Pack Light on the Move: Exploitation and Exploration in a Dynamic Environment

Marco LiCalzi and Davide Marchiori

Abstract This paper revisits a recent study by Posen and Levinthal (Manag Sci 58:587–601, 2012) on the exploration/exploitation tradeoff for a multi-armed bandit problem, where the reward probabilities undergo random shocks. We show that their analysis suffers two shortcomings: it assumes that learning is based on stale evidence, and it overlooks the steady state. We let the learning rule endogenously discard stale evidence, and we perform the long run analyses. The comparative study demonstrates that some of their conclusions must be qualified.

1 Introduction

In many situations, an agent must simultaneously make decisions to maximize its rewards while learning the process that generates these rewards. This leads to a tradeoff between exploration versus exploitation. Exploratory actions gather information and attempt to discover profitable actions. Exploitative actions aim to maximize the current reward based on the present state of knowledge. When the agent diverts resources towards exploration, he sacrifices the current reward in exchange for the hope of higher future rewards.

The dilemma between exploration and exploitation is well-known in machine learning, where the agent is an algorithm; see f.i. Cesa-Bianchi and Lugosi [2]. Within this field, the simplest and most frequent example is the multi-armed bandit problem, extensively studied in statistics as well (~~Barry~~ and Fristedt 1985). However, in the literature on organizational studies, the exploration/exploitation trade-off has come to be associated mostly with a seminal contribution by March [5], that introduced a peculiar model of his own.

AQ1

M. LiCalzi (✉) · D. Marchiori
Department of Management, Università Ca' Foscari Venezia, Venice, Italy
e-mail: licalzi@unive.it; davide.marchiori@unive.it

S. Leitner and F. Wall (eds.), *Artificial Economics and Self Organization*, Lecture Notes in Economics and Mathematical Systems 669, DOI 10.1007/978-3-319-00912-4_16,

The popularity of March [5], as witnessed by more than 10,000 citations on Google Scholar, has firmly placed the exploration/exploitation trade-off among the methodological toolbox of organizational studies, but the peculiarity of his modeling choice has shifted attention away from the multi-armed bandit problem as a modeling tool. This shortcoming was recently addressed by Posen and Levinthal [6], that explicitly discuss some similarities between the bandit problem and the March model.

Their paper inquires about the implications of the exploration/exploitation trade-off for organizational learning when the environment changes dynamically or, more precisely, when the process generating the rewards is not stationary. Using the bandit problem as a workhorse, they challenge the conventional view that an increasingly turbulent (i.e., non-stationary) environment should necessarily elicit more exploration.

We believe that Posen and Levinthal [6] make two very important contributions. First, they raise fundamental questions (as well as providing convincing answers) about the impact of turbulence in an environment for organizational learning. Second, they implicitly make a strong methodological case for a revival of the bandit problem as a modeling tool.

On the other hand, we argue that two (apparently minor) of their modeling choices are potentially misleading. The first one is the length of the horizon over which the study is carried out: this is too short to provide information about the steady state. The second one is that learning is based on the whole past evidence (including what turbulence has made obsolete): this makes it too slow to detect shocks, and hence ineffective.

This paper sets out to discuss and correct these flaws, revisiting their analysis over the short and the long run. We propose two (nested) learning models that endogenously recognize and shed away stale evidence, and compare their performance with the original model by Posen and Levinthal [6]. We check several of their conclusions, and show how a few of these need to be qualified. Paraphrasing the title of their paper, our major result demonstrates the importance of packing light (evidence) when chasing a moving target. Shedding away obsolete information is crucial to attain a superior performance as well as making learning resilient to shocks.

2 The Model

We summarize the model proposed in Posen and Levinthal [6]; then, we present the crucial tweaks we advocate. At each period t, an organization must choose among $N = 10$ alternatives. Each alternative $i = 1, \ldots, 10$ has two possible outcomes: $+1$ (success) or –1 (failure). These are generated as a (Bernoullian) random reward R_t^i in $\{-1, 1\}$, with probability p_t^i of success. Thus, the state of the environment in period t is summarized by the vector $P_t = [p_t^1, \ldots, p_t^{10}]$.

In the standard bandit problem, the environment is stationary and $P_t = P$ for all t. Posen and Levinthal [6]—from now, PL for brevity—relax this assumption and introduce environmental turbulence as follows. Each alternative i is given an initial probability p_0^i randomly drawn from a Beta distribution with $\alpha = \beta = 2$. This has a unimodal and symmetric density, with expected value $1/2$ and variance $1/20$. The turbulence in the environment follows from a probabilistic shock that may occur in each period with probability η. When $\eta = 0$, the environment is stationary; increasing η raises the level of turbulence. For $\eta > 0$, PL assume $\eta = 0.005 \times 2^k$ with k being an integer between 0 and 6. When a shock occurs, each of the payoff probabilities is independently reset with probability $1/2$ by an independent draw from the same Beta distribution.

At each period t, the organization holds a propensity q_t^i for each alternative that is formally similar (and proportional to) its subjective probability assessment that the i-th alternative yields success, and thus leads to a reward of 1. At time t, its propensities over the 10 available alternatives are summarized by the vector $Q_t = [q_t^1, \ldots, q_t^{10}]$. Propensities are updated using a simple rule, akin to similar treatments in reinforcement learning; see Duffy [3].

Let n_t^i be the number of successes and the total number of plays for the i-th alternative up to (and including) period t. PL define the propensities recursively by

$$q_{t+1}^i = \left(\frac{n_t^i}{n_t^i + 1}\right) q_t^i + \left(\frac{1}{n_t^i + 1}\right) \frac{R_t^i + 1}{2} \tag{1}$$

with the initial condition $q_0^i = 1/2$ for each i. As n_t^i increases, the weight associated to the most recent outcome declines.

This paper follows PL's assumption about propensities to facilitate comparison. However, we notice that Eq. (1), while certainly reasonable, is a reduced form that omits the specification of the relationship between q_t^i and the number of successes and failures experienced with the i-th alternative. A more explicit formulation might have been the following. Let s_t^i and n_t^i be respectively the number of successes and the total number of plays for the i-th alternative up to (and including) period t. Let us define the propensities by $q_{t+1}^i = (1 + s_t^i)/(2 + n_t^i)$, with the initial condition $s_0^i = n_0^i = 0$ to ensure $q_1^i = 1/2$ for each i. Then the updating rule for propensities would read

$$q_{t+1}^i = \left(\frac{n_t^i + 1}{n_t^i + 2}\right) q_t^i + \left(\frac{1}{n_t^i + 2}\right) \frac{R_t^i + 1}{2}$$

The choice behavior in each period depends on the distribution of propensities and on the intensity of the search strategy. More precisely, PL assume a version of the *softmax* algorithm; see f.i. Sutton and Barto [7]. In period t, the organization picks alternative i with probability

$$m_t^i = \frac{\exp\left(10 q_t^i/\tau\right)}{\sum_{j=1}^{10} \exp\left(10 q_t^j/\tau\right)}$$

where the parameter τ in $\{0.02, 0.25, 0.50, 0.75, 1\}$ directly relates to the intensity of the exploration motive. For $\tau = 0.02$, the organization picks with very high probability the alternative with the highest current propensity; this is an exploitative action. As τ increases, the choice probability shifts towards other alternatives and exploratory actions become more likely.

We argue that the evolution of propensities in (1) is not plausible for dynamic environments, because it is implicitly based on a cumulative accrual of evidence. When $\eta > 0$ and a shock displaces alternative i, the past outcomes for i become uninformative about the new value of p_t^i. However, Rule (1) keeps cumulating such stale evidence when computing the propensity for i. Moreover, since the weight for a new piece of evidence decreases as $1/(n_t^i + 1)$, the marginal impact of more recent information is decreasing; that is, the cumulative effect of past history tends to overwhelm fresh evidence. For instance, suppose that alternative i has had a long history of successes; if a negative shock makes p_i drop, the firm would take in a substantial streak of failures before its propensity q_t^i is brought back in line with the new value of p_i.

This bias may be partially corrected by a higher τ, because increasing exploration speeds up the alignment process between the propensity vector Q_t and the actual probabilities in P_t. However, this is inefficient because it takes ever longer streaks of experiments to overturn the cumulated past evidence. One of our goals is to demonstrate the advantages for an organization to shed away stale evidence in a turbulent environment.

Formally, the root of the problem in PL's setup is that the *marginal impact* of the last observation in Eq. (1) declines as $1/(n_t^i + 1)$. Among many different ways to correct this problem, an optimal choice should depend on η. However, the exact value of this parameter is unlikely to be known to the organization. Therefore, we opt for a simple rule that is robust to such lack of quantitative information about η. Its robustness comes from a built-in mechanism that modulates the intensity with which past evidence is shed away as a function of the degree η of turbulence in the environment.

We advocate two modifications to PL's learning model. Both refresh evidence endogenously. The first one deals with the possibility that the current choice may have been made unfavorable by a negative shock. When alternative i is chosen and $n_t^i \geq \bar{n}$, we split its past history into two segments of equal length: the first and the second half. (When n_t^i is odd, we include the median event in both histories.) We aim to drop from consideration the initial segment when a shock might have occurred and past evidence turned stale. To do so, we compute the average performances $\bar{R}_i^1$ and $\bar{R}_i^2$ over the first and the second segment, respectively. Then, with probability equal to $\left|\bar{R}_i^1 - \bar{R}_i^2\right|/2$, a *refresh* takes place: we delete the initial segment and recompute q_t^i accordingly. Since we only act when $n_t^i \geq \bar{n}$, the length of the past history after a deletion never goes below $\bar{n}/2$. For $\bar{n} \uparrow \infty$, we recover the model in PL. For demonstration purposes, in this paper we set $\bar{n} = 30$.

The second modification recognizes that alternatives that have not been tried in a long time may have been reset by a shock. In particular, whenever a refresh takes place, we reset the propensity for each alternative that has never been explored since the previous refresh to $1/2$.

In short, the first modification reduces the risk of staying with an alternative that has turned into a "false positive"; the second recovers forgone alternatives that might have changed into "false negatives". We refer to the model dealing only with false positives as M1, and to the full model as M2. We were surprised to discover how much M2 improves over M1 in a dynamic environment. Each of the values reported below is an average based on 5,000 simulations with different seeds.

3 The Stationary Environment

Our benchmark is the stationary environment, when $\eta = 0$. PL consider four indicators. *Performance* in PL is the cumulated value of rewards; for ease of comparison, we report the average performance $(\sum_{\tau=1}^{t} R^*_\tau)/t$ per period, where R^*_t is the reward associated with the choice made at period t. *Knowledge* embodies the ability of Q_t to track P_t and is measured by $1-\sum_i (p^i_t - q^i_t)^2$. The *Opinion* indicator $\sum_i (q^i_t - \bar{q}_t)^2$ is the sample variance of propensities; the higher it is, the more diverse the propensities and therefore the probabilities of choosing each alternative. Finally, the *Exploration* indicator computes the probability that the choice at time t is different from the choice at time $t-1$.

PL report the values of these four indicators at $t = 500$. As it turns out, this horizon is too short to take into account the onset of the steady state and thus PL's analysis is limited to the short run. (They do not mention a rationale for this choice.) We replicate their short-run analysis at $t = 500$ and extend it to the long-run at $t = 5{,}000$. The short- and long-run values for PL are shown on the left-hand side of Table 1, respectively on the first and second line of each box. With a few exceptions (notably, when $\tau = 0.02$), differences in values between short- and long-run hover around 10 %. The working paper provides a visual representation of the data, that we omit for brevity.

The left-hand side of Table 1 confirms and extends the short-run results in PL's Sect. 3.1. Exploratory behavior is increasing in τ, and the optimal level of search intensity τ is around 0.5. Except for $\tau = 0.02$, the long-run performance is about 10 % higher than PL's short-run estimate: since the search intensity never abates, this increase is not due to "cashing in" from reducing the searching efforts but instead stems from the long-run stationarity.

Knowledge and Opinion are similarly higher, as an immediate consequence of the larger cumulated number of experiences. The increase in Exploration is due to a little known property: in the short run, the softmax algorithm tends to ignore an alternative that has failed on the first few attempts, regardless of its actual probability of success. Any of such false negatives contributes towards making the algorithm focus on very few alternatives in its early stages. However, given enough time, the algorithm eventually returns to such alternatives and, if it finds them valuable, puts them back in the explorable basket. To gauge the extent of this effect, Table 2

Table 1 Performance, knowledge, opinions, and choices in the stationary environment

		PL					M1					M2				
	τ	0.02	0.25	0.50	0.75	1.0	0.02	0.25	0.50	0.75	1.0	0.02	0.25	0.50	0.75	1.0
Performance	$t = 500$	0.48	0.52	0.56	0.54	0.50	0.53	0.55	0.56	0.53	0.50	0.54	0.55	0.55	0.51	0.47
	$t = 5{,}000$	0.49	0.58	0.61	0.59	0.55	0.59	0.61	0.61	0.59	0.54	0.61	0.61	0.60	0.55	0.50
Knowledge	$t = 500$	0.55	0.56	0.59	0.65	0.72	0.54	0.56	0.59	0.65	0.71	0.59	0.61	0.64	0.70	0.76
	$t = 5{,}000$	0.56	0.57	0.64	0.77	0.89	0.52	0.54	0.61	0.73	0.83	0.61	0.63	0.66	0.73	0.79
Opinion	$t = 500$	0.16	0.21	0.32	0.44	0.50	0.21	0.23	0.31	0.40	0.46	0.11	0.12	0.16	0.24	0.31
	$t = 5{,}000$	0.17	0.23	0.42	0.53	0.55	0.23	0.27	0.37	0.47	0.50	0.11	0.12	0.16	0.23	0.29
Exploration	$t = 500$	0.00	0.02	0.18	0.39	0.54	0.00	0.02	0.14	0.33	0.48	0.01	0.05	0.17	0.38	0.54
	$t = 5{,}000$	0.00	0.04	0.26	0.46	0.60	0.00	0.01	0.13	0.33	0.51	0.02	0.04	0.15	0.37	0.54

Table 2 Percentage of (almost) unexplored alternatives

	PL					M1					M2				
τ	0.02	0.25	0.50	0.75	1.0	0.02	0.25	0.50	0.75	1.0	0.02	0.25	0.50	0.75	1.0
$t = 500$	0.89	0.86	0.79	0.69	0.60	0.83	0.80	0.74	0.65	0.57	0.81	0.78	0.71	0.62	0.51
$t = 5{,}000$	0.89	0.83	0.67	0.48	0.27	0.75	0.72	0.59	0.41	0.23	0.63	0.54	0.31	0.05	0.00

provides estimates for the percentage of alternatives that are explored less than $\bar{n}/2 = 15$ times in the whole period.

The rest of Table 1 provides data for our models M1 and M2, where old evidence may be discarded. One would expect PL to perform better in a stationary environment, because P_t is constant over time and thus evidence never gets stale. However, by forgetting stale evidence, both M1 and M2 refresh propensities and have an endogenous bias towards more search. Such bias overcomes the "false negatives" trap of the softmax algorithm and makes their performance competitive with (and often marginally better than) PL. In particular, both M1 and M2 achieve their superior performance with a lower level for the Exploration indicator: compared to PL, they are less likely to switch the current choice.

Instead of PL's five-points grid, we computed the optimal search intensity over a finer 100-points grid and found the following optimal values: $\tau = 0.56$ (0.48) for PL when $t = 500$ ($t = 5{,}000$, respectively); $\tau = 0.45$ (0.36) for M1; and $\tau = 0.40$ (0.24) for M2. The sharp reduction in the optimal search intensity from PL to M1 to M2 stems from their search bias. Within each model, the optimal τ decreases when going from the short- to the long-run because steady state learning is more effective.

We summarise our comparative evaluation of the three learning rules. The search intensity τ is not easy to tune in practice, but our models are more robust: they deliver a tighter range across different values of τ. This comes with less switches in choice and tighter opinions (for the same level of τ), and an overall comparable performance. Thus, although the three learning models are roughly comparable in a static environment, ours are more robust.

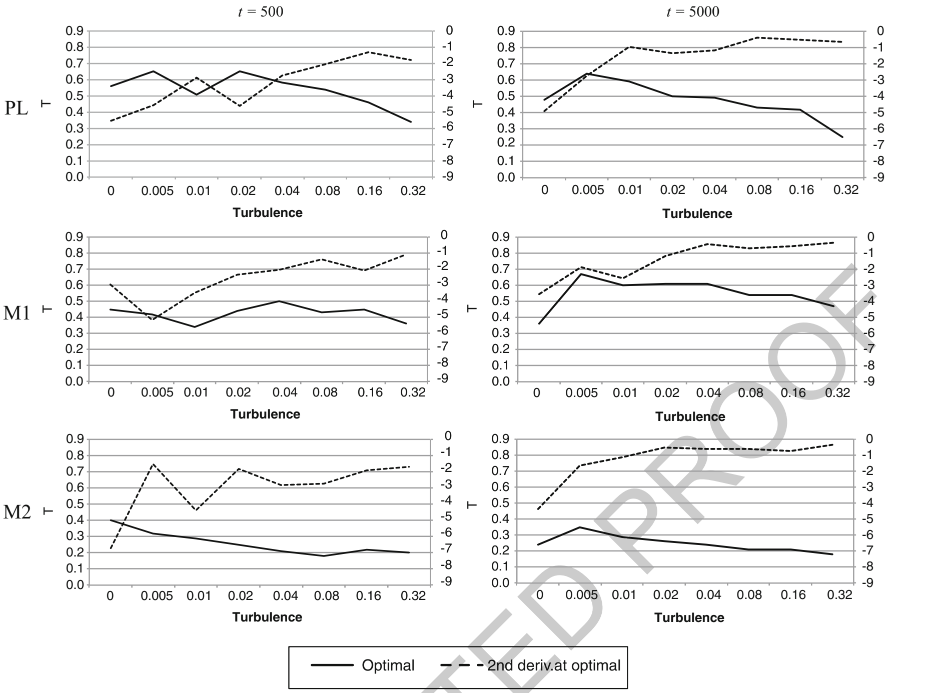

Fig. 1 Optimal exploration strategy across turbulence levels

4 The Dynamic Environment

In a dynamic environment, turbulence is represented by the probability $\eta > 0$ that in each period a shock resets the actual probabilities in P_t. Following PL, we consider $\eta = 0.005 \times 2^k$ for $k = 0, 1, \ldots, 6$. The main result in PL is that the optimal level of search intensity has an inverse U-shaped form that is right skewed. We found that this statement must be qualified as follows.

PL derive the curve by "fitting a third order polynomial to the results" (p. 593), but no details are provided and the available points are just five. Therefore, we opted for a brute force approach and did and extensive search over [0.02, 2.00] using a grid with mesh 0.01. Figure 1 illustrates the results, reporting data for $t = 500$ and $t = 5{,}000$ on the left- and right-hand side, respectively.

Let us begin with the long run ($t = 5{,}000$), as represented on the right-hand side of Fig. 1. For $\eta > 0$, the optimal search intensity is actually decreasing in the turbulence level. The inverse U-shaped form is a visual artefact created by the inclusion of the first datapoint ($\eta = 0$) corresponding to zero turbulence. In a dynamic environment, an organisation with a sufficiently long horizon has an optimal search intensity that is decreasing in the level of turbulence.

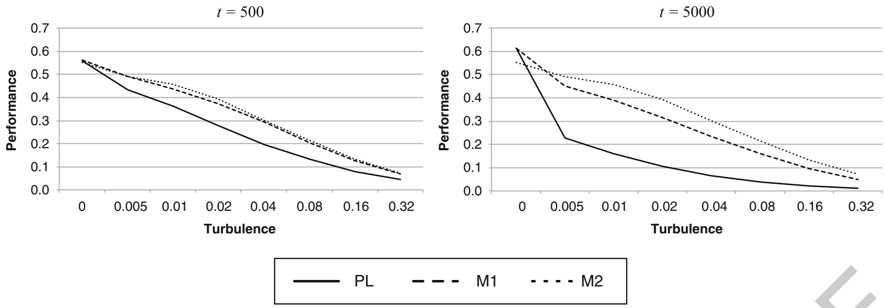

Fig. 2 Performance across turbulence levels ($\tau = 0.5$)

Consider now the short run ($t = 500$). We find a dip at $\tau = 0.01$ for PL, but this might be due to noise. On the other hand, the optimal search intensity for M1 stays pretty flat, and for M2 is decreasing overall. We argue below that M2 is superior to PL; thus, for an organisation with an appropriate learning model and a short run horizon, the optimal search intensity is *decreasing* in the level of turbulence. We conclude that, under an appropriate model specification, turbulence has a systematic negative effect on the optimal search intensity: the inverse U-shaped form claimed by PL is not an accurate depiction of this result.

PL discuss how the value of the second derivative of the performance at the optimal τ can be used as a proxy for the intensity of the tradeoff between exploration and exploitation. While generally negative, the closer to zero, the flatter the curve $f(\tau)$; and hence the less important pinning down the right τ is. Lack of details in PL prevented us from replicating their work, so we decided to compute our approximation to the second derivative in two steps. First, for each point τ on our grid, we computed the second-order central difference $D(\tau) = [f(\tau + 0.01) - 2f(\tau) + f(\tau - 0.01)]/h^2$. Second, we performed a simple smoothing by replacing $D(\tau)$ with the weighted mean

$$\overline{D}(\tau) = \frac{D(\tau - 0.02) + 2D(\tau - 0.01) + 4D(\tau) + 2D(\tau + 0.01) + D(\tau + 0.02)}{10}$$

The graph for the (approximated) second derivative is superimposed as a dashed line on the panels in Fig. 1. After cautioning the reader not to put much weight on the first datapoint ($\eta = 0$), we find that in most cases the second derivative is increasing in the turbulence level, confirming PL's claim that pinpointing the optimal τ matters less to performance when turbulence is higher.

Coming to performance, we were puzzled by the contrast between PL's extensive discussion of it for the stationary environment ($\eta = 0$) and the complete lack of data for $\eta > 0$. A primary element in evaluating the plausibility of the learning rule under turbulence should be its performance. Figure 2 provide a visual representation of the data for $\tau = 0.5$. (The working paper provides tables with the numerical values for this figure as well as for the following ones.) Here, as in PL, we leave the search intensity τ constant. Alternatively, one might consider the optimal performance using the best search intensity for each η. We report the outcome of this exercise in the working paper: we found qualitatively similar results that are even more favourable to the claim we advance below. Hence, fixing $\tau = 0.5$ avoids biasing the graphs against PL.

Except when $\eta = 0$, the performance for M1 and M2 is consistently and significantly better than for PL over both horizons. In the long run, the degradation in performance for PL is much stronger and, if one ignores the data point for $\eta = 0$, fairly disastrous: PL scores about 20 % when turbulence is minimal ($\eta = 0.005$) and drops to virtually 0 %—equivalent to random choice—under intense turbulence ($\eta = 0.32$). It is hard to claim that PL's rule captures effective learning in a turbulent environment.

To the contrary, both of our models deal with intense turbulence reasonably well. The decline in performance when η increases is not as abrupt as PL and, even under intense turbulence, they manage to rake up a performance that is small but significantly higher than the 0 % associated with random choice. Moreover, by explicitly dealing with the foregone alternatives that shocks might have turned into false negatives, M2 performs significantly better than M1 in the long run. Therefore, when a shock is deemed to have occurred, one should not only drop evidence about the (potentially) false positive as in M1, but also about the (potentially) false negatives as in M2. This is worth pointing out because many studies about the representativeness heuristic suggest that people are less prone to review evidence about false negatives than about false positives. The main conclusion is that shedding stale evidence makes the search process in a dynamic environment perform better as well as exhibit resilience to turbulence.

PL convincingly argue that turbulence erodes performance by two effects: it alters the future value of existing knowledge and reduces the payoff from efforts to generate new knowledge. To disentangle these two effects, they use a differences-in-differences analysis assuming a search intensity $\tau = 0.5$. (See PL for details.) Their approach separately estimates the accretion of new knowledge and the erosion of existing knowledge for different levels of turbulence. These two effects jointly determine the net change in knowledge. We replicated their short-run analysis ($t = 500$), and extended it to the long-run ($t = 5{,}000$) using propensities at $t = 4{,}000$ and $t = 5{,}000$. As before, the choice $\tau = 0.5$ fits PL better than our models; but, again, we redrew the graphs using the optimal value of τ for each turbulence level, and found no qualitative differences. Using PL's setting for ease of comparability, the results are shown in Fig. 3.

We found again that the details in some of PL's statements need amendments. Looking at the short-run, all models exhibit the same behaviour; namely, both accretion and erosion have an inverse U-shaped form and the net effect on knowledge is overall positive across all levels of turbulence. The size of the two effects, however, is quite different: in PL none of the two effects brings about a change greater than 8 % in absolute, while in M1 and M2 this can go as high as 14 %. The vertical dilation in the graphs as we move downwards from PL to M2

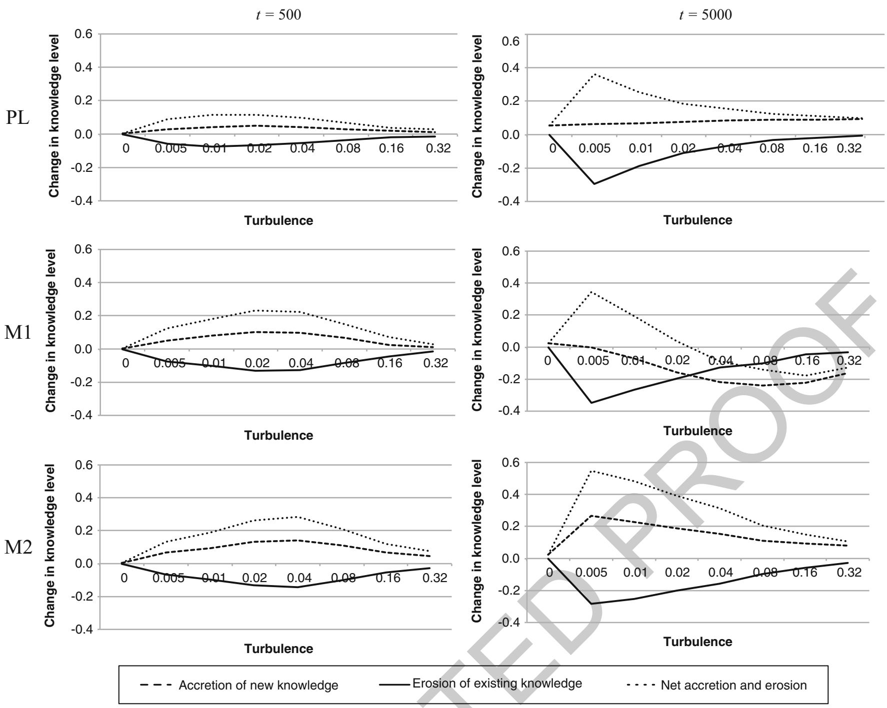

Fig. 3 Knowledge accretion and erosion across turbulence levels ($\tau = 0.5$)

on either side of Fig. 3 is apparent. Shedding evidence magnifies both the positive accretion effect and the negative erosion effect.

Over the long run, the two effects change shape for all models, after discarding the insignificant datapoint at $\eta = 0$. Conforming to intuition, one would expect knowledge accretion and knowledge erosion to be respectively decreasing and increasing in turbulence. This occurs only for M2, while PL and M1 match the pattern for knowledge accretion only partially. PL exhibits knowledge accretion that is increasing in turbulence. M1's knowledge accretion is decreasing over most of the range, but eventually starts climbing up generating a U-shape. Given that M2 is superior in what regards both performance and the net effect on knowledge, it is reassuring to see that the pattern of its knowledge accretion effect matches intuition.

AQ3 Our last batch of work replicates and extends PL's Fig. 6 reporting the accuracy of knowledge, the strength of opinions, and the probability of exploration at $\tau = 0.5$ in Fig. 4. Over the short run, the three models exhibits the same qualitative shapes for the three indicators and these are consistent with intuition. With respect to turbulence levels, knowledge is decreasing, strength of opinions is decreasing, and probability of switching choice is increasing.

Moving to the long run reveals a few hidden patterns. First, the knowledge indicator goes almost flat for PL, suggesting that the knowledge generated within

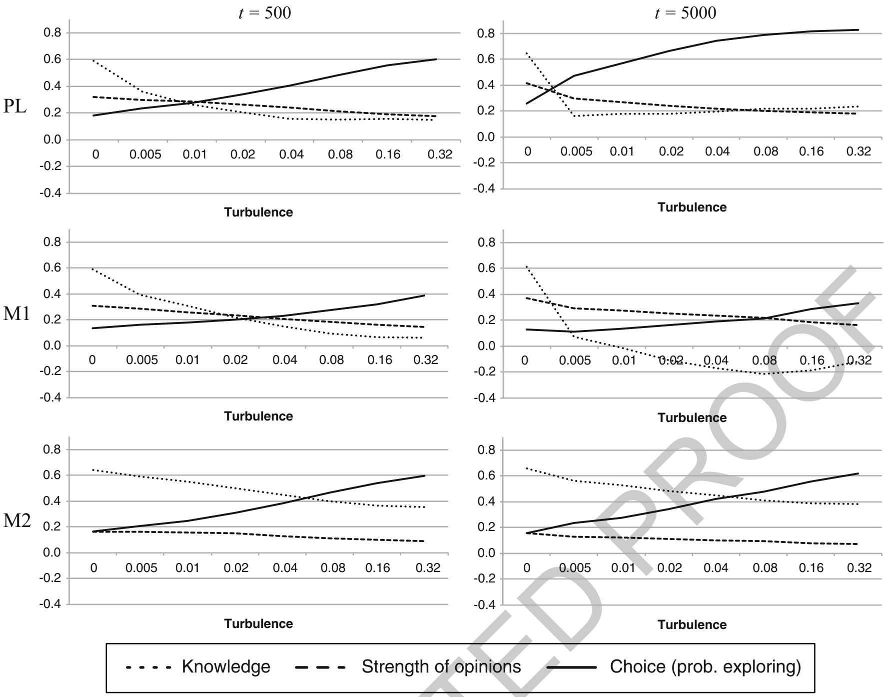

Fig. 4 Knowledge, opinions, and choices across turbulence levels ($\tau = 0.5$)

this model in the long run is unaffected by the level of turbulence. (Differently put, once we enter the steady state, the level of turbulence has a negligible effect on knowledge.) With little variation in opinions, PL ends up with very similar propensities across all alternatives and, accordingly, the probability of switching becomes much bigger: in practice, PL ends up being close to (randomly) wandering across alternatives. M1 generates even less knowledge in the long run, but its strength of opinions is bigger: in other words, propensities are more polarised (which helps focusing choice and reduces the probability of switching) but on the wrong alternatives (which adversely affects knowledge).

Finally, M2 is very effective in the long run: its knowledge indicator is small and decreasing with respect to turbulence, because in a dynamic environment it is ineffective to strive for high levels of knowledge. Keeping knowledge small ("pack light") allows opinions to change swiftly and track shocks accurately; hence, their strength resists homogenisation and stays around 0.4 even when turbulence is intense. Finally, the probability of switching choice increases less than PL and more than M1: in other words, the action bias of M2 is intermediate. This is necessary to balance two opposing effects: the risk of wandering choices (as in PL) against the possibility that exploration cannot keep with the flow of incoming shocks. Notably enough, M2 achieves this balance endogenously: our models have not been calibrated for maximum performance.

5 Conclusions

We revisit a recent study by Posen and Levinthal [6] about learning under turbulence. We claim that their analysis overlooks the long run and posits a learning model that puts too much weight on stale evidence. This leads us to suggest two learning models that incorporate an endogenous mechanism to spot and shed away obsolete evidence. M1 deals only with the possibility that some shock may have made the current choice a false positive, while M2 adds a concern for foregone alternatives that may have become false negatives. PL is nested into M1, and M1 is nested into M2.

The comparative analysis shows that M2 offers a significantly superior performance, making PL an implausible candidate for an effective learning model. Even under intense turbulence, its ability to "pack light evidence" makes it properly responsive to shocks, and allows it to deliver a performance that is both robust and resilient. We believe that clarifying the importance of giving up on obsolete evidence is the major contribution of this paper.

Finally, we carry out a comparative analysis for several claims in Posen and Levinthal [6], both over the short and run long run and across the three models. While their main insights survive, we find and point out which qualifications are needed for their validity. In particular, some of the (somewhat unintuitive) non-monotone relationships they discover using PL in the short run disappear when the analysis is carried out in the long run using M2.

References

1. Berry D, Fristedt B (1985) Bandit problems. Chapman and Hall, London
2. Cesa-Bianchi N, Lugosi G (2006) Prediction, learning, and games. Cambridge University Press, New York
3. Duffy J (2006) Agent-based models and human subject experiments. In: Tesfatsion L, Judd KL (eds) Handbook of computational economics, vol 2. North-Holland, Amsterdam/New York, pp 949–1011
4. LiCalzi M, Marchiori D (2013) Pack light on the move: exploitation and exploration in a dynamic environment. Working Paper 4/2013, Department of Management, Università Ca' Foscari Venezia,
5. March JG (1991) Exploration and exploitation in organizational learning. Organ Sci 1:71–87
6. Posen HE, Levinthal DA (2012) Chasing a moving target: exploitation and exploration in dynamic environments. Manag Sci 58:587–601
7. Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. The MIT University Press, Cambridge

AUTHOR QUERIES

AQ1. Please provide the details of “Barry and Fristedt (1985)” in reference list.
AQ2. Please check “Sect. 3.1” in the sentence starting “The left-hand side of Table 1...”.
AQ3. Please check “Fig. 6” in the sentence starting “Our last batch of work replicates...”.
AQ4. Please cite Refs. [1, 4] in text.