ISSN 2279-9362

Collegio Carlo Alberto

Rational Inattention and Rate Distortion Theory: A Teaching Note

Tommaso Denti, Massimo Marinacci and Luigi Montrucchio

No. 574
December 2018

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

Rational Inattention and Rate Distortion Theory: A Teaching Note*

Tommaso Denti[a] Massimo Marinacci[b] Luigi Montrucchio[c]

[a]Department of Economics, Cornell University

[b]Department of Decision Sciences and IGIER, Università Bocconi

[c]Collegio Carlo Alberto, Università di Torino

October 23, 2018

Abstract

In this teaching note we discuss the relation between rational inattention and a major branch of information theory called "rate distortion theory." Focusing on methods, we translate tools from rate distortion theory into the language of rational inattention. These tools provide an alternative, more primitive, approach to the study of optimal attention allocation.

1 Introduction

Rational inattention, due to Sims (2003), borrows ideas from information theory to study how economic agents allocate attention. It became one of the leading models of costly information acquisition in economics.[1] Cost of information is measured by Shannon's mutual information, the main measure of informational content used in information theory.

This note discusses the relation between rational inattention and a branch of information theory called "rate distortion theory." We explain how tools from rate distortion theory can be used to study rational inattention. Discussions about rational inattention and information theory usually focus on foundational aspects, e.g., does information theory provide a foundation for rational inattention? Here we abstract away from those concerns and focus on methods.

The formal relation between rational inattention and rate distortion theory is easy to explain: the two fields study optimization problems that are dual to each other. In rational inattention, the allocation of attention is represented by the maximization of expected utility subject to a cost proportional to mutual information. Rate distortion theory studies lossy data compression: the minimization of mutual information subject to an upper bound on the expected loss from compressing the data. By thinking of loss functions as the negative of utility functions, we see that the problems of optimal attention allocation and lossy data compression are dual to each other.

This note analyzes an abstract version of these optimization problems. In particular, we borrow from Gallager (1968, 1972) and, especially, Csiszar (1974a, 1974b). Their methods lead to a primitive approach not based on Lagrange multipliers, which are the main technique used in applications.

*We wish to thank Roberto Corrao for some very useful comments and the financial support of ERC (grant INDIMACRO).

[1]See Mackowiak et al. (2018) for a recent survey of rational inattention.

Avoiding Lagrange multipliers is particularly useful for settings with continuous variables where the regularity conditions required by infinite-dimensional convex programming may be hard to verify.

The relation to rate distortion theory is known to researchers in rational inattention. Methods from rate distortion theory, however, have not been fully integrated into the rational inattention literature yet. The contribution of this note is to make these methods more accessible to economists, in particular to graduate students, interested in rational inattention.

We organize this note in two sections. Section 2 presents a purely mathematical analysis, abstracting away from applications, while Section 3 explains how the mathematical results apply to rational inattention and rate distortion theory.

2 Theory

2.1 Preliminaries

Spaces Let $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ be Polish spaces with Borel sigma-algebras $\mathcal{X}$ and $\mathcal{Y}$, respectively. Unless otherwise stated, we denote by x and y generic elements of X and Y, and by A and B generic elements of the sigma algebras $\mathcal{X}$ and $\mathcal{Y}$, respectively.

We endow $(X, \mathcal{X})$ with a (Borel) probability measure $\mu : \mathcal{X} \to [0, 1]$. We denote by $\Delta(Y)$ the set of all probability measures on $(Y, \mathcal{Y})$, with generic element $\nu : \mathcal{Y} \to [0, 1]$. We equip $\Delta(Y)$ with the weak topology and the corresponding Borel sigma-algebra. For $\nu, \nu' \in \Delta(Y)$, we write $\nu \ll \nu'$ when ν is *absolutely continuous* with respect to ν', i.e., $\nu'(B) = 0$ implies $\nu(B) = 0$ for all $B \in \mathcal{Y}$.

Barycenters and kernels A *stochastic kernel* $\sigma : X \to \Delta(Y)$ is a measurable map from $(X, \mathcal{X})$ into the Polish space $\Delta(Y)$. We denote by Σ the collection of all these kernels; its elements will be written as $\sigma(\cdot|x)$. Define $\mu \otimes \sigma \in \Delta(X \times Y)$ by

$$(\mu \otimes \sigma)(A \times B) = \int_A \sigma(B|x) d\mu(x).$$

The *barycenter* $\nu_\sigma \in \Delta(Y)$ of kernel σ is the probability measure

$$\nu_\sigma(B) = \int_X \sigma(B|x) d\mu(x),$$

So, it is the marginal of $\mu \otimes \sigma$ on Y. We denote by $\mu \times \nu$ the product measure on $X \times Y$ of marginals μ and ν.

Let $k : X \times Y \to \mathbb{R}$ be a measurable function. Throughout the paper we assume that:

1. k is bounded above, that is, $\sup_{x,y} k(x, y) < \infty$,
2. $k(\cdot, y)$ is μ-integrable for some $y \in Y$.[2]

The kernel σ^ν associated with a probability measure $\nu \in \Delta(Y)$ is defined by

$$\sigma^\nu(B|x) = K(x; \nu) \int_B e^{k(x,y)} d\nu(y),$$

where $K(x; \nu) = 1/\int_Y e^{k(x,y)} d\nu(y)$ is a normalizing constant. Kernel σ^ν is well defined: by Tonelli's Theorem, for every B the function $x \mapsto \int_B e^{k(x,y)} d\nu(y)$ is measurable, and so $x \mapsto \sigma^\nu(B|x)$ is a measurable map. For every x the probability measures $\sigma^\nu(\cdot|x)$ and ν are equivalent, with density $d\sigma^\nu(\cdot|x)/d\nu = K(x; \nu) e^{k(x,\cdot)}$.

[2] As it will be seen later, this condition ensures that $\sup_\sigma V(\sigma) > -\infty$.

Entropy For $\nu, \nu' \in \Delta(Y)$, the *relative entropy* $D(\nu\|\nu')$ is defined by

$$D(\nu\|\nu') = \begin{cases} \int_Y \left(\log \frac{d\nu}{d\nu'}\right) d\nu & \text{if } \nu \ll \nu' \\ +\infty & \text{else} \end{cases}$$

where the convention $\log 0 = -\infty$ is adopted. For kernels $\sigma, \sigma' \in \Sigma$, we have (see Lemma 1.4.3 of Dupuis and Ellis, 1997)

$$D(\mu \otimes \sigma\|\mu \otimes \sigma') = \int_X D(\sigma(\cdot|x)\|\sigma'(\cdot|x))d\mu(x).$$

This rule, known in information theory as the "chain rule," relies on both spaces X and Y being Polish.

2.2 The theorem

We consider the concave functional $U : \Sigma \to \mathbb{R} \cup \{-\infty\}$ given by[3]

$$U(\sigma) = \int_{X\times Y} k(x,y)d(\mu \otimes \sigma)(x,y) - D(\mu \otimes \sigma\|\mu \times \nu_\sigma)$$

as well as the concave functional $V : \Delta(Y) \to \mathbb{R} \cup \{-\infty\}$ given by

$$V(\nu) = \int_X \left(\log \int_Y e^{k(x,y)} d\nu(y)\right) d\mu(x) = -\int_X \log K(x;\nu) d\mu(x)$$

Next we state the main result of the note, which relates the variational problems based on these two functionals.

Theorem 1 *We have*

$$\sup_\sigma U(\sigma) = \sup_\nu V(\nu).$$

If kernel $\hat\sigma$ solves $\sup_\sigma U(\sigma)$, then its barycenter $\nu_{\hat\sigma}$ solves $\sup_\nu V(\nu)$ and $\hat\sigma(\cdot|x) = \sigma^{\nu_{\hat\sigma}}(\cdot|x)$ for μ-almost all x. Conversely, if probability measure $\hat\nu$ solves $\sup_\nu V(\nu)$, then its kernel $\sigma^{\hat\nu}$ solves $\sup_\sigma U(\sigma)$ and $\hat\nu = \nu_{\sigma^{\hat\nu}}$.

According to this theorem, we thus have

$$\sup_\sigma \int_{X\times Y} k\, d(\mu \otimes \sigma) - D(\mu \otimes \sigma\|\mu \times \nu_\sigma) = \sup_\nu \int_X \left(\log \int_Y e^k d\nu\right) d\mu.$$

Therefore, the two variational problems are equivalent in terms of value attainment. Moreover, if we identify kernels that are equal μ-a.e., then the *barycenter map* $\sigma \mapsto \nu_\sigma$ is a bijection between the solution sets $\arg\max_\sigma U(\sigma)$ and $\arg\max_\nu V(\nu)$, and its inverse is the *kernel map* $\nu \mapsto \sigma^\nu$.[4]

We can diagram these relations between the solution sets as follows:

$$\arg\max_\sigma U(\sigma) \quad \underset{\sigma=\sigma^\nu}{\overset{\nu=\nu_\sigma}{\underset{\longleftarrow}{\overset{\longrightarrow}{\text{1:1}}}}} \quad \arg\max_\nu V(\nu).$$

Summing up, this beautiful theorem proves that two, *prima facie* altogether different, variational problems are equivalent in terms of value attainment and that their solutions are in a one-to-one correspondence via two natural maps, one the inverse of the other.

[3] Concavity is a consequence of the convexity of the mapping $(\nu, \nu') \to D(\nu\|\nu')$.

[4] We have not proved yet that the solutions sets are nonempty. We will do it later on in Proposition 7 under additional regularity assumptions.

2.3 Proof

We consider the functional $W : \Sigma \times \Delta(Y) \to \mathbb{R} \cup \{-\infty\}$ given by

$$W(\sigma, \nu) = \int_{X \times Y} k(x, y) d(\mu \otimes \sigma)(x, y) - D(\mu \otimes \sigma \| \mu \times \nu).$$

The functional W is related to U and V by the following key identities.

Lemma 2 *If $\sigma(\cdot|x) \ll \nu$ for μ-almost all x, then*

$$W(\sigma, \nu) = U(\sigma) - D(\nu_\sigma \| \nu) \tag{1}$$

$$= V(\nu) - D(\mu \otimes \sigma \| \mu \otimes \sigma^\nu). \tag{2}$$

Proof Let $W(\sigma, \nu) = a - b$, where $b = D(\mu \otimes \sigma \| \mu \times \nu)$ and

$$a = \int_{X \times Y} k(x, y) d(\mu \otimes \sigma)(x, y) = \int_X \left(\int_Y k(x, y) d\sigma(y|x) \right) d\mu(x).$$

Since $\sigma(\cdot|x) \ll \nu$ for μ-almost all x, we have

$$b = \int_X D(\sigma(\cdot|x) \| \nu) d\mu(x) = \int_X \left(\int_Y \log \left(\frac{d\sigma(\cdot|x)}{d\nu} \right) d\sigma(\cdot|x) \right) d\mu(x).$$

Identity (1) follows from

$$\begin{aligned}
b &= \int_X \left(\int_Y \log \left(\frac{d\sigma(\cdot|x)}{d\sigma^\nu(\cdot|x)} \frac{d\sigma^\nu(\cdot|x)}{d\nu} \right) d\sigma(\cdot|x) \right) d\mu(x) \\
&= \int_X \left(\int_Y \log \left(\frac{d\sigma(\cdot|x)}{d\sigma^\nu(\cdot|x)} \right) d\sigma(\cdot|x) \right) d\mu(x) + \int_X \left(\int_Y \log \left(\frac{d\sigma^\nu(\cdot|x)}{d\nu} \right) d\sigma(\cdot|x) \right) d\mu(x) \\
&= \int_X D(\sigma(\cdot|x) \| \sigma^\nu(\cdot|x)) d\mu(x) + \int_X \left(\int_Y \log \left(K(x; \nu) e^{k(x, \cdot)} \right) d\sigma(\cdot|x) \right) d\mu(x) \\
&= D(\mu \otimes \sigma \| \mu \otimes \sigma^\nu) - V(\nu) + a.
\end{aligned}$$

Identity (2) follows from

$$\begin{aligned}
b &= \int_X \left(\int_Y \log \left(\frac{d\sigma(\cdot|x)}{d\nu_\sigma} \frac{d\nu_\sigma}{d\nu} \right) d\sigma(\cdot|x) \right) d\mu(x) \\
&= \int_X \left(\int_Y \log \left(\frac{d\sigma(\cdot|x)}{d\nu_\sigma} \right) d\sigma(\cdot|x) \right) d\mu(x) + \int_X \left(\int_Y \log \left(\frac{d\nu_\sigma}{d\nu} \right) d\sigma(\cdot|x) \right) d\mu(x) \\
&= \int_X D(\sigma(\cdot|x) \| \nu_\sigma) d\mu(x) + \int_Y \log \left(\frac{d\nu_\sigma}{d\nu} \right) d\nu_\sigma \\
&= D(\mu \otimes \sigma \| \mu \times \nu_\sigma) + D(\nu_\sigma \| \nu).
\end{aligned}$$

■

Building on (1) and (2), the next lemma generalizes a classic formula.

Lemma 3 *We have*

$$V(\nu) = W(\sigma^\nu, \nu) = \sup_\sigma W(\sigma, \nu). \tag{3}$$

If probability measure $\hat{\nu}$ solves $\sup_\nu V(\nu)$, then $\hat{\nu} = \nu_{\sigma^{\hat{\nu}}}$.

This lemma generalizes the well-known variational formula for relative entropy (cf. p. 27 of Dupuis and Ellis, 1997):

$$\log \int_Y e^{f(y)} d\nu(y) = \sup_{\nu'} \left\{ \int_Y f(y) d\nu'(y) - D(\nu' \| \nu) \right\}$$

for every bounded Borel measurable function $f : Y \to \mathbb{R}$. Indeed, (3) reduces to this classic formula when X is a singleton, so that we can identify Σ with $\Delta(Y)$ and k with f. This variational formula plays an important role in decision theory (see, e.g., Hansen and Marinacci, 2016).

Proof If the set $\{x : \sigma(\cdot|x) \not\ll \nu\}$ is not μ-null, then $D(\mu \otimes \sigma \| \mu \times \nu) = \infty$, and so $V(\nu) \geq W(\sigma, \nu) = -\infty$. If instead, μ-a.e., $\sigma(\cdot|x) \ll \nu$, then by (1) we have

$$V(\nu) = W(\sigma, \nu) + D(\mu \otimes \sigma \| \mu \otimes \sigma^{\nu}) \geq W(\sigma, \nu).$$

Overall, $V(\nu) \geq \sup_{\sigma} W(\sigma, \nu)$ and the supremum is attained by the kernels for which, μ-a.e., $\sigma(\cdot|x) = \sigma^{\nu}(\cdot|x)$. Actually, $W(\sigma^{\nu}, \nu) = V(\nu)$.

Now let $\hat{\nu}$ solve $\sup_{\nu} V(\nu)$. We observe that $V(\hat{\nu})$ is finite. Indeed, if ν puts probability 1 on the y such that $k(\cdot, y)$ is μ-integrable, then

$$-\infty < \int_X k(x, y) d\mu(x) = V(\nu) \leq V(\hat{\nu}).$$

By (1) and (2), we can write

$$\begin{aligned} V(\hat{\nu}) = W(\sigma^{\hat{\nu}}, \hat{\nu}) &= U(\sigma^{\hat{\nu}}) - D(\nu_{\sigma^{\hat{\nu}}} \| \hat{\nu}) = W(\sigma^{\hat{\nu}}, \nu_{\sigma^{\hat{\nu}}}) - D(\nu_{\sigma^{\hat{\nu}}} \| \hat{\nu}) \\ &= V(\nu_{\sigma^{\hat{\nu}}}) - D(\nu_{\sigma^{\hat{\nu}}} \| \hat{\nu}) - D(\mu \otimes \sigma^{\hat{\nu}} \| \mu \otimes \sigma^{\nu_{\sigma^{\hat{\nu}}}}). \end{aligned}$$

Since $V(\hat{\nu}) \geq V(\nu_{\sigma^{\hat{\nu}}})$ and $V(\hat{\nu})$ is finite, we must have $D(\nu_{\sigma^{\hat{\nu}}} \| \hat{\nu}) = 0$, that is, $\hat{\nu} = \nu_{\sigma^{\hat{\nu}}}$ (see Lemma 1.4.1 of Dupuis and Ellis, 1997). ■

The next lemma mirrors the last one.

Lemma 4 *We have*

$$U(\sigma) = W(\sigma, \nu_{\sigma}) = \sup_{\nu} W(\sigma, \nu).$$

If kernel $\hat{\sigma}$ solves $\sup_{\sigma} U(\sigma)$, then $\hat{\sigma}(\cdot|x) = \sigma^{\nu_{\hat{\sigma}}}(\cdot|x)$ for μ-almost all x.

Proof If the set $\{x : \sigma(\cdot|x) \not\ll \nu\}$ is not μ-null, then $D(\mu \otimes \sigma \| \mu \times \nu) = \infty$, and so $U(\sigma) \geq W(\sigma, \nu) = -\infty$. If instead, μ-a.e., $\sigma(\cdot|x) \ll \nu$, then by (1) we have

$$U(\sigma) = W(\sigma, \nu) + D(\nu_{\sigma} \| \nu) \geq W(\sigma, \nu).$$

Overall, $U(\sigma) \geq \sup_{\nu} W(\sigma, \nu)$ and the supremum is attained by the barycenter of σ and so $W(\sigma, \nu_{\sigma}) = U(\sigma)$.

Now let $\hat{\sigma}$ solve $\sup_{\sigma} U(\sigma)$. We observe that $U(\hat{\sigma})$ is finite. Indeed, if $\mu \otimes \sigma = \mu \times \nu_{\sigma}$ and ν_{σ} puts probability 1 on the y such that $k(\cdot, y)$ is μ-integrable, then

$$-\infty < \int_X k(x, y) d\mu(x) = U(\sigma) \leq U(\hat{\sigma}).$$

In particular, $D(\mu \otimes \hat{\sigma} \| \mu \times \nu_{\hat{\sigma}}) < \infty$, which implies $\hat{\sigma}(\cdot|x) \ll \nu_{\hat{\sigma}}$ for μ-almost all x. By (1) and (2) we can write (being $\hat{\sigma}(\cdot|x) \ll \nu_{\hat{\sigma}}$, μ-a.e.)

$$\begin{aligned} U(\hat{\sigma}) = W(\hat{\sigma}, \nu_{\hat{\sigma}}) = V(\nu_{\hat{\sigma}}) - D(\mu \otimes \hat{\sigma} \| \mu \otimes \sigma^{\nu_{\hat{\sigma}}}) &= W(\sigma^{\nu_{\hat{\sigma}}}, \nu_{\hat{\sigma}}) - D(\mu \otimes \hat{\sigma} \| \mu \otimes \sigma^{\nu_{\hat{\sigma}}}) \\ &= U(\sigma^{\nu_{\hat{\sigma}}}) - D(\nu_{\sigma^{\nu_{\hat{\sigma}}}} \| \nu_{\hat{\sigma}}) - D(\mu \otimes \hat{\sigma} \| \mu \otimes \sigma^{\nu_{\hat{\sigma}}}). \end{aligned}$$

Since $U(\hat{\sigma}) \geq U(\sigma^{\nu_{\hat{\sigma}}})$ and $U(\hat{\sigma})$ is finite, we must have $D(\mu \otimes \hat{\sigma} \| \mu \otimes \sigma^{\nu_{\hat{\sigma}}}) = 0$, that is, $\hat{\sigma}(\cdot|x) = \sigma^{\nu_{\hat{\sigma}}}(\cdot|x)$ for μ-almost all x (see Lemma 1.4.1 of Dupuis and Ellis, 1997). ■

We can now easily prove the theorem.

Proof of Theorem 1 It follows immediately from Lemmas 3 and 4. ■

2.4 Optimality condition

In this section we characterize the solutions of the variational problem $\sup_\nu V(\nu)$.

Proposition 5 *A probability measure $\hat{\nu}$ solves $\sup_\nu V(\nu)$ if and only if*

$$\int_X K(x;\hat{\nu})e^{k(x,y)}d\mu(x) \leq 1 \quad \forall y \in Y. \tag{4}$$

Proof "If". Let $\hat{\nu}$ satisfy condition (4). It is enough to apply the inequality $\log t \leq t-1$ to prove that $\hat{\nu}$ solves $\sup_\nu V(\nu)$. For each $\nu \in \Delta(Y)$, we have

$$\begin{aligned}
V(\nu) &= \int_X \left(\log \int_Y e^{k(x,y)}d\nu(y)\right) d\mu(x) = \int_X \left(\log\left(\frac{\int_Y K(x;\hat{\nu})e^{k(x,y)}d\nu(y)}{K(x;\hat{\nu})}\right)\right) d\mu(x) \\
&= \int_X \left(\log\left(\int_Y K(x;\hat{\nu})e^{k(x,y)}d\nu(y)\right)\right) d\mu(x) - \int_X \log K(x;\hat{\nu})d\mu(x) \\
&\leq \int_X \left(\int_Y K(x;\hat{\nu})e^{k(x,y)}d\nu(y) - 1\right) d\mu(x) + V(\hat{\nu}) \\
&= \int_Y \left(\int_X K(x;\hat{\nu})e^{k(x,y)}d\mu(x) - 1\right) d\nu(y) + V(\hat{\nu}) \leq V(\hat{\nu}).
\end{aligned}$$

"Only if". Let $V(\hat{\nu}) = \sup_\nu V(\nu)$. If ν is any element of the convex set $\Delta(Y)$, we have

$$V((1-t)\hat{\nu} + t\nu) = V(\hat{\nu} + t(\nu - \hat{\nu})) \leq V(\hat{\nu})$$

for every $t \in [0,1]$. This implies that the directional derivative is negative, i.e.,

$$V'(\hat{\nu};\nu-\hat{\nu}) = \lim_{t\to 0^+} \frac{V(\hat{\nu} + t(\nu-\hat{\nu})) - V(\hat{\nu})}{t} \leq 0.$$

Since V is concave, this limit exists (finite or infinite) because the scalar function $t \longmapsto t^{-1}\left[V(\hat{\nu} + t(\nu-\hat{\nu})) - V(\hat{\nu})\right]$ increases as $t \downarrow 0$.

Hence, $V'(\hat{\nu};\nu-\hat{\nu}) \leq 0$ for all $\nu \in \Delta(Y)$ is a necessary condition. On the other hand,

$$\begin{aligned}
&V'(\hat{\nu};\nu-\hat{\nu}) = \lim_{t\to 0^+} \frac{V((1-t)\hat{\nu} + t\nu) - V(\hat{\nu})}{t} \\
&= \lim_{t\to 0^+} \frac{\int_X \left(\log\left((1-t)\int_Y e^k d\hat{\nu} + t\int_Y e^k d\nu\right)\right) d\mu - \int_X \left(\log \int_Y e^k d\hat{\nu}\right) d\mu}{t} \\
&= \lim_{t\to 0^+} \int_X \frac{\log\left((1-t)\int_Y e^k d\hat{\nu} + t\int_Y e^k d\nu\right) - \log\int_Y e^k d\hat{\nu}}{t} d\mu \\
&= \int_X \lim_{t\to 0^+} \frac{\log\left((1-t)\int_Y e^k d\hat{\nu} + t\int_Y e^k d\nu\right) - \log\int_Y e^k d\hat{\nu}}{t} d\mu
\end{aligned}$$

where the last step is justified by the monotone convergence theorem because the integrand function increases as $t \downarrow 0$. A straightforward computation leads to

$$V'(\hat{\nu};\nu-\hat{\nu}) = \int_X \left[\frac{\int_Y e^k d\nu}{\int_Y e^k d\hat{\nu}}\right] d\mu - 1 \leq 0$$

If ν puts probability 1 on a generic point $y \in Y$, we get

$$\int_X K(x,\hat{\nu})\, e^{k(x,y)} d\mu \leq 1$$

which is the desired necessary condition. ■

Condition (4) has the following useful equivalent formulation.

Corollary 6 *A probability measure $\hat{\nu}$ satisfies (4) if and only if there is a set $B \in \mathcal{Y}$ of full $\hat{\nu}$-measure such that*

$$\int_X K(x;\hat{\nu})e^{k(x,y)}d\mu(x) = 1 \quad \forall y \in B,$$
$$\int_X K(x;\hat{\nu})e^{k(x,y)}d\mu(x) \leq 1 \quad \textit{otherwise.}$$

Proof Of course, these conditions imply (4). The converse implication relies on the following routine claim.

Claim. If $\nu \in \Delta(Y)$ and f is integrable and such that $f(y) \leq \int f d\nu$ for all y, then $f(y) = \int f d\nu$ for ν-almost all y.

Proof of the Claim. By contradiction, we suppose there are $\nu \in \Delta(Y)$ and f such that $f(y) \leq \int f d\nu$ for all y but the set $B = \{y : f(y) = \int f d\nu\}$ has not full ν-measure. Thus we have

$$\begin{aligned}\int f d\nu &= \int_{\{y:f(y)<\int f d\nu\}} f d\nu + \int_{\{y:f(y)=\int f d\nu\}} f d\nu \\ &< (1-\mu(B))\int f d\nu + \mu(B)\int f d\nu = \int f d\nu\end{aligned}$$

which is a contradiction. Therefore, if $f(y) \leq \int f d\nu$ for all y, then we must have $f(y) = \int f d\nu$ for ν-almost all y. This completes the proof of the Claim. □

In view of the Claim, to conclude the proof is enough to observe that, by the Tonelli Theorem,

$$\int_Y \left(\int_X K(x;\hat{\nu})e^{k(x,y)}d\mu(x)\right) d\hat{\nu}(y) = \int_X \left(K(x;\hat{\nu})\int_Y e^{k(x,y)}d\hat{\nu}(y)\right) d\mu(x) = 1.$$

Hence, condition (4) is equivalent to

$$\int_X K(x;\hat{\nu})e^{k(x,y)}d\mu(x) \leq \int_Y \left(\int_X K(x;\hat{\nu})e^{k(x,y)}d\mu(x)\right) d\hat{\nu}(y)$$

for all $y \in Y$ and so the Claim yields the desired result. ■

2.5 Existence

Under additional regularity conditions, the next proposition verifies that the solution set $\arg\max_\nu V(\nu)$ is nonempty.

Proposition 7 *Let Y be compact and $k(x,\cdot)$ upper semi-continuous on Y for μ-almost all x. The functional V is weakly upper semicontinuous and there exists at least an element $\hat{\nu}$ of $\Delta(Y)$ such that $V(\hat{\nu}) = \sup_\nu V(\nu)$.*

Proof Let $\nu_n \to \nu$. The functions $e^{k(x,\cdot)}$ are upper semicontinuous and bounded for μ-almost all x. Therefore

$$\limsup_{n\to\infty} \int_Y e^{k(x,y)}d\nu_n(y) \leq \int_Y e^{k(x,y)}d\nu(y)$$

for μ-almost all x (see Theorem 15.5 of Aliprantis and Border, 2006). This in turn implies that

$$\limsup_{n\to\infty} \log \int_Y e^{k(x,y)}d\nu_n(y) \leq \log \int_Y e^{k(x,y)}d\nu(y)$$

for μ-almost all x. As a consequence, Fatou's lemma implies

$$\begin{aligned}\limsup_{n\to\infty}\int_X\left(\log\int_Y e^{k(x,y)}d\nu_n(y)\right)d\mu(x) &\leq \int_X \limsup_{n\to\infty}\left(\log\int_Y e^{k(x,y)}d\nu_n(y)\right)d\mu(x)\\ &\leq \int_X\left(\log\int_Y e^{k(x,y)}d\nu(y)\right)d\mu(x).\end{aligned}$$

The weak upper semi-continuity of the functional is thus proved. The weak compactness of $\Delta(Y)$ concludes the proof (see Theorem 2.43 of Aliprantis and Border, 2006). ■

We conclude that, by Proposition 7 and Theorem 1, if Y is compact and $k(x,\cdot)$ is upper semi-continuous on Y for μ-almost all x, then we have

$$\max_\sigma U(\sigma) = \max_\nu V(\nu).$$

and so the solutions sets of the two variational problems are nonempty.

2.6 A parametric version

In applications, a simple parametric version of Theorem 1 is often useful. Specifically, given a non-zero scalar parameter λ, define $V_\lambda : \Delta(Y) \to \mathbb{R}$ by

$$V_\lambda(\nu) = \int_X\left(\lambda\log\int_Y e^{\frac{k(x,y)}{\lambda}}d\nu(y)\right)d\mu(x)$$

and $U_\lambda : \Sigma \to \mathbb{R}$ by

$$U_\lambda(\sigma) = \int_{X\times Y} k(x,y)d(\mu\otimes\sigma)(x,y) - \lambda D(\mu\otimes\sigma\|\mu\times\nu_\sigma). \tag{5}$$

The parametric kernel σ_λ^ν associated to probability measure $\nu \in \Delta(Y)$ is given by

$$\sigma_\lambda^\nu(B|x) = K_\lambda(x;\nu)\int_B e^{\frac{k(x,y)}{\lambda}}d\nu(y),$$

where $K_\lambda(x;\nu) = 1/\int_Y e^{\frac{k(x,y)}{\lambda}}d\nu(y)$ is a normalizing constant.

We can now state and prove the parametric version of Theorem 1. Here we need the function k to be bounded to consider also negative values of the parameter λ.

Corollary 8 *If k is bounded, we have*

$$\begin{aligned}\lambda > 0 &\Longrightarrow \sup_\sigma U_\lambda(\sigma) = \sup_\nu V_\lambda(\nu),\\ \lambda < 0 &\Longrightarrow \inf_\sigma U_\lambda(\sigma) = \inf_\nu V_\lambda(\nu).\end{aligned}$$

If we identify kernels that are equal μ-a.e., then for each $\lambda \neq 0$ the barycenter map $\sigma \mapsto \nu_\sigma$ is a bijection between the solution sets, and the kernel map $\nu \mapsto \sigma_\lambda^\nu$ is its inverse.

Proof Let $\lambda > 0$. We have:

$$\begin{aligned}&\sup_\sigma\int_{X\times Y} k(x,y)d(\mu\otimes\sigma)(x,y) - \lambda D(\mu\otimes\sigma\|\mu\times\nu_\sigma)\\ =&\lambda\sup_\sigma\int_{X\times Y}\frac{k(x,y)}{\lambda}d(\mu\otimes\sigma)(x,y) - D(\mu\otimes\sigma\|\mu\times\nu_\sigma)\\ =&\lambda\sup_\nu\int_X\left(\log\int_Y e^{\frac{k(x,y)}{\lambda}}d\nu(y)\right)d\mu(x)\\ =&\sup_\nu\int_X\left(\lambda\log\int_Y e^{\frac{k(x,y)}{\lambda}}d\nu(y)\right)d\mu(x)\end{aligned}$$

For $\lambda < 0$,

$$\begin{aligned}
&\inf_{\sigma} \int_{X\times Y} k(x,y)d(\mu\otimes\sigma)(x,y) - \lambda D(\mu\otimes\sigma\|\mu\times\nu_\sigma)\\
=&\lambda \sup_{\sigma} \int_{X\times Y} \frac{k(x,y)}{\lambda} d(\mu\otimes\sigma)(x,y) - D(\mu\otimes\sigma\|\mu\times\nu_\sigma)\\
=&\lambda \sup_{\nu} \int_X \left(\log \int_Y e^{\frac{k(x,y)}{\lambda}} d\nu(y)\right) d\mu(x)\\
=&\inf_{\nu} \int_X \left(\lambda \log \int_Y e^{\frac{k(x,y)}{\lambda}} d\nu(y)\right) d\mu(x).
\end{aligned}$$

■

So far we have assumed that $\lambda \neq 0$. If k is continuous, by the Laplace method we have

$$\lim_{\lambda\to 0^-} V_\lambda(\nu) = \int_X \min_{y\in\text{supp}(\nu)} k(x,y)d\mu(x) \quad \text{and} \quad \lim_{\lambda\to 0^+} V_\lambda(\nu) = \int_X \max_{y\in\text{supp}(\nu)} k(x,y)d\mu(x)$$

for all ν that have compact support (see Cerreia-Vioglio et al., 2018). The functional V_λ has thus no meaningful value at $\lambda = 0$ because at that point occurs a transition between maximization and minimization standpoints.

3 Applications

3.1 Rational inattention

A decision maker (DM, for short) faces a choice under uncertainty. Let X be a finite set of *states*. The DM has to choose an *action* from some finite set Y and each action has state-dependent utility. Let $u : X \times Y \to \mathbb{R}$ be her *utility function*. Uncertainty about the state is represented by a subjective probability measure μ on X.

Before choosing what action to take, the DM can run an experiment to reduce the uncertainty she faces. Let Z be a finite set of *signals*. An *experiment* is a stochastic kernel $\tau : X \to \Delta(Z)$ mapping states into probability measures on signals. We denote by $\tau(z|x)$ the probability of observing signal z in state x. A *decision function* $f : Z \to \Delta(Y)$ specifies a (mixed) action for every possible signal realization.

To maximize expected utility, the DM can run any experiment she desires subject to a cost proportional to its informational content. Shannon's mutual information is used to quantify such content: given experiment τ, the quantity

$$D(\mu\otimes\tau\|\mu\times\nu_\tau) = \sum_{x,z} \left(\log \frac{\tau(z|x)}{\sum_{x'}\tau(z|x')\mu(x')}\right)\tau(z|x)\mu(x)$$

is the called *mutual information* of signal and state. It is the relative entropy between $\mu\otimes\tau$ and the product of the marginals $\mu\times\nu_\tau$.[5]

Overall, the DM solves the *rational inattention* problem

$$\max_{\tau,f} \sum_{x,y,z} u(x,y)f(y|z)\tau(z|x)\mu(x) - \lambda D(\mu\otimes\tau\|\mu\times\nu_\tau), \tag{6}$$

[5] Here ν_τ is the barycenter of τ. A standard textbook reference for mutual information, entropy, and related notions is chapter 2 of Cover and Thomas (2006).

where $\lambda > 0$ is a scale factor parametrizing the marginal cost of information (see Matejka and McKay, 2015). Rational inattention was introduced in economics by Sims (2003) to model optimal attention allocation. Stripped of this interpretation, (6) is a costly information acquisition problem with two distinctive features: many experiments are available and the cost of information has a particular functional form based on mutual information.

In rational inattention, a revelation principle is often invoked to bypass decision functions (see, e.g., Corollary 1 in Matejka and McKay, 2015).

Proposition 9 *Suppose Z has more elements than Y. Given $\hat{\sigma} : X \to \Delta(Y)$, the following statements are equivalent:*

(i) $\hat{\sigma}$ solves problem

$$\max_{\sigma} \sum_{x,y} u(x,y)\sigma(y|x)\mu(x) - \lambda D(\mu \otimes \sigma \| \mu \times \nu_\sigma). \tag{7}$$

(ii) There are $\hat{\tau} : X \to \Delta(Z)$ and $\hat{f} : Z \to \Delta(Y)$ solving (6) such that

$$\hat{\sigma}(y|x) = \sum_{z} \hat{f}(y|z)\hat{\tau}(z|x) \quad \forall x \in X, \forall y \in Y.$$

In view of this revelation principle, problem (7) is equivalent to

$$\max_{\sigma} U_\lambda(\sigma)$$

where U_λ is given by (5), with $k = u$. The analysis of Section 2 can therefore be applied to study optimal attention allocation.

3.2 Rate distortion theory

Rate distortion theory is a major branch of information theory, due to Shannon (1959), that studies lossy data compression.[6] The basic problem in data compression is to represent a given message using the minimum amount of information without incurring any distortion, that is, without losing any data. This is called *lossless data compression*. Sometimes, however, a perfect representation would require too much information and losing some data is unavoidable. *Lossy data compression* is about representing a given message using the minimum amount of information without exceeding a certain level of distortion.

For every level of distortion $d > 0$, the *rate distortion function* specifies the exact amount of information necessary to represent a given message without exceeding (on average) d. To illustrate, let a message be an element of the finite set X with probability measure μ. The message has to be represented by an element of the finite set Y. Representing x by y leads to distortion or *loss* $l(x,y) \geq 0$. The rate distortion function $R : (0,\infty) \to [0,\infty)$ is the value function of the following optimization problem:

$$R(d) = \min_{\sigma} D(\mu \otimes \sigma \| \mu \times \nu_\sigma) \quad \text{sub} \sum_{x,y} l(x,y)\sigma(y|x)\mu(x) \leq d. \tag{8}$$

The Lagrangian L associated with this convex program is:

$$\begin{aligned} L(\sigma,\lambda) &= D(\mu \otimes \sigma \| \mu \times \nu_\sigma) + \lambda \left(\sum_{x,y} l(x,y)\sigma(y|x)\mu(x) - d \right) \\ &= D(\mu \otimes \sigma \| \mu \times \nu_\sigma) + \lambda \sum_{x,y} l(x,y)\sigma(y|x)\mu(x) - \lambda d, \end{aligned}$$

[6] Textbook references for rate distortion theory are Berger (1971) and chapter 10 of Cover and Thomas, 2006. A comprehensive survey is Berger and Gibson (1998).

where $\lambda \geq 0$ is the multiplier. Under the hypothesis that problem (8) has solutions, the saddle property of the Lagrangian L leads to the following result.

Proposition 10 *We have* $R(d) = \max_{\lambda \geq 0} \{Q(\lambda) - \lambda d\}$, *where*

$$Q(\lambda) = \min_{\sigma} \left[D(\mu \otimes \sigma \| \mu \times \nu_\sigma) + \lambda \sum_{x,y} l(x,y)\sigma(y|x)\mu(x) \right]. \tag{9}$$

Proof The min-max property of the Lagrange function yields $\max_{\lambda \geq 0} \min_{\sigma} L(\sigma, \lambda) = R(d)$. In view of the definition of L, we have $\min_{\sigma} L(\sigma, \lambda) = Q(\lambda) - \lambda d$, as desired. ■

To compute the rate distortion function, therefore, we can first study the unconstrained optimization problem (9). We have the obvious relation

$$Q(\lambda) = \lambda \min_{\sigma} U_{-\frac{1}{\lambda}}(\sigma) \qquad \text{for } \lambda > 0$$

with $k = l$. The analysis of Section 2 can therefore be applied to study lossy data compression.

References

[1] C. D. Aliprantis and K. C. Border, *Infinite dimensional analysis*, 3rd ed., Springer, 2006.

[2] T. Berger, *Rate distortion theory: A mathematical basis for data compression*, Prentice-Hall, 1971.

[3] T. Berger and J. D. Gibson, Lossy source coding, *IEEE Transactions on Information Theory*, 44, 2693-2723, 1998.

[4] S. Cerreia-Vioglio, F. Maccheroni and M. Marinacci, A characterization of probabilities with full support in metric spaces and its implications for Laplace method, mimeo, 2018.

[5] T. M. Cover and J. A. Thomas, *Elements of information theory*, 3nd ed., Wiley, 2006.

[6] I. Csiszar, On an extremum problem of information theory, *Studia Scientiarum Mathematicarum Hungarica*, 9, 57-71, 1974a.

[7] I. Csiszar, On the computation of rate-distortion functions, *IEEE Transactions on Information Theory*, 20, 122-124, 1974b.

[8] I. Csiszar and J. Korner, *Information theory: coding theorems for discrete memoryless systems*, Cambridge University Press, Cambridge, 2011.

[9] P. Dupuis and R. S. Ellis, *A weak convergence approach to the theory of large deviations*, Wiley, 1997.

[10] R. G. Gallager, *Information theory and reliable communication*, Wiley, 1968.

[11] R. G. Gallager, *Information theory and reliable communication (Udine course)*, Springer Verlag, 1972.

[12] L. P. Hansen and M. Marinacci, Ambiguity aversion and model misspecification: an economic perspective, *Statistical Science*, 31, 511-515, 2016.

[13] B. Mackowiak, F. Matejka, and M. Wiederholt, Rational inattention: A disciplined behavioral model, mimeo, 2018.

[14] F. Matejka and A. McKay, Rational inattention to discrete choices: A new foundation for the multinomial logit model, *American Economic Review*, 105, 272-298, 2015.

[15] C. E. Shannon, Coding theorems for a discrete source with a fidelity criterion, *IRE Nat. Conv. Record*, 4, 142-163, 1959.

[16] C. A. Sims, Implications of rational inattention, *Journal of Monetary Economics*, 50, 665-690, 2003.