Target Tracking for Contextual Bandits: Application to Demand Side Management

Margaux Brégère, Pierre Gaillard, Yannig Goude, Gilles Stoltz

Target Tracking for Contextual Bandits: Application to Demand Side Management

Margaux Brégère [1 2 3] **Pierre Gaillard** [3] **Yannig Goude** [1 2] **Gilles Stoltz** [2]

Abstract

We propose a contextual-bandit approach for demand side management by offering price incentives. More precisely, a target mean consumption is set at each round and the mean consumption is modeled as a complex function of the distribution of prices sent and of some contextual variables such as the temperature, weather, and so on. The performance of our strategies is measured in quadratic losses through a regret criterion. We offer $\sqrt{T}$ upper bounds on this regret (up to poly-logarithmic terms), for strategies inspired by standard strategies for contextual bandits (like LinUCB, see Li et al., 2010). Simulations on a real data set gathered by UK Power Networks, in which price incentives were offered, show that our strategies are effective and may indeed manage demand response by suitably picking the price levels.

1. Introduction

Electricity management is classically performed by anticipating demand and adjusting accordingly production. The development of smart grids, and in particular the installation of smart meters (see Yan et al., 2013; Mallet et al., 2014), come with new opportunities: getting new sources of information, offering new services. For example, demand-side management (also called demand-side response; see Albadi & El-Saadany, 2007; Siano, 2014 for an overview) consists of reducing or increasing consumption of electricity users when needed, typically reducing at peak times and encouraging consumption of off-peak times. This is good to adjust to intermittency of renewable energies and is made possible by the development of energy storage devices such as batteries or even electric vehicles (see Fischer et al., 2015; Kikusato et al., 2018); the storages at hand can take place at a convenient moment for the electricity provider. We will consider such a demand-side management system, based on price incentives sent to users via their smart meters. We propose here to adapt contextual bandit algorithms to that end, which are already used in online advertising. Other such systems were based on different heuristics (Shareef et al., 2018; Wang et al., 2015).

The structure of our contribution is to first provide a modeling of this management system, in Section 2. It relies on making the mean consumption as close as possible to a moving target by sequentially picking price allocations. The literature discussion of the main ingredient of our algorithms, contextual bandit theory, is postponed till Section 2.4. Then, our main results are stated and discussed in Section 3: we control our cumulative loss through a $T^{2/3}$ regret bound with respect to the best constant price allocation. A refinement as far as convergence rates are concerned is offered in Section 4. A section with simulations based on a real data set concludes the paper: Section 5. For the sake of length, most of the proofs are provided in appendix.

Notation. Without further indications, $\|x\|$ denotes the Euclidean norm of a vector x. For the other norms, there will be a subscript: e.g., the supremum norm of x is is denoted by $\|x\|_\infty$.

2. Setting and model

Our setting consists of a modeling of electricity consumption and of an aim—tracking a target consumption. Both rely on price levels sent out to the customers.

2.1. Modeling of the electricity consumption

We consider a large population of customers of some electricity provider and assume it homogeneous, which is rather reasonable (Mei et al., 2017). The consumption of each customer at each instance t depends, among others, on some exogenous factors (temperature, wind, season, day of the week, etc.), which will form a context vector $x_t \in \mathcal{X}$, where $\mathcal{X}$ is some parametric space. The electricity provider aims to manage demand response: it sets a target mean consumption c_t for each time instance. To achieve it, it

[1]EDF R&D, Palaiseau, France [2]Laboratoire de mathématiques d'Orsay, Université Paris-Sud, CNRS, Universit Paris-Saclay, Orsay, France [3]Inria Paris, France. Correspondence to: Margaux Brégère <margaux.bregere@edf.fr>.

Date: January 28, 2019
This article is not published yet, it is only a HAL / arXiv preprint.

changes electricity prices accordingly (by making it more expensive to reduce consumption or less expensive to encourage customers to consume more now rather than in some hours). We assume that $K \geqslant 2$ price levels (tariffs) are available. The individual consumption of a given customer getting tariff $j \in \{1, \ldots, K\}$ is assumed to be of the form $\varphi(x_t, j) +$ white noise, where the white noise models the variability due to the customers, and where φ is some function associating with a context x_t and a tariff j an expected consumption $\varphi(x_t, j)$. Details on and examples of φ are provided below. At instance t, the electricity provider sends tariff j to a share $p_{t,j}$ of the customers; we denote by p_t the convex vector $(p_{t,1}, \ldots, p_{t,K})$. As the population is rather homogeneous, it is unimportant to know to which specific customer a given signal was sent; only the global proportions $p_{t,j}$ matter. The mean consumption observed equals

$$Y_{t,p_t} = \sum_{j=1}^{K} p_{t,j}\, \varphi(x_t, j) + \text{noise}\,.$$

The noise term is to be further discussed below; we first focus on the φ function by means of examples.

Example 1. The simplest approach consists in considering a linear model per price level, i.e., parameters $\theta_1, \ldots, \theta_K \in \mathbb{R}^{\dim(\mathcal{X})}$ with $\varphi(x_t, j) = \theta_j^{\mathrm{T}} x_t$. We denote $\theta = (\theta_j)_{1 \leqslant j \leqslant K}$ the vector formed by aggregating all vectors θ_j.

This approach can be generalized by replacing x_t by a vector-valued function $b(x_t)$. This corresponds to the case where it is assumed that the $\varphi(\,\cdot\,, j)$ belong to some set $\mathcal{H}$ of functions $h : \mathcal{X} \to \mathbb{R}$, with a basis composed of $b_1, \ldots, b_q$. Then, $b = (b_1, \ldots, b_q)$. For instance, $\mathcal{H}$ can be given by histograms on a given grid of $\mathcal{X}$. □

Example 2. Generalized additive models (Wood, 2006) form a powerful and efficient semi-parametric approach to model electricity consumption (see, among others, Goude et al., 2014; Gaillard et al., 2016). It models the load as a sum of independent exogenous variable effects. In our simulations, see (12), we will consider a mean expected consumption of the form $\varphi(x_t, j) = \varphi(x_t, 0) + \xi_j$, that is, the tariff will have a linear impact on the mean consumption, independently of the contexts.

The baseline mean consumption $\varphi(x_t, 0)$ will be modeled as a sum of simple $\mathbb{R} \to \mathbb{R}$ functions, each taking as input a single component of the context vector:

$$\varphi(x_t, 0) = \sum_{i=1}^{Q} f^{(i)}(x_{t,h(i)})\,,$$

where $Q \geqslant 1$ and where each $h(i) \in \big\{1, \ldots, \dim(\mathcal{X})\big\}$. Some components $h(i)$ may be used several times.

When the considered component $x_{t,h(i)}$ takes continuous values, these functions $f^{(i)}$ are so-called cubic splines: $\mathcal{C}^2$–smooth functions made up of sections of cubic polynomials joined together at points of a grid (the knots). Choosing the number q_i of knots (points at which the sections join) and their locations is sufficient to determine (in closed form) a linear basis $\big(b_1^{(i)}, \ldots, b_{q_i}^{(i)}\big)$ of size q_i, see Wood (2006) for details. The function $f^{(i)}$ can then be represented on this basis by a vector of length q_i, denoted by $\theta^{(i)}$:

$$f^{(i)} = \sum_{j=1}^{q_i} \theta_j^{(i)} b_j^{(i)}\,.$$

When the considered component $x_{t,h(i)}$ takes finitely many values, we write $f^{(i)}$ as a sum of indicator functions:

$$f^{(i)} = \sum_{j=1}^{q_i} \theta_j^{(i)} \mathbf{1}_{\{v_j^{(i)}\}}\,,$$

where the $v_j^{(i)}$ are the q_i modalities for the component $h(i)$.

All in all, $\varphi(x_t, j)$ can be represented by a vector of dimension $K + q_1 + \ldots + q_Q$ obtained by aggregating the ξ_j and the vectors $\theta^{(i)}$ into a single vector. □

Both examples above show that it is reasonable to assume that there exists some unknown $\theta \in \mathbb{R}^d$ and some known transfer function ϕ such that $\varphi(x_t, j) = \phi(x_t, j)^{\mathrm{T}}\theta$.

By linearly extending ϕ in its second component, we get

$$Y_{t,p_t} = \phi(x_t, p_t)^{\mathrm{T}}\theta + \text{noise}\,.$$

We will actually not use in the sequel that $\phi(x, p)$ is linear in p: the dependency of $\phi(x, p)$ in p could be arbitrary.

We now move on to the noise term. We first recall that we assumed that our population is rather homogeneous, which is a natural feature as soon as it is large enough. Therefore, we may assume that the variabilities within the group of customers getting the same tariff j can be combined into a single random variable $\varepsilon_{t,j}$. We denote by ε_t the vector $(\varepsilon_{t,1}, \ldots, \varepsilon_{t,K})$. All in all, we will mainly consider the following model.

Model 1: tariff-dependent noise. When the electricity provider picks the convex vector p, the mean consumption obtained at time instance t equals

$$Y_{t,p} = \phi(x_t, p)^{\mathrm{T}}\theta + p^{\mathrm{T}}\varepsilon_t\,.$$

The noise vectors $\varepsilon_1, \varepsilon_2, \ldots$ are ρ–sub-Gaussian[1] i.i.d. random variables with $\mathbb{E}[\varepsilon_1] = (0, \ldots, 0)^{\mathrm{T}}$. We denote by $\Gamma = \mathrm{Var}(\varepsilon_1)$ their covariance matrix.

No assumption is made on Γ in the model above. However, when it is proportional to the $K \times K$ matrix $[1]$, the noises associated with each group can be combined into a global noise, leading to the following model. It is less realistic in practice, but we discuss it because regret bounds may be improved in the presence of a global noise.

Model 2: global noise. When the electricity provider picks the convex vector p, the mean consumption obtained at time instance t equals

$$Y_{t,p} = \phi(x_t, p)^{\mathrm{T}}\theta + e_t\,.$$

[1] A d–dimensional random vector ε is ρ–sub-Gaussian, where $\rho > 0$, if for all $\nu \in \mathbb{R}^d$, one has $\mathbb{E}\big[\mathrm{e}^{\nu^{\mathrm{T}}\varepsilon}\big] \leqslant \mathrm{e}^{\rho^2\|\nu\|^2/2}$.

The scalar noises $e_1, e_2, \dots$ are ρ–sub-Gaussian i.i.d. random variables, with $\mathbb{E}[e_1] = 0$. We denote by $\sigma^2 = \mathrm{Var}(e_1)$ the variance of the random noises e_t.

2.2. Tracking a target consumption

We now move on to the aim of the electricity provider. At each time instance t, it picks an allocation of price levels p_t and wants the observed mean consumption Y_{t,p_t} to be as close as possible to some target mean consumption c_t. This target is set in advance by another branch of the provider and p_t is to be picked based on this target: our algorithms will explain how to pick p_t given c_t but will not discuss the choice of the latter. In this article we will measure the discrepancy between the observed Y_{t,p_t} and the target c_t via a quadratic loss: $(Y_{t,p_t} - c_t)^2$.

We may set some restrictions on the convex combinations p that can be picked: we denote by $\mathcal{P}$ the set of legible allocations of price levels. This models some operational or marketing constraints that the electricity provider may encounter. We will see that whether $\mathcal{P}$ is a strict subset of all convex vectors or whether it is given by the set of all convex vectors plays no role in our theoretical analysis.

As explained in Section 3.1 and as is standard in online learning theory, to minimize the cumulative loss suffered we will minimize some regret.

2.3. Summary: online protocol

After picking an allocation of price levels p_t, the electricity provider only observes Y_{t,p_t}: it thus faces a bandit monitoring. Because of the contexts x_t, the problem considered falls under the umbrella of contextual bandits. No stochastic assumptions are made on the sequences x_t and c_t: the contexts x_t and c_t will be considered as picked by the environment. Finally, mean consumptions are assumed to be bounded between 0 and C, where C is some known maximal value.

The online protocol described in Sections 2.1 and 2.2 is stated in Protocol 1. We see that the choices x_t, c_t and p_t need to be $\mathcal{F}_{t-1}$–measurable, where

$$\mathcal{F}_{t-1} \stackrel{\text{def}}{=} \sigma(\varepsilon_1, \dots, \varepsilon_{t-1}) .$$

2.4. Literature discussion: contextual bandits

In many bandit problems the learner has access to additional information at the beginning of each round. Several settings for this side information may be considered. The adversarial case was introduced in Auer et al. (2002, Section 7, algorithm Exp4): and subsequent improvements were suggested in Beygelzimer et al. (2011) and McMahan & Streeter (2009). The case of i.i.d. contexts with rewards depending on contexts through an unknown parametric model was introduced by Wang et al. (2005b) and generalized to the non-i.i.d. setting in Wang et al. (2005a), then to the multivariate and nonparametric case in Perchet & Rigollet (2013). Hybrid versions (adversarial contexts but stochastic dependencies of the rewards on the contexts, usually in a linear fashion) are the most popular ones. They were introduced by Abe & Long (1999) and further studied in Auer (2002). A key technical ingredient to deal with them is confidence ellipsoids on the linear parameter; see Dani et al. (2008), Rusmevichientong & Tsitsiklis (2010) and Abbasi-Yadkori et al. (2011). The celebrated UCB algorithm of Lai & Robbins (1985) was generalized in this hybrid setting as the LinUCB algorithm, by Li et al. (2010) and Chu et al. (2011). Later, Filippi et al. (2010) extended it to a setting with generalized additive models and Valko et al. (2013) proposed a kernalized version of UCB. Other approaches, not relying on confidence ellipsoids, consider sampling strategies (see Gopalan et al., 2014) and are currently extended to bandit problems with complicated dependency in contextual variables (Mannor, 2018). Our model falls under the umbrella of hybrid versions considering stochastic linear bandit problems given a context. The main difference of our setting lies in how we measure performance: not directly with the rewards or their analogous quantities Y_{t,p_t} in our setting, but through how far away they are from the targets c_t.

Protocol 1 Target Tracking for Contextual Bandits

Input
Parametric context set $\mathcal{X}$
Set of legible convex weights $\mathcal{P}$
Bound on mean consumptions C
Transfer function $\phi : \mathcal{X} \times \mathcal{P} \to \mathbb{R}^d$
Unknown parameters
Transfer parameter $\theta \in \mathbb{R}^d$
Covariance matrix Γ of size $K \times K$ (Model 1)
Variance σ^2 (Model 2)
for $t = 1, 2, \dots$ **do**
Observe a context $x_t \in \mathcal{X}$ and a target $c_t \in (0, C)$
Choose an allocation of price levels $p_t \in \mathcal{P}$
Observe a resulting mean consumption

$$Y_{t,p_t} = \phi(x_t, p_t)^{\mathrm{T}}\theta + p_t^{\mathrm{T}}\varepsilon_t \qquad \text{(Model 1)}$$
$$Y_{t,p_t} = \phi(x_t, p_t)^{\mathrm{T}}\theta + e_t \qquad \text{(Model 2)}$$

Suffer a loss $(Y_{t,p_t} - c_t)^2$
end for
Aim
Minimize the cumulative loss $L_T = \sum_{t=1}^{T} (Y_{t,p_t} - c_t)^2$

3. Main result, with Model 1

This section considers Model 1.

We take inspiration from LinUCB (Li et al., 2010; Chu et al., 2011): given the form of the observed mean consumption, the key is to estimate the parameter θ. Denoting by I_d the $d \times d$ identity matrix and picking $\lambda > 0$, we classically do so according to

$$\widehat{\theta}_t \stackrel{\text{def}}{=} V_t^{-1} \sum_{s=1}^{t} Y_{s,p_s} \phi(x_s, p_s) \tag{1}$$

where $V_t \stackrel{\text{def}}{=} \lambda \mathrm{I}_d + \sum_{s=1}^{t} \phi(x_s, p_s)\phi(x_s, p_s)^{\mathrm{T}}$.

A straightforward adaptation of earlier results (see Theorem 2 of Abbasi-Yadkori et al., 2011 or Theorem 20.2 in the monograph by Lattimore & Szepesvári, 2018) yields the following deviation inequality; details are provided in Appendix A.

Lemma 1. *No matter how the provider picks the p_t, we have, for all $t \geqslant 1$ and all $\delta \in (0,1)$,*

$$\sqrt{\big(\widehat{\theta}_t - \theta\big)^{\mathrm{T}} V_t \big(\widehat{\theta}_t - \theta\big)} \stackrel{\text{def}}{=} \big\| V_t^{1/2}\big(\widehat{\theta}_t - \theta\big) \big\| \leqslant \sqrt{\lambda}\|\theta\| + \rho\sqrt{2\ln\frac{1}{\delta} + d\ln\frac{1}{\lambda} + \ln\det(V_t)}\,,$$

with probability at least $1-\delta$.

3.1. Regret as a proxy for minimizing losses

We are interested in the cumulative sum of the losses, but under suitable assumptions (e.g., bounded noise) the latter is close to the sum of the conditionally expected losses (e.g., through Hoeffding's inequality). Typical statements are of the form: for all strategies of the provider and of the environment,

$$L_T = \sum_{t=1}^{T} (Y_{t,p_t} - c_t)^2 \leqslant \sum_{t=1}^{T} \mathbb{E}\big[(Y_{t,p_t} - c_t)^2 \,\big|\, \mathcal{F}_{t-1}\big] + O\big(\sqrt{T\ln(1/\delta)}\big).$$

All regret bounds in the sequel will involve the sum of conditionally expected losses

$$\overline{L}_T \stackrel{\text{def}}{=} \sum_{t=1}^{T} \mathbb{E}\big[(Y_{t,p_t} - c_t)^2 \,\big|\, \mathcal{F}_{t-1}\big]$$

but up to adding a deviation term to all these regret bounds, we get from them a bound on the true cumulative loss L_T.

Now, the choices x_t, c_t and p_t are $\mathcal{F}_{t-1}$–measurable, where $\mathcal{F}_{t-1} = \sigma(\varepsilon_1, \ldots, \varepsilon_{t-1})$. Therefore, under Model 1,

$$\begin{aligned}
\mathbb{E}\big[(Y_{t,p_t} - c_t)^2 \,\big|\, \mathcal{F}_{t-1}\big] &= \mathbb{E}\Big[\big(\phi(x_t,p_t)^{\mathrm{T}}\theta + p_t^{\mathrm{T}}\varepsilon_t - c_t\big)^2 \,\Big|\, \mathcal{F}_{t-1}\Big] \\
&= \big(\phi(x_t,p_t)^{\mathrm{T}}\theta - c_t\big)^2 + \mathbb{E}\big[(p_t^{\mathrm{T}}\varepsilon_t)^2 \,\big|\, \mathcal{F}_{t-1}\big] \\
&\qquad + \mathbb{E}\Big[2\big(\phi(x_t,p_t)^{\mathrm{T}}\theta - c_t\big)p_t^{\mathrm{T}}\varepsilon_t \,\Big|\, \mathcal{F}_{t-1}\Big] \\
&= \big(\phi(x_t,p_t)^{\mathrm{T}}\theta - c_t\big)^2 + p_t^{\mathrm{T}}\Gamma p_t\,.
\end{aligned} \tag{2}$$

We got the rewriting

$$\overline{L}_T = \sum_{t=1}^{T} \big(\phi(x_t,p_t)^{\mathrm{T}}\theta - c_t\big)^2 + p_t^{\mathrm{T}}\Gamma p_t$$

and we therefore introduce the regret

$$\overline{R}_T = \sum_{t=1}^{T} \big(\phi(x_t,p_t)^{\mathrm{T}}\theta - c_t\big)^2 + p_t^{\mathrm{T}}\Gamma p_t - \sum_{t=1}^{T} \min_{p\in\mathcal{P}} \Big\{ \big(\phi(x_t,p)^{\mathrm{T}}\theta - c_t\big)^2 + p^{\mathrm{T}}\Gamma p \Big\}\,.$$

This will be the quantity of interest in the sequel.

3.2. Optimistic algorithm: all but the estimation of Γ

We assume that in the first n rounds an estimator $\widehat{\Gamma}_n$ of the covariance matrix Γ was obtained; details are provided in the next subsection. We explain here how the algorithm plays for rounds $t \geqslant n+1$.

We assumed that the transfer function ϕ and the bound $C > 0$ on the target mean consumptions were known. We use the notation $[x]_C = \min\big\{\max\{x, 0\}, C\big\}$ for the clipped part of a real number x (clipping between 0 and C).

We then estimate the instantaneous losses (2)

$$\ell_{t,p} \stackrel{\text{def}}{=} \mathbb{E}\big[(Y_{t,p} - c_t)^2\big] = \big(\phi(x_t,p)^{\mathrm{T}}\theta - c_t\big)^2 + p^{\mathrm{T}}\Gamma p$$

associated with each choice $p \in \mathcal{P}$ by:

$$\widehat{\ell}_{t,p} = \Big(\big[\phi(x_t,p)^{\mathrm{T}}\widehat{\theta}_{t-1}\big]_C - c_t\Big)^2 + p^{\mathrm{T}}\widehat{\Gamma}_n p\,.$$

We also denote by $\alpha_{t,p}$ deviation bounds, to be set by the analysis.

The optimistic algorithm picks, for $t \geqslant n+1$:

$$p_t \in \arg\min_{p\in\mathcal{P}} \big\{\widehat{\ell}_{t,p} - \alpha_{t,p}\big\}\,. \tag{3}$$

Comment: In linear contextual bandits, rewards are linear in θ and to maximize global gain, LinUCB (Li et al., 2010) picks a vector p which maximizes a sum of the form $\phi(x_t,p)^{\mathrm{T}}\widehat{\theta}_{t-1} + \tilde{\alpha}_{t,p}$. Here, as we want to track the target, we slightly change this expression by substituting the target c_t and taking a quadratic loss. But the spirit is similar.

3.3. Optimistic algorithm: estimation of Γ

The estimation of the covariance matrix Γ is hard to perform (on the fly and simultaneously) as the algorithm is running. We leave this problem for future research and devote here the first n rounds to this estimation. We created from scratch the estimation of Γ proposed below and studied in Lemma 2, as we could find no suitable result in the literature.

For each pair

$$(i,j) \in E \stackrel{\text{def}}{=} \big\{(i,j) \in \{1,\dots,K\}^2 : 1 \leqslant i \leqslant j \leqslant K\big\}$$

we define the weight vector $p^{(i,j)}$ as: for $k \in \{1,\dots,K\}$,

$$p_k^{(i,j)} = \begin{cases} 1 & \text{if } k = i = j, \\ 1/2 & \text{if } k \in \{i,j\} \text{ and } i \neq j, \\ 0 & \text{if } k \notin \{i,j\}. \end{cases}$$

These correspond to all weights vectors that either assign all the mass to a single component, like the $p^{(i,i)}$, or share the mass equally between two components, like the $p^{(i,j)}$ for $i \neq j$. There are $K(K+1)/2$ different weight vectors considered. We order these weight vectors, e.g., in lexicographic order, and pull them one after the other, in order. This implies that in the initial exploration phase of length n, all vectors indexed by E are selected at least

$$n_0 \stackrel{\text{def}}{=} \Big\lfloor \tfrac{2n}{K(K-1)} \Big\rfloor \geqslant \tfrac{2n}{K^2}$$

times. At the end of the exploration period, we define $\widehat{\theta}_n$ as in (1) and the estimator

$$\widehat{\Gamma}_n \in \operatorname*{arg\,min}_{\widehat{\Gamma} \in \mathcal{M}_K(\mathbb{R})} \sum_{t=1}^{n} \big(\widehat{Z}_t^2 - p_t^{\mathrm{T}} \widehat{\Gamma} p_t\big)^2 , \qquad (4)$$

where $\widehat{Z}_t \stackrel{\text{def}}{=} Y_{t,p_t} - \big[\phi(x_t,p_t)^{\mathrm{T}} \widehat{\theta}_n\big]_C$. Note that $\widehat{\Gamma}_n$ can be computed efficiently by solving a linear system as soon as K is small enough.

3.4. Statement of our main result

Theorem 1. *Fix a risk level $\delta \in (0,1)$ and a time horizon $T \geqslant 1$. Assume that the boundedness assumptions (5) hold. The optimistic algorithm (3) with an initial exploration of length $n = O(T^{2/3})$ rounds satisfies*

$$\overline{R}_T = O\Big(T^{2/3} \ln^2\big(\tfrac{T}{\delta}\big) \sqrt{\ln \tfrac{1}{\delta}}\Big)$$

with probability at least $1-\delta$.

3.5. Analysis: structure

We first indicate the boundedness assumptions that will be useful in the proof of Theorem 1 and will then provide the structure of the analysis.

Boundedness assumptions. They are all linked to the knowledge that the mean consumption lies in $(0,C)$ and indicate some normalization of the modeling:

$$\|\phi\|_\infty \leqslant 1\,, \qquad \|\theta\|_\infty \leqslant C\,, \qquad \phi^{\mathrm{T}}\theta \in [0,C]\,. \qquad (5)$$

As a consequence, $\|\theta\| \leqslant \sqrt{d}\,C$ and all eigenvalues of V_t lie in $[\lambda, \lambda+t]$, thus $\ln\big(\det(V_t)\big) \in \big[d\ln\lambda, d\ln(\lambda+t)\big]$.

The deviation bound of Lemma 1 plays a key role in the algorithm. We introduce the following upper bound on it:

$$B_t(\delta) \stackrel{\text{def}}{=} \sqrt{\lambda d}\,C + \rho\sqrt{2\ln\tfrac{1}{\delta} + d\ln\big(1+\tfrac{t}{\lambda}\big)}\,. \qquad (6)$$

Finally, we also assume that a bound G is known, such that

$$\forall p \in \mathcal{P}, \qquad p^{\mathrm{T}}\Gamma p \leqslant G\,.$$

A last consequence of all these boundedness assumptions is that $L \stackrel{\text{def}}{=} C^2 + G$ upper bounds the (conditionally) expected losses $\ell_{t,p} = \big(\phi(x_t,p)^{\mathrm{T}}\theta - c_t\big)^2 + p^{\mathrm{T}}\Gamma p$.

Structure of the analysis. The analysis exploits how well the $\widehat{\theta}_t$ estimate θ and how well $\widehat{\Gamma}_n$ estimates Γ. The regret bound, as is clear from Proposition 1 below, also consists of these two parts. The proof is to be found in Appendix B.

Proposition 1. *Fix a risk level $\delta \in (0,1)$ and an exploration budget $n \geqslant 1$. Assume that the boundedness assumptions (5) hold. Consider an estimator $\widehat{\Gamma}_n$ of Γ such that $\sup_{p\in\mathcal{P}} \big|p^{\mathrm{T}}(\Gamma - \widehat{\Gamma}_n)p\big| \leqslant \gamma$ with probability at least $1-\delta/2$, for some $\gamma > 0$.*

Then choosing $\lambda > 0$ and

$$a_{t,p} = \min\Big\{L,\ 2C\,B_{t-1}(\delta t^{-2})\,\big\|V_{t-1}^{-1/2}\phi(x_t,p)\big\|\Big\}, \qquad \alpha_{t,p} = \gamma + a_{t,p}\,, \qquad (7)$$

the optimistic algorithm (3) ensures that w.p. $1-\delta$,

$$\textstyle\sum_{t=n+1}^{T} \ell_{t,p_t} - \sum_{t=n+1}^{T} \min_{p\in\mathcal{P}} \ell_{t,p} \leqslant 2\sum_{t=n+1}^{T} \alpha_{t,p_t}\,.$$

Comment: Li et al. (2010) pick $\alpha(t,p)$ proportional to $\big\|V_{t-1}^{-1/2}\phi(x_t,p)\big\|$ only, but we need an additional term to account for the covariance matrix.

We are thus left with studying how well $\widehat{\Gamma}_n$ estimates Γ and with controlling the sum of the $a_{t,p}$. The next two lemmas take care of these issues. Their proofs are to be found in Appendices C and D.

Lemma 2. *For all $\delta \in (0,1)$, the estimator (4) satisfies: with probability at least $1-\delta$,*

$$\sup_{p\in\mathcal{P}} \Big|p^{\mathrm{T}}\big(\widehat{\Gamma}_n - \Gamma\big)p\Big| \leqslant K^2\Big(\tfrac{K}{2}+4\Big)\kappa_n\sqrt{\tfrac{1}{n}} = O\Big(\frac{1}{\sqrt{n}}\ln^2(n/\delta)\sqrt{\ln(1/\delta)}\Big)\,,$$

where $\kappa_n = \big(C + 2M_n\big)B_n(\delta/3) + M_n'$ with $M_n = \rho/2 + \ln(6n/\delta)$ and $M_n' = M_n^2\sqrt{2\ln(3K^2/\delta)} + 2\sqrt{\exp(2\rho)\delta/6}$.

Comment: We derived the estimator of Γ as well as Lemma 2 from scratch: we could find no suitable result in the literature for estimating Γ in our context.

Lemma 3. *No matter how the environment and provider pick the x_t and p_t,*

$$\textstyle\sum_{t=n+1}^{T} a_{t,p_t} \leqslant \sqrt{\big(2C\overline{B}\big)^2 + \frac{L^2}{2}}\sqrt{dT\ln\frac{\lambda+T}{\lambda}} = O\big(\sqrt{T\ln T\ln(T/\delta)}\big)\,,$$

where $\overline{B} \stackrel{\text{def}}{=} \sqrt{d\lambda}\,C + \rho\sqrt{2\ln(T^2/\delta) + d\ln(1+T/\lambda)}$.

Comment: This lemma follows from a straightforward adaptation/generalization of Lemma 19.1 of the monograph

by Lattimore & Szepesvári (2018); see also a similar result in Lemma 3 by Chu et al. (2011).

We are now ready to conclude the proof of Theorem 1. Indeed, using for the first n rounds that $L = C^2 + G$ upper bounds the (conditionally) expected losses ℓ_{t,p_t}, Proposition 1 and Lemmas 2 and 3 show that, w.p. $1-\delta$

$$\overline{R}_T \leqslant nL + T\gamma + \textstyle\sum_{t=n+1}^{T} a_{t,p_t}$$
$$\leqslant nL + O\Big(T \ln^2\big(\tfrac{n}{\delta}\big)\sqrt{\tfrac{\ln(1/\delta)}{n}} + \sqrt{T\,\ln T \ln(T/\delta)}\Big).$$

Picking n of order $T^{2/3}$ concludes the proof.

Comment: The algorithm of Theorem 1 considered above depends on δ via the tuning (7) of α. But we can also have a result in expectation, i.e., a regret defined with $\mathbb{E}\big[\ell_{t,p_t}\big]$ and $\min \mathbb{E}\big[\ell_{t,p}\big]$, in which case the algorithm can be made independent of δ. Only Step 3 of the proof of Proposition 1 is to be modified. The same rates in T are obtained.

4. Fast rates, with Model 2

In this section, we consider Model 2 and show that under an attainability condition stated below, the order of magnitude of the regret bound in Theorem 1 can be reduced to a poly-logarithmic rate. This result is in strong contrast with the typical results for contextual bandits. We underline in the proof the key step where we gain orders of magnitude in the regret bound. Before doing so, we note that similarly to Section 3.1,

$$\mathbb{E}\big[(Y_{t,p} - c_t)^2\big] = \big(\phi(x_t,p)^{\mathrm{T}}\theta - c_t\big)^2 + \sigma^2\,, \qquad (8)$$

which leads us to introduce a regret $\overline{R}_T$ defined by $\overline{R}_T = \displaystyle\sum_{t=1}^{T}\big(\phi(x_t,p_t)^{\mathrm{T}}\theta - c_t\big)^2 - \sum_{t=1}^{T}\min_{p\in\mathcal{P}}\Big\{\big(\phi(x_t,p)^{\mathrm{T}}\theta - c_t\big)^2\Big\}\,.$

Thus, as far as the minimization of the regret is concerned, Model 2 is a special case of Model 1, corresponding to a matrix Γ that can be taken as the null matrix $[0]$. Of course, as explained in Section 2.1, the covariance matrix Γ of Model 2 is $\sigma^2[1]$ in terms of real modeling, but in terms of regret-minimization it can be taken as $\Gamma = [0]$. Therefore, all results established above for Model 1 extend to Model 2, but under an additional assumption stated below, the $\sqrt{T}$ rates (up to poly-logarithmic terms) obtained above can be reduced to poly-logarithmic rates only.

Assumption 2: Attainability. For each time instance $t \geqslant 1$, the expected mean consumption is attainable, i.e.,

$$\exists p \in \mathcal{P}: \quad \phi(x_t,p)^{\mathrm{T}}\theta = c_t\,.$$

We denote by $p_t^\star$ such an element of $\mathcal{P}$.

In Model 2 and under this assumption, the expected losses $\ell_{t,p}$ defined in (8) are such that, for all $t \geqslant 1$ and all $x_t \in \mathcal{X}$,

$$\min_{p\in\mathcal{P}} \ell_{t,p} = \ell_{t,p_t^\star} = \sigma^2\,. \qquad (9)$$

As in Model 2 the variance terms σ^2 cancel out when considering the regret, the variance σ^2 does not need to be estimated. Our optimistic algorithm thus takes a simpler form. For each $t \geqslant 2$ and $p \in \mathcal{P}$ we consider the same estimators (1) of θ as before and then define

$$\widetilde{\ell}_{t,p} = \big(\phi(x_t,p)^{\mathrm{T}}\widehat{\theta}_{t-1} - c_t\big)^2$$

(no clipping needs to be considered in this case). We set

$$\beta_{t,p} = B_{t-1}(\delta t^{-2})^2\,\big\|\,V_{t-1}^{-1/2}\phi(x_t,p)\,\big\|^2 \qquad (10)$$

and then pick: $$p_t \in \arg\min_{p\in\mathcal{P}}\Big\{\widetilde{\ell}_{t,p} - \beta_{t,p}\Big\} \qquad (11)$$

for $t \geqslant 2$ and p_1 arbitrarily. The tuning parameter $\lambda > 0$ is hidden in $B_{t-1}(\delta t^{-2})^2$. We get the following theorem, whose proof is deferred to Appendix E and re-uses many parts of the proofs of Proposition 1 and Lemma 3.

Theorem 2. *In Model 2, assume that the boundedness assumptions (5) hold. Then, the optimistic algorithm (11), tuned with $\lambda > 0$, ensures that for all $\delta \in (0,1)$,*

$$\overline{R}_T \leqslant d\big(4\overline{B}^2 + \tfrac{C^2}{2}\big)\ln\tfrac{\lambda+T}{\lambda} = O\big(\ln^2(T)\big)\,,$$

w.p. at least $1-\delta$, where $\overline{B}$ is defined as in Lemma 3.

5. Simulations

Our simulations rely on a real data set of residential electricity consumption, in which different tariffs were sent to the customers according to some policy. But of course, we cannot test an alternative policy on historical data (we only observed the outcome of the tariffs sent) and therefore need to build a data simulator. This is what we explain first.

5.1. The underlying real data set / The simulator

We consider the data set "*SmartMeter Energy Consumption Data in London Households*[2]". These open data are published by UK Power Networks and contain energy consumption (in kWh per half hour) at half hourly intervals of a thousand customers subjected to dynamic energy prices. A single tariff (among High–1, Normal–2 or Low–3) is offered to all the population for each half hour interval. The tariffs were announced in advance. The report by Schofield et al. (2014) provides a full description of this experimentation and an exhaustive analysis of results. We only kept customers with more than 95% of data available (980 clients) and considered their mean consumption. (Such a level of aggregation enables a proper estimation of the load whereas individual consumptions are erratic, see, e.g., Sevlian & Rajagopal, 2018.) As far as contexts are concerned, we considered half-hourly temperatures τ_t in London, obtained from the NOAA[3]. We also created calendar variables: the

[2] https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households

[3] https://www.noaa.gov/ – We managed missing data by interpolating them linearly.

day of the week w_t (equal to 1 for Monday, 2 for Tuesday, etc.), the half-hour of the day $h_t \in \{1, \dots, 48\}$, and the position in the year: $y_t \in [0, 1]$, linear values between $y_t = 0$ on January 1st at 00:00 and $y_t = 1$ on December the 31st at 23:59.

Realistic simulator. It is based on the following additive model, which breaks down time by half hours:

$$\varphi(x_t, j) = \sum_{h=1}^{48} \left[f_h^\tau(\tau_t) + f_h^y(y_t) + \eta_h \right] \mathbf{1}_{\{h_t = h\}} + \sum_{w=1}^{7} \zeta_w \mathbf{1}_{\{w_t = w\}} + \xi_j \,, \quad (12)$$

where the f_h^τ and f_h^y are functions catching the effect of the temperature and of the yearly seasonality. As explained in Example 2, the transfer parameter θ gathers coordinates of the f_h^τ and the f_h^y in bases of splines, as well as the coefficients η_h, ζ_w and ξ_j. Here, we work under the assumption that exogenous factors do not impact customers' reaction to tariff changes (which is admittedly a first step, and more complex models could be considered). Our algorithms will have to sequentially estimate the parameter, but we also need to set it to get our simulator in the first place. We do so by exploiting historical data together with the allocations of prices picked, of the form $(0, 1, 0)$, $(1, 0, 0)$ and $(0, 0, 1)$ only on these data (all customers were getting the same tariff), and apply the formula (1) through the R–package `mgcv` (which replaces the λ identity matrix with a slightly more complex definite positive matrix S, see Wood, 2006). The deterministic part of the obtained model is realistic enough: its adjusted R-square on historical observations equals 92% while its mean absolute percentage error equals 8.82%. Now, as far as noise is concerned, we take multivariate Gaussian noise vectors ε_t, where the covariance matrix Γ was built again based on realistic values. The diagonal coefficients $\Gamma_{j,j}$ are given by the empirical variance of the residuals associated with tariff j, while non-diagonal coefficients $\Gamma_{j,j'}$ are given by the empirical covariance of between residuals of tariffs j and j' at times t and $t + 48$, and times t and $t - 48$.

5.2. Design of our experiment

Target creation. We focus on attainable targets which stay in the convex envelope of the mean consumption associated with the High–1 and Low–3 tariffs, namely, $\varphi(x_t, 1) \leqslant c_t \leqslant \varphi(x_t, 3)$. To smooth consumption, we pick c_t near $\varphi(x_t, 3)$ during the night and near $\varphi(x_t, 1)$ in the evening. These hypothesis can be seen as an ideal configuration where targets and customers portfolio are in a way compatible.

$\mathcal{P}$ **restriction.** We assume that the electricity provider cannot send Low and High tariffs at the same round and that population can be split in $N = 100$ equal parts. Thus, $\mathcal{P}$ is restricted to the grid

$$\left\{ \left(\tfrac{i}{N}, 1 - \tfrac{i}{N}, 0\right),\ (0, 1, 0),\ \left(0, \tfrac{i}{N}, 1 - \tfrac{i}{N}\right),\ i = 1, \dots, N \right\}$$

Training period, testing period. We create one year of data using historical contexts and assuming that only Normal tariffs are picked: $p_t = (0, 1, 0)$; this is a training period. Then the provider starts exploring the effects of tariffs for an additional month (a January month, based on the historical contexts) and freely picks the p_t according to our algorithm; this is the testing period. The rationale is that this is how electricity providers do and then, θ is better estimated. Its estimation is still performed via the formula (1) and as indicated above (with the `mgcv` package), including the year when only $p_t = (0, 1, 0)$ allocations were picked. To simplify the analysis we assume that the algorithm knows the covariance matrix used by the simulator. To make sure that learning focuses on the parameters ξ_j, as other parameters were decently estimated in the training period, we modify the exploration term $\alpha_{t,p}$ of (3) into

$$\alpha_{t,p} = 2CB_{t-1}(\delta t^{-2}) \| \tilde{V}_{t-1}^{-1/2} p_t \| \,,$$

with $\tilde{V}_{t-1} = \lambda I_d + \sum_{s=1}^{t-1} p_s p_s^{\mathrm{T}}$. We pick a convenient value for λ.

5.3. Results

Algorithms were run 200 times each. The simplest set of results is provided in Figure 3: the regrets suffered on each run are compared to the theoretical orders of magnitude of the regret bounds. As expected, we observe a lower regrets for Model 2.

The bottom parts of Figures 1–2 indicate, for a single run, which allocation vectors p_t were picked over time. During the first day of the testing period, the algorithms explore the effect of tariffs by sending the same tariff to all customers (the p_t vectors are Dirac masses) while at the end of the testing period, they cleverly exploit the possibility to split the population in two groups of tariffs. Note that, over the first iterations, the exploration term for Model 2 is much larger than the exploitation term (but quickly vanishes), which leads to an initiak quasi-deterministic exploration and an erratic consumption (unlike in Model 1).

We obtain an approximation of the expected mean consumption $\varphi(x_t, p_t)$ by averaging the 200 observed consumptions, and this is the main (black, solid) line to look at in the top parts of Figures 1–2. Four plots are depicted depending on the day of the testing period (first, last) and of the model considered. These (approximated) expected mean consumptions may be compared to the targets set (dashed red line). The algorithms seem to perform better on the last day of the testing period for Model 2 than for Model 1 as the expected mean consumption seems closer to the target. However, in Model 1, the algorithm has to pick tariffs leading to the best bias-variance trade-off (the expected loss features a variance term). This is why the average consumption does not overlap the target as in Model 2. This results in a slightly biased

estimator of the mean consumption in Model 1.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Abe, N. and Long, P. M. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pp. 3–11, 1999.

Albadi, M. H. and El-Saadany, E. F. Demand response in electricity markets: An overview. In *2007 IEEE Power Engineering Society General Meeting*, pp. 1–5, June 2007. doi: 10.1109/PES.2007.385728.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26, 2011.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS'11)*, pp. 208–214, 2011.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. *21st Annual Conference on Learning Theory*, 2008.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23*, pp. 586–594. Curran Associates, Inc., 2010.

Fischer, D., Scherer, J., Flunk, A., Kreifels, N., Byskov-Lindberg, K., and Wille-Haussmann, B. Impact of hp, chp, pv and evs on households' electric load profiles. In *2015 IEEE Eindhoven PowerTech*, pp. 1–6, June 2015. doi: 10.1109/PTC.2015.7232784.

Gaillard, P., Goude, Y., and Nedellec, R. Additive models and robust aggregation for gefcom2014 probabilistic electric load and electricity price forecasting. *International Journal of forecasting*, 32(3):1038–1050, 2016.

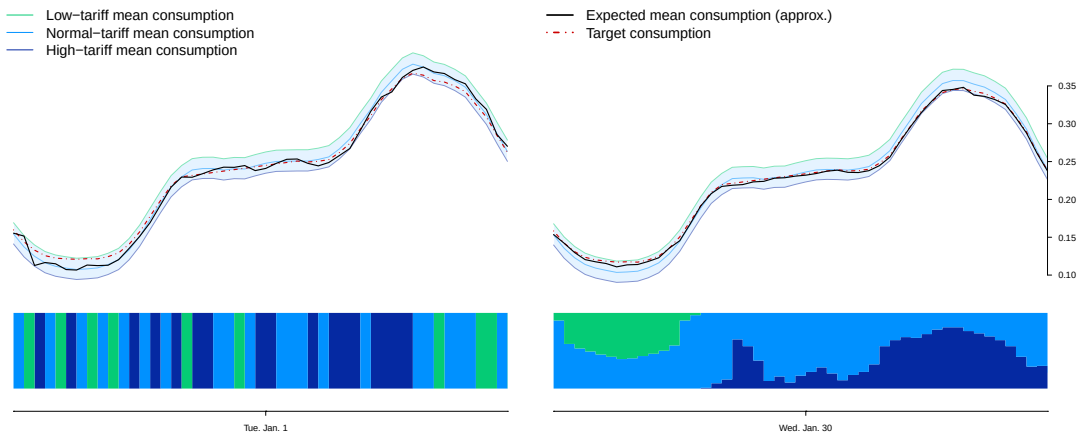

Figure 1. Left: January 1st (first day of the testing set). *Right*: January 31st (last day of the testing set).

Top: 200 runs are considered. Plot: average of mean consumptions over 200 runs for the algorithm associated with Model 1 (full black line); target consumption (dashed red line); mean consumption associated with each tariff (Low–1 in green, Normal–2 in blue and High–3 in navy). The envelope of attainable targets is in pastel blue.

Bottom: A single run is considered. Plot: proportions p_t used over time.

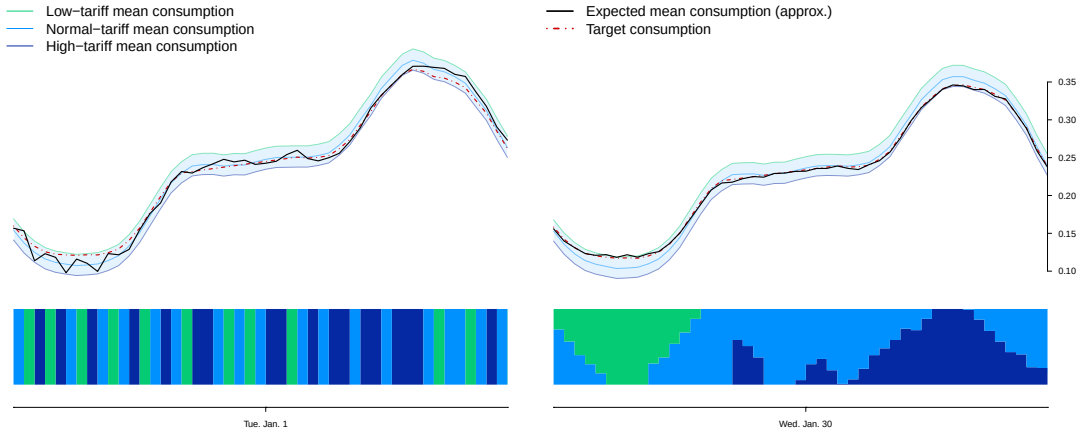

Figure 2. Same legend, but with Model 2 (full black line).

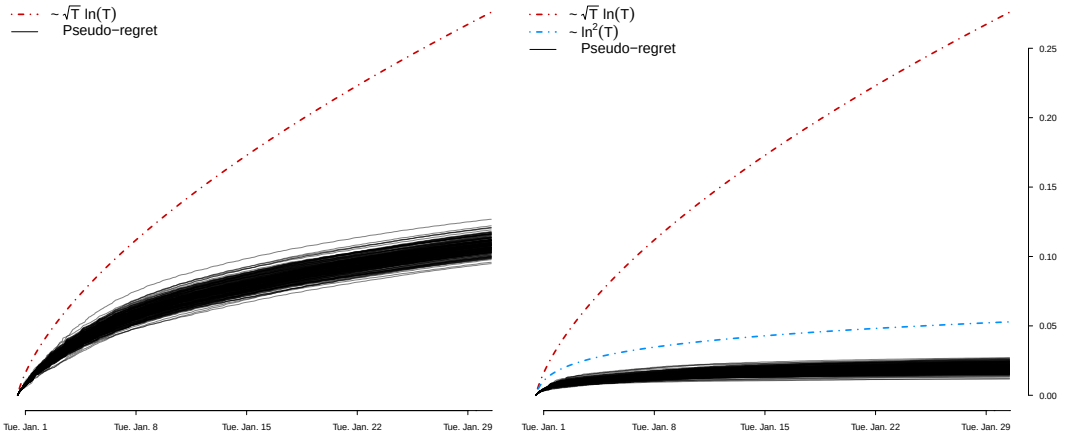

Figure 3. Regret curves for each of the 200 runs for Model 1 (*left*) and Model 2 (*right*). We also provide plots of $c\sqrt{T}\ln T$ and $c'\ln^2(T)$ for some well-chosen constants $c, c' > 0$.

Gopalan, A., Mannor, S., and Mansour, Y. Thompson sampling for complex online problems. In *International Conference on Machine Learning*, pp. 100–108, 2014.

Goude, Y., Nedellec, R., and Kong, N. Local short and middle term electricity load forecasting with semi-parametric additive models. *IEEE transactions on smart grid*, 5(1): 440–446, 2014.

Kikusato, H., Mori, K., Yoshizawa, S., Fujimoto, Y., Asano, H., Hayashi, Y., Kawashima, A., Inagaki, S., and Suzuki, T. Electric vehicle charge-discharge management for utilization of photovoltaic by coordination between home and grid energy management systems. *IEEE Transactions on Smart Grid*, pp. 1–1, 2018. ISSN 1949-3053. doi: 10.1109/TSG.2018.2820026.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22, 1985.

Lattimore, T. and Szepesvári, C. Bandit algorithms. *preprint*, 2018.

Li, L., Chu, W., Langford, J., and Schapire, R. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*, pp. 661–670, 2010.

Mallet, P., Granstrom, P. O., Hallberg, P., Lorenz, G., and Mandatova, P. Power to the people!: European perspectives on the future of electric distribution. *IEEE Power and Energy Magazine*, 12(2):51–64, March 2014. ISSN 1540-7977. doi: 10.1109/MPE.2013.2294512.

Mannor, S. Misspecified and complex bandits problems, 2018. Talk at 50èmes Journées de Statistique, EDF Lab Paris Saclay, May 31. 2018.

McMahan, H. B. and Streeter, M. J. Tighter bounds for multi-armed bandits with expert advice. In *COLT*, 2009.

Mei, J., De Castro, Y., Goude, Y., and Hébrail, G. Nonnegative matrix factorization for time series recovery from a few temporal aggregates. In *International Conference on Machine Learning*, pp. 2382–2390, 2017.

Perchet, V. and Rigollet, P. The multi-armed bandit problem with covariates. *The Annals of Statistics*, pp. 693–721, 2013.

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

Schofield, J., Carmichael, R., Tindemans, S., Woolf, M., Bilton, M., and Strbac, G. Residential consumer responsiveness to time-varying pricing. 2014.

Sevlian, R. and Rajagopal, R. A scaling law for short term load forecasting on varying levels of aggregation. 98: 350–361, 06 2018.

Shareef, H., Ahmed, M. S., Mohamed, A., and Al Hassan, E. Review on home energy management system considering demand responses, smart technologies, and intelligent controllers. *IEEE Access*, 6:24498–24509, 2018.

Siano, P. Demand response and smart grids?a survey. *Renewable and Sustainable Energy Reviews*, 30:461 – 478, 2014. ISSN 1364-0321. doi: https://doi.org/10.1016/j.rser.2013.10.022. URL http://www.sciencedirect.com/science/article/pii/S1364032113007211.

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. Finite-time analysis of kernelised contextual bandits. *arXiv preprint arXiv:1309.6869*, 2013.

Wang, C.-C., Kulkarni, S. R., and Poor, H. V. Arbitrary side observations in bandit problems. *Advances in Applied Mathematics*, 34(4):903–938, 2005a.

Wang, C.-C., Kulkarni, S. R., and Poor, H. V. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3):338–355, 2005b.

Wang, Y., Chen, Q., Kang, C., Zhang, M., Wang, K., and Zhao, Y. Load profiling and its application to demand response: A review. *Tsinghua Science and Technology*, 20(2):117–129, 2015.

Wood, S. *Generalized Additive Models: An Introduction with R*. CRC Press, 2006.

Yan, Y., Qian, Y., Sharif, H., and Tipper, D. A survey on smart grid communication infrastructures: Motivations, requirements and challenges. *IEEE Communications Surveys Tutorials*, 15(1):5–20, First 2013. ISSN 1553-877X. doi: 10.1109/SURV.2012.021312.00034.

Target Tracking for Contextual Bandits: Application to Demand Side Management

Appendices

We provide the proofs in order of appearance of the corresponding result:

- The proof of Lemma 1 in Appendix A
- The proof of Proposition 1 in Appendix B
- The proof of Lemma 2 in Appendix C
- The proof of Lemma 3 in Appendix D
- The proof of Theorem 2 in Appendix E

A. Proof of Lemma 1

The proof below relies on Laplace's method on super-martingales, which is a standard argument to provide confidence bounds on a self-normalized sum of conditionally centered random vectors. See Theorem 2 of Abbasi-Yadkori et al. (2011) or Theorem 20.2 in the monograph by Lattimore & Szepesvári (2018).

Under Model 1 and given the definition of V_t, we have the rewriting

$$\begin{aligned}\widehat{\theta}_t &= V_t^{-1}\sum_{s=1}^{t}\phi(x_s,p_s)Y_{s,p_s}\\ &= V_t^{-1}\sum_{s=1}^{t}\phi(x_s,p_s)\big(\phi(x_s,p_s)^{\mathrm{T}}\theta + p_s^{\mathrm{T}}\varepsilon_s\big)\\ &= V_t^{-1}\big((V_t-\lambda \mathrm{I}_d)\theta + M_t\big) = \theta - \lambda V_t^{-1}\theta + V_t^{-1}M_t\,,\end{aligned}$$

where we introduced

$$M_t = \sum_{s=1}^{t}\phi(x_s,p_s)p_s^{\mathrm{T}}\varepsilon_s\,,$$

which is a martingale with respect to $\mathcal{F}_t = \sigma(\varepsilon_1,\dots,\varepsilon_t)$. Therefore, by a triangle inequality,

$$\begin{aligned}\big\|\, V_t^{1/2}\big(\widehat{\theta}_t-\theta\big)\,\big\| &= \big\|-\lambda V_t^{-1/2}\theta + V_t^{-1/2}M_t\big\|\\ &\leqslant \lambda\big\|\, V_t^{-1/2}\theta\,\big\| + \big\|\, V_t^{-1/2}M_t\,\big\|\,.\end{aligned}$$

On the one hand, given that all eigenvalues of the symmetric matrix V_t are larger than λ (given the $\lambda \mathrm{I}_d$ term in its definition), all eigenvalues of $V_t^{-1/2}$ are smaller than $1/\sqrt{\lambda}$ and thus,

$$\lambda\big\|\, V_t^{-1/2}\theta\,\big\| \leqslant \lambda\,\frac{1}{\sqrt{\lambda}}\|\theta\| = \sqrt{\lambda}\|\theta\|\,.$$

We now prove, on the other hand, that with probability at least $1-\delta$,

$$\big\|\, V_t^{-1/2}M_t\,\big\| \leqslant \rho\sqrt{2\ln\frac{1}{\delta} + d\ln\frac{1}{\lambda} + \ln\det(V_t)}\,,$$

which will conclude the proof of the lemma.

Step 1: Introducing super-martingales. For all $\nu \in \mathbb{R}^d$, we consider

$$S_{t,\nu} = \exp\left(\nu^{\mathrm{T}} M_t - \frac{\rho^2}{2}\nu^{\mathrm{T}} V_t \nu\right)$$

and now show that it is an $\mathcal{F}_t$–super-martingale. First, note that since the common distribution of the $\varepsilon_1, \varepsilon_2, \ldots$ is ρ–sub-Gaussian, then for all $\mathcal{F}_{t-1}$–measurable random vectors ν_{t-1},

$$\mathbb{E}\left[\mathrm{e}^{\nu_{t-1}^{\mathrm{T}}\varepsilon_t} \,\middle|\, \mathcal{F}_{t-1}\right] \leqslant \mathrm{e}^{\rho^2\|\nu_{t-1}\|^2/2}. \tag{13}$$

Now,

$$S_{t,\nu} = S_{t-1,\nu}\,\exp\left(\nu^{\mathrm{T}}\phi(x_t,p_t)p_t^{\mathrm{T}}\varepsilon_t - \frac{\rho^2}{2}\nu^{\mathrm{T}}\phi(x_t,p_t)\phi(x_t,p_t)^{\mathrm{T}}\nu\right)$$

where, by using the sub-Gaussian assumption (13) and the fact that $\sum_j p_{j,t}^2 \leqslant 1$ for all convex weight vectors p_t,

$$\mathbb{E}\left[\exp\left(\nu^{\mathrm{T}}\phi(x_t,p_t)p_t^{\mathrm{T}}\varepsilon_t\right) \,\middle|\, \mathcal{F}_{t-1}\right] \leqslant \exp\left(\frac{\rho^2}{2}\nu^{\mathrm{T}}\phi(x_t,p_t)\underbrace{p_t^{\mathrm{T}}p_t}_{\leqslant 1}\phi(x_t,p_t)^{\mathrm{T}}\nu\right).$$

This implies $\mathbb{E}\left[S_{t,\nu}\middle|\mathcal{F}_{t-1}\right] \leqslant S_{t-1,\nu}$.

Note that the rewriting of $S_{t,\nu}$ in its vertex form is, with $m = V_t^{-1}M_t/\rho^2$:

$$\begin{aligned} S_{t,\nu} &= \exp\left(\frac{1}{2}(\nu-m)^{\mathrm{T}}\,\rho^2 V_t\,(\nu-m) + \frac{1}{2}m^{\mathrm{T}}\rho^2 V_t\,m\right) \\ &= \exp\left(\frac{1}{2}(\nu-m)^{\mathrm{T}}\,\rho^2 V_t\,(\nu-m)\right) \\ &\qquad \times \exp\left(\frac{1}{2\rho^2}\left\|\,V_t^{-1/2}M_t\,\right\|^2\right). \end{aligned}$$

Step 2: Laplace's method—integrating $S_{t,\nu}$ over $\nu \in \mathbb{R}^d$. The basic observation behind this method is that (given the vertex form) $S_{t,\nu}$ is maximal at $\nu = m = V_t^{-1}M_t/\rho^2$ and then equals $\exp\left(\left\|\,V_t^{-1/2}M_t\,\right\|^2/(2\rho^2)\right)$, which is (a transformation of) the quantity to control. Now, because the exp function quickly vanishes, the integral over $\nu \in \mathbb{R}^d$ is close to this maximum. We therefore consider

$$\overline{S}_t = \int_{\mathbb{R}^d} S_{t,\nu}\,\mathrm{d}\nu\,.$$

We will make repeated uses of the fact that the Gaussian density functions,

$$\nu \longmapsto \frac{1}{\sqrt{\det(2\pi C)}}\,\exp\left((\nu-m)^{\mathrm{T}}C^{-1}(\nu-m)\right),$$

where $m \in \mathbb{R}^d$ and C is a (symmetric) positive-definite matrix, integrate to 1 over $\mathbb{R}^d$. This gives us first the rewriting

$$\overline{S}_t = \sqrt{\det\left(2\pi\rho^{-2}V_t^{-1}\right)}\,\exp\left(\frac{1}{2\rho^2}\left\|\,V_t^{-1/2}M_t\,\right\|^2\right).$$

Second, by the Fubini-Tonelli theorem and the super-martingale property

$$\mathbb{E}\left[S_{t,\nu}\right] \leqslant \mathbb{E}\left[S_{0,\nu}\right] = \exp\left(-\lambda\rho^2\|\nu\|^2/2\right),$$

we also have

$$\begin{aligned} \mathbb{E}\left[\overline{S}_t\right] &\leqslant \int_{\mathbb{R}^d}\exp\left(-\lambda\rho^2\|\nu\|^2/2\right)\mathrm{d}\nu \\ &= \sqrt{\det\left(2\pi\rho^{-2}\lambda^{-1}\mathrm{I}_d\right)}\,. \end{aligned}$$

Combining the two statements, we proved

$$\mathbb{E}\left[\exp\left(\frac{1}{2\rho^2}\left\|\,V_t^{-1/2}M_t\,\right\|^2\right)\right] \leqslant \sqrt{\frac{\det\left(V_t\right)}{\lambda^d}}\,.$$

Step 3: Markov-Chernov bound. For $u > 0$,

$$\begin{aligned} &\mathbb{P}\left[\left\|\,V_t^{-1/2}M_t\,\right\| > u\right] \\ &= \mathbb{P}\left[\frac{1}{2\rho^2}\left\|\,V_t^{-1/2}M_t\,\right\|^2 > \frac{u^2}{2\rho^2}\right] \\ &\leqslant \exp\left(-\frac{u^2}{2\rho^2}\right)\mathbb{E}\left[\exp\left(\frac{1}{2\rho^2}\left\|\,V_t^{-1/2}M_t\,\right\|^2\right)\right] \\ &\leqslant \exp\left(-\frac{u^2}{2\rho^2} + \frac{1}{2}\ln\frac{\det\left(V_t\right)}{\lambda^d}\right) = \delta \end{aligned}$$

for the claimed choice

$$u = \rho\sqrt{2\ln\frac{1}{\delta} + d\ln\frac{1}{\lambda} + \ln\det(V_t)}\,.$$

B. Proof of Proposition 1

Comment: The main difference with the regret analysis of LinUCB provided by Chu et al. (2011) or Lattimore & Szepesvári (2018) is in the first part of *Step 1*, as we need to deal with slightly more complicated quantities: not just with linear quantities of the form $\phi(x_t,p)^{\mathrm{T}}\theta$. Steps 2 and 3 are easy consequences of Step 1.

We show below (*Step 1*) that for all $t \geqslant 2$, if

$$\left\| V_{t-1}^{1/2}\big(\widehat{\theta}_{t-1}-\theta\big)\right\| \leqslant B_{t-1}(\delta t^{-2}) \qquad \text{and} \qquad \left\|\Gamma-\widehat{\Gamma}_t\right\|_{\infty} \leqslant \gamma\,, \quad (14)$$

then

$$\forall p\in\mathcal{P}, \qquad \left|\ell_{t,p}-\widehat{\ell}_{t,p}\right| \leqslant \alpha_{t,p}\,. \quad (15)$$

Property (15), for those t for which it is satisfied, entails (*Step 2*) that the corresponding instantaneous regrets are bounded by

$$r_t \stackrel{\text{def}}{=} \ell_{t,p_t}-\min_{p\in\mathcal{P}}\ell_{t,p} \leqslant 2\alpha_{t,p_t}\,.$$

It only remains to deal (*Step 3*) with the rounds t when (15) does not hold; they account for the $1-\delta$ confidence level.

Step 1: Good estimation of the losses. When the two events (14) hold, we have

$$\begin{aligned}
&\left|\ell_{t,p}-\widehat{\ell}_{t,p}\right| \\
&= \Bigg|\big(\phi(x_t,p)^{\mathrm{T}}\theta-c_t\big)^2+p^{\mathrm{T}}\Gamma p \\
&\qquad -\Big(\big[\phi(x_t,p)^{\mathrm{T}}\widehat{\theta}_{t-1}\big]_C-c_t\Big)^2+p^{\mathrm{T}}\widehat{\Gamma}_t p\Bigg| \\
&\leqslant \left|p^{\mathrm{T}}\Gamma p-p^{\mathrm{T}}\widehat{\Gamma}_t p\right| \\
&\qquad +\left|\big(\phi(x_t,p)^{\mathrm{T}}\theta-c_t\big)^2-\Big(\big[\phi(x_t,p)^{\mathrm{T}}\widehat{\theta}_{t-1}\big]_C-c_t\Big)^2\right|.
\end{aligned}$$

On the one hand, $\left|p^{\mathrm{T}}\Gamma p-p^{\mathrm{T}}\widehat{\Gamma}_t p\right| \leqslant \gamma$ while on the other hand,

$$\begin{aligned}
&\left|\big(\phi(x_t,p)^{\mathrm{T}}\theta-c_t\big)^2-\Big(\big[\phi(x_t,p)^{\mathrm{T}}\widehat{\theta}_{t-1}\big]_C-c_t\Big)^2\right| \\
&= \left|\phi(x_t,p)^{\mathrm{T}}\theta-\big[\phi(x_t,p)^{\mathrm{T}}\widehat{\theta}_{t-1}\big]_C\right| \\
&\qquad \times\left|\phi(x_t,p)^{\mathrm{T}}\theta+\big[\phi(x_t,p)^{\mathrm{T}}\widehat{\theta}_{t-1}\big]_C-2c_t\right|,
\end{aligned}$$

where by the boundedness assumptions (5), all quantities in the final inequality lie in $[0,C]$, thus

$$\left|\phi(x_t,p)^{\mathrm{T}}\theta+\big[\phi(x_t,p)^{\mathrm{T}}\widehat{\theta}_{t-1}\big]_C-2c_t\right| \leqslant 2C\,.$$

Finally,

$$\begin{aligned}
&\left|\phi(x_t,p)^{\mathrm{T}}\theta-\big[\phi(x_t,p)^{\mathrm{T}}\widehat{\theta}_{t-1}\big]_C\right| \\
&\leqslant \left|\phi(x_t,p)^{\mathrm{T}}\theta-\phi(x_t,p)^{\mathrm{T}}\widehat{\theta}_{t-1}\right| \\
&\leqslant \left\|V_{t-1}^{1/2}\big(\theta-\widehat{\theta}_{t-1}\big)\right\| \left\|V_{t-1}^{-1/2}\phi(x_t,p)\right\|, \qquad (16)
\end{aligned}$$

where we used the Cauchy-Schwarz inequality for the second inequality, and the fact that $\left|y-[x]_C\right| \leqslant |y-x|$ when $y\in[0,C]$ and $x\in\mathbb{R}$ for the first inequality. Collecting all bounds together, we proved

$$\begin{aligned}
&\left|\big(\phi(x_t,p)^{\mathrm{T}}\theta-c_t\big)^2-\Big(\big[\phi(x_t,p)^{\mathrm{T}}\widehat{\theta}_{t-1}\big]_C-c_t\Big)^2\right| \\
&\qquad \leqslant 2C\underbrace{\left\|V_{t-1}^{1/2}\big(\theta-\widehat{\theta}_{t-1}\big)\right\|}_{\leqslant B_{t-1}(\delta t^{-2})}\left\|V_{t-1}^{-1/2}\phi(x_t,p)\right\|,
\end{aligned}$$

but of course, this term is also bounded by the quantity L introduced in Section 3.5. This concludes the proof of the claimed inequality (15).

Step 2: Resulting bound on the instantaneous regrets. We denote by

$$p_t^\star \in \operatorname*{arg\,min}_{p\in\mathcal{P}}\big\{\ell_{t,p}+p^{\mathrm{T}}\Gamma p\big\} \quad (17)$$

an optimal convex vector to be used at round t. By definition (3) of the optimistic algorithm, we have that the played p_t satisfies

$$\begin{aligned}
&\widehat{\ell}_{t,p_t}-\alpha_{t,p_t} \leqslant \widehat{\ell}_{t,p_t^\star}-\alpha_{t,p_t^\star}\,, \\
\text{that is,}\qquad &\widehat{\ell}_{t,p_t}-\widehat{\ell}_{t,p_t^\star} \leqslant \alpha_{t,p_t}-\alpha_{t,p_t^\star}\,.
\end{aligned}$$

Now, for those t for which both events (14) hold, the property (15) also holds and yields, respectively for $p=p_t$ and $p=p_t^\star$:

$$\ell_{t,p_t}-\widehat{\ell}_{t,p_t} \leqslant \alpha_{t,p_t} \qquad \text{and} \qquad \widehat{\ell}_{t,p_t^\star}-\ell_{t,p_t^\star} \leqslant \alpha_{t,p_t^\star}\,.$$

Combining all these three inequalities together, we proved

$$\begin{aligned}
r_t &= \ell_{t,p_t}-\ell_{t,p_t^\star} \\
&= \big(\ell_{t,p_t}-\widehat{\ell}_{t,p_t}\big)+\big(\widehat{\ell}_{t,p_t}-\widehat{\ell}_{t,p_t^\star}\big)+\big(\widehat{\ell}_{t,p_t^\star}-\ell_{t,p_t^\star}\big) \\
&\leqslant \alpha_{t,p_t}+\big(\alpha_{t,p_t}-\alpha_{t,p_t^\star}\big)+\alpha_{t,p_t^\star} = 2\alpha_{t,p_t}\,,
\end{aligned}$$

as claimed. This yields the $2\sum\alpha_{t,p_t}$ in the regret bound, where the sum is for $t\geqslant n+1$.

Step 3: Special cases. We conclude the proof by dealing with the time steps $t\geqslant n+1$ when at least one of the events (14) does not hold. By a union bound, this happens for some $t\geqslant n+1$ with probability at most

$$\frac{\delta}{2}+\delta\sum_{t\geqslant n+1}t^{-2} \leqslant \frac{\delta}{2}+\delta\int_2^\infty\frac{1}{t^2}\,\mathrm{d}t = \delta\,.$$

These special cases thus account for the claimed $1-\delta$ confidence level.

C. Proof of Lemma 2

We derived the proof scheme below from scratch as we could find no suitable result in the literature for estimating Γ in our context.

We first consider the following auxiliary result.

Lemma 4. *Let $n \geqslant 1$. Assume that the common distribution of the $\varepsilon_1, \varepsilon_2, \dots$ is ρ–sub-Gaussian. Then, no matter how the provider picks the p_t, we have, for all $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left\| \sum_{t=1}^{n} p_t p_t^{\mathrm{T}} \big(\widehat{\Gamma}_n - \Gamma\big) p_t p_t^{\mathrm{T}} \right\|_{\infty} \leqslant \kappa_n \sqrt{n}\,,$$

where the quantities κ_n, M_n and M_n' are defined as in Lemma 2:

$$\begin{aligned} M_n &\stackrel{\text{def}}{=} \rho/2 + \ln(6n/\delta) \\ M_n' &\stackrel{\text{def}}{=} M_n^2 \sqrt{2\ln(3K^2/\delta)} + 2\sqrt{\exp(2\rho)\delta/6} \\ \kappa_n &\stackrel{\text{def}}{=} \big(C + 2M_n\big) B_n(\delta/3) + M_n' \end{aligned}$$

Proof of Lemma 4. We can show that $\widehat{\Gamma}_n$ defined in (4) satisfies

$$\sum_{t=1}^{n} p_t p_t^{\mathrm{T}} \widehat{\Gamma}_n p_t p_t^{\mathrm{T}} = \sum_{t=1}^{n} \widehat{Z}_t^2 p_t p_t^{\mathrm{T}}\,, \tag{18}$$

where we recall that $\widehat{Z}_t \stackrel{\text{def}}{=} Y_{t,p_t} - \big[\phi(x_t,p_t)^{\mathrm{T}}\widehat{\theta}_n\big]_C$. Indeed, with,

$$\Phi(\Gamma) \stackrel{\text{def}}{=} \sum_{t=1}^{n} \Big(\widehat{Z}_t^2 - p_t^{\mathrm{T}}\Gamma p_t\Big)^2 = \sum_{t=1}^{n} \Big(\widehat{Z}_t^2 - \mathrm{Tr}\big(\Gamma p_t p_t^{\mathrm{T}}\big)\Big)^2,$$

using $\nabla_A \mathrm{Tr}(AB) = B$, we get

$$\nabla_\Gamma \Phi(\Gamma) = \sum_{t=1}^{n} 2 p_t p_t^{\mathrm{T}} \Big(\widehat{Z}_t^2 - p_t^{\mathrm{T}}\Gamma p_t\Big),$$

which leads to (18) by canceling the gradient and keeping in mind that $p_t^{\mathrm{T}}\Gamma p_t$ is a scalar value.

Let us denote $Z_t \stackrel{\text{def}}{=} Y_{t,p_t} - \phi(x_t,p_t)^{\mathrm{T}}\theta$ for all $t \geqslant 1$. To prove the lemma, we replace $\widehat{\Gamma}_n$ by using (18) and apply a triangular inequality:

$$\begin{aligned} &\left\| \sum_{t=1}^{n} p_t p_t^{\mathrm{T}} \big(\widehat{\Gamma}_n - \Gamma\big) p_t p_t^{\mathrm{T}} \right\|_{\infty} \qquad (19) \\ &\leqslant \left\| \sum_{t=1}^{n} (\widehat{Z}_t^2 - Z_t^2) p_t p_t^{\mathrm{T}} \right\|_{\infty} + \left\| \sum_{t=1}^{n} Z_t^2 p_t p_t^{\mathrm{T}} - p_t p_t^{\mathrm{T}} \Gamma p_t p_t^{\mathrm{T}} \right\|_{\infty} \end{aligned}$$

We will consecutively provide bounds for each of the two terms in the right-hand side of the above inequality, each holding with probability at least $1-\delta/3$. To do so, we focus on the event defined below where all Z_t are bounded:

$$\mathcal{E}_n(\delta) \stackrel{\text{def}}{=} \big\{\forall t = 1,\dots n, \quad |Z_t| \leqslant M_n\big\}, \tag{20}$$

with M_n defined in the statement of the lemma. We will show below that $\mathcal{E}_n(\delta)$ takes place with probability at least $1-\delta/3$. All in all, our obtained global bound will hold with probability at least $1-\delta$, as stated in the lemma.

Recall that p_t is $\mathcal{F}_{t-1} = \sigma(\varepsilon_1,\dots,\varepsilon_{t-1})$ measurable. For $t \in \{1,\dots,n\}$, as ε_t is a ρ–sub-Gaussian variable independent of $\mathcal{F}_{t-1}$,

$$\mathbb{E}\Big[\exp(p_t^{\mathrm{T}}\varepsilon_t) \,\Big|\, \mathcal{F}_{t-1}\Big] \leqslant \exp\left(\frac{\rho\|p_t\|^2}{2}\right) \leqslant \exp\Big(\frac{\rho}{2}\Big).$$

Using the Markov-Chernov inequality, we obtain

$$\begin{aligned} \mathbb{P}\Big(Z_t \geqslant M_n \,\Big|\, \mathcal{F}_{t-1}\Big) &\leqslant \mathbb{E}\Big[\exp(Z_t) \,\Big|\, \mathcal{F}_{t-1}\Big]\, \exp(-M_n) \\ &\leqslant \exp\Big(\frac{\rho}{2} - M_n\Big) = \frac{\delta}{6n}. \end{aligned} \tag{21}$$

Symmetrically, we get that $\mathbb{P}(Z_t \leqslant -M_n) \leqslant \delta/6n$. Combining all these bounds for $t = 1,\dots,n$, the event $\mathcal{E}_n(\delta)$ happens with probability at least $1-\delta/3$.

Since $\phi(x_t,p_t)^{\mathrm{T}}\theta \in [0,C]$ by Assumption (5), we have

$$|\widehat{Z}_t - Z_t| = \Big|\phi(x_t,p_t)^{\mathrm{T}}\theta - \big[\phi(x_t,p_t)^{\mathrm{T}}\widehat{\theta}_n\big]_C\Big| \leqslant C\,,$$

and therefore, on $\mathcal{E}_n(\delta)$,

$$\big|\widehat{Z}_t + Z_t\big| \leqslant \big|\widehat{Z}_t - Z_t\big| + \big|Z_t\big| \leqslant C + 2M_n \stackrel{\text{def}}{=} M_n''\,.$$

Upper-bound of the first term in (19). Noting that all components of $p_t p_t^{\mathrm{T}}$ are upper-bounded by 1, we have

$$\begin{aligned} \left\| \sum_{t=1}^{n} (\widehat{Z}_t^2 - Z_t^2) p_t p_t^{\mathrm{T}} \right\|_{\infty} &\leqslant \sum_{t=1}^{n} \big|\widehat{Z}_t^2 - Z_t^2\big| \\ &= \sum_{t=1}^{n} \big|(\widehat{Z}_t - Z_t)(\widehat{Z}_t + Z_t)\big| \\ &\leqslant M_n'' \sqrt{n \sum_{t=1}^{n} (\widehat{Z}_t - Z_t)^2}\,, \end{aligned}$$

where the last inequality was obtained by $|\widehat{Z}_t + Z_t| \leqslant M_n''$ together with the Cauchy-Schwarz inequality. Using that $\big|y - [x]_C\big| \leqslant |y-x|$ when $y \in [0,C]$ and $x \in \mathbb{R}$, we note that

$$\big|\widehat{Z}_t - Z_t\big| \leqslant \Big|\phi(x_t,p_t)^{\mathrm{T}}(\widehat{\theta}_n - \theta)\Big|\,,$$

All in all, we proved so far

$$\begin{aligned}
&\Big\|\sum_{t=1}^{n}(\widehat{Z}_t^2 - Z_t^2)p_t p_t^{\mathrm{T}}\Big\|_\infty \\
&\leqslant M_n''\sqrt{n(\widehat{\theta}_n-\theta)^{\mathrm{T}}\left(\sum_{t=1}^{n}\phi(x_t,p_t)\phi(x_t,p_t)^{\mathrm{T}}\right)(\widehat{\theta}_n-\theta)} \\
&= M_n''\sqrt{n(\widehat{\theta}_n-\theta)^{\mathrm{T}}\left(V_n-\lambda I\right)(\widehat{\theta}_n-\theta)} \\
&\leqslant M_n''\sqrt{n(\widehat{\theta}_n-\theta)^{\mathrm{T}}V_n(\widehat{\theta}_n-\theta)} \\
&= M_n''\big\|V_n^{1/2}\big(\theta-\widehat{\theta}_n\big)\big\|\sqrt{n}\,,
\end{aligned}$$

where $V_n = \lambda I + \sum_{t=1}^{n}\phi(x_t,p_t)\phi(x_t,p_t)^{\mathrm{T}}$ was used for the last steps.

From Lemma 1, we finally obtain that with probability at least $1-\delta/3$,

$$\Big\|\sum_{t=1}^{n}(\widehat{Z}_t^2 - Z_t^2)p_t p_t^{\mathrm{T}}\Big\|_\infty \leqslant M_n'' B_n(\delta/3)\sqrt{n}\,. \tag{22}$$

Upper-bound of the second term in (19). Recall that p_t is $\mathcal{F}_{t-1}$ measurable. In Model 1, we have $Z_t = Y_{t,p_t} - \phi(x_t,p_t)^{\mathrm{T}}\theta = p_t^{\mathrm{T}}\varepsilon_t$. These two observations yield

$$\begin{aligned}
\mathbb{E}\big[Z_t^2 p_t p_t^{\mathrm{T}}\,\big|\,\mathcal{F}_{t-1}\big] &= \mathbb{E}\big[p_t Z_t^2 p_t^{\mathrm{T}}\,\big|\,\mathcal{F}_{t-1}\big] \\
&= \mathbb{E}\big[p_t p_t^{\mathrm{T}}\varepsilon_t\varepsilon_t^{\mathrm{T}}p_t p_t^{\mathrm{T}}\,\big|\,\mathcal{F}_{t-1}\big] \\
&= p_t p_t^{\mathrm{T}}\,\mathbb{E}\big[\varepsilon_t\varepsilon_t^{\mathrm{T}}\,\big|\,\mathcal{F}_{t-1}\big]\,p_t p_t^{\mathrm{T}} = p_t p_t^{\mathrm{T}}\Gamma p_t p_t^{\mathrm{T}}\,.
\end{aligned} \tag{23}$$

We wish to apply the Hoeffding–Azuma inequality to each component of $Z_t^2 p_t p_t^{\mathrm{T}}$, however, we need some boundedness to do so. Therefore, we consider instead $Z_t^2\mathbf{1}_{\{|Z_t|\leqslant M_n\}}$. The indicated inequality, together with a union bound, entails that with probability at least $1-\delta/3$,

$$\begin{aligned}
&\Big\|\sum_{t=1}^{n} Z_t^2\mathbf{1}_{\{|Z_t|\leqslant M_n\}}p_t p_t^{\mathrm{T}} \\
&\qquad - \sum_{t=1}^{n}\mathbb{E}\Big[Z_t^2\mathbf{1}_{\{|Z_t|\leqslant M_n\}}p_t p_t^{\mathrm{T}}\,\Big|\,\mathcal{F}_{t-1}\Big]\Big\|_\infty \\
&\leqslant M_n^2\sqrt{2n\ln(3K^2/\delta)}\,.
\end{aligned} \tag{24}$$

Over $\mathcal{E}_n(\delta)$, using (23) and applying a triangular inequality, we obtain

$$\begin{aligned}
&\Big\|\sum_{t=1}^{n} Z_t^2 p_t p_t^{\mathrm{T}} - p_t p_t^{\mathrm{T}}\Gamma p_t p_t^{\mathrm{T}}\Big\|_\infty \\
&= \Big\|\sum_{t=1}^{n} Z_t^2\mathbf{1}_{\{|Z_t|\leqslant M_n\}}p_t p_t^{\mathrm{T}} - \sum_{t=1}^{n}\mathbb{E}\big[Z_t^2 p_t p_t^{\mathrm{T}}\,\big|\,\mathcal{F}_{t-1}\big]\Big\|_\infty \\
&\leqslant \Big\|\sum_{t=1}^{n} Z_t^2\mathbf{1}_{\{|Z_t|\leqslant M_n\}}p_t p_t^{\mathrm{T}} \\
&\qquad\qquad - \sum_{t=1}^{n}\mathbb{E}\big[Z_t^2 p_t p_t^{\mathrm{T}}\mathbf{1}_{\{|Z_t|\leqslant M_n\}}\,\big|\,\mathcal{F}_{t-1}\big]\Big\|_\infty \\
&\qquad + \sum_{t=1}^{n}\Big\|\mathbb{E}\big[Z_t^2 p_t p_t^{\mathrm{T}}\mathbf{1}_{\{|Z_t|>M_n\}}\,\big|\,\mathcal{F}_{t-1}\big]\Big\|_\infty\,.
\end{aligned} \tag{25}$$

We just need to bound the last term of the inequality above to conclude this part. Using that $x^2 \leqslant \exp(x)$ for $x \geqslant 0$, we get

$$\begin{aligned}
&\mathbb{E}\Big[Z_t^2\mathbf{1}_{\{|Z_t|>M_n\}}\,\Big|\,\mathcal{F}_{t-1}\Big] \\
&\leqslant \mathbb{E}\Big[\exp(Z_t)\mathbf{1}_{\{Z_t>M_n\}}\,\Big|\,\mathcal{F}_{t-1}\Big] \\
&\qquad + \mathbb{E}\Big[\exp(-Z_t)\mathbf{1}_{\{Z_t<-M_n\}}\,\Big|\,\mathcal{F}_{t-1}\Big]\,.
\end{aligned}$$

Applying the Cauchy-Schwarz inequality yields

$$\begin{aligned}
&\mathbb{E}\Big[\exp(Z_t)\mathbf{1}_{\{Z_t>M_n\}}\,\Big|\,\mathcal{F}_{t-1}\Big] \\
&\leqslant \sqrt{\mathbb{E}\big[\exp(2Z_t)\,\big|\,\mathcal{F}_{t-1}\big]\;\mathbb{E}\big[\mathbf{1}_{\{Z_t>M_n\}}\,\big|\,\mathcal{F}_{t-1}\big]}\,.
\end{aligned}$$

Now, thanks to the sub-Gaussian property of ε_t with $\nu = p_t$, we have $\mathbb{E}\big[\exp(2Z_t)\,\big|\,\mathcal{F}_{t-1}\big] \leqslant \exp(2\rho)$. Combining with (21), we proved

$$\mathbb{E}\Big[\exp(Z_t)\mathbf{1}_{\{Z_t>M_n\}}\,\Big|\,\mathcal{F}_{t-1}\Big] \leqslant \sqrt{\exp(2\rho)\frac{\delta}{6n}}\,.$$

Symmetrically,

$$\mathbb{E}\Big[\exp(-Z_t)\mathbf{1}_{\{Z_t<-M_n\}}\,\Big|\,\mathcal{F}_{t-1}\Big] \leqslant \sqrt{\exp(2\rho)\frac{\delta}{6n}}\,.$$

Thus, we have $\mathbb{E}\big[Z_t^2\mathbf{1}_{\{|Z_t|>M_n\}}\,\big|\,\mathcal{F}_{t-1}\big] \leqslant 2\sqrt{\exp(2\rho)\delta/6n}$ and as all components of the $p_t p_t^{\mathrm{T}}$ are in $[0,1]$,

$$\Big\|\mathbb{E}\big[Z_t^2\mathbf{1}_{\{|Z_t|>M_n\}}p_t p_t^{\mathrm{T}}\,\big|\,\mathcal{F}_{t-1}\big]\Big\|_\infty \leqslant 2\sqrt{\exp(2\rho)\frac{\delta}{6n}}. \tag{26}$$

Finally , combining (25) with (24) and (26), we get with probability $1-\delta/3$

$$\begin{aligned}
&\Big\|\sum_{t=1}^{n} Z_t^2 p_t p_t^{\mathrm{T}} - p_t p_t^{\mathrm{T}}\Gamma p_t p_t^{\mathrm{T}}\Big\|_\infty \\
&\leqslant M_n^2\sqrt{2n\ln(3K^2/\delta)} + 2n\sqrt{\exp(2\rho)\delta/6n} = M_n'\sqrt{n}\,,
\end{aligned}$$

where M'_n is defined in the statement of the lemma.

Combining the two upper-bounds into (19). Combining the above upper-bound with (19) and (22), it yields with probability $1-\delta$

$$\Big\| \sum_{t=1}^{n} p_t p_t^{\mathrm{T}} \Big(\widehat{\Gamma}_n - \Gamma \Big) p_t p_t^{\mathrm{T}} \Big\|_{\infty} \leqslant M'_n \sqrt{n} + M''_n B_n(\delta/3) \sqrt{n} \,,$$

which concludes the proof. □

Conclusion of the proof of Lemma 2

Remember from Section 3.3 that all vectors $p^{(i,j)}$ are played at least $n_0 \geqslant 2n/K^2$ times in the n exploration rounds.

Proof of Lemma 2. Applying Lemma 4 together with

$$\begin{aligned} p_t p_t^{\mathrm{T}} \big(\widehat{\Gamma}_n - \Gamma \big) p_t p_t^{\mathrm{T}} &= p_t \mathrm{Tr}\Big(p_t^{\mathrm{T}} \big(\widehat{\Gamma}_n - \Gamma \big) p_t \Big) p_t^{\mathrm{T}} \\ &= \mathrm{Tr}\Big(\big(\widehat{\Gamma}_n - \Gamma \big) p_t p_t^{\mathrm{T}} \Big) p_t p_t^{\mathrm{T}} \end{aligned} \quad (27)$$

we have, with probability at least $1-\delta$, that for all pairs of coordinates $(i,j) \in E$,

$$\left| \sum_{t=1}^{n} \mathrm{Tr}\Big(\big(\widehat{\Gamma}_n - \Gamma \big) p_t p_t^{\mathrm{T}} \Big) \big[p_t p_t^{\mathrm{T}} \big]_{i,j} \right| \leqslant \kappa_n \sqrt{n} \,. \quad (28)$$

Consider first the off-diagonal elements $1 \leqslant i < j \leqslant K$ (and remember that in the set E considered in Section 3.3, we only have pairs (i,j) with $i \leqslant j$). Remark that since p_t is of the form $p^{(i',j')}$ for all $t \geqslant 1$ we have

$$\big[p_t p_t^{\mathrm{T}} \big]_{i,j} = \begin{cases} \frac{1}{4} & \text{if } p_t = p^{(i,j)} \\ 0 & \text{otherwise} \end{cases} \,. \quad (29)$$

Using that $p_t = p^{(i,j)}$ at least for n_0 rounds, Inequality (28) entails

$$\frac{n_0}{4} \left| \mathrm{Tr}\Big(\big(\widehat{\Gamma}_n - \Gamma \big) p^{(i,j)} p^{(i,j)^{\mathrm{T}}} \Big) \right| \leqslant \kappa_n \sqrt{n}$$

which using $n_0 \geqslant 2n/K^2$ becomes

$$\left| \mathrm{Tr}\Big(\big(\widehat{\Gamma}_n - \Gamma \big) p^{(i,j)} p^{(i,j)^{\mathrm{T}}} \Big) \right| \leqslant 2\kappa_n K^2 \sqrt{\frac{1}{n}} \,. \quad (30)$$

Now, let us consider the diagonal elements. Let $1 \leqslant i \leqslant K$. We have

$$\big[p_t p_t^{\mathrm{T}} \big]_{i,i} = \begin{cases} 1 & \text{if } p_t = p^{(i,i)} \\ \frac{1}{4} & \text{if } p_t = p^{(i,j)} \text{ for some } i < j \\ \frac{1}{4} & \text{if } p_t = p^{(j,i)} \text{ for some } j < i \\ 0 & \text{otherwise} \end{cases} \quad (31)$$

but note that, defining $p^{(i,j)}$ in an obvious way for all pairs (i,j), even the ones with $i \geqslant j$, we have

$$\sum_{i<j} p^{(i,j)} p^{(i,j)^{\mathrm{T}}} + \sum_{j<i} p^{(j,i)} p^{(j,i)^{\mathrm{T}}} = \sum_{j \neq i} p^{(i,j)} p^{(i,j)^{\mathrm{T}}} \,.$$

Therefore, Inequality (28) yields

$$\left| \mathrm{Tr}\left(\big(\widehat{\Gamma}_n - \Gamma \big) \Big(p^{(i,i)} p^{(i,i)^{\mathrm{T}}} + \frac{1}{4} \sum_{j \neq i} p^{(i,j)} p^{(i,j)^{\mathrm{T}}} \Big) \right) \right| \leqslant \frac{\kappa_n K^2}{2} \sqrt{\frac{1}{n}} \,. \quad (32)$$

Now, let $q \in \mathcal{P}$. Straightforward calculation shows that

$$q q^{\mathrm{T}} = \sum_{i=1}^{K} \sum_{j=1}^{K} u(i,j)\, p^{(i,j)} p^{(i,j)^{\mathrm{T}}} \quad (33)$$

with $u(i,j) = 2 q_i q_j$ if $i \neq j$ and $u(i,i) = 2q_i^2 - q_i$.

Indeed, using (29) and (31), for $1 \leqslant k \leqslant k' \leqslant K$, as $p^{(k,k')} p^{(k,k')^{\mathrm{T}}} = p^{(k',k)} p^{(k',k)^{\mathrm{T}}}$, we get

$$\begin{aligned} q_k q_{k'} = \big[q q^{\mathrm{T}} \big]_{k,k'} &= \sum_{i=1}^{K} \sum_{j=1}^{K} u(i,j) \, \big[p^{(i,j)} p^{(i,j)^{\mathrm{T}}} \big]_{k,k'} \\ &= 2u(k,k') \end{aligned}$$

Thus, we obtain $u(k,k') = 2 q_k q'_k$. Now, let us calculate the diagonal elements:

$$\begin{aligned} q_k^2 = \big[q q^{\mathrm{T}} \big]_{k,k} &= \sum_{i=1}^{K} \sum_{j=1}^{K} u(i,j) \, \big[p^{(i,j)} p^{(i,j)^{\mathrm{T}}} \big]_{k,k} \\ &= u(k,k) + \sum_{i \neq k} \frac{2u(i,k)}{4} = u(k,k) + \sum_{i \neq k} q_k q_i \\ &= u(k,k) + \sum_{i=1}^{K} q_k q_i - q_k^2 = u(k,k) + q_k - q_k^2 \,, \end{aligned}$$

which leads to $u(k,k) = 2q_k - q_k$. We can rewrite (33) to make appear the terms of (30) and (32),

$$\begin{aligned} q q^{\mathrm{T}} = \sum_{i=1}^{K} \Big(\sum_{j \neq i} \Big(u(i,j) - \frac{u(i,i)}{4} \Big) p^{(i,j)} p^{(i,j)^{\mathrm{T}}} \Big) \\ + u(i,i) \Big(p^{(i,i)} p^{(i,i)^{\mathrm{T}}} + \frac{1}{4} \sum_{j \neq i} p^{(i,j)} p^{(i,j)^{\mathrm{T}}} \Big) \end{aligned}$$

Therefore, substituting $q q^{\mathrm{T}}$ and using the linearity of the

Trace, we obtain

$$\begin{aligned}
&\Big|q^{\mathrm{T}}\big(\widehat{\Gamma}_n-\Gamma\big)q\Big| = \Big|\mathrm{Tr}\Big(\big(\widehat{\Gamma}_n-\Gamma\big)qq^{\mathrm{T}}\Big)\Big| \\
&= \Bigg|\sum_{i=1}^{K}\mathrm{Tr}\Bigg(\big(\widehat{\Gamma}_n-\Gamma\big)\Bigg(\sum_{j\neq i}\Big(u(i,j)-\frac{u(i,i)}{4}\Big)p^{(i,j)}p^{(i,j)^{\mathrm{T}}} \\
&\qquad + u(i,i)\Big(p^{(i,i)}p^{(i,i)^{\mathrm{T}}}+\frac{1}{4}\sum_{j\neq i}p^{(i,j)}p^{(i,j)^{\mathrm{T}}}\Big)\Bigg)\Bigg)\Bigg| \\
&\leqslant \sum_{i=1}^{K}\sum_{j\neq i}\Big|u(i,j)-\frac{u(i,i)}{4}\Big|\Big|\mathrm{Tr}\Big(\big(\widehat{\Gamma}_n-\Gamma\big)p^{(i,j)}p^{(i,j)^{\mathrm{T}}}\Big)\Big| \\
&+\sum_{i=1}^{K}|u(i,i)| \\
&\qquad\Big|\mathrm{Tr}\Bigg(\big(\widehat{\Gamma}_n-\Gamma\big)\Big(p^{(i,i)}p^{(i,i)^{\mathrm{T}}}+\frac{1}{4}\sum_{j\neq i}p^{(i,j)}p^{(i,j)^{\mathrm{T}}}\Big)\Bigg)\Big|.
\end{aligned}$$

Then, using the upper-bounds (30) and (32) this entails

$$\begin{aligned}
&\Big|q^{\mathrm{T}}\big(\widehat{\Sigma}_n-\Sigma\big)q\Big| \\
&\leqslant 2\kappa_n K^2\sqrt{\frac{1}{n}}\sum_{i=1}^{K}\Big(\frac{|u(i,i)|}{4}+\sum_{j\neq i}\Big(u(i,j)+\frac{|u(i,i)|}{4}\Big)\Big) \\
&= 2\kappa_n K^2\sqrt{\frac{1}{n}}\sum_{i=1}^{K}\Big(\frac{K|u(i,i)|}{4}+\sum_{j\neq i}u(i,j)\Big) \\
&= 2\kappa_n K^2\sqrt{\frac{1}{n}}\sum_{i=1}^{K}\Big(\frac{Kq_i}{4}+2q_i(1-q_i)\Big) \\
&= 2\kappa_n K^2\sqrt{\frac{1}{n}}\sum_{i=1}^{K}q_i\Big(\frac{K}{4}+2\Big) \leqslant K^2\Big(\frac{K}{2}+4\Big)\kappa_n\sqrt{\frac{1}{n}},
\end{aligned}$$

where the last two inequalities are by definition of $u(i,j) = 2q_iq_j$ and $|u(i,i)| \leqslant q_i$. This concludes the proof of Lemma 2. □

D. Proof of Lemma 3

We recall that this lemma is a straightforward adaptation/generalization of Lemma 19.1 of the monograph by Lattimore & Szepesvári (2018); see also a similar result in Lemma 3 by Chu et al. (2011).

We consider the worst case when all summations would start at $n+1=2$.

By definition, the quantity $\overline{B}$ upper bounds all the $B_{t-1}(\delta t^{-2})$. It therefore suffices to upper bound

$$\begin{aligned}
&\sum_{t=2}^{T}\min\Big\{L,\ 2C\overline{B}\,\big\|V_{t-1}^{-1/2}\phi(x_t,p_t)\big\|\Big\} \\
&\leqslant \sqrt{T}\sqrt{\sum_{t=2}^{T}\min\Big\{L^2,\ \big(2C\overline{B}\big)^2\,\big\|V_{t-1}^{-1/2}\phi(x_t,p_t)\big\|^2\Big\}} \\
&= \sqrt{T}\sqrt{\sum_{t=2}^{T}\min\left\{L^2,\ \big(2C\overline{B}\big)^2\left(\frac{\det(V_t)}{\det(V_{t-1})}-1\right)\right\}}
\end{aligned}$$

where we applied first the Cauchy-Schwarz inequality and used second the equality

$$\begin{aligned}
&1+\big\|V_{t-1}^{-1/2}\phi(x_t,p_t)\big\|^2 \\
&\qquad = 1+\phi(x_t,p_t)^{\mathrm{T}}V_{t-1}^{-1}\phi(x_t,p_t) = \frac{\det(V_t)}{\det(V_{t-1})}\,,
\end{aligned}$$

that follows from a standard result in online matrix theory, namely, Lemma 5 below.

Now, we get a telescoping sum with the logarithm function by using the inequality

$$\forall b>0,\quad \forall u>0,\qquad \min\{b,u\} \leqslant b\,\frac{\ln(1+u)}{\ln(1+b)}\,, \tag{34}$$

which is proved below. Namely, we further bound the sum above by

$$\begin{aligned}
&\sum_{t=2}^{T}\min\left\{L^2,\ \big(2C\overline{B}\big)^2\left(\frac{\det(V_t)}{\det(V_{t-1})}-1\right)\right\} \\
&\leqslant \big(2C\overline{B}\big)^2\sum_{t=2}^{T}\min\left\{\frac{L^2}{\big(2C\overline{B}\big)^2},\ \frac{\det(V_t)}{\det(V_{t-1})}-1\right\} \\
&\leqslant \big(2C\overline{B}\big)^2\sum_{t=2}^{T}\frac{L^2/\big(2C\overline{B}\big)^2}{\ln\Big(1+L^2/\big(2C\overline{B}\big)^2\Big)}\ln\left(\frac{\det(V_t)}{\det(V_{t-1})}\right) \\
&= \frac{L^2}{\ln\Big(1+L^2/\big(2C\overline{B}\big)^2\Big)}\ln\left(\frac{\det(V_T)}{\det(V_2)}\right) \\
&\leqslant \frac{L^2}{\ln\Big(1+L^2/\big(2C\overline{B}\big)^2\Big)}\,d\ln\frac{\lambda+T}{\lambda}
\end{aligned}$$

where we used (5) and one of its consequences to get the last inequality.

Finally, we use $1/\ln(1+u) \leqslant 1/u + 1/2$ for all $u \geqslant 0$ to get a more readable constant:

$$\frac{L^2}{\ln\Big(1 + L^2/\big(2C\overline{B}\big)^2\Big)} \leqslant \big(2C\overline{B}\big)^2 + \frac{L^2}{2}\,.$$

The proof is concluded by collecting all pieces. □

Finally, we now provide the proofs of two either straightforward or standard results used above.

D.1. A standard result in online matrix theory

The following result is extremely standard in online matrix theory (see, among many others, Lemma 11.11 in Cesa-Bianchi & Lugosi, 2006 or the proof of Lemma 19.1 in the monograph by Lattimore & Szepesvári, 2018).

Lemma 5. *Let M a $d \times d$ full-rank matrix, let u, $v \in \mathbb{R}^d$ be two arbitrary vectors. Then*

$$1 + v^{\mathrm{T}} M^{-1} u = \frac{\det\big(M + uv^{\mathrm{T}}\big)}{\det(M)}\,.$$

The proof first considers the case $M = \mathrm{I}_d$. We are then left with showing that $\det\big(\mathrm{I}_d + uv^{\mathrm{T}}\big) = 1 + v^{\mathrm{T}}u$, which follows from taking the determinant of every term of the equality

$$\begin{bmatrix} \mathrm{I}_d & 0 \\ v^{\mathrm{T}} & 1 \end{bmatrix} \begin{bmatrix} \mathrm{I}_d + uv^{\mathrm{T}} & u \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathrm{I}_d & 0 \\ -v^{\mathrm{T}} & 1 \end{bmatrix} = \begin{bmatrix} \mathrm{I}_d & u \\ 0 & 1 + v^{\mathrm{T}}u \end{bmatrix}\,.$$

Now, we can reduce the case of a general M to this simpler case by noting that

$$\begin{aligned} \det\big(M + uv^{\mathrm{T}}\big) &= \det(M)\ \det\Big(\mathrm{I}_d + \big(M^{-1}u\big)v^{\mathrm{T}}\Big) \\ &= \det(M)\ \big(1 + v^{\mathrm{T}}M^{-1}u\big)\,. \end{aligned}$$

D.2. Proof of Inequality (34)

This inequality is used in Lemma 19.1 of the monograph by Lattimore & Szepesvári (2018), in the special case $b = 1$. The extension to $b > 0$ is straightforward.

We fix $b > 0$. We want to prove that

$$\forall u > 0, \qquad \min\{b, u\} \leqslant b\,\frac{\ln(1+u)}{\ln(1+b)}\,. \tag{35}$$

We first note that

$$\min\{b, u\} = b\,\frac{\ln(1+u)}{\ln(1+b)} \quad \text{for } u = b$$

and that $\min\{b, u\} = b$ for $u \geqslant b$, with the right-hand side of (35) being an increasing function of u. Therefore, it suffices to prove (35) for $u \in [0, b]$, where $\min\{b, u\} = u$. Now,

$$u \longmapsto b\,\frac{\ln(1+u)}{\ln(1+b)} - u$$

is a concave and (twice) differentiable function, vanishing at $u = 0$ and $u = b$, and is therefore non-negative on $[0, b]$. This concludes the proof.

E. Proof of Theorem 2

Comment: The key observation lies in Step 1 (and is tagged as such); the rest is standard maths.

Because of the expression for the expected losses (8) and the consequence (9) of attainability, the regret can be rewritten as

$$R_T = \sum_{t=1}^{T} \ell_{t,p_t} = \sum_{t=1}^{T} \big(\phi(x_t,p_t)^{\mathrm{T}}\theta - c_t\big)^2 .$$

We first successively prove (*Step 1*) that for $t \geqslant 2$, if the bound of Lemma 1 holds, namely,

$$\left\| V_{t-1}^{1/2}\big(\theta - \widehat{\theta}_{t-1}\big) \right\| \leqslant B_{t-1}(\delta t^{-2}) , \tag{36}$$

then

$$\ell_{t,p_t} \leqslant 2\beta_{t,p_t} + 2\widetilde{\ell}_{t,p_t} , \tag{37}$$

$$\widetilde{\ell}_{t,p_t} \leqslant \beta_{t,p_t} + \widetilde{\ell}_{t,p_t^\star} - \beta_{t,p_t^\star} , \tag{38}$$

$$\widetilde{\ell}_{t,p_t^\star} \leqslant \beta_{t,p_t^\star} . \tag{39}$$

These inequalities collectively entail the bound $\ell_{t,p_t} \leqslant 4\beta_{t,p_t}$. Of course, because of the boundedness assumptions (5), we also have $\ell_{t,p_t} \leqslant C^2$. It then suffices to bound the sum (*Step 2*) of the ℓ_{t,p_t} by the sum of the $\min\big\{C^2, 4\beta_{t,p_t}\big\}$ and control for the probability of (36).

Step 1: Proof of (37)–(39). Inequality (38) holds by definition of the algorithm. For (39) and (37), we re-use the inequality (16) proved earlier: for all $p \in \mathcal{P}$,

$$\begin{aligned}
&\Big(\phi(x_t,p)^{\mathrm{T}}\big(\theta - \widehat{\theta}_{t-1}\big)\Big)^2 \\
&\leqslant \left\| V_{t-1}^{1/2}\big(\theta - \widehat{\theta}_{t-1}\big) \right\|^2 \left\| V_{t-1}^{-1/2}\phi(x_t,p) \right\|^2 && (40)\\
&\leqslant B_{t-1}(\delta t^{-2})^2 \left\| V_{t-1}^{-1/2}\phi(x_t,p) \right\|^2 \stackrel{\text{def}}{=} \beta_{t,p} , && (41)
\end{aligned}$$

where we used the bound (36) for the last inequality. This inequality directly yields (39) by taking $p = p_t^\star$.

Now comes the specific improvement and our key observation: using that $(u+v)^2 \leqslant 2u^2 + 2v^2$, we have

$$\begin{aligned}
\ell_{t,p_t} &= \Big(\phi(x_t,p_t)^{\mathrm{T}}\theta - \phi(x_t,p_t)^{\mathrm{T}}\widehat{\theta}_{t-1} \\
&\qquad + \phi(x_t,p_t)^{\mathrm{T}}\widehat{\theta}_{t-1} - c_t\Big)^2 \\
&\leqslant 2\big(\phi(x_t,p_t)^{\mathrm{T}}\theta - \phi(x_t,p_t)^{\mathrm{T}}\widehat{\theta}_{t-1}\big)^2 \\
&\qquad + 2\underbrace{\big(\phi(x_t,p_t)^{\mathrm{T}}\widehat{\theta}_{t-1} - c_t\big)^2}_{=\widetilde{\ell}_{t,p_t}} ,
\end{aligned}$$

which yields (37) via (41) used with $p = p_t$.

Step 2: Summing the bounds. First, the bound (36) holds, by Lemma 1, with probability at least $1-\delta t^{-2}$ for a given $t \geqslant 2$. By a union bound, it holds for all $t \geqslant 2$ with probability at least $1-\delta$. By bounding ℓ_{t,p_t} by C^2 and the $B_{t-1}(\delta t^{-2})$ by $\overline{B}$, we therefore get, from Step 1, that with probability at least $1-\delta$,

$$\overline{R}_T \leqslant C^2 + \sum_{t=2}^{T} \min\Big\{C^2,\ 4\overline{B}^2 \left\| V_{t-1}^{-1/2}\phi(x_t,p) \right\|^2\Big\} .$$

Now, as in the proof of Lemma 3 above (Appendix D),

$$\begin{aligned}
&\sum_{t=2}^{T} \min\Big\{C^2,\ 4\overline{B}^2 \left\| V_{t-1}^{-1/2}\phi(x_t,p) \right\|^2\Big\} \\
&= \sum_{t=2}^{T} \min\left\{C^2,\ 4\overline{B}^2\left(\frac{\det(V_T)}{\det(V_1)} - 1\right)\right\} \\
&\leqslant 4\overline{B}^2 \sum_{t=2}^{T} \frac{C^2/\big(4\overline{B}^2\big)}{\ln\Big(1 + C^2/\big(4\overline{B}^2\big)\Big)} \ln\left(\frac{\det(V_t)}{\det(V_{t-1})}\right) \\
&= \frac{C^2}{\ln\Big(1 + C^2/\big(4\overline{B}^2\big)\Big)} \ln\left(\frac{\det(V_T)}{\det(V_1)}\right) \\
&\leqslant \left(4\overline{B}^2 + \frac{C^2}{2}\right) d \ln\frac{\lambda + T}{\lambda} .
\end{aligned}$$

This concludes the proof.