being invested in risky trees (note that riskless tree's share does not enter the expression, thus even the scales of three summands are not pinned down here). This introduces a sort of endogeneity, circularity to the choice of what shares to use: a contemplated portfolio allocation generates respective shares, thus inducing the distortion function and the corresponding aproximating distribution, which in turn will lead to another candidate portfolio choice and updated set of shares that differ from those used for the latest iteration of an approximating distribution.

To resolve this dilemma, we motivate the choice of wealth shares to use by the minimization of maximum loss, robustness to "worst-case" scenarios (cf. Hansen and Sargent, 2007). This approach will allow us to get rid of the connection to actual portfolio allocations or net supplies altogether. Results of its application are condensed in the following Proposition. (Appendix §K discusses alternative resolutions of this circularity dilemma.)

Proposition 3.2 (Robust Strategy)**.** *Let player Agent solve problem $\mathcal{P}_{\mathcal{I}}$ with disortion function defined as in the statement of Proposition 3.1:*

$$\min_{f_r(\boldsymbol{r},\hat{\boldsymbol{r}})} \mathrm{E}^f[d(\boldsymbol{r}_{t+1},\hat{\boldsymbol{r}}_{t+1})] \iff \min_{f_r(\boldsymbol{r},\hat{\boldsymbol{r}})} \mathrm{E}^f\left[\left(\boldsymbol{\omega}_t^\intercal(\boldsymbol{r}_{t+1}-\hat{\boldsymbol{r}}_{t+1}+\check{\boldsymbol{\mu}}_r)\right)^2\right]$$

subject to standard information constraint and technical restrictions (i.e., λ-, $\mu(\boldsymbol{x})$- and

$\nu(\boldsymbol{x}, \hat{\boldsymbol{x}})$-constraints). Let player Nature maximize the same objective function with respect to $\boldsymbol{\omega}_{1K,t}$, that is solve the following problem:

$$\max_{\boldsymbol{\omega}_t} \mathrm{E}^f[d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1})] \iff \max_{\boldsymbol{\omega}_t} \mathrm{E}^f\left[\left(\boldsymbol{\omega}_t^\intercal(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r)\right)^2\right]$$

subject to

$$\begin{aligned} \boldsymbol{\omega}_t^\intercal \mathbf{1} &\leq \widehat{\omega}_t^N & \forall \widehat{\omega}_t^N &\in \mathbb{R}_+, \\ \omega_{k,t} &\geq 0 & \forall k &\in \{1, \ldots, K\}. \end{aligned}$$

Then a simultaneous-move game posesses a Nash-equilibrium characterized by the following strategies ($\forall \widehat{\omega}_t^A, \widehat{\omega}_t^N \in \mathbb{R}_+$): Agent plays

$$f_r^\star(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}) := \arg\min_{f_r(\boldsymbol{r}, \hat{\boldsymbol{r}})} \mathrm{E}^f\left[\sum_k \frac{1}{K}\left(\widehat{\omega}_t^A\right)^2\left(r_{k,t+1} - \hat{r}_{k,t+1} + \check{\mu}_{r,k}(\widehat{\omega}_t^A)\right)^2\right],$$

where

$$\check{\mu}_{r,k}(\widehat{\omega}_t^A) := \frac{1}{2}(\sigma_{r,k}^2 - \hat{\sigma}_{r,k}^2)(1 - \widehat{\omega}_t^A) \qquad \forall k \in \{1, \ldots, K\};$$

Nature plays a mixed strategy

$$\begin{aligned} \omega_{k,t} &:= \widehat{\omega}_t^N, & k &= k^\star, \\ \omega_{k,t} &:= 0, & k &\neq k^\star, \end{aligned}$$

where

$$k^\star \sim \mathcal{U}(\{1, \ldots, K\}).$$

A game where player Nature is a second-mover posesses a Nash-equilibrium characterized by the same strategies.

Proof. See Appendix §F.2.2.

□

Fixed scalars $\widehat{\omega}_t^A$ and $\widehat{\omega}_t^N$ in the Proposition above serve the role of unknown constant parameters denoting some chosen value of total share of wealth invested in risky assets in Agent's and Nature's problems, respectively. For instance, $\widehat{\omega}_t^A$ may be thought as implicit borrowing/collateral constraint.

Also note that constraint $\omega_{k,t} \geq 0 \ \forall k \in \{1, \ldots, K\}$ in Nature's problem ensures consistency with general equilibrium.

Proposition 3.2 is of interest to us not in itself, but as a means of further refining of the distoriton function to be used. Such a distortion function is formulated in Proposition 3.3.

Proposition 3.3 (Robust Approximate Distortion Function)**.** *Robust strategy given by the statement of Proposition 3.2, and assuming total share of wealth invested in risky assets agrees with the definition in Proposition 3.1, i.e.*

$$\widehat{\omega}_t := \mathbf{1}^\intercal \boldsymbol{\omega}_t = \mathbf{1}^\intercal \mathrm{diag}(\boldsymbol{P}_t)\boldsymbol{q}_t \,/\, W_t,$$

induces the following distortion function:

$$d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}) := (\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r(\widehat{\omega}_t))^\intercal (\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r(\widehat{\omega}_t)),$$

where

$$\begin{aligned}\check{\boldsymbol{\mu}}_r(\widehat{\omega}_t) &:= \frac{1}{2}\mathrm{diag}^{-1}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r)(1 - \widehat{\omega}_t) = \\ &= \frac{1}{2}\mathrm{diag}(\sigma_{r,1}^2 - \hat{\sigma}_{r,1}^2, \cdots, \sigma_{r,K}^2 - \hat{\sigma}_{r,K}^2)\mathbf{1}(1 - \widehat{\omega}_t).\end{aligned}$$

Proof. See Appendix §F.2.3.

□

F.2.1 Proof of Proposition 3.1

Proof. Consider continuous-time dynamics of dividend and price processes from problem $\mathcal{P}_{\mathcal{Q}}$ (their counterparts from problem $\mathcal{P}_{\mathcal{QI}}$ are analogous to ones below):

$$\begin{aligned}
d\boldsymbol{D}_t &:= \operatorname{diag}(\boldsymbol{D}_t)\boldsymbol{\mu}_D dt + \operatorname{diag}(\boldsymbol{D}_t)\boldsymbol{\sigma}_D d\boldsymbol{B}_t, \\
dP_{0,t} &:= r_{0,t} P_{0,t} dt, \\
d\boldsymbol{P}_t &:= \operatorname{diag}(\boldsymbol{P}_t)(\boldsymbol{\mu}_P + \frac{1}{2}\operatorname{diag}^{-1}(\boldsymbol{\sigma}_P \boldsymbol{\sigma}_P^\intercal))dt + \operatorname{diag}(\boldsymbol{P}_t)\boldsymbol{\sigma}_P d\boldsymbol{B}_t,
\end{aligned}$$

where $\boldsymbol{D}_t$ is the dividend process, $\boldsymbol{\mu}_D$ is a K-sized constant vector, $\boldsymbol{\sigma}_D$ is a $K \times K$ constant matrix, $\boldsymbol{B}_t$ is a standard K-dimensional Brownian motion, $P_{0,t}$ is thought as the money account process with stochastic instantaneous interest rate $r_{0,t}$ (which is equivalent to rolling over just maturing zero coupon bonds), and $\boldsymbol{P}_t$ is the price process corresponding to assets with the dividend process given above.

By Itō's lemma, for $\boldsymbol{P}_t := \boldsymbol{P}(\boldsymbol{D}_t)$ we have:

$$\begin{aligned}
\boldsymbol{\mu}_P &= \operatorname{diag}(\boldsymbol{P}_t)^{-1}\frac{\partial \boldsymbol{P}}{\partial \boldsymbol{D}^\intercal}\operatorname{diag}(\boldsymbol{D}_t)\boldsymbol{\mu}_D + \frac{1}{2}\operatorname{diag}(\boldsymbol{P}_t)^{-1}\boldsymbol{\sigma}_D^\intercal \operatorname{diag}(\boldsymbol{D}_t)\frac{\partial^2 \boldsymbol{P}}{\partial \boldsymbol{D}\partial \boldsymbol{D}^\intercal}\operatorname{diag}(\boldsymbol{D}_t)\boldsymbol{\sigma}_D - \\
&\quad - \frac{1}{2}\operatorname{diag}^{-1}(\boldsymbol{\sigma}_P\boldsymbol{\sigma}_P^\intercal), \\
\boldsymbol{\sigma}_P &= \operatorname{diag}(\boldsymbol{P}_t)^{-1}\frac{\partial \boldsymbol{P}}{\partial \boldsymbol{D}^\intercal}\operatorname{diag}(\boldsymbol{D}_t)\boldsymbol{\sigma}_D.
\end{aligned}$$

We also have, deducing the (ex-dividend) wealth dynamics from the budget constraint:

$$\begin{aligned}
d\ln \boldsymbol{P}_t &= \boldsymbol{\mu}_P dt + \boldsymbol{\sigma}_P d\boldsymbol{B}_t, \\
dW_t &= W_t\left(\boldsymbol{\omega}_t^\intercal(\boldsymbol{\mu}_P + \frac{1}{2}\operatorname{diag}^{-1}(\boldsymbol{\sigma}_P\boldsymbol{\sigma}_P^\intercal) + \operatorname{diag}(\boldsymbol{P}_t)^{-1}\boldsymbol{D}_t - r_{0,t}\mathbf{1}) + r_{0,t} - \frac{C_t}{W_t}\right)dt + \\
&\quad + W_t\boldsymbol{\omega}_t^\intercal\boldsymbol{\sigma}_P d\boldsymbol{B}_t, \\
d\ln W_t &= \left(\boldsymbol{\omega}_t^\intercal(\boldsymbol{\mu}_P + \frac{1}{2}\operatorname{diag}^{-1}(\boldsymbol{\sigma}_P\boldsymbol{\sigma}_P^\intercal) + \operatorname{diag}(\boldsymbol{P}_t)^{-1}\boldsymbol{D}_t - r_{0,t}\mathbf{1}) + r_{0,t} - \frac{C_t}{W_t} - \right. \\
&\quad \left. - \frac{1}{2}\boldsymbol{\omega}_t^\intercal\boldsymbol{\sigma}_P\boldsymbol{\sigma}_P^\intercal\boldsymbol{\omega}_t\right)dt + \boldsymbol{\omega}_t^\intercal\boldsymbol{\sigma}_P d\boldsymbol{B}_t,
\end{aligned}$$

where

$$\begin{aligned} W_t &:= P_{0,t}q_{0,t} + \boldsymbol{P}_t^\intercal \boldsymbol{q}_t, \\ \omega_{0,t} &:= P_{0,t}q_{0,t} \,/\, W_t, \\ \boldsymbol{\omega}_t &:= \frac{1}{W_t}\mathrm{diag}(\boldsymbol{P}_t)\boldsymbol{q}_t. \end{aligned}$$

Increasing time intervals to $dt = 1$ and setting

$$\begin{aligned} \boldsymbol{\Sigma}_r &:= \boldsymbol{\sigma}_P \boldsymbol{\sigma}_P^\intercal, \\ \boldsymbol{r}_{t+1} &:= \boldsymbol{\mu}_P + \boldsymbol{\sigma}_P(\boldsymbol{B}_{t+1} - \boldsymbol{B}_t) =: \boldsymbol{\mu}_r + \mathcal{N}(0, \boldsymbol{\Sigma}_r) \end{aligned}$$

produces a continuous-time approximation to a discrete-time case:

$$\begin{aligned} \ln W_{t+1} - \ln W_t \approx \Bigg(&\boldsymbol{\omega}_t^\intercal(\mathrm{diag}(\boldsymbol{P}_t)^{-1}\boldsymbol{D}_t - r_{0,t}\mathbf{1}) + r_{0,t} - \frac{C_t}{W_t} + \\ &+\frac{1}{2}\boldsymbol{\omega}_t^\intercal \mathrm{diag}^{-1}(\boldsymbol{\Sigma}_r) - \frac{1}{2}\boldsymbol{\omega}_t^\intercal \boldsymbol{\Sigma}_r \boldsymbol{\omega}_t\Bigg) + \boldsymbol{\omega}_t^\intercal \boldsymbol{r}_{t+1}, \end{aligned}$$

with its constrained counterpart being (will verify later in Proposition 5 that log-normality of constrained random variables is admissible, and deduce the value of $\hat{\boldsymbol{\mu}}_r$ shortly)

$$\begin{aligned} \ln \hat{W}_{t+1} - \ln W_t \approx \Bigg(&\boldsymbol{\omega}_t^\intercal(\mathrm{diag}(\boldsymbol{P}_t)^{-1}\boldsymbol{D}_t - r_{0,t}\mathbf{1}) + r_{0,t} - \frac{C_t}{W_t} + \\ &+\frac{1}{2}\boldsymbol{\omega}_t^\intercal \mathrm{diag}^{-1}(\hat{\boldsymbol{\Sigma}}_r) - \frac{1}{2}\boldsymbol{\omega}_t^\intercal \hat{\boldsymbol{\Sigma}}_r \boldsymbol{\omega}_t\Bigg) + \boldsymbol{\omega}_t^\intercal \hat{\boldsymbol{r}}_{t+1}. \end{aligned}$$

Thus,

$$\begin{aligned} d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}) &= (\ln W_{t+1} - \ln \hat{W}_{t+1})^2 \approx \\ &\approx \left(\boldsymbol{\omega}_t^\intercal(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1}) + \frac{1}{2}\boldsymbol{\omega}_t^\intercal \mathrm{diag}^{-1}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r) - \frac{1}{2}\boldsymbol{\omega}_t^\intercal(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r)\boldsymbol{\omega}_t\right)^2. \end{aligned}$$

Since wealth process follows a geometric Brownian motion, equalizing expected growth

rates of true $\ln W_{t+1}$ and approximate $\ln \hat{W}_{t+1}$ when volatility $\boldsymbol{\Sigma}_r$ is replaced with $\hat{\boldsymbol{\Sigma}}_r$ necessitates an adjustment to the mean of approximating random variable, $\hat{\boldsymbol{\mu}}_r$. It is easy to see that the correct mean has to be

$$\hat{\boldsymbol{\mu}}_r := \boldsymbol{\mu}_r + \frac{1}{2}\mathrm{diag}^{-1}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r) - \frac{1}{2}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r)\boldsymbol{\omega}_t =: \boldsymbol{\mu}_r + \check{\boldsymbol{\mu}}_r.$$

Which also is, by the usual mean-as-a-minimum-MSE-estimator logic, the minimizer of

$$\mathrm{E}^f[d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1})] \approx \mathrm{E}^f\left[\left(\boldsymbol{\omega}_t^\intercal(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1}) + \frac{1}{2}\boldsymbol{\omega}_t^\intercal\mathrm{diag}^{-1}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r) - \frac{1}{2}\boldsymbol{\omega}_t^\intercal(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r)\boldsymbol{\omega}_t\right)^2\right],$$

producing as a result

$$\mathrm{E}^f[d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1})]\Big|_{\hat{\boldsymbol{\mu}}_r = \boldsymbol{\mu}_r + \check{\boldsymbol{\mu}}_r} \approx \boldsymbol{\omega}_t^\intercal \boldsymbol{\Psi}_r \boldsymbol{\omega}_t,$$

where we have introduced (in accordance with Proposition 5) notation

$$\boldsymbol{\Psi}_r := \mathrm{E}^f[(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r)(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r)^\intercal].$$

Now approximate distortion function can also be formulated as

$$\begin{aligned} d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}) &\approx \left(\boldsymbol{\omega}_t^\intercal(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1}) + \frac{1}{2}\boldsymbol{\omega}_t^\intercal\mathrm{diag}^{-1}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r) - \frac{1}{2}\boldsymbol{\omega}_t^\intercal(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r)\boldsymbol{\omega}_t\right)^2 = \\ &= \left(\boldsymbol{\omega}_t^\intercal(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r)\right)^2. \end{aligned}$$

Lastly, keeping time intervals infinitesimally short would leave us in continuous time framework, with the above expressions being exact.

□

F.2.2 Proof of Proposition 3.2

Proof. Consider simultaneous-move game first. The argument proceeds in 4 steps.

1. For a given probability density function $f_r(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1})$, we can write the objective function as

$$\mathrm{E}^f[d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1})] = \mathrm{E}^f\left[\left(\boldsymbol{\omega}_t^\intercal(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r)\right)^2\right] = \boldsymbol{\omega}_t^\intercal \boldsymbol{\Psi}_r \boldsymbol{\omega}_t,$$

   where we use (in accordance with Proposition 5) notation

$$\boldsymbol{\Psi}_r := \mathrm{E}^f[(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r)(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r)^\intercal].$$

2. Nature solves

$$\max_{\boldsymbol{\omega}_t} \mathrm{E}^f[d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1})] \iff \max_{\boldsymbol{\omega}_t} \boldsymbol{\omega}_t^\intercal \boldsymbol{\Psi}_r \boldsymbol{\omega}_t$$

   subject to

$$\begin{aligned} &\boldsymbol{\omega}_t^\intercal \mathbf{1} \le \widehat{\omega}_t^N && \forall \widehat{\omega}_t^N \in \mathbb{R}_+, \\ &\omega_{k,t} \ge 0 && \forall k \in \{1, \ldots, K\}. \end{aligned}$$

   (i) Perfect information case: $\boldsymbol{\Psi}_r$ is known. (Benchmark case.)

   Notice that

$$\begin{aligned} \mathrm{V}\left(\boldsymbol{\omega}_t^\intercal(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r)\right) &= \sum_k \omega_{k,t}^2 \psi_{r,k}^2 + 2\sum_{k \ne l} \omega_{k,t}\omega_{l,t}\psi_{r,kl} \le \\ &\le \sum_k \omega_{k,t}^2 \psi_{r,k}^2 + 2\sum_{k \ne l} \omega_{k,t}\omega_{l,t}\psi_{r,k}\psi_{r,l} \le \\ &\le \left(\widehat{\omega}_t^N\right)^2 \max_{k \in \{1,\ldots,K\}} \psi_{r,k}^2. \end{aligned}$$

Thus, Nature chooses corner solution:

$$\omega_{k,t} := \widehat{\omega}_t^N, \qquad k = k^\star := \arg \max_{k \in \{1,\ldots,K\}} \psi_{r,k}^2 \text{ (randomize if multiplicity)},$$
$$\omega_{k,t} := 0, \qquad k \neq k^\star.$$

(ii) Imperfect information case: $\boldsymbol{\Psi}_r$ unknown. (Simultaneous-move case.)

Still, Nature would want to choose corner solution. The principle of indifference (principle of insufficient reason) entails uniform distribution for the corner solution's candidate:

$$k^\star \sim \mathcal{U}(\{1, \ldots, K\}).$$

Thus, Nature's move is:

$$\omega_{k,t} := \widehat{\omega}_t^N, \qquad k = k^\star \sim \mathcal{U}(\{1, \ldots, K\}),$$
$$\omega_{k,t} := 0, \qquad k \neq k^\star.$$

3. Agent anticipates Nature's strategy and formulates the objective function to solve ($\forall \widehat{\omega}_t^A \in \mathbb{R}_+$; notice that $\widehat{\omega}_t^A \neq \widehat{\omega}_t^N$ in general, so coordination between players on exact share is not necessary for achieving an equilibrium):

$$\min_{f_r(\boldsymbol{r},\hat{\boldsymbol{r}})} \sum_k \frac{1}{K} \mathrm{E}^f [d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}) \mid \omega_{k,t} = \widehat{\omega}_t^A;\, \omega_{l,t} = 0,\, \forall l \neq k] \iff$$
$$\min_{f_r(\boldsymbol{r},\hat{\boldsymbol{r}})} \mathrm{E}^f \left[\sum_k \frac{1}{K} d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}) \middle| \omega_{k,t} = \widehat{\omega}_t^A;\, \omega_{l,t} = 0,\, \forall l \neq k \right] \iff$$
$$\min_{f_r(\boldsymbol{r},\hat{\boldsymbol{r}})} \mathrm{E}^f \left[\sum_k \frac{1}{K} \left(\boldsymbol{\omega}_t^\intercal (\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r) \right)^2 \middle| \omega_{k,t} = \widehat{\omega}_t^A;\, \omega_{l,t} = 0,\, \forall l \neq k \right] \iff$$
$$\min_{f_r(\boldsymbol{r},\hat{\boldsymbol{r}})} \mathrm{E}^f \left[\sum_k \frac{1}{K} \left(\boldsymbol{\omega}_t^\intercal (\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \frac{1}{2} \mathrm{diag}^{-1}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r) - \frac{1}{2} (\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_r) \boldsymbol{\omega}_t) \right)^2 \middle| \omega_{k,t} = \widehat{\omega}_t^A;\, \omega_{l,t} = 0,\, \forall l \neq k \right] \iff$$
$$\min_{f_r(\boldsymbol{r},\hat{\boldsymbol{r}})} \mathrm{E}^f \left[\sum_k \frac{1}{K} \left(\widehat{\omega}_t^A \right)^2 \left(r_{k,t+1} - \hat{r}_{k,t+1} + \frac{1}{2} (\sigma_{r,k}^2 - \hat{\sigma}_{r,k}^2) - \frac{1}{2} (\sigma_{r,k}^2 - \hat{\sigma}_{r,k}^2) \widehat{\omega}_t^A \right)^2 \right] \iff$$
$$\min_{f_r(\boldsymbol{r},\hat{\boldsymbol{r}})} \mathrm{E}^f \left[\sum_k \frac{1}{K} \left(\widehat{\omega}_t^A \right)^2 \left(r_{k,t+1} - \hat{r}_{k,t+1} + \check{\mu}_{r,k}(\widehat{\omega}_t^A) \right)^2 \right]$$

subject to standard informational constraints.

4. In Nash-equilibrium of this simultaneous-move game, players' strategies are as follows ($\forall \widehat{\omega}_t^A, \widehat{\omega}_t^N \in \mathbb{R}_+$):

   - Agent plays

   $$f_r^\star(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}) := \arg \min_{f_r(\boldsymbol{r}, \hat{\boldsymbol{r}})} \mathrm{E}^f \left[\sum_k \frac{1}{K} \left(\widehat{\omega}_t^A \right)^2 \left(r_{k,t+1} - \hat{r}_{k,t+1} + \check{\mu}_{r,k}(\widehat{\omega}_t^A) \right)^2 \right];$$

   - Nature plays a mixed strategy

   $$\begin{aligned} \omega_{k,t} &:= \widehat{\omega}_t^N, & k &= k^\star, \\ \omega_{k,t} &:= 0, & k &\neq k^\star, \end{aligned}$$

   where

   $$k^\star \sim \mathcal{U}(\{1, \ldots, K\}).$$

For a game with Nature as a second-mover, the proof requires only a slight modification in step 2, where the perfect information case applies. (Note that step 3 remains unchanged due to previous step's randomization in situation of multiplicity.)

□

F.2.3 Proof of Proposition 3.3

Proof. Plugging $\widehat{\omega}_t^A := \mathbf{1}^\intercal \boldsymbol{\omega}_t =: \widehat{\omega}_t$ into Agent's equilibrium strategy expression in Proposition 3.2, we immediately have

$$\sum_k \frac{1}{K} \widehat{\omega}_t^2 (r_{k,t+1} - \hat{r}_{k,t+1} + \check{\mu}_{r,k}(\widehat{\omega}_t))^2 \propto (\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r(\widehat{\omega}_t))^\intercal (\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r(\widehat{\omega}_t)) =: d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}),$$

utilizing the fact that for the purpose of extremization, distance metrics are defined only up to a constant of proportionality.

□

F.3 Proof of Proposition B.1

Proof. Substitute from (3.7), (3.8) and (3.9) into $d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1})$ and $\mathcal{I}(g_r(\boldsymbol{r}_{t+1}); h_r(\hat{\boldsymbol{r}}_{t+1}))$:

$$\begin{aligned} d(\boldsymbol{r}_{t+1}, \hat{\boldsymbol{r}}_{t+1}) &= (\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r(\widehat{\omega}_t))^{\mathsf{T}}(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r(\widehat{\omega}_t)) = \\ &= (\boldsymbol{\Xi}\boldsymbol{x} - \boldsymbol{\Xi}\hat{\boldsymbol{x}} + \boldsymbol{\Xi}\check{\boldsymbol{\mu}}(\widehat{\omega}_t))^{\mathsf{T}}(\boldsymbol{\Xi}\boldsymbol{x} - \boldsymbol{\Xi}\hat{\boldsymbol{x}} + \boldsymbol{\Xi}\check{\boldsymbol{\mu}}(\widehat{\omega}_t)) = \\ &= (\boldsymbol{x} - \hat{\boldsymbol{x}} + \check{\boldsymbol{\mu}}(\widehat{\omega}_t))^{\mathsf{T}}\boldsymbol{\Xi}^{\mathsf{T}}\boldsymbol{\Xi}(\boldsymbol{x} - \hat{\boldsymbol{x}} + \check{\boldsymbol{\mu}}(\widehat{\omega}_t)) = \\ &= (\boldsymbol{x} - \hat{\boldsymbol{x}} + \check{\boldsymbol{\mu}}(\widehat{\omega}_t))^{\mathsf{T}}(\boldsymbol{x} - \hat{\boldsymbol{x}} + \check{\boldsymbol{\mu}}(\widehat{\omega}_t)) =: d(\boldsymbol{x}, \hat{\boldsymbol{x}}), \end{aligned}$$

$$\begin{aligned} \mathcal{I}(g_r(\boldsymbol{r}_{t+1}); h_r(\hat{\boldsymbol{r}}_{t+1})) &= \mathcal{E}(g_r(\boldsymbol{r}_{t+1})) + \mathcal{E}(h_r(\hat{\boldsymbol{r}}_{t+1})) - \mathcal{E}(g_r(\boldsymbol{r}_{t+1}), h_r(\hat{\boldsymbol{r}}_{t+1})) = \\ &= \mathcal{E}(g_r(\boldsymbol{\Xi}\boldsymbol{x})) + \mathcal{E}(h_r(\boldsymbol{\Xi}\hat{\boldsymbol{x}})) - \mathcal{E}(g_r(\boldsymbol{\Xi}\boldsymbol{x}), h_r(\boldsymbol{\Xi}\hat{\boldsymbol{x}})) = \\ &= \mathcal{E}(g(\boldsymbol{x})) + \ln(\mathrm{abs}|\boldsymbol{\Xi}|) + \mathcal{E}(h(\hat{\boldsymbol{x}})) + \ln(\mathrm{abs}|\boldsymbol{\Xi}|) - \\ &\quad - \mathcal{E}(g(\boldsymbol{x}), h(\hat{\boldsymbol{x}})) - \ln(\mathrm{abs}(|\boldsymbol{\Xi}||\boldsymbol{\Xi}|)) = \\ &= \mathcal{E}(g(\boldsymbol{x})) + \mathcal{E}(h(\hat{\boldsymbol{x}})) - \mathcal{E}(g(\boldsymbol{x}), h(\hat{\boldsymbol{x}})) =: \mathcal{I}(g(\boldsymbol{x}); h(\hat{\boldsymbol{x}})). \end{aligned}$$

□

F.4 Proof of Proposition 4

Proof. We derive (or simply guess some parts of) the specific solution to the informational problem, then verify that it satisfies the necessary conditions for optimality. Given that objective function is convex, it has a unique minimum, so locating one candidate solution that satisfies the Karush–Kuhn–Tucker conditions suffices.

(a) Interior solution ("large" κ, "small" λ).

The following proof essentially replicates Sims's (2003, 2006) arguments; but this is a classic result in information theory, e.g. see Berger (1971) or Cover and Thomas (2006).

1. $\boldsymbol{x} \in \mathbb{R}^K$. Guess (and verify later) that $\hat{\boldsymbol{x}} \in \mathbb{R}^K$. Hence,

$$\nu(\boldsymbol{x}, \hat{\boldsymbol{x}}) = 0 \quad \forall \boldsymbol{x}, \hat{\boldsymbol{x}} \in \mathbb{R}^K.$$

2. Show that $e^{-\frac{1}{\lambda}\mu(\boldsymbol{x})} = (\pi\lambda)^{-\frac{K}{2}}$ is a valid element of the solution:

$$\begin{aligned}
1 &= \int_{\mathbb{R}^K} f(\boldsymbol{x}|\hat{\boldsymbol{x}})\, d\boldsymbol{x} = \int_{\mathbb{R}^K} e^{\frac{1}{\lambda}\nu(\boldsymbol{x},\hat{\boldsymbol{x}}) - \frac{1}{\lambda}\mu(\boldsymbol{x}) - \frac{1}{\lambda}(\boldsymbol{x}-\hat{\boldsymbol{x}}+\check{\boldsymbol{\mu}}(\widehat{\omega}_t))^\intercal(\boldsymbol{x}-\hat{\boldsymbol{x}}+\check{\boldsymbol{\mu}}(\widehat{\omega}_t))} =: \\
&=: \int_{\mathbb{R}^K} b(\boldsymbol{x})(2\pi)^{-\frac{K}{2}} \left|\frac{\lambda}{2}\boldsymbol{I}_K\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{x}}+\check{\boldsymbol{\mu}}(\widehat{\omega}_t))^\intercal \left(\frac{\lambda}{2}\boldsymbol{I}_K\right)^{-1}(\boldsymbol{x}-\hat{\boldsymbol{x}}+\check{\boldsymbol{\mu}}(\widehat{\omega}_t))}\, d\boldsymbol{x} = \\
&= \int_{\mathbb{R}^K} b(\boldsymbol{x})\phi(\hat{\boldsymbol{x}} - \check{\boldsymbol{\mu}}(\widehat{\omega}_t) - \boldsymbol{x} \mid \frac{\lambda}{2}\boldsymbol{I}_K)\, d\boldsymbol{x} =: (b * \phi)\,(\hat{\boldsymbol{x}} - \check{\boldsymbol{\mu}}(\widehat{\omega}_t)).
\end{aligned}$$

Applying Fourier transform to the convolution above, we get:

$$\delta(\boldsymbol{\xi}) = \widetilde{1} = \widetilde{b}(\boldsymbol{\xi})\widetilde{\phi}(\boldsymbol{\xi}|\frac{\lambda}{2}\boldsymbol{I}_K),$$

$$\widetilde{b}(\boldsymbol{\xi}) = \frac{\delta(\boldsymbol{\xi})}{\widetilde{\phi}(\boldsymbol{\xi}|\frac{\lambda}{2}\boldsymbol{I}_K)} = \delta(\boldsymbol{\xi})e^{2\pi i(\hat{\boldsymbol{x}}-\check{\boldsymbol{\mu}}(\widehat{\omega}_t))\cdot\boldsymbol{\xi} + \pi^2\lambda\boldsymbol{\xi}^\intercal\boldsymbol{\xi}} = \delta(\boldsymbol{\xi}) = \begin{cases} 1 & \text{if } \boldsymbol{\xi} = \mathbf{0}, \\ 0 & \text{if } \boldsymbol{\xi} \neq \mathbf{0}. \end{cases}$$

Inverse Fourier transform gives

$$b(\boldsymbol{x}) = \int_{\mathbb{R}^K} \widetilde{b}(\boldsymbol{\xi})e^{2\pi i\boldsymbol{x}\cdot\boldsymbol{\xi}}\, d\boldsymbol{\xi} = \int_{\mathbb{R}^K} \delta(\boldsymbol{\xi})e^{2\pi i\boldsymbol{x}\cdot\boldsymbol{\xi}}\, d\boldsymbol{\xi} = 1;$$

therefore,

$$e^{-\frac{1}{\lambda}\mu(\boldsymbol{x})} = (2\pi)^{-\frac{K}{2}} \left|\frac{\lambda}{2}\boldsymbol{I}_K\right|^{-\frac{1}{2}} = (\pi\lambda)^{-\frac{K}{2}}.$$

This means

$$\mu(\boldsymbol{x}) = \lambda \frac{K}{2} \ln(\pi\lambda).$$

3. Denoting

$$\boldsymbol{\epsilon} := \boldsymbol{x} - \hat{\boldsymbol{x}} + \check{\boldsymbol{\mu}}(\widehat{\omega}_t) \sim \mathcal{N}(\mathbf{0}, \frac{\lambda}{2}\boldsymbol{I}_K),$$

we can represent $\boldsymbol{x}$ as a sum of two different terms, $(\hat{\boldsymbol{x}} - \check{\boldsymbol{\mu}}(\widehat{\omega}_t))$ and "approximation error" $\boldsymbol{\epsilon}$:

$$\boldsymbol{x} = \hat{\boldsymbol{x}} - \check{\boldsymbol{\mu}}(\widehat{\omega}_t) + \boldsymbol{\epsilon}.$$

A convolution of independently distributed $(\hat{\boldsymbol{x}} - \check{\boldsymbol{\mu}}(\widehat{\omega}_t))$ with independent $\mathcal{N}$-distributed $\boldsymbol{\epsilon}$ that results in $\mathcal{N}$-distributed $\boldsymbol{x}$ implies $\mathcal{N}$ distribution for $(\hat{\boldsymbol{x}} - \check{\boldsymbol{\mu}}(\widehat{\omega}_t))$ too, and hence also for $\hat{\boldsymbol{x}}$:

$$\hat{\boldsymbol{x}} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}(\widehat{\omega}_t), \hat{\boldsymbol{\Sigma}}).$$

Therefore,

$$\boldsymbol{\Psi} := \mathrm{E}^f[(\boldsymbol{x} - \hat{\boldsymbol{x}} + \check{\boldsymbol{\mu}}(\widehat{\omega}_t))(\boldsymbol{x} - \hat{\boldsymbol{x}} + \check{\boldsymbol{\mu}}(\widehat{\omega}_t))^\intercal] = \mathrm{E}^\phi[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal] = \frac{\lambda}{2}\boldsymbol{I}_K,$$

and we also have

$$\boldsymbol{\Sigma} = \hat{\boldsymbol{\Sigma}} + \boldsymbol{\Psi}$$

(it can be seen here why we needed uncorrelated random variables: diagonal $\boldsymbol{\Sigma}$ ensures that resulting $\hat{\boldsymbol{\Sigma}}$ is positive semi-definite).

Lastly, conformity of the means (accounting for the bias term) as well as conformity of the variances in case of $\mathcal{N}$ densities $f(\boldsymbol{x}|\hat{\boldsymbol{x}})$ and $h(\hat{\boldsymbol{x}})$ necessarily leads to their product $f(\boldsymbol{x}, \hat{\boldsymbol{x}})$ satisfying the constraint $\int_{\mathrm{supp}(h)} f(\boldsymbol{x}, \hat{\boldsymbol{x}})\, d\hat{\boldsymbol{x}} = g(\boldsymbol{x})\ \forall \boldsymbol{x} \in \mathrm{supp}(g)$ (the $\mu(\boldsymbol{x})$-constraint).

4. Information processing capacity constraint reduces to

$$\kappa = \frac{1}{2}(\ln|\boldsymbol{\Sigma}| - \ln|\boldsymbol{\Psi}|) = \frac{1}{2}\left(\ln|\boldsymbol{\Sigma}| - \ln\left|\frac{\lambda}{2}\boldsymbol{I}_K\right|\right),$$

which means that

$$\lambda = 2\left(e^{-2\kappa}|\boldsymbol{\Sigma}|\right)^{\frac{1}{K}}.$$

(b) Boundary solution ("small" κ, "large" λ).

The essense of this proof is known in information theory as "reverse water-filling" solution, or more accurately "reverse water-filling on the eigenvalues" for cases like ours, e.g. see Berger (1971). This procedure is more general than what was used for interior solution.

0. If the condition $\sigma_k^2 > \frac{\lambda}{2} \quad \forall k \in \{1,\dots,K\}$ doesn't hold, i.e. $\exists k \in \{1,\dots,K\}: \ \sigma_k^2 \leq \frac{\lambda}{2}$, the interior solution violates the constraint $\int_{\text{supp}(h)} f(\boldsymbol{x},\hat{\boldsymbol{x}})\,d\hat{\boldsymbol{x}} = g(\boldsymbol{x}) \ \forall \boldsymbol{x} \in \text{supp}(g)$ (the $\mu(\boldsymbol{x})$-constraint).

1. $\boldsymbol{x} \in \mathbb{R}^K$. Look for solution with $\hat{\boldsymbol{x}} \in \mathbb{R}^{k^*} \times \{\hat{\mu}_{k^*+1}(\widehat{\omega}_t),\cdots,\hat{\mu}_K(\widehat{\omega}_t)\}$ (otherwise, $\epsilon_{k^*+1},\cdots,\epsilon_K$ will have non-zero means in order to satisfy the $\mu(\boldsymbol{x})$-constraint). Hence,

$$\nu(\boldsymbol{x},\hat{\boldsymbol{x}}) = 0 \quad \forall \boldsymbol{x} \in \mathbb{R}^K, \forall \hat{\boldsymbol{x}} \in \mathbb{R}^{k^*} \times \{\hat{\mu}_{k^*+1}(\widehat{\omega}_t),\cdots,\hat{\mu}_K(\widehat{\omega}_t)\}.$$

2. Show that $e^{-\frac{1}{\lambda}\mu(\boldsymbol{x})} = (2\pi)^{-\frac{K}{2}}\left(\left(\frac{\lambda}{2}\right)^{k^*}\sigma_{k^*+1}^2 \times \cdots \times \sigma_K^2\right)^{-\frac{1}{2}}$ is a valid element of the solution:

$$\begin{aligned}
1 &= \int_{\mathbb{R}^K} f(\boldsymbol{x}|\hat{\boldsymbol{x}})\,d\boldsymbol{x} = \int_{\mathbb{R}^K} e^{\frac{1}{\lambda}\nu(\boldsymbol{x},\hat{\boldsymbol{x}})-\frac{1}{\lambda}\mu(\boldsymbol{x})-\frac{1}{\lambda}(\boldsymbol{x}-\hat{\boldsymbol{x}}+\check{\boldsymbol{\mu}}(\widehat{\omega}_t))^\intercal(\boldsymbol{x}-\hat{\boldsymbol{x}}+\check{\boldsymbol{\mu}}(\widehat{\omega}_t))}\,d\boldsymbol{x} =: \\
&=: \int_{\mathbb{R}^K} b(\boldsymbol{x})(2\pi)^{-\frac{K}{2}}\,|\boldsymbol{\Psi}|^{-\frac{1}{2}}\,e^{-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{x}}+\check{\boldsymbol{\mu}}(\widehat{\omega}_t))^\intercal\boldsymbol{\Psi}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{x}}+\check{\boldsymbol{\mu}}(\widehat{\omega}_t))}\,d\boldsymbol{x} = \\
&= \int_{\mathbb{R}^K} b(\boldsymbol{x})\phi(\hat{\boldsymbol{x}}-\check{\boldsymbol{\mu}}(\widehat{\omega}_t)-\boldsymbol{x} \mid \boldsymbol{\Psi})\,d\boldsymbol{x} =: (b*\phi)\,(\hat{\boldsymbol{x}}-\check{\boldsymbol{\mu}}(\widehat{\omega}_t)).
\end{aligned}$$

Applying Fourier transform to the convolution above, we get:

$$\delta(\boldsymbol{\xi}) = \widetilde{1} = \widetilde{b}(\boldsymbol{\xi})\widetilde{\phi}(\boldsymbol{\xi}|\boldsymbol{\Psi}),$$

$$\widetilde{b}(\boldsymbol{\xi}) = \frac{\delta(\boldsymbol{\xi})}{\widetilde{\phi}(\boldsymbol{\xi}|\boldsymbol{\Psi})} = \delta(\boldsymbol{\xi})e^{2\pi i(\hat{\boldsymbol{x}}-\check{\boldsymbol{\mu}}(\widehat{\omega}_t))\cdot\boldsymbol{\xi}+\pi^2(\lambda\sum_1^{k^*}\xi_k^2+2\sum_{k^*+1}^{K}\sigma_k^2\xi_k^2)} =$$
$$= \delta(\boldsymbol{\xi}) = \begin{cases} 1 & \text{if } \boldsymbol{\xi} = \mathbf{0}, \\ 0 & \text{if } \boldsymbol{\xi} \neq \mathbf{0}. \end{cases}$$

Inverse Fourier transform gives

$$b(\boldsymbol{x}) = \int_{\mathbb{R}^K} \widetilde{b}(\boldsymbol{\xi})e^{2\pi i\boldsymbol{x}\cdot\boldsymbol{\xi}}\, d\boldsymbol{\xi} = \int_{\mathbb{R}^K} \delta(\boldsymbol{\xi})e^{2\pi i\boldsymbol{x}\cdot\boldsymbol{\xi}}\, d\boldsymbol{\xi} = 1,$$

therefore,

$$e^{-\frac{1}{\lambda}\mu(\boldsymbol{x})} = (2\pi)^{-\frac{K}{2}}\,|\boldsymbol{\Psi}|^{-\frac{1}{2}} = (2\pi)^{-\frac{K}{2}}\left(\left(\frac{\lambda}{2}\right)^{k^*}\sigma_{k^*+1}^2\times\cdots\times\sigma_K^2\right)^{-\frac{1}{2}}.$$

This means

$$\mu(\boldsymbol{x}) = \lambda\left(\frac{K}{2}\ln(2\pi) + \frac{1}{2}\ln\left(\left(\frac{\lambda}{2}\right)^{k^*}\sigma_{k^*+1}^2\times\cdots\times\sigma_K^2\right)\right).$$

3. Denoting

$$\boldsymbol{\epsilon} := \boldsymbol{x} - \hat{\boldsymbol{x}} + \check{\boldsymbol{\mu}}(\widehat{\omega}_t) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}),$$

we can represent $\boldsymbol{x}$ as a sum of two different terms, $(\hat{\boldsymbol{x}} - \check{\boldsymbol{\mu}}(\widehat{\omega}_t))$ and "approximation error" $\boldsymbol{\epsilon}$:

$$\boldsymbol{x} = \hat{\boldsymbol{x}} - \check{\boldsymbol{\mu}}(\widehat{\omega}_t) + \boldsymbol{\epsilon}.$$

A convolution of independently distributed $(\hat{\boldsymbol{x}} - \check{\boldsymbol{\mu}}(\widehat{\omega}_t))$ with independent $\mathcal{N}$-distributed

$\boldsymbol{\epsilon}$ that results in $\mathcal{N}$-distributed $\boldsymbol{x}$ implies $\mathcal{N}$ distribution for $(\hat{\boldsymbol{x}} - \check{\boldsymbol{\mu}}(\widehat{\omega}_t))$ too, and hence also for $\hat{\boldsymbol{x}}$:

$$\hat{\boldsymbol{x}} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}(\widehat{\omega}_t), \hat{\boldsymbol{\Sigma}}).$$

Therefore (with $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K\}$ being the standard basis of $\mathbb{R}^K$),

$$\boldsymbol{\Psi} := \mathrm{E}^f[(\boldsymbol{x} - \hat{\boldsymbol{x}} + \check{\boldsymbol{\mu}}(\widehat{\omega}_t))(\boldsymbol{x} - \hat{\boldsymbol{x}} + \check{\boldsymbol{\mu}}(\widehat{\omega}_t))^\intercal] = \mathrm{E}^\phi[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal] = \begin{bmatrix} \frac{\lambda}{2}\boldsymbol{I}_{k^*} & \boldsymbol{0} \\ \boldsymbol{0} & \sum_{k=1}^{K-k^*} \boldsymbol{e}_k^\intercal \boldsymbol{\Sigma}_{k^*+1K} \boldsymbol{e}_k \end{bmatrix},$$

and we also have

$$\boldsymbol{\Sigma} = \hat{\boldsymbol{\Sigma}} + \boldsymbol{\Psi}$$

(it can be seen here why we needed uncorrelated random variables: diagonal $\boldsymbol{\Sigma}$ ensures that resulting $\hat{\boldsymbol{\Sigma}}$ is positive semi-definite).

Lastly, conformity of the means (accounting for the bias term) as well as conformity of the variances in case of $\mathcal{N}$ densities $f(\boldsymbol{x}|\hat{\boldsymbol{x}})$ and $h(\hat{\boldsymbol{x}})$ necessarily leads to their product $f(\boldsymbol{x}, \hat{\boldsymbol{x}})$ satisfying the constraint $\int_{\mathrm{supp}(h)} f(\boldsymbol{x}, \hat{\boldsymbol{x}})\, d\hat{\boldsymbol{x}} = g(\boldsymbol{x})\ \forall \boldsymbol{x} \in \mathrm{supp}(g)$ (the $\mu(\boldsymbol{x})$-constraint).

4. Information processing capacity constraint reduces to

$$\kappa = \frac{1}{2}(\ln|\boldsymbol{\Sigma}| - \ln|\boldsymbol{\Psi}|) = \frac{1}{2}\left(\ln|\boldsymbol{\Sigma}| - \ln\left|\frac{\lambda}{2}\boldsymbol{I}_{k^*}\right| - \ln\left|\sum_{k=1}^{K-k^*} \boldsymbol{e}_k^\intercal \boldsymbol{\Sigma}_{k^*+1K} \boldsymbol{e}_k\right|\right),$$

which means that

$$\lambda = 2\left(e^{-2\kappa}\sigma_{k^*+1}^{-2} \cdots \sigma_K^{-2}|\boldsymbol{\Sigma}|\right)^{\frac{1}{k^*}}.$$

□

F.5 Proof of Proposition 5

Proof. Premultiplying with $\boldsymbol{\Xi}$ equation $\boldsymbol{x} = \hat{\boldsymbol{x}} - \check{\boldsymbol{\mu}}(\widehat{\omega}_t) + \boldsymbol{\epsilon}$ produces

$$\boldsymbol{r}_{t+1} = \hat{\boldsymbol{r}}_{t+1} - \check{\boldsymbol{\mu}}_r(\widehat{\omega}_t) + \boldsymbol{\epsilon}_{r,t+1},$$

using (3.7), (3.8), (3.9), and defining $\boldsymbol{\epsilon}_{r,t+1} := \boldsymbol{\Xi}\boldsymbol{\epsilon}$.

Premultiplying with $\boldsymbol{\Xi}$ and postmultiplying with $\boldsymbol{\Xi}^{-1}$ equation $\boldsymbol{\Sigma} = \hat{\boldsymbol{\Sigma}} + \boldsymbol{\Psi}$ produces

$$\boldsymbol{\Sigma}_r = \hat{\boldsymbol{\Sigma}}_r + \boldsymbol{\Psi}_r,$$

as from formulas in $\mathcal{P}_{\diagdown}$, $\hat{\boldsymbol{\Sigma}}_r := \boldsymbol{\Xi}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Xi}^{-1}$ and $\boldsymbol{\Psi}_r := \boldsymbol{\Xi}\boldsymbol{\Psi}\boldsymbol{\Xi}^{-1}$.

Given that $\hat{\boldsymbol{x}} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}(\widehat{\omega}_t), \hat{\boldsymbol{\Sigma}})$, $\hat{\boldsymbol{r}}_{t+1} = \boldsymbol{\Xi}\hat{\boldsymbol{x}}$ (from equation 3.8) is distributed as $\mathcal{N}(\boldsymbol{\Xi}\hat{\boldsymbol{\mu}}(\widehat{\omega}_t), \boldsymbol{\Xi}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Xi}^{\intercal})$. Since before the decorrelating transformation mean of $\hat{\boldsymbol{r}}_{t+1}$ was $\hat{\boldsymbol{\mu}}_r(\widehat{\omega}_t)$ by Proposition 3.1, and also using relationships in $\mathcal{P}_{\diagdown}$, we have $\hat{\boldsymbol{r}}_{t+1} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_r(\widehat{\omega}_t), \hat{\boldsymbol{\Sigma}}_r)$.

Given that $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi})$, $\boldsymbol{\epsilon}_{r,t+1} = \boldsymbol{\Xi}\boldsymbol{\epsilon}$ (by the definition from above) is distributed as $\mathcal{N}(\boldsymbol{\Xi}\boldsymbol{0}, \boldsymbol{\Xi}\boldsymbol{\Psi}\boldsymbol{\Xi}^{\intercal})$. Using relationships in $\mathcal{P}_{\diagdown}$ we get $\boldsymbol{\epsilon}_{r,t+1} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_r)$.

□

F.6 Proof of Proposition C.1

Proof. Let $\boldsymbol{\theta}_{\iota-1} := \{q_{0,t,\iota-1}, \boldsymbol{q}_{t,\iota-1}\}$ be the proposed parameter choice that has been accepted at iteration $(\iota - 1)$. By construction, it approaches the optimal parameter choice $\boldsymbol{\theta}^*$, i.e. $||\boldsymbol{\theta}_{\iota-1} - \boldsymbol{\theta}^*||_2^2 \to 0$ as ι increases. The corresponding update of the bias of the mean can

then be represented by

$$\begin{aligned}
\left|\left|\check{\boldsymbol{\mu}}_{r,\iota-1}-\check{\boldsymbol{\mu}}_r\right|\right|_2^2 &= \Big|\Big|\frac{1}{2}\text{diag}^{-1}(\boldsymbol{\Sigma}_r-\hat{\boldsymbol{\Sigma}}_{r,\iota-1})-\frac{1}{2}(\boldsymbol{\Sigma}_r-\hat{\boldsymbol{\Sigma}}_{r,\iota-1})\boldsymbol{\omega}_{t,\iota-1}- \\
&\quad -\frac{1}{2}\text{diag}^{-1}(\boldsymbol{\Sigma}_r-\hat{\boldsymbol{\Sigma}}_r)+\frac{1}{2}(\boldsymbol{\Sigma}_r-\hat{\boldsymbol{\Sigma}}_r)\boldsymbol{\omega}_t\Big|\Big|_2^2= \\
&= \Big|\Big|\frac{1}{2}\text{diag}^{-1}(\boldsymbol{\Psi}_{r,\iota-1})-\frac{1}{2}\boldsymbol{\Psi}_{r,\iota-1}\boldsymbol{\omega}_{t,\iota-1}-\frac{1}{2}\text{diag}^{-1}(\boldsymbol{\Psi}_r)+\frac{1}{2}\boldsymbol{\Psi}_r\boldsymbol{\omega}_t\Big|\Big|_2^2= \\
&= \Big|\Big|\frac{1}{2}\text{diag}^{-1}(\boldsymbol{\Psi}_r)-\frac{1}{2}\boldsymbol{\Psi}_r\boldsymbol{\omega}_{t,\iota-1}-\frac{1}{2}\text{diag}^{-1}(\boldsymbol{\Psi}_r)+\frac{1}{2}\boldsymbol{\Psi}_r\boldsymbol{\omega}_t\Big|\Big|_2^2= \\
&= \frac{1}{2}\left|\left|\boldsymbol{\Psi}_r\left(\boldsymbol{\omega}_{t,\iota-1}-\boldsymbol{\omega}_t\right)\right|\right|_2^2= \\
&= \frac{1}{2}\left|\left|\boldsymbol{\Psi}_r\left(\frac{1}{W_t}\text{diag}(\boldsymbol{P}_t)\boldsymbol{q}_{t,\iota-1}-\frac{1}{W_t}\text{diag}(\boldsymbol{P}_t)\boldsymbol{q}_t^*\right)\right|\right|_2^2= \\
&= \frac{1}{2}\left|\left|\boldsymbol{\Psi}_r\frac{1}{W_t}\text{diag}(\boldsymbol{P}_t)\left(\boldsymbol{q}_{t,\iota-1}-\boldsymbol{q}_t^*\right)\right|\right|_2^2\leq \\
&\leq \frac{1}{2}\left|\left|\boldsymbol{\Psi}_r\frac{1}{W_t}\text{diag}(\boldsymbol{P}_t)\right|\right|_2^2\left|\left|\boldsymbol{q}_{t,\iota-1}-\boldsymbol{q}_t^*\right|\right|_2^2,
\end{aligned}$$

where the third equality is due to invariance of the solution to informational problem (modulo specific values of the bias term), and the weak inequality is due to consistency of induced matrix norm. Clearly, improvement in $\boldsymbol{q}_{t,\iota-1}$ directly leads to improvement in $\check{\boldsymbol{\mu}}_{r,\iota-1}$.

We are dealing with a well-posed consumption-investment maximization problem that posesses a unique maximum to which the above iterative optimization procedure converges. Covergence at some iteration ι by definition means that (making the dependence of $\varphi(\cdot|\boldsymbol{\theta})$ on $\hat{\boldsymbol{\mu}}_r$ and in turn on $\check{\boldsymbol{\mu}}_r$ explicit)

$$\left|\left|\varphi(\boldsymbol{\Xi}^\intercal\hat{\boldsymbol{r}}_{t+1}|\boldsymbol{\theta}_\iota,\hat{\boldsymbol{\mu}}_r(\check{\boldsymbol{\mu}}_{r,\iota-1}))-\varphi(\boldsymbol{\Xi}^\intercal\hat{\boldsymbol{r}}_{t+1}|\boldsymbol{\theta}^*,\hat{\boldsymbol{\mu}}_r(\check{\boldsymbol{\mu}}_r))\right|\right|_2^2<\varepsilon_\varphi$$

for some pre-specified $\varepsilon_\varphi>0$. This corresponds to covergence in chosen parameters:

$$\left|\left|\boldsymbol{q}_{t,\iota}-\boldsymbol{q}_t^*\right|\right|_2^2\leq||\boldsymbol{\theta}_\iota-\boldsymbol{\theta}^*||_2^2<\varepsilon_\theta$$

for some small $\varepsilon_\theta > 0$, as well as in bias of the mean:

$$||\check{\boldsymbol{\mu}}_{r,\iota} - \check{\boldsymbol{\mu}}_r||_2^2 \leq \frac{1}{2}\left|\left|\boldsymbol{\Psi}_r \frac{1}{W_t}\mathrm{diag}(\boldsymbol{P}_t)\right|\right|_2^2 \left|\left|\boldsymbol{q}_{t,\iota} - \boldsymbol{q}_t^*\right|\right|_2^2 < M\varepsilon_\theta =: \varepsilon_\mu$$

for some $0 < M < \infty$ (finiteness of M is guaranteed by stationarity of the environment).

According to optimization procedure considered here, upon reaching the optimum we have

$$\boldsymbol{q}_{t,\iota} = \boldsymbol{q}_{t,\iota-1},$$

with a related implication for updating of the bias term:

$$\begin{aligned}
\check{\boldsymbol{\mu}}_{r,\iota} &= \frac{1}{2}\mathrm{diag}^{-1}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_{r,\iota}) - \frac{1}{2}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_{r,\iota})\boldsymbol{\omega}_{t,\iota} = \\
&= \frac{1}{2}\mathrm{diag}^{-1}(\boldsymbol{\Psi}_r) - \frac{1}{2}\boldsymbol{\Psi}_r \frac{1}{W_t}\mathrm{diag}(\boldsymbol{P}_t)\boldsymbol{q}_{t,\iota} = \\
&= \frac{1}{2}\mathrm{diag}^{-1}(\boldsymbol{\Psi}_r) - \frac{1}{2}\boldsymbol{\Psi}_r \frac{1}{W_t}\mathrm{diag}(\boldsymbol{P}_t)\boldsymbol{q}_{t,\iota-1} = \\
&= \frac{1}{2}\mathrm{diag}^{-1}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_{r,\iota-1}) - \frac{1}{2}(\boldsymbol{\Sigma}_r - \hat{\boldsymbol{\Sigma}}_{r,\iota-1})\boldsymbol{\omega}_{t,\iota-1} = \check{\boldsymbol{\mu}}_{r,\iota-1}.
\end{aligned}$$

Which is immediately reftected in update of the mean, verifying the convergence:

$$\begin{aligned}
||\varphi(\boldsymbol{x}|\boldsymbol{\theta}_{\iota-1}) - \varphi(\boldsymbol{x}|\boldsymbol{\theta}^*)||_2^2 &= \left|\left|\varphi(\boldsymbol{\Xi}^\intercal \hat{\boldsymbol{r}}_{t+1}|\boldsymbol{\theta}_\iota, \hat{\boldsymbol{\mu}}_r(\check{\boldsymbol{\mu}}_{r,\iota})) - \varphi(\boldsymbol{\Xi}^\intercal \hat{\boldsymbol{r}}_{t+1}|\boldsymbol{\theta}^*, \hat{\boldsymbol{\mu}}_r(\check{\boldsymbol{\mu}}_r))\right|\right|_2^2 = \\
&= \left|\left|\varphi(\boldsymbol{\Xi}^\intercal \hat{\boldsymbol{r}}_{t+1}|\boldsymbol{\theta}_\iota, \hat{\boldsymbol{\mu}}_r(\check{\boldsymbol{\mu}}_{r,\iota-1})) - \varphi(\boldsymbol{\Xi}^\intercal \hat{\boldsymbol{r}}_{t+1}|\boldsymbol{\theta}^*, \hat{\boldsymbol{\mu}}_r(\check{\boldsymbol{\mu}}_r))\right|\right|_2^2 < \varepsilon_\varphi.
\end{aligned}$$

Finalizing the argument, at every iteration Propositions 4–5 hold with bias term $\check{\boldsymbol{\mu}}_{r,\iota}$ and mean $\hat{\boldsymbol{\mu}}_{r,\iota}$. Moreover, at the optimum $\check{\boldsymbol{\mu}}_{r,\iota}$, $\hat{\boldsymbol{\mu}}_{r,\iota}$ and $\boldsymbol{\omega}_{t,\iota}$ are consistent with definitions for $\check{\boldsymbol{\mu}}_r$, $\hat{\boldsymbol{\mu}}_r$ and $\boldsymbol{\omega}_t$ in Proposition 3.1. This gives the stated result.

□

F.7 Proof of Corollary 2

Proof. For $\boldsymbol{r}_{t+1} \sim \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ (equation 3.4), with $\boldsymbol{\Sigma}_r = \boldsymbol{\Xi}\boldsymbol{\Sigma}\boldsymbol{\Xi}^{-1}$ and $\boldsymbol{\Sigma}$ diagonal (using formulas in $\mathcal{P}_{\boxslash}$), we have

$$\rho_{r,kl} := \frac{\sum_{m=1}^{K} \xi_{km}\xi_{lm}\sigma_m^2}{\left(\sum_{m=1}^{K} \xi_{km}^2\sigma_m^2\right)^{1/2}\left(\sum_{m=1}^{K} \xi_{lm}^2\sigma_m^2\right)^{1/2}}.$$

From Proposition 5 we obtain $\hat{\boldsymbol{r}}_{t+1} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_r, \hat{\boldsymbol{\Sigma}}_r)$, where $\hat{\boldsymbol{\Sigma}}_r = \boldsymbol{\Xi}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Xi}^{-1}$ with diagonal $\hat{\boldsymbol{\Sigma}}$, hence

$$\begin{aligned}\hat{\rho}_{r,kl} :=& \frac{\sum_{m=1}^{K} \xi_{km}\xi_{lm}\hat{\sigma}_m^2}{\left(\sum_{m=1}^{K} \xi_{km}^2\hat{\sigma}_m^2\right)^{1/2}\left(\sum_{m=1}^{K} \xi_{lm}^2\hat{\sigma}_m^2\right)^{1/2}} = \\ =& \frac{\sum_{m=1}^{K} \xi_{km}\xi_{lm}(\sigma_m^2 - \psi_m^2)}{\left(\sum_{m=1}^{K} \xi_{km}^2(\sigma_m^2 - \psi_m^2)\right)^{1/2}\left(\sum_{m=1}^{K} \xi_{lm}^2(\sigma_m^2 - \psi_m^2)\right)^{1/2}}.\end{aligned}$$

In the interior solution case, $\psi_m^2 = \psi_n^2, \forall m, n \in \{1, \ldots, K\}$, thus

$$\begin{aligned}\hat{\rho}_{r,kl} =& \frac{\sum_{m=1}^{K} \xi_{km}\xi_{lm}(\sigma_m^2 - \psi_1^2)}{\left(\sum_{m=1}^{K} \xi_{km}^2(\sigma_m^2 - \psi_1^2)\right)^{1/2}\left(\sum_{m=1}^{K} \xi_{lm}^2(\sigma_m^2 - \psi_1^2)\right)^{1/2}} = \\ =& \frac{\sum_{m=1}^{K} \xi_{km}\xi_{lm}\sigma_m^2}{\left(\sum_{m=1}^{K} \xi_{km}^2\sigma_m^2 - \psi_1^2\right)^{1/2}\left(\sum_{m=1}^{K} \xi_{lm}^2\sigma_m^2 - \psi_1^2\right)^{1/2}},\end{aligned}$$

with the last equality following from orthogonality of the matrix of eigenvectors $\boldsymbol{\Xi}$, i.e. from the fact that $\sum_{m=1}^{K} \xi_{km}\xi_{lm} = \delta_{kl}$, where δ_{kl} is Kronecker delta function returning 1 when $k = l$ and 0 otherwise. Simple algebraic manipulations deliver the stated relationship:

$$\hat{\rho}_{r,kl} = \rho_{r,kl} \times \frac{\left(\sum_{m=1}^{K} \xi_{km}^2\sigma_m^2\right)^{1/2}\left(\sum_{m=1}^{K} \xi_{lm}^2\sigma_m^2\right)^{1/2}}{\left(\sum_{m=1}^{K} \xi_{km}^2\sigma_m^2 - \psi_1^2\right)^{1/2}\left(\sum_{m=1}^{K} \xi_{lm}^2\sigma_m^2 - \psi_1^2\right)^{1/2}}, \quad \forall k, l \in \{1, \ldots, K\},$$

with the fraction term on the right clearly being larger than or equal to 1, thus pushing $|\hat{\rho}_{r,kl}|$ from $|\rho_{r,kl}|$ towards 1. Lastly, Proposition 4 provides the value of ψ_1^2.

In the boundary solution case, $\psi_m^2 \neq \psi_n^2, \forall m, n \in \{1, \ldots, K\}$ in general, so a simple

argument from above does not go through. The proof is based instead on providing two contrasting examples. Given set $\{\sigma_k^2\}_1^K$ sorted in descending order, for each $k \in \{1, \ldots, K\}$ take

$$\psi_k^2 := \begin{cases} \sigma_K^2 & \text{if } k = K, \\ \sigma_K^2 + \psi_+^2 & \text{if } k \neq K, \end{cases}$$

where $\sigma_k^2 - \sigma_K^2 > \psi_+^2 \geq 0$ for every $k \in \{1, \ldots, K-1\}$. Then we have:

$$\begin{aligned} \hat{\rho}_{r,kl} =& \frac{\sum_{m=1}^K \xi_{km}\xi_{lm}(\sigma_m^2 - \sigma_K^2) - \sum_{m=1}^{K-1} \xi_{km}\xi_{lm}\psi_+^2}{\left(\sum_{m=1}^K \xi_{km}^2(\sigma_m^2 - \sigma_K^2) - \sum_{m=1}^{K-1} \xi_{km}^2\psi_+^2\right)^{1/2}\left(\sum_{m=1}^K \xi_{lm}^2(\sigma_m^2 - \sigma_K^2) - \sum_{m=1}^{K-1} \xi_{lm}^2\psi_+^2\right)^{1/2}} = \\ =& \frac{\sum_{m=1}^K \xi_{km}\xi_{lm}\sigma_m^2 + \xi_{kK}\xi_{lK}\psi_+^2}{\left(\sum_{m=1}^K \xi_{km}^2\sigma_m^2 - \sigma_K^2 - \psi_+^2 + \xi_{kK}^2\psi_+^2\right)^{1/2}\left(\sum_{m=1}^K \xi_{lm}^2\sigma_m^2 - \sigma_K^2 - \psi_+^2 + \xi_{lK}^2\psi_+^2\right)^{1/2}}, \end{aligned}$$

with the last step using orthogonality of the matrix of eigenvectors $\mathbf{\Xi}$. Now consider the limit of $\hat{\rho}_{r,kl}$ when σ_K^2 becomes negligibly small:

$$\hat{\rho}_{r,kl,\text{-}K} := \lim_{\sigma_K^2 \to 0+} \hat{\rho}_{r,kl} = \frac{\sum_{m=1}^{K-1} \xi_{km}\xi_{lm}\sigma_m^2 + \xi_{kK}\xi_{lK}\psi_+^2}{\left(\sum_{m=1}^{K-1} \xi_{km}^2\sigma_m^2 - \psi_+^2 + \xi_{kK}^2\psi_+^2\right)^{1/2}\left(\sum_{m=1}^{K-1} \xi_{lm}^2\sigma_m^2 - \psi_+^2 + \xi_{lK}^2\psi_+^2\right)^{1/2}}.$$

Evaluated at $\psi_+^2 := 0$, this limit for $\hat{\rho}_{r,kl}$ would coincide with its counterpart for $\rho_{r,kl}$ irrespective of the values of ξ_{kK} and ξ_{lK}:

$$\begin{aligned} \hat{\rho}_{r,kl,\text{-}K}\Big|_{\psi_+^2=0} = \lim_{\sigma_K^2 \to 0+} \hat{\rho}_{r,kl}\Big|_{\psi_+^2=0} &= \frac{\sum_{m=1}^{K-1} \xi_{km}\xi_{lm}\sigma_m^2}{\left(\sum_{m=1}^{K-1} \xi_{km}^2\sigma_m^2\right)^{1/2}\left(\sum_{m=1}^{K-1} \xi_{lm}^2\sigma_m^2\right)^{1/2}} = \\ &= \lim_{\sigma_K^2 \to 0+} \rho_{r,kl}\Big|_{\psi_+^2=0} = \rho_{r,kl,\text{-}K}\Big|_{\psi_+^2=0}. \end{aligned}$$

Assume $\rho_{r,kl,\text{-}K} > 0$. Choose $\xi_{kK} = \pm 1/\sqrt{2}$ and $\xi_{lK} = \mp 1/\sqrt{2}$, which produces

$$\hat{\rho}_{r,kl,\text{-}K} = \frac{\sum_{m=1}^{K-1} \xi_{km}\xi_{lm}\sigma_m^2 - \frac{1}{2}\psi_+^2}{\left(\sum_{m=1}^{K-1} \xi_{km}^2\sigma_m^2 - \frac{1}{2}\psi_+^2\right)^{1/2}\left(\sum_{m=1}^{K-1} \xi_{lm}^2\sigma_m^2 - \frac{1}{2}\psi_+^2\right)^{1/2}}.$$

Given that expression $\hat{\rho}_{r,kl,\text{-}K}$ is continuous and differentiable on the whole interval of ad-

missible ψ_+^2, we can take respective derivative at point $\psi_+^2 = 0$, obtaining (after some rearrangement):

$$\begin{aligned}\left.\frac{\partial\hat{\rho}_{r,kl,\text{-}K}}{\partial\psi_+^2}\right|_{\psi_+^2=0} &= \hat{\rho}_{r,kl,\text{-}K} \times \left(\frac{1}{2}\frac{1}{\sum_{m=1}^{K-1}\xi_{km}^2\sigma_m^2 - \frac{1}{2}\psi_+^2} + \frac{1}{2}\frac{1}{\sum_{m=1}^{K-1}\xi_{lm}^2\sigma_m^2 - \frac{1}{2}\psi_+^2}\right)\Bigg|_{\psi_+^2=0} - \\ &\quad - \left.\frac{1}{\left(\sum_{m=1}^{K-1}\xi_{km}^2\sigma_m^2 - \frac{1}{2}\psi_+^2\right)^{1/2}\left(\sum_{m=1}^{K-1}\xi_{lm}^2\sigma_m^2 - \frac{1}{2}\psi_+^2\right)^{1/2}}\right|_{\psi_+^2=0} = \\ &= \rho_{r,kl,\text{-}K} \times \left(\frac{1}{2}\frac{1}{\sum_{m=1}^{K-1}\xi_{km}^2\sigma_m^2} + \frac{1}{2}\frac{1}{\sum_{m=1}^{K-1}\xi_{lm}^2\sigma_m^2}\right) - \\ &\quad - \frac{1}{\left(\sum_{m=1}^{K-1}\xi_{km}^2\sigma_m^2\right)^{1/2}\left(\sum_{m=1}^{K-1}\xi_{lm}^2\sigma_m^2\right)^{1/2}}.\end{aligned}$$

By Young's inequality for products,

$$\frac{1}{2}\frac{1}{\sum_{m=1}^{K-1}\xi_{km}^2\sigma_m^2} + \frac{1}{2}\frac{1}{\sum_{m=1}^{K-1}\xi_{lm}^2\sigma_m^2} \geq \frac{1}{\left(\sum_{m=1}^{K-1}\xi_{km}^2\sigma_m^2\right)^{1/2}\left(\sum_{m=1}^{K-1}\xi_{lm}^2\sigma_m^2\right)^{1/2}},$$

with equality holding if and only if $\sum_{m=1}^{K-1}\xi_{km}^2\sigma_m^2 = \sum_{m=1}^{K-1}\xi_{lm}^2\sigma_m^2$. Unless the equality does hold, there always exists $\rho_{r,kl,\text{-}K} \in (0,1]$ such that $\partial\hat{\rho}_{r,kl,\text{-}K}/\partial\psi_+^2\big|_{\psi_+^2=0} > 0$, in which case $\hat{\rho}_{r,kl}$ moves away from $\rho_{r,kl}$ towards 1; and even without invocation of Young's inequality, it's trivial to find $\rho_{r,kl,\text{-}K} \in (0,1]$ such that $\partial\hat{\rho}_{r,kl,\text{-}K}/\partial\psi_+^2\big|_{\psi_+^2=0} < 0$, in which case $\hat{\rho}_{r,kl}$ moves away from $\rho_{r,kl}$ towards 0. If $\rho_{r,kl,\text{-}K} = 0$, we immediately get $\partial\hat{\rho}_{r,kl,\text{-}K}/\partial\psi_+^2\big|_{\psi_+^2=0} < 0$, in which case $\hat{\rho}_{r,kl}$ moves away from $\rho_{r,kl} = 0$ towards -1. (The situation under assumption of $\rho_{r,kl,\text{-}K} < 0$ can be handled similarly.)

□

APPENDIX G

ALGORITHM FOR DECISION-MAKING UNDER RISK

G.1 Scalar random variable case

Consider the primitive construction block of the process of decision-making under uncertainty: evaluation of a simple lottery. We will present the detailed steps of such an evaluation using an illustrative example.

Let x be a discretely distributed scalar random variable with probability mass function $\ddot{g}(x)$:

$$x \sim \ddot{g}(x), \tag{G.1}$$

with $\ddot{g}(x)$ defined, say, as in Figure G.0.

$$\ddot{g}(x) = \begin{cases} 1/8 & \text{if } x = 7, \\ 1/8 & \text{if } x = 9, \\ 1/4 & \text{if } x = 11, \\ 1/4 & \text{if } x = 13, \\ 1/8 & \text{if } x = 19, \\ 1/8 & \text{if } x = 21. \end{cases}$$

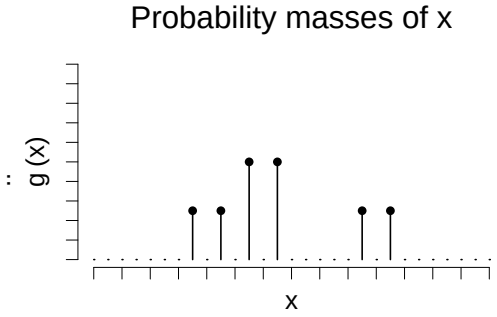


Figure G.0: Description of $\ddot{g}(x)$.

Define (Shannon) entropy of a random variable x (or rather of the corresponding probability distribution) as

$$\mathcal{E}(\ddot{g}(x)) := -\sum_{i=1}^{n_q} \ddot{g}(x_i) \log \ddot{g}(x_i), \tag{G.2}$$

where

$$n_q := |\text{supp}(\ddot{g})|. \tag{G.3}$$

Entropy is a measure of the average uncertainty in a random variable. Note the important feature of the entropy functional (and its relatives) in that it does not depend on the actual

values taken by a random variable, $\{x_i\}_1^{n_q}$, and is only a function of the probability masses, $\{\ddot{g}(x_i)\}_1^{n_q}$. In our case,

$$\mathcal{E}(\ddot{g}(x)) = -\left(4 \times \frac{1}{8}\log\frac{1}{8} + 2 \times \frac{1}{4}\log\frac{1}{4}\right) = 2.5 \text{ bits.} \tag{G.4}$$

Step 1: Simplification of discrete distribution.

Let

$$\hat{x} \sim \ddot{h}(\hat{x}) \tag{G.5}$$

be a simplified version of x, described in Figure G.1.

$$\ddot{h}(\hat{x}) = \begin{cases} 1/4 & \text{if } \hat{x} = 8, \\ 1/2 & \text{if } \hat{x} = 12, \\ 1/4 & \text{if } \hat{x} = 20. \end{cases}$$

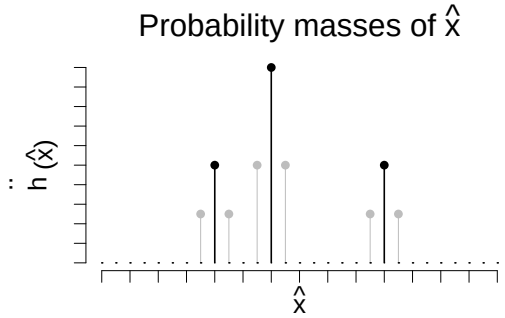


Figure G.1: Description of $\ddot{h}(\hat{x})$.

(Specific form of $\ddot{h}(\cdot)$, in particular its relation to $\ddot{g}(\cdot)$, is a matter of choice that is discussed later.)

Probability mass function $\ddot{h}(\hat{x})$ is coarser than $\ddot{g}(x)$, i.e. is characterized by lower entropy:

$$\begin{aligned}\mathcal{E}(\ddot{h}(\hat{x})) &= -\sum_{j=1}^{\hat{n}_q} \ddot{h}(\hat{x}_j)\log\ddot{h}(\hat{x}_j) = \\ &= -\left(2 \times \frac{1}{4}\log\frac{1}{4} + \frac{1}{2}\log\frac{1}{2}\right) = 1.5 \text{ bits,}\end{aligned} \tag{G.6}$$

where

$$\hat{n}_q := |\text{supp}(\ddot{h})|. \tag{G.7}$$

Step 2: Generating codebook.

Assume uniform discretization of the range of $\ddot{G}(x)$ into n_d quantiles, where the number of discretization points is defined by

$$n_d := \frac{1}{\delta}$$

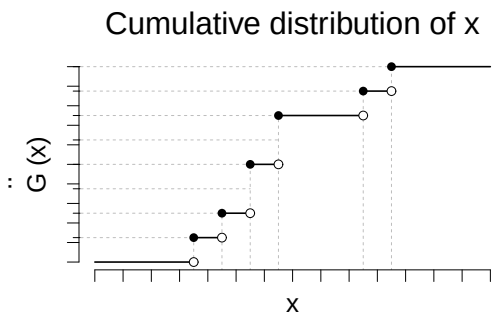
for cell size

$$\delta := \gcd(\{\ddot{g}(x_i)\}_1^{n_q}),$$

with $\gcd(a_1, \ldots, a_n)$ returning the greatest common divisor of $a_1, \ldots, a_n \in \mathbb{R}$, so that

$$\ddot{G}(x_i) \geq \ddot{G}(x_{i-1}) + \delta.$$

Left panel of Figure G.2.1 visualizes the process of discretization.



Codebook for $\ddot{G}(x)$

x	$\ddot{G}(x)$	$\ddot{g}(x)$	Native code	Optimal code
7	$^1/_8$	$^1/_8$	000111	000
9	$^2/_8$	$^1/_8$	001001	001
11	$^4/_8$	$^1/_4$	001011	10
13	$^6/_8$	$^1/_4$	001101	11
19	$^7/_8$	$^1/_8$	010011	010
21	$^8/_8$	$^1/_8$	010101	011
E[$L(x)$]	-	-	6	2.5

Figure G.2.1: Generating codebook for $\ddot{G}(x)$.

Right panel of Figure G.2.1 exhibits a codebook, a map from source alphabet to output alphabet, which in our case summarizes optimal code for probability distribution $\ddot{G}(x)$. We will use it for exchanging information that sources from this distribution (like a binary Morse code for encoding English letters and Arabic numerals, or genetic code in the DNA with four nucleotides for encoding different amino acids). It is constructed using the most basic algorithm for optimal coding, Huffman procedure.

Given some alphabet, expected length of any instantaneous (this term is defined later)

code for random variable x is bounded below by the entropy of x:

$$\mathrm{E}[L(x)] = \sum_{i=1}^{n_q} \ddot{g}(x_i) L(x_i) \geq \mathcal{E}(\ddot{g}(x)), \tag{G.8}$$

where $L(x_i)$ is the length of the codeword associated with x_i.

Optimal coding guarantees an expected codeword length within 1 bit of the lower bound, i.e. $\mathcal{E}(\ddot{g}(x)) \leq \mathrm{E}[L^*(x)] < \mathcal{E}(\ddot{g}(x)) + 1$, and expected length of Huffman code is at least as small as that of any other optimal code. In our simple example, Huffman code achieves the lower bound: $\mathrm{E}[L^*(x)] = \mathcal{E}(\ddot{g}(x)) = 2.5$ bits, c.f. equation (G.2). (Incidentally, Huffman coding here is equivalent to another optimal coding procedure, so called Shannon code, which assigns codeword lengths of $\lceil \log {}^{1}\!/_{\ddot{g}(x)} \rceil$ and overshoots the lower bound by less than 1 bit.)

Additonaly, Figure G.2.1 provides what we call "native" code for x. It is constructed using alphabet $\mathcal{A} := \{\texttt{000000}, \ldots, \texttt{111111}\}$, which is assumed to be the default code that the decision-making agent is endowed with at the outset and that is used for all operations (such as communication, computation, etc.). It is not optimized for probability distribution of x, or to be more accurate, assumes uniform distribution of a random variable over the domain $\{0, \ldots, 63\}$, and uses fixed 6-bit-long codewords. (This distinguishment is without loss of generality, and native code may coincide with optimal code.)

Similarly, assume uniform discretization of the range of $\ddot{H}(\hat{x})$ into $\hat{n}_d$ quantiles, where

$$\hat{n}_d := \frac{1}{\hat{\delta}}$$

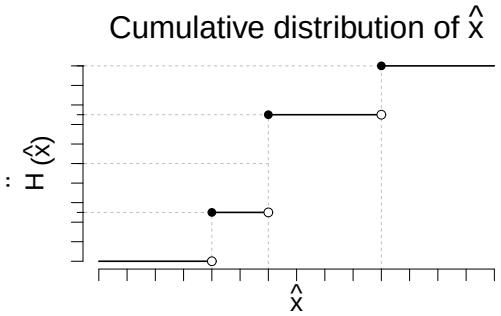
for cell size

$$\hat{\delta} := \gcd(\{\ddot{h}(\hat{x}_j)\}_1^{\hat{n}_q}),$$

so that

$$\ddot{H}(\hat{x}_j) \geq \ddot{H}(\hat{x}_{j-1}) + \hat{\delta}.$$

Figure G.2.2 presents discretization and coding results for probabiltiy distribution $\ddot{H}(\hat{x})$.



Codebook for $\ddot{H}(\hat{x})$

$\hat{x}$	$\ddot{H}(\hat{x})$	$\ddot{h}(\hat{x})$	Native code	Optimal code
8	$1/4$	$1/4$	001000	00
12	$3/4$	$1/2$	001100	1
20	$4/4$	$1/4$	010100	01
$\mathrm{E}[L(\hat{x})]$	-	-	6	1.5

Figure G.2.2: Generating codebook for $\ddot{H}(\hat{x})$.

Again, the optimal code reaches lower bound $\mathrm{E}[L^*(\hat{x})] = \mathcal{E}(\ddot{h}(\hat{x})) = 1.5$ bits, as in equation (G.6), using 1 to 2 bits per codeword; while the native code still uses 6 bits as before.

In general, native codebook differs from optimal codebook, which leads to certain code redundancy. Such redundancy, or what we call code overhead, can take one of two different forms (leading the discussion below in terms of $\ddot{h}(\hat{x})$, which in this connection is more relevant):

(i) D-overhead is an artefact of divergence of native codebook from optimal one and arises when using the former instead of the latter. This requires (applying "Wrong code" theorem in Cover and Thomas, 2006):

$$\begin{aligned}\text{D-overhead} = \hat{n}_d \times \mathcal{D}\left(\ddot{h}(\hat{x}) \,\middle\|\, \frac{1}{|\mathcal{A}|}\right) = \hat{n}_d \times \sum_{j=1}^{\hat{n}_q} \ddot{h}(\hat{x}_j) \log \frac{\ddot{h}(\hat{x}_j)}{1/|\mathcal{A}|} = \\ = \hat{n}_d \times \left(\log|\mathcal{A}| - \mathcal{E}(\ddot{h}(\hat{x}))\right), \qquad \text{(G.9)}\end{aligned}$$

which is bounded by

$$\text{D-overhead} \in [0, \hat{n}_d \log|\mathcal{A}|] \qquad \text{(G.10)}$$

(for $\mathrm{supp}(\ddot{h}(\hat{x})) \subseteq \mathcal{A}$, and since uniform distribution is a maximum entropy distribution for a given finite support set).

Here, relative entropy (or Kullback–Leibler divergence) $\mathcal{D}(\ddot{\pi}_1 \| \ddot{\pi}_2)$ for some probability

mass functions $\ddot{\pi}_1$ and $\ddot{\pi}_2$ is defined as follows:

$$\mathcal{D}\left(\ddot{\pi}_1(\chi)\|\ddot{\pi}_2(\chi)\right) := \sum_{i=1}^{|\mathrm{supp}(\ddot{\pi}_1)|} \ddot{\pi}_1(\chi_i) \log \frac{\ddot{\pi}_1(\chi_i)}{\ddot{\pi}_2(\chi_i)}. \tag{G.11}$$

In our example D-overhead amounts to

$$\text{D-overhead} = \hat{n}_d \times \left(\log|\mathcal{A}| - \mathcal{E}(\ddot{h}(\hat{x}))\right) = 4 \times (6 - 1.5) = 18 \text{ bits}. \tag{G.12}$$

(ii) P-overhead comprises the preamble that provides the description of the optimal codebook. In our case it amounts to:

$$\begin{aligned}\text{P-overhead} = \text{“ } &\texttt{001000 000010 00 001100 000001 1} \\ &\underbrace{\texttt{010100}}_{\substack{\text{value}}} \underbrace{\texttt{000010}}_{\substack{\text{length of}\\ \text{codeword}}} \underbrace{\texttt{01}}_{\text{codeword}} \underbrace{\texttt{111111}}_{\substack{\text{EOM}\\ \text{symbol}}}\text{”} = \\ &= 3 \times (6+6) + 5 + 6 = 47 \text{ bits}, \end{aligned} \tag{G.13}$$

and in general (allowing for maximum length of a codeword to be 64 bits) this requires:

$$\begin{aligned}\text{P-overhead} &= |\mathrm{supp}(\ddot{h})| \times (\log|\mathcal{A}| + \log|\mathcal{A}|) + \sum_{j=1}^{|\mathrm{supp}(\ddot{h})|} L^*(\hat{x}_j) + \log|\mathcal{A}| \geq \\ &\geq |\mathrm{supp}(\ddot{h})| \times 2\log|\mathcal{A}| - \sum_{j=1}^{|\mathrm{supp}(\ddot{h})|} \log \ddot{h}(\hat{x}_j) + \log|\mathcal{A}| \geq \\ &\geq |\mathrm{supp}(\ddot{h})| \times 2\log|\mathcal{A}| + |\mathrm{supp}(\ddot{h})| \times \log|\mathrm{supp}(\ddot{h})| + \log|\mathcal{A}| = \\ &= \hat{n}_q \times \left(2\log|\mathcal{A}| + \log\hat{n}_q\right) + \log|\mathcal{A}|, \end{aligned} \tag{G.14}$$

with equality signs for $\ddot{h}(\hat{x})$ that is both dyadic and uniform (here the first inequality is due to bound on the optimal code length). However, P-overhead is unbounded from above even for fixed $\hat{n}_q$ as optimal $L^*(\hat{x}_j) \to \infty$ for $\hat{x}_j$ such that $\ddot{h}(\hat{x}_j) \to 0$ (in which

case D-overhead also blows up owing to $\hat{n}_d \to \infty$).

The choice of smallest overhead is straight-forward:

$$\begin{aligned} \text{overhead}^* &:= \min\{\text{D-overhead}, \text{P-overhead}\} = \\ &= \min\{18, 47\} = 18 \text{ bits.} \end{aligned} \tag{G.15}$$

It is clear that for finite alphabet $\mathcal{A}$, D-overhead $\in \mathcal{O}(\hat{n}_d)$, while P-overhead is at best an $\mathcal{O}(\hat{n}_q \log \hat{n}_q)$ expression. Given that $\hat{n}_q \leq \hat{n}_d$, asymptotically P-overhead may become more attractive, but only in cases when resolution of discretization is much higher than resolution of quantization (more on which later).

In information theory arguments it is often assumed for reasons of simplicity that the alphabet used for coding is the optimal one and achieves the lower bound of expected code length. Within our framework, this situation can be understood as asymptotics in terms of number of entire algorithm's uses while holding other parameters such as $|\text{supp}(\ddot{h})|$, $|\mathcal{A}|$, $\max_i L^*(\hat{x}_i)$ finite, i.e. as effectively changing the alphabet used by paying fixed but asymptotically negligible costs of the necessary P-overhead. However, asymptotic point of view is sensible only when dealing with stationary probability distributions, hence it would loose the generality with respect to potential applications.

Step 3: Description of simplified distribution.

After codebook generation step, the next three steps are best illustrated by the flowchart on Figure G.3.

Description of the simplified distribution is conducted by sending values of its domain at each discretization interval via communication channel. But first, consider the task of sending via communication channel just a single realization of random variable.

In information theory, communication channels are characterized by their channel ca-

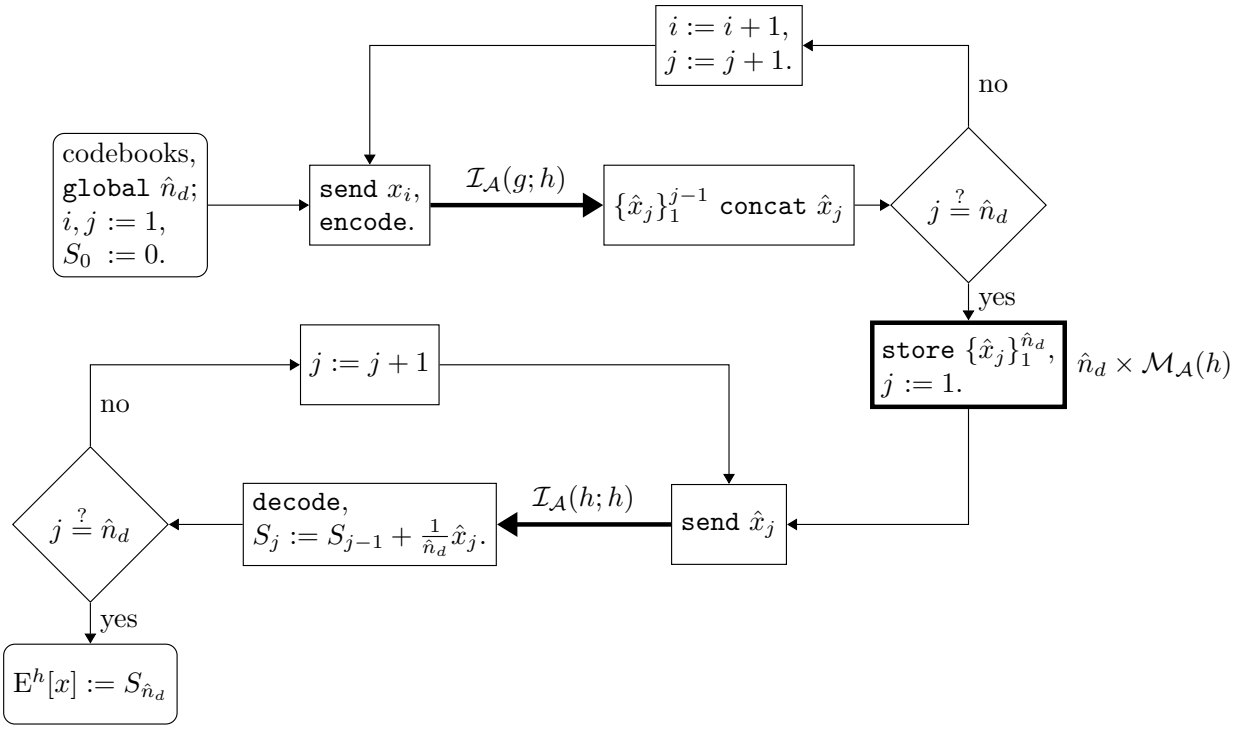


Figure G.3: Flowchart of description, storage and computation steps (scalar random variable case, computation of the mean).

pacity, which is formally defined as the maximum mutual information between input and output random variables:

$$\max_{\ddot{g}(x)} \left\{ \mathcal{I}(\ddot{g}(x); \ddot{h}(\hat{x})) \right\}. \tag{G.16}$$

In turn, mutual information between random variables x and $\hat{x}$ is defined as the relative entropy between their joint probability mass function $\ddot{f}(x, \hat{x})$ and the product of their marginal probability mass functions $\ddot{g}(x)\ddot{h}(\hat{x})$, which makes it

$$\mathcal{I}(\ddot{g}(x); \ddot{h}(\hat{x})) := \mathcal{D}\left(\ddot{f}(x, \hat{x}) \| \ddot{g}(x)\ddot{h}(\hat{x})\right) = \sum_{i=1}^{n_q} \sum_{j=1}^{\hat{n}_q} \ddot{f}(x_i, \hat{x}_j) \log \left(\frac{\ddot{f}(x_i, \hat{x}_j)}{\ddot{g}(x_i)\ddot{h}(\hat{x}_j)} \right). \tag{G.17}$$

It can be shown that

$$\begin{aligned} \mathcal{I}(\ddot{g}(x); \ddot{h}(\hat{x})) &= \mathcal{E}(\ddot{g}(x)) + \mathcal{E}(\ddot{h}(\hat{x})) - \mathcal{E}(\ddot{g}(x), \ddot{h}(\hat{x})) = \\ &= \mathcal{E}(\ddot{g}(x)) - \mathcal{E}(\ddot{f}(x|\hat{x})) = \mathcal{E}(\ddot{h}(\hat{x})) - \mathcal{E}(\ddot{f}(\hat{x}|x)), \end{aligned} \tag{G.18}$$

thus mutual information measures expected reduction in uncertainty about a random variable after observing realization of another random variable.

Operationally, channel capacity is defined as the highest rate (in bits per channel use, or per unit of time) at which information can be sent with arbitrarily low error probability. (To be pedantic, bandwidth is one of the factors determining channel capacity.) That is, channel capacity imposes a constraint on the complexity of messages that can be sent via it without error. Message is a codeword representing realization of a random variable. Complexity of a message is measured by its entropy, i.e. by entropy of the underlying random variable, because one of the crucial features of the entropy functional is that it does not depend on the actual values taken by a random variable, but only on the corresponding probability masses. Therefore, channel capacity limits the admissible entropy of the output, and whenever the entropy of the desired input is higher than available capacity, sending information without error requires an adjustment to the input. This leads us to using input from the simplified probability distribution $\ddot{h}(\hat{x})$ rather than from the original probability distribution $\ddot{g}(x)$. Note that for transparency reasons, we choose to deal here with bearing losses at the time of encoding but with a lossless communication itself, which implies $\mathcal{E}(\ddot{f}(\hat{x}|x)) = 0$ in (G.18).

Proceeding further, we now consider usage of the communication channel to transmit the description of the simplified distribution by sending values of its domain, $\hat{x}$, at each discretization interval. The construction that follows will be based on Proposition G.1.

Proposition G.1 (Description of Probability Distribution)**.** *Let χ be random variable distributed according to probability mass function $\ddot{\pi}(\chi)$. Assume discretization of the range of the corresponding cumulative distribution function $\ddot{\Pi}(\chi)$, assume uniform discretization, assume using instantaneous code and assume receiving unordered (i.e., randomly ordered) codewords. Then optimal description of the probability distribution $\ddot{\Pi}(\chi)$ takes*

$$\frac{1}{\delta} \times \mathcal{E}(\ddot{\pi}(\chi)) \text{ bits},$$

where

$$\delta := \gcd\left(\{\ddot{\pi}(\chi_i)\}_1^{|\mathrm{supp}(\ddot{\pi})|}\right).$$

Proof. See §G.3.

□

Basically, this Proposition establishes a useful correspondence between the entropy of random variable χ (really, the entropy of probability mass function $\ddot{\pi}(\chi)$) and the length of the description of its probability distribution $\ddot{\Pi}(\chi)$.

The assumption about uniform discretization can be motivated by, first, economizing on the need to also describe the discretization rule (i.e., the method of allocating discretization cell widths) itself and thus confining to rules that are fixed or that can be readily deduced from the very structure of a description sequence, and, second, by uniform prior belief on the locations of probability masses over the domain or more generally by principle of insufficient reason.

An instantaneous (or prefix) code is defined as a code system such that no codeword in it is a prefix for any other codeword in the system.

An alternative approach to description of probability distributions is taken in quantization literature (an encyclopedic overview of this literature is available in Gray and Neuhoff, 1998). State of the art procedure is to construct a finite set of reproducible probability distributions, find the element of this set that is closest in some metric to the desired distribution, and to pass on just the unique index of this closest set element. This modern type-based scheme was developed by Reznik (2010), who also proposed an asymptotically optimal algorithm of its implementation. Expanding the set of fully reproducible distributions—by increasing the density of reproduction lattice (grid)—allows for more accurate description of the distribution in focus. It is worth noting, however, that procedures implementing a classic tree-based scheme still dominate the practice and are used e.g. for data compression in ZIP and JPEG standards. In contrast to this quantization-motivated approach, which boils

down to describing the range of the probability distribution in target, our approach reduces to description of its domain. This is done for the sake of simplicity, so that formalization of the description step conformed with that of the following two steps as much as possible.

Lastly, recall that our decision-making agent is endowed with a native code based on alphabet $\mathcal{A}$, which is used for all operations including transmission via the communication channel of the simplified probability distribution's desription sequence. A channel that transmits information from input random variable x into output random variable $\hat{x}$ using alphabet $\mathcal{A}$ will be denoted by $\mathcal{I}_{\mathcal{A}}(\ddot{g}(x); \ddot{h}(\hat{x}))$.

Therefore, in our example we have:

- Effective channel capacity demands per codeword [in information-theoretic terms; or equivalently in engineering terms, per one channel activation] equal

$$\begin{aligned} \mathcal{I}(\ddot{g}(x); \ddot{h}(\hat{x})) = \mathcal{E}(\ddot{h}(\hat{x})) - \mathcal{E}(\ddot{f}(\hat{x}|x)) = \\ = \mathcal{E}(\ddot{h}(\hat{x})) - 0 = 1.5 \text{ bits}; \end{aligned} \tag{G.19}$$

- Physical channel capacity demands per codeword equal

$$\begin{aligned} \mathcal{I}_{\mathcal{A}}(\ddot{g}(x); \ddot{h}(\hat{x})) = \mathcal{E}_{\mathcal{A}}(\ddot{h}(\hat{x})) - \mathcal{E}_{\mathcal{A}}(\ddot{f}(\hat{x}|x)) = \\ = \mathcal{E}(\ddot{h}(\hat{x})) + \mathcal{D}\left(\ddot{h}(\hat{x}) \middle\| \frac{1}{|\mathcal{A}|}\right) - 0 = 1.5 + 4.5 = 6 \text{ bits} \end{aligned} \tag{G.20}$$

  (i.e., the length of native code's codeword).

Above, we again used the relative entropy property previously applied by the "Wrong code" theorem:

$$\mathcal{E}_{\mathcal{A}}(\ddot{h}(\hat{x})) = \mathcal{E}(\ddot{h}(\hat{x})) + \mathcal{D}\left(\ddot{h}(\hat{x}) \middle\| \frac{1}{|\mathcal{A}|}\right). \tag{G.21}$$

Sending every domain value corresponding to each discretization cell takes $\hat{n}_d$ codewords, or channel transmission operations. Aggregating, we get channel capacity demands per full

code sequence (per full transmission), i.e. $\hat{n}_d \times$(channel capacity per codeword) [equivalently, $\hat{n}_d \times$ (channel capacity per activation)]:

- Effective channel capacity demands per full code sequence [per full transmission] equal

$$\hat{n}_d \times \mathcal{I}(\ddot{g}(x); \ddot{h}(\hat{x})) = 4 \times 1.5 = 6 \text{ bits;} \tag{G.22}$$

- Physical channel capacity demands per full code sequence equal

$$\hat{n}_d \times \mathcal{I}_{\mathcal{A}}(\ddot{g}(x); \ddot{h}(\hat{x})) = 4 \times 6 = 24 \text{ bits.} \tag{G.23}$$

Thus, our desired input "`000 001 10 10 11 11 010 011`" would take

$$n_d \times \mathcal{E}(\ddot{g}(x)) = 8 \times 2.5 = 20 \text{ bits,} \tag{G.24}$$

our net input "`00 1 1 01`" takes only

$$\hat{n}_d \times \mathcal{E}(\ddot{h}(\hat{x})) = 4 \times 1.5 = 6 \text{ bits,} \tag{G.25}$$

the sent message "`001000 001100 001100 010100`" takes

$$\hat{n}_d \times \mathcal{I}(\ddot{g}(x); \ddot{h}(\hat{x})) + \text{overhead} = 4 \times 1.5 + 4 \times 4.5 = 6 + 18 = 24 \text{ bits,} \tag{G.26}$$

and net output is the same as net input in (G.25), also taking 6 bits, due to the fact that communication not exceeding channel capacity is lossless.

To sum up, it takes 24 bits to fully describe our simplified probability distribution $\ddot{h}(\hat{x})$. (Alternative state of the art procedure due to Reznik (2010) would use 20 bits instead.)

Step 4: Storage in working memory.

Next, the description of the simplified distribution is loaded into "working memory". Memory storage is a resource that will be used for performing computations in Step 5.

In our example, distinguishing between operations in optimal and native code, we have:

- Effective working memory capacity demands per codeword [per one memory writing operation]

$$\mathcal{M}(\ddot{h}(\hat{x})) = \mathrm{E}[L^*(\hat{x})] = \mathcal{E}(\ddot{h}(\hat{x})) = 1.5 \text{ bits;} \tag{G.27}$$

- Physical working memory capacity demands per codeword

$$\mathcal{M}_{\mathcal{A}}(\ddot{h}(\hat{x})) = L_{\mathcal{A}}(\hat{x}) = \mathcal{E}_{\mathcal{A}}(\ddot{h}(\hat{x})) = 6 \text{ bits.} \tag{G.28}$$

On aggregate, this makes:

- Effective working memory capacity demands per full code sequence [per full storage session]:

$$\hat{n}_d \times \mathcal{M}(\ddot{h}(\hat{x})) = 4 \times 1.5 = 6 \text{ bits;} \tag{G.29}$$

- Physical working memory capacity demands per full code sequence:

$$\hat{n}_d \times \mathcal{M}_{\mathcal{A}}(\ddot{h}(\hat{x})) = 4 \times 6 = 24 \text{ bits.} \tag{G.30}$$

Step 5: Computation of statistic.

Taking mean of $\hat{x}$ as a statistic of interest, it is computed as a partial sum of $\hat{x}_j$, each weighted by factor $1/\hat{n}_d$, for $j = 1, 2, \ldots, \hat{n}_d$. In order to conduct $\hat{n}_d$ of these summation operations, every time an $\hat{x}_j$ has to be retrieved from the working memory (costlessly, but this is WLOG) and sent via communication channel to the "arithmetic unit" (whose operation

costs we abstract away from for simplicity reasons). Thus, the computation process utilizes the communication channel similarly to description process.

Effective channel capacity demands per codeword $\mathcal{I}(\ddot{h}(\hat{x}); \ddot{h}(\hat{x}))$ equal 1.5 bits, while physical channel capacity demands per codeword $\mathcal{I}_{\mathcal{A}}(\ddot{h}(\hat{x}); \ddot{h}(\hat{x}))$ equal 6 bits (note the replacement of $\ddot{g}(x)$ with $\ddot{h}(\hat{x})$ in the mutual information functionals' arguments). Effective channel capacity demands per full code sequence $\hat{n}_d \times \mathcal{I}(\ddot{h}(\hat{x}); \ddot{h}(\hat{x}))$ equal 6 bits, and physical channel capacity demands per full code sequence $\hat{n}_d \times \mathcal{I}_{\mathcal{A}}(\ddot{h}(\hat{x}); \ddot{h}(\hat{x}))$ equal 24 bits. Our net input are the same 6 bits as in (G.25), the sent message repeats 24 bits from (G.26), and net output equals net input.

Digression: In the interest of generality we should emphasize that the case when x is a continuously distributed random variable does not pose any practical difficulties (although some care must be taken nevertheless, as continuous entropy is not invariant to transformations such as change of variables). Previous formulas can be readily converted into their continuous counterpats, so in principle we can treat the presented algorithm as applicable to both cases. However, for our specific coding examples to remain valid (and/or basing on findings of neurophysiological nature), we may wish to convert a continous random variable into its discrete analog as a preliminary step.

Step 0: Quantization of continuous distribution.

Let x be a continuously distributed scalar random variable with probabiliy density function $\bar{g}(x)$,

$$x \sim \bar{g}(x), \tag{G.31}$$

as depicted in the left panel of Figure G.0*.

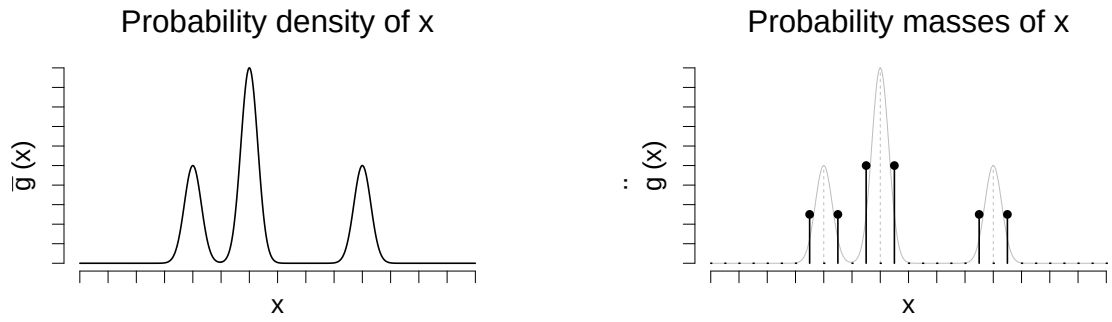


Figure G.0*: Depiction of $\bar{g}(x)$ and $\ddot{g}(x)$.

Define differential (or continuous) entropy of random variable x as

$$\mathcal{E}(\bar{g}(x)) := -\int_{\mathrm{supp}(\bar{g})} \bar{g}(x) \log \bar{g}(x)\, dx, \tag{G.32}$$

(in this case, a continuous alphabet is presumed).

Assume uniform quantization of $\mathrm{supp}(\bar{g})$ into identical cells of width Δ not optimized for particular distribution. This is optimal under fairly general conditions for high resolution quantization (i.e., "small" Δ and "large" n_q) case. As a result, we obtain quantized probability density function $\ddot{g}(x)$, defined by a sequence of equations

$$\ddot{g}(x_i) := \int_{\Delta i}^{\Delta(i+1)} \bar{g}(x)\, dx = \bar{g}(x_i)\Delta \tag{G.33}$$

for $i = 1, 2, \ldots, n_q$. Here, n_q is the number of quantization points, determined by covering (tiling) the set $\mathrm{supp}(\bar{g}(x))$ with Δ-cells:

$$n_q := \frac{m(\mathrm{supp}(\bar{g}))}{\Delta} \tag{G.34}$$

for some appropriate measure m on σ-algebra over $\mathrm{supp}(\bar{g})$. This also gives the relationship

$$|\mathrm{supp}(\ddot{g})| := n_q, \tag{G.35}$$

which we have already used earlier.

The process of quantization is visualized by the right panel of Figure G.0*. Using cell width $\Delta := 2$, we thus produced the probability mass function $\ddot{g}(x)$ from our previously discussed example.

For Riemann-integrable $\bar{g}(x)$, the following correspondence between continuous and discrete entropy holds as $\Delta \to 0$ (Cover and Thomas, 2006):

$$\mathcal{E}(\ddot{g}(x)) \simeq \mathcal{E}(\bar{g}(x)) + |\text{supp}(\ddot{g})| := \mathcal{E}(\bar{g}(x)) + n_q, \tag{G.36}$$

which in our example boils down to

$$2.5 \text{ bits} \simeq \mathcal{E}(\bar{g}(x)) + 6 \text{ bits}. \tag{G.37}$$

Wrapping-up remarks: To recap, the information processing algorithm consists of the following steps:

[**0.** Quantization of continuous distribution.]

1. Simplification of discrete distribution.

2. Generating codebook.

3. Description of simplified distribution.

4. Storage in working memory.

5. Computation of statistic.

The last three steps contain potential "bottlenecks" on the way of information flow (shown in bold on Figure G.3), formally represented by the following, potentially binding, physical constraints:

- for the Description step, communication channel capacity demands $\hat{n}_d \times \mathcal{I}_{\mathcal{A}}(\ddot{g}(x); \ddot{h}(\hat{x}))$ are bounded by the available full description channel capacity $\mathcal{K}_D$,

$$\hat{n}_d \times \mathcal{I}_{\mathcal{A}}(\ddot{g}(x); \ddot{h}(\hat{x})) \leq n_{\mathcal{I}D} \times \widehat{\mathcal{I}}_{\mathcal{A}D} =: \mathcal{K}_D, \tag{G.38}$$

  where available channel capacity is formed by $n_{\mathcal{I}D}$ number of $\widehat{\mathcal{I}}_{\mathcal{A}D}$-bit wide physical communication channels;

- for the Storage step, working memory capacity demands $\hat{n}_d \times \mathcal{M}_{\mathcal{A}}(\ddot{h}(\hat{x}))$ are bounded by the available full storage memory capacity $\mathcal{K}_S$,

$$\hat{n}_d \times \mathcal{M}_{\mathcal{A}}(\ddot{h}(\hat{x})) \leq n_{\mathcal{M}} \times \widehat{\mathcal{M}}_{\mathcal{A}} =: \mathcal{K}_S, \tag{G.39}$$

  where available memory capacity is formed by $n_{\mathcal{M}}$ number of $\widehat{\mathcal{M}}_{\mathcal{A}}$-bit large physical working memory cells;

- for the Computation step, communication channel capacity demands $\hat{n}_d \times \mathcal{I}_{\mathcal{A}}(\ddot{h}(\hat{x}); \ddot{h}(\hat{x}))$ are bounded by the available full computation channel capacity $\mathcal{K}_C$,

$$\hat{n}_d \times \mathcal{I}_{\mathcal{A}}(\ddot{h}(\hat{x}); \ddot{h}(\hat{x})) \leq n_{\mathcal{I}C} \times \widehat{\mathcal{I}}_{\mathcal{A}C} =: \mathcal{K}_C, \tag{G.40}$$

  where available channel capacity is formed by $n_{\mathcal{I}C}$ number of $\widehat{\mathcal{I}}_{\mathcal{A}C}$-bit wide physical communication channels.

Taking a wider, computational complexity perspective, the above three steps can be viewed as particular examples of, respectively, communication, space and time complexity concepts (see Arora and Barak, 2009).

Utilized in the manner of the described algorithm, binding constraints on description channel capacity, on working memory capacity, or on computation channel capacity demands are operationally equivalent to each other.

Define $\mathcal{K}^*$ as the full physical capacity bound implied by the tightest constraint:

$$\mathcal{K}^* := \min\{\mathcal{K}_D,\ \mathcal{K}_S,\ \mathcal{K}_C\}. \tag{G.41}$$

(In our coding example, $\mathcal{K}^* = 24$ bits.)

Conditional on the value of $\mathcal{K}^*$, we can assume without loss of generality any one of the three steps is the physical "bottleneck" at play (essentially, this is a matter of taste).

Lastly, the same results can be obtained by replacing lossless description of the simplified distribution (Step 3) with Monte Carlo sampling from it [or with its numerical quadrature approximation]. Specifically, instead of computing the exact statistic for the simplified distribution described fully, we can compute its approximation using the simplified distribution's empirical counterpart [or using weighted qudrature nodes]. Rather than taking $\{\hat{x}_j\}_1^{\hat{n}_d}$, we can take instead a sample of $\hat{x}_j \sim \ddot{h}(\hat{x}_j)$, for $j = 1, 2, \ldots, \hat{n}_{mc}$ [or nodes $\{\hat{x}_j\}_1^{\hat{n}_{nq}}$]. The only material difference is that $\hat{n}_{mc}$ (in general, $\hat{n}_{mc} \neq \hat{n}_d$) [$\hat{n}_{nq}$, in general $\hat{n}_{nq} \neq \hat{n}_d$] would now reflect the size of Monte Carlo sample [number of quadrature nodes].

Overall, additionally taking Monte Carlo sampling [numerical quadrature approximation] as a possible replacement for description of the simplified distribution step, along with storage and computation steps, there are 5 alternative mechanisms that generate the same communication-based calculus, modulo change of interpretation and possibly redefinition of $\hat{n}_d$ parameter.

As a remark on formal notation, from a computational standpoint the presented algorithm constitutes an ancillary procedure $\mathcal{P}_{\int}$ for evaluation of the expectation operator $\mathrm{E}^h[x]$.

G.2 Vector random variable case

We are still not ready to apply the approach presented above even to the most basic task of choice between two simple lotteries, as in Figure 2.1. For instance, recall full physical

capacity bound $\mathcal{K}^*$ defined in equation (G.41). It is important to realize that even allocating $2\mathcal{K}^*$ for processing 2 lotteries does not mean just repeating the previous algorithm 2 times.

To begin with, we need to account for the well-known result that the problem of encoding several—even independent—random variables simultaneously is not equivalent to a combination of the corresponding stand-alone problems (this is evident from the so called sphere-covering argument used in geometry, see Cover and Thomas, 2006).

Also, handling functions of two or more random variables requires some care. For $K \times K$ matrix $\boldsymbol{A}$, $\mathcal{E}(\pi(\boldsymbol{A}\boldsymbol{\chi}))$ works smoothly, but there are difficulties with non-square $\boldsymbol{A}$, i.e. with entropy of a convolution of random variables: even $\mathcal{E}(\ddot{\pi}(\sum_k \chi_k))$ is cumbersome (cumbersome enough to warrant works by Tao, 2010; Tao and Vu, 2006). Going further, for bijective function $\varphi : \mathbb{R}^K \mapsto \mathbb{R}^K$, $\mathcal{E}(\pi(\varphi(\boldsymbol{\chi})))$ works relatively straight-forwardly, but result is not available in general for $\varphi : \mathbb{R}^K \mapsto \mathbb{R}^{K'}, K > K'$, i.e. for entropy of "non-linear convolution".

Therefore, we will proceed as follows. Firstly, we use the fact that $\mathcal{E}(\cdot)$ and $\mathcal{I}(\cdot\,;\cdot)$ functionals conveniently generalize to K-dimensional random vectors $\boldsymbol{\chi}$ in the desired manner.

Also, we apply transformations defined by $\mathbb{R}^K \mapsto \mathbb{R}^{K'}$ functions (including those defined by $K' \times K$ matrices) after rather than before any communication and storage processes, i.e. in the Computation rather than Description step.

Lastly, optimization (i.e., maximizing $\varphi(\boldsymbol{\chi})$ or equating it to 0) is executed as several repetitions of the Computation step iterating on function $\varphi(\cdot|\boldsymbol{\theta})$ for different choices of parameter vector $\boldsymbol{\theta}$ (e.g., number of iterations $n_\iota = 2$ for binary choice from our motivating example, $n_\iota = n$ for weighted sums on an n-points grid).

The generalized algorithm is presented in the flowchart on Figure G.4. It still deals with a simple lottery, but now the lottery is not scalar- but vector-valued, and the computed statistic is not just the average but any parameterized function, whose parameters are allowed to vary in the course of an iterative optimization procedure.

Notice, the optimization process here is constrained by available capacity, but such a constraint is by construction invariant to specific choices of parameter $\boldsymbol{\theta}$ or specification of

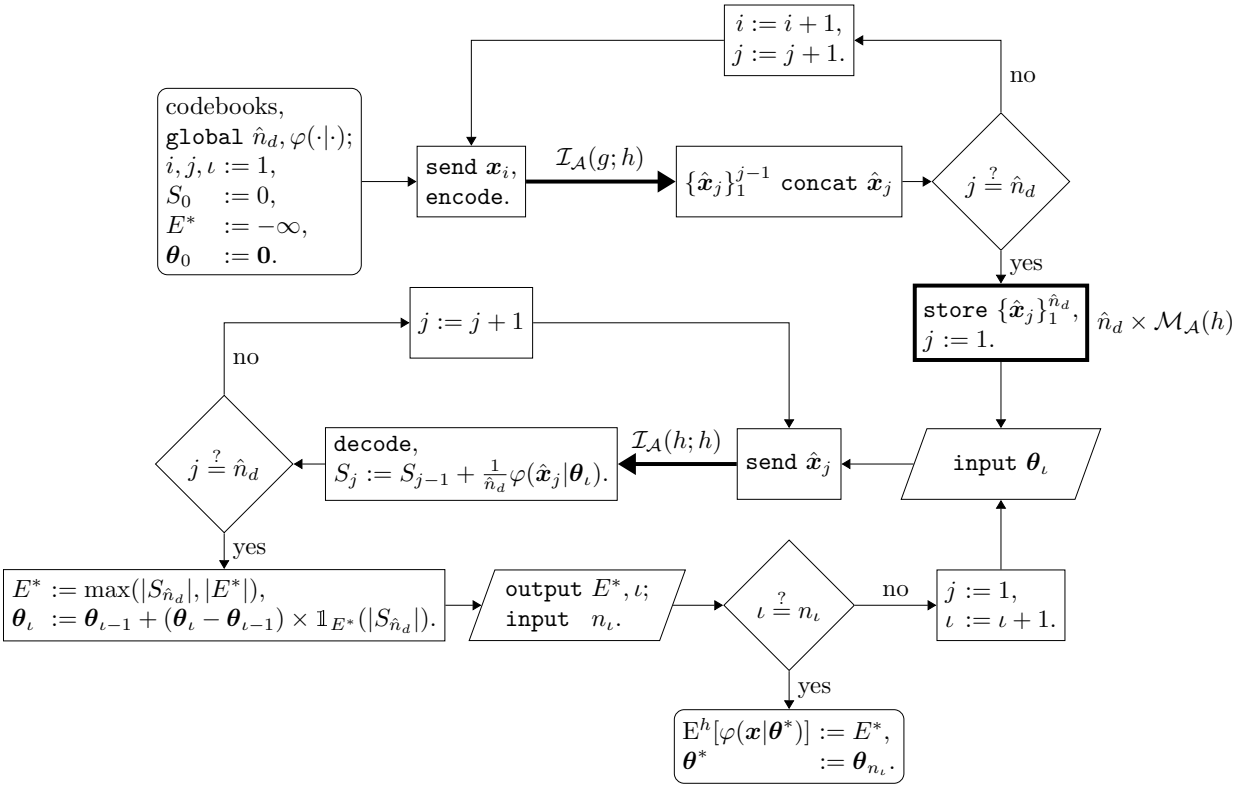


Figure G.4: Flowchart of description, storage and computation steps (vector random variable case, computation of some predetermined statistic, allowing for iterations on the latter).

$\varphi(\cdot|\cdot)$. This will allow us to segregate the actual decision-making problem from background problem of the optimal utilization of available capacity, and to solve them separately.

Finally, note that from a computational standpoint, now the algorithm solves a given optimization problem using the presented above (ancillary) expectation operator evaluation procedure $\mathcal{P}_{\int}$ at each iteration of the optimization procedure.

Capacity accounting: Next, let us carefully consider the issue of communication channel (or storage memory) capacity accounting.

Redefine the full physical capacity bound implied by the tightest (per iteration, in case of computation step) constraint $\mathcal{K}^*$ more generally as:

$$\mathcal{K}^* := \min\{\mathcal{K}_D,\ \mathcal{K}_S,\ \mathcal{K}_C \,/\, n_\iota\}. \tag{G.42}$$

(For instance, if $n_\iota = 2$, to be able to use in the full procedure 6 bits per codeword for each of 4 discretization cells as in our coding example requires: $\mathcal{K}_D \geq 24$, $\mathcal{K}_S \geq 24$, and $\mathcal{K}_C \geq 48$.)

Without loss of generality, assume the binding constraint is the computation step:

$$\mathcal{K}^* = \mathcal{K}_C \,/\, n_\iota, \tag{G.43}$$

while demands for full per iteration computation channel capacity yield (in the string of equalities below, we are moving backwards along the algorithm's path):

$$\begin{aligned}
\mathcal{K}_C \,/\, n_\iota = \mathcal{K}^* = \\
&= \hat{n}_d \times \mathcal{I}_{\mathcal{A}}(\ddot{h}(\hat{\boldsymbol{x}}); \ddot{h}(\hat{\boldsymbol{x}})) = \\
&= \hat{n}_d \times \mathcal{M}_{\mathcal{A}}(\ddot{h}(\hat{\boldsymbol{x}})) = \\
&= \hat{n}_d \times \mathcal{I}_{\mathcal{A}}(\ddot{g}(\boldsymbol{x}); \ddot{h}(\hat{\boldsymbol{x}})) = \\
&= \hat{n}_d \times \mathcal{I}(\ddot{g}(\boldsymbol{x}); \ddot{h}(\hat{\boldsymbol{x}})) + \text{overhead}.
\end{aligned} \tag{G.44}$$

Rearranging,

$$\mathcal{I}(\ddot{g}(\boldsymbol{x}); \ddot{h}(\hat{\boldsymbol{x}})) = \frac{\mathcal{K}_C}{\hat{n}_d \times n_\iota} - \frac{\text{overhead}}{\hat{n}_d} =: \kappa, \tag{G.45}$$

where κ denotes effective capacity bound per codeword (binding, as we assumed above).

Clearly, effective capacity κ is not equivalent to available full computation channel capacity $\mathcal{K}_C$, or more generally to available full physical capacity $\mathcal{K}^*$. It might, say, fall not just because of reduction in computation channel capacity, but also when using more discretization points and more iterations, as well as due to larger per-codeword overhead. This demonstrates why effective capacity κ, measured as implied information processing capacity constraint's bound, differs substantially from full physical capacity $\mathcal{K}^*$ (see the main text for further elaboration).

In general, effective capacity κ bounds the tightest constraint from above (also expanding the mutual information functional below):

$$\begin{aligned}\kappa \geq \mathcal{I}(\ddot{g}(\boldsymbol{x}); \ddot{h}(\hat{\boldsymbol{x}})) = \mathcal{E}(\ddot{g}(\boldsymbol{x})) + \mathcal{E}(\ddot{h}(\hat{\boldsymbol{x}})) - \mathcal{E}(\ddot{g}(\boldsymbol{x}), \ddot{h}(\hat{\boldsymbol{x}})) =: \mathcal{E}(\ddot{g}(\boldsymbol{x})) + \mathcal{E}(\ddot{h}(\hat{\boldsymbol{x}})) - \mathcal{E}(\ddot{f}(\boldsymbol{x}, \hat{\boldsymbol{x}})) = \\ = -\sum_{i=1}^{n_q} \ddot{g}(\boldsymbol{x}_i) \log \ddot{g}(\boldsymbol{x}_i) - \sum_{j=1}^{\hat{n}_q} \ddot{h}(\hat{\boldsymbol{x}}_j) \log \ddot{h}(\hat{\boldsymbol{x}}_j) + \sum_{i=1}^{n_q} \sum_{j=1}^{\hat{n}_q} \ddot{f}(\boldsymbol{x}_i, \hat{\boldsymbol{x}}_j) \log \ddot{f}(\boldsymbol{x}_i, \hat{\boldsymbol{x}}_j),\end{aligned} \tag{G.46}$$

with its continuous counterpart being

$$\begin{aligned}\kappa \geq \mathcal{I}(\bar{g}(\boldsymbol{x}); \bar{h}(\hat{\boldsymbol{x}})) = \mathcal{E}(\bar{g}(\boldsymbol{x})) + \mathcal{E}(\bar{h}(\hat{\boldsymbol{x}})) - \mathcal{E}(\bar{g}(\boldsymbol{x}), \bar{h}(\hat{\boldsymbol{x}})) =: \mathcal{E}(\bar{g}(\boldsymbol{x})) + \mathcal{E}(\bar{h}(\hat{\boldsymbol{x}})) - \mathcal{E}(\bar{f}(\boldsymbol{x}, \hat{\boldsymbol{x}})) = \\ = -\int_{\text{supp}(\bar{g})} \bar{g}(\boldsymbol{x}) \log \bar{g}(\boldsymbol{x})\, d\boldsymbol{x} - \int_{\text{supp}(\bar{h})} \bar{h}(\hat{\boldsymbol{x}}) \log \bar{h}(\hat{\boldsymbol{x}})\, d\hat{\boldsymbol{x}} + \\ + \int_{\text{supp}(\bar{g})} \int_{\text{supp}(\bar{h})} \bar{f}(\boldsymbol{x}, \hat{\boldsymbol{x}}) \log \bar{f}(\boldsymbol{x}, \hat{\boldsymbol{x}})\, d\hat{\boldsymbol{x}}\, d\boldsymbol{x}.\end{aligned} \tag{G.47}$$

Note that in the main text, we focus on continuously distributed random variables exclusively (unless stated otherwise), so differentiation between the "double-dot" and "bar" probability distributions is irrelevant and we refrain from using these accents.

G.3 Proof of Proposition G.1, with additional comments

Proof. The range of CDF $\ddot{\Pi}(\chi)$ is exhaustively and efficiently (i.e. without intersections) tiled by δ-cells. Discretization of the range (once $n_d := 1/\delta$ is known) leaves only the domain of $\ddot{\Pi}(\chi)$ to describe.

Description of the CDF domain is done by specifying a sequence of domain values corresponding to each discretization cell, that is $\{\chi_i\}_1^{n_d}$ for $\chi_i = \ddot{\Pi}^{-1}(i\delta)$. Expected length of respective optimal codewords using instantaneous code equals $\mathcal{E}(\ddot{\pi}(\chi))$, which pins down the lower bound for average length of codewords in the description sequence. This can be shown by adjusting the argument about optimal instantaneous code and the bound on its expected length (e.g., see Cover and Thomas, 2006).

Specifically, denoting by n_i the number of discretization points per each quantization point i, we wish to minimize total description length:

$$\min_{\{L(\chi_i)\}_{i=1}^{n}} \sum_{i=1}^{n_d} L(\chi_i) = \sum_{i=1}^{n_q} n_i L(\chi_i) = n_d \sum_{i=1}^{n_q} \frac{n_i}{n_d} L(\chi_i) \tag{G.48}$$

subject to Kraft's inequality (that necessarily holds for any instantaneous code and guarantees the existence of such code)

$$\sum_{i=1}^{n_q} 2^{-L(\chi_i)} \leq 1$$

which yields

$$L^*(\chi_i) := -\log \frac{n_i}{n_d}.$$

At the same time, in the process of discretization we allocated n_i-s according to the rule

$$n_i := \frac{\ddot{\Pi}(\chi_i) - \ddot{\Pi}(\chi_{i-1})}{\delta} =: \frac{\ddot{\pi}(\chi_i)}{\delta},$$

hence

$$\frac{n_i}{n_d} = \ddot{\pi}(\chi_i).$$

Substituting into objective function (G.48) produces

$$\sum_{i=1}^{n_d} L^*(\chi_i) = -n_d \sum_{i=1}^{n_q} \frac{n_i}{n_d} \log \frac{n_i}{n_d} = -n_d \sum_{i=1}^{n_q} \ddot{\pi}(\chi) \log \ddot{\pi}(\chi) =: n_d \, \mathcal{E}(\ddot{\pi}(\chi)),$$

which is the claimed result.

Lastly, receiving χ_i as a sequence ordered, say, from lower to higher cumulative probability mass $\ddot{\Pi}(\chi_i)$ would have allowed shorter codeword lengths by ruling out the subset of the support corresponding to $\chi_{i'} < \chi_i$ with each new χ_i received. This possibility is assumed away.

□

Some additional comments are warranted. Note that minimizing $\sum_{i=1}^{n_d} L(\chi_i) = \sum_{i=1}^{n_q} n_i L(\chi_i)$ (i.e., number-of-occurences-weighted sum of codeword lengths) is equivalent to minimizing $\mathrm{E}[L(\chi)] = \sum_{i=1}^{n_q} \ddot{\pi}(\chi_i) L(\chi_i)$ (probability-weighted sum of lengths), hence the code assignment we obtain can also be written as

$$L^*(\chi_i) = -\log \ddot{\pi}(\chi_i),$$

which is a classic result for optimal instantaneous codes, and in expectation achieves the theoretical lower bound:

$$\mathrm{E}[L^*(\chi_i)] = -\sum_{i=1}^{n_q} \ddot{\pi}(\chi_i) \log \ddot{\pi}(\chi_i) =: \mathcal{E}(\ddot{\pi}(\chi)).$$

In some cases suggested theoretically optimal codeword lengths may be achievable only asymptotically, but using Shannon code assignment $L(\chi_i) := \lceil -\log \ddot{\pi}(\chi_i) \rceil$, the following bound is always achievable in practice:

$$\frac{1}{\delta} \times (\mathcal{E}(\ddot{\pi}(\chi)) + 1) \text{ bits.}$$

APPENDIX H

ALGORITHM FOR DECISION-MAKING UNDER RISK: A TOY PRIMER

Consider a fruit-tree bearing some random number of fruits every year, with known probability distribution guiding possible fruit harvests: say, the harvest may be 7 or 9 fruits each with probability $1/8$, 11 or 13 fruits with respective probabilities $1/4$, and 19 or 21 fruits with probabilities $1/8$. (The corresponding tables and plots are provided above in §G.)

There is also a potential investor. Essentially, he/she would first wish to assess the value of tree as a capital good. For instance, how productive the tree is on average: say, its mean harvest is $7 \times 1/8 + 9 \times 1/8 + 11 \times 1/4 + \ldots + 21 \times 1/8 = 13$ fruits; perhaps investor also cares about variance of the harvest: the latter can be calculated to be 20 here. Secondly, he would then make the decision about buying or selling the tree on the market: say, deciding to buy 3 trees.

Operationally, this involves sending a message, i.e., a sequence of signals, describing the above probability disribution from one, perceptive part of investor's brain to another, calculating part, and computing the relevant statistic of interest. However, if the probability distribution is "too complex" given investor's information processing limitations (we can think of them as individual "bandwidth"), the above procedure becomes infeasible and has to be modified.

In such a case the distribution is "simplified": investor chooses another probability distribution that roughly approximates the original one but possesses lower "complexity", and uses this subjective simplified distribution in computations of the value of the tree. Say, calculating the average harvest as $8 \times 1/4 + 12 \times 1/2 + 20 \times 1/4 = 13$ fruits, also relying on this new distribution in calculating the variance of just 19. Then, he decides how many trees to buy or sell: the decision is likely to be affected and differ from preceding 3 trees.

Moreover, investor may wish to bias the subjective distribution downward and calculate

the average as $7 \times 1/4 + 11 \times 1/2 + 19 \times 1/4 = 12$ fruits, with the corresponding variance calculation of 19 still; and then choosing to buy, say, 3 trees again, as lower productivity offsets lower risk. Taking this toy example even further, investor's identical (i.e., having the same information processing limits) twin employed outside of investment business—perhaps as an econometrician in academia—would likely stick to the unbiased average of 13 fruits and think that buying, say, 4 trees would have been a more sensible decision, being rather puzzled by his brother's cautious investment choice.

Formally, the investor here solves the following 2-period consumption-investment problem (the notation below is the same as in the main text, so specific details are omitted):

$$\max_{\{C_t, q_t\}} \{u(C_t) + \beta \mathrm{E}_t^h [u(C_{t+1})]\} = \{u(C_t) + \beta \int_{\mathbb{R}_+} u(C_{t+1})\, h(\hat{D}_{t+1})\, d\hat{D}_{t+1}\}$$

subject to budget constraints

$$C_t + P_t q_t = (P_t + D_t) q_{t-1} =: W_t,$$
$$C_{t+1} = \hat{D}_{t+1} q_t =: \hat{W}_{t+1}$$

with q_{t-1} and D_t given; and also where

$$h(\hat{D}_{t+1}) := \int_{\mathrm{supp}(g)} f(D_{t+1}, \hat{D}_{t+1})\, dD_{t+1},$$
$$f(D_{t+1}, \hat{D}_{t+1}) := \arg \left\{ \min_{f(\cdot,\cdot)} \mathrm{E}^f \left[|u(C(D_{t+1})) - u(C(\hat{D}_{t+1}))|^2 \right] \text{ s.t. } \mathcal{I}(g(D_{t+1}); h(\hat{D}_{t+1})) \leq \kappa \right\},$$
$$g(D_{t+1}) \text{ is given;}$$

plus the necessary technical restrictions. Basically, he chooses the controls as if facing a proxy variable $\hat{D}_{t+1}$ (i.e., D_{t+1} with some noise).

In the context of our example above, canonical rational inattention theory (Sims, 2003, 2006) would essentially envisage a consumer receving a nature's signal that informs about the

fruit harvest which has just ripened (e.g., that this turned out to be a bad year, yielding only 7 or 9 fruits with probability 1/2 each). The consumer will then decide how many fruits to eat (say, 5), and how many to commit selling on the market for reinvestment into trees (say, the rest). In anticipation, before the harvest ripens, the consumer chooses the probability distribution of such signal in every contingency taking into account the complexity of the fruit harvest's distribution (i.e., an optimal information acquisition strategy). If the latter is too high, the consumer needs to restrict the accuracy of the signal he is able to receive, thus behaving rationally inattentively. Clearly, this is a different problem from the one presented above.

Formally, the consumer there solves the following 2-period consumption-investment problem:

$$\max_{f(\cdot,\cdot)} \mathrm{E}_t^f \left[u(C_{t+1}) + \beta u(C_{t+2})\right] = \\ = \int_{\mathbb{R}^3_+} \left[u(C_{t+1}) + \beta u(C_{t+2})\right] \, f(\{C_{t+1}, q_{t+1}\}, D_{t+1}) \, d\{C_{t+1}, q_{t+1}\} \, dD_{t+1}$$

subject to budget constraints

$$C_{t+1} + P_{t+1} q_{t+1} = (P_{t+1} + D_{t+1}) q_t =: W_{t+1},$$
$$C_{t+2} = D_{t+2} q_{t+1} =: W_{t+2}$$

with q_t and D_{t+2} given; and also subject to

$$\mathcal{I}(g(D_{t+1}); h(\{C_{t+1}, q_{t+1}\})) \leq \kappa,$$
$$h(\{C_{t+1}, q_{t+1}\}) := \int_{\mathrm{supp}(g)} f(\{C_{t+1}, q_{t+1}\}, D_{t+1}) \, dD_{t+1},$$
$$g(D_{t+1}) \text{ is given;}$$

plus the necessary technical restrictions. Basically, he ends up choosing the controls while

observing D_{t+1} with some noise (i.e., a proxy variable $\hat{D}_{t+1}$).

APPENDIX I

NEUROFOUNDATIONS

This work is explicitly grounded on existing knowledge about the human brain and cognition, however patchy that knowledge is. Unfortunately, a disproportionate amount of current understanding is based on non-human evidence that can be gathered using not just relatively crude functional magnetic resonance imaging, but also high-resolution invasive methods. Naturally, vision takes a key spot by virtue of being one of the most complex neural systems observed in non-humans. To put things into perspective, primary visual cortex is one of the most intensively studied region of the brain, yet according to the estimate of Olshausen and Field (2005) we understand only 15% of its function.

Our framework is built not at the level of individual neurons, but rather functional subsystems (really, populations of neurons); and it is most directly related to computational neuroscience, less so to cognitive or behavioral neuroscience/psychology, and only indirectly to cellular neuroscience or neurophysiology (omitting altogether the discussion of specific anatomical regions of the central nervous system, chemical neuromodulators, etc.). However, Eliasmith and Anderson (2003), Eliasmith (2013) as well as Rao (2010) show how such computational model can be implemented with neurologically plausible apparatus.

Functional subsystems included in our framework are linked in a centralized serial (hierarchical, sequential) order, but each such module allows for distributed and/or parallel processing within the subsystem. Indeed, it is a well established fact that human brain utilizes diverse forms of computation, e.g. a hierarchical organization has been found in the visual cortex and distributed architecture seems to be utilized in the memory system (but even these example systems are not fully specialized and exhibit different types of processing); see Squire et al. (2008) for details.

We incorporate some key neuroscientific elements and concepts:

- Limited capacity communication channel, e.g. the optic nerve may be thought of as a

limited-capacity channel (Burton, 2000).

- Working memory, defined as a limited capacity system that temporarily maintains and stores information to support human thought processes by providing an interface between perception, long-term memory and action (Baddeley, 2003); a good primer on the operations with working memory necessary for on-line computations or solving more abstract Tower of Hanoi problem and Raven's Progressive Matrices test are provided by Smith and Jonides (2003).

- Bottleneck, arising when a downstream subsystem (output) has tighter capacity limits that an upstream one (input), for instance in communication channels serving auditory and visual processing (Baddeley et al., 1997; Zhaoping, 2006) as well as due to working memory restrictions (Reynolds et al., 2008); ultimately, such physical capacity limits reflect energy costs and a pursuit for metabolic efficiency (see Laughlin et al., 2000).

- "Native" (default, fixed, internal) code, which particular brain subsystem uses for its operations and which is a given (e.g., see Campbell and Epp, 2005); it is optimal with respect to *both* physiological characteristics of the brain subsystem and the "average" stimulus it encodes (i.e., in environmental sense), but in general it is not efficient (in the sense of Shannon) for the "average" stimulus, and channel capacity related arguments should account for this fact (as Laming, 2010, puts it, "the human subject can be viewed as a communication system with fixed 'coding'," see his paper for more details); taking a different perspective, as long as "native" codebook reflects prior beliefs about the stimulus to be encoded, then this prior should arguably be understood in the sense of empirical Bayes methodology.

- Efficient code, which is optimal with respect to the "average" stimulus the corresponding brain subsystem encodes (i.e., efficient in Shannon sense);[1] though it may not be

1. Unfortunately, under the influence of the so called "efficient code hypothesis" (Barlow, 1961), the convention in neuroscience and psychology does not discriminate between the concepts of "native" (fixed) and efficient codes, usually assuming that the latter applies. Laming (2010) offers a comprehensive critique.

optimal for some subset—say, most current—stimuli, an ad hoc improvement and additional reduction in capacity requirements of the brain subsystem involved can still be achieved, albeit at a cost to auxiliary subsystems (Bor et al., 2003); ensembles and summary statistics, as parts of scene processing, are also the concepts relevant here (e.g., see Cohen et al., 2016).

- Interface for encoding and decoding, which in information-theoretic sense are abstract notions defined as a map ("codebook") from input, source alphabet to output, target alphabet and, respectively, its reverse map (Cover and Thomas, 2006; Campbell and Epp, 2005; also see Hertz et al., 1991), while in neurosience they take concrete forms of mapping an input (say, sensory) stimulus to a neural response in terms of spike sequence and, respectively, its inverse (Squire et al., 2008; Dayan and Abbott, 2001; Doya et al., 2007); with Doya et al. (2007), Dayan and Abbott (2001), as well as Borst and Theunissen (1999) providing a bridge between abstract information-theoretic and applied neural treatments of these concepts.

- Value (of some reward) and value function (recognizing its dynamic recursive forward-looking aspects) are established notions in neuroscience, e.g. see Rangel et al. (2008), as well as Schultz et al. (1997), McClure et al. (2004) and Doya (2008) for laboratory evidence related to rewards; value fiunction is understood in terms similar to classical dynamic programming, e.g. see Lee et al. (2012) and Doya (2007) for short reviews, as well as Dayan and Abbott (2001) or Glimcher et al. (2008) for textbook treatment, with Bertsekas and Tsitsiklis (1996) as well as Sutton and Barto (1998) offering theoretical underpinnings.

- Probabilistic objects, such as probability distributions (including prior or posterior), risk (variance or entropy) and ambiguity, or likelihoods and mathematical expectations (e.g., expected value of the reward) are accepted components of brain activity, see Schultz et al. (2008) for an excellent review, as well as Ma and Jazayeri (2014), Doya

(2008), Platt and Huettel (2008), Rushworth and Behrens (2008), Knill and Pouget (2004), Pouget et al. (2013), Yang and Shadlen (2007), Vilares et al. (2012), Barber et al. (2003).

- Prediction errors, which is the discrepancy between new information and some "reference frame" (some predicted value or prior distribution), is another important concept, for instance used explicitly or implicitly in information coding (in the form of "predictive coding", which amounts to encoding and transmitting only residual errors in prediction, see Huang and Rao, 2011, Spratling, 2012, Summerfield and Egner, 2009, Zhaoping, 2006, Dayan and Abbott, 2001) or learning (in the form of "reward prediction errors" in reinforcement learning, see Lee et al., 2012, Doya, 2007, Dayan and Abbott, 2001, Schultz et al., 1997, Daw et al., 2011); also see below with regard to the process of learning.[2]

- Numerical/quantity information is treated here in correspondence with that in mathematics, and although by no means taken for granted, there actually is neuroscientific evidence that in important issues this may be a valid approach; see Dehaene(2009), Nieder and Dehaene (2009) and Nieder (2005) for reviews, specific instances are (i) discrete (numerocity, e.g. number of items) and continuous (extent, e.g. length) quantities (they are supported by functionally overlapping populations of neurons bolstering the idea of abstract quantity, a generalized magnitude system in the brain, as shown by Tudusciuc and Nieder, 2007, 2009), (ii) symbolic signs (e.g., Arabic digit "3" or written word "three") and analog iconic signs (e.g., sets of dots) representing quantities (they are supported by the same neural populations expressing some shared abstract code, see Piazza et al., 2007) as well as (iii) correspondence between concrete

2. Predictive coding may be playing an important role in implementing Bayesian inference (i.e., only differences from prior are encoded in the process of updating some relevant posterior distribution), for example see Huang and Rao (2011), also see Kwisthout and van Rooij (2013). Reward prediction errors as a concept are consistent with the mechanism of reference-dependent valuation, e.g. see Kahneman and Tversky (1979, 1992).

quantities and abstract formal symbols ("variables") used in computations (this is a natural feature of neural networks, e.g. see Dehaene and Changeux, 2003, for a simple neural model primer), hence we freely interchange within each pair as we often implicitly do in applied mathematics (and in line with common practice in computational neuroscience, see Dayan and Abbott, 2001).

The framework also utilizes commonly accepted neural mechamisms and processes:

- Non-linear computation and function approximation potentially performed by neural networks as well as machinery for training neural network parameters that provide great flexibility in terms of plausible neural mechanisms available to execute—and explain—human behavior (Hertz et al., 1991; Dayan and Abbott, 2001; Cybenko, 1989; also see Bullinaria, 2000, regarding the benefits and pitfalls of such flexibility).[3]

- Learning, which in neuroscience boils down to adjusting neural network synapses/connection weights, is usually differentiated into unsupervised learning and supervised (most commonly, reinforcement learning based on minimizing (reward) prediction errors), and may apply to learning entropy-reducing (redundancy-reducing in the neuroscientific literature) approximations, auxiliary transformations, but most importantly the model itself, including the value function; Hertz et al. (1991), Dayan and Abbott (2001) focus on the relevant mechanisms in artificial, while Lee et al. (2012), Rangel et al. (2008), Doya (2007) discuss such mechanisms from the perpective of biological neural networks; by way of clarification, (i) we focus on networks that have converged to the solution (i.e., solution to our dynamic programming and informational problems are already known—this may alternatively be interpreted as if fixed time and mental costs have been incurred at the outset— in abstract symbolic form, and subsequently any new parameter values such as variance-covariance matrix are just "plugged into"

3. Hertz et al. (1991) shows how even the simplest one-layer feed-forward artificial neural network containing a finite number of neurons can perform PCA.

the optimal solution), and (ii) so called active learning and evidence accumulation is beyond the scope of our work and is assumed away.

- Tranformations are applied to input information in order to reduce redundancy (decompositions, decorrelation) and/or dimensionality/entropy (filters); e.g. in visual processing the neuroscientifically accepted procedures are principal component analysis for Gaussian and independent component analysis or wavelet-like basis function representation for non-Gaussian heavy-tailed inputs (Dayan and Abbott, 2001; Simoncelli and Olshausen, 2001; Zhaoping, 2006; Doi et al., 2012), with a number of studies demonstrating emergence of receptive/projective fields similar to those observed in animal cortical areas and at the same time consistent with the result of implementing the above transformation procedures when neural models are trained on natural images (for example, see Olshausen and Field, 1996, or Spratling, 2012); such redundancy/dimension-reducing transformations are a form of efficient coding implemented in terms of structural ("causal" in neuroscience) higher-level objects (in case of wavelet-like transforms the conventional term in neuroscience is "sparse coding"), and constitute intermediate stages of predictive coding procedures, for more details see Huang and Rao (2011), Zhaoping (2006), Dayan and Abbott (2001).

- An important instance of such transformations is coordinate transformation and subsequent adjustment of the mean (as a form of "compensation") stated in Propositions 4–5 as well as C.1, which effectively reduces to linear rotation and translation operations; it could conceivably be implemented by a mechanism similar to the one responsible for maintaining organism's internal image of the world in spatial cognition, i.e. determination of the head-centered coordinates of some target given retinal coordinates that must account for any eye movement, which is evidently performed (once initial non-linear decomposition has been computed) online in real time via simple linear operations (see Pouget and Sejnowski, 1997, or Dayan and Abbott, 2001); also, this

mechanism may be related to neuroscientific findings that in primates' brains expected (mean) reward is clearly discriminated from reward uncertainty, i.e. measures of its risk (variance/entropy) and ambiguity (see Schultz et al., 2008, for a survey).

- Probabilistic computations, required to operate with probabilistic objects, seem to be one of the supported animal brain processes; literature includes examples of manipulations with probabilities/probability densities, priors and posteriors (Barber et al., 2003; Vilares et al., 2012; Kepecs and Mainen, 2012; Ma, 2012) as well as with likelihood functions and their ratios (Yang and Shadlen, 2007), performing integration/marginalization of joint probability densities (Beck et al., 2011), optimally combining several models (O'Reilly et al., 2013; Knill and Pouget, 2004), model inversion (Dayan and Abbott, 2001; Botvinick and Toussaint, 2012), implementing Monte Carlo sampling (Griffiths et al., 2012; Buesing et al., 2011; Hoyer and Hyvärinen, 2003) and stochastic mental simulations (Battaglia et al., 2013), etc. (for a broader overview, see Doya et al., 2007; Pouget et al., 2013; Dayan and Daw, 2008; Ma and Jazayeri, 2014).

- Recursive processing of information, which is necessary for iterative optimization (as well as for evaluating a menu of available options in serial rather than parallel manner within each optimizing iteration), is a feasible neural mechanism as a number of closed loops passing across a sequence of different brain areas has already been identified, according to Miller and Wallis (2008), Doya and Kimura (2008), Doya (2007), Daw et al. (2005) and Tanaka et al. (2004); this should not be confused with the standard feedforward, recurrent and feedback connections (though the role of feedback connections being less well understood) within the same area/between functionally similar areas of the brain, e.g. see Dayan and Abbott (2001) as well as Zhaoping (2006).

The machinery comprising the above neuroscientific elements and processes resonates strongly with the proposals of Chris Eliasmith's group that underlie the functional architecture and implementation principles of the large-scale computational prototype of human

brain constructed and reported in Eliasmith et al. (2012), with (some aspects of) the general approach of the neural model for decision making under uncertainty by Rao (2010), with the "levels of understanding" framework discussed and updated in Poggio (2012), with the theoretical treatment of "early vision" in Zhaoping (2006), among others.

Note that neuroscience often relies on Bayesian formalism for dealing with probabilistic material (e.g., see Doya et al., 2007; Pouget et al., 2013; Ma and Jazayeri, 2014; Dayan and Daw, 2008); while information theory uses both classical and Bayesian approaches (Cover and Thomas, 2006, is inclined to the former, but MacKay, 2003, emphasizes the latter); for the sake of simpler exposition, we follow the steps of Cover and Thomas (2006) and do not adopt Bayesian formalism explicitly, but our approach is reconciled with the Bayesian one by restricting ourselves to uninformative flat priors and ever-recurrent informational dynamics that precludes learning/updating. Additionally, note that we abstract away from neuron noise, which may play an important role in neural systems (Simoncelli and Olshausen, 2001; Eliasmith and Anderson, 2003; Cordes et al., 2007; Pouget et al., 2013; Ma and Jazayeri, 2014; though also see the sampling argument of Hoyer and Hyvarinen, 2003).

Further supporting details organized in a more systematic book format can be found in Squire et al. (2008), Dayan and Abbott (2001), Doya et al. (2007) that particularly emphasizes Bayesian approach, Baddeley et al. (2000) that stresses information-theoretic foundations, and Glimcher et al. (2008) that focuses on economic decision-making.

APPENDIX J

GENERALIZED OPTIMIZATION PROBLEM

More explicitly, in place of optimization problem

$$\max_{\boldsymbol{\theta}} \mathrm{E}_t^g \left[\varphi^\sharp(\boldsymbol{x}|\boldsymbol{\theta})\right] = \int_{\mathrm{supp}(g)} \varphi^\sharp(\boldsymbol{x}|\boldsymbol{\theta})\, g(\boldsymbol{x})\, d\boldsymbol{x}, \qquad \{\mathcal{P}_\theta\}$$

where

$$g(\boldsymbol{x}) \text{ is given,}$$

we consider generalized optimization problem

$$\max_{\boldsymbol{\theta}} \mathrm{E}^h \left[\varphi(\hat{\boldsymbol{x}}|\boldsymbol{\theta})\right] = \int_{\mathrm{supp}(h)} \varphi(\hat{\boldsymbol{x}}|\boldsymbol{\theta})\, h(\hat{\boldsymbol{x}})\, d\hat{\boldsymbol{x}}, \qquad \{\mathcal{P}_{\theta\mathcal{I}}\}$$

where

$$\begin{aligned} h(\hat{\boldsymbol{x}}) &:= \int_{\mathrm{supp}(g)} f(\boldsymbol{x}, \hat{\boldsymbol{x}})\, d\boldsymbol{x}, \\ f(\boldsymbol{x}, \hat{\boldsymbol{x}}) &:= \arg \left\{ \min_{f(\boldsymbol{x},\hat{\boldsymbol{x}})} \mathrm{E}^f \left[d(\varphi^\sharp(\boldsymbol{x}|\boldsymbol{\theta}), \varphi(\hat{\boldsymbol{x}}|\boldsymbol{\theta})) \right] \text{ subject to } \mathcal{I}(g(\boldsymbol{x}); h(\hat{\boldsymbol{x}})) \leq \kappa \right\}, \\ & g(\boldsymbol{x}) \text{ is given.} \end{aligned}$$

For a given $\boldsymbol{\theta}$, objective function $\varphi^\sharp(\cdot)$ by construction weakly envelops $\varphi(\cdot)$ from above. Distortion function may be chosen with the purpose of producing objective function approximation in the sense of L^p norm (to power p):

$$d(\varphi^\sharp(\boldsymbol{x}|\boldsymbol{\theta}), \varphi(\hat{\boldsymbol{x}}|\boldsymbol{\theta})) = ||\varphi^\sharp(\boldsymbol{x}|\boldsymbol{\theta}) - \varphi(\hat{\boldsymbol{x}}|\boldsymbol{\theta})||_p^p = |\varphi^\sharp(\boldsymbol{x}|\boldsymbol{\theta}) - \varphi(\hat{\boldsymbol{x}}|\boldsymbol{\theta})|^p =: d(\boldsymbol{x}, \hat{\boldsymbol{x}}). \qquad \text{(J.1)}$$

It is zero at partitions of $\mathrm{supp}(g(\boldsymbol{x})) \times \mathrm{supp}(h(\hat{\boldsymbol{x}}))$ such that $\varphi^\sharp(\boldsymbol{x}|\boldsymbol{\theta}) = \varphi(\hat{\boldsymbol{x}}|\boldsymbol{\theta})$, and

positive otherwise. Pointwise, non-integrated approximation error for $p = 2$ equals

$$\begin{aligned} d(\boldsymbol{x}, \hat{\boldsymbol{x}}) &= |\varphi^{\sharp}(\boldsymbol{x}|\boldsymbol{\theta}) - \varphi(\hat{\boldsymbol{x}}|\boldsymbol{\theta})|^2 = \\ &= \left| \varphi^{\sharp}(\boldsymbol{x}|\boldsymbol{\theta}) + \frac{\partial \varphi^{\sharp}(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \boldsymbol{x}^{\intercal}} (\hat{\boldsymbol{x}} - \boldsymbol{x}) + o(|\boldsymbol{x} - \hat{\boldsymbol{x}}|) - \varphi^{\sharp}(\boldsymbol{x}|\boldsymbol{\theta}) \right|^2 = \\ &= \left(\frac{\partial \varphi^{\sharp}(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \boldsymbol{x}^{\intercal}} (\hat{\boldsymbol{x}} - \boldsymbol{x}) + o(|\boldsymbol{x} - \hat{\boldsymbol{x}}|) \right)^2 . \end{aligned} \tag{J.2}$$

Thus, in regions with "large" $\partial \varphi^{\sharp}(\boldsymbol{x}|\boldsymbol{\theta}) / \partial \boldsymbol{x}^{\intercal}$, values of $\hat{\boldsymbol{x}}$ "far" from $\boldsymbol{x}$ will be penalized and ensured as "low" probability mass $f(\boldsymbol{x}, \hat{\boldsymbol{x}})$ as possible.

The generalized optimization problem can be expressed more condensely by defining a non-linear (information-processing capacity) "constrained expectations" operator

$$\begin{aligned} \mathrm{E}^{\kappa,p}\left[\varphi^{\sharp}(\boldsymbol{x}|\boldsymbol{\theta})\right] := \mathrm{E}^{h}\left[\varphi(\hat{\boldsymbol{x}}|\boldsymbol{\theta})\right] \ \text{ s.t. } \qquad & h(\hat{\boldsymbol{x}}) := \int_{\mathrm{supp}(g)} f(\boldsymbol{x}, \hat{\boldsymbol{x}}) \, d\boldsymbol{x}, \\ & f(\boldsymbol{x}, \hat{\boldsymbol{x}}) := \arg \Bigg\{ \min_{f(\boldsymbol{x}, \hat{\boldsymbol{x}})} \mathrm{E}^{f}\left[d(\varphi^{\sharp}(\boldsymbol{x}|\boldsymbol{\theta}), \varphi(\hat{\boldsymbol{x}}|\boldsymbol{\theta})) \right] \\ & \qquad\qquad \text{s.t. } \mathcal{I}(g(\boldsymbol{x}); h(\hat{\boldsymbol{x}})) \leq \kappa \Bigg\}, \\ & g(\boldsymbol{x}) \text{ is given.} \end{aligned} \tag{J.3}$$

Then, in place of solving

$$\max_{\boldsymbol{\theta}} \mathrm{E}^{g}\left[\varphi^{\sharp}(\boldsymbol{x}|\boldsymbol{\theta})\right],$$

we say that we are solving

$$\max_{\boldsymbol{\theta}} \mathrm{E}^{\kappa,p}\left[\varphi^{\sharp}(\boldsymbol{x}|\boldsymbol{\theta})\right].$$

Original optimization problem is a special case of the generalized formulation. For any $g(\boldsymbol{x})$ such that $\mathcal{E}(g(\boldsymbol{x})) < \infty$, there exists $\kappa^{\sharp} < \infty$ such that for any $\kappa \geq \kappa^{\sharp}$ and any chosen norm L^p (to power p), we have $\mathcal{P}_{\theta\mathcal{I}} \iff \mathcal{P}_{\theta}$, i.e.

$$\max_{\boldsymbol{\theta}} \mathrm{E}^{\kappa,p}\left[\varphi^{\sharp}(\boldsymbol{x}|\boldsymbol{\theta})\right] \iff \max_{\boldsymbol{\theta}} \mathrm{E}^{g}\left[\varphi^{\sharp}(\boldsymbol{x}|\boldsymbol{\theta})\right]. \tag{J.4}$$

APPENDIX K

CHOICE OF WEALTH SHARES INVESTED: RESOLVING THE DILEMMA OF CIRCULARITY

K.1 Alternative approaches to resolution

In addition to rather agnostic approach based on the robustness argument, which is actually taken in the dissertation, there are alternative ways of dealing with the dilemma.

One obvious method to resolve it would be to appeal to a fixed-point argument and use optimal allocations, in equilibrium leading to the use of net supplies of each tree as corresponding shares. But this would require (i) iterative updating of shares and of the distortion function, imposing unrealistic information processing demands (or continuous updating with some analogue of abstract formal symbolic-like calculations); or else (ii) making "schizophrenic" assumptions about the agent's information set: that he knows (nominal) net supplies before making allocation choices that will affect market prices and thus net supplies, but does not use that knowledge about net supplies to back out optimal allocations directly, without the need for any optimization (dismissing this last criticism by assuming our representative agent is formed by a continuum of identical agents is possible only if the latter are not aware of their identicality). Another alternative would be to use agent's previous-period holdings as shares; but in a stationary problem like ours this suffers from the same criticism of "schizophrenia" as the previous approach.

In the interest of completeness, §K.2 shows that the iterative/continuous updating approach yields results qualitatively similar to (in some sense actually a degenerate version of) those stemming from a more agnostic approach taken in the main body of the dissertation.

K.2 Iterative/continuous updating approach

Here we sketch the core features of the solution to feasible consumption and portfolio choice problem $\mathcal{P}_{\mathcal{QI}}$ with the correspoding informational sub-problem based on the distortion function from Proposition 3.1 in its unaltered form.

Right away, Propositions 3.2, 3.3 and 4 become redundant. As a result, so does Proposition C.1.

Next, Proposition 5 is modified in the following way. For random vector $\boldsymbol{r}_{t+1} \sim \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$, specific solution to the informational problem takes the form (note the absence of the boundary solution case)

$$f(\boldsymbol{\omega}_t^\intercal \boldsymbol{r}_{t+1} | \boldsymbol{\omega}_t^\intercal \hat{\boldsymbol{r}}_{t+1}) = (2\pi)^{-\frac{1}{2}} \left|\frac{\lambda}{2}\right|^{-\frac{1}{2}} \exp\left(-\frac{\left(\boldsymbol{\omega}_t^\intercal(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r)\right)^2}{2(\lambda/2)}\right), \quad \forall \hat{\boldsymbol{r}}_{t+1} \in \mathbb{R}^K;$$

resulting in decomposition

$$\boldsymbol{\omega}_t^\intercal \boldsymbol{r}_{t+1} = \boldsymbol{\omega}_t^\intercal \hat{\boldsymbol{r}}_{t+1} - \boldsymbol{\omega}_t^\intercal \check{\boldsymbol{\mu}}_r + \epsilon_{r,t+1},$$

also producing

$$\boldsymbol{\omega}_t^\intercal \boldsymbol{\Sigma}_r \boldsymbol{\omega}_t = \boldsymbol{\omega}_t^\intercal \hat{\boldsymbol{\Sigma}}_r \boldsymbol{\omega}_t + \Psi_r,$$

where

$$\epsilon_{r,t+1} \sim \mathcal{N}(0, \Psi_r), \qquad \Psi_r = \frac{\lambda}{2},$$
$$\boldsymbol{\omega}_t^\intercal \hat{\boldsymbol{r}}_{t+1} \sim \mathcal{N}(\boldsymbol{\omega}_t^\intercal \hat{\boldsymbol{\mu}}_r, \boldsymbol{\omega}_t^\intercal \hat{\boldsymbol{\Sigma}}_r \boldsymbol{\omega}_t), \qquad \boldsymbol{\omega}_t^\intercal \hat{\boldsymbol{\Sigma}}_r \boldsymbol{\omega}_t = \boldsymbol{\omega}_t^\intercal \boldsymbol{\Sigma}_r \boldsymbol{\omega}_t - \Psi_r;$$

$$\lambda = 2e^{-2\kappa} |\boldsymbol{\omega}_t^\intercal \boldsymbol{\Sigma}_r \boldsymbol{\omega}_t|.$$

Or more conveniently (signifying Moore–Penrose pseudo-inverse and pseudo-determinant in the sense of the product of all non-zero eigenvalues with + superscript and subscript,

respectively),

$$f(\boldsymbol{r}_{t+1}|\hat{\boldsymbol{r}}_{t+1}) = (2\pi)^{-\frac{K}{2}} \, |\boldsymbol{\Psi}_r|_+^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r)^\intercal \boldsymbol{\Psi}_r^+ (\boldsymbol{r}_{t+1} - \hat{\boldsymbol{r}}_{t+1} + \check{\boldsymbol{\mu}}_r)\right), \\ \forall \hat{\boldsymbol{r}}_{t+1} \in \mathbb{R}^K;$$

resulting in decomposition

$$\boldsymbol{r}_{t+1} = \hat{\boldsymbol{r}}_{t+1} - \check{\boldsymbol{\mu}}_r + \boldsymbol{\epsilon}_{r,t+1},$$

also producing

$$\boldsymbol{\Sigma}_r = \hat{\boldsymbol{\Sigma}}_r + \boldsymbol{\Psi}_r,$$

where

$$\begin{aligned} \boldsymbol{\epsilon}_{r,t+1} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_r), & \boldsymbol{\Psi}_r &= \frac{\lambda}{2}(\boldsymbol{\omega}_t^\intercal \boldsymbol{\omega}_t)^{-2} \boldsymbol{\omega}_t \boldsymbol{\omega}_t^\intercal, \\ \hat{\boldsymbol{r}}_{t+1} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_r, \hat{\boldsymbol{\Sigma}}_r), & \hat{\boldsymbol{\Sigma}}_r &= \boldsymbol{\Sigma}_r - \boldsymbol{\Psi}_r; \end{aligned}$$

$$\lambda = 2e^{-2\kappa} \boldsymbol{\omega}_t^\intercal \boldsymbol{\omega}_t |\boldsymbol{\Sigma}_r|.$$

(Note that

$$|\boldsymbol{\omega}_t^\intercal \boldsymbol{\Sigma}_r \boldsymbol{\omega}_t| \lesseqgtr \boldsymbol{\omega}_t^\intercal \boldsymbol{\omega}_t |\boldsymbol{\Sigma}_r|$$

in general, as can be easily shown using Rayleigh–Ritz theorem, and hence λ multipliers above are different for the same values of κ. In short, this stems from the fact that differential entropy is not invariant to transformations such as rescaling of random variables involved.) This latter, more convenient formulation implies a rank-one matrix $\boldsymbol{\Psi}_r$, so that elements of vector $\boldsymbol{\epsilon}_{r,t+1}$ are perfectly correlated (random vector $\boldsymbol{\epsilon}_{r,t+1}$ is then said to have singular Normal distribution). It is also noteworthy that, counter to intuition, elements of approximation error variance-covariance matrix $\boldsymbol{\Psi}_r$ that correspond to bigger components of $\boldsymbol{\omega}_t$ are relatively larger and, hence, respective approximation precisions are smaller.

Finally, Proposition 2 in its part applicable to interior soluton case holds without change, i.e. subjective correlations between elements of $\hat{\boldsymbol{r}}_{t+1}$ are inflated versions of their objective counterparts (as can be seen from the direction of change of generic correlation coefficient $\hat{\rho}_{r,kl}$ for $k, l \in \{1, \ldots, K\}$ when λ increases).

Solution to consumption and investment sub-problem is adjusted inasmuch as to account for changes in $\hat{\boldsymbol{\Sigma}}_r$.

To summarize, the main features of the solution under iterative/continuous updating approach are similar to those based on the robustness argument. Solutions to informational sub-problem have the same broad form, which in turn determines the extent of differences to the solutions to consumption and investment sub-problems, and key results such as categorization hold under both approaches.

APPENDIX L

SHRINKAGE

Another alternative way of implementing variance counter-adjustment is variance-covariance matrix "shrinkage", which involves shrinking it towards some "well-behaved" target matrix, or equivalently shrinking the variance-covariance matrix's eigenvalues in desired direction in the sense of Stein (see Stein, 1956; James and Stein 1961). From the perspective of our approach, such shrinkage can also be interpreted as an attempt to recover the original variance-covariance matrix $\mathbf{\Sigma}_r$ from simplified matrix $\hat{\mathbf{\Sigma}}_r$.

In terms of notation used here, a popular method of Ledoit and Wolf (2004a) uses as shrunk and thus regularized variance-covariance matrix $\left((1-\alpha)\hat{\mathbf{\Sigma}}_r + \alpha\eta\boldsymbol{I}_K\right)$, for $\alpha \in [0,1]$ and some positive constant η; note that replacing $\eta\boldsymbol{I}_K$ with $\left(\hat{\mathbf{\Sigma}}_r + \alpha^{-1}\mathbf{\Psi}_r\right)$ would recover $\mathbf{\Sigma}_r$ precisely. While Jagannathan and Ma (2003) use instead $\left(\hat{\mathbf{\Sigma}}_r + \mathbf{\Upsilon}_r\right)$ for some reduced-rank matrix $\mathbf{\Upsilon}_r$; again, replacing $\mathbf{\Upsilon}_r$ with $\mathbf{\Psi}_r$ would recover $\mathbf{\Sigma}_r$ exactly.

Conventional rationale for shrinkage is regularizing eigenvalues by squeezing them together (reducing the largest and amplifying the smallest ones), which improves matrix conditioning and makes easier its inversion and hence usage in, say, portfolio optimization. There is an established result in random matrix theory that eigenvalues of sample valiance-covariance matrix are overdispersed: the largest sample eigenvalue asymptotically overestimates the largest population eigenvalue, and the smallest sample eigenvalue underestimates its population counterpart. This result is based on Marchenko-Pastur and Wigner's semicircle distribution laws, for reference see Stein (1975 or 1986) and Johnstone (2001), also see Ledoit and Wolf (2013).

Note that Proposition 4 implies an increase in the condition number (defined as the ratio of the largest and the smallest eigenvalues) for matrix $\hat{\mathbf{\Sigma}}$ relatively to the condition number of $\mathbf{\Sigma}$. The parallel between simplified variance-covariance matrix that we obtain and the sample variance-covariance matrix tackled in random matrix theory is interesting in its own right.

(In the interest of completeness, practical applications have extended Stein-type shrinkage also to estimates of the mean (Jorion, 1985; 1986), which resonates with our earlier arguments regarding mean adjustment.)

APPENDIX M

MACHINE-AIDED INFORMATION PROCESSING

Our treatment is general enough to also account for information processing performed with the aid of machines.

When presenting in part §2.2 (together with §G) the process of decision-making under risk that underlies our framework, we illustrated the reduction of probability distribution's entropy by amalgamation of several low-probability rare events into one, as well as by complete omission of events not observed in sample. The algorithm discussed there also focused on mental computations as the practical implementation of information processing we are concerned about.

However, we can push this logic further and argue that it also holds for machine-aided computations. For example, this is trivially true when the costs—in terms of information processing capacity demands—of formalization and coding machine instructions are proportional to entropy of the distribution concerned (as was the case with mental computations). (Such a proportional relationship can be established by way of renowned Solomonoff-Kolmogorov-Chaitin complexity notion (see Rissanen, 2007), e.g. employing the arguments of Leung-Yan-Cheong and Cover (1978).) Therefore, the same algorithm of §G applies to real-world situations in which machine-computing technologies are used heavily. The only technical difference pertains to implementation and interpretation: while in the case of mental computations the most natural information processing "bottleneck" is working memory, or perhaps the computation of optimal decision, in the case of machine computations it is the summarization of given information, i.e. expressing everything in unambiguous formal language (which, as a matter of fact, admits "back-of-the-envelope" approximations, but precludes "hand-waving"). The operational outcome of extending our logic to account also for machine computations again boils down to reduction in the entropy of the probability distributions used, because the high-entropy true distribution has to be simplified to become amenable for being coded up.

In this sense of machine-aided information processing, a computationally cheap way of summarizing a complex true probability distribution that saves on formalization and coding costs is using observed sample data, which thus serve as a simplified distribution. (Supposing subsequent mental computations above and beyond that, "low" information processing capacity κ would require further simplification of the sample distribution, as well as further compensating mean-adjustment.)

If (representative) investor was the leading example of an agent using mental computations, an econometrician processing sample data is the most intuitive way of thinking about an agent that uses machine computations. The corresponding effective information processing capacity κ is then easiest understood as effective capacity of an econometrician, rather than as effective capacity of investor relying exclusively (i.e., no more and no less) on sample data. Indeed, econometricians usually utilize sample distribution as is without imposing any (pessimistic) adjustments to the mean, in contrast to (optimally behaving) investors. This fits perfectly into our formal framework (although not contemplating it ex ante, econometrician's behavior nevertheless falls under its logic ex post): according to Proposition 3.1, optimal magnitude of mean adjusment is measured from the origin point of zero wealth shares invested in risky assets $\boldsymbol{\omega}_t$, with said point implying matching expected values of returns and at the same time presenting a sensible way to think about an impartial non-market-participating econometrician.

REFERENCES

[1] Abdellaoui, Mohammed. (2000) "Parameter-Free Elicitation of Utility and Probability Weighting Functions", *Management Science*, 46(11): 1497–1512.

[2] Abdellaoui, Mohammed, Frank Vossmann and Martin Weber. (2005) "Choice-Based Elicitation and Decomposition of Decision Weights for Gains and Losses under Uncertainty", *Management Science*, 51(9): 1384–1399.

[3] Abel, Andrew B. (2002) "An Exploration of the Effects of Pessimism and Doubt on Asset Returns." *Journal of Economic Dynamics and Control*, 26 (7–8): 1075–1092.

[4] Allais, Maurice. (1953) "Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine", *Econometrica*, 21(4): 503–546.

[5] Alvarez, Fernando and Urban J. Jermann. (2005) "Using Asset Prices to Measure the Persistence of the Marginal Utility of Wealth", *Econometrica*, 73(6): 1977–2016.

[6] Anderson, Anders. (2013) "Trading and Under-Diversification", *Review of Finance*, 17: 1699–1741.

[7] Andersson, Magnus, Elizaveta Krylova and Sami Vähämaa. (2008) "Why Does the Correlation Between Stock and Bond Returns Vary Over Time?", *Applied Financial Economics*, 18: 139–151.

[8] Aït-Sahalia, Yacine and Andrew W. Lo. (2000) "Nonparametric Risk Management and Implied Risk Aversion", *Journal of Econometrics*, 94: 9–51.

[9] Arora, Sanjeev and Boaz Barak. (2009) *Computational Complexity: A Modern Approach*, New York, NY: Cambridge University Press.

[10] Backus, David, Mikhail Chernov and Stanley Zin. (2014) "Sources of Entropy in Representative Agent Models", *Journal of Finance*, 69(1): 51–99.

[11] Baddeley, Alan. (2003) "Working Memory: Looking Back and Looking Forward." *Nature Reviews Neuroscience*, 4: 829–839.

[12] Baddeley, Roland, L. F. Abbott, Michael C.A. Booth, Frank Sengpiel, Tobe Freeman, Edward A. Wakeman and Edmund T. Rolls. (1997) "Responses of neurons in primary and inferior temporal visual cortices to natural scenes." *Proceedings of the Royal society B*, 264(1389): 1775–1783.

[13] Baddeley, R., P. Hancock and P. Földiák (eds.). (2000) *Information Theory and the Brain*, New York, NY: Cambridge University Press.

[14] Bansal, Ravi and Amir Yaron. (2004) "Risks for the Long-Run: A Potential Resolution of Asset Pricing Puzzles." *Journal of Finance*, 59(4), 1481–1509.

[15] Barber, Brad M., Terrance Odean, Ning Zhu. (2009) "Systematic Noise", *Journal of Financial Markets*, 12: 547–569.

[16] Barber, M. J., J.W. Clark and C. H. Anderson. (2003) "Neural Representation of Probabilistic Information." *Neural Computation*, 15: 1843–1864.

[17] Barberis, Nicholas and Ming Huang. (2008) "Stocks as Lotteries: The Implications of Probability Weighting for Security Prices", *American Economic Review*, 98(5): 2066–2100.

[18] Barberis, Nicholas and Andrei Shleifer. (2003) "Style investing", *Journal of Financial Economics*, 68: 161–199.

[19] Barberis, Nicholas, Andrei Shleifer and Jeffrey Wurgler. (2005) "Comovement", *Journal of Financial Economics*, 75: 283–317.

[20] Barberis, Nicholas C. (2013a) "The Psychology of Tail Events: Progress and Challenges", *American Economic Review: Papers and Proceedings*, 103(3): 611–616.

[21] Barberis, Nicholas C. (2013b) "Thirty Years of Prospect Theory in Economics: A Review and Assessment", *Journal of Economic Perspectives*, 27(1): 173–196.

[22] Barlow, Horace B. (1961) "Possible Principles Underlying the Transformation of Sensory Messages." In: W. Rosenblith (ed.), *Sensory Communication*, Cambridge, MA: MIT Press.

[23] Barro, Robert J. (2006) "Rare Disasters and Asset Markets in the Twentieth Century." *The Quarterly Journal of Economics*, 121(3): 823–866.

[24] Battaglia, Peter W., Jessica B. Hamrick and Joshua B. Tenenbaum. (2013) "Simulation as an Engine of Physical Scene Understanding." *Proceedings of the National Academy of Sciences*, 110(45): 18327–18332.

[25] Bayer, Hannah. (2008) "Focus on Decision Making", *Nature Neuroscience*, 11(4): 387.

[26] Beck, Jeffrey M., Peter E. Latham and Alexandre Pouget. (2011) "Marginalization in Neural Circuits with Divisive Normalization." *Journal of Neuroscience*, 31(43):15310–15319.

[27] Berger, Toby. (1971) *Rate-distortion Theory: A Mathematical Basis for Data Compression*,

Englewood Cliffs, NJ: Prentice-Hall.

[28] Bertsekas, Dimitri P. and John N. Tsitsiklis. (1996) *Neuro-Dynamic Programming*, Belmont, MA: Athena Scientific.

[29] Bickel, Peter J. and Bo Li (2006) "Regularization in Statistics." *Test*, 15(2): 271-344.

[30] Bishop, Christopher M. (2006) *Pattern Recognition and Machine Learning*, Springer.

[31] Bleichrodt, Han, Jaco van Rijn and Magnus Johannesson. (1999) "Probability Weighting and Utility Curvature in QALY-Based Decision Making", *Journal of Mathematical Psychology*, 43: 238–260.

[32] Bleichrodt, Han, Jose Luis Pinto. (2000) "A Parameter-Free Elicitation of the Probability Weighting Function in Medical Decision Analysis", *Management Science*, 46(11): 1485–1496.

[33] Blume, Marshall E. and Irwin Friend (1975) "The Asset Structure of Individual Portfolios and Some Implications for Utility Functions", *The Journal of Finance*, 30(2): 585–603.

[34] Bor, Daniel, John Duncan, Richard J. Wiseman and Adrian M. Owen. (2003) "Encoding Strategies Dissociate Prefrontal Activity from Working Memory Demand", *Neuron* 37: 361–367.

[35] Bor, Daniel and Adrian M. Owen. (2006) "Working Memory: Linking Capacity with Selectivity", *Current Biology*, 16(4): 136–138.

[36] Bordalo, Pedro, Nicola Gennaioli and Andrei Shleifer. (2016) "Diagnostic Expectations and Credit Cycles." Manuscript.

[37] Borst, Alexander and Frederic E. Theunissen. (1999) "Information Theory and Neural Coding." *Nature Neuroscience*, 2(11): 947–957.

[38] Bossaerts, Peter, Kerstin Preuschoff and Ming Hsu. (2008). "The Neurobiological Foundations of Valuation in Human Decision Making Under Uncertainty." In: Paul W. Glimcher, Colin F. Camerer, Ernst Fehr, Russell A. Poldrack (eds.), *Neuroeconomics: Decision Making and the Brain*, Academic Press.

[39] Botvinick, Matthew and Marc Toussaint. (2012) "Planning as Inference." *Trends in Cognitive Sciences*, 16(10): 485–488.

[40] Boyer, Brian H. (2011) "Style-Related Comovement: Fundamentals or Labels?", *The Journal of Finance*, 66(1): 307–332.

[41] Breeden, Douglas T. (1979) "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities." *Journal of Financial Economics*, 7: 265–296.

[42] Brown, Stephen J. and William N. Goetzmann. (1997) "Mutual Fund Styles", *Journal of Financial Economics*, 43: 373–399.

[43] Bruhin, Adrian, Helga Fehr-Duda and Thomas Epper. (2010) "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion", *Econometrica*, 78(4): 1375–1412.

[44] Brunnermeier, Markus K., Christian Gollier, and Jonathan A. Parker. (2007) "Optimal Beliefs, Asset Prices, and the Preference for Skewed Returns." *American Economic Review (Papers and Proceedings)*, 97 (2): 159–165.

[45] Brunnermeier, Markus K. and Jonathan A. Parker. (2005) "Optimal Expectations." *American Economic Review*, 95 (4): 1092–1118.

[46] Brunnermeier, Markus K., Filippos Papakonstantinou and Jonathan A. Parker. (2013) "Optimal Time-Inconsistent Beliefs: Misplanning, Procrastination, and Commitment." Manuscript.

[47] Buesing, Lars, Johannes Bill, Bernhard Nessler and Wolfgang Maass. (2011) "Neural Dynamics as Sampling: A Model for Stochastic Computation in Recurrent Networks of Spiking Neurons." *PLoS Computational Biology*, 7(11): 1–22.

[48] Bullinaria, John A. (2000) "Free Gifts from Connectionist Modelling." In: R. Baddeley, P. Hancock and P. Foldiak (eds.), *Information Theory and the Brain*, pages 221–240, New York, NY: Cambridge University Press.

[49] Bunge, Silvia A. (2005) "Foreword", *Cognitive Brain Research*, 23: 1.

[50] Burton, Brian G. (2000) "Problems and Solutions in Early Visual Processing." In: R. Baddeley, P. Hancock and P. Foldiak (eds.), *Information Theory and the Brain*, pages 25–40, New York, NY: Cambridge University Press.

[51] Camerer, Colin F. (1995) "Individual Decision Making." In: John H. Kagel and Alvin E. Roth (eds.), *Handbook of Experimental Economics*, chapter 8, pages 587–703, Princeton, NJ: Princeton University Press.

[52] Camerer, Colin F. and Teck-Hua Ho. (1994) "Violations of the Betweenness Axiom and Nonlinearity in Probability", *Journal of Risk and Uncertainty*, 8(2): 167–196.

[53] Campbell, Jamie I. D. and Lynette J. Epp. (2005) "Architectures for Arithmetic." In: Jamie I.D. Campbell (ed.), *Handbook of Mathematical Cognition*, pages 347–360, Psychology Press.

[54] Campbell, John Y. (2003) "Consumption-Based Asset Pricing." In: George M. Constantinides, Milton Harris and Rene M. Stulz (eds.), *Handbook of the Economics of Finance*, volume 1, part B, chapter 13, pages 803–887, Elsevier.

[55] Campbell, John Y. and John Ammer. (1993) "What Moves the Stock and Bond Markets? A Variance Decomposition for Long-Term Asset Returns", *The Journal of Finance*, 48(1): 3–37.

[56] Campbell, John Y. and John Cochrane. (1999) "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior." *Journal of Political Economy*, 107: 205–251.

[57] Campbell, John Y. and Luis M. Viceira. (2002a) *Appendix for "Strategic Asset Allocation: Portfolio Choice for Long-Term Investors"*, Oxford University Press.

[58] Campbell, John Y. and Luis M. Viceira. (2002b) *Strategic Asset Allocation: Portfolio Choice for Long-Term Investors*, Oxford University Press.

[59] Campbell, John Y., Yeung Lewis Chan and Luis M. Viceira. (2003) "A Multivariate Model of Strategic Asset Allocation", *Journal of Financial Economics*, 67: 41–80.

[60] Canner, Niko, N. Gregory Mankiw and David N. Weil. (1997) "An Asset Allocation Puzzle", *American Economic Review*, 87(1): 181–191.

[61] Caplin, Andrew and Mark Dean. (2013) "The Behavioral Implications of Rational Inattention with Shannon Entropy." Manuscript.

[62] Caplin, Andrew and Mark Dean. (forthcoming) "Revealed Preference, Rational Inattention, and Costly Information Acquisition." *American Economic Review*.

[63] Carroll, Christopher D. (2003) "Macroeconomic Expectations of Households and Professional Forecasters." *Quarterly Journal of Economics*, 118 (1): 269–298.

[64] Cecchetti, Stephen G., Pok-sang Lam and Nelson C. Mark. (2000) "Asset Pricing with Distorted Beliefs: Are Equity Returns Too Good to Be True?" *American Economic Review*, 90 (4): 787–805.

[65] Chan, Louis K.C., Jason Karceski, and Josef Lakonishok. (2000) "New Paradigm or Same

Old Hype in Equity Investing?", *Financial Analysts Journal*, 56(4): 23–36.

[66] Chen, Hui, Winston Wei Dou and Leonid Kogan. (2015) "Measuring the "Dark Matter" in Asset Pricing Models." Manuscript.

[67] Chen, Zhe and Simon Haykin. (2002). "On Different Facets of Regularization Theory." *Neural Computation*, 14: 2791–2846.

[68] Choi, Nicole and Richard W. Sias. (2009) "Institutional Industry Herding", *Journal of Financial Economics*, 94: 469–491.

[69] Cohen, Michael A., Daniel C. Dennett and Nancy Kanwisher. (2016) "What is the Bandwidth of Perceptual Experience?" *Trends in Cognitive Sciences*, 20(5): 324–335.

[70] Connolly, Robert, Chris Stivers and Licheng Sun. (2005) "Stock Market Uncertainty and the Stock-Bond Return Relation", *Journal of Financial and Quantitative Analysis*, 40(1): 161–194.

[71] Cooper, Michael J., Huseyin Gulen and P. Raghavendra Rau. (2005) "Changing Names with Style: Mutual Fund Name Changes and Their Effects on Fund Flows", *The Journal of Finance*, 60(6): 2825–2858.

[72] Cordes, Sara, C. R. Gallistel, Rochel Gelman and Peter Latham. (2007) "Nonverbal Arithmetic in Humans: Light from Noise." *Perception & Psychophysics*, 69(7): 1185–1203.

[73] Cover, Thomas M. and Joy A. Thomas. (2006) *Elements of Information Theory*, 2nd ed., New York: Wiley-Interscience.

[74] Cowan, Nelson. (2000) "The Magical Number 4 in Short-term Memory: A Reconsideration of Mental Storage Capacity", *Behavioral and Brain Sciences*, 24(1): 87–185.

[75] Cybenko, George. (1989) "Approximations by Superpositions of Sigmoidal Functions." *Mathematics of Control, Signals, and Systems*, 2(4): 303–314.

[76] Damper, Robert I. and Stevan Harnad. (2000) "Neural Network Modeling of Categorical Perception", *Perception & Psychophysics*, 62(4): 843–867.

[77] Daw, Nathaniel D., Samuel J. Gershman, Ben Seymour, Peter Dayan, and Raymond J. Dolan. (2011) "Model-Based Influences on Humans' Choices and Striatal Prediction Errors." *Neuron*, 69: 1204–1215.

[78] Daw, Nathaniel D., Yael Niv and Peter Dayan. (2005) "Uncertainty-Based Competition

Between Prefrontal and Dorsolateral Striatal Systems for Behavioral Control." *Nature Neuroscience*, 8(12): 1704–1711.

[79] Dayan, Peter and Laurence F. Abbott. (2001) *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*, The MIT Press.

[80] Dayan, Peter and Nathaniel D. Daw. (2008) "Decision Theory, Reinforcement Learning, and the Brain." *Cognitive, Affective, & Behavioral Neuroscience*, 8(4): 429–453.

[81] DeGiorgi, Enrico G. and Shane Legg. (2012) "Dynamic portfolio choice and asset pricing with narrow framing and probability weighting", *Journal of Economic Dynamics and Control*, 36: 951–972.

[82] Dehaene, Stanislas. (2009) "Origins of Mathematical Intuitions: The Case of Arithmetic." *Annals of the New York Academy of Sciences*, 1156: 232–259.

[83] Dehaene, Stanislas and Jean-Pierre Changeux. (2003) "Development of Elementary Numerical Abilities: A Neuronal Model." *Journal of Cognitive Neuroscience*, 5(4): 390–407.

[84] Dierkes, Maik. (2013) "Probability Weighting and Asset Prices." Manuscript.

[85] Dimson, Elroy, Paul Marsh and Mike Staunton. (2003) "Global Evidence on the Equity Risk Premium". *LBS Institute of Finance and Accounting Working Paper*, No. IFA 385.

[86] Doeswijk, Ronald Q., Trevin W. Lam and Laurens Swinkels. (2014) "The Global Multi-Asset Market Portfolio 1959–2012." Manuscript.

[87] Doi, Eizaburo, Jeffrey L. Gauthier, Greg D. Field, Jonathon Shlens, Alexander Sher, Martin Greschner, Timothy A. Machado, Lauren H. Jepson, Keith Mathieson, Deborah E. Gunning, Alan M. Litke, Liam Paninski, E. J. Chichilnisky and Eero P. Simoncelli. (2012) "Efficient Coding of Spatial Information in the Primate Retina." *Journal of Neuroscience*, 32(46):16256–16264.

[88] Doya, Kenji. (2008) "Modulators of Decision Making." *Nature Neuroscience*, 11(4): 410–416.

[89] Doya, Kenji. (2007) "Reinforcement Learning: Computational Theory and Biological Mechanisms." *HFSP Journal*, 1(1): 30–40.

[90] Doya, Kenji, Shin Ishii, Alexandre Pouget and Rajesh P.N. Rao (eds.). (2007) *Bayesian Brain: Probabilistic Approaches to Neural Coding*, The MIT Press.

[91] Doya, Kenji and Minoru Kimura. (2008) "The Basal Ganglia and the Encoding of Value." In:

Paul W. Glimcher, Colin F. Camerer, Ernst Fehr, Russell A. Poldrack (eds.), *Neuroeconomics: Decision Making and the Brain*, pages 407–416, Academic Press.

[92] Eliasmith, Chris. (2013) *How to Build a Brain: A Neural Architecture for Biological Cognition*, New York: Oxford University Press.

[93] Eliasmith, Chris and Charles H. Anderson. (2003) *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*, The MIT Press.

[94] Eliasmith,Chris, Terrence C. Stewart, Xuan Choo, Trevor Bekolay, Travis DeWolf, Yichuan Tang, Daniel Rasmussen. (2012) "A Large-Scale Model of the Functioning Brain." *Science*, 338(6111): 1202–1205.

[95] Epstein, Larry G. and Stanley E. Zin. (1989) "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework." *Econometrica*, 57: 937–969.

[96] Epstein, Larry G. and Stanley E. Zin. (1990) "'First-Order' Risk Aversion and the Equity Premium Puzzle." *Journal of Monetary Economics*, 26(3): 387–407.

[97] Epstein, Larry G. and Stanley E. Zin. (1991) "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis." *Journal of Political Economy*, 99(2): 263–286.

[98] Fama, Eugene F. (1963) "Mandelbrot and the Stable Paretian Hypothesis." *The Journal of Business*, 36(4): 420–429.

[99] Fama, Eugene F. and Kenneth R. French. (1993) "Common Risk Factors in the Returns on Stocks and Bonds", *Journal of Financial Economics*, 33: 3–56.

[100] Fama, Eugene F. and Kenneth R. French. (2002) "The Equity Premium", *The Journal of Finance*, 57: 637–659.

[101] Fehr-Duda, Helga and Thomas Epper. (2012) "Probability and Risk: Foundations and Economic Implications of Probability-Dependent Risk Preferences", *Annual Review of Economics*, 4: 567–593.

[102] Fleming, Jeff, Chris Kirby and Barbara Ostdiek. (1998) "Information and Volatility Linkages in the Stock, Bond, and Money Markets", *Journal of Financial Economics*, 49: 111–137.

[103] Fleming, Stephen M., Laurence T. Maloney and Nathaniel D. Daw. (2013) "The Irrationality

of Categorical Perception", *Journal of Neuroscience*, 33(49): 19060–19070.

[104] Fox, Craig R., Brett A. Rogers and Amos Tversky. (1996) "Options Traders Exhibit Subadditive Decision Weights", *Journal of Risk and Uncertainty*, 13(1): 5–17.

[105] Fox, Craig R. and Russell A. Poldrack. (2008). "Prospect Theory and the Brain." In: Paul W. Glimcher, Colin F. Camerer, Ernst Fehr, Russell A. Poldrack (eds.), *Neuroeconomics: Decision Making and the Brain*, Academic Press.

[106] French, Kenneth R. and James M. Poterba. (1991) "Investor Diversification and International Equity Markets", *American Economic Review*, 81(2): 222–226.

[107] Friston, Karl. (2009) "The Free-energy Principle: A Rough Guide to the Brain?" *Trends in Cognitive Sciences*, 13(7): 293–301.

[108] Friston, Karl. (2010) "The Free-energy Principle: A Unified Brain Theory?" *Nature Reviews Neuroscience*, 11: 127–138.

[109] Froot, Kenneth and Melvyn Teo. (2008) "Style Investing and Institutional Investors", *Journal of Financial and Quantitative Analysis*, 43(4): 883–906.

[110] Fuster, Andreas, Benjamin Hebert, and David Laibson. (2012) "Natural Expectations, Macroeconomic Dynamics, and Asset Pricing." *NBER Macroeconomics Annual*, 26 (1): 1–48.

[111] Gabaix, Xavier. (2014a) "A Sparsity-Based Model of Bounded Rationality." *Quarterly Journal of Economics*, 129 (4): 1661–1710.

[112] Gabaix, Xavier. (2014b) "Sparse Dynamic Programming and Aggregate Fluctuations." Manuscript.

[113] Gabaix, Xavier and David Laibson. (2000) "A Boundedly Rational Decision Algorithm." *American Economic Review (Papers and Proceedings)*, 90 (2): 433–438.

[114] Gabaix, Xavier and David Laibson. (2001) "The 6D Bias and the Equity-Premium Puzzle." *NBER Macroeconomics Annual*, 16: 257–312.

[115] Gabaix, Xavier and David Laibson. (2005) "Bounded Rationality and Directed Cognition." Manuscript.

[116] Gabaix, Xavier, David Laibson, Guillermo Moloche, and Stephen Weinberg. (2006) "Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model." *American Economic Review*, 96 (4): 1043–1068.

[117] Ghosh, Anisha, Christian Julliard, Alex P. Taylor. (2013) "What is the Consumption-CAPM missing? An Information-Theoretic Framework for the Analysis of Asset Pricing Models." Manuscript.

[118] Gigerenzer, Gerd and Reinhard Selten (Eds.). (2001) *Bounded Rationality: The Adaptive Toolbox*, Cambridge, MA: MIT Press.

[119] Gigerenzer, Gerd, Peter M. Todd, and the ABC Research Group. (2000) *Simple Heuristics that Make Us Smart*, New York, NY: Oxford University Press.

[120] Gilboa, Itzhak and David Schmeidler. (1995) "Case-Based Decision Theory." *Quarterly Journal of Economics*, 110 (3): 605–639.

[121] Gilovich, Thomas, Dale W. Griffin, and Daniel Kahneman (Eds.). (2002) *Heuristics and Biases: The Psychology of Intuitive Judgment*, New York, NY: Cambridge University Press.

[122] Glimcher, Paul W., Colin F. Camerer, Ernst Fehr, Russell A. Poldrack (eds.). (2008) *Neuroeconomics: Decision Making and the Brain*, Academic Press.

[123] Goetzmann, William N. and Alok Kumar. (2008) "Equity Portfolio Diversification", *Review of Finance*, 12: 433–463.

[124] Goetzmann, William N., Lingfeng Li, and K. Geert Rouwenhorst. (2005) "Long-Term Global Market Correlations", *The Journal of Business*, 78(1): 1–38.

[125] Goldstone, Robert L. and Andrew T. Hendrickson. (2010) "Categorical perception", *WIREs Cognitive Science*, 1(1): 69–78.

[126] Gonzalez, Richard and George Wu. (1999) "On the Shape of the Probability Weighting Function", *Cognitive Psychology*, 38: 129–166.

[127] Gray, Robert M. and David L. Neuhoff. (1998) "Quantization." *IEEE Transactions on Information Theory*, 44(6): 2325–2383.

[128] Griffiths, Thomas L., Edward Vul and Adam N. Sanborn. (2012) "Bridging Levels of Analysis for Probabilistic Models of Cognition." *Current Directions in Psychological Science*, 21(4): 263–268.

[129] Guidolin, Massimo and Allan Timmermann. (2007) "Asset Allocation under Multivariate Regime Switching", *Journal of Economic Dynamics and Control*, 31: 3503–3544.

[130] Gul, Faruk. (1991) "A Theory of Disappointment Aversion", *Econometrica*, 59(3): 667–686.

[131] Gurevich, Gregory, Doron Kliger and Ori Levy. (2009) "Decision-making Under Uncertainty — A Field Study of Cumulative Prospect Theory", *Journal of Banking and Finance*, 33: 1221–1229.

[132] Hansen, Lars P. (2007) "Beliefs, Doubts and Learning: Valuing Macroeconomic Risk." *American Economic Review*, 97 (2): 1–30.

[133] Hansen, Lars P. (2014) "Nobel Lecture: Uncertainty Outside and Inside Economic Models", *Journal of Political Economy*, 122(5): 945–987.

[134] Hansen, Lars P., John Heaton, Junghoon Lee and Nikolai Roussanov. (2007) "Intertemporal Substitution and Risk Aversion." In: James J. Heckman and Edward E. Leamer (eds.), *Handbook of Econometrics*, volume 6, part A, chapter 61, pages 3967–4056, Elsevier.

[135] Hansen, Lars P., John C. Heaton and Nan Li. (2008) "Consumption Strikes Back?: Measuring Long-Run Risk." *Journal of Political Economy*, 116(2): 260–302.

[136] Hansen, Lars Peter and Ravi Jagannathan. (1991) "Implications of Security Market Data for Models of Dynamic Economies", *Journal of Political Economy*, 99(2): 225–262.

[137] Hansen, Lars P. and Scott F. Richard. (1987) "The Role of Conditioning Information in Deducing Testable Restrictions Implied by DynamicAsset Pricing Models." *Econometrica*, 55 (3): 587–613.

[138] Hansen, Lars P. and Thomas J. Sargent. (2007a) *Robustness*, Princeton, NJ: Princeton University Press.

[139] Hansen, Lars P. and Thomas J. Sargent. (2007b) "Recursive Robust Estimation and Control Without Commitment." *Journal of Economic Theory*, 136 (1): 1–27.

[140] Hansen, Lars P. and Thomas J. Sargent. (2010) "Fragile Beliefs and the Price of Uncertainty", *Quantitative Economics*, 1: 129–162.

[141] Hansen, Lars P. and Kenneth J. Singleton. (1983) "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns", *Journal of Political Economy*, 91(2): 249–265.

[142] Hastie, Trevor, Robert Tibshirani, Jerome H. Friedman. (2009) *Elements of Statistical Learning: Data Mining, Inference and Prediction*, New York: Springer-Verlag.

[143] He, Xue Dong and Xun Yu Zhou. (2011) "Portfolio Choice Under Cumulative Prospect Theory: An Analytical Treatment", *Management Science*, 57(2): 315–331.

[144] Hens, Thorsten and Christian Reichlin. (2013) "Three Solutions to the Pricing Kernel Puzzle", *Review of Finance*, 17: 1065–1098.

[145] Hertz, John, Anders Krogh and Richard G. Palmer. (1991) *Introduction to the Theory of Neural Computation*, Reading, MA: Perseus.

[146] Hoyer, Patrik O. and Aapo Hyvärinen. (2003) "Interpreting Neural Response Variability as Monte Carlo Sampling of the Posterior." Manuscript.

[147] Hsu, Ming, Ian Krajbich, Chen Zhao, and Colin F. Camerer. (2009) "Neural Response to Reward Anticipation under Risk Is Nonlinear in Probabilities", *Journal of Neuroscience*, 29(7): 2231–2237.

[148] Huang, Yanping and Rajesh P. N. Rao. (2011) "Predictive Coding." *WIREs Cognitive Science*, 2: 580–593.

[149] Jackwerth, Jens Carsten. (2000) "Recovering Risk Aversion from Option Prices and Realized Returns", *Review of Financial Studies*, 13(2): 433–451.

[150] Jagannathan, Ravi and Tongshu Ma (2003). "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps." *Journal of Finance*, 58(4): 1651–1684.

[151] Jame, Russell and Qing Tong. (2014) "Industry-based Style Investing", *Journal of Financial Markets*, 19: 110–130.

[152] James, W. and Charles M. Stein (1961), "Estimation with Quadratic Loss", Proc. Fourth Berkeley Symp. Math. Statist. Prob. 1: 361–379.

[153] Jaynes, Edwin T. (2003) *Probability Theory: The Logic of Science*, Cambridge University Press.

[154] Jorion, Philippe. (1985). "International Portfolio Diversification with Estimation Risk." *Journal of Business*, 58(3): 259–278.

[155] Jorion, Philippe. (1986). "Bayes-Stein Estimation for Portfolio Analysis." *Journal of Financial and Quantitative Analysis*, 21(3): 279–292.

[156] Jehiel, Philippe. (2005) "Analogy-Based Expectation Equilibrium." *Journal of Economic Theory*, 123 (2): 81–104.

[157] Johnstone, Ian M. (2001) "On the Distribution of the Largest Eigenvalue in Principal Components Analysis." *Annals of Statistics*, 29(2): 295–327.

[158] Kahneman, Daniel and Amos Tversky. (1972) "Subjective Probability: A Judgment of Representativeness", *Cognitive Psychology*, 3: 430–454.

[159] Kahneman, Daniel and Amos Tversky. (1979) "Prospect Theory: An Analysis of Decision under Risk", *Econometrica*, 47(2): 263–291.

[160] Kelly, Morgan. (1995) "All Their Eggs in One Basket: Portfolio Diversification of US Households", *Journal of Economic Behavior and Oraanization*, 27: 87–96.

[161] Kepecs, Adam and Zachary F. Mainen. (2012) "A Computational Framework for the Study of Confidence in Humans and Animals." *Philosophical Transactions of the Royal Society B*, 367: 1322–1337.

[162] Kleinberg, Joel and Herbert Kaufman. (1971) "Constancy in Short-term Memory: Bits and Chunks", *Journal of Experimental Psychology*, 90(2): 326–333.

[163] Kliger, Doron and Ori Levy. (2009) "Theories of Choice Under Risk: Insights from Financial Markets", *Journal of Economic Behavior and Organization*, 71: 330–346.

[164] Knill, David C. and Alexandre Pouget. (2004) "The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation." *Trends in Neurosciences*, 27(12): 712–719.

[165] Kumar, Alok. (2009) "Dynamic Style Preferences of Individual Investors and Stock Returns", *Journal of Financial and Quantitative Analysis*, 44(3): 607–640.

[166] Kumar, Alok and Charles M. C. Lee. (2006) "Retail Investor Sentiment and Return Comovements", *The Journal of Finance*, 61(5): 2451–2486.

[167] Kwisthout, Johan and Iris van Rooij. (2013) "Predictive Coding and the Bayesian Brain: Intractability Hurdles That Are Yet To Be Overcome." Manuscript.

[168] Laming, Donald. (2010) "Statistical Information and Uncertainty: A Critique of Applications in Experimental Psychology." *Entropy*, 12: 720–771.

[169] Laughlin, Simon B., John C. Anderson, David O'Carroll and Rob De Ruyter Van Steveninck. (2000) "Coding Efficiency and the Metabolic Cost of Sensory and Neural Information." In: R. Baddeley, P. Hancock and P. Foldiak (eds.), *Information Theory and the Brain*, pages 41–61, New York, NY: Cambridge University Press.

[170] Ledoit, Olivier and Michael Wolf (2003) "Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portofolio Selection." *Journal of Empirical Finance*,

10(5): 603–621.

[171] Ledoit, Olivier and Michael Wolf (2004a) "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices." *Journal of Multivariate Analysis*, 88(2): 365–411.

[172] Ledoit, Olivier and Michael Wolf (2004b) "Honey, I Shrunk the Sample Covariance Matrix." *Journal of Portfolio Management*, 30(4): 110–119.

[173] Ledoit, Olivier and Michael Wolf. (2013) "Spectrum Estimation: A Unified Framework for Covariance Matrix Estimation and PCA in Large Dimensions." Manuscript.

[174] Lee, Daeyeol, Hyojung Seo and Min Whan Jung. (2012) "Neural Basis of Reinforcement Learning and Decision Making." *Annual Review of Neuroscience*, 35: 287–308.

[175] Leung-Yan-Cheong, Sik K. and Thomas M. Cover. (1978) "Some Equivalences Between Shannon Entropy and Kolmogorov Complexity", *IEEE Transactions on Information Theory*, IT-24 (3): 331–338.

[176] Li, Lingfeng. (2002) "Macroeconomic Factors and the Correlation of Stock and Bond Returns". *Yale International Center for Finance Working Paper*, No. 02-46.

[177] Li, Lingfeng. (2003) "An Economic Measure of Diversification Benefits". *Yale International Center for Finance Working Paper*, No. 03-11.

[178] List, John A. (2004) "Neoclassical Theory Versus Prospect Theory: Evidence from the Marketplace", *Econometrica*, 72(2): 615–625.

[179] List, John A.and Michael S. Haigh. (2005) "A Simple Test of Expected Utility Theory Using Professional Traders", *Proceedings of the National Academy of Sciences*, 102(3): 945–948.

[180] Lucas, Robert E., Jr. (1978) "Asset Prices in an Exchange Economy." *Econometrica*, 46 (6): 1429–1445.

[181] Luce, R. Duncan. (1988) "Rank-Dependent, Subjective Expected-Utility Representations", *Journal of Risk and Uncertainty*, 1: 305–332.

[182] Ludvigson, Sydney C. (2013) "Advances in Consumption-Based Asset Pricing: Empirical Tests." In: George M. Constantinides, Milton Harris and Rene M. Stulz (eds.), *Handbook of the Economics of Finance*, volume 2, part B, chapter 12, pages 799–906, Elsevier.

[183] Ma, Wei Ji. (2012) "Organizing Probabilistic Models of Perception." *Trends in Cognitive Sciences*, 16(10): 511–518.

[184] Ma, Wei Ji and Mehrdad Jazayeri. (2014) "Neural Coding of Uncertainty and Probability." *Annual Review of Neuroscience*, 37: 205–220.

[185] MacKay, David J. C. (2003) *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press.

[186] Malloy, Christopher J., Tobias J. Moskowitz and Annette Vissing-Jørgensen (2009). "Long-Run Stockholder Consumption Risk and Asset Returns." *Journal of Finance*, 64(6): 2427–2479.

[187] Mandelbrot, Benoit. (1963) "The Variation of Certain Speculative Prices." *The Journal of Business*, 36(4): 394–419.

[188] Manski, Charles F. (2017) "Survey Measurement of Probabilistic Macroeconomics Expectations: Progress and Promise." Manuscript.

[189] Marschak, Jacob. (1959) "Remarks on the Economics of Information." *Cowles Foundation Discussion Paper*, No. 70.

[190] Marschak, Jacob. (1968) "Economics of Inquiring, Communicating, Deciding." *American Economic Review*, 58(2): 1–18.

[191] Marschak, Jacob. (1971) "Economics of Information Systems." *Journal of the American Statistical Association*, 66(333): 192–219.

[192] Matějka, Filip. (2014a) "Rigid Pricing and Rationally Inattentive Consumer." Manuscript.

[193] Matějka, Filip. (2014b) "Rationally Inattentive Seller: Sales and Discrete Pricing." Manuscript.

[194] Matějka, Filip and Alisdair McKay. (2014) "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model." Manuscript.

[195] Matějka, Filip and Christopher A. Sims. (2010) "Discrete Actions in Information-Constrained Tracking Problems." Manuscript.

[196] McClure, Samuel M., David I. Laibson, George Loewenstein, Jonathan D. Cohen. (2004) "Separate Neural Systems Value Immediate and Delayed Monetary Rewards." *Science*, 306: 503–507.

[197] McGrattan, Ellen R. and Edward C. Prescott (2003) "Average Debt and Equity Returns: Puzzling?" *American Economic Review*, 93: 392–397.

[198] McGrattan, Ellen R. and Edward C. Prescott (2005) "Taxes, Regulations, and the Value of U.S. and U.K. Corporations", *Review of Economic Studies*, 92: 767–796.

[199] Menzly, Lior, Tano Santos and Pietro Veronesi. (2004) "Understanding Predictability." *Journal of Political Economy*, 112(1): 1–47.

[200] Merton, Robert C. (1969) "Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case." *The Review of Economics and Statistics*, 51(3): 247–257.

[201] Migliore, Michele, Gaspare Novara and Domenico Tegolo. (2008) "Single Neuron Binding Properties and the Magical Number 7", *Hippocampus*, 18: 1122–1130.

[202] Miller, George A. (1956). "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review*, 63(2): 81–97.

[203] Miller, Earl and Jonathan Wallis. (2008) "The Prefrontal Cortex and Executive Brain Functions." In: Larry R. Squire et al. (eds.), *Fundamental Neuroscience*, pages 1199–1222, 3rd ed., Academic Press.

[204] Mullainathan, Sendhil. (2002a) "A Memory-Based Model of Bounded Rationality." *Quarterly Journal of Economics*, 117 (3): 735–774.

[205] Mullainathan, Sendhil. (2002b) "Thinking Through Categories." Manuscript.

[206] Murphy, Kevin P. (2012) *Machine Learning — a Probabilistic Perspective*, MIT Press.

[207] Nieder, Andreas. (2005) "Counting on Neurons: The Neurobiology of Numerical Competence." *Nature Reviews Neuroscience*, 6: 177–190.

[208] Nieder, Andreas and Stanislas Dehaene. (2009) "Representation of Number in the Brain." *Annual Review of Neuroscience*, 32: 185–208.

[209] Noussair, Charles N. and Bodo Vogt. (2013) "The Influence of Probability Format on Elicited Certainty Equivalents", *Progress in Brain Research*, 202: 151–171.

[210] Olshausen, Bruno A. and Field, David J. (1996) "Emergence of Simple-CellReceptive Field Properties by Learning a Sparse Code for Natural Images." *Nature*, 381: 607–609.

[211] Olshausen, Bruno A. and Field, David J. (2005) "How Close AreWe to Understanding V1?" *Neural Computation*, 17: 1665–1699.

[212] O'Reilly, Jill X., Saad Jbabdi, Matthew F. S. Rushworth, Timothy E. J. Behrens. (2013) "Brain Systems for Probabilistic and Dynamic Prediction: Computational Specificity and

Integration." *PLOS Biology*, 11(9): 1–14.

[213] Ortega, Pedro A. and Daniel A. Braun. (2013) "Thermodynamics as a Theory of Decision-Making with Information-Processing Costs", *Proceedings of the Royal Society A*, 469: 20120683.

[214] Owen, Adrian M. (2004) "Working Memory: Imaging the Magic Number Four", *Current Biology*, 14: 573–574.

[215] Padoa-Schioppa, Camillo, Aldo Rustichini. (2014) "Rational Attention and Adaptive Coding: A Puzzle and a Solution." *American Economic Review*, 104(5): 507–513.

[216] Pammi, V.S. Chandrasekhar and Narayanan Srinivasan. (2013) "Preface", *Progress in Brain Research*, 202: xi.

[217] Parker, Jonathan A. and Christian Julliard. (2005) "Consumption Risk and the Cross Section of Expected Returns." *Journal of Political Economy*, 113(1): 185–222.

[218] Peng, Lin and Wei Xiong. (2006) "Investor Attention, Overconfidence and Category Learning." *Journal of Financial Economics*, 80 (3): 563–602.

[219] Piazza, Manuela, Philippe Pinel, Denis Le Bihan and Stanislas Dehaene. (2007) "A Magnitude Code Common to Numerosities and Number Symbols in Human Intraparietal Cortex." *Neuron*, 53: 293–305.

[220] Pindyck, Robert S. and Julio J. Rotemberg. (1990) "The Excess Co-Movement of Commodity Prices", *The Economic Journal*, 100(403): 1173–1189.

[221] Platt, Michael L. and Scott A. Huettel. (2008) "Risky Business: The Neuroeconomics of Decision Making Under Uncertainty." *Nature Neuroscience*, 11(4): 398–403.

[222] Poggio, Tomaso. (2012) "The Levels of Understanding framework, revised." *Perception*, 41(9): 1017–1023.

[223] Pouget, Alexandre, Jeffrey M. Beck, Wei Ji Ma and Peter E. Latham. (2013) "Probabilistic Brains: Knowns and Unknowns." *Nature Neuroscience*, 16(9): 1170–1178.

[224] Pouget, Alexandre and Terrence J. Sejnowski. (1997) "Spatial Transformations in the Parietal Cortex Using Basis Functions." *Journal of Cognitive Neuroscience*, 9(2): 222–237.

[225] Polkovnichenko, Valery. (2005) "Household Portfolio Diversification: A Case for Rank-Dependent Preferences", *Review of Financial Studies*, 18(4): 1467–1502.

[226] Polkovnichenko, Valery and Feng Zhao. (2013) "Probability Weighting Functions Implied in Options Prices", *Journal of Financial Economics*, 107: 580–609.

[227] Post, Thierry and Haim Levy. (2005) "Does Risk Seeking Drive Stock Prices? A Stochastic Dominance Analysis of Aggregate Investor Preferences and Beliefs", *Review of Financial Studies*, 18(3): 925–953.

[228] Prelec, Drazen. (1998) "The Probability Weighting Function", *Econometrica*, 66(3): 497–527.

[229] Quiggin, John. (1982) "A Theory of Anticipated Utility", *Journal of Economic Behavior and Organization*, 3: 323–343.

[230] Rangel, Antonio, Colin Camerer and P. Read Montague. (2008) "A Framework for Studying the Neurobiology of Value-Based Decision Making." *Nature Reviews Neuroscience*, 9: 545–556.

[231] Rao, Rajesh P. N. (2010) "Decision Making Under Uncertainty: a Neural Model Based on Partially Observable Markov Decision Processes." *Frontiers in Computational Neuroscience*, 4(146): 1–18.

[232] Ravid, Doron. (2016) "Bargaining with Rational Inattention." Manuscript.

[233] Reynolds, John H., Jacqueline P. Gottlieb, and Sabine Kastner. (2008) "Attention." In: Larry R. Squire et al. (eds.), *Fundamental Neuroscience*, pages 1113–1132, 3rd ed., Academic Press.

[234] Reznik, Yuriy A. (2010) "Quantization of Discrete Probability Distributions." Manuscript.

[235] Rietz, Thomas. (1988) "The Equity Risk Premium: A Solution." *Journal of Monetary Economics*, 22: 117–131.

[236] Rissanen, Jorma. (2007) *Information and Complexity in Statistical Modeling*, Springer.

[237] Rosenberg, Joshua V. and Robert F. Engle. (2002) "Empirical Pricing Kernels", *Journal of Financial Economics*, 64: 341–372.

[238] Routledge, Bryan R. and Stanley E. Zin. (2010) "Generalized Disappointment Aversion and Asset Prices", *Journal of Finance*, 65(4): 1303–1332.

[239] Rushworth, Matthew F. S. and Timothy E. J. Behrens. (2008) "Choice, Uncertainty and Value in Prefrontal and Cingulate Cortex." *Nature Neuroscience*, 11(4): 389–397.

[240] Rustichini, Aldo and Camillo Padoa-Schioppa. (2015) "A Neuro-Computational Model of Economic Decisions." *Journal of Neurophysiology*, 114(5): 1382–1398.

[241] Samuelson, Paul A. (1969) "Lifetime Portfolio Selection By Dynamic Stochastic Programming." *The Review of Economics and Statistics*, 51(3): 239–246.

[242] Scheinkman, José A. and Wei Xiong. (2003) "Overconfidence and Speculative Bubbles." *Journal of Political Economy*, 111 (6): 1183–1219.

[243] Schmeidler, David. (1989) "Subjective Probability and Expected Utility without Additivity", *Econometrica*, 57(3): 571–587.

[244] Schultz, Wolfram. (2008) "Introduction. Neuroeconomics: the Promise and the Profit", *Philosophical Transactions of the Royal Society B*, 363: 3767–3769.

[245] Schultz, Wolfram, Peter Dayan and P. Read Montague. (1997) "A Neural Substrate of Prediction and Reward." *Science*, 275: 1593–1599.

[246] Schultz, Wolfram, Kerstin Preuschoff, Colin Camerer, Ming Hsu, Christopher D. Fiorillo, Philippe N. Tobler and Peter Bossaerts. (2008) "Explicit Neural Signals Reflecting Reward Uncertainty." *Philosophical Transactions of the Royal Society B*, 363: 3801–3811.

[247] Shannon, Claude E. (1948) "A Mathematical Theory of Communication." *Bell System Technical Journal*, 27: 379–423, 623–656.

[248] Sharpe, William F. (1992) "Asset Allocation: Management Style and Performance Measurement", *Journal of Portfolio Management*, 18(2): 7–19.

[249] Shiffrin, Richard M. and Robert M. Nosofsky. (1994) "Seven Plus or Minus Two: A Commentary On Capacity Limitations", *Psychological Review*, 101(2): 357–361.

[250] Shiller, Robert J. (1989) "Comovements in Stock Prices and Comovements in Dividends", *The Journal of Finance*, 44(3): 719–729.

[251] Shiller, Robert J. and Andrea E. Beltratti. (1992) "Stock Prices and Bond Yields: Can their comovements be explained in terms of present value models?", *Journal of Monetary Economics*, 30: 25–46.

[252] Simon, Herbert A. (1997) *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations*, 4th edn, New York: Free Press.

[253] Simon, Herbert A. (1957) *Models of Man: Social and Rational*, New York: John Wiley and Sons.

[254] Simoncelli, Eero P. and Bruno A. Olshausen. (2001) "Natural Image Statistics and Neural

Representation." *Annual Review of Neuroscience*, 24: 1193–1216.

[255] Sims, Christopher A. (1998) "Stickiness." *Carnegie-Rochester Conference Series on Public Policy*, 49(1): 317–356.

[256] Sims, Christopher A. (2003) "Implications of Rational Inattention." *Journal of Monetary Economics*, 50: 665–690.

[257] Sims, Christopher A. (2006) "Rational Inattention: A Research Agenda." Manuscript.

[258] Sims, Christopher A. (2010) "Rational Inattention and Monetary Economics." In: Benjamin M. Friedman and Michael Woodford (eds.), *Handbook of Monetary Economics*, edition 1, volume 3, chapter 4, pages 155–181, Elsevier.

[259] Smith, Edward E. and John Jonides. (2003) "Executive Control and Thought." In: Larry R. Squire, James L. Roberts, Nicholas C. Spitzer, Michael J. Zigmond, Susan K. McConnell and Floyd E. Bloom (eds.), *Fundamental Neuroscience*, pages 1377–1394, 2nd ed., Academic Press.

[260] Spratling, Michael W. (2012) "Unsupervised Learning of Generative and Discriminative Weights Encoding Elementary Image Components in a Predictive Coding Model of Cortical Function." *Neural Computation*, 24: 60–103.

[261] Squire, Larry R., Floyd Bloom, Nicholas C. Spitzer, Sascha du Lac, Anirvan Ghosh, Darwin Berg (eds.). (2008) *Fundamental Neuroscience*, 3rd ed., Academic Press.

[262] Statman, Meir. (1987) "How Many Stocks Make a Diversified Portfolio?", *The Journal of Financial and Quantitative Analysis*, 22(3): 353–363.

[263] Stein, Charles M. (1956), "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Distribution", Proc. Third Berkeley Symp. Math. Statist. Prob. 1: 197–206.

[264] Stein, Charles M. (1975) "Estimation of a Covariance Matrix", Rietz Lecture, 39th Annual Meeting of the Institute of Mathematical Statistics. Atlanta, GA.

[265] Stein, Charles M. (1986) "Lectures on the Theory of Estimation of Many Parameters." *Journal of Soviet Mathematics*, 34(1): 1373–1403.

[266] Steiner, Jakub and Colin Stewart (2016) "Perceiving Prospects Properly." Manuscript.

[267] Still, Susanne, David A. Sivak, Anthony J. Bell and Gavin E. Crooks. (2012) "Thermodynamics of Prediction", *Physical Review Letters*, 109(120604).

[268] Stott, Henry P. (2006) "Cumulative Prospect Theory's Functional Menagerie", *Journal of Risk and Uncertainty*, 32(2): 101–130.

[269] Stutzer, Michael. (1995) "A Bayesian Approach to Diagnosis of Asset Pricing Models", *Journal of Econometrics*, 68: 367–397.

[270] Stutzer, Michael. (1996) "A Simple Nonparametric Approach to Derivative Security Valuation", *Journal of Finance*, 51(5): 1633–1652.

[271] Summerfield, Christopher and Tobias Egner. (2009) "Expectation (and Attention) in Visual Cognition." *Trends in Cognitive Sciences*, 13(9): 403–409.

[272] Sutton, Richard S. and Andrew G. Barto. (1998) *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press.

[273] Tanaka, Saori C., Kenji Doya, Go Okada, Kazutaka Ueda,Yasumasa Okamoto and Shigeto Yamawaki. (2004) "Prediction of Immediate and Future Rewards Differentially Recruits Cortico-Basal Ganglia Loops." *Nature Neuroscience*, 7(8): 887–893.

[274] Tao, Terence. (2010) "Sumset and Inverse Sumset Theory for Shannon Entropy." *Combinatorics, Probability and Computing*, 19: 603–639.

[275] Tao, Terence and Van Vu. (2006) "Entropy Methods." Manuscript.

[276] Teo, Melvyn, Sung-Jun Woo. (2004) "Style Effects in the Cross-Section of Stock Returns", *Journal of Financial Economics*, 74: 367–398.

[277] Tijsseling, Adriaan and Harnad, Stevan. (1997) "Warping Similarity Space in Category Learning by Backprop Nets." In: Ramscar, M., Hahn, U., Cambouropolos, E. & Pain, H. (eds.) Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization. Department of Artificial Intelligence, Edinburgh University: 263–269.

[278] Tsai, Jerry and Jessica A. Wachter. (2015) "Disaster Risk and Its Implications for Asset Pricing", *Annual Review of Financial Economics*, 7: 219–252.

[279] Tsai, Jerry and Jessica A. Wachter. (2016) "Rare Booms and Disasters in a Multisector Endowment Economy", *Review of Financial Studies*, 29(5): 1113–1169.

[280] Tudusciuc, Oana and Andreas Nieder. (2007) "Neuronal Population Coding of Continuous and Discrete Quantity in the Primate Posterior Parietal Cortex." *Proceedings of the National Academy of Sciences*, 104(36): 14513–14518.

[281] Tudusciuc, Oana and Andreas Nieder. (2009) "Contributions of Primate Prefrontal and Posterior Parietal Cortices to Length and Numerosity Representation." *Journal of Neurophysiology*, 101: 2984–2994.

[282] Tversky, Amos and Daniel Kahneman. (1971) "Belief in the Law of Small Numbers", *Psychological Bulletin*, 76(2): 105–110.

[283] Tversky, Amos and Daniel Kahneman. (1974) "Judgment under Uncertainty: Heuristics and Biases", *Science*, 185(4157): 1124–1131.

[284] Tversky, Amos and Daniel Kahneman. (1992) "Advances in Prospect Theory: Cumulative Representation of Uncertainty", *Journal of Risk and Uncertainty*, 5: 297–323.

[285] Van Nieuwerburgh, Stijn and Laura Veldkamp. (2009) "Information Immobility and the Home Bias Puzzle", *Journal of Finance*, 64(3): 1187–1215.

[286] Van Nieuwerburgh, Stijn and Laura Veldkamp. (2010) "Information Acquisition and Under-Diversification", *Review of Economic Studies*, 77: 779–805.

[287] Veldkamp, Laura L. (2006) "Information Markets and the Comovement of Asset Prices", *Review of Economic Studies*, 73: 823–845.

[288] Veronesi, Pietro. (2004) "The Peso Problem Hypothesis and Stock Market Returns", *Journal of Economic Dynamics and Control*, 28: 707–725.

[289] Vilares, Iris, James D. Howard, Hugo L. Fernandes, Jay A. Gottfried and Konrad P. Kording. (2012) "Differential Representations of Prior and Likelihood Uncertainty in the Human Brain." *Current Biology*, 22(18): 1641–1648.

[290] Wachter, Jessica A. (2010) "Asset Allocation", *Annual Review of Financial Economics*, 2: 175–206.

[291] Wahal, Sunil and M. Deniz Yavuz. (2013) "Style Investing, Comovement and Return Predictability", *Journal of Financial Economics*, 107: 136–154.

[292] Weil, Philippe. (1989) "The Equity Premium Puzzle and the Risk-Free Rate Puzzle." *Journal of Monetary Economics*, 24: 401–421.

[293] Weitzman, Martin L. (2007) "Subjective Expectations and Asset-Return Puzzles." *American Economic Review*, 97 (4): 1102–1130.

[294] Wilson, Andrea. (2014) "Bounded Memory and Biases in Information Processing." *Econo-*

metrica, 82 (6): 2257–2294.

[295] Won, Joong-Ho, Johan Lim, Seung-Jean Kim and Bala Rajaratnam. (2012) "Condition Number Regularized Covariance Estimation." Manuscript.

[296] Woodford, Michael. (2014) "An Optimizing Neuroeconomic Model of Discrete Choice." Manuscript.

[297] Woodford, Michael. (2012) "Inattentive Valuation and Reference-Dependent Choice." Manuscript.

[298] Wu, George and Richard Gonzalez. (1996) "Curvature of the Probability Weighting Function", *Management Science*, 42(12): 1676–1690.

[299] Xia, Jianming and Xun Yu Zhou. (2012) "Arrow-Debreu Equilibria for Rank-Dependent Utilities." Manuscript.

[300] Yaari, Menahem E. (1987) "The Dual Theory of Choice under Risk", *Econometrica*, 55(1): 95–115.

[301] Yang, Jian, Yinggang Zhou, Zijun Wang. (2009) "The Stock-Bond Correlation and Macroeconomic Conditions: One and a Half Centuries of Evidence", *Journal of Banking and Finance*, 33: 670–680.

[302] Yang, Tianming and Michael N. Shadlen. (2007) "Probabilistic Reasoning by Neurons." *Nature*, 447: 1075–1082.

[303] Zhaoping, Li. (2006) "Theoretical Understanding of the Early Visual Processes by Data Compression and Data Selection." *Network: Computation in Neural Systems*, 17: 301–334.

[304] Ziegler, Alexandre. (2007) "Why Does Implied Risk Aversion Smile?" *The Review of Financial Studies*, 20(3): 859–904.