Hiring Decisions with Knapsack Bandits

1061261

March 2023

Preface

The following research proposal is of type (ii) "economic theory drawing on ideas from machine learning" as stated in the syllabus of the course. The topic of the paper is strongly connected to the notions of Online Learning and (stochastic) Multi-Armed Bandits discussed in class (Topics 8, 9, 11 in [1]).

1 Introduction

This line of research explores the macro and micro modelling potential of Bandits with Knapsacks (`BwK`, [2]) in the context of structural Public Policy Design and, more specifically, the hiring decisions in labor markets with rolling admission systems. I ambition to provide an organic online learning framework to understand firms' behaviour and their response to structural shocks including minimum wage policies, mergers, prices caps, productivity shocks and others. These tools can be incorporated within larger macro / micro settings to model the agency of firms whenever information asymmetries are deemed relevant like in Acharya and Wee [3].

`BwK` is an active area of research which connects standard notions of exploration-exploitation trade-offs in multi-arm bandits (MAB) with "resource capacity constraints". Despite the inherent economic nature of the problem, where agents need to make sequential decisions in the presence of uncertainty while resourced-constrained, there have been few (if not any) attempts at introducing these results within standard economic modelling. In particular, I conjecture that current research in macro and micro mod-

elling could benefit from incorporating the notions of BwK as heuristic tools for the derivation of useful theory and policy relevant algorithms. I use the problem of the budget/space constrained wage-setting firm as an operating example.

2 Literature Review

Bandits with Knapsacks - BwK. Literature in MAB is both wide and deep ([4], [5]). However, the idea of resource constrained bandits or bandits with knapsacks (BwK) was only (recently) introduced by Badanidiyuru et al [2]. The idea is quite simple. Consider a fixed known set of m arms and d resources such that, at every time step i the agent plays action x_i. At the end of period i a reward r_{x_i} and a consumption vector C_{x_i} are realized. Define $v := (r, c_1, \ldots, c_d) \in [0,1] \times [0,1]^d$ which we assume follows a stochastic process with $\mathbb{E}[v_i \mid x_i] = \mu_{x_i}$. Crucially, there exists a hard bound vector $B : (B_1, \ldots, B_d)$ such that $\sum_i^\rho c_{ij} \leq B_j$ for all resources j, where ρ is the $\arg\min_\tau \sum_i^\tau c_{ij} = B_j$ given a policy π. This is, it is the earliest stopping time for which one of the resources is depleted. In this context, time K is just one resource which is consumed deterministically at rate 1 for all arms i, and consequently $\rho \leq K$. Note that there is no loss in replacing $B_j = B$ for all j, as one could just define $C'_j = C_j \cdot B/B_j$.

Concave Bandits with Convex Knapsacks - BwCR. Shortly after the publication of [2], Agrawal and Devanur [6] released an extension of BwK for a generalized reward function $f(v) : [0,1]^{d+1} \to [0,1]$ with f concave and L-Lipschitz, and an arbitrary convex set S representing the resource constrains (including reward r). Additionally, their framework allowed for some notion of regret minimization in terms of distance with respect to the target set S rather than strict hard bounds as proposed by [2]. They proposed near-optimal algorithms of $\tilde{O}\Big(L \mid\mid \mathbf{1}_d \mid\mid \sqrt{\frac{m}{K} \ln\left(\frac{mKd}{\delta}\right)}\Big)$ for the reward regret and $\tilde{O}\Big(\mid\mid \mathbf{1}_d \mid\mid \sqrt{\frac{m}{K} \ln\left(\frac{mKd}{\delta}\right)}\Big)$ for the resource regret, where $\mid\mid \mathbf{1}_d \mid\mid$ is the norm of the d column vector of ones induced by the same notion of distance that we used in the characterization of L-Lipschitzness.

Crucially, regret is measured with respect to OPT, the "optimal dynamic policy". Unfortunately, OPT might be very difficult to characterize even with known DGP. As a result, OPT is frequently replaced by $\texttt{OPT}_{\text{LP}}$ which characterizes a static probability mixture of arms rather than a single arm x^* as in the standard MAB problem. $\texttt{OPT}_{\text{LP}}$ is defined as the solution of a tractable LP, such that $\texttt{OPT}_{\text{LP}} > \texttt{OPT}$. Regret is shown to be

sublinear wrt to $\texttt{OPT}_{\text{LP}}$, as K-sublinear regret could be an excessively weak requirement in cases where $B << K$. This model is referred as Bandits with Concave Rewards and Convex Knapsacks (BwCR) and it will be our workhorse for economic modelling in Section 3.

Adaptive Policy Design. This research proposal builds upon the recent literature on Adaptive Policy Design. As early as 2003, Kleinberg et al. [7] introduced bandit notions within the firm context, more specifically, by characterizing the monopolist price-setting problem. More recently, Cesa-Bianchi et al. [8] characterized a strictly more difficult problem, the bilateral-trade problem, using similar machinery. Finally, Cesa-Bianchi et al. [9] presented an intermediate problem (adaptive welfare maximization) with interesting regret guarantees in the context of tax setting. These ideas were further extended to standard adverse selection contexts with two potentially correlated latent variables by Gonzalez [10]. In this paper, Gonzalez shows regret bounds of $\tilde{O}(K^{2/3})$, matching the results of the weakly easier problem in [9].

Our problem is most similar in terms of structure to the context presented by Gonzalez, but the tools and heuristics should be easily accommodated to previous literature. In fact, [2] show BwK guarantees for the monopolist price setting problem under additive linear reward function f and hard upper-bounded resource constraints. In this sense, we ambition to apply the more general prescriptions of [6] to provide similar flavor results for the adaptive welfare maximization problem with constrained resources and diminishing returns. Additionally, we look forward to incorporating these heuristics within mainstream macro and micro modelling environments.

Macro and (Micro) Modelling. As an example of the economic modelling potential of the tools derived in this paper, we import some intuitions from Acharya and Wee [3]. In this paper, the authors point at the limiting information processing capacity of firms as a potential explanation for the slow employment recovery following a shock in general productivity. Although their model conveys interesting intuitions, the limited information processing capacity is not properly micro founded. In addition, the processing capacity of firms is not endogenously characterized within the model, but exogenously calibrated. In this paper, I ambition to obtain similar results to the ones delivered by the authors, while modelling the information and learning process in a more transparent, endogenous and flexible manner.

Connection to Class Material. The topic in this paper is strongly connected to the notions of Online Learning and exploration-exploitation trade-off taught in class. In particular, `BwCR` is simply a generalization of the standard (stochastic) MAB framework. Moreover, the suggested algorithm can be interpreted as a variation of well-known `UCB` algorithms [11].

3 Theoretical Framework

Setting. Consider the following framework. A wage-setting firm is presented with a sequence of $\{1, 2, \ldots, K\}$ employees, which are perfectly characterized by a duple $(u_i, v_i) \in \mathbb{R}^2_+$, where z_i represents the ith realization of the random variable Z. We assume that the sequence of duples follow an *iid* stochastic process for some unknown distribution function $F_{U,V}$. The problem of the firm can be summarized as

$$\begin{aligned} \max_{\{x_i\}_1^\rho \in \Omega_x} & f\big(\frac{1}{K}\sum_i^\rho \mathbb{1}(x_i \geq v_i)\cdot u_i\big) - \frac{1}{K}\sum_i^\rho \mathbb{1}(x_i \geq v_i)\cdot x_i \\ s.t. & \sum_i^\rho \mathbb{1}(x_i \geq v_i)\cdot x_i \leq B_1 \\ & \sum_i^\rho \mathbb{1}(x_i \geq v_i) \leq B_2 \end{aligned} \tag{1}$$

where f is a known concave L-Lipschtiz production function, which takes the average productivity of employees as input, B_1 is a wage budget, and B_2 is a space (number of employees) cap for the company. Intuitively, the firm wants to maximize its profit (assume price has been normalized to 1) by optimally selecting the number of employees, who present heterogeneous productivity U and reservation wage V, subject to budget and/or space considerations. I refer to this problem as Bandits with Knapsacks for the wage setting firm (`BwW`).

Why `BwW`? `BwW` allows us to incorporate transparently the notion of (constrained) sequential decision making in the presence of uncertainty. Classic results will frame this problem by relying on (posterior) expectations. However, Online Learning allows us to infer optimal decisions without prior knowledge nor prior assumptions on the DGP. This feature is particularly attractive within the realm of firm decisions, as firms

can be envisioned as agents which are born, learn, and exit the market conditional on not suitable learning (among others). In addition, BwW conveys very powerful economic intuitions to the Online Learning process, by (after suitable exploration) selecting the most "cost-effective" probability mixture of arms. Classic frameworks can be seen as particular instances of BwW, where $F_{U,V}$ belongs to the information history of the firm.

Mapping to BwCR. To map our problem into BwCR notation define vector $\boldsymbol{v} = (\mathbb{1}(x \geq v) \cdot u, \mathbb{1}(x \geq v) \cdot x, \mathbb{1}(x \geq v))$, where the stochastic behaviour of $\boldsymbol{v}$ given x is entirely characterized by the joint distribution $F_{U,V}$ in its domain. Now consider function $g : [0,1]^3 \to [0,1]$, defined via $\frac{1}{K}\sum \boldsymbol{v} \mapsto 1/2(1 + f(1/K \sum v_1) - 1/K \sum v_2)$.

Claim. g is concave in $\frac{1}{K}\sum \boldsymbol{v}$. *Proof.* By concavity of f it is immediate that $h_1 : \boldsymbol{v} \mapsto f(1/K \sum v_1) - 1/K \sum v_2$ is concave. Moreover, $g_2 : \boldsymbol{w} \mapsto 1/2 \cdot (1 + \boldsymbol{w})$ is a linear monotonically increasing transformation of w. The composition of concave monotonically increasing functions is concave. It follows that $g : v \mapsto h_2(h_1(v))$ is concave. □

Following, define the convex set S as the polytope defined via $\{1/K \sum v_1 \in [0,1], 1/K \sum v_2 \leq B, B_1/B_2 K \sum v_3 \leq B\}$, where $B = B_1/K$. The set S is not empty with probability 1 provided $\mathbf{0} \in S$. Finally, to ensure strict bounds with high probability (as opposed to BwCR regret minimization wrt to distance to S), we may define the set $S^\epsilon = \{y(1-\epsilon), \forall y \in S\}$, which satisfies equation (5) in [6]. This result holds because boundaries are additively linear in resources (despite f not being linear as in the BwK reduction of BwCR). We may conclude that solving the non-linear problem in equation (1), whose solution is given as a mixture of arms in simplex Δ_m, is equivalent (with high probability) to minimize the following two notions of regret,

$$\begin{aligned} \mathcal{R}_1 &= g(Vp^*) - g(1/K \sum_i^K \boldsymbol{v}_i) \\ \mathcal{R}_2 &= d(1/K \sum_i^K \boldsymbol{v}_i, S^\epsilon) \end{aligned} \tag{2}$$

Where $g(Vp^*) = \texttt{OPT}_{\text{LP}}$ is the solution to the fractional relaxation of the linear program induced by the g transformation of the maximization problem, and $d()$ is the L_∞ distance. I refer the reader to [2] and [6] for a more detailed explanation of the

mathematical tools behind these results.

Algorithm. I conclude this section by providing a near-optimal algorithm for `BwW`, which is implementable in polynomial time. [6] provide as well more efficient algorithms for `BwCR` at the expense of some regret. The intuition for the main algorithm is based on solving a linear program for the optimal mixing probabilities subject to picking matrices M whose elements M_{xj} are simultaneously in the [Lower Confidence Bound, Upper Confidence Bound] interval for all arms x and all resources j. We denote this set in period i as $\mathcal{H}_i$.

Algorithm 1 UCB Algorithm for `BwW`

Input K, δ, m, B, d
Set $\gamma = \log(\frac{mKd}{\delta}), \epsilon = \sqrt{\frac{\gamma m}{B}} + \log(K)\frac{\gamma m}{B}$
for $i = 1, 2, \ldots, K$
  select $p_i = \arg\max_{p \in \Delta_m} \max_{M' \in \mathcal{H}_i} g(M'p)$
    s.t. $\min_{M'' \in \mathcal{H}_i} d(M''p, S^{\epsilon}) \leq 0$
play $x_i \sim p_i$
Recover v and update $\mathcal{H}_{i+1}$
end for

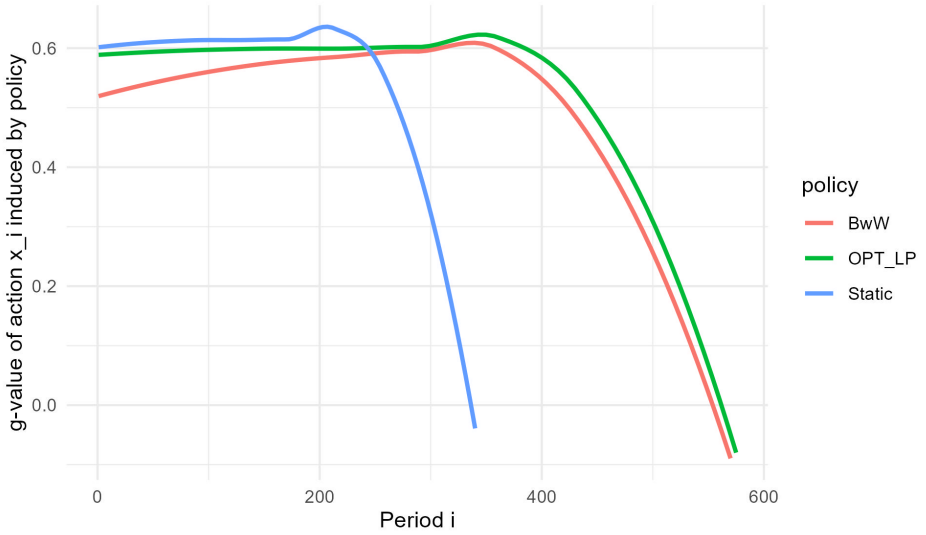
where $\mathcal{H}_i := \{M : M_{xj} \in \mathcal{H}_{i,xj} = [LCB_{i,xj}, UCB_{i,xj}]$. For each term in matrix M_{xj} define $\text{UCB}_{i,xj} = \min\{1, \hat{\nu}_{xj} + 2\text{rad}(\hat{\nu}_{xj}, k_{i,x} + 1)$, where $k_{i,x}$ is the number of times arm x was played before i, $\hat{\nu}_{xj}$ is the empirical average of ν_x and $\text{rad}(\nu, N) = \sqrt{\frac{\gamma\nu}{N}} + \frac{\gamma}{N}$. Intuition for the rad() operator is very similar to inequality bounds in Hoeffding-Azuma inequality derivation. LCB is characterized analogously.

4 Expected Contributions

4.1 Theoretical Contributions

Upper Bounds Characterization. Within the theory realm, I ambition to derive optimal bounds for the special `BwW` case. In terms of generality, `BwW` lies between `BwK` and `BwCR` for which optimal bounds have been characterized. Moreover, the `BwCR` reduction to `BwK` present the same regret bounds as taylored `BwK` bounds, thus I am positive towards the characterization of upper bounds in this particular case.

Figure 1: $g(x_i)$ given $U \sim \mathcal{U}[0,1]$, $v = \frac{1}{2}u$



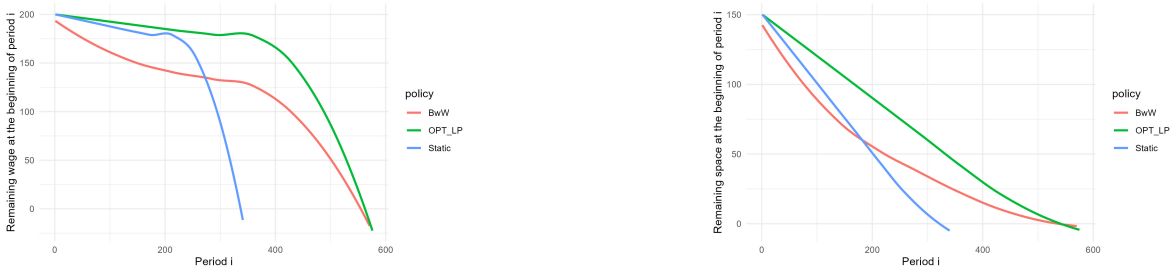
Parameters: $K = 1000$, $B_1 = 200, B_2 = 150, \delta = 0.05, \epsilon = 0.05, \gamma = 1.26$. Averages across 100 replications.

Simulation Exercise. Additionally, I would like to show the empirical potential of the theoretical results above through a simulation exercise. As an illustration, consider the special case where $f(v_1) = \sqrt{v_1}$, $U = \mathcal{U}[0,1]$, $V = 1/2U$.

As we can see in Figure (1), the g-value (or reward function value) induced by `BwW` converges to the value induced by $\texttt{OPT}_{\text{LP}}$. This is, a policy which samples x_i from the integer relaxation LP with $\Omega_X = (0, 1/4, 1/2, 3/4, 1))$. In this case, optimal solution is given by the probability vector $(0.6, 0.4, 0, 0, 0)$. On the other hand, the "greedy" cost unaware best arm which selects $x = 1/4$ with probability 1 obtains higher initial returns but it consumes its resources too fast leading to depletion and, consequently, early termination of the algorithm. These intuitions are confirmed in Figure (2).

Constrained MDPs. Recently, Efroni et al. [12] characterized a similar problem to `BwCR` within the realm of Markov Decision Processes (MDPs). This framework provides extra flexibility to the model, by accommodating for decision dependent states. In particular, this scenario could be of great interest in the context of the problem of the firm, as it enables us to define the wage budget of the firm as a function of previous profits. Consider,

Figure 2: Remaining wage budget/space given $U \sim \mathcal{U}[0,1]$, $v = \frac{1}{2}u$



(a) Average Remaining Wage

(b) Average Remaining Space

Parameters: $K = 1000$, $B_1 = 200, B_2 = 150, \delta = 0.05, \epsilon = 0.05, \gamma = 1.26$. Averages across 100 replications.

$$B_{1i} = B_1 - \sum_{t=1}^{i-1} \mathbb{1}(x_t \geq v_t) \cdot x_t + \phi(g(\sum_{t=1}^{i-1} \boldsymbol{v}_t)) \tag{3}$$

where ϕ is a function of the reward value up to period $i-1$ (i.e. a share of profits). The implications of such adoption are not obvious to me. On the one hand, MDPs compensate even more the adoption of "cost aware" policies as opposed to reward-maximizing actions (best static decision), provided sustainable firms are able to live longer. On the other hand, the greedy pursue of statics first best (conditional on some initial exploration) can be seen as cost effective policy because it will increase budget in subsequent periods. As a future line of research, I would like to leverage this extension to the model to gain further insights on (i) the behaviour of firms, and (ii) the trade-offs involved in optimal adaptive policies with (stochastic) decision-dependent environments.

4.2 Applied Contributions

Macro and (Micro) Modelling. In my opinion, the most relevant implication of `BwW` is its economic modelling potential. `BwW` provides a very realistic framework for understanding the optimal behaviour of price-setting firms in the presence of uncertainty. In this case, `BwW` deals with hiring decisions, but similar algorithms can be derived for alternative firm decisions like optimal price setting with constrained demand and unknown demand elasticities. In this section, we suggest two applications of our work: (i) structural policy analysis and (ii) embedded modelling.

Structural Policy Analysis. `BwW` generates a tractable scenario for understanding the impact of exogenous policy shocks. Even in our stylized model, one could easily analyze the production and employment implications of increases in minimum wage by either restricting the policy space Ω_X or by setting a floor on firms' (average) wage expenditure of the kind $1/K \sum_i \mathbb{1}(x_i \geq v_i) \cdot x_i \geq B_{\text{MW}}$. In addition, we may analyze the change in composition of the employed and unemployed in terms of productivity (and/or reservation wage) following a structural shock. This topic has gathered much attention in the literature in the last few years (see Mulligan [13]).

Similarly, one could look at the impact of mergers across firms Q_1, Q_2 on production and employment by adding up budget constraints (subject to some inference on the joined production technology), i.e. $B_{\text{Merge}} = B_{Q_1} + B_{Q_2}$. Finally, under alternative characterizations of `BwW`, one could further analyze the impact of price caps, industrial policy and others. All these forecasts are susceptible to be used as oracle estimates of firms' optimal behavior in comparison with empirical performance.

Learning and Information. Beyond the standard economic predictive power of our model, Online Learning introduces an additional variable which is frequently ignored in mainstream economic analysis: Learning Rates. There are different ways information, learning and uncertainty can be accounted for in this model. Explicitly, one may consider γ as the learning rate, and model it as a function of contextual indicators (under the premise that smaller γ increases the aggressiveness of learning, at the expense of optimal exploration). However, one may consider additional channels to alter the learning process of the firm like establishing a cap on the number of periods firm Q may use to bound $\mathscr{H}_i$. This is, substitute $k_{i,x}$ and $\hat{v}_{xj}$ by the constrained analogs $k^T_{i,x} = \sum_{t=i-T}^{i} \mathbb{1}(x_i = x)$ and $\hat{v}^T_{xj} = 1/k^T_{i,x} \sum_{t=i-T}^{i} \nu_{xj}$. In this scenario, firms may only rely on a finite number of periods T to forecast the reward/cost ratio of playing arm x, similar to "limited information capacity constrain χ" in [3] or imperfect information recall in standard microeconomics literature.

Eventually, one could analyze the welfare implications of information loss in this model, either coming from limiting information capacity or from firms bankruptcy, which would lead to complete "bounds reset" as in Coen [14]. Despite its simplicity, this model of information is much more flexible than the one introduced in [3] where no firm heterogeneity is allowed, and information parameters do not respond to environment change, nor to firm decisions. In other words, productivity realizations (extra

information) have no impact on the future decisions of the company.

Embedded Modelling. Finally, it is important to highlight that the analytical tractability of this model, may allow the applied (macro)economist to embed `BwW` within broader economic models. Remember, that `BwW` has been shown to converge to $\texttt{OPT}_{\text{LP}}$ which is the solution of a very simple (and transparent) LP. As a result, $\texttt{OPT}_{\text{LP}}$ comparative statics can be incorporated analytically to bigger (general) equilibrium models with applications in IO, Macro-finance, Macro-Labor and Public Economics.

5 Conclusion

This research proposal ambitions to build a new theoretical environment for the analysis of the problem of the firm in the presence of uncertainty by relying on `BwW` like algorithms. To do so, future lines of research should confirm the regret-guarantees of `BwW` and `BwW`-MPDs. Once performance results have been established, there is virtually no limit to the theory implications which can be extracted from this sort of models. From direct structural policy analysis (minimum wage, mergers, industrial policy, etc.) to embedded modelling in larger general equilibrium frameworks which rely on $\texttt{OPT}_{\text{LP}}$ statics.

References

[1] Maximilian Kasy. Foundations of machine learning, 2023.

[2] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

[3] Sushant Acharya and Shu Lin Wee. Rational inattention in hiring decisions. *American Economic Journal: Macroeconomics*, 12(1):1–40, 2020.

[4] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[5] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[6] Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006, 2014.

[7] Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 594–605. IEEE, 2003.

[8] Nicolò Cesa-Bianchi, Tommaso R Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. A regret analysis of bilateral trade. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 289–309, 2021.

[9] Nicolo Cesa-Bianchi, Roberto Colomboni, and Maximilian Kasy. Adaptive maximization of social welfare. 2022.

[10] Carlos Gonzalez Perez. Adverse selection in adaptive settings. 2023.

[11] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

[12] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

[13] Casey B Mulligan. Rising labor productivity during the 2008-9 recession. Technical report, National Bureau of Economic Research, 2011.

[14] Patrick Coen. Information loss over the business cycle. 2021.