A "quantized" approach to rational inattention

Gilles Saint-Paul*
Toulouse School of Economics

September 22, 2011

1 Introduction

In a series of seminal papers, Sims (2003,2006) has proposed a novel approach to bounded rationality. It is based on the view that people face information capacity constraints defined using Shannon's (1948) theory of information. More precisely, the number of bits that one can use to process the exogenous variables (like income) into the endogenous ones (like consumption) is limited. That informational requirement is defined by Shannon's mutual information concept, which tells us the amount of information obtained on a variable when one observes another, correlated one.

As a consequence of that information constraint, the endogenous variable is noisy compared to the optimal behaviour that would prevail absent an informational constraint. Thus, is many applications (such as Luo (2008), who studies a consumption problem, and the papers cited below on price-setting) the agents rationally allocates this noise so as to maximize its utility subject to the information capacity constraint. The more noisy is the endogenous

*I am indebted to Enduardo Engel, Giuseppe Moscarini, Filip Mateka, Christopher Sims, and seminar participants at Sciences Po, Paris, CREI, Universitat Pompeu Frabra, Barcelona, Universidade do Minho, Braga, Tilburg, Princeton, Yale and IMT Lucca, for helpful comments and suggestions.

variable in a given zone of the distribution of exogenous variables, the less the agent pays attention to that zone and the greater the informational capacity left for processing other zones.

As pointed out by Sims, the reason why noise must inevitably arise is that if the distribution of the exogenous variables is continuous, then an infinite amount of information would be needed to process a deterministic mapping from the exogenous variables into the endogenous ones.

In some settings, the noise is inherent to the problem of measuring a signal and the agents' informational capacity is used to reduce such a noise. The rational inattention theory then tells us, in some sense, how to optimally design the noise so as to get the highest possible welfare subject to the information capacity constraint.

In other settings, though, the result that behaviour adds noise to the exogenous variables is unpalatable. If the realization of the latter is perfectly observed then the agent would have to generate the noise artificially, but then it is problematic to ignore the information needed to generate such noise. It would then be more reasonable to assume that the behaviour of the agent remains deterministic while the information processing constraint prevents it from targeting the optimal behaviour. However, that is not what is happening in the rational inattention literature.

In this paper, I propose an alternative approach to that issue. The idea is that the choice variable may be a deterministic function of the exogenous one and still make use of a finite amount of information if the choice variable is discrete rather than continuous; that is, the mapping from the realization of the exogenous variables to the endogenous ones is piece-wise constant, reflecting the fact that the agent can only elect a finite number of values for the choice variable, because of the informational constraint.

Thus, limited information is now a source of lumpiness in behavior, rather than a source of noise. The state space faced by the agent is partitioned into clusters and all points in the same cluster yield the same action. Of course, limited information is not the only source of lumpy behavior; it is well known

that there are other sources, such as fixed or linear adjustment costs. But the approach proposed here yields many potentially testable predictions: In general, we expect that the greater the information processing ability of an economic entity, the less lumpy its behavior.

Another central result (Section 2) is that the mutual information between the exogenous variable and the endogenous one is simply equal to the entropy, in the usual discrete sense, of the endogenous variable. That is, the mutual information does not depend on the exact mapping from the exogenous variable to the endogenous one but only on the probability weights of the (discrete) distribution of the latter. This remedies some weaknesses of the continuous notions of entropy which is used in the literature, which makes it impossible to separate the probability weights of the variables from their actual values. More results regarding the convexity of clusters and the properties of the solution when shocks can be aggregated into composite ones are derived in section 3.

Sections 4 and 5 illustrate the kind of results that my approach would deliver by applying it to two simple examples: a general linear-quadratic problem with a uniform distribution, and a simple static model of price-setting where individual price setters face aggregate monetary shocks and idiosyncratic productivity shocks. This model delivers a lumpy price-setting behavior where the number and size of clusters depends on the dispersion of shocks and the firm's information processing capacity. It is consistent with recent micro-level evidence on reference prices found by Eichenbaum et al (2008).

The literature that has studied this issue (in particular, Mackowiak and Wiederholt (2009a,b), Paciello (2007)) uses Sim's noisy approach and has shown that under rational inattention prices were "sticky" in the sense that the aggregate price level was not reacting one for one to the aggregate money stock[1]. However, prices are not lumpy: even a small monetary shock will

[1]The same result is reached by Saint-Paul (2005) in a world where firms are irrational and experiment alternative price-setting rules, while exerting local spillovers on each other.

generate a small (but non neutral) response of individual prices (one notable exception is Moscarini (2004))[2]. Here, in contrast, rational inattention leads to lumpy price-setting behavior; for prices to change, the shocks faced by a firm must be large enough to trigger a move to a different cluster. As discussed in Section 6, this makes a substantial difference. In the noisy approach, it is optimal to react less than one for one to monetary shocks because one can only react to a noisy measure of those shocks, as in the Lucas (1972) misperception model. Consequently, such underreaction also prevails at the aggregate level. Here, though, no noise is introduced and stickiness arises at the individual level because the same price is charged within a cluster of realizations of the individual price setter's relevant shock. But, as in Caplin and Spulber (1986), such stickiness is greatly reduced in the aggregate because of the contribution of the firms which move across clusters as a result of a monetary shock.

It is important to note that while lumpiness is the only way to reconcile limited mutual information with deterministic behavior, the converse is not true. Depending on the objective function and the distribution of noise, the support of the endogenous variable may either be discrete or continuous, as recently found by Matejka (2008) and Matejka and Sims (2010), who provide some partial but powerful sufficient conditions for discreteness to arise. Yet, even in this case, it is always optimal to have a noisy behaviour, so that the implications for aggregate price stickiness will resemble those of Mackowiak and Wiederholt (2009a), rather than those derived here.

[2]In that paper, lumpiness arises for different reasons than here. Time is continuous and there is a constraint on the flow of information processed by the agent. the exogenous variable follows a diffusion process. A noisy signal of that variable can be obtained at a cost. The cost structure of information is such that the signal will be drawn infrequently, at discrete dates. Thus there is lumpiness "in time" rather than in the state space. Consequently, the model is similar to that of Mankiw and Reis's (2002) sticky information paper.

2 Continous and discrete entropy and mutual information

It is somewhat important to realize that there are two different concepts of entropy. For a discrete distribution with n outcomes and probabilities p_i, $i = 1, ..., n$, we may define entropy as[3]

$$S = -\sum_{i=1}^{n} p_i \log p_i.$$

On the other hand, for a continuous distribution with density $f(x)$, we may define entropy as

$$H(f) = -\int f(x) \log f(x).dx.$$

The reason why the two concepts do not coincide is as follows. A discrete distribution is always the limit of a sequence of continuous distributions, as they become more concentrated around the discrete outcomes. However, the continuous entropy H of those approximations does not converge to the corresponding S. Instead, it converges to $-\infty$.

Take for example the extreme case where $x = 0$ with probability 1. Clearly, $S(x) = 0$. This discrete distribution is the limit of the continuous one defined by density $f_\varepsilon(x) = f(x/\varepsilon)/\varepsilon$, for any density $f()$ over $(-\infty, +\infty)$, which is continuous and such that $f(0) > 0$, as ε goes to zero (in terms of distribution theory, these distributions converge to a Dirac function $\delta(x)$). Furthermore,

$$H(f_\varepsilon) = H(f) + \log \varepsilon,$$

so that

$$\lim_{\varepsilon \to 0} H(f_\varepsilon) = -\infty.$$

Entropy is lower, the more concentrated the distribution. For both discrete and continuous distributions, the most concentrated one is when all the

[3] In the usual definition, the logarithm is in base 2, but I will more conveniently use natural logarithms instead.

mass is at a single point. But the lower bound of S is zero, while that of H is -∞.

Let us now turn to mutual information, which plays a key role in the theory of rational inattention. We consider two random variables x and y. Their densities are $g()$ and $f()$, respectively. For any realization of x, we denote by $f(y \mid x)$ the conditional distribution of y and its entropy is

$$H(y \mid x) = -\int_y f(y \mid x) \log f(y \mid x).dy.$$

This can be averaged over x, which allows to define the conditional entropy of y :

$$H_x(y) = \int_x H(y \mid x)g(x)dx.$$

Now it can be easily shown that the entropy of the joint distribution of x and y, $H(x, y)$, is such that[4]

$$H(x, y) = H(x) + H_x(y) = H(y) + H_y(x).$$

Consequently, we have that

$$H(y) - H_x(y) = H(x) - H_y(x) = M(x, y),$$

which is the *mutual information* between x and y. This quantity tells us how much knowledge of one variable reduces the entropy of the other, on average. If the two variables are independent, then $M(x, y) = 0$. On the other hand, if one had $y = x$, then the joint distribution is degenerate and $f(y \mid x)$ becomes equal to the Dirac function $\delta(y - x)$. Hence all the $H(y \mid x)$ are equal to $-\infty$ and so is $H_x(y)$. We then have that $M(x, y) = +\infty$. This means that knowledge of x gives us an infinite amount of information about y. The same conclusion would be reached if instead of $y = x$, there was any other mapping which allowed to retreive one variable from the other.

[4]In fact, that property is one of the axioms imposed by Shannon to derive his functional form for entropy.

The theory of rational inattention, as proposed by Sims, assumes that an agent receives a signal y (say, income), which must be processed into a decision variable x (say, consumption). The agent's ability to process information is limited and that limit takes the form of a constraint on the mutual information between the two variables:

$$M(x, y) \leq K.$$

Since $M(x, y) = +\infty$ if x is a deterministic function of y, this constraint cannot be matched. The endogenous variable must be related to the exogenous one in a noisy fashion for the information capacity constraint to be matched. In other words, processing a continuum of real values with perfect precision requires an infinite amount of information.

I now show that there is an important exception to that principle, and this is the case when x, while being a deterministic function of y, is "quantized" in that it only takes a finite number of values. There is then no longer a mapping from y to x. While x can be retreived from y, the converse is not true. In such a case, the mutual information between x and y remains finite, and is in fact equal to the discrete entropy S of the random variable x.

Let us consider a collection of values of x, $\Omega = \{x_1, ..., x_n\}$, and assume that any y is assigned to one of those values, called $x(y)$. For any $i \in \{1, ...n\}$, we define $T_i = \{y, x(y) = x_i\}$. To avoid manipulating infinite quantities, I will consider my deterministic assignment as the limit, for $\varepsilon \to 0$, of the random variable x defined by its conditional distribution:

$$f_\varepsilon(x \mid y) = \frac{1}{\varepsilon} \hat{f}\left(\frac{x - x(y)}{\varepsilon}\right). \tag{1}$$

Here again, $\hat{f}()$ is any density such that $\hat{f}(0) > 0$. We are again in a situation where the conditional of x is a Dirac, now around $x(y)$, and we approximate it by a density which becomes increasingly concentrated around $x(y)$. To fix ideas, one can just take the standard normal density for $\hat{f}()$.

Then the following can be proved:

Theorem 1 – Let $M(\varepsilon)$ be the mutual information between x and y if x is distributed according to (1). Let $S(X) = -\sum_i P_i \log P_i$ be the discrete entropy of the random variable whose realization is x_i, with corresponding probability $P_i = \int_{T_i} dF$. Then

$$\lim_{\varepsilon \to 0} M(\varepsilon) = S(X).$$

Proof- See Appendix

Theorem 1 tells us that if a policy function is quantized, then the mutual information between the exogenous variable and the policy variable is equal to the entropy, in the discrete sense, of the policy variable. In particular, it only depends on its probability weights and not on the actual values of that variable. Nor does it depend on how the values of y are assigned to the different clusters, as long as the probability weights are unchanged.[5]

3 Some additional results

In this section I provide some additional results that are likely to be relevant in many practical applications.

Assume the value function is $V(y, x)$, where $y \in R^n$ is the vector of exogenous variables and $x \in R^p$ is the vector of endogenous choice variables. That is, one is maximizing

$$EV = \int \int_{x,y} V(x, y) dF(x, y).$$

Assume V is continuous. Assume we have a finite partition of R^n, $\{T_1, ..., T_i, ..., T_N\}$, and let x_i the common value of x assigned to any $y \in T_i$. Assume also that

[5] On the other hand, if the policy function were discrete but stochastic, the mutual information would depend on the assignment rules. For example, assume that y is distributed uniformly over [0,1] and there are two clusters, each with probability 1/2. Then in the deterministic case, $M(x, y) = \ln 2$. In the stochastic case, let $p(y)$ the probability density that y is assigned to cluster 1. Then $m(x, y) = \ln 2 + \int_0^1 (\ln p(y) + \ln(1 - p(y))p(y)dy < \ln 2$.

y has a density $f()$ with full support[6], implying that

$$EV = \sum_i \int_{T_i} V(x_i, y) f(y) dy.$$

For any set S, denote by $\mathring{S}$ its interior.

Then a necessary condition for the assignment to be optimal is the *no swapping condition:*

$$\forall i, j, \forall y_i \in \mathring{T}_i, y_j \in \mathring{T}_j, V(x_i, y_i) + V(x_j, y_j) \geq V(x_j, y_i) + V(x_i, y_j,).$$

If this failed to hold, we could take to small neighborhoods around y_i and y_j of equal probability weights and intervert them, reallocating the first one to T_j and the first one to T_i. This would leave the total entropy of the partition unchanged but achieve a higher value of the objective.

This condition allows us to proven a number of results. I first consider whether the clusters T_i may be convex, a natural property to seek.

3.1 Cluster convexity

For any set S, denote by $C(S)$ its convex envelope and by $\bar{S}$ its closure. For any real-valued function $g(v_1, ..., v_n)$, where v_i is a vector of dimension q_i, denote by $\nabla_i g$ its gradient with respect to v_i, which is also a vector of dimension q_i. Denote by a dot $(\cdot)$ the scalar product.

Clearly, given an optimal clustering, any other clustering which only differs by sets of measure zero and assigns the same values of x almost surely is also optimal. To get around this problem, I will restrict the choice of clusters to sets with a dense interior, i.e. we must have $T_i \subset \overline{\mathring{T}_i}$, and that have the same measure as their interior. Since V is continuous, an immediate consequence is that the no swapping condition must also hold for $y_i \in T_i, y_j \in T_j$, not just over the interiors of those sets.

[6]The proofs of the following results can be adapted without difficulty if $f()$ does not have a full support.

Theorem 2 – Assume V is such that:

$\forall x, x', y$, the following property $P(x, x', y; y')$ holds for $y' \in R^n - M(x, x', y)$, where $M(x, x', y)$ is a manifold:

$P(x, x', y; y') : ((y - y') \cdot (\nabla_2 V(x, z) - \nabla_2 V(x', z))$ has a constant non zero sign for $z \in [y, y']$.

Then $C(T_i) \subset \bar{T}_i$, for all i.

Proof – See Appendix

This theorem tells us that if we restrict our choice to usual sets that have a dense interior with full measure, and if P holds except on a set of small dimension, then the optimal clusters have a convex closure, which essentially means that they are convex except perhaps along their boundary, which is of measure zero and irrelevant.

In the quadratic case, $V = -\|y - x\|^2$, and $(y - y') \cdot (\nabla_2 V(x, z) - \nabla_2 V(x', z)) = 2(x' - x) \cdot (y - y')$ and (P) clearly holds since $2(x' - x) \cdot (y - y')$ does not depend on z and is nonzero except over a hyperplane. As a corollary, in the one dimensional case, the partition would consist of intervals.

Property P is a generalized single-crossing condition. If for example $p = 1$, then it boils down to monotonicity of $\partial V(x, y)/\partial y - \partial V(x', y)/\partial y$, which is a standard single-crossing condition.

3.2 Aggregating shocks

Suppose that the multidimensional vector y of exogenous variables only comes into play through a single-dimensional aggregate $u(y)$. Do we get an equivalent solution by reformulating the problem in terms of the composite aggregate u? The following results provide an answer.

Theorem 3 – Assume there exists a continuous mapping $u() : R^p \to R$, which is regular everywhere i.e. all its partial derivatives cannot simultane-

ously vanish, and such that $V(x,y) = V(x,u(y))$. *Assume the single-crossing property P' holds:*

$$\forall x \neq x' \; V(x,u) - V(x',u) \text{ is strictly monotonic in } u. \text{ (P')}$$

Let $\{T_i\}$ *be an optimal partition and* $U_i = u(\mathring{T}_i)$, *Then for all* $i, j, i \neq j$, $U_i \cap U_j = \varnothing$.

Proof – See Appendix

This result shows us that the optimal policy remains essentially deterministic in terms of the composite shock: a given value of the composite shock u is assigned to a single T_i, except for realizations of y that lie on the frontier of a cluster, which is typically of measure zero. An immediate consequence is that nothing is lost by restricting oneself to the "projected" problem, i.e. restating the orginial problem in terms of the composite shock, using its distribution. This is stated in Theorem 4.

Theorem 4 – Assume the assumptions of Theorem 3 hold. Let $m()$ *be the measure over* R *defined by* $m(S) = \mu(u^{-1}(S)) = \int_{u^{-1}(S)} f(y)dy$. *Let a partition* $\{U_i\}$ *of* R *with assignment* $\{x_i\}$ *which solves solves problem* $P2$:

$$\max_{\{U_i\}, x()|x(U_i)=x_i} EV(x(u),u) \;\; s.t. \;\; -\sum_{i=1}^{N} m(U_i) \ln m(U_i) \leq K \; . \qquad \text{(P2)}$$

Then the partition of R^p *given by* $\{T_i = u^{-1}(U_i)\}$ *with the same assignment solves problem* $P1$:

$$\max_{T, x()|x(T_i)=x_i} EV(x(y),u(y),) \;\; s.t. \;\; -\sum_{i=1}^{N} F(T_i) \ln F(T_i) \leq K \qquad \text{(P1)}$$

Proof: See Appendix.

It would be interesting to be able to predict which zones of the space one will pay more attention to. With the preceding notations, cluster i eats

$-F(T_i)\ln F(T_i)$ units of informational capacity. That is $-\ln F(T_i)$ per unit of probability mass. Therefore, and intuitively, more attention is paid to cluster i, the smaller its weight $F(T_i)$. Unfortunately, not much can be said in general from the first-order conditions about which zones of the distribution of y will get more attention. Suppose now that $n = 1$. Then if the single crossing condition holds T_i is an interval $[y_i, y_{i+1}]$ and the first-order condition for setting the boundary optimally is

$$V(x_i, y_i) - V(x_{i-1}, y_i) = -\lambda \ln \frac{F(T_i)}{F(T_{i-1})},$$

where λ is the Lagrange multiplier of the information capacity constraint. Therefore, if a cluster i is smaller than its neighbor $i-1$ (hence paid more attention to), then the RHS is positive, implying that total utility would go up if the common policy applied to cluster i were also applied to the values of y that are close to its boundary y_i in the neighboring cluster $i-1$. Absent the information capacity constraint, one would like to widen cluster i, but doing so would increase the entropy of X. At the optimum, the shadow cost of the extra information needed to widen cluster i is equal to the marginal utility gain of doing so, as captured by the preceding equation.

This illustrates how optimization is done better in the smaller clusters: Since a smaller cluster is more costly in terms of information, it must in turn deliver greater benefits in terms of optimization. This rules out some configurations. For example, if V is quadratic, i.e. $V(x, y) = -(x-y)^2$, then a wide cluster delivers low utility at its borders, and therefore cannot be paid too much attention, meaning it must have a large value of $F(T_i)$. In particular, if the density of $F(), f()$ is very small over a wide area, such an area must typically be included in an even larger cluster to ensure that it has sufficiently high probability weight. Apparently, little more can be said beyond these considerations.

4 The linear-quadratic case

I now apply these ideas to the linear-quadratic case. In its simplest case, the agent receives a continuous signal $y \in R$ with density $f(y)$ and associated measure $F(M) = \int_M f(y)dy$, and wants to approximate it (in the least squares sense) by a deterministic function $x(y)$ which is constant over each subset of a finite partition of the domain of y. Thus, using the preceding derivation for mutual information in the discrete case, we can formulate the problem as follows (in the sequel I will use natural logarithms in the definition of entropy. Thus K is expressed in bits / $\ln 2$).

$$\min_{n,\Omega=(x_1,...,x_n),x():\mathbb{R}\to\Omega} E(x(y)-y)^2 \quad s.t. \quad -\sum_{i=1}^{n} F(x^{-1}(x_n)) \ln F(x^{-1}(x_n)) \leq K \tag{P3}$$

As pointed out in the previous section, property (P) is satisfied, so that the optimal clusters are intervals. Given that mutual information does not depend on the actual values of x, we then have:

Lemma 1 – $x_n = E(y \mid y \in T_n)$. *Therefore* $x(y)$ *is non decreasing.*

Proof: the first part is straightforward from the optimal choice of x_n. The second derives from the fact that the T_ns are intervals except for subsets of measure zero.

I now focus on the case where $f()$ is uniform over $[0, 1]$. It is then possible to fully characterize the equilibrium:

Theorem 5 – *Assume* $f()$ *is uniform. Then an optimal policy is such that*

(i) The interval $[0, 1]$ *is partitioned into* N *adjacent intervals* $[y_n, y_{n+1}], y_0 = 0, y_N = 1$.

(ii) $N = INT^+(e^K)$, *where* $INT^+(z)$ *is the smallest integer* m *such that* $z \leq m$.

(iii) $N-1$ *intervals have the same length* Δ*, where* Δ *is the smallest solution to*

$$-(N-1)\Delta \ln \Delta - (1-(N-1)\Delta)\ln(1-(N-1)\Delta) = K,$$

while the remaining interval has length $1-(N-1)\Delta$.

(iv)

$$\Delta < 1/N < 1-(N-1)\Delta$$

(v) For $y \in [y_n, y_{n+1}], x(y) = x_n = \frac{y_n+y_{n+1}}{2}$

(v) The resulting value function is $V = (N-1)\Delta^3 + (1-(N-1)\Delta)^3$

(vi) The arrangement of those intervals is irrelevant.

Proof — See Appendix.

Note that if capacity K is such that there is an integer number of bits, then $K = k \ln 2$ with k integer, and $e^K = 2^k$. In this important special case, the optimal solution, quite naturally, consists in splitting the interval into 2^k intervals, since one needs exactly k bits to encode the actual interval to which y is assigned. Furthermore, in this limit case where the capacity constraint is marginally binding for $N = 2^k$, all intervals will have the same length $1/2^k$.

If $K/\ln 2$ is not integer, then partitioning into equal intervals is not optimal. Instead, we have one more interval than the largest number of intervals that would allow us to have an equal partition while meeting the informational constraint. We pick $N-1$ equally sized intervals of length Δ, and the remaining one has length $\Delta' = 1-(N-1)\Delta > \Delta$. Δ is such that the informational constraint binds with equality.

5 An illustration with price-setting

I now discuss the implications of the approach derived above for the problem of price setting and the effects of monetary policy.

Let us consider the following static version of the standard new Keynesian model[7]. There is a continuum of consumers-yeoman farmers of total mass 1.

[7] See Weitzman (1985), Blanchard and Kiyotaki (1987).

They are indexed by i and they monopolistically supply an atomistic good with the same index i. Thus there is also a continuum of goods of mass 1. The utility function for individual j is

$$V_j = E \ln \left[\left(\int_0^1 c_{ij}^{\alpha} di \right)^{\frac{1}{2\alpha}} \left(\frac{m_j}{p} \right)^{1/2} X^{-\psi} - z_j x_j^{1+\mu} \right],$$

where E is the expectations operator, c_{ij} consumption of good i, m_j money holdings, p the aggregate price level, z_j an idiosyncratic supply shock and x_j the supply of good j. The term in $X^{-\psi}$ is a negative congestion externality, where X is aggregate real output (defined below) and $\psi \geq 0$. This will allow me to pick the value of ψ so as to focus on a special case which is computationally much simpler, while what is lost by doing so is independent of the point being illustrated here.

For simplicity, the aggregate price level that deflates money holdings in the utility function is assumed to be equal to the price index that is dual to the aggregate consumption index

$$c_j = \left(\int_0^1 c_{ij}^{\alpha} di \right)^{\frac{1}{\alpha}}$$

:

$$p = \left[\int_0^1 p_i^{-\frac{\alpha}{1-\alpha}} di \right]^{-\frac{1-\alpha}{\alpha}}.$$

The usual derivations concerning demand functions and aggregation are made in the Appendix. We can show that each yeoman farmer maximizes the indirect utility function given by:

$$E \ln \left[p_j^{-\frac{\alpha}{1-\alpha}} - \phi_j p_j^{-\frac{1+\mu}{1-\alpha}} \right], \tag{2}$$

where ϕ_j is a composite shock defined by

$$\phi_j = z_j M^{\mu+\psi} p^{1-\psi+\frac{\alpha\mu}{1-\alpha}}. \tag{3}$$

From now on, I will assume that ψ is such that the composite shock does not depend on the aggregate price level: $\psi = 1 + \alpha\mu/(1-\alpha)$. Thus, $\phi_j =$

$z_j M^{\frac{\mu+1-\alpha}{1-\alpha}}$; spillovers in price formation across firms are shut down, which greatly simplifies the analysis. It is then useful to define γ as $\gamma = \frac{\mu+1-\alpha}{1-\alpha}$.

As a benchmark, we can derive the flexible price equilibrium with no informational constraint where a different price is set for each realization of ϕ_j. The FOC for price-setting is equivalent to

$$\begin{aligned} p_j &= \left(\frac{(1+\mu)\phi_j}{\alpha}\right)^{1/\gamma} \\ &= \left(\frac{1+\mu}{\alpha}\right)^{1/\gamma} z_j^{1/\gamma} M. \end{aligned} \tag{4}$$

Integrating we get the aggregate price level:

$$p = M\tilde{z}^{1/\gamma}\left(\frac{1+\mu}{\alpha}\right)^{1/\gamma},$$

where $\tilde{z}$ is an aggregate of z defined as

$$\tilde{z} = \left[\int_0^1 z_j^{-\frac{\alpha}{1-\alpha+\mu}} dj\right]^{-\frac{1-\alpha+\mu}{\alpha}} = E(z^{-\frac{\alpha}{1-\alpha+\mu}})^{-\frac{1-\alpha+\mu}{\alpha}}.$$

Thus money is neutral, the aggregate price level is proportional to M, and real aggregate output is constant and equal to

$$X = \frac{Y}{p} = \frac{M}{p} = \left(\tilde{z}\frac{1+\mu}{\alpha}\right)^{-1/\gamma}.$$

Output is lower, the larger the aggregate cost index $\tilde{z}$, the larger the elasticity of the disutility of effort μ, and the lower the elasticity of demand for the individual goods, i.e. the larger the markup over marginal cost $1/\alpha$.

The New Keynesian literature takes this framework and imposes some nominal price rigidity. I now introduce capacity constraints in processing information along the lines discussed above and derive the associated behaviour of output and the price level.

Under rational inattention, people do not have the information processing ability to pursue a rule like (4) for any value of ϕ_j. Instead they are going to pursue a rule such that the mutual information between p_j and ϕ_j satisfies

a capacity constraint. Let us assume that, as in the above analysis, they pursue a discrete deterministic rule and partition the support of ϕ_j into intervals $I_k = [\bar{\phi}_k, \bar{\phi}_{k+1}]$ such that a constant value of p_j, denoted by $\bar{p}_k$, is pursued within each interval. We assume k varies between 0 and $N+1$, with $\bar{\phi}_0 = 0$ and $\bar{\phi}_{N+1} = +\infty$.

The distribution of the composite shock ϕ has density

$$g(\phi) = \int_0^{+\infty} f(M) M^{-\gamma} h(\phi M^{-\gamma}) dM. \tag{5}$$

Individuals select the number of intervals, their bounds and their associated price levels so as to maximize:

$$\max_{N,\{\bar{\phi}_k, k=1,...,N\},\{\bar{p}_k, k=0,...,N\}} U = \sum_{k=0}^{N} \int_{\bar{\phi}_k}^{\bar{\phi}_{k+1}} g(\phi) \ln \left[\bar{p}_k^{-\frac{\alpha}{1-\alpha}} - \phi \bar{p}_k^{-\frac{1+\mu}{1-\alpha}} \right] d\phi, \tag{6}$$

subject to the information capacity constraint

$$-\sum_{k=0}^{N} \left(\int_{\bar{\phi}_k}^{\bar{\phi}_{k+1}} g(\phi) d\phi \right) \ln \left(\int_{\bar{\phi}_k}^{\bar{\phi}_{k+1}} g(\phi) d\phi \right) \leq K. \tag{7}$$

An equilibrium is therefore a set $\{N, \{\bar{\phi}_k, k = 1, ..., N\}, \{\bar{p}_k, k = 0, ..., N\}\}$ which maximizes (6) subject to (7). The solution to this problem then delivers the aggregate price level as a function $p(M)$ of the realization of the aggregate money stock. Given M, a price-setter j is in interval I_k iff $\bar{\phi}_k \leq z_j M^{\gamma} < \bar{\phi}_{k+1}$, which occurs with probability $H(\bar{\phi}_{k+1} M^{-\gamma}) - H(\bar{\phi}_k M^{-\gamma})$. Therefore, the aggregate price level $p(M)$ is given by

$$p(M) = \left(\sum_{k=0}^{N} \left(H(\bar{\phi}_{k+1} M^{-\gamma}) - H(\bar{\phi}_k M^{-\gamma}) \right) \bar{p}_k^{-\frac{\alpha}{1-\alpha}} \right)^{-\frac{1-\alpha}{\alpha}}, \tag{8}$$

where by convention $H(+\infty) = 1$. This in turn allows us to compute output $X = M/p(M)$. Note that the assumption made on ψ guarantees that the environment faced by each price-setter only depends on the exogenous vari-

ables and not on the prices set by other agents[8]. This greatly simplifies the computations.

I solve for such an equilibrium numerically, performing global optimization on all the possible partitions of the domain of ϕ into a finite number of intervals which match the informational capacity constraint. To keep things tractable the possible values for the jump points have been discretized[9].

Table 1 reports some summary statistics for the simulations. I start from a benchmark numerical exercise where both $f()$ and $h()$ are log-normal[10], with $E \ln M = E \ln z = 0$ and $\text{Var}(\ln M) = \text{Var}(\ln z) = 1$. The other parameters were $\mu = 1$ and $\alpha = 0.5$.

I first start by simulating this economy for $K = 1.2$ and I gradually loosen the information capacity constraint by increasing K. Table 1 reports the corresponding number of clusters along with the variance of log output. Figure 1 reports the behavior of output as a function of the monetary shock M. We see that for a wide range of values of the money stock the curve is quite flat: despite the small number of clusters, heterogeneity due to idiosyncratic shocks is enough to yield near neutrality at the aggregate level, a not unusual result (Caplin and Spulber (1987), Caballero and Engel (1993), Burstein and Hellwig (2007)). The curve is tilda-shaped: at small (resp. large) values of M, most firms charge their minimum (resp. maximum) price, and an increase in M boosts output. For intermediate values, a composition effect creates a force in the opposite direction, as some firms move to a cluster with a higher price. This composition effect creates a zone where money growth is

[8] Otherwise, the shock ϕ and its distribution $g()$ would themselves depend on the aggregate price level, and there would be no closed-form formula such as (8) for the latter–one would then need to search for a fixed point equilibrium rather than just an optimum.

[9] More precisely, there are $\bar{N}$ possible values of $\bar{\phi}_k$ separated by a probability weight of $1/(\bar{N}+1)$, i.e. if those eligible critical values are denoted by $\tilde{\phi}_j$, $j = 1, ..., \bar{N}$, then $\int_{\tilde{\phi}_j}^{\tilde{\phi}_{j+1}} g(\phi) d\phi = 1/(\bar{N}+1)$.

In the simulations, one has picked $\bar{N} = 20$.

[10] In the simulations, the distributions are truncated to eliminate the zone where utility is not defined, i.e. $\left[p_j^{-\frac{\alpha}{1-\alpha}} - \phi_j p_j^{-\frac{1+\mu}{1-\alpha}} \right] \leq 0$. The parameters are such that the truncated zone has a very small probability weight.

contractionary, which also happens in other models of price rigidity.

Figure 2 compares the flatter portion of the output curve between a low information ($K = 1.2$) and a high information ($K = 1.5$) regime. We see that output is substantially flatter in the latter case. Nevertheless, as Table 1 shows, for local increases in capacity, the variance of output may well go up.

It is also interesting to look at the distribution of individual prices. They are reported in Figures 3 (for $K = 1.2$) and 4 (for $K = 1.6$). The dimension of each rectangle along the y-axis is the price and along the x-axis it is the probability weight associated with the corresponding interval of values of ϕ. We see that the probability weights on each price are decreasing with the price, meaning that price-setters are devoting more attention to situations where the required price is higher. This is presumably due to the marginal disutility of labor schedule: the utility cost of not paying attention to these states is high because if one charges too low a price the labor input must be very high[11].

Entropy	#of clusters	Variance of output
1.2	4	0.17
1.3	5	0.18
1.4	5	0.187
1.5	5	0.1
1.6	6	0.1

Table 1.

Table 2 analyses the effect of an increase in the variance of monetary shocks on the distribution of individual prices for $K = 1.4$. We compare the benchmark situation (Table 2a) to one such that $\text{Var}(\ln M) = 2$ and $E(\ln M) = -0.5$ (Implying that $E(M)$ is the same as in the benchmark) (Table 2b). We see that the increase in the variance of money shocks compells price-setters to devote more attention to high realization of those shocks[12]:

[11] This clearly rests on my assumption that demand must be met; this might not remain realistic for very high realizations of the demand shock.

[12] That is because of the skewness of the log-normal distribution along with the increasing

the upper-tail of the distribution of the composite shocks is split into more, and finer, clusters, while the first interval is coarser. Also, the variance of log output increases from 0.19 to 0.49.

Table 3c performs the reverse exercise of dividing the variance of monetary shocks by 2, while again adjusting the mean log of M to hold $E(M)$ constant. We see that the number of clusters is the same, and so is their size, but the order is changed: the second cluster gets the biggest weight, while more attention is paid to low realizations of the shock than before. The intuition for this result is unclear.

Cluster	Price	Weight
1	0.95	0.43
2	1.96	0.24
3	3.74	0.19
4	7.0	0.095
5	23.63	0.048

Table 2a – $K = 1.4$, benchmark.

Cluster	Price	Weight
1	0.76	0.52
2	1.67	0.19
3	2.84	0.095
4	4.99	0.095
5	8.66	0.048
6	47.4	0.048

Table 2b – $K = 1.4$, $\text{Var}(\ln M) = 2$ and $E(\ln M) = -0.5$.

Cluster	Price	Weight
1	0.81	0.19
2	1.93	0.43
3	3.65	0.24
4	6.02	0.095
5	13.45	0.048

Table 2c – $K = 1.4$, $\text{Var}(\ln M) = 0.5$ and $E(\ln M) = 0.25$

marginal disutility of labor property. But for even larger increases in the variance of money shocks, the price setters will also spend information capacity on the lower tail of the distribution. Thus, for $\text{Var}(\ln M) = 4$, cluster 1 has a minimal weight of 0.048.

6 Implications for rigidity

As the preceding section makes clear, the quantized model only implies a moderate degree of price rigidity at the aggregate level. By contrast, in a model where noisy behavior is allowed for, as that of Mackoviak and Wiederholt (2009a,b), the aggregate price level reacts less than one for one to monetary shocks throughout the whole distribution of those shocks. The reason is that this class of models is similar to the Lucas (1972, 1973) misperception model. Information capacity constraints preclude price-setters from reacting one for one to the monetary shock. Instead, they can only react to a noisy signal of the monetary shock. Given that, their optimal inference about the true realization of the money stock will react less than one-for-one to that money stock, as implied by Bayes's Law. The only difference with the Lucas misperception model is that the noise is now designed optimally by the

price-setters so as to meet the information capacity constraint.[13] In the aggregate, all individual prices react less than one-for-one to the money shock,

[13] As an aside, it is interesting to note that the reaction of prices to monetary shocks is optimal conditional on the existence of noise. In other words, it is not the information capacity constraint which is constraining that reaction to be suboptimally low, but rather the noise itself (of course the noise is also a by-product of the information capacity constraint). To see this, let us get back to the standard Gaussian linear-quadratic problem:

$$V = \min E(x-y)^2.$$

Assume $y \sim N(0, \sigma_y^2)$ and $x = ay + \varepsilon$, where a is a reaction coefficient and ε the endogenous noise, assumed normal with zero mean and variance σ_ε^2 and orthogonal to y.

An optimality condition for x is

$$E(y \mid x) = x. \tag{9}$$

This optimality condition pins down the correlation between x and y and it would hold if one observed an exogenous noisy signal of y. In our context, we have $E(y \mid x) = \frac{a\sigma_y^2}{\sigma_\varepsilon^2 + a^2\sigma_y^2} x$. Therefore the optimality condition (9) is equivalent to

$$a(1-a) = \frac{\sigma_\varepsilon^2}{\sigma_y^2}. \tag{10}$$

The value of the objective function is $V = (a-1)^2\sigma_y^2 + \sigma_\varepsilon^2$; The mutual information between x and y is

$$M(x,y) = \frac{1}{2}(\log(a^2\sigma_y^2 + \sigma_\varepsilon^2) - \log(\sigma_\varepsilon^2)). \tag{11}$$

Thus our problem is equivalent to maximizing V subject to

$$\frac{a^2\sigma_y^2}{\sigma_\varepsilon^2} \leq K. \tag{12}$$

Given this constraint, which involves a, it is not a priori obvious that the optimality condition (9) should hold. In contrast, if x and y have a discrete distribution, $M(x,y)$ only depends on the probability weights of their joint distribution, and it is always possible to pick the values of x while leaving $M(x,y)$ unchanged so as to match (9). That $M(x,y)$ is not independent of the values of x because of the presence of a in (11) is a weakness of the entropy concept applied to continuous distributions.

Nevertheless, since one picks both σ_ε and a optimally given the constraint (12), one has one degree of freedom left to match the optimality condition (9)-(10), which turns out to hold at the optimum. Indeed, at the optimum $a = \frac{K}{K+1}$ and $\sigma_\varepsilon^2 = \frac{K}{(K+1)^2}\sigma_\varepsilon^2$, implying that (10) holds.

This proves that the underreaction of x to y does not come from a failure of (9) that would be the price to pay for matching the information capacity constraint. And, if this were the case, it would be an artifact of the use of continuous entropy. Instead, this underreaction is optimal given the presence of (endogenous) noise, exactly as if the noise were exogenous instead.

and so does the aggregate price level. These considerations apply to any model where agents are allowed to introduce noise in their policy functions in order to save on information capacity. In particular, underreaction would also arise if the distribution of the exogenous variable (y) were purely discrete or if that of the endogenous variable (x) turned out to be discrete yet noisy conditional on y as in Matejka and Sims (2010).

On the other hand, in the quantized model developed here, noisy policies are precluded. The information capacity constraint is matched by lumping the realizations of y in clusters within which the same policy is pursued. The actual value of x within each cluster does not affect the mutual information between x and y, since, as we have seen, it only depends on the probability weights of the discrete random variable x. Therefore, within each cluster one will pick the optimal value of x conditional on being in that cluster, ignoring the information capacity constraint. Consequently, if, in the absence of information constraints, it is optimal for x to react one-for-one to y, this will remain so in the quantized solution when one moves across clusters. In the aggregate, money neutrality tends to arise in a similar fashion as in Caplin and Spulber (1987): the large price adjustment of firms that are near the frontier between two clusters tends to offset the price inertia of those firms that remain in the same cluster. Thus the model, resembles a menu cost model rather than the Lucas misperception model, and has much less aggregate price stickiness.

7 Discussion

The general message of this paper is that information processing constraints yield lumpy behavior. Thus, when the exogenous variables change, inattention results in inaction, while in the standard approach it is associated with inadequacy, i.e. embodies excess noise. In both cases, the endogenous variable does not react enough to the exogenous one, although here there will be a jump if one crosses the frontier between clusters.

Perhaps the most straightforward example of discreteness in the economic sphere is the system used by the credit rating agencies: instead of providing a continuous default probability, they rely on lumpy letter grades such as AAA, B+, and so forth. Thus, investors can save on information processing by applying simple rules that treat each credit rating category uniformly.

The existence of lumpiness in the adjustment of economic variables has been documented in a number of areas. For example, Doms and Dunne (1998), studying investment at the plant level, find that "Many plants occasionally alter their capital stocks in lumpy fashions. Of the plants in a balanced panel, over half experience a capital adjustment of at least 37 % in one year, and by 50% in two consecutive years". In the area of price setting, Klenow and Kryvstov (2008) find (table III) that individual price changes are usually large, with a mean size of 14 %. Dhyne et al. (2006) report similar findings, along with substantial heterogeneity in the degree of lumpiness of price adjustment across sectors. More recently, Eichenbaum et al. (2008), using scanner data from the retail trade sector, find that firms pick their prices among a number of finite "reference" prices, and that the evidence is consistent with the view that it is not costly to change prices as long as the new price remains a reference price. This is exactly what happens in the model described above. On the other hand, reference prices change infrequently, which may be interpreted as the outcome of a costly reoptimization process in light of perceived changes in the underlying distribution of shocks or in the technology for processing information.[14]

Finally, evidence of lumpiness in employment can be found in Davis et al (1996) or Caballero et al. (1997). The latter, in particular, found that the distribution of employment changes is typically bimodal.

Of course, rational inattention is not the only reason why there could be lumpiness. The above literature has mostly focused on fixed and linear adjustment costs and rational inattention and adjustment costs are not mu-

[14] Analysing this reoptimization process in the quantized case is an important topic for further research. It would allow to construct truly dynamic models of price setting in the fashion of the one spelled out in Section 5.

tually exclusive mechanisms. The rational inattention mechanism may be of particular interest when large adjustment costs are implausible, as in the area of price setting. Furthermore, a range of novel predictions may be generated regarding the determinants of lumpiness: The greater an economic agent's ability to process information, the less lumpy its behaviour. Thus one may speculate that advances in information technologies have reduced lumpiness[15], or that firms with a greater fraction of highly skilled workers have less lumpy behavior – this may help explain, for example, the finding by Doms and Dunne (1998) that smaller plants have a more lumpy adjustement, if one is willing to believe that smaller plants employ fewer skilled workers, or by Dhyne et al (2006, fig. 1) that some sectors (like gasoline) have much less lumpy price adjustment than others (like haircuts).

[15]This is the message of the empirical study by Bartel et al. (2005) in the dimension of product diversity.

REFERENCES

Bartel, Ann, Ann P., Casey Ichniowski and Kathryn L. Shaw (2005), "How Does Information Technology Really Affect Productivity? Plant-Level Comparisons of Product Innovation, Process Improvement and Worker Skills"NBER WP #11773.

Blanchard, Olivier, and Nobuhiro Kiyotaki (1987), "Monopolistic Competition and the Effects of Aggregate Demand Monopolistic Competition and the Effects of Aggregate Demand", *American Economic Review*, 77, 4, 647-666

Burstein, Ariel and Christian Hellwig (2007), "Prices and Market Shares in a Menu Cost Model", NBER working paper.

Caballero, Ricardo and Eduardo M. R. A. Engel (1993), "Heterogeneity and Output Fluctuations in a Dynamic Menu-Cost Economy" *Review of Economic Studies*, 60, 1, 95-119.

Caballero, Ricardo, Engel, Eduardo and John Haltiwanger (1997), "Aggregate Employment Dynamics: Building from Microeconomic Evidence", *American Economic Review,* 87, 1, 115-137

Caplin, Andrew S. and Daniel F. Spulber (1987), "Menu Costs and the Neutrality of Money" *Quarterly Journal of Economics*, 102, 4, 703-726

Davis, Steven J., John Haltiwanger and Scott Schuh (1996), *Job creation and destruction.* Cambridge MA: MIT Press

Dhyne, Emmanuel, Luis J. Alvarez, Hervé Le Bihan, Giovanni Veronese, Daniel Dias, Johannes Hoffmann, Nicole Jonker, Patrick Lünnemann, Fabio Rumler and Jouko Vilmunen (2006), "Price Changes in the Euro Area and the United States: Some Facts from Individual Consumer Price Data" *Journal of Economic Perspectives,* 20, 2, 171-192

Doms, Mark E. and Timothy Dunne (1998), "Capital Adjustment Patterns in Manufacturing Plants" *Review of Economic Dynamics*, 1(2), 409-429.

Eichenbaum, Martin, Nir Jaimovich and Sergio Rebelo (2008), "Reference Prices and Nominal Rigidities", NBER Working Paper 13829.

Klenow, Peter and Oleksiy Kryvtsov (2008), "State-dependent or time-

dependent pricing: does it matter for recent US inflation?", *Quarterly Journal of Economics,* 123, 3, 863-904

Lucas, Robert E. (1972), "Expectations and the neutrality of money", *Journal of Economic Theory*

——————— (1973), "International evidence on the output-inflation trade-off", *American Economic Review*

Luo, Yulei (2008): "Consumption Dynamics under Information Processing Constraints."

Review of Economic Dynamics, 11: 366-385.

Mackowiak, Bartosz and Mirko Wiederholt (2009a): "Optimal Sticky Prices under Rational

Inattention." *American Economic Review*, 99, 769-803.

Mackowiak, Bartosz and Mirko Wiederholt (2009b), "Business Cycle Dynamics under Rational Inattention", mimeo, ECB.

Mankiw, G., Reis, R., 2002. "Sticky information versus sticky prices: A proposal to replace the New Keynesian Phillips curve", *Quarterly Journal of Economics* 117 (4), 1295-1328.

Matejka, Filip (2008), "Rigid Pricing and rationally Inattentive Consumer", Working Paper, Princeton University

Matejka, Filip and Christopher A. Sims (2010), "Discrete Actions in Information Constrained Tracking Problems", Working Paper, Princeton University

Moscarini, G., (2004) "Limited information capacity as a source of inertia", *Journal of Economic Dynamics and Control* 28(10), 2003-2035.

Paciello, Luigi (2007), "The Response of Prices to Technology and Monetary Policy Shocks under Rational Inattention", mimeo, Einaudi Institute for Economics and Finance

Saint-Paul, Gilles (2005) "Some evolutionary foundations for price level rigidity", *American Economic Review.*

Shannon, Claude (1948), "A Mathematical Theory of Communication", *Bell System Technical Journal*, vol. 27, p. 379-423

Sims, Christopher A. (2003): "Implications of Rational Inattention." *Journal of Monetary*

Economics, 50(3), 665-690.

Sims, Christopher A. (2006): "Rational Inattention: Beyond the Linear-Quadratic

Case." *American Economic Review*, 96(2): 158-163.

Weitzman, Martin L. "The Simple Macroeconomics of Profit Sharing", *American Economic Review*, 75, 5 (1985), 937-953

APPENDIX

Proof of Theorem 1.

We have that

$$M(\varepsilon) = H(Y;\varepsilon) - H_X(Y;\varepsilon).$$

Let us compute $H_x(Y;\varepsilon)$. We denote by $p(y)$ the distribution of y and by $g(x) = \int_y f_\varepsilon(x \mid y)p(y)dy$ the unconditional distribution of x. By Bayes's law the conditional distribution of y is $f_\varepsilon(y \mid x) = \frac{f_\varepsilon(x|y)p(y)}{g(x)}$. Therefore we have that

$$\begin{aligned}
H_x(Y;\varepsilon) &= \int_x \left(-\int_y \frac{f_\varepsilon(x \mid y)p(y)}{g(x)} \log \frac{f_\varepsilon(x \mid y)p(y)}{g(x)} \right) g(x)dx \\
&= \int_x \left(-\int_y f_\varepsilon(x \mid y)p(y) \log \frac{f_\varepsilon(x \mid y)p(y)}{g(x)} \right) dx \\
&= \int_x \left(-\int_y \frac{1}{\varepsilon}\hat{f}\left(\frac{x-x(y)}{\varepsilon}\right) p(y) \left[\begin{array}{c} \log\left(\frac{1}{\varepsilon}\hat{f}\left(\frac{x-x(y)}{\varepsilon}\right)p(y)\right) \\ -\log\left(\int_u \frac{1}{\varepsilon}\hat{f}\left(\frac{x-x(u)}{\varepsilon}\right)p(u)du\right) \end{array} \right] dy \right) dx \\
&= I_1 + I_2,
\end{aligned}$$

where

$$I_1 = -\int_x \left(\int_y \frac{1}{\varepsilon}\hat{f}\left(\frac{x-x(y)}{\varepsilon}\right) p(y) \log\left(\frac{1}{\varepsilon}\hat{f}\left(\frac{x-x(y)}{\varepsilon}\right)p(y)\right) dy \right) dx$$

and

$$I_2 = \int_x \left(\int_y \frac{1}{\varepsilon}\hat{f}\left(\frac{x-x(y)}{\varepsilon}\right) p(y)dy \right) \log\left(\int_u \frac{1}{\varepsilon}\hat{f}\left(\frac{x-x(u)}{\varepsilon}\right)p(u)du \right) dx.$$

Let $H_n(Y) = -\int_{T_n} p(y)\log p(y)dy$. Clearly, $\sum_n H_n(Y) = H(Y)$. Furthermore,

$$\begin{aligned}
&\int_y \frac{1}{\varepsilon}\hat{f}\left(\frac{x-x(y)}{\varepsilon}\right) p(y) \log\left(\frac{1}{\varepsilon}\hat{f}\left(\frac{x-x(y)}{\varepsilon}\right)p(y)\right) dy \\
= &\sum_n \int_{T_n} \frac{1}{\varepsilon}\hat{f}\left(\frac{x-x_n}{\varepsilon}\right) p(y) \log\left(\frac{1}{\varepsilon}\hat{f}\left(\frac{x-x_n}{\varepsilon}\right)p(y)\right) dy \\
= &\sum_n \frac{1}{\varepsilon}\hat{f}\left(\frac{x-x_n}{\varepsilon}\right) \left(\log\left(\frac{1}{\varepsilon}\hat{f}\left(\frac{x-x_n}{\varepsilon}\right)\right) P_n + H_n(Y)\right).
\end{aligned}$$

Therefore,

$$\begin{aligned} I_1 &= -\int_x \left[\sum_n \frac{1}{\varepsilon} \hat{f}\left(\frac{x-x_n}{\varepsilon}\right) \left(\log\left(\frac{1}{\varepsilon} \hat{f}\left(\frac{x-x_n}{\varepsilon}\right)\right) P_n + H_n(Y)\right)\right] dx \\ &= -\sum_n P_n \int_x \frac{1}{\varepsilon} \hat{f}\left(\frac{x-x_n}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon} \hat{f}\left(\frac{x-x_n}{\varepsilon}\right)\right) dx - \sum_n H_n(Y) \int_x \frac{1}{\varepsilon} \hat{f}\left(\frac{x-x_n}{\varepsilon}\right) dx \\ &= \log\varepsilon + H_{\hat{f}} + H(Y), \end{aligned}$$

where $H_{\hat{f}} = -\int_z \hat{f}(z) \log \hat{f}(z) dz$ is the entropy of distribution $\hat{f}()$ and the last equality can be obtained straightforwardly by considering the variable change $z = \frac{x-x_n}{\varepsilon}$.

Next, we have that

$$\int_y \frac{1}{\varepsilon} \hat{f}\left(\frac{x-x(y)}{\varepsilon}\right) p(y) dy = \sum_n \frac{1}{\varepsilon} \hat{f}\left(\frac{x-x_n}{\varepsilon}\right) P_n.$$

Therefore,

$$\begin{aligned} I_2 &= \int_x \sum_n \frac{1}{\varepsilon} \hat{f}\left(\frac{x-x_n}{\varepsilon}\right) P_n \log\left(\sum_k \frac{1}{\varepsilon} \hat{f}\left(\frac{x-x_k}{\varepsilon}\right) P_k\right) dx \\ &= -\log\varepsilon + \sum_n P_n \int_x \frac{1}{\varepsilon} \hat{f}\left(\frac{x-x_n}{\varepsilon}\right) \log\left(\sum_k \hat{f}\left(\frac{x-x_k}{\varepsilon}\right) P_k\right) dx \\ &= -\log\varepsilon + \sum_n P_n \int_z \hat{f}(z) \log\left(\sum_k \hat{f}\left(z + \frac{x_n - x_k}{\varepsilon}\right) P_k\right) dz. \end{aligned}$$

Consider the function $g_n(z;\varepsilon) = \hat{f}(z) \log\left(\sum_k \hat{f}\left(z + \frac{x_n - x_k}{\varepsilon}\right) P_k\right)$. Clearly, as $\varepsilon \to 0$, it converges simply to $g_n(z) = \hat{f}(z) \log\left(\hat{f}(z) P_n\right)$. Furthermore, $|g_n(z;\varepsilon)| \leq \hat{f}(z) \left| \log \hat{M} \right|$, where $\hat{M} = \max \hat{f}$. According to the Dominated Convergence Theorem, it follows that $\lim_{\varepsilon\to 0} \int_z g_n(z;\varepsilon) dz = \int_z g_n(z) dz$.

From this we get that

$$\begin{aligned} \lim_{\varepsilon\to 0} (I_2 + \log\varepsilon) &= \sum_n P_n \int_{-\infty}^{+\infty} \hat{f}(z) \log\left(\hat{f}(z) P_n\right). \\ &= -H_{\hat{f}} - S(X). \end{aligned}$$

Therefore:

$$\begin{aligned}\lim_{\varepsilon\to 0} H_X(Y;\varepsilon) &= \lim_{\varepsilon\to 0} \log\varepsilon + H_{\hat{f}} + H(Y) + I_2 \\ &= H_{\hat{f}} + H(Y) - H_{\hat{f}} - S(X) \\ &= H(Y) - S(X).\end{aligned}$$

Consequently, $M(\varepsilon) = H(Y) - H_X(Y;\varepsilon)$ converges to $S(X)$ as $\varepsilon \to 0$. QED.

Proof of Theorem 2.

Assume $i = 1$ to fix ideas. Let $y, y' \in \mathring{T}_1$ Let $\theta \in [0,1]$ and $y'' = \theta y + (1-\theta)y'$. Assume $y'' \in T_k$.

Then by the no-swapping condition we must have

$$V(x_k, y'') + V(x_1, y) \geq V(x_1, y'') + V(x_k, y), \tag{13}$$

and similarly for y' replacing y :

$$V(x_k, y'') + V(x_1, y') \geq V(x_1, y'') + V(x_k, y'), \tag{14}$$

First, consider the case where $P(x_1, x_k, y; y')$ holds. Consider the function

$$h(\phi) = V(x_k, \phi y + (1-\phi)y') - V(x_1, \phi y + (1-\phi)y').$$

Then, applying the NSC to y and y', we get that

$$h(0) \leq h(\theta) \tag{15}$$

and

$$h(1) \leq h(\theta). \tag{16}$$

However, $h'(\phi) = (y-y')\cdot(\nabla_2 V(x_k, \phi y + (1-\phi)y') - \nabla_2 V(x_1, \phi y + (1-\phi)y'))$. Suppose $x_k \neq x_1$. Since $P(x_1, x_1, y; y')$ holds, $h'()$ has a constant nonzero sign over $[0,1]$, so that $h()$ is strictly monotonous and (15) and (16) cannot hold simultaneously. Therefore, it must be that $x_k = x_1$, i.e. $y'' \in T_1$.

Assume now that $P(x_1, x_k, y; y')$ does not hold. Since $y' \in \mathring{T}_1$, there exists a bowl $B(y', r) \subset T_1$. Furthermore,$U = \bigcup_{i=2}^{N} M(x_1, x_i, y)$ is a finite reunion of

manifolds, i.e. is at most of dimension $n-1$, implying that $B(y',r)-U$ is dense in $B(y',r)$. Thus there exists a sequence $\{\tilde{y}_i, i=1,...\}$ such that $\tilde{y}_i \to y'$ and $\tilde{y}_i \in B(y',r)-U$. By construction, $\tilde{y}_i \in T_1$ and $P(x_1, x_l, y, \tilde{y}_i)$ holds for any i and $l \neq 1$. In particular, this holds for $l(i)$ such that $\hat{y}_i = \theta y + (1-\theta)\tilde{y}_i \in T_{l(i)}$, if $l(i) \neq 1$. In such a case, we can apply the above reasoning with $\tilde{y}_i$ replacing y' and $\hat{y}_i$ replacing y'' and show that $l(i)=1$. Therefore, $\hat{y}_i \in T_1$ for all i. Since, by continuity, $\hat{y}_i \to y''$, it follows that $y'' \in \bar{T}_1$.

Thus we have proved that $C(\mathring{T}_1) \subset \bar{T}_1$. Furthermore, since $\mathring{T}_1$ is dense, $C(T_1) \subset C\left(\overline{\mathring{T}_1}\right) \subset \overline{C(\mathring{T}_1)} \subset \bar{T}_1$.

QED

Proof of Theorem 3.

Take $i=1$ and $j=2$ to fix ideas. Let B_1,B_2 be two open bowls such that $B_1 \subset T_1$ and $B_2 \subset T_2$. Since B_1 and B_2 are connected and $u()$ is continuous, $u(B_1)$ and $u(B_2)$ are intervals and so is $S = u(B_1) \cap u(B_2)$. Assume $\mathring{S} = (a,b)$, with $a < b$. Assume $V(x_1,u) - V(x_2,u)$ is strictly increasing in u. This is without loss of generality since u can always be redefined as $-u$. Let $\mu_1(z) = \mu(u^{-1}([a,z]) \cap B_1)$, where $\mu()$ denotes the measure on R^p defined by $f()$. For any interval $(c,d) \subset (a,b)$, we have that $(c,d) \subset u(B_1)$, implying $u^{-1}((c,d)) \cap B_1 \neq \varnothing$. Since $u^{-1}((c,d))$ and B_1 are open sets, so is their intersection, which contains a bowl. Since $f()$ is a density with full support, $\mu(u^{-1}((c,d)) \cap B_1) > 0$. It follows that μ_1 is strictly increasing in z, and furthermore since f is a density, $\mu(u^{-1}((c,d)) \cap B_1)$ is arbitrarily small for d arbitrarily close to c. Consequently, μ_1 is also continuous. Similarly we define $\mu_2(z) = \mu(u^{-1}(z,b) \cap B_2)$, which is continuous and strictly decreasing. By continuity, we can pick a' and b' such that $a < a' < b' < b$ and $\mu_1(a') = \mu_2(b') > 0$. The set $S_1 = u^{-1}([a,z]) \cap B_1$ is such that $\mu(S_1) > 0$ and $S_1 \subset T_1$. Similarly, $S_2 = u^{-1}([z,b]) \cap B_2$ is such that $\mu(S_2) = \mu(S_1) > 0$ and $S_2 \subset T_2$. Consider a new assignment such that T_1 is replaced by $(T_1 - S_1) \cup S_2$, and still assigned to x_1, while T_2 is replaced by $(T_2 - S_2) \cup S_1$, and still assigned to x_2, while all the other clusters are unchanged. Clearly,

the new assignment has the same entropy as the initial one, and therefore matches the information capacity constraint. Furthermore, the change in the objective function induced by such a swap is $\Delta = \int_{S_1}(V(x_2,u(y)) - V(x_1,u(y)))f(y)dy + \int_{S_2}(V(x_1,u(y)) - V(x_2,u(y)))f(y)dy$. By construction $u(y) > u(y')$ for $y \in S_2$ and $y' \in S_1$, implying that $V(x_1,u(y)) - V(x_2,u(y)) > V(x_1,u(y')) - V(x_2,u(y'))$. Since $\int_{S_1} f(y)dy = \int_{S_2} f(y)dy$, it follows that $\Delta > 0$, which is impossible since the initial assignment is assumed optimal.

Therefore, we must have $\mathring{S} = \varnothing$, implying that S has at most one point. Consider now $z \in u(\mathring{T}_1) \cap u(\mathring{T}_2)$. By definition there exist two open bowls B_1 and B_2 such that $y_1 \in B_1 \subset T_1, y_2 \in B_2 \subset T_2$, and $u(y_1) = u(y_2) = z$. The preceding argument implies that $u(B_1) \cap u(B_2) = \{z\}$. But, since $u(B_1)$ and $u(B_2)$ are both intervals, this can only happen if z is at the boundary of each. Assume for example that $u(B_1) = (a,z]$. Since B_1 is open and $y_1 \in B_1$, for any unitary vector v there exists $\lambda > 0$ such that $[y_1 - \lambda v, y_1 + \lambda v] \subset B_1$. Over that interval, $u()$ must reach a local extremum at y_1, implying that it is singular at y_1, which is ruled out by assumption. Thus we have a contradiction and it must be that $u(\mathring{T}_1) \cap u(\mathring{T}_2) = \varnothing$. QED.

Proof of Theorem 4.

Below I will denote by $\{X_i, x_i\}$ any quantized solution with partition $\{X_i\}$ such that $x = x_i$ over X_i, and by $V\{X_i, x_i\}$ the corresponding value of the objective function. Consider $\{T_i, \tilde{x}_i\}$ which solves problem $P1$. Then theorem 3 implies that $\{u(\mathring{T}_i)\}$ form a partition of $\cup_i u(\mathring{T}_i)$. Furthermore, we have that $\mu(T_i) = \mu(\mathring{T}_i)$ and $\mathring{T}_i \subset u^{-1}(u(\mathring{T}_i))$, so that $\mu(\mathring{T}_i) \leq m(u(\mathring{T}_i))$. Since $\sum_i \mu(\mathring{T}_i) = \sum_i \mu(T_i) = 1$, it must be that $\mu(\mathring{T}_i) = m(u(\mathring{T}_i))$ for all i and that $\sum_i m(u(\mathring{T}_i)) = 1$. Therefore, $\{u(\mathring{T}_i)\}$ is almost surely a partition of $u(R^p)$ and $V\{u(\mathring{T}_i), \tilde{x}_i\} \leq V\{U_i, x_i\}$. Furthermore, since $\mu(T_i) = m(u(\mathring{T}_i))$, $\{T_i\}$ and $\{u(\mathring{T}_i)\}$ have the same mutual information. Finally, $V\{T_i, \tilde{x}_i\} = \sum_i \int_{T_i} V(\tilde{x}_i, u(y))f(y)dy = \sum_i \int_{\mathring{T}_i} V(\tilde{x}_i, u(y))f(y)dy = \sum_i \int_{u^{-1}(u(\mathring{T}_i))} V(\tilde{x}_i, u(y))f(y)dy = \sum_i \int_{u(\mathring{T}_i)} V(\tilde{x}_i, u)dm(u)$, where the equality again comes from the fact that $\mu(T_i) = m(u(\mathring{T}_i)) = \mu(u^{-1}(u(\mathring{T}_i))$. Thus $V\{T_i, \tilde{x}_i\} = V\{u(\mathring{T}_i), \tilde{x}_i\} \leq V\{U_i, x_i\}$. By a similar argument we can prove that $\{u^{-1}(U_i), x_i\}$ has the same mutual

information as $\{U_i, x_i\}$ and that $V\{U_i, x_i\} = V\ \{u^{-1}(U_i), x_i\} \leq V\{T_i, \tilde{x}_i\}$. Therefore, $V\{U_i, x_i\} = V\{T_i, \tilde{x}_i\}$, implying that $\{u^{-1}(U_i), x_i\}$ solves $P1$.

QED.

Proof of Theorem 5.

Lemma 1 implies that any optimum must be a partition by intervals, up to a set of measure zero. It follows that one cannot improve on such a partition. By Lemma 2, for any partition the optimal x_n must be $\frac{y_n+y_{n+1}}{2}$, which proves (v). Next, computing the value function for such a configuration, we get that

$$E(x(y) - y)^2 = \frac{1}{12}\sum_{n=0}^{N-1}(y_{n+1} - y_n)^3. \tag{17}$$

For a given N, we minimize (17) subject to

$$\begin{aligned} y_0 &= 0, \\ y_N &= 1, \\ -\sum_{n=0}^{N-1}(y_{n+1} - y_n)\ln(y_{n+1} - y_n) &\leq K. \end{aligned}$$

The FOCs are:

$$(y_n - y_{n-1})^2 - (y_{n+1} - y_n)^2 = \lambda(\ln(y_n - y_{n-1}) - \ln(y_{n+1} - y_n)),\ 0 < n < N. \tag{18}$$

Note that with N fixed but absent the capacity constraint, optimality would imply that $y_n - y_{n-1} = y_{n+1} - y_n$. All intervals would then be of constant length $1/N$ and the resulting entropy would be $\ln N$. Thus, if $\ln N < K$, then $\lambda = 0$ and the optimal solution is the unconstrained one. However, one can always improve on this by picking a larger N, since the initial configuration can always be replicated by collapsing the additional interval to a set of measure zero by equating their bounds. Therefore the optimal N will be such that $\ln N \geq K$, i.e. the capacity constraint will be binding. Let us then

consider such an N. Call Δ_n the length of interval n. The FOC (18) implies that $\Delta_n^2 - \lambda \ln \Delta_n$ is invariant across intervals. Since the function $X^2 - \lambda \ln X$ is U-shaped, Δ_n can at most have two values, let us call them Δ and Δ'. Clearly, the invariance property is then satisfied for $\lambda = \frac{\Delta'^2 - \Delta^2}{\ln \Delta' - \ln \Delta}$. Without loss of generality, assume $\Delta \leq \Delta'$. Let q the number of intervals of length Δ. Since the whole [0,1] interval must be partitioned, it must be that

$$q\Delta + (N - q)\Delta' = 1$$

and

$$-q\Delta \ln \Delta - (N - q)\Delta' \ln \Delta' = K.$$

Eliminating Δ', we get

$$\Delta' = \frac{1 - q\Delta}{N - q},$$

and we see that Δ must solve

$$\phi(\Delta) = -q\Delta \ln \Delta - (1 - q\Delta) \ln \left(\frac{1 - q\Delta}{N - q} \right) = K. \tag{19}$$

The function $\phi(\Delta)$ is increasing and then decreasing and reaches its maximum at $\Delta = 1/N$, at which point we also have $\Delta' = 1/N$. Therefore, there is at most one solution Δ such that $\Delta \leq \Delta'$. Furthermore, $\phi(0) = \ln(N - q)$ and $\phi(1/N) = \ln N$. Therefore, there exists a solution for Δ provided

$$\ln(N - q) < K \leq \ln N.$$

In particular, for any N such that $\ln N \geq K$ the set of values of q for which this holds is non empty, since $\ln(N - q) = 0$ for $q = N - 1$.

Despite that q is integer, equation (19) also defines a value of Δ for any real number q. Furthermore,

$$\frac{\partial \phi}{\partial q} = \Delta(\ln \Delta' - \ln \Delta) + \Delta - \Delta' < 0.^{16}$$

Since $\phi'(\Delta) > 0$, it follows that $\frac{d\Delta}{dq} > 0$.

[16] It can be checked that this expression is always negative by noting that it would be equal to zero at $\Delta = \Delta'$ and that its derivative with respect to Δ' is $\Delta/\Delta' - 1 < 0$.

Next, note that the resulting loss function, up to a positive multiplicative constant, is equal to $V = q\Delta^3 + (N-q)\Delta'^3$. Differentiating, we get

$$\begin{aligned} dV &= (\Delta^3 - \Delta'^3)dq - 3q\Delta'^2 dq + 3\Delta'^3 dq + 3q\Delta^2 d\Delta - 3q\Delta'^2 d\Delta \\ &= (\Delta^3 - \Delta'^3)dq + 3\Delta'^2(\Delta' - q)dq + 3q(\Delta^2 - \Delta'^2)d\Delta. \end{aligned}$$

Since $\frac{d\Delta}{dq} > 0$, $\Delta < \Delta'$, and $\Delta' < 1 \leq q$, all terms are negative if $dq > 0$. Therefore, V is a decreasing function of q; given N, the optimal value of q is the largest possible one, i.e. $q = N - 1$. The resulting loss function is then

$$V = (N-1)\Delta^3 + (1-(N-1)\Delta)^3, \tag{20}$$

and Δ now solves

$$\tilde{\phi}(\Delta) = -(N-1)\Delta \ln \Delta - (1-(N-1)\Delta)\ln(1-(N-1)\Delta) = K. \tag{21}$$

What is the optimal value of N? First of all, differentiating $\tilde{\phi}$ with respect to N and Δ we get

$$\frac{d\Delta}{dN} = -\frac{\Delta}{N-1}\left(1 + \frac{1}{\ln \Delta' - \ln \Delta}\right) < 0. \tag{22}$$

Next, differentiating (20) and using (22) we get that

$$\frac{dV}{dN} = \Delta^3 - 3\Delta\Delta'^2 + 3\Delta(\Delta'^2 - \Delta^2)\left(1 + \frac{1}{\ln \Delta' - \ln \Delta}\right).$$

This expression is positive if and only if

$$2\Delta^2 < \frac{3(\Delta + \Delta')(\Delta' - \Delta)}{\ln \Delta' - \ln \Delta}.$$

Calling $\theta = \Delta'/\Delta > 1$, this is equivalent to $\ln \theta < 3(\theta^2 - 1)/2$, which is always true.

Thus $dV/dN > 0$. Consequently, the optimal value of N is the smallest one such that $\ln N \geq K$, i.e. $N = INT(e^K)$.

QED

Derivation of (2)-(3).

The budget constraint of the individual is

$$\int_0^1 p_i c_{ij} + m_j \leq y_j + s_j,$$

where $y_j = p_j x_j$ is his income and s_j is rebated seignioriage. In equilibrium the total money stock is $M = \int_0^1 m_j dj$ and we assume for simplicity that seignoriage is rebated proportionally to the value of output produced by the individual:

$$s_j = M\frac{y_j}{Y}, \forall j,$$

where

$$Y = \int_0^1 y_j dj$$

is GDP. Aggregate real output is defined as $X = \left(\int_0^1 x_j^\alpha dj\right)^{1/\alpha}$.

We assume that the money stock is drawn from a distribution with density $f(M)$ and c.d.f $F(M)$. We also assume that the idiosyncratic shock is drawn from a distribution with density $h(z)$ and cumulative $H(z)$.

Solving for the consumer's optimal consumption and money holdings yields, after a few steps, the following relationship:

$$c_{ij} = \frac{m_j}{p_i^{\frac{1}{1-\alpha}} p^{-\frac{\alpha}{1-\alpha}}}. \tag{23}$$

Aggregating across individuals, this gives the demand curve for good i :

$$C_i = \frac{M}{p_i^{\frac{1}{1-\alpha}} p^{-\frac{\alpha}{1-\alpha}}}. \tag{24}$$

We assume that all producers meet demand. Therefore, $x_j = C_j$.Next,

$$\begin{aligned} Y &= \int p_j x_j dj \\ &= \int p_j C_j dj \\ &= M. \end{aligned}$$

We can also check that $X = \left(\int C_i^\alpha di\right)^{1/\alpha} = M/p$.

Furthermore, aggregating (23) across goods we see that the aggregate consumption index for individual j is equal to

$$c_j = \frac{m_j}{p}.$$

We also have that $\int_0^1 p_i c_{ij} = m_j = pc_j$. Substituting into the budget constraint, we get that

$$\begin{aligned} m_j &= \frac{y_j + s_j}{2}; \\ c_j &= \frac{y_j + s_j}{2p}. \end{aligned}$$

Noting that $s_j = M\frac{y_j}{Y}$ and $y_j = p_j x_j$ we get an indirect utility function

$$\begin{aligned} V_j &= E \ln \left[c_j^{1/2} \left(\frac{m_j}{p} \right)^{1/2} \left(\frac{M}{p} \right)^{-\psi} - z_j x_j^{1+\mu} \right] \\ &= E \ln \left[\frac{p_j x_j (1 + M/Y)}{2p} \left(\frac{M}{p} \right)^{-\psi} - z_j x_j^{1+\mu} \right] \\ &= .E \ln \left[\frac{p_j x_j}{p} \left(\frac{M}{p} \right)^{-\psi} - z_j x_j^{1+\mu} \right] \end{aligned} \tag{25}$$

It is this quantity that the individual maximizes when setting his price p_j subject to the demand curve (24). Substituting this demand curve into (25) we can rewrite the objective function of the producer as

$$\begin{aligned} V_j &= E \ln \left[\left(\frac{p_j}{p} \right)^{-\frac{\alpha}{1-\alpha}} \frac{M}{p} \left(\frac{M}{p} \right)^{-\psi} - z_j \frac{M^{1+\mu}}{p_j^{\frac{1+\mu}{1-\alpha}} p^{-\frac{\alpha(1+\mu)}{1-\alpha}}} \right] \\ &= E \ln \left[p_j^{-\frac{\alpha}{1-\alpha}} - \phi_j p_j^{-\frac{1+\mu}{1-\alpha}} \right] + E \ln \left[p^{\frac{2\alpha-1}{1-\alpha}+\psi} M^{1-\psi} \right], \end{aligned}$$

where ϕ_j is defined by (3). This clearly amounts to maximizing (2).

Figure 1

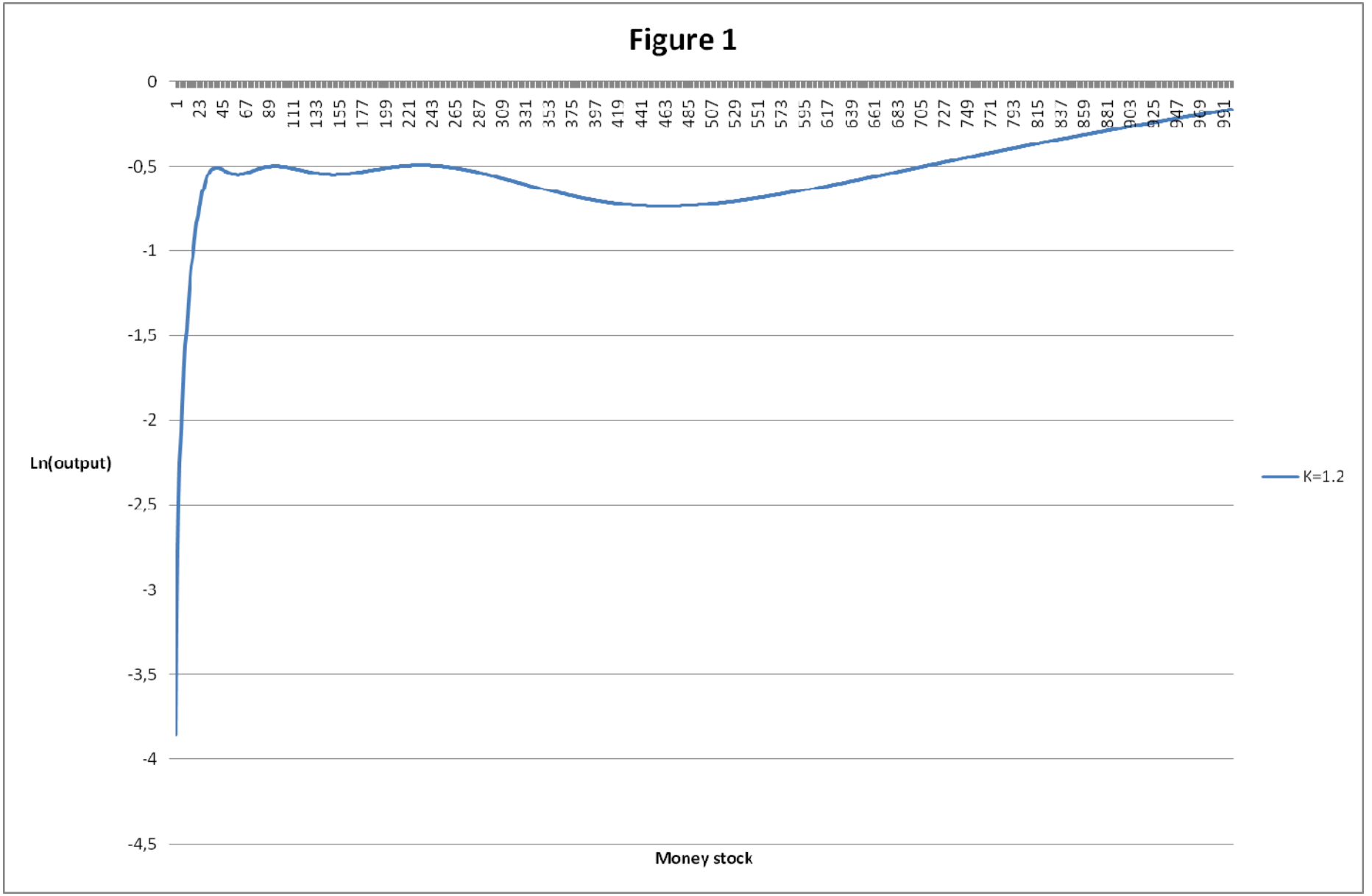
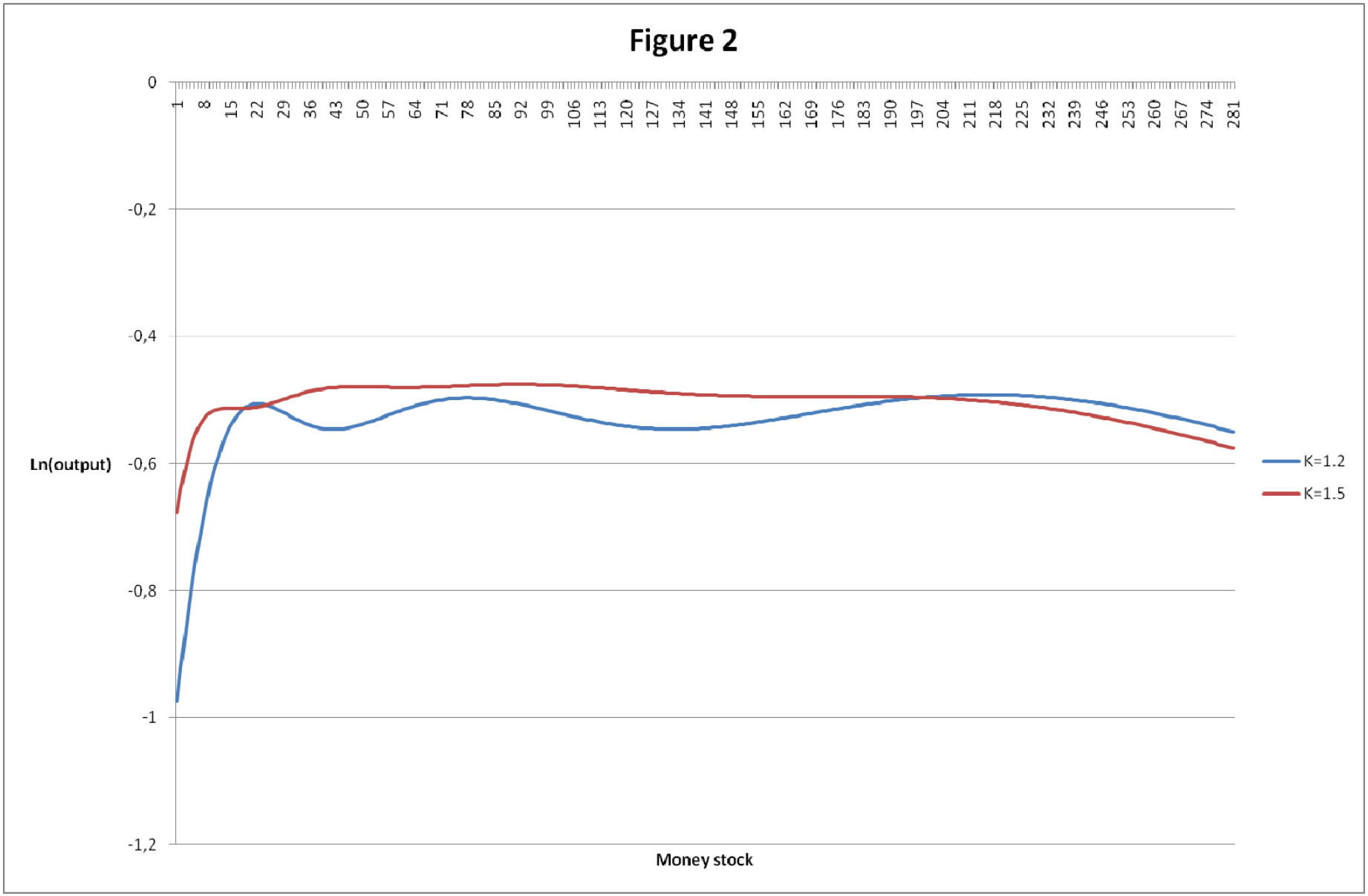

0
-0,5
-1
-1,5
-2
-2,5
-3
-3,5
-4
-4,5
Ln(output)
1
23
45
67
89
111
133
155
177
199
221
243
265
287
309
331
353
375
397
419
441
463
485
507
529
551
573
595
617
639
661
683
705
727
749
771
793
815
837
859
881
903
925
947
969
991
Money stock
K=1.2

Figure 2


0
-0,2
-0,4
-0,6
-0,8
-1
-1,2
Ln(output)
1
8
15
22
29
36
43
50
57
64
71
78
85
92
99
106
113
120
127
134
141
148
155
162
169
176
183
190
197
204
211
218
225
232
239
246
253
260
267
274
281
Money stock
K=1.2
K=1.5

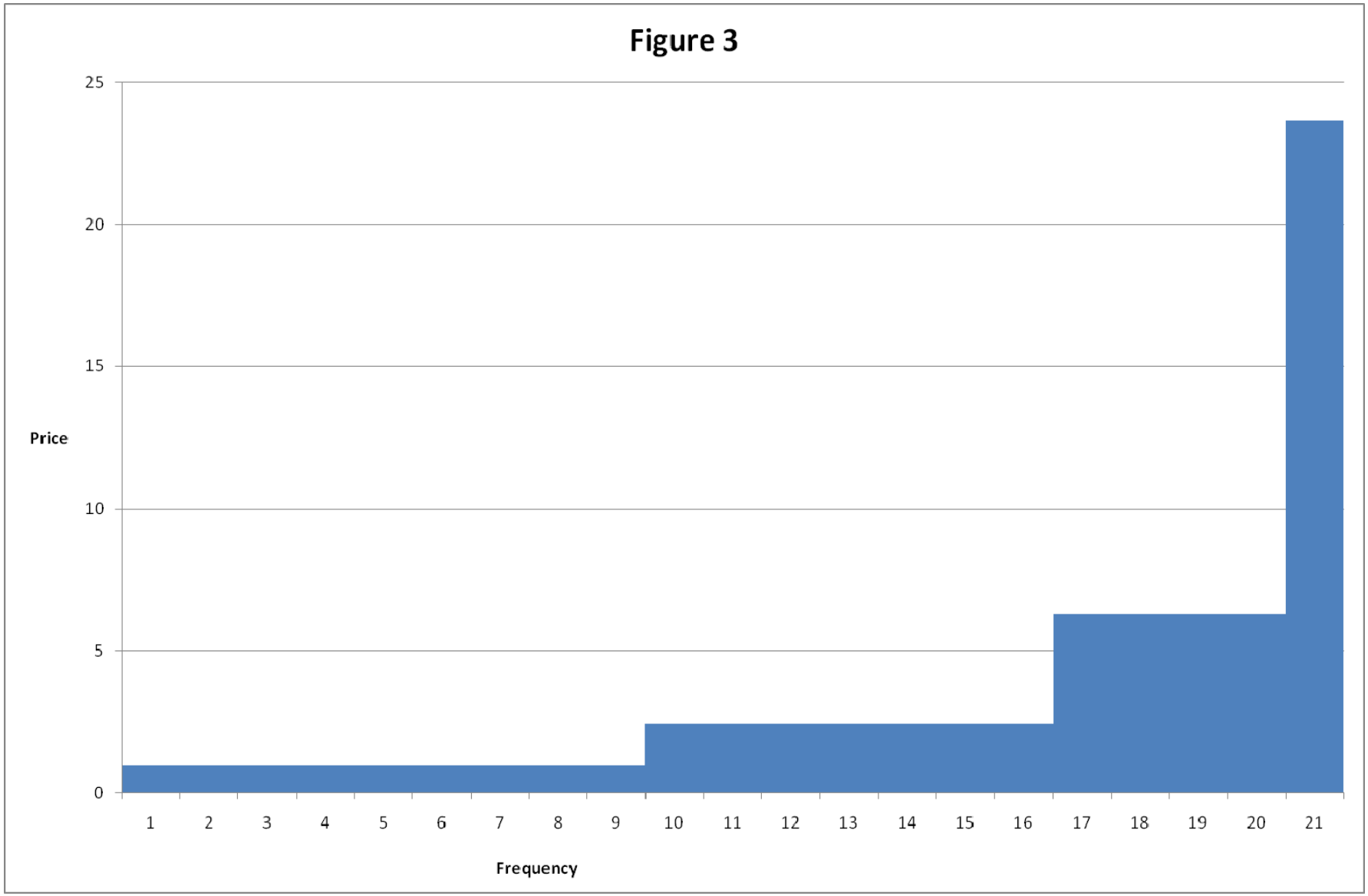


Figure 3
25
20
15
10
5
0
Price
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
Frequency

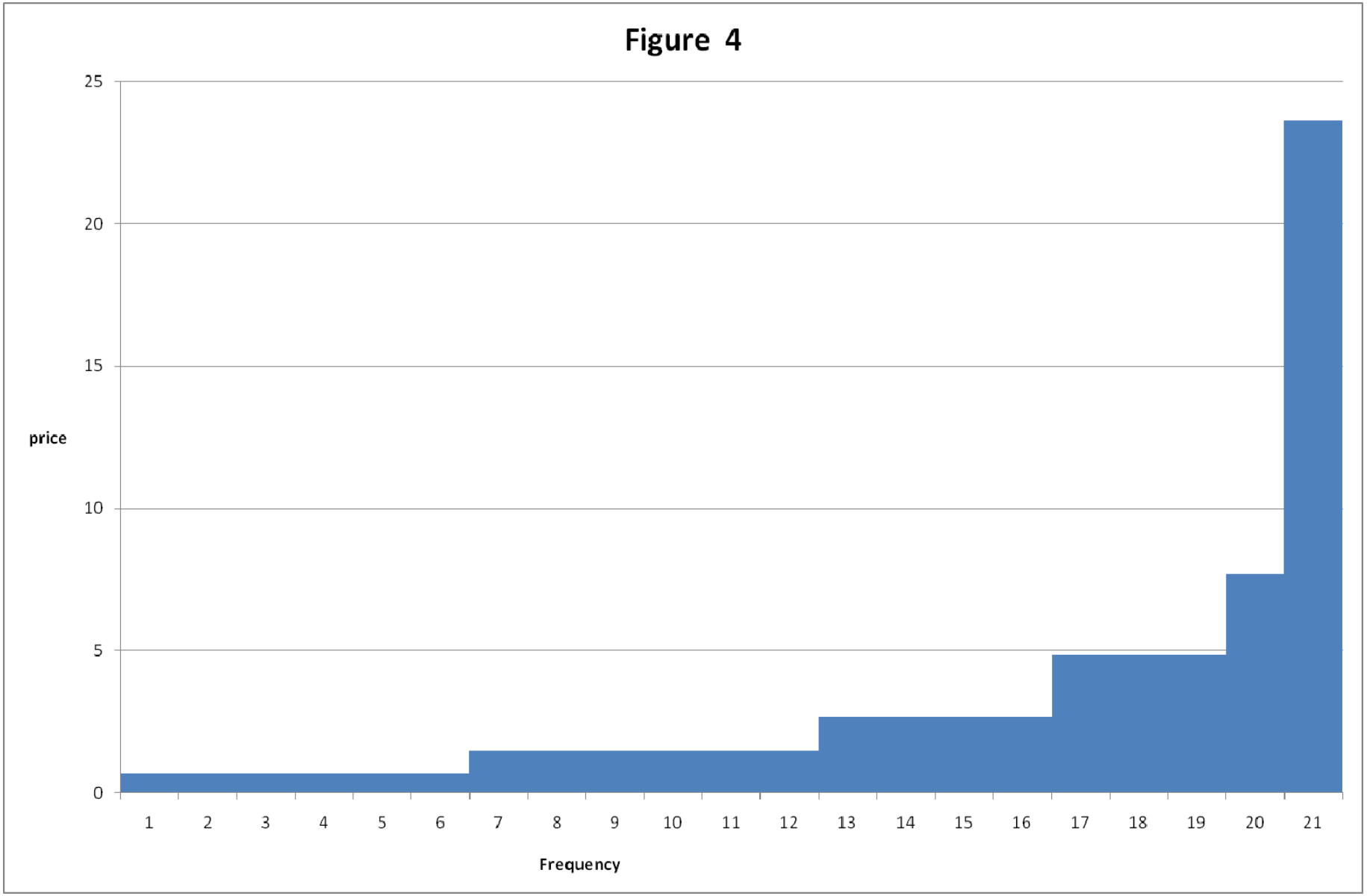


Figure 4
25
20
15
10
5
0
price
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
Frequency