A Technical Note on Non-Stationary Parametric Bandits: Existing Mistakes and Preliminary Solutions

Louis Faury, Yoan Russac, Marc Abeille, Clément Calauzènes

A Technical Note on Non-Stationary Parametric Bandits: Existing Mistakes and Preliminary Solutions

Louis Faury [*] L.FAURY@CRITEO.COM
LTCI TélécomParis, Criteo AI Lab
Yoan Russac [*] YOAN.RUSSAC@ENS.FR
ENS Paris, Université PSL, CNRS, Inria
Marc Abeille M.ABEILLE@CRITEO.COM
Criteo AI Lab
Clément Calauzènes C.CALAUZENES@CRITEO.COM
Criteo AI Lab

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

In this note[1] we identify several mistakes appearing in the existing literature on non-stationary parametric bandits. More precisely, we study Generalized Linear Bandits (GLBs) in drifting environments, where the level of non-stationarity is characterized by a general metric known as the variation-budget. Existing methods to solve such problems typically involve forgetting mechanisms, which allow for a fine balance between the learning and tracking requirements of the problem. We uncover two significant mistakes in their theoretical analysis. The first arises when bounding the tracking error suffered by forgetting mechanisms. The second emerges when considering non-linear reward models, which requires extra care to balance the learning and tracking guarantees. We introduce a geometrical assumption on the arm set, sufficient to overcome the aforementioned technical gaps and recover minimax-optimality. We also share preliminary attempts at fixing those gaps under general configurations. Unfortunately, our solution yields degraded rates (w.r.t to the horizon), which raises new open questions regarding the optimality of forgetting mechanisms in non-stationary parametric bandits.

Keywords: Stochastic Bandits, Generalized Linear Model, Non-Stationarity.

1. Motivation and Setting

1.1. Motivation

Linear Bandits and non-stationarity. The Linear Bandit (LB) framework has proven to be an important paradigm for sequential decision making under uncertainty. It notably extends the Multi-Arm Bandit (MAB) framework to address the exploration-exploitation dilemma when the arm-set is large (potentially infinite) or changing over time. While the LB has now been extensively studied (Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011; Abeille and Lazaric, 2017) in its original formulation, a recent strand of research studies its adaptation to non-stationary environments. Notable are the contributions of Cheung et al. (2019b); Russac et al. (2019); Zhao et al. (2020) which prove that under appropriate algorithmic changes, existing LB concepts can be leveraged to handle a drift of the reward model. Aside their theoretical interests, these results further anchor the spectrum of potential applications of the LB framework to real-world problems, where non-stationarity is commonplace.

[*] Equal contribution.

1. This technical note was not peer-reviewed by ALT's Program Committee. An extended version can be found in the companion arXiv paper (Faury et al., 2021)

Extensions to Generalized Linear Bandits. Perhaps the main limitation of LB resides in its inability to model specific (e.g binary, discrete) rewards. One axis of research to operate beyond linearity was initiated with the introduction of Generalized Linear Bandit (GLBs) by Filippi et al. (2010). This framework allows to handle rewards which (in expectation) can be expressed as a generalized linear model. Notable members of this family are the logistic and Poisson models. Given the remarkable importance and widespread use of such models in practice, ensuring their resilience to non-stationarity stands as a crucial milestone in the parametric bandit literature.

1.2. Setting

We consider the stochastic contextual bandit setting under parameter-drift. The environment starts by picking a sequence of parameters $\{\theta_\star^t\}_{t=1}^\infty$. A repeated game then begins between the environment and an learning agent. At each round t, the environment presents the agent with a set of actions $\mathcal{X}_t$ (potentially contextual, large or even infinite). The agent selects an action $x_t \in \mathcal{X}_t$ and receives a (stochastic) reward r_{t+1}. In this note, we work under the fundamental assumption that there exists a structural relationship between actions and their associated reward in the form of:

$$\mathbb{E}\left[r_{t+1} \,|\, \mathcal{F}_t, x_t\right] = \mu\left(\langle x_t, \theta_\star^t \rangle\right). \tag{1}$$

The filtration $\mathcal{F}_t := \sigma(\{x_s, r_{s+1}\}_{s=1}^{t-1})$ represents the information acquired up to round t, and μ is a strictly increasing, continuously differentiable real-valued function most often referred to as the inverse link function. Note that this general framework includes both the LB ($\mu(z) = z$) as well as non-linear parametric models such as the Logistic Bandit ($\mu(z) = 1/(1+e^{-z})$). The goal of the agent is to minimize the cumulative pseudo-regret:

$$R_T := \sum_{t=1}^{T} \mu(\langle x_\star^t, \theta_\star^t \rangle) - \mu(\langle x_t, \theta_\star^t \rangle) \text{ where } x_\star^t = \arg\max_{x \in \mathcal{X}_t} \mu(\langle x, \theta_\star^t \rangle)\,.$$

We recall the following assumptions, commonly adopted in the study of parametric bandits:

Assumption 1 (Bounded decision set) *For all $t \geq 1$, $\left\|\theta_\star^t\right\|_2 \leq S$. Further, the actions have bounded norms: $\|x\|_2 \leq L$ for all $x \in \mathcal{X}_t$.*

Assumption 2 (Bounded reward) *There exists $\sigma > 0$ s.t $0 \leq r_t \leq 2\sigma$ holds almost surely.*

We will denote $\Theta = \{\theta : \|\theta\|_2 \leq S\}$ the set of admissible parameters and $\mathcal{X} = \{x : \|x\|_2 \leq L\}$. We assume that the quantities L, S and σ are known to the agent. For a given inverse link function μ, we will follow the notation from Filippi et al. (2010) and denote:

$$k_\mu = \sup_{x \in \mathcal{X}, \theta \in \Theta} \dot{\mu}\left(\langle x, \theta \rangle\right), \qquad c_\mu = \inf_{x \in \mathcal{X}, \theta \in \Theta} \dot{\mu}\left(\langle x, \theta \rangle\right)\,.$$

Note that in the linear case, we obtain $k_\mu = c_\mu = 1$. The true parameters $\{\theta_\star^t\}_{t=1}^\infty$ are unknown, and their drift is quantified by the variation *variation-budget*, which characterizes the magnitude of the non-stationarity in the environment:

$$B_{T,\star} := \sum_{t=1}^{T-1} \left\|\theta_\star^{t+1} - \theta_\star^t\right\|_2.$$

Naturally $B_{T,\star}$ is unknown, but we will assume that the agent has knowledge of an upper-bound $B_T \geq B_{T,\star}$. Such assumption is common in non-stationary bandits (Besbes et al., 2014; Cheung et al., 2019a; Zhao et al., 2020).

1.3. Existing Approaches

Principle. All existing works that tackle non-stationary GLBs (including the linear case) rely on the same conceptual approach, *i.e* addressing the reward's drift by progressively forgetting past data (Cheung et al., 2019b; Russac et al., 2019; Zhao et al., 2020). This is achieved by maintaining estimations based on a truncated history of the data, judging that old observations no longer carry valuable signal about the current ground truth $\theta_\star^t$. Formally, the learning is canonically performed through the *quasi-maximum likelihood* principle, albeit equipped with a forgetting mechanism. Let b be a primitive of μ, $\lambda > 0$, $\{w_{s,t}\}$ the sequence of forgetting weights, and define:

$$\hat{\theta}_t = \underset{\theta \in \mathbb{R}^d}{\arg\min} \sum_{s=1}^{t-1} w_{s,t} \left[b(\langle x_s, \theta \rangle) - r_{s+1} \langle x_s, \theta \rangle \right] + \frac{\lambda c_\mu}{2} \|\theta\|_2^2 \,, \tag{2}$$

This formulation covers the sliding-window approach of Cheung et al. (2019b) with $w_{s,t} = \mathbb{1}(t-s \leq D)$ (D being the length of the sliding window) and the exponential-weights of Russac et al. (2019) with $\omega_{s,t} = \gamma^{t-1-s}$ and $\gamma \in (0,1)$. The exploration is conducted according to the optimism-in-face-of-uncertainty principle; confidence regions for the ground-truth parameters are build around $\hat{\theta}_t$ and leveraged to ensure that the learner plays optimistic arms.

Linear Bandit. In the linear case, all approaches discussed hereinbefore follow this general path and announce regret rates of the form:

$$R_T = \tilde{\mathcal{O}} \left(B_T^{1/3} T^{2/3} \right) \,.$$

Anticipating Section 2.1, it turns out that this rate can only hold with a relatively strong assumption on the geometry of the arm sets $\{\mathcal{X}_t\}_t$ (at least with the existing analysis). In the general case, a correct analysis yields a degraded rate, (*c.f* Touati and Vincent, 2021), which does not match the lower-bound of Cheung et al. (2019b).

Generalized Linear Bandits. (Cheung et al., 2019b; Zhao et al., 2020) extended their LB analysis to GLBs. With the exception of an inflation of the exploration bonus to account for non-linearity, their algorithm remains the same. They claim the following regret upper-bound:

$$R_T = \tilde{\mathcal{O}} \left(k_\mu c_\mu^{-1} B_T^{1/3} T^{2/3} \right) \,,$$

which stands as a natural extension of the existing stationary bounds from Filippi et al. (2010). Not only does the previous remark regarding the validity of the rates (w.r.t T and B_T) passes on to this setting, their approach also disregards the fundamental non-linear aspect of GLBs. This is the topic of Section 2.2, and is extensively discussed in the extended version of this note (Faury et al., 2021).

2. Identification and First Solutions to Existing Mistakes

2.1. LBs: Technical Gaps and Degraded Rates

Some preliminary intuition. A strong conceptual advantage (at least from an analysis point of view) of forgetting strategies is that it allows for a natural decoupling of the *learning* and *tracking* aspects of non-stationary bandit problems. At each round t, the learning aspect is rooted in the noisy nature of the environment, which blurs the sequence of $\{\theta_s^\star\}_{s=1}^{t-1}$ that generated observed rewards.

The learning guarantees of forgetting policies can be extended from existing stationary analyses (*e.g*, Abbasi-Yadkori et al., 2011). This fundamentally requires an on-policy approach: deviations can only be measured in the directions that were played. Practically speaking, this means that the right metric to derive confidence intervals is $\|\cdot\|_{\mathbf{V_t}}$ where $\mathbf{V_t} = \sum_{s=1}^{t-1} w_{s,t} x_s x_s^\top + \lambda \mathbf{I}_d$. On the other hand, the tracking aspect is inherited from the drift of $\theta_\star^{t-1}$ to $\theta_\star^t$ which induces an incompressible estimation error. It is therefore fundamentally tied to the variation-budget B_T, which is an off-policy metric (*i.e* independent of the trajectory that was played) characterized by the ℓ_2 norm. Both aspects are conflicting sources of regret; reaching optimality requires finding the correct balance between the two of them.

A flaw in the control of the tracking error. Naturally, the tracking error can only be observed (at least at analysis time) in the directions that were actually played by the algorithm and for which rewards were collected. Henceforth, the main challenge when controlling the tracking error lies in converting its on-policy version to its off-policy counterpart (which is B_T). This is where current approaches make a mistake by claiming that this can be done at no cost on the regret. If specializing to the sliding-window mechanism, the error can be traced back to the following statement:

$$\forall t \leq T, \quad \left\| \mathbf{V}_t^{-1} \sum_{s=t-D}^{t-1} x_s x_s^\top (\theta_\star^s - \theta_\star^t) \right\|_2 \leq \sum_{s=t-D}^{t-1} \|\theta_\star^s - \theta_\star^{s+1}\| \,, \tag{3}$$

which links the deviation between $\theta_\star^t$ and $\theta_\star^s$ in each direction x_s (on-policy) to the variation-budget over the length of the sliding window (off-policy). Such a statement appears several times in the literature, for instance in (Cheung et al., 2019b, Appendix B), (Russac et al., 2019, Appendix B.3) and (Zhao et al., 2020, Appendix A). Unfortunately, this is in general false. The approach followed by previous works ties the left-hand side of Equation (3) to $\lambda_{\max}$, the highest eigenvalue of the matrix $\mathbf{V}_t^{-1} \sum_{s=t-D}^{t-1} x_s x_s^\top$. They then proceed to show that the latter is smaller than a universal constant (one which does not depend on the dimension d or the sliding-window's length D). The first step of this reasoning is false; indeed, $\mathbf{V}_t^{-1} \sum_{s=t-D}^{t-1} x_s x_s^\top$ being <u>not</u> a symmetric matrix, its operator norm cannot be bounded by its larger eigenvalue; actually, one can easily design counter-examples where the two are arbitrarily different. This indicates that the impact of this mistake on the validity of the regret bound is significant; the matrix $\mathbf{V}_t^{-1} \sum_{s=t-D}^{t-1} x_s x_s^\top$ being dependent of the algorithm's behavior, we cannot, in all generality, discard *a-priori* the events that such counter-examples arise.

Preserving the rates. We can however look at *sufficient* conditions for the current analysis to hold. In particular, it is sufficient that $\mathbf{V}_t^{-1} \sum_{s=t-D}^{t-1} x_s x_s^\top$ is a *symmetric* matrix. Equivalently, we can require for the two positive semidefinite matrices $\mathbf{V}_t^{-1}$ and $\sum_{s=t-D}^{t-1} x_s x_s^\top$ to share the same basis of eigenvectors. This is a strong requirement; not only should it hold for all $t \leq T$, but furthermore such matrices are generated by the algorithm itself. This co-diagonalizability requirement must therefore hold for virtually *any* sequence of arms $\{x_s\}$! The only reasonable situation where this can be verified arises when it is *de-facto* imposed by the geometry of the action set $\mathcal{X}$; for instance, when $\mathcal{X}$ lies along an orthogonal basis.

Proposition 1 *Let $\{e_i\}_{i=1}^d$ be an orthonormal basis of $\mathbb{R}^d$ and $\mathcal{X}$ be such that for all $x \in \mathcal{X}$, there exists $\alpha \in \mathbb{R}$, $i \in [1, d]$ such that $x = \alpha e_i$. Then on the non-stationary LB problem, forgetting strategies achieve a regret upper-bound of the form $R_T = \tilde{\mathcal{O}}(B_T^{1/3} T^{2/3})$.*

Fixing the analysis. A correct treatment of the tracking error was recently proposed by (Touati and Vincent, 2021, Section 5). They showed that a correct bounding of the left-hand side of Equation (3) leads to the following control of the tracking error:

$$\forall t \leq T, \quad \left\| \mathbf{V}_t^{-1} \sum_{s=t-D}^{t-1} x_s x_s^\top (\theta_\star^s - \theta_\star^t) \right\|_2 \leq \sqrt{dD} \sum_{s=t-D}^{t-1} \|\theta_\star^s - \theta_\star^{s+1}\| \, .$$

The apparition of the sliding-window's length D in this bound eventually shifts its optimal value (in terms of learning v.s tracking regret balance) and yields degraded rates.

Proposition 2 (Touati and Vincent (2021)) *Under general arm-set geometry, forgetting strategies achieve a regret upper-bound of the form* $R_T = \tilde{\mathcal{O}}\left(B_T^{1/4} T^{3/4}\right)$ *on the non-stationary LB problem.*

Note that this upper-bound lags behind the lower-bound of (Cheung et al., 2019a). We discuss potential reasons for this gap in Section 3.

2.2. GLBs: Troublesome Non-Linearity

We now turn to non-stationary GLBs, and ask the following question: putting aside the aforementioned technical gaps, can we directly extend the algorithms designed for the linear case?

Failure of direct LB extensions. Most of the existing works on non-stationary LB propose relatively straight-forward extensions of their algorithms to handle GLBs (Cheung et al., 2019a; Zhao et al., 2020). Unfortunately, existing analyses suffer from important caveats because they overlook a crucial feature of GLBs inherited from the non-linearity of the link function μ. Following Filippi et al. (2010), they rely on a linearization of the reward function around $\hat{\theta}_t$. Naturally, the linear approximation must accurately describe the *effective* behavior of the reward signal (characterized by the ground-truth $\theta_\star^t$). From Assumption 2, this translates in the structural constraint $\hat{\theta}_t \in \Theta$, which is implicitly assumed to hold in previous attempts. Unfortunately, there exists no proof guaranteeing that $\hat{\theta}_t \in \Theta$ could hold. Actually, existing deviation bounds (Abbasi-Yadkori et al., 2011, Theorem 1) rather suggest that in some directions, *even in the stationary case*, $\hat{\theta}_t$ can grow to be $\sqrt{d \log(t)}$ far from Θ. The situation is worse under non-stationarity since $\hat{\theta}_t$ can be B_t far from Θ. This flaw in the analysis is critical and cannot be easily fixed without severely degrading the regret guarantee. When $\hat{\theta}_t \notin \Theta$, this impacts the ratio $k_\mu c_\mu^{-1}$ which captures the degree of non-linearity of the inverse link function. For the highly non-linear logistic function, easy computations show that $c_\mu^{-1} \geq e^{SL}$. If we were to inflate the radius of the admissible set Θ from S to $S + \delta_S$ (so that it contains $\hat{\theta}_t$), the estimated non-linearity of the reward function would be even stronger and R_μ would be multiplied by a factor $e^{L\delta_S}$! Because the regret bound scales linearly with this quantity, this exponential growth would lead to prohibitively deficient performance guarantees.

Challenges. Filippi et al. (2010) countered the aforementioned difficulty by introducing a *projection* step, mapping $\hat{\theta}_t$ back to an admissible parameter $\tilde{\theta}_t \in \Theta$. The latter is then used to predict the performance of the available actions. Their projection step essentially incorporates the prior knowledge $\theta_\star \in \Theta$ (Assumption 2) without degrading the learning guarantees of the maximum likelihood estimator. The situation is different here as under parameter drift one needs to preserve both the learning and tracking guarantees of $\hat{\theta}_t$. However, and as previously discussed, both mechanisms

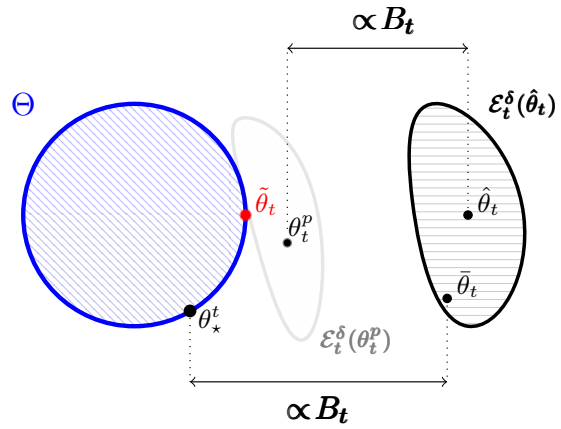

Figure 1: Illustration of the generalized projection step. The confidence set $\mathcal{E}_t^\delta(\hat{\theta}_t)$ is obtained by leveraging the learning guarantees of the estimator $\hat{\theta}_t$. It concentrates around $\bar{\theta}_t$, which tracks the sequence $\{\theta_\star^s\}_{s=1}^{t-1}$ and contrary to the stationary case can lie outside Θ. Our solution involves finding a translation θ_t^p of $\hat{\theta}_t$ such that the confidence set centered at this new reference point effectively intersects Θ, say at $\tilde{\theta}_t$. This translation occurs in a metric that preserves the tracking guarantees of $\bar{\theta}_t$. The deviations $(\theta_t^p \leftrightarrow \hat{\theta}_t)$ and $(\bar{\theta}_t \leftrightarrow \theta_\star^t)$ are linked to the parameter-drift B_t. On the other hand, the deviations $(\hat{\theta}_t \leftrightarrow \bar{\theta}_t)$ and $(\tilde{\theta}_t \leftrightarrow \theta_t^p)$ are characterized by the stochastic nature of the problem.

have different dynamics and are characterized by different metrics. This leads to a tension in the design of the projection as this requires to incorporate the knowledge $\{\theta_\star^t\} \in \Theta$, without degrading neither the learning nor the tracking guarantees. This situation therefore calls for a generalization of the projection step of Filippi et al. (2010), in order to adapt to both sources of deviation.

A compatible projection. We provide such a generalization in a companion paper to this technical note (Faury et al., 2021). The main idea is to compute the optimal translation under the tracking metric of the confidence set characterized by the learning metric. We illustrate this idea in Figure 1, and refer the interested reader to the companion paper for detailed derivations. As in the linear case, we can recover the minimax rates with sufficient assumptions on the arm-set geometry.

Proposition 3 *Let $\{e_i\}_{i=1}^d$ be an orthonormal basis of $\mathbb{R}^d$ and $\mathcal{X}$ be such that for all $x \in \mathcal{X}$, there exists $\alpha \in \mathbb{R}$, $i \in [1, d]$ such that $x = \alpha e_i$. Then forgetting strategies achieve a regret upper-bound of the form $R_T = \tilde{\mathcal{O}}(B_T^{1/3} T^{2/3})$ on the non-stationary GLB problem.*

For general arm-set geometry, sub-linear rates can still be recovered however with a sensibly more serious degradation than in the linear case. The culprit for this deterioration remains conceptually the same: the transfer from the on-policy to the off-policy tracking error comes at an additional cost due to the non-linearity of the reward function.

Proposition 4 *Under general arm-set geometry, forgetting strategies achieve a regret upper-bound of the form $R_T = \tilde{\mathcal{O}}\left(B_T^{1/5} T^{4/5}\right)$ on the non-stationary GLB problem.*

Reward Model	Assumption	Regret Upper Bound
Linear	Orthogonal action sets	$\widetilde{\mathcal{O}}\left(B_T^{1/3}T^{2/3}\right)$
	✗	$\widetilde{\mathcal{O}}\left(B_T^{1/4}T^{3/4}\right)$
Generalized Linear	Orthogonal action sets	$\widetilde{\mathcal{O}}\left(B_T^{1/3}T^{2/3}\right)$
	✗	$\widetilde{\mathcal{O}}\left(B_T^{1/5}T^{4/5}\right)$

Table 1: Non-stationary parametric bandits: regret upper-bounds.

3. Summary and Open Questions

Summary. We summarize in Table 1 the different upper-bounds available for forgetting policies for both the LB and GLB setting. In the linear case, we can only show that known rates hold up to a strong assumption on the problem's geometry. In general, the rate is degraded to $T^{3/4}$. In the GLB setting, one must be particularly careful regarding the treatment of non-linearity and existing algorithms must resort to adequate projection steps. Again, we retrieve $T^{2/3}$ rates under specific geometry but in all generality, the regret bound suffers from an even stronger degradation to $T^{4/5}$.

Open questions. This note raises several open questions. The first one concerns the nature of the difference between the LB and GLB regret upper-bounds. We postulate that this is an artefact of the proof and that an improved analysis should yield the same rates. The second question regards the optimality of forgetting strategies. Indeed, the only existing lower-bound for non-stationary parametric bandits was obtained by Cheung et al. (2019b) in the linear case, and scales as $\Omega(B_T^{1/3}T^{2/3})$. The observed gap with the upper-bounds obtained by a correct analysis could potentially be explained by a fundamental sub-optimality of the forgetting principle. We see several ways of answering this question: **(1)** by providing an improved analysis of forgetting strategies in the general case, matching the lower bound or **(2)** proving lower-bounds for forgetting policies which establish their sub-optimality. Finally, this note raises the question of the true minimax rates behind the non-stationary parametric bandit problem. Indeed, we are not aware of existing methods matching the lower-bound of Cheung et al. (2019b)[2] in the general case (*i.e* without any geometric assumption). This might be explained by the nature of this lower-bound, which is obtained on a very specific problem instance (*i.e* piece-wise stationary) and might be too specific to cover harder non-stationary problems.[3] We believe that establishing new lower-bounds under generic dynamic scenarios (*e.g* where the ground truth evolves at every round) therefore stands as a crucial missing piece in the non-stationary parametric bandit literature.

2. (Chen et al., 2019) do obtain the desired rates, however for a different adversary of the regret. It is not straight-forward to adapt their guarantees to the more challenging setting discussed here.
3. Actually, if the environment is known to be piece-wise stationary the proof strategy can be adapted to avoid the difficulties presented in this note - (*c.f* Russac et al., 2021).

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic Multi-Armed-Bandit problem with Non-Stationary Rewards. In *Advances in Neural Information Processing Systems*, pages 199–207, 2014.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A New Algorithm for Non-Stationary Contextual Bandits: Efficient, Optimal, and Parameter-Free. *arXiv preprint arXiv:1902.00980*, 2019.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Hedging the Drift: Learning to Optimize under Non-Stationarity. *arXiv preprint arXiv:1903.01461*, 2019a.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to Optimize under Non-Stationarity. In *Proceedings of the 22rd International Conference on Artificial Intelligence and Statistics*, pages 1079–1087, 2019b.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic Linear Optimization under Bandit Feedback. In *COLT*, 2008.

Louis Faury, Yoan Russac, Marc Abeille, and Clement Calauzenes. Regret Bounds for Generalized Linear Bandits under Parameter Drift. *arXiv*, 2021.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric Bandits: The Generalized Linear Case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted Linear Bandits for Non-Stationary Environments. In *Advances in Neural Information Processing Systems*, pages 12017–12026, 2019.

Yoan Russac, Louis Faury, Olivier Cappé, and Aurélien Garivier. Self-Concordant Analysis of Generalized Linear Bandits with Forgetting. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 2021.

Ahmed Touati and Pascal Vincent. Efficient Learning in Non-Stationary Linear Markov Decision Processes. *arXiv preprint arXiv:2010.12870*, 2021.

Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 2020, 2020.