Multi-Agent Learning in Contextual Games under Unknown Constraints

Anna M. Maddux
EPFL Lausanne

Maryam Kamgarpour
EPFL Lausanne

Abstract

We consider the problem of learning to play a repeated contextual game with unknown reward and unknown constraints functions. Such games arise in applications where each agent's action needs to belong to a feasible set, but the feasible set is a priori unknown. For example, in constrained multi-agent reinforcement learning, the constraints on the agents' policies are a function of the unknown dynamics and hence, are themselves unknown. Under kernel-based regularity assumptions on the unknown functions, we develop a no-regret, no-violation approach that exploits similarities among different reward and constraint outcomes. The no-violation property ensures that the time-averaged sum of constraint violations converges to zero as the game is repeated. We show that our algorithm referred to as c.z.AdaNormalGP, obtains kernel-dependent regret bounds, and the cumulative constraint violations have sublinear kernel-dependent upper bounds. In addition, we introduce the notion of constrained contextual coarse correlated equilibria (c.z.CCE) and show that ϵ-c.z.CCEs can be approached whenever players follow a no-regret no-violation strategy. Finally, we experimentally demonstrate the effectiveness of c.z.AdaNormalGP on an instance of multi-agent reinforcement learning.

Several real-world problems such as auctions, traffic routing, and multi-robot applications involve multiple self-interested agents that repeatedly interact with each other. Such problems can be described as *repeated games* (Cesa-Bianchi and Lugosi, 2006). In most cases, the underlying game is unknown to the agents, consider, for example, the outcome of an auction or travel times in a transportation network. Thus, a significant body of research has focused on deriving algorithms that allow agents to learn a strategy that is aligned with their goals.

The goal of a given learning agent in a repeated game is to maximize her cumulative reward. However, without making any assumptions about other agents' learning approach, such a goal is difficult to attain. In this setting, a non-trivial yet attainable performance measure is no-regret. A learning agent aims to learn a no-regret strategy by repeatedly playing an action and observing the corresponding reward. Existing no-regret algorithms for repeated games (Littlestone and Warmuth, 1994; Freund and Schapire, 1995) are based on the assumption that agents can play any action.

In many practical scenarios, however, agents' actions are subject to a variety of constraints, and a priori the set of feasible actions may not be known. These constraints can represent various requirements, such as safety regulations (Sun et al., 2017; Usmanova et al., 2020), fairness considerations, or task-specific limitations (Guo et al., 2022). For example, in thermal control of buildings, the occupants' comfort temperature and energy cost are functions of the temperature controller parameters and unknown if the dynamics of the controller are unknown. The problem of no-regret learning under unknown constraints has been gaining increasing attention (see related work below). However, this problem has not been sufficiently explored in a game setting.

In this paper, we consider the class of *contextual games with unknown constraints. Repeated contextual games* (Sessa et al., 2020) generalize the class of static games, where agents additionally observe contextual information at each round. In routing games, for instance, agents may observe the network's capacity or weather conditions. Leveraging the context information, the agents can learn a no-regret policy, which maps the observed context to their actions which results in a stronger notion of regret. In constrained games, how-

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

ever, the agents need to operate within the boundaries of their unknown constraints. We thus address learning a no-regret no-violation strategy for a constrained contextual game, where no-violation implies a sublinear bound on the cumulative constraint violations.

Related work. No-regret learning under unknown constraints has been extensively studied by various research fields including online convex optimization (Guo et al., 2022; Yi et al., 2021), black-box optimization (Xu et al., 2023), reinforcement learning (RL) (Wei et al., 2022), and multi-agent reinforcement learning (MARL) (Chen et al., 2022). In safety-critical applications, a line of research (Sui et al., 2015; Turchetta et al., 2019) has focused on developing methods, where no constraint violations are tolerated. These methods are restricted to a single-player setting and require knowing an initially feasible point. For many real-world problems, however, it may be sufficient to restrict the amount of constraint violations. In designing a controller to regulate the temperature in a building (Di Natale et al., 2022), for instance, deviating from the occupants' comfort temperature is tolerable but should be avoided. Several works (Sun et al., 2017; Yu and Neely, 2020; Wei et al., 2022; Zhou and Ji, 2022) consider the so-called soft constraints, in which the goal is to limit the sum of constraint function values over the rounds. Thus, violations at one round can be compensated in a different round. In our paper, we consider a stronger notion of constraint satisfaction over rounds. In particular, we aim to limit the cumulative constraint *violations* rather than values (Yi et al., 2021; Guo et al., 2022; Xu et al., 2023). Perhaps closest to ours is the setup of constrained online optimization (Yi et al., 2021; Guo et al., 2022), where the unknown time-varying loss and constraint functions are convex. In our setup of repeated games with unknown constraints, the reward of each agent is static and the time-variations are due to the dependence of the reward function on the actions of other agents. Our goal is to leverage this structure of the game to achieve improved no-regret no-violations, in the absence of convexity and with the inclusion of contextual information.

Incorporating contextual information into no-regret learning has been extensively studied in the bandit literature (Slivkins, 2014) for reward functions that depend linearly on context features (Li et al., 2010; Chu et al., 2011; Lin et al., 2022) and more generally for reward functions that are an arbitrary linear function of the contexts' images in some corresponding reproducing kernel Hilbert space (RKHS) (Krause and Ong, 2011; Valko et al., 2013). The RKHS assumption has been considered in contextual games (Sessa et al., 2020) and contextual no-regret algorithms have been derived for this class of games. While Sessa et al. (2020) leveraged the game structure to improve the regret rate compared to an online learning scenario, it did not incorporate unknown constraints and thus, did not come with a provable violation bound.

Contributions.

We propose a novel no-regret, no-violation algorithm c.z.AdaNormalGP for playing unknown contextual games with constraints. For the handling of constraints in no-regret learning, we establish a connection between playing a (contextual) game with unknown constraints and the sleeping expert problem (Blum, 1995; Freund et al., 1997). The sleeping expert problem was initially formulated to address dynamically available action sets rather than constraints on the action sets. While our feasible action set is static, the challenge in our setting is that this set is a priori unknown. To derive a no-regret algorithm with bounded rate of constraint violation, we exploit the assumption that similar contexts and agents' actions lead to similar reward and constraint values and similarly to Sessa et al. (2020) and Xu et al. (2023) use the RKHS framework to learn the reward and constraint functions online based on past game data. In particular, we view an action as an asleep expert if it is infeasible with respect to the estimate of the constraint function at a given round.

We establish a regret rate and a cumulative constraint violation rate for our algorithms c.z.AdaNormalGP. Concretely, c.z.AdaNormalGP with K actions obtains the following bounds after T rounds:

- For finite context spaces, c.z.AdaNormalGP achieves $\mathcal{O}(\sqrt{|\mathcal{Z}|T\log K} + \gamma_0^T\sqrt{T})$ regret and $\mathcal{O}(\gamma_m^T\sqrt{T})$ cumulative constraint violation for each $m \in [M]$.

- For infinite context spaces, c.z.AdaNormalGP achieves $\mathcal{O}(T^{\frac{d+1}{d+2}}\sqrt{\log(K)} + \gamma_0^T\sqrt{T})$ regret and $\mathcal{O}(\gamma_m^T\sqrt{T})$ cumulative constraint violation for each $m \in [M]$.

Our regret rate is the same, up to logarithmic factors in T and K, as those for unconstrained contextual games. Our violation rate matches that of Xu et al. (2023), which only considers a static online optimization problem without any context. We further define a new notion of constrained contextual Coarse Correlated Equilibria (c.z.CCE) for constrained contextual games and show that c.z.CCEs can be approached whenever agents follow a no-regret, no-violation algorithm such as c.z.AdaNormalGP.

1 Problem Setup

We consider a repeated contextual game among N agents or players, where the two terms are used interchangeably. At every round, each player selects a feasible action and receives a payoff which depends on the action profile chosen by all players and the *context* at that round. More formally, let $\mathcal{A}_i$ denote the action set of player i and let $\mathcal{Z}$ represent the set of possible contexts. We assume $\mathcal{A}_i^f(z) := \{a_i \in \mathcal{A}_i \mid g_{i,m}(a_i, z) \leq 0, \quad \forall m \in [M]\}$ to be the context-dependent feasible action set, where $z \in \mathcal{Z}$ and $\{g_{i,m} : \mathcal{A}_i \times \mathcal{Z} \to \mathbb{R}\}_{m\in[M]}$ is a set containing M constraint functions. Furthermore, define $r_i : \mathcal{A} \times \mathcal{Z} \to [0, 1]$ to be the reward function of each player i, where $\mathcal{A} := \mathcal{A}_1 \times \ldots \times \mathcal{A}_N$ is the joint action space. We assume that each agent's reward function r_i and constraint functions $g_{i,m}$, $m \in [M]$, are unknown to her, and thus also her context-dependent feasible action set $\mathcal{A}_i^f(z)$ for all $z \in \mathcal{Z}$. Then, a repeated contextual game with unknown constraints proceeds as follows. At every round t, context z^t is revealed. The players observe z^t and, based on it, simultaneously each player $i \in [N]$ selects an action $a_i^t \in \mathcal{A}_i$ which needs to be feasible, i.e., $a_i^t \in \mathcal{A}_i^f(z^t)$. Then, players obtain rewards $r_i(a_i^t, a_{-i}^t, z^t)$ and constraints $g_{i,m}(a_i^t, z^t)$, where $m \in [M]$ and $a_{-i}^t := (a_1^t, \ldots, a_{i-1}^t, a_{i+1}^t, \ldots, a_N^t)$.

We define Π_i to be the set of all policies $\pi_i : \mathcal{Z} \to \mathcal{A}_i$ mapping contexts to actions. After T rounds, the performance of player i is measured by the *constrained contextual regret*:

$$\begin{aligned} R_i^T = \max_{\pi_i \in \Pi_i} \sum_{t=1}^{T} r_i(\pi_i(z^t), a_{-i}^t, z^t) - \sum_{t=1}^{T} r_i(a_i^t, a_{-i}^t, z^t) \\ \text{s.t.} \quad g_{i,m}(\pi_i(z^t), z^t) \leq 0, \quad \forall m \in [M], \forall t \geq 1. \end{aligned} \tag{1}$$

We define the cumulative constraint violation for each constraint $g_{i,m}$ with $m \in [M]$ as:

$$\mathcal{V}_{i,m}^T = \sum_{t=1}^{T} [g_{i,m}(a_i^t, z^t)]_+, \tag{2}$$

where $[x]_+ := \max\{0, x\}$. The constrained contextual regret, short regret, measures the gain player i could have achieved by following her best feasible fixed policy in hindsight had the sequence of other players' actions $\{a_{-i}\}_{t=1}^T$, the reward $r_i(\cdot)$ and the constraint $g_{i,m}(\cdot)$ functions been known to her. The cumulative constraint violation measures the accumulated violations over all rounds T for each constraint function. A strategy is *no-regret* for player i if $R_i^T/T \to 0$ as $T \to \infty$. Similarly, we call a strategy *no-violation* for player i if $\mathcal{V}_{i,m}^T/T \to 0$ for all $m \in [M]$ as $T \to \infty$.

Repeated contextual games with unknown constraints generalize the class of repeated static games in two ways: Firstly, as discussed in Sessa et al. (2020) and Valko et al. (2013), the game may change from round to round due to a potentially different context z^t. Secondly and so far unaddressed by the literature, the action set $\mathcal{A}_i$ of each player may be subject to unknown (context-dependent) constraints. This makes this generalization challenging as a priori a player does not know whether an action is feasible or not.

The problem we address in this paper is the design of an algorithm that is simultaneously *no-regret* and *no-violation*. To ensure that the learning problem is meaningful, we assume that a feasible policy exists for each player.

Assumption 1 (Feasibility assumption)**.** *We assume that for each agent $i \in \mathcal{N}$ the following problem*

$$\begin{aligned} \max_{\pi_i \in \Pi_i} \quad & \sum_{t=1}^{T} r_i(\pi_i(z^t), a_{-i}^t, z^t) \\ \text{s.t.} \quad & g_{i,m}(\pi_i(z^t), z^t) \leq 0, \quad \forall m \in [M], \forall t \geq 1 \end{aligned} \tag{3}$$

is feasible with optimal constrained solution $\overline{\pi}_i^*$.

Observe that this is a natural assumption for requiring no-cumulative violation. In particular, in a non-contextual game setting this assumption is equivalent to assuming that a feasible action exists.

1.1 Feedback model and regularity assumption

At the end of each round, player i receives feedback information that she can use to update her strategy and thereby improve her performance in terms of regret and constraint violations. In a full-information feedback model player i observes her reward for any action $a_i \in \mathcal{A}_i$ given the other players' actions a_{-i}^t and the observed context z^t, i.e., $\mathbf{r}_i^t = [r_i^t(a_i, a_{-i}^t, z^t)]_{a_i \in \mathcal{A}_i}$. A more realistic feedback model is bandit feedback, where player i observes her reward $r_i^t(a_i^t, a_{-i}^t, z^t)$ for the played action profile a^t. In the constrained setting, we extend the term bandit feedback such that after each round player i additionally observes her constraint values $g_{i,m}(a_i^t, z^t)$, for all $m \in [M]$, for her played action a_i^t. Additionally, in many practical settings player i can observe the other players' played actions a_{-i}^t or some aggregative function $\gamma(a_{-i}^t)$ thereof, e.g. in routing games, agents may observe the total occupancy of each edge in the road network, or in electricity markets, agents may observe the aggregate load.

Assumption 2 (Feedback assumption)**.** *We consider a noisy bandit feedback model, where, at every round t, player i observes a noisy measurement of the reward $\tilde{r}_i^t = r_i(a_i^t, a_{-i}^t, z^t) + \epsilon_i^t$ and a noisy measurement of each constraint $\tilde{g}_{i,m}^t = g_{i,m}(a_i^t) + \epsilon_{i,m}^t$. The noise ϵ_i^t is $\sigma_{i,0}$-sub-Gaussian and $\epsilon_{i,m}^t$ is $\sigma_{i,m}$-sub-Gaussian for*

each $m \in [M]$. Furthermore, we assume player i also observes the played actions a^t_{-i} of the other players.

Attaining *no-regret* and *no-violation* is impossible without any regularity assumptions on the reward and constraint functions (Srinivas et al., 2010). In many practical settings, including routing games and thermal control, similar inputs lead to similar rewards and constraint values which implies some regularity in the reward and constraint functions. In the application of thermal control, for instance, similar choices of control input parameters and set point temperatures result in similar energy consumption and room temperature.

Assumption 3 (Regularity assumption)**.** *We assume the action set $\mathcal{A}_i$ is finite with $|\mathcal{A}_i| = K$ and consider a general context set $\mathcal{Z} \subseteq \mathbb{R}^d$. Let $\mathcal{X} = \mathcal{A} \times \mathcal{Z}$. We assume the unknown reward function $r_i : \mathcal{A} \times \mathcal{Z} \to [0, 1]$ has a bounded norm $\|r_i\|_{k_{i,0}} = \sqrt{\langle r_i, r_i \rangle_{k_{i,0}}} \leq B_{i,0}$ in a reproducing kernel Hilbert space (RKHS, see (Rasmussen and Williams, 2005)) associated with a positive semi-definite kernel function $k_{i,0}(\cdot, \cdot)$. The RKHS is denoted by $\mathcal{H}_{k_{i,0}}(\mathcal{X})$. We further assume bounded variance by restricting $k_{i,0}(x, x') \leq 1$ for all $x, x' \in \mathcal{X}$. Let $\mathcal{X}_i = \mathcal{A}_i \times \mathcal{Z}$. Similarly, we assume each unknown constraint function $g_{i,m} : \mathcal{A}_i \times \mathcal{Z} \to [0, 1]$, $m \in [M]$, has a bounded norm $\|g_{i,m}\|_{k_{i,m}} \leq B_{i,m}$ in a RKHS associated with a positive semi-definite kernel function $k_{i,m}(\cdot, \cdot)$. The RKHS is denoted by $\mathcal{H}_{k_{i,m}}(\mathcal{X}_i)$ and we assume that $k_{i,m}(x_i, x'_i) \leq 1$ for all $x_i, x'_i \in \mathcal{X}_i$.*

For general RKHS $\mathcal{H}_k(\mathcal{X})$ with positive semi-definite kernel function $k(\cdot, \cdot)$, the RKHS norm $\|f\|_k$ measures smoothness[1] of $f \in \mathcal{H}_k(\mathcal{X})$ with respect to the kernel function $k(\cdot, \cdot)$, while the kernel $k(\cdot, \cdot)$ encodes similarities between different points $x, x' \in \mathcal{X}$. Assuming that some unknown function has a bounded norm in an RKHS is standard in black-box optimization (Srinivas et al., 2010; Chowdhury and Gopalan, 2017) and was recently also exploited in repeated (contextual) games (Sessa et al., 2019, 2020). It is analogous to assuming bounded weights in linear parametric bandits (Chu et al., 2011; Abbasi-Yadkori et al., 2011). The bound of 1 on $k(x, x')$ is only required to obtain scale-free bounds. Otherwise, if $k(x, x') \leq n$ for all $x, x' \in \mathcal{X}$ or $\|f\|_k \leq nB$, our bounds increase by a factor of n (Agrawal and Goyal, 2013). Typical kernel choices are the *Squared Exponential*, *Matérn* and the *polynomial* kernel:

$$k_{SE}(x, x') = \exp\left(-\frac{s^2}{2l^2}\right),$$
$$k_{\text{Matérn}}(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{s\sqrt{2\nu}}{l}\right)^{\nu} B_\nu\left(\frac{s\sqrt{2\nu}}{l}\right)^{\nu},$$
$$k_{\text{poly}}(x, x') = \left(b + \frac{x^T x'}{l}\right)^d,$$

where $l, \nu, b > 0$ are kernel hyperparameters, $s = \|x - x'\|_2$ encodes similarities between points $x, x' \in \mathcal{X}$, and $B_\nu(\cdot)$ is the modified Bessel function. We emphasize that Assumption 2 and 3 allow player i to learn about her unknown reward r_i and constraint $g_{i,m}$ functions.

The Gaussian process (GP) framework can be used to learn unknown functions $f \in \mathcal{H}_k(\mathcal{X})$ with $\|f\|_k < \infty$ since such functions can be modeled as a sample from a GP (Rasmussen and Williams, 2005, Section 6.2). A GP over $\mathcal{X}$ is a probability distribution over functions $f(x) \sim \mathcal{GP}(\mu(x), k(x, x'))$, specified by its mean and covariance functions $\mu(\cdot)$ and $k(\cdot, \cdot)$, respectively. A GP prior $\mathcal{GP}(0, k(\cdot, \cdot))$ over the initial distribution of the unknown function f, where $k(\cdot, \cdot)$ is the kernel function associated with the RKHS $\mathcal{H}_k(\mathcal{X})$, is used to capture the uncertainty over f. Then, the GP framework can be used to predict function values $f(x)$ for any points $x \in \mathcal{X}$ based on a history of measurements $\{y^\tau\}_{\tau=1}^t$ at points $\{x^\tau\}_{\tau=1}^t$ with $y^\tau = f(x^\tau) + \epsilon^\tau$ and $\epsilon^\tau \sim \mathcal{N}(0, \sigma^2)$. Conditioned on the history of measurements, the posterior distribution over f is a GP with mean and variance functions:

$$\mu^t(x) = \mathbf{k}^t(x)^\top (\mathbf{K}^t + \sigma^2 \mathbf{I}^t)^{-1} \mathbf{y}^t \tag{4}$$
$$(\sigma^t)^2(x) = k(x, x) - \mathbf{k}^t(x)^\top (\mathbf{K}^t + \sigma^2 \mathbf{I})^{-1} \mathbf{k}^t(x), \tag{5}$$

where $\mathbf{k}^t(x) = [k(x^\tau, x)]_{\tau=1}^t$, $\mathbf{y}^t = [y^\tau]_{\tau=1}^t$, and $\mathbf{K}^t = [k(x^\tau, x^{\tau'})]_{\tau,\tau'=1}^t$ is the kernel matrix.

In the next section, we present our algorithm which exploits the feedback model and the regularity assumption to derive a no-regret no-violation algorithm. We focus on the perspective of a single player and drop the subscript i wherever this is possible without causing confusion.

2 The c.z.AdaNormalGP Algorithm

In our approach, we build upper confidence bounds on the reward function and lower confidence bounds on the constraint functions using the GP framework proposed in the previous section. We thereby obtain optimistic estimates of the reward and the constraints which enable us to derive bounds for the regret and cumulative constraint violations. Particularly, for the

[1] This can be seen from $|f(x) - f(x')| = |\langle f, k(x, \cdot) - k(x', \cdot)\rangle| \leq \|f\|_k \|k(x, \cdot) - k(x', \cdot)\|_k$ for any $f \in \mathcal{H}_k(\mathcal{X})$, using the reproducing property and Cauchy-Schwarz inequality.

Algorithm 1 c.z.AdaNormalGP algorithm

1: **for** $t = 1, \dots, T$ **do**
2: `Observe context` z^t.
3: **if** $\exists m \in [M]$ such that $\min_{\pi_i \in \Pi_i} \mathrm{LCB}_m^t(\pi_i(z^t), z^t) > 0$ **then**
4: `Declare infeasibility.`
5: **end if**
6: `Compute distribution` $p^t(z^t)$ `using` z^t `and`
7: $[a_i^\tau, a_{-i}^\tau, z^\tau, \tilde{r}^\tau]_{\tau=1}^{t-1}$
8: `Sample` $a_i^t \sim \overline{p}^t(z^t)$, `where`

$$\overline{p}_{a_i}^t(z^t) = \begin{cases} \frac{p_{a_i}^t(z^t)}{\sum_{\overline{a}_i \in \mathcal{A}_i} \overline{p}_{\overline{a}_i}^t(z^t)}, & \text{if } \mathrm{LCB}_m^t(a_i, z^t) \leq 0 \ \forall m \in [M], \\ 0, & \text{otherwise.} \end{cases}$$

9: `Observe noisy reward` $\tilde{r}^t$ `and constraints`
10: $\tilde{g}_m^t$ `and action profile` a_{-i}^t.
11: `Append` $(a_i^t, a_{-i}^t, \tilde{r}^t, z^t)$ `and` $(a_i^t, \tilde{g}_m^t, z^t)$ `to the`
12: `history of play.`
13: `Update` μ_0^t, σ_0^t `and` μ_m^t, σ_m^t `via` (4)-(5).
14: **end for**

constraints, we use lower confidence bounds to ensure that even in early rounds when we have poor estimates of the constraints, we can find feasible actions. Considering optimistic estimates of the constraints is not limiting since as the amount of past game data increases over the rounds uncertainty about the unknown constraints decreases and thus the confidence bounds shrink towards the true constraint functions

Making use of past game data $[(a_i^\tau, a_{-i}^\tau, \tilde{r}^\tau, \tilde{g}_m^\tau)]_{\tau=1}^{t-1}$, at round t, the posterior mean $\mu_m^t(\cdot)$ and posterior variance $(\sigma_m^t)^2(\cdot)$, for $m \in \{0\} \cup [M]$, can be used to obtain an upper and a lower confidence bound on r and g_m, respectively:

$$\mathrm{UCB}_m^t(x) := \mu_m^{t-1}(x) + \beta_m^t \sigma_m^{t-1}(x), \quad \forall x \in \mathcal{X} \quad (6)$$

$$\mathrm{LCB}_m^t(x) := \mu_m^{t-1}(x) - \beta_m^t \sigma_m^{t-1}(x), \quad \forall x \in \mathcal{X}, \quad (7)$$

where $\mathcal{X} = \mathcal{A} \times \mathcal{Z}$ for the reward function r and $\mathcal{X}_i = \mathcal{A}_i \times \mathcal{Z}$ for each constraint function g_m. Parameter β_m^t controls the width of the confidence bound. Importantly, if β_m^t is specified adequately, the reward r and the constraint functions g_m are bounded from below and above by the lower LCB_m^t and the upper UCB_m^t confidence bound with high probability. For the reward function, for example, if β_0^t is specified as $\beta_0^t = B_0 + \sigma_0 \sqrt{2(\gamma_0^{t-1} + 1 + \log(1/\delta))}$, then $\mathrm{LCB}_0^t(x) \leq r(x) \leq \mathrm{UCB}_0^t(x)$, for all $x \in \mathcal{X}$ and $t \geq 1$, with probability at least $1 - \delta$ (Chowdhury and Gopalan, 2017). Here, γ_0^t, the *maximum information gain* (Srinivas et al., 2010), is a kernel-dependent quantity which for the reward r is defined as:

$$\gamma_0^t := \max_{A \subset \mathcal{X} : |A| = t} I(y_A; r_A).$$

Quantity $I(y_A, r_A)$ denotes the mutual information between the true reward values $\mathbf{r}_A = [r(x)]_{x \in A}$ and the noisy reward measurements $\mathbf{y}_A = [r(x) + \epsilon]_{x \in A}$. It quantifies the maximal reduction in uncertainty about r after observing points $A \subset \mathcal{X}$ and their corresponding noisy measurements $\mathbf{y}_A$. The maximum information gain γ_m^t of each constraint function g_m is defined analogously.

We now introduce constrained contextual AdaNormalGP (c.z.AdaNormalGP) algorithm for repeated contextual games with a priori unknown constraints on the action set of each player. At each round, c.z.AdaNormalGP observes context z^t and checks whether a feasible action exists with respect to any LCB_m^t estimate of constraint g_m, since the constraints g_m are unknown. The algorithm then computes a context-dependent distribution $p^t(z^t) \in \Delta_K$ over the action set $\mathcal{A}_i$ based on the UCB_0^t estimate of the reward r and the observed context z^t. Here Δ_K denotes the K-dimensional probability simplex. The distribution is renormalized to assign zero weight to an action $a_i \in \mathcal{A}_i$ which is infeasible with respect to any LCB_m^t estimate, $m \in [M]$. A feasible action a_i^t is sampled from the renormalized distribution. Once the computation rule of $p^t(z^t)$ (line 6 of Algorithm 1) is specified, c.z.AdaNormalGP is well-defined.

Consider a fixed context $z \in \mathcal{Z}$. We establish that playing a repeated game with unknown constraints is equivalent to playing a sleeping expert problem (Blum, 1995; Freund et al., 1997). In the sleeping expert problem, at each round, each expert (action a_i) reports whether it is awake (feasible) or asleep (infeasible). Since the constraint functions g_m are unknown, c.z.AdaNormalGP maintains an optimistic estimate of the set of feasible actions at each round, defined as the set of actions $\mathcal{A}_i$ for which, for all $m \in [M]$, $\mathrm{LCB}_m^t(a_i, z^t) \leq 0$. Thus, in a repeated game with unknown constraints, the set of awake experts at round t corresponds to the set of feasible actions with respect to the LCB_m^t estimate of each constraint g_m for all $m \in [M]$. Leveraging this connection, we specify a computation rule of $p^t(z^t)$ based on the update rule of the sleeping expert algorithm AdaNormalHedge (Luo and Schapire, 2015). In the following, we specify $p^t(z^t)$ when the context space $\mathcal{Z}$ is finite (Algorithm 2). In Appendix A.2.2 we specify a computation rule $p^t(z^t)$ for an infinite context space and provide bounds on the regret and the cumulative constraint violations for this setting. (Algorithm 3).

Remark 1. *The above-established connection implies that for a fixed context $z \in \mathcal{Z}$, any update rule that corresponds to an update rule of a sleeping expert problem can be used to obtain high probability regret bounds. Moreover, since any sleeping expert problem can be re-*

duced to a regular expert problem (Freund et al., 1997), any update rule that corresponds to the update rule of the reduced expert problem can be used to obtain high probability regret bounds for repeated games with unknown constraints. These bounds depend on the regret obtained by the expert algorithm. We formally introduce and study general regret bounds in Appendix A.2 for the case of finite and infinite context spaces.

2.1 Finite number of contexts

When the context set $\mathcal{Z}$ is finite, on a high level the distribution $p^t(z^t) \in \Delta_K$ (line 6 of Algorithm 1) is computed by maintaining a distribution for each context $z \in \mathcal{Z}$ which is updated whenever z is observed. Thus, playing a repeated contextual game reduces to playing $|\mathcal{Z}|$ repeated static games. For a specific context $z \in \mathcal{Z}$ the update rule is defined as the general update rule of AdaNormalHedge (Luo and Schapire, 2015). Under this computation rule of $p^t(z^t)$, summarized in Algorithm 2, c.z.AdaNormalGP achieves the following high-probability bounds.

Theorem 1. *Fix $\delta \in (0,1)$. Under Assumptions 1-3, if a player plays according to c.z.AdaNormalGP with $p^t(z^t)$ computed according to Algorithm 2 and $\beta_m^t = B_m + \sigma_m\sqrt{2(\gamma_m^{t-1} + 1 + \log(2(M+1)/\delta))}$ for all $m \in \{0\} \cup [M]$, then with probability at least $1-\delta$:*

$$R^T = \mathcal{O}\big(\sqrt{|\mathcal{Z}|T(\log(K) + \log(B) + \log(1+\log(K))} + \sqrt{T\log(2/\delta)} + \beta_0^T\sqrt{T\gamma_0^T}\big) \tag{8}$$

$$\mathcal{V}_m^T = \mathcal{O}\big(\beta_m^T\sqrt{T\gamma_m^T}\big), \quad \forall m \in [M], \tag{9}$$

where $B = 1 + \frac{3}{2}\frac{1}{K}\sum_{a_i=1}^{K}(1+\log(1+C_i^t(a_i))) \leq \frac{5}{2} + \frac{3}{2}\log(1+T)$.

We provide a detailed proof in Appendix A.1.1.

Corollary 1. *Under the same assumptions as in Theorem 1, if a player plays according to c.z.AdaNormalGP with $|\mathcal{Z}| = 1$, then with probability $1-\delta$ the following regret[2] bound for repeated static games with unknown constraints holds:*

$$R^T = \mathcal{O}\big(\sqrt{T(\log(K) + \log(B) + \log(1+\log(K))} + \sqrt{T\log(2/\delta)} + \beta_0^T\sqrt{T\gamma_0^T}\big).$$

The high probability bound on the cumulative constraint violations is the same as in Theorem 1.

In repeated contextual games, c.GPMW (Sessa et al., 2020) algorithm achieves $\mathcal{O}(\sqrt{|\mathcal{Z}|T\log K} + \sqrt{T\log(2/\delta)} + \beta_0^T\sqrt{T\gamma_0^T})$ which is the same as our regret bound up to logarithmic terms. Following the same arguments as in the proof of Theorem 1, AdaNormalHedge can be extended to the contextual setting which implies a $\mathcal{O}(\sqrt{|\mathcal{Z}|T(\log(K) + \log(B) + \log(1+\log K))} + \sqrt{T\log(2/\delta)})$ bound on the contextual sleeping expert regret. Note that, unlike AdaNormalHedge, c.z.AdaNormalGP does not consider full-information, but obtains a regret bound that is equal up to the term $\mathcal{O}(\beta_0^T\sqrt{T\gamma_0^T})$. Furthermore, both c.GPMW and AdaNormalHedge do not provide bounds on the cumulative constraint violations. In constrained black-box optimization, CONFIG (Xu et al., 2023) algorithm obtains a bound on the cumulative constraint violations which is the same as ours.

2.2 Game equilibria

In the previous section, we focused on the perspective of a single player and did not make any assumptions about how the other players select their actions. It is known that if all players follow a no-regret dynamics a coarse correlated equilibrium (CCE) can be approached (Hart and Mas-Colell, 2000). A CCE of a game is a probability distribution $\rho \in \Delta^{|\mathcal{A}|}$ such that

$$\mathbb{E}_{a\sim\rho}\big[r_i(a)\big] \geq \mathbb{E}_{a\sim\rho}\big[r_i(a_i', a_{-i})\big], \quad \forall i \in \mathcal{N}, a_i' \in \mathcal{A}_i.$$

In other words, if an action profile is sampled from a CCE, then in expectation each player is better off following the sampled action than playing any other action. To deal with repeated contextual games, Sessa et al. (2020) introduced the notion of contextual CCEs (z.CCE) and extended the results of Hart and Mas-Colell (2000) to this setting. In the following, we introduce the notion of constrained contextual CCEs (c.z.CCE) to address the fact that only a subset of a player's action set is feasible.

Definition 1. *Let $z^1, \dots, z^T$ be the revealed sequence of contexts. A constrained contextual coarse correlated equilibrium is a joint policy $\rho : \mathcal{Z} \to \Delta^{|\mathcal{A}|}$ such that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{a\sim\rho_{z^t}}\big[[g_{i,m}(a_i, z^t)]_+\big] \leq 0$$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{a\sim\rho_{z^t}}\big[r_i(a, z^t)\big] \geq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{a\sim\rho_{z^t}}\big[r_i(\pi_i(z^t), a_{-i}, z^t)\big],$$

for all $i \in \mathcal{N}$, $m \in [M]$ and for all feasible $\pi_i \in \Pi_i$, i.e., $g_{i,m}(\pi_i(z^t), z^t) \leq 0$ for all $t = [T]$. An ϵ-c.z.CCE is a joint policy $\rho : \mathcal{Z} \to \Delta^{|\mathcal{A}|}$ such that the above two inequalities are satisfied up to $\epsilon > 0$ accuracy.

The above definition requires that for a fixed context an action profile sampled from a c.z.CCE is feasible in expectation. To illustrate, suppose ρ is a c.z.CCE and

[2] The static games setup is recovered by assuming $z^t = z^0$ for all t and the regret is defined as $R^T = \max_{a_i\in\mathcal{A}_i}\sum_{t=1}^{T} r_i(a_i, a_{-i}^t) - \sum_{t=1}^{T} r_i(a_i^t, a_{-i}^t)$ subject to $g_{i,m}(a_i, z^t) \leq 0, \quad \forall m \in [M], \forall t \geq 1$.

Algorithm 2 Update rule for finite context space $\mathcal{Z}$

Set $R^1_{a_i}(z) = 0$, $C^1_{a_i}(z) = 0$ and let $p^1(z)$ be the uniform distribution $\forall z \in \mathcal{Z}$.

for $t = 2, \dots, T$ **do**

  `Compute reward estimate` $\hat{r}^t$ `with` $\hat{r}^t(a_i) = \min\{1, \texttt{UCB}^t_0(a_i, a^t_{-i}, z^t)\}, \quad \forall a_i \in \mathcal{A}_i$.

  `Set for` z^t `and every` $a_i \in \mathcal{A}_i$ `with` $\mathbb{E}[\hat{r}^t(a^t_i)] = \overline{p}^t(z^t)^\top \hat{r}^t$:

$$R^t_{a_i}(z^t) = \begin{cases} R^{t-1}_{a_i}(z^t) + (\hat{r}^t(a_i) - \mathbb{E}[\hat{r}^t(a^t_i)]), & \texttt{if LCB}^t_m(a_i, z^t) \leq 0, \ \forall m \in [M], \\ R^{t-1}_{a_i}(z^t), & \texttt{otherwise}. \end{cases} \tag{10}$$

$$C^t_{a_i}(z^t) = \begin{cases} C^{t-1}_{a_i}(z^t) + |\hat{r}^t(a_i) - \mathbb{E}[\hat{r}^t(a^t_i)]|, & \texttt{if LCB}^t_m(a_i, z^t) \leq 0, \ \forall m \in [M], \\ C^{t-1}_{a_i}(z^t), & \texttt{otherwise}. \end{cases} \tag{11}$$

  `Update` $p^{t+1}_{a_i}(z^t) = \frac{w(R^t_{a_i}(z^t), C^t_{a_i}(z^t))}{\sum_{\overline{a}_i \in \mathcal{A}_i} w(R^t_{\overline{a}_i}(z^t), C^t_{\overline{a}_i}(z^t))}$, `where` $w(R, C) = \frac{1}{2}(\exp([R+1]^2_+/3(C+1)) - \exp([R-1]^2_+/3(C+1)))$.

end for

suppose there is a trusted device that, for any context z^t, samples a joint action from $\rho(z^t)$. When complying with such a device in expectation each player's action is feasible and each player is better off than when using any other feasible policy π_i.

Next, we show that a c.z.CCE can be approached when all players minimize their constrained contextual regret. We first define the empirical joint policy ρ^T at round T as

$$\rho^T_z(a) = \begin{cases} \frac{1}{T_z}|\{t \in [T] : z^t = z, a^t = a\}|, & \text{if } z \in \mathcal{Z}^T \\ \frac{1}{|\mathcal{A}|}, & \text{ow.}, \end{cases} \tag{12}$$

where $T_z = |\{t \in [T] : z^t = z\}|$ is the number of rounds that context z was revealed and $\mathcal{Z}^T \subseteq \mathcal{Z}$ is the set of revealed contexts. Note that for unobserved contexts ρ^T_z can be any arbitrary distribution.

Proposition 1. *After T rounds, the empirical joint policy ρ^T is an ϵ-c.z.CCE of the played constrained contextual game with $\epsilon \leq \max\{\epsilon_1, \epsilon_2\}$, where $\epsilon_1 = \max_{i \in \mathcal{N}} R^T_i/T$ and $\epsilon_2 = \max_{i \in \mathcal{N}, m \in [M]} \mathcal{V}^T_{i,m}/T$.*

We provide a detailed proof in Appendix A.2.3. Proposition 1 implies that if all players follow no-regret, no-violation dynamics then the empirical joint policy ρ^T converges to a *c.z.CCE* of the constrained contextual game, as $T \to \infty$.

3 Experiments

Here, we focus on the application of our approach to a realistic multi-building temperature control design problem. In Appendix A.3.1 we benchmark our algorithm against existing approaches for randomly generated contextual N-player games.

3.1 Temperature controller design

Optimizing a building's energy consumption is of prime importance, both for ecological and economical reasons, as nearly 40% of the total energy is consumed by the building sector (Laustsen, 2008) and the price of electricity has been surging over the last few years. Several recent works have looked at this problem from a single agent perspective, using Bayesian optimization (Fiducioso et al., 2019), control or reinforcement learning (Di Natale et al., 2022). These approaches do not account for the dependence of the electricity price on the aggregate consumption of energy of multiple buildings. Our work particularly accounts for this dependence, generalizing the single-player approaches to a multi-player game setting.

We tune temperature controllers for multiple buildings using c.AdaNormalGP[3] with the goal of reducing the buildings' energy cost while satisfying operational constraints, such as pleasant room temperature.

We abstract this problem as follows: We consider three buildings $i \in \{1, 2, 3\}$ with a single thermal zone which we simulate using **Energym** (Scharnhorst et al., 2021) and **Modelica** (Fritzson and Engelson, 1998). Each building is equipped with a heat pump which, similarly to Xu et al. (2023), is regulated by a proportional controller (Ang et al., 2005). The decision variable of each building corresponds to the parameters $a_i = (K_i, \text{SP}_i, \text{ST}_i)$ of the building's proportional controller, where K_i is the controller gain, SP_i is the set point temperature during the day, and ST_i the switching time from day- to nighttime set point temperature. Fixing the selected control parameters a_i, we simulate the energy consumption and temperature of the building based on the Energym and Modelica mod-

[3] We do not consider contexts in this application and thus refer to Algorithm 1 as c.AdaNormalGP.

els for a fixed period. After this period, the building observes its electricity cost $J_i(a_i, a_{-i})$ which is an aggregate function of the total energy consumption, and observes the maximum temperature deviation $g_i(a_i)$ from the setpoint temperature. Based on this feedback each building aims to tune its control parameters to reduce its energy cost while at the same time ensuring that the temperature stays within some occupant-specific comfort range.

In our experiments[4], we discretize the space of control parameters and let each building select its control parameters a_i according to c.AdaNormalGP. We simulate the energy consumption and temperature for two days. Similarly to Hall et al. (2022), we model the cost function as an affine function of the aggregate consumption:

$$J_i(a_i, a_{-i}) = \sum_{h=1}^{48} \left(c_1 \sum_{i=1}^{3} l_i^h(a_i) + c_2 \right) l_i^h(a_i),$$

where $l_i^h(a_i)$ denotes the energy consumption of building i at hour h. The positive constants c_1 and c_2 represent different price rates which depend on whether the energy is consumed during peak or off-peak hours. We further require that the maximum temperature deviation from the setpoint temperature during the night remains below an occupant-specific threshold:

$$\max_{h \in 1,\dots,48} g_i^h(a_i) - g_{i,\text{thr}} \leq 0.$$

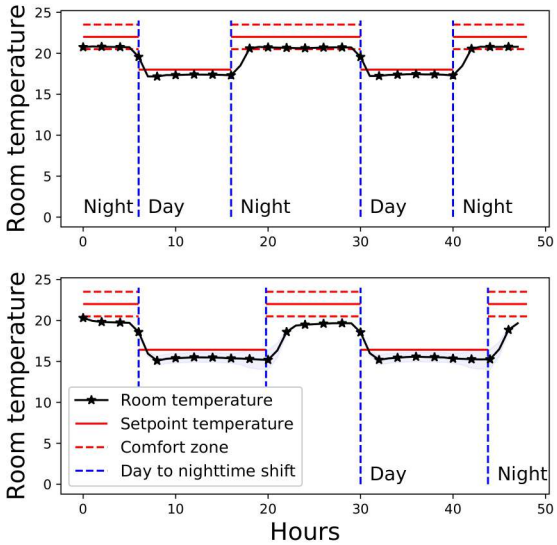

Figure 1: Mean temperature over 48 hours, where the control inputs are sampled from the weights learned by c.AdaNormalGP (*top*) and GPMW (*bottom*).

[4]The code is made available at Github.

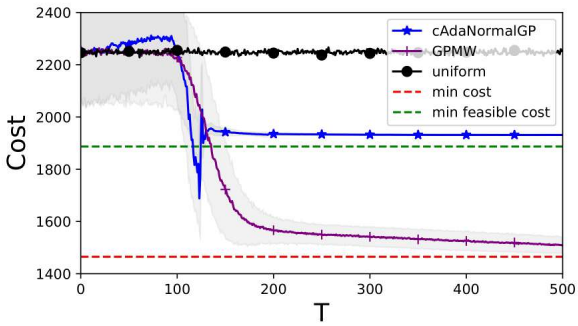

Figure 2: Mean energy cost achieved by c.AdaNormalGP, GPMW, and uniformly at random sampled control inputs for each round $t = 1, \dots T$. The minimum feasible- and the minimum cost are found exhaustively over the entire action space.

We compare the performance of c.AdaNormalGP with GPMW (Sessa et al., 2019), which neglects constraints on the maximum temperature deviation from the occupant's comfort range. To run c.AdaNormalGP and GPMW, we use the polynomial kernel for the cost function and the squared exponential kernel for the constraint function and run both algorithms for $T = 500$ rounds. For example, for building 1, Figure 1 shows the mean room temperature over 48 hours resulting from 1000 control inputs a_1 that are sampled according to c.AdaNormalGP (*top*) and GPMW (*bottom*). Following c.AdaNormalGP the mean temperature stays within the occupant's comfort range during the nighttime whereas under GPMW the deviation of the mean temperature exceeds the threshold and thus is below the occupant's comfort range. Furthermore, following c.AdaNormalGP and GPMW each building learns to select control parameters that, given the other buildings' energy demand, reduce the building's energy cost compared to the case where the control parameters are selected uniformly at random. This is shown in Figure 2 for building 1. Although the energy cost is lower when following GPMW than when following c.AdaNormalGP such a cost is not attainable when also respecting the constraints on the occupant's comfort temperature range. In conclusion, following c.AdaNormalGP each building not only learns to select feasible control parameters but also attains nearly the minimal cost achievable under the temperature constraints.

4 Conclusion and further discussion

We considered the problem of repeated contextual games under unknown constraints. Establishing a connection between this problem and the sleeping expert problem, we proposed the constrained contextual

c.z.AdaNormalGP algorithm, a no-regret no-violation algorithm for playing such games with finite and infinite context spaces. Our algorithm attains high probability regret and cumulative constraint violation bounds, where the former is comparable to bounds obtained in the unconstrained setting. We further showed that if all players follow no-regret, no-violation dynamics then the empirical joint policy converges to a *c.z.CCE* of the constrained contextual game as the game is repeated infinitely often. We demonstrated the effectiveness of c.z.AdaNormalGP on a multi-building temperature control design problem. Future research directions include learning in repeated games with coupling constraints.

Acknowledgements

This research is gratefully supported by the Swiss National Science Foundation under the NCCR Automation Grant.

References

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

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*. PMLR.

Ang, K. H., Chong, G., and Li, Y. (2005). Pid control system analysis, design, and technology. *IEEE transactions on control systems technology*.

Blum, A. (1995). Empirical support for winnow and weighted-majority based algorithms: results on a calendar scheduling domain. In *Machine Learning Proceedings 1995*.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press, USA.

Chen, Z., Ma, S., and Zhou, Y. (2022). Finding correlated equilibrium of constrained markov game: A primal-dual approach. In *Advances in Neural Information Processing Systems*.

Chowdhury, S. R. and Gopalan, A. (2017). On kernelized multi-armed bandits. In *International Conference on Machine Learning*. PMLR.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. PMLR.

Clarkson, K. L. (2005). Nearest-neighbor searching and metric space dimensions.

Di Natale, L., Lian, Y., Maddalena, E. T., Shi, J., and Jones, C. N. (2022). Lessons learned from data-driven building control experiments: Contrasting gaussian process-based mpc, bilevel deepc, and deep reinforcement learning. In *2022 IEEE 61st Conference on Decision and Control (CDC)*.

Fiducioso, M., Curi, S., Schumacher, B., Gwerder, M., and Krause, A. (2019). Safe contextual bayesian optimization for sustainable room temperature pid control tuning. *arXiv preprint arXiv:1906.12086*.

Freund, Y. and Schapire, R. E. (1995). A decision-theoretic generalization of online learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*. Springer-Verlag.

Freund, Y., Schapire, R. E., Singer, Y., and Warmuth, M. K. (1997). Using and combining predictors that specialize. In *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing*. Association for Computing Machinery.

Fritzson, P. and Engelson, V. (1998). Modelica—a unified object-oriented language for system modeling and simulation. In *ECOOP'98—Object-Oriented Programming: 12th European Conference Brussels, Belgium, July 20–24, 1998 Proceedings 12*. Springer.

Guo, H., Liu, X., Wei, H., and Ying, L. (2022). Online convex optimization with hard constraints: Towards the best of two worlds and beyond. In *Advances in Neural Information Processing Systems*.

Hall, S., Belgioioso, G., Liao-McPherson, D., and Dorfler, F. (2022). Receding horizon games with coupling constraints for demand-side management. In *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE.

Hart, S. and Mas-Colell, A. (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica*.

Hazan, E. and Megiddo, N. (2007). Online learning with prior knowledge. In *Proceedings of the 20th Annual Conference on Learning Theory*, COLT'07. Springer-Verlag.

Krause, A. and Ong, C. (2011). Contextual gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*.

Krauthgamer, R. and Lee, J. R. (2004). Navigating nets: Simple algorithms for proximity search. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics.

Laustsen, J. (2008). Energy efficiency requirements in building codes, energy efficiency policies for new buildings. iea information paper.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*. Association for Computing Machinery.

Lin, Y., Wang, Y., and Zhou, E. (2022). Risk-averse contextual multi-armed bandit problem with linear payoffs. *Journal of Systems Science and Systems Engineering*.

Littlestone, N. and Warmuth, M. (1994). The weighted majority algorithm. *Information and Computation*.

Luo, H. and Schapire, R. E. (2014). A drifting-games analysis for online learning and applications to boosting. In *Advances in Neural Information Processing Systems*.

Luo, H. and Schapire, R. E. (2015). Achieving all with no parameters: AdaNormalHedge. In *Proceedings of The 28th Conference on Learning Theory*. PMLR.

Mourtada, J. and Gaïffas, S. (2019). On the optimality of the hedge algorithm in the stochastic regime. *J. Mach. Learn. Res.*

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Scharnhorst, P., Schubnel, B., Fernández Bandera, C., Salom, J., Taddeo, P., Boegli, M., Gorecki, T., Stauffer, Y., Peppas, A., and Politi, C. (2021). Energym: A building model library for controller benchmarking. *Applied Sciences*.

Sessa, P. G., Bogunovic, I., Kamgarpour, M., and Krause, A. (2019). No-regret learning in unknown games with correlated payoffs. *Advances in Neural Information Processing Systems*.

Sessa, P. G., Bogunovic, I., Krause, A., and Kamgarpour, M. (2020). Contextual games: Multi-agent learning with side information. In *Advances in Neural Information Processing Systems*.

Slivkins, A. (2014). Contextual bandits with similarity information. *Journal of Machine Learning Research*.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *In Proceedings International Conference on Machine Learning (ICML)*.

Sui, Y., Gotovos, A., Burdick, J., and Krause, A. (2015). Safe exploration for optimization with gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning*. PMLR.

Sun, W., Dey, D., and Kapoor, A. (2017). Safety-aware algorithms for adversarial contextual bandit. In *Proceedings of the 34th International Conference on Machine Learning*. PMLR.

Turchetta, M., Berkenkamp, F., and Krause, A. (2019). Safe exploration for interactive machine learning. In *Advances in Neural Information Processing Systems*.

Usmanova, I., Krause, A., and Kamgarpour, M. (2020). Safe non-smooth black-box optimization with application to policy search. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, Proceedings of Machine Learning Research. PMLR.

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. (2013). Finite-time analysis of kernelised contextual bandits. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

Wei, H., Liu, X., and Ying, L. (2022). A provably-efficient model-free algorithm for infinite-horizon average-reward constrained markov decision processes. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xu, W., Jiang, Y., Svetozarevic, B., and Jones, C. (2023). Constrained efficient global optimization of expensive black-box functions. In *International Conference on Machine Learning*. PMLR.

Yi, X., Li, X., Yang, T., Xie, L., Chai, T., and Johansson, K. H. (2021). Regret and cumulative constraint violation analysis for online convex optimization with long term constraints.

Yu, H. and Neely, M. J. (2020). A low complexity algorithm with $\mathcal{O}(\sqrt{T})$ regret and $\mathcal{O}(1)$ constraint violations for online convex optimization with long term constraints. *Journal of Machine Learning Research*.

Zhou, X. and Ji, B. (2022). On kernelized multi-armed bandits with constraints. In *Advances in Neural Information Processing Systems*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes.]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes.]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. []

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes.]
   (b) Complete proofs of all theoretical results. [Yes.]
   (c) Clear explanations of any assumptions. [Yes.]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes, in the final version, if accepted, we will provide a URL to the code and data.]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes, in the URL we include the code for tuning the hyperparameters of the kernels.]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes.]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No.]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes, we reference Energym and Modelica.]
   (b) The license information of the assets, if applicable. [Not Applicable.]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable.]
   (d) Information about consent from data providers/curators. [Not Applicable.]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable.]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable.]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable.]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable.]

A Appendix

A.1 Supplementary Material for Section 2

For proving Theorems 1, 3, 4 and 5 we make use of the following lemma and theorem.

Lemma 1. *(Chowdhury and Gopalan, 2017, Theorem 2) Let $\mathcal{H}_k$ be the RKHS of real-valued functions on $\mathcal{X} \subset \mathbb{R}^d$ with underlying kernel function k. Consider an unknown function $f : \mathcal{X} \to \mathbb{R}$ in $\mathcal{H}_k$ such that $\|f\|_k \leq D$, and the sampling model $y^t = f(x^t) + \epsilon^t$, where ϵ^t is σ-sub-Gaussian (with independence between times). By setting*

$$\beta^t = D + \sigma\sqrt{2(\gamma^{t-1} + 1 + \log(1/\delta))}$$

the following holds with probability at least $1 - \delta$:

$$|\mu^{t-1}(x) - f(x)| \leq \beta^t \sigma^{t-1}(x), \quad \forall x \in \mathcal{X},\ \forall t \geq 1,$$

where $\mu^{t-1}(\cdot)$ and $\sigma^{t-1}(\cdot)$ are given in (4)-(5). Here, γ^{t-1} denotes the maximum information gain, a kernel-dependent quantity defined in Section 2.

Theorem 2. *(Luo and Schapire, 2015, Theorem 3) For the sleeping expert problem, the regret of AdaNormalHedge[5] is bounded as follows:*

$$\begin{aligned} R^T(a_i) = \sum_{t=1}^{T} 1_{a_i,t}(\mathbb{E}[l^t] - l^t(a_i)) &\leq \sqrt{3C_{a_i}^T(\log(K) + \log(B) + \log(1 + \log(K)))}, \\ &\leq \sqrt{3T(\log(K) + \log(B) + \log(1 + \log(K)))} \end{aligned} \tag{13}$$

where $l^t(a_i)$ is the loss of action (expert) $a_i \in \mathcal{A}_i$, $l^t = [l^t(a_i)]_{a_i \in \mathcal{A}_i} \in [0,1]^K$ is the loss vector of all actions (experts), $C_{a_i}^T = \sum_{t=1}^T |\mathbb{E}[l^t(a_i^t)] - l^t(a_i)|$ and $B = 1 + \frac{3}{2}\frac{1}{K}\sum_{a_i \in \mathcal{A}_i}(1 + \log(1 + C_{a_i}^T)) \leq \frac{5}{2} + \frac{3}{2}\log(1 + T)$. Here, $1_{a_i,t}$ denotes the indicator function defined as:

$$1_{a_i,t} = \begin{cases} 1, & \text{if action (expert) } a_i \text{ is available (awake) at round } t, \\ 0, & \text{otherwise.} \end{cases}$$

A.1.1 Finite number of contexts: Proof of Theorem 1

In this section, we provide a proof for Theorem 1. By making use of Lemma 1, we first show that upper and lower confidence bounds hold simultaneously for the reward and all constraint functions. Next, we show that c.z.AdaNormalGP does not declare infeasibility with high probability if the underlying game is feasible. We then prove an upper bound on the cumulative constraint violations making use of the previous two results and techniques from GP optimization Srinivas et al. (2010) to account for not knowing the true constraint functions. To obtain an upper bound on the regret we leverage that for each context $z \in \mathcal{Z}$ a separate distribution is maintained and analyze the regret obtained by each context $z \in \mathcal{Z}$. Fixing a context $z \in \mathcal{Z}$, the proof further follows by decomposing the regret into the sum of two terms. The first term corresponds to the regret that a player incurs with respect to the optimistic reward estimates, i.e., the upper confidence bounds, when the set of feasible actions is a priori unknown. This term can be upper bounded by Theorem 2 thanks to the connection between constrained games and the sleeping expert problem. The second term stems from not knowing the true reward function and can be upper bounded by using techniques from GP optimization.

Lemma 2. *With probability $1 - \frac{\delta}{2}$ the following holds simultaneously:*

$$\begin{aligned} &LCB_0^t(a,z) \leq r(a,z) \leq \min\{1, UCB_0^t(a,z)\}, \quad \forall a \in \mathcal{A}, \forall z \in \mathcal{Z}, \forall t \geq 1, \\ &LCB_m^t(a_i,z) \leq g_m(a_i,z) \leq UCB_m^t(a_i,z), \quad \forall a_i \in \mathcal{A}_i, \forall z \in \mathcal{Z}, \forall t \geq 1, \forall m \in [M], \end{aligned} \tag{14}$$

with $\beta_m^t = B_m + \sigma\sqrt{2(\gamma_m^{t-1} + 1 + \log(2(M+1)/\delta))}$ for $m \in \{0\} \cup [M]$.

[5] We consider AdaNormalHedge (Algorithm 1 in Luo and Schapire (2015)), where the $p_{a_i}^t$ is predicted via equation (3) in Luo and Schapire (2015).

Proof. By Lemma 1 and since the true unknown reward function r lies in $[0,1]$, with probability at least $1-\frac{\delta}{2(M+1)}$ the reward function can be upper and lower bounded as follows:

$$\mathrm{LCB}_0^t(a,z) \leq r(a,z) \leq \min\{1, \mathrm{UCB}_0^t(a,z)\}, \quad \forall a \in \mathcal{A},\ \forall z \in \mathcal{Z},\ \forall t \geq 1, \tag{15}$$

where the upper- and lower confidence bounds $\mathrm{UCB}_0^t(a,z)$ and $\mathrm{LCB}_0^t(a,z)$ are defined as in (6)-(7) and $\beta_0^t = B_0 + \sigma\sqrt{2(\gamma_0^{t-1} + 1 + \log(2(M+1)/\delta))}$. Analogously, with probability at least $1-\frac{\delta}{2(M+1)}$ each unknown constraint function g_m, $m \in [M]$, can be upper and lower bounded as follows:

$$\mathrm{LCB}_m^t(a_i,z) \leq g_m(a_i,z) \leq \mathrm{UCB}_m^t(a_i,z), \quad \forall a_i \in \mathcal{A}_i,\ \forall z \in \mathcal{Z},\ \forall t \geq 1, \tag{16}$$

where $\mathrm{UCB}_m^t(a_i,z)$ and $\mathrm{LCB}_m^t(a_i,z)$ are defined as in (6)-(7) and $\beta_m^t = B_m + \sigma\sqrt{2(\gamma_m^{t-1} + 1 + \log(2(M+1)/\delta))}$.

Note that the confidence bounds $\mathrm{LCB}_0^t(a,z)$, $\mathrm{UCB}_0^t(a,z)$, $\mathrm{LCB}_m^t(a_i,z)$ and $\mathrm{UCB}_m^t(a_i,z)$, $m \in [M]$, are random variables since they depend on the random sampling process of the algorithm (line 7 of Algorithm 1) and on noisy observations of the reward $\tilde{r}^t$ and constraints $\tilde{g}_m^t$, $m \in [M]$. Define $E_0 := \cap_{a\in\mathcal{A}} \cap_{z\in\mathcal{Z}} \cap_{t\in[T]} \{\mathrm{LCB}_0^t(a,z) \leq r(a,z) \leq \min\{1, \mathrm{UCB}_0^t(a,z)\}\}$ and define $E_m := \cap_{a_i\in\mathcal{A}_i} \cap_{z\in\mathcal{Z}} \cap_{t\in[T]} \{\mathrm{LCB}_m^t(a_i,z) \leq g_m(a_i,z) \leq \mathrm{UCB}_m^t(a_i,z)\}$ for all $m \in [M]$. Then, using De Morgan's law, the union bound and (15)-(16) we have

$$\begin{aligned}
\mathbb{P}(\cap_{m=0}^M E_m) &= 1 - \mathbb{P}((\cap_{m=0}^M E_m)^C) = 1 - \mathbb{P}(\cup_{m=0}^M E_m^C) \\
&\geq 1 - \sum_{m=0}^M \mathbb{P}(E_m^C) \\
&\geq 1 - \sum_{m=0}^M \frac{\delta}{2(M+1)} = 1 - \frac{\delta}{2}.
\end{aligned}$$

□

Lemma 3. *With probability $1-\frac{\delta}{2}$, c.z.AdaNormalGP (line 3 of Algorithm 1) does not declare infeasibility.*

Proof. By Assumption 1 the problem given in (3) is feasible and has optimal solution $\overline{\pi}_i^*$, i.e.,

$$\begin{aligned}
\overline{\pi}_i^* \in \arg\max_{\pi_i\in\Pi_i} &\sum_{t=1}^T r(\pi_i(z^t), a_{-i}^t, z^t) \\
\text{s.t. } & g_m(\pi_i(z^t), z^t) \leq 0, \quad \forall m \in [M], \forall t \geq 1.
\end{aligned}$$

By Lemma 2, with probability $1-\frac{\delta}{2}$, $\mathrm{LCB}_m^t(a_i,z) \leq g_m(a_i,z)$ for all $a_i \in \mathcal{A}_i$, $z \in \mathcal{Z}$, $m \in [M]$, and $t \geq 1$. Thus, by setting $a_i = \overline{\pi}_i^*(z^t)$ with probability $1-\frac{\delta}{2}$ we have that $\mathrm{LCB}_m^t(\overline{\pi}_i^*(z^t), z^t) \leq 0$ for all $m \in [M]$ and $t \geq 1$ and infeasibility is not declared. □

We proceed with upper-bounding the cumulative constraint violations for each constraint.

Lemma 4. *With probability at least $1-\frac{\delta}{2}$ the cumulative constraint violations $\mathcal{V}_m^T$ for each constraint g_m, $m \in [M]$, are upper-bounded as follows:*

$$\mathcal{V}_m^T = \mathcal{O}\left(\beta_m^T \sqrt{T\gamma_m^T}\right), \quad \forall m \in [M].$$

Proof. With probability at least $1-\frac{\delta}{2}$ the following holds:

$$\begin{aligned}
\mathcal{V}_m^T &= \sum_{t=1}^{T}[g_m(a_i^t, z^t)]_+ && (17)\\
&= \sum_{t=1}^{T}[g_m(a_i^t, z^t) - \mathrm{LCB}_m^t(a_i^t, z^t) + \mathrm{LCB}_m^t(a_i^t, z^t)]_+ && (18)\\
&\leq \sum_{t=1}^{T}[g_m(a_i^t, z^t) - \mathrm{LCB}_m^t(a_i^t, z^t)]_+ + \sum_{t=1}^{T}[\mathrm{LCB}_m^t(a_i^t, z^t)]_+ && (19)\\
&\leq \sum_{t=1}^{T}[\mathrm{UCB}_m^t(a_i^t, z^t) - \mathrm{LCB}_m^t(a_i^t, z^t)]_+ + \sum_{t=1}^{T}[\underbrace{\mathrm{LCB}_m^t(a_i^t, z^t)}_{\leq 0}]_+ && (20)\\
&\leq \sum_{t=1}^{T}[\mathrm{UCB}_m^t(a_i^t, z^t) - \mathrm{LCB}_m^t(a_i^t, z^t)]_+ && (21)\\
&\leq 2\beta_m^T \sum_{t=1}^{T}\sigma_m^{t-1}(a_i^t, z^t) && (22)\\
&\leq C_1\beta_m^T\sqrt{T\gamma_m^T}, && (23)
\end{aligned}$$

where equation (20) follows from Lemma 2 and holds with probability at least $1-\frac{\delta}{2}$. Equation (21) follows from the fact that the sampled action a_i^t at each round t is feasible with respect to the lower confidence bound $\mathrm{LCB}_m^t(\cdot)$ for all $m \in [M]$ (line 7 of Algorithm 1) and that with probability at least $1-\frac{\delta}{2}$ infeasibility is not declared. Equation (22) follows firstly from the definition of $\mathrm{UCB}_m^t(\cdot)$ in (6) and $\mathrm{UCB}_m^t(\cdot)$ in (7) and the fact that β_m^t is increasing in t.The last equation makes use of (Srinivas et al., 2010, Lemma 5.4) which provides an upper bound on the sum of posterior standard deviations, where $C_1 = \frac{8}{\log(1+\sigma^{-2})}$. □

Next, we proceed with upper-bounding the constrained contextual regret, short regret.

Lemma 5. *Conditioned on equation* (14) *in Lemma 2 holding true, the constrained contextual regret* R^T *is upper-bounded with probability* $1-\frac{\delta}{2}$ *by:*

$$R^T = \mathcal{O}\left(\sqrt{|\mathcal{Z}|T(\log(K)+\log(B)+\log(1+\log(K))} + \sqrt{T\log(2/\delta)} + \beta_0^T\sqrt{T\gamma_0^T}\right).$$

Proof. Recall the definition of regret:

$$\begin{aligned}
R^T &= \max_{\pi_i\in\Pi_i}\sum_{t=1}^{T} r(\pi_i(z^t), a_{-i}^t, z^t) - \sum_{t=1}^{T} r(a_i^t, a_{-i}^t, z^t)\\
&\quad \text{s.t. } g_m(\pi_i(z^t), z^t) \leq 0, \quad \forall m \in [M], \forall t \geq 1,\\
&= \sum_{t=1}^{T} r(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t) - \sum_{t=1}^{T} r(a_i^t, a_{-i}^t, z^t).
\end{aligned}$$

Note that for a finite context space $\mathcal{Z}$, c.z.AdaNormalGP maintains a separate distribution $p^t(z)$ for each context $z \in \mathcal{Z}$, where $p^t(z^t)$ is computed according to Algorithm 2. Thus, the regret notion can be rewritten as:

$$R^T = \sum_{z\in\mathcal{Z}}\sum_{t:z^t=z} r(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t) - r(a_i^t, a_{-i}^t, z^t).$$

The reward and thus R^T can be further upper bounded as follows:

$$\begin{aligned}
R^T &= \sum_{z\in\mathcal{Z}}\sum_{t:z^t=z} r(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t) - r(a_i^t, a_{-i}^t, z^t) \\
&\le \sum_{z\in\mathcal{Z}}\sum_{t:z^t=z} \min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t)\} - (\mathrm{LCB}_0^t(a_i^t, a_{-i}^t, z^t)) \qquad (24)\\
&\le \sum_{z\in\mathcal{Z}}\sum_{t:z^t=z} \min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t)\} - (\mathrm{UCB}_0^t(a_i^t, a_{-i}^t, z^t) - 2\beta_0^t\sigma_0^{t-1}(a_i^t, a_{-i}^t, z^t)) \\
&\le \sum_{z\in\mathcal{Z}}\sum_{t:z^t=z} \min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t)\} - \min\{1, \mathrm{UCB}_0^t(a_i^t, a_{-i}^t, z^t)\} \\
&\quad + \sum_{t=1}^T 2\beta_0^t\sigma_0^{t-1}(a_i^t, a_{-i}^t, z^t) \\
&\le \sum_{z\in\mathcal{Z}}\sum_{t:z^t=z} (\min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t)\} - \min\{1, \mathrm{UCB}_0^t(a_i^t, a_{-i}^t, z^t)\}) + C_1\beta_0^T\sqrt{T\gamma_0^T}, \qquad (25)
\end{aligned}$$

where in equation (24) we used equation (14) from Lemma 2 and in equation (25) we again used (Srinivas et al., 2010, Lemma 5.4) with $C_1 = \frac{8}{\log(1+\sigma^{-2})}$.

To further upper-bound R^T we rely on the Hoeffding-Azuma inequality (Cesa-Bianchi and Lugosi, 2006, Lemma A.7). Note that the collection of random variables $\{V^t\}_{t=1}^T$ defined as $V^t := \min\{1, \mathrm{UCB}_0^t(a_i^t, a_{-i}^t, z^t)\} - \sum_{a_i\in\mathcal{A}_i}\overline{p}_{a_i}^t(z^t)\min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\}$ forms a martingale difference sequence[6]. Then, the Hoeffding-Azuma inequality is applicable and the following holds with probability at least $1-\delta/2$:

$$\sum_{z\in\mathcal{Z}}\sum_{t:z^t=z} \Big|\min\{1, \mathrm{UCB}_0^t(a_i^t, a_{-i}^t, z^t)\} - \sum_{a_i\in\mathcal{A}_i}\overline{p}_{a_i}^t(z^t)\min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\}\Big| \le \sqrt{T/2\log(2/\delta)}. \qquad (26)$$

Here, we used that $V^t \in [0,1]$ and $\sum_{z\in\mathcal{Z}}\sum_{t:z^t=z} 1 = T$. Plugging equation (26) into equation (25), with probability at least $1-\delta/2$, we obtain the following upper bound on R^T:

$$\begin{aligned}
R^T \le &\sum_{z\in\mathcal{Z}}\sum_{t:z^t=z} \Big(\min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t)\} - \sum_{a_i\in\mathcal{A}_i}\overline{p}_{a_i}^t(z^t)\min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\}\Big) \\
&+ \sqrt{T/2\log(2/\delta)} + C_1\beta_0^T\sqrt{T\gamma_0^T}.
\end{aligned} \qquad (27)$$

Define the function $f^t(\cdot) = \min\{1, \mathrm{UCB}_0^t(\cdot, a_{-i}^t, z^t)\}$. Since we conditioned on equation (14) in Lemma 2 holding true, $\mathrm{UCB}_0^t(\cdot) \ge 0$ since $r(\cdot) \ge 0$ and thus $f^t(\cdot) \in [0,1]^K$. For a fixed $z \in \mathcal{Z}$, the first term in equation (27) can be rewritten as:

$$\begin{aligned}
&\sum_{t:z^t=z} \min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t)\} - \sum_{a_i\in\mathcal{A}_i}\overline{p}_{a_i}^t(z^t)\min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\} \\
&= \sum_{t:z^t=z} f^t(\overline{\pi}_i^*(z^t)) - \sum_{a_i\in\mathcal{A}_i}\overline{p}_{a_i}^t(z^t) f^t(a_i^t) \qquad (28)\\
&= \sum_{t:z^t=z} 1_{\overline{\pi}_i^*(z^t),t}\Big(f^t(\overline{\pi}_i^*(z^t)) - \sum_{a_i\in\mathcal{A}_i}\overline{p}_{a_i}^t(z^t) f^t(a_i^t)\Big), \qquad (29)
\end{aligned}$$

where for all $a_i \in \mathcal{A}_i$ we define:

$$1_{a_i,t} := \begin{cases} 1, & \text{if } \mathrm{LCB}_m^t(a_i, z^t) \le 0 \text{ for every } m \in [M], \\ 0, & \text{otherwise.} \end{cases} \qquad (30)$$

Since we conditioned on equation (14) in Lemma 2 holding true, $1_{\overline{\pi}_i^*(z^t),t}$ equals 1 for all $t \ge 1$ and therefore equation (28) and equation (29) are equivalent. Now observe that equation (29) is precisely the regret which a

[6] $\{V^t\}_{t=1}^T$ forms a martingale difference sequence with respect to $U^1, \ldots, U^T$, where $U^t \in [\sum_{j=1}^{k-1} p_j^t(z^t), \sum_{j=1}^{k} p_j^t(z^t))$ if and only if $a_i^t = k$.

player with reward function $f^t(\cdot) \in [0,1]$ incurs in a sleeping experts problem after $T_z = \sum_{t=1}^{T} 1_{\{z^t=z\}}$ repetitions of the game. Here T_z denotes the number of times context z^t is revealed. As mentioned previously, for each context $z \in \mathcal{Z}$ a distribution $p^t(z)$ is maintained which is updated whenever the revealed context z^t equals z. The update rule for each distribution (Line $3-4$ in Algorithm 2) corresponds exactly to that of general AdaNormalHedge proposed in Luo and Schapire (2014). Thus, at each round t action a_i^t is chosen according to general AdaNormalHedge with the set of sleeping experts $[1_{a_i,t}]_{a_i \in \mathcal{A}_i}$ defined as in equation (30) and which receives the full information feedback $\hat{r}^t = [f^t(a_i)]_{a_i \in \mathcal{A}_i}$. Concretely, a_i^t is sampled from a subset of awake experts, respective available actions, where an action $a_i \in \mathcal{A}_i$ is available at round t, if $1_{a_i,t} = 1$, and unavailable if $1_{a_i,t} = 0$.

In summary, equation (29) corresponds to the regret incurred by the general AdaNormalHedge algorithm after T_z repetitions and can therefore be upper-bounded by Theorem 2:

$$\sum_{t:z^t=z} 1_{\overline{\pi}_i^*(z^t),t} \left(f^t(\overline{\pi}_i^*(z^t)) - \sum_{a_i \in \mathcal{A}_i} \overline{p}_{a_i}^t(z^t) f^t(a_i^t) \right) \leq \sqrt{3T_z(\log(K) + \log(B) + \log(1+\log(K)))}, \tag{31}$$

where $B \leq \frac{5}{2} + \frac{3}{2}\log(1+T_z)$.

Now summing over all contexts $z \in \mathcal{Z}$ and using the Cauchy-Schwarz inequality, we can further upper-bound equation (27) with probability at least $1-\delta/2$:

$$\begin{aligned}
R^T \leq & \sum_{z \in \mathcal{Z}} \sum_{t:z^t=z} \min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t)\} - \sum_{a_i \in \mathcal{A}_i} \overline{p}_{a_i}^t(z^t) \min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\} \\
& + \sqrt{T/2\log(2/\delta)} + C_1 \beta_0^T \sqrt{T\gamma_0^T} \\
& \leq \sum_{z \in \mathcal{Z}} \sqrt{3T_z(\log(K) + \log(B) + \log(1+\log(K)))} + \sqrt{T/2\log(2/\delta)} + C_1 \beta_0^T \sqrt{T\gamma_0^T} \\
& \leq \sqrt{|\mathcal{Z}| \sum_{z \in \mathcal{Z}} 3T_z(\log(K) + \log(B) + \log(1+\log(K)))} + \sqrt{T/2\log(2/\delta)} + C_1 \beta_0^T \sqrt{T\gamma_0^T} \\
& = \sqrt{|\mathcal{Z}| 3T(\log(K) + \log(B) + \log(1+\log(K)))} + \sqrt{T/2\log(2/\delta)} + C_1 \beta_0^T \sqrt{T\gamma_0^T}
\end{aligned}$$

□

Proof. (of Theorem 1) By combining Lemma 2-5 Theorem 1 follows. In particular it follows by standard probability arguments:

$$\begin{aligned}
\mathbb{P}(\tilde{E}_1 \cap \tilde{E}_2) &= 1 - \mathbb{P}((\tilde{E}_1 \cap \tilde{E}_2)^C) \\
&= 1 - \mathbb{P}(\tilde{E}_1^C \cup \tilde{E}_2^C) \\
&\geq 1 - (\mathbb{P}(\tilde{E}_1^C) + \mathbb{P}(\tilde{E}_2^C)) \\
&\geq 1 - (\frac{\delta}{2} + \frac{\delta}{2}) = 1 - \delta,
\end{aligned}$$

where $\tilde{E}_1$ corresponds to equation (14) in Lemma 2 and $\tilde{E}_2$ corresponds to equation (26). By Lemma 1 and the Hoeffding-Azuma inequality, respectively, the following holds $\mathbb{P}(\tilde{E}_1^C) \leq \delta/2$ and $\mathbb{P}(\tilde{E}_2^C) \leq \delta/2$. □

A.1.2 Infinite (large) number of contexts

If the context space $\mathcal{Z}$ is large or infinite, we additionally assume that under the optimal feasible policy similar contexts lead to a similar performance Sessa et al. (2020). In the design of a thermal controller, contexts could refer to weather conditions which forms an infinitely larges set. Then, similar weather condition, e.g. the outdoor temperature, result in similar control input parameters under the optimal feasible policy.

Assumption 4. *The optimal feasible policy in hindsight $\overline{\pi}_i^*(\cdot)$ in (3) is L_p-Lipschitz, i.e.,*

$$|\overline{\pi}_i^*(z) - \overline{\pi}_i^*(z')| \leq L_p \|z - z'\|_1, \quad \forall z, z' \in \mathcal{Z}.$$

Algorithm 3 Update rule for infinite context space $\mathcal{Z}$

Set $\epsilon > 0$, $\mathcal{C} = \{z^1\}$, $R^1_{a_i}(z^1) = 0$, $C^1_{a_i}(z^1) = 0$, and let $p^1(z^1)$ be the uniform distribution.
for $t = 2, \dots, T$ **do**
  Set $\overline{z}^t = \arg\min_{z \in \mathcal{C}} \|z^t - z\|_1$.
  if $\|z^t - \overline{z}^t\|_1 > \epsilon$ **then**
    Add z^t to $\mathcal{C}$, set $\overline{z}^t = z^t, R^1_{a_i}(\overline{z}^t) = 0, C^1_{a_i}(\overline{z}^t) = 0$, and $p^t(\overline{z}^t)$ as the uniform distribution.
  else
    Compute reward estimate $\hat{r}^t$ with $\hat{r}^t(a_i) = \min\{1, \mathtt{UCB}^t_0(a_i, a^t_{-i}, z^t)\}, \quad \forall a_i$.
    Compute for $\overline{z}^t$ and all $a_i \in \mathcal{A}_i$, $R^t_{a_i}(\overline{z}^t)$ and $C^t_{a_i}(\overline{z}^t)$ via eq. (10) and (11) in Alg. 2.
    Update $p^{t+1}_{a_i}(\overline{z}^t)$ via eq. (4) in Alg. 2 and set $p^{t+1}(z^t) = p^{t+1}(\overline{z}^t)$.
  end if
end for

Moreover, the reward function $r_i(\cdot)$ is L_r-Lipschitz with respect to the first component, i.e.,

$$|r_i(a_i, a_{-i}, z) - r_i(a'_i, a_{-i}, z)| \leq L_r \|a_i - a'_i\|_1,$$

for all $a_i, a'_i \in \mathcal{A}_i$, all $a_{-i} \in \mathcal{A}_{-i}$ and $z \in \mathcal{Z}$.

Furthermore, we assume $\mathcal{Z} \subseteq [0,1]^d$ to obtain a scale-free regret bound and for simplicity, we consider context-independent constraints $g_m : \mathcal{A}_i \to \mathbb{R}$, $m \in [M]$. Assumption 1 then implies that $g_m(\overline{\pi}^*_i(z^t)) \leq 0$ for all $m \in [M]$.

Similarly to Hazan and Megiddo (2007); Sessa et al. (2020), the distribution $p^t(z^t) \in \Delta_K$ (line 6 of Algorithm 1) is computed by building an ϵ-net Clarkson (2005); Krauthgamer and Lee (2004) of the context space $\mathcal{Z}$ in a greedy fashion as new contexts are revealed. Concretely, at each round t, either a new L_1-ball centered at z^t is created, if z^t is more than ϵ away from the closest ball, or z^t is assigned to the closest L_1-ball. In the latter case, $p^t(z^t)$ is computed via the general AdaNormalHedge Luo and Schapire (2015) update rule using only past game data of those rounds $\tau < t$ for which z^τ belongs to the same L_1-ball as z^t. Under this computation rule of $p^t(z^t)$, summarized in Algorithm 3, c.z.AdaNormalGP achieves the following high-probability bounds.

Theorem 3. *Fix $\delta \in (0,1)$. Under Assumptions 1-4, if a player plays according to c.z.AdaNormalGP with $p^t(z^t)$ computed according to Algorithm 3 and $\beta^t_m = B_m + \sigma_m \sqrt{2(\gamma^{t-1}_m + 1 + \log(2(M+1)/\delta))}$ for all $m \in \{0\} \cup [M]$, then with probability at least $1 - \delta$:*

$$R^T = \mathcal{O}\big((L_r L_p)^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}} \sqrt{\log(K) + \log(B) + \log(1 + \log(K))} + \sqrt{T/2 \log(2/\delta)} + 2\beta^T_0 \sqrt{T\gamma^T_0}\big)$$
$$\mathcal{V}^T_m = \mathcal{O}\big(\beta^T_m \sqrt{T\gamma^T_m}\big), \quad \forall m \in [M],$$

where $B = 1 + \frac{3}{2}\frac{1}{K}\sum_{a_i=1}^K (1 + \log(1 + C^t_i(a_i))) \leq \frac{5}{2} + \frac{3}{2}\log(1+T)$.

In repeated contextual games, c.GPMW Sessa et al. (2020) algorithm also considers infinite context spaces and achieves similar contextual regret bound for this setting, namely, $\mathcal{O}((L_r L_p)^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}} \sqrt{\log K} + \sqrt{T \log(2/\delta)} + \beta^T_0 \sqrt{T\gamma^T_0})$. In the following, we provide a proof for Theorem 3.

Proof. Lemmas 2-4 in Appendix A.1.1 remain valid for infinite context space $\mathcal{Z}$ and context-independent constraints $g_m : \mathcal{A}_i \to \mathbb{R}$, where Assumption 1 implies $g_m(\overline{\pi}^*_i(z^t)) \leq 0$ for all $m \in [M]$. Thus, to prove Theorem 3 it remains to show an upper bound on the regret. We leverage that $|\mathcal{C}|$ separate distributions for all ϵ-close contexts $z \in \mathcal{Z}$ are maintained, where $|\mathcal{C}|$ is the total number of L_1-balls created up to round T. Thus, similarly to the proof of Theorem 1, we analyze the regret obtained by each L_1-ball separately. For each L_1-ball (fixed $z \in \mathcal{C}$), the proof further follows by decomposing the regret into the sum of two terms. The first term corresponds to the regret that a player incurs by considering the same distribution for ϵ-close contexts rather than a separate distribution for each context $z \in \mathcal{Z}$. However, this term can be upper-bounded by making use of Lipschitzness (Assumption 4). The second term can be bounded using the same proof techniques as in Theorem 1. The theorem then follows from the same standard probability arguments as in the proof of Theorem 1.

Lemma 6. *Conditioned on equation* (14) *in Lemma 2 holding true, the constrained contextual regret* R^T *is upper-bounded with probability* $1-\frac{\delta}{2}$ *by:*

$$R^T = \mathcal{O}\Big((L_r L_p)^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}} \sqrt{\log(K)+\log(B)+\log(1+\log(K))} + \sqrt{T/2\log(2/\delta)} + 2\beta_0^T \sqrt{T\gamma_0^T} \Big).$$

Proof. Note that for infinite (large) context space $\mathcal{Z}$, c.z.AdaNormalGP builds an ϵ-net of the context space by creating new L_1-balls in a greedy fashion to compute $p^t(z^t)$ (Algorithm 3). Thus, after T game rounds, the ϵ-net consists of the set of contexts $z \in \mathcal{Z}$, denoted by $\mathcal{C}$, that form the centers of the balls created so far. Moreover, the variable $\overline{z}^t$ indicates the center of the ball that context z^t belongs to. Thus, the regret can be rewritten as:

$$\begin{aligned} R^T &= \sum_{z\in\mathcal{C}} \sum_{t:\overline{z}^t=z} r(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t) - r(a_i^t, a_{-i}^t, z^t) \\ &= \underbrace{\sum_{z\in\mathcal{C}} \sum_{t:\overline{z}^t=z} r(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t) - r(\overline{\pi}_i^*(z), a_{-i}^t, z^t)}_{L-R^T} + \underbrace{\sum_{z\in\mathcal{C}} \sum_{t:\overline{z}^t=z} r(\overline{\pi}_i^*(z), a_{-i}^t, z^t) - r(a_i^t, a_{-i}^t, z^t)}_{R-R^T}, \end{aligned}$$

where in the last equality we add and subtract the term $r(\overline{\pi}_i^*(z), a_{-i}^t, z^t)$.

The first term $L-R^T$ can be bounded by making use of Assumption 4, namely, the L_r-Lipschitzness of $r(\cdot)$ in its first argument and L_p-Lipschitzness of the the optimal feasible policy $\overline{\pi}_i^*(\cdot)$:

$$L - R^T \le \sum_{z\in\mathcal{C}} \sum_{t:\overline{z}^t=z} L_r \|\overline{\pi}_i^*(z^t) - \overline{\pi}_i^*(z)\|_1 \le \sum_{z\in\mathcal{C}} \sum_{t:\overline{z}^t=z} L_r L_p \|z^t - z\|_1 \le L_r L_p T\epsilon \tag{32}$$

where we used that $\|z^t - z\|_1 \le \epsilon$ by definition of the ϵ-net (line 3 and 6 Algorithm 3).

Next, we proceed with bounding $R-R^T$. The reward and thus $R-R^T$ can be further upper bounded by the same arguments and steps (see equation (24) and (25)) as in the proof of Lemma 5:

$$\begin{aligned} R - R^T &\le \sum_{z\in\mathcal{C}} \sum_{t:\overline{z}^t=z} r(\overline{\pi}_i^*(z), a_{-i}^t, z^t) - r(a_i^t, a_{-i}^t, z^t) \\ &\le \sum_{z\in\mathcal{C}} \sum_{t:\overline{z}^t=z} (\min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z), a_{-i}^t, z^t)\} - \min\{1, \mathrm{UCB}_0^t(a_i^t, a_{-i}^t, z^t)\}) + C_1 \beta_0^T \sqrt{T\gamma_0^T}, \end{aligned} \tag{33}$$

where $C_1 = \frac{8}{\log(1+\sigma^{-2})}$.

The Hoeffding-Azuma inequality (Cesa-Bianchi and Lugosi, 2006, Lemma A.7) in (26) remains applicable in the case where the context set $\mathcal{Z}$ is infinite. Thus, with probability at least $1-\delta/2$ we obtain the following upper bound on $R-R^T$:

$$\begin{aligned} R - R^T \le &\sum_{z\in\mathcal{C}} \sum_{t:\overline{z}^t=z} \big(\min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z), a_{-i}^t, z^t)\} - \sum_{a_i\in\mathcal{A}_i} \overline{p}_{a_i}^t(z^t) \min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\} \big) \\ &+ \sqrt{T/2\log(2/\delta)} + C_1 \beta_0^T \sqrt{T\gamma_0^T}. \end{aligned} \tag{34}$$

Define the function $f^t(\cdot) = \min\{1, \mathrm{UCB}_0^t(\cdot, a_{-i}^t, z^t)\}$. Since we conditioned on equation (14) in Lemma 2 holding true, $\mathrm{UCB}_0^t(\cdot) \ge 0$ since $r(\cdot) \ge 0$ and thus $f^t(\cdot) \in [0,1]^K$. Similarly to equation (28) and (29), for a fixed $z \in \mathcal{C}$, the first term in equation (34) can be rewritten as:

$$\begin{aligned} &\sum_{t:\overline{z}^t=z} \min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z), a_{-i}^t, z^t)\} - \sum_{a_i\in\mathcal{A}_i} \overline{p}_{a_i}^t(z^t) \min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\} \\ &= \sum_{t:\overline{z}^t=z} f^t(\overline{\pi}_i^*(z)) - \sum_{a_i\in\mathcal{A}_i} \overline{p}_{a_i}^t(z^t) f^t(a_i^t) \qquad (35) \\ &= \sum_{t:\overline{z}^t=z} 1_{\overline{\pi}_i^*(z),t} \Big(f^t(\overline{\pi}_i^*(z)) - \sum_{a_i\in\mathcal{A}_i} \overline{p}_{a_i}^t(z^t) f^t(a_i^t) \Big), \qquad (36) \end{aligned}$$

where for all $a_i \in \mathcal{A}_i$ we define:

$$1_{a_i,t} := \begin{cases} 1, & \text{if } \mathrm{LCB}_m^t(a_i) \leq 0 \text{ for every } m \in [M], \\ 0, & \text{otherwise.} \end{cases} \tag{37}$$

Since we conditioned on equation (14) in Lemma 2 holding true, $1_{\overline{\pi}_i^*(z),t}$ equals 1 for all $t \geq 1$ since $z \in \{z^1, \cdots, z^T\}$ and thus $\mathrm{LCB}_m^t(\overline{\pi}_i^*(z)) \leq g_m(\overline{\pi}_i^*(z)) \leq 0$. Therefore, equation (35) and equation (36) are equivalent. Observe that equation (36) is precisely the regret with respect to $\overline{\pi}_i^*(z)$ which a player with reward functions $f_i^t(\cdot) \in [0,1]$ incurs in a sleeping experts problem after $T_z = \sum_{t=1}^T 1_{\{\overline{z}^t = z\}}$ repetitions of the game. Here T_z denotes the number of times the revealed context z^t belonged to the ball centered at $z \in \mathcal{C}$. Note, furthermore, that for each context $z \in \mathcal{C}$ a distribution $p^t(z)$ is maintained which is updated whenever context z^t belongs to the ball with center z. The update rule for each distribution (Line $7-8$ in Algorithm 3) corresponds exactly to that of general AdaNormalHedge proposed in Luo and Schapire (2014). Thus, at each round t, action a_i^t is chosen according to general AdaNormalHedge with the set of sleeping experts $[1_{a_i,t}]_{a_i \in \mathcal{A}_i}$ defined as in equation (37) and which receives the full information feedback $\hat{r}^t = [f^t(a_i)]_{a_i \in \mathcal{A}_i}$.

In summary, equation (36) corresponds to the regret incurred by the general AdaNormalHedge algorithm with respect to expert $\overline{\pi}_i^*(z) \in \mathcal{A}_i$ after T_z repetitions and can therefore be upper-bounded by Theorem 2 as follows:

$$\sum_{t:\overline{z}^t = z} 1_{\overline{\pi}_i^*(z),t}(f^t(\overline{\pi}_i^*(z)) - f^t(a_i^t)) \leq \sqrt{3T_z(\log(K) + \log(B) + \log(1 + \log(K)))}, \tag{38}$$

where $B =\leq \frac{5}{2} + \frac{3}{2}\log(1 + T_z)$.

Now summing over all the contexts $z \in \mathcal{C}$ and using the Cauchy-Schwarz inequality, we can further upper-bound equation (34) with probability at least $1 - \delta/2$:

$$\begin{aligned} R - R^T &\leq \sum_{z \in \mathcal{C}} \sum_{t:\overline{z}^t = z} \min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z), a_{-i}^t, z^t)\} - \sum_{a_i \in \mathcal{A}_i} \overline{p}_{a_i}^t(z^t) \min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\} \\ &+ \sqrt{T/2\log(2/\delta)} + C_1 \beta_0^T \sqrt{T\gamma_0^T} \\ &\leq \sum_{z \in \mathcal{C}} \sqrt{3T_z(\log(K) + \log(B) + \log(1 + \log(K)))} + \sqrt{T/2\log(2/\delta)} + C_1 \beta_0^T \sqrt{T\gamma_0^T} \\ &\leq \sqrt{|\mathcal{C}| \sum_{z \in \mathcal{C}} 3T_z(\log(K) + \log(B) + \log(1 + \log(K)))} + \sqrt{T/2\log(2/\delta)} + C_1 \beta_0^T \sqrt{T\gamma_0^T} \\ &= \sqrt{|\mathcal{C}| 3T(\log(K) + \log(B) + \log(1 + \log(K)))} + \sqrt{T/2\log(2/\delta)} + C_1 \beta_0^T \sqrt{T\gamma_0^T} \\ &\leq \epsilon^{-d/2} \sqrt{3T(\log(K) + \log(B) + \log(1 + \log(K)))} + \sqrt{T/2\log(2/\delta)} + C_1 \beta_0^T \sqrt{T\gamma_0^T}. \end{aligned} \tag{39}$$

In the last inequality we used that $\mathcal{C} \leq (1/\epsilon)^d$, because the contexts space $\mathcal{Z} \subseteq [0,1]^d$ can be covered by at most $(1/\epsilon)^d$ balls of radius ϵ such that the distance between their centers is at least ϵ Clarkson (2005).

Combining the bounds of $L - R^T$ in (32) and $R - R^T$ in (39), with probability at least $1 - \delta/2$ the regret is bounded by:

$$\begin{aligned} R^T &= L - R^T + R - R^T \\ &\leq L_r L_p T\epsilon + \epsilon^{-d/2} \sqrt{3T(\log(K) + \log(B) + \log(1 + \log(K)))} \\ &+ \sqrt{T/2\log(2/\delta)} + C_1 \beta_0^T \sqrt{T\gamma_0^T} \\ &= (L_r L_p)^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}} (1 + \sqrt{3(\log(K) + \log(B) + \log(1 + \log(K)))}) \\ &+ \sqrt{T/2\log(2/\delta)} + C_1 \beta_0^T \sqrt{T\gamma_0^T}, \end{aligned}$$

where in the last equality ϵ is set as $(L_r L_p)^{-\frac{2}{d+2}} T^{-\frac{1}{d+2}}$. □

□

A.2 Supplementary Material for Remark 1

In this section, we show that we can obtain high probability bounds on the regret and the cumulative constraint violations for Algorithm 1 presented in Section 2, when other computation rules of $p^t(z^t)$ (line 6 of Algorithm 1) than the update rule of AdaNormalHedge Luo and Schapire (2015) are used. Recall that once $p^t(z^t)$ is specified, Algorithm 1 is well-defined.

A.2.1 Expert algorithm bounds for finite context space

In this subsection, we present general regret bounds that depend on the specified computation rule $p^t(z^t)$ for the case when the context space $\mathcal{Z}$ is finite. As in Section 2.1, the distribution $p^t(z^t)$ is computed by maintaining a distribution for each context $z \in \mathcal{Z}$. For a fixed context $z \in \mathcal{Z}$, however, any update rule that corresponds to an update rule of a sleeping expert problem with regret bounds can be used to obtain a high probability regret bound. This follows from the connection between playing a repeated game with unknown constraints and playing a sleeping expert problem. In particular, since any regular expert algorithm such as Hedge Freund and Schapire (1995) can be reformulated as a sleeping expert algorithm Freund et al. (1997) with the same regret bound, for a fixed context $z \in \mathcal{Z}$, any update rule that corresponds to the reformulated update rule of an expert problem can be used to obtain a high probability regret bound. This bound then depends on the regret obtained by the expert algorithm. Algorithm 4 provides a computation rule for $p^t(z^t)$ which is based on the reformulation of the update rule of any no-regret expert algorithm $\mathcal{E}$ to obtain a sleeping expert algorithm. The bounds on the cumulative constraint violations obtained in Theorem 1 are independent of the update rule and therefore remain valid.

Theorem 4. *Fix $\delta \in (0,1)$. Under the same assumptions as in Theorem 1, if a player plays according to Algorithm 1 with $p^t(z^t)$ computed according to Algorithm 4 then with probability at least $1-\delta$:*

$$R^T = \mathcal{O}\left(\sqrt{|\mathcal{Z}| \sum_{z\in\mathcal{Z}} (R^{T_z}(\mathcal{E}))^2} + \sqrt{T\log(2/\delta)} + \beta_0^T \sqrt{T\gamma_0^T}\right)$$
$$\mathcal{V}_m^T = \mathcal{O}\left(\beta_m^T \sqrt{T\gamma_m^T}\right), \quad \forall m \in [M],$$

where $R^{T_z}(\mathcal{E})$ is the regret obtained by an expert algorithm $\mathcal{E}$ after T_z rounds and $T_z = \sum_{t=1}^T 1_{\{z^t=z\}}$. We denote Algorithm 1 with $p^t(z^t)$ computed according to Algorithm 4 as c.z.$\mathcal{E}$GP.

For example, when using the update rule of the well-known no-regret algorithm Hedge Freund and Schapire (1995) to compute $p^t(z^t)$ in Algorithm 4, we obtain the following bound on the regret.

Corollary 2. *If the expert algorithm $\mathcal{E}$ corresponds to the Hedge algorithm with step size $\eta^t = 2\sqrt{\log K / \sum_{\tau=1}^t 1_{\{z^\tau=z^t\}}}$ and prediction rule:*

$$p_{a_i}^{t+1}(z) = p_{a_i}^t(z) \exp(\eta^t UCB_0^t(a_i, a_{-i}^t, z^t) 1_{\{z^t=z\}}, \quad \forall a_i \in \mathcal{A}_i,\ z \in \mathcal{Z},$$

then playing according to the so-called c.z.HedgeGP algorithm[7] yields the following high-probability regret bound:

$$R^T = \mathcal{O}(\sqrt{|\mathcal{Z}| T \log K} + \sqrt{T\log(2/\delta)} + \beta_0^T \sqrt{T\gamma_0^T}).$$

Proof. The proof follows from Theorem 4 and (Mourtada and Gaïffas, 2019, Proposition 1) which states that with η^t defined as in Corollary 2, the regret of Hedge which receives full-information feedback $\mathbf{r}^t \in [0,1]^K$ is bounded as follows $\sum_{t=1}^T \left(r^t(a_i^*) - \sum_{a_i\in\mathcal{A}_i} p_{a_i}^t r^t(a_i^t)\right) \leq \sqrt{T\log K}$. □

Proof. (of Theorem 4) Lemma 2-4 in Appendix A.1.1 remain valid since the same assumptions as in Theorem 1 are made. Thus, to prove Theorem 4 it is left to show that conditioned on equation (14) in Lemma 2 holding

[7] Here, we consider the case where $p^t(z^t)$ is computed according to Algorithm 4.

Algorithm 4 Update rule for finite context space $\mathcal{Z}$

Input: Full-information **no-regret algorithm** $\mathcal{E}$. Let $p^1(z)$ be the uniform distribution $\forall z \in \mathcal{Z}$.

1: **for** $t = 2, \dots, T$ **do**

2: `Compute reward estimate` $\hat{\mathbf{r}}^t$ `with`

$$\hat{r}^t(a_i) = \begin{cases} \min\{1, \mathtt{UCB}_0^t(a_i, a_{-i}^t, z^t)\}, & \mathtt{if\ LCB}_m^t(a_i, z^t) \leq 0,\ \forall m \in [M], \\ \langle \overline{p}^t(z^t), \hat{r}^t \rangle, & \mathtt{otherwise}. \end{cases} \tag{40}$$

3: `Pass` $\hat{\mathbf{r}}^t \in \mathbb{R}^K$ `and` $p^t(z^t) \in \mathbb{R}^K$ `to algorithm` $\mathcal{E}$.

4: `Let` $p^{t+1}(z^t)$ `be the prediction of algorithm` $\mathcal{E}$.

5: **end for**

true, the regret R^T is upper-bounded with probability at least $1 - \frac{\delta}{2}$ by:

$$R^T \leq \sqrt{|\mathcal{Z}| \sum_{z \in \mathcal{Z}} (R^{T_z}(\mathcal{E}))^2} + \sqrt{T/2 \log(2/\delta)} + C_1 \beta_0^T \sqrt{T \gamma_0^T}. \tag{41}$$

The theorem then follows from the same standard probability arguments as in the proof of Theorem 1.

The main difference between Lemma 5 used to prove Theorem 1 and the statement in equation (41) is that the update rule $p^t(z^t)$ is computed according to Algorithm 2 for the former and according to Algorithm 4 for the latter. Thus, to show the statement in equation (41), any steps and arguments from Lemma 5 that are independent of the computation rule of $p^t(z^t)$ can be repeated.

Since c.z.$\mathcal{E}$GP (Algorithm 1 with $p^t(z^t)$ computed according to Algorithm 4) maintains a separate distribution $p^t(z)$ for each context $z \in \mathcal{Z}$ the regret notion can be rewritten as:

$$R^T = \sum_{z \in \mathcal{Z}} \sum_{t: z^t = z} r(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t) - r(a_i^t, a_{-i}^t, z^t).$$

The regret R^T can be further upper-bounded by the same arguments and steps as in equations (24), (25) and (26). Then, with probability at least $1 - \frac{\delta}{2}$ the following upper bound holds:

$$\begin{aligned} R^T &\leq \sum_{z \in \mathcal{Z}} \sum_{t: z^t = z} r(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t) - r(a_i^t, a_{-i}^t, z^t) \\ &\leq \sum_{z \in \mathcal{Z}} \sum_{t: z^t = z} (\min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t)\} - \min\{1, \mathrm{UCB}_0^t(a_i^t, a_{-i}^t, z^t)\}) + C_1 \beta_0^T \sqrt{T \gamma_0^T} \\ &\leq \sum_{z \in \mathcal{Z}} \sum_{t: z^t = z} \big(\min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t)\} - \sum_{a_i \in \mathcal{A}_i} \overline{p}_{a_i}^t(z^t) \min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\} \big) \\ &+ \sqrt{T/2 \log(2/\delta)} + C_1 \beta_0^T \sqrt{T \gamma_0^T}. \end{aligned} \tag{42}$$

where $C_1 = \frac{8}{\log(1+\sigma^{-2})}$.

Recall the function $\hat{\mathbf{r}}^t$ (line 2 of Algorithm 4) defined as:

$$\hat{r}^t(a_i) = \begin{cases} \min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\}, & \text{if } \mathrm{LCB}_m^t(a_i, z^t) \leq 0,\ \forall m \in [M], \\ \langle \overline{p}^t(z^t), \hat{r}^t \rangle, & \text{otherwise}. \end{cases} \tag{43}$$

For a fixed $z \in \mathcal{Z}$, the first term in equation (42) can be rewritten as:

$$\sum_{t:z^t=z} \min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t)\} - \sum_{a_i \in \mathcal{A}_i} \overline{p}_{a_i}^t(z^t) \min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\} \tag{44}$$

$$= \sum_{t:z^t=z} \hat{r}^t(\overline{\pi}_i^*(z^t)) - \sum_{a_i \in \mathcal{A}_i} \overline{p}_{a_i}^t(z^t) \hat{r}^t(a_i) \tag{45}$$

$$= \sum_{t:z^t=z} 1_{\overline{\pi}_i^*(z^t),t} \left(\hat{r}^t(\overline{\pi}_i^*(z^t)) - \sum_{a_i \in \mathcal{A}_i} \overline{p}_{a_i}^t(z^t) \hat{r}^t(a_i) \right), \tag{46}$$

$$= \sum_{t:z^t=z} \left(\hat{r}^t(\overline{\pi}_i^*(z^t)) - \sum_{a_i \in \mathcal{A}_i} p_{a_i}^t(z^t) \hat{r}^t(a_i) \right), \tag{47}$$

where for all $a_i \in \mathcal{A}_i$ the indicator function $1_{a_i,t}$ is defined as in equation (30)

Equation (45) follows by definition of $\hat{\mathbf{r}}$ (line 2 Algorithm 4). Concretely, since we conditioned on equation (14) in Lemma 2 holding true, $\mathrm{LCB}_m^t(\overline{\pi}_i^*(z^t), z^t) \leq 0$ for all $m \in [M]$ and thus $\hat{r}^t(\overline{\pi}_i^*(z^t)) = \min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t)\}$. Furthermore, note that $\overline{p}_{a_i}^t(z^t) = 0$ if there exists an $m \in [M]$ such that $\mathrm{LCB}_m^t(a_i, z^t) > 0$ and in this case $\overline{p}_{a_i}^t(z^t) \min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\} = \overline{p}_{a_i}^t(z^t)\hat{r}^t(a_i) = 0$. Since we conditioned on equation (14) in Lemma 2 holding true, $1_{\overline{\pi}_i^*(z^t),t}$ equals 1 for all $t \geq 1$ and therefore equation (45) and equation (46) are equivalent. Equation (47) follows from the fact that any expert problem can be reformulated as a sleeping expert problem and vice versa Freund et al. (1997). Specifically, if we are given a no-regret expert algorithm $\mathcal{E}$ with prediction $p^t(z^t)$ on round t then the prediction of the corresponding sleeping expert algorithm is given by $\overline{p}_{a_i}^t(z^t) \propto 1_{a_i,t} p_{a_i}^t(z^t)$ on round t. Furthermore, the reward vector $\hat{\mathbf{r}}^t$ which is given to $\mathcal{E}$ as full-information feedback is designed such that:

$$\sum_{a_i : 1_{a_i,t}=1} p_{a_i}^t(z^t)\hat{r}^t(a_i) + \sum_{a_i : 1_{a_i,t}=0} p_{a_i}^t(z^t)\hat{r}^t(a_i) = \sum_{a_i : 1_{a_i,t}=1} \overline{p}_{a_i}^t(z^t)\hat{r}^t(a_i).$$

thus, by design the equivalence between equation (46) and equation (47) follows. Note furthermore, that since we conditioned on equation (14) in Lemma 2 holding true, $\mathrm{UCB}_0^t(\cdot) \geq 0$ since $r(\cdot) \geq 0$ and thus $\hat{\mathbf{r}}^t \in [0,1]^K$. Now observe that equation (47) is precisely the regret which a player with reward function $\hat{r}^t(\cdot) \in [0,1]$ incurs in an expert problem after $T_z = \sum_{t=1}^T 1_{\{z^t=z\}}$ repetitions of the game. Here T_z denotes the number of times context z^t is revealed. The prediction $p^t(z^t)$ for each context $z \in \mathcal{Z}$ (Line 4 in Algorithm 4) is given by the no-regret algorithm $\mathcal{E}$ which receives full-information feedback $\hat{\mathbf{r}}^t = [\hat{r}^t(a_i)]_{a_i \in \mathcal{A}_i}$. Thus, equation (47) corresponds to the regret incurred by the algorithm $\mathcal{E}$ after T_z repetitions, i.e.,

$$\sum_{t:z^t=z} \left(\hat{r}^t(\overline{\pi}_i^*(z^t)) - \sum_{a_i \in \mathcal{A}_i} p_{a_i}^t(z^t)\hat{r}^t(a_i^t) \right) \leq R^{T_z}(\mathcal{E}). \tag{48}$$

Now summing over all contexts $z \in \mathcal{Z}$ and using the Cauchy-Schwarz inequality, we can further upper-bound equation (42) with probability at least $1 - \delta/2$:

$$\begin{aligned} R^T \leq & \sum_{z \in \mathcal{Z}} \sum_{t:z^t=z} \min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t)\} - \sum_{a_i \in \mathcal{A}_i} \overline{p}_{a_i}^t(z^t) \min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\} \\ & + \sqrt{T/2 \log(2/\delta)} + C_1 \beta_0^T \sqrt{T \gamma_0^T} \\ & \leq \sum_{z \in \mathcal{Z}} R^{T_z}(\mathcal{E}) + \sqrt{T/2 \log(2/\delta)} + C_1 \beta_0^T \sqrt{T \gamma_0^T} \\ & \leq \sqrt{|\mathcal{Z}| \sum_{z \in \mathcal{Z}} (R^{T_z}(\mathcal{E}))^2} + \sqrt{T/2 \log(2/\delta)} + C_1 \beta_0^T \sqrt{T \gamma_0^T} \end{aligned}$$

□

A.2.2 Expert algorithm bounds for infinite context space

In this subsection, we present general regret bounds that depend on the specified computation rule $p^t(z^t)$ for the case when the context space $\mathcal{Z}$ is infinite. As in Appendix A.1.2, the distribution $p^t(z^t)$ is computed by

Algorithm 5 Update rule for infinite (large) context space $\mathcal{Z}$

Input: Full-information **no-regret algorithm** $\mathcal{E}$.
Set: $\epsilon > 0$ and $\mathcal{C} = \{z^1\}$ and let $p^1(z^1)$ be the uniform distribution.

for $t = 2, \dots, T$ **do**
  Set $\overline{z}^t = \arg\min_{z \in \mathcal{C}} \|z^t - z\|_1$.
  if $\|z^t - \overline{z}^t\|_1 > \epsilon$ **then**
    Add z^t to $\mathcal{C}$, set $\overline{z}^t = z^t$ and let $p^t(\overline{z}^t)$ be the uniform distribution.
  else
    Compute reward estimate $\hat{\mathtt{r}}^t$ with

$$\hat{r}^t(a_i) = \begin{cases} \min\{1, \mathtt{UCB}_0^t(a_i, a_{-i}^t, z^t)\}, & \text{if } \mathtt{LCB}_m^t(a_i) \leq 0, \ \forall m \in [M], \\ \langle \overline{p}^t(z^t), \hat{r}^t \rangle, & \text{otherwise.} \end{cases} \tag{49}$$

    Pass $\hat{\mathtt{r}}^t \in \mathbb{R}^K$ and $p^t(\overline{z}^t) \in \mathbb{R}^K$ to algorithm $\mathcal{E}$.
    Let $p^{t+1}(\overline{z}^t)$ be the prediction of $\mathcal{E}$ and set $p^{t+1}(z^t) = p^{t+1}(\overline{z}^t)$.
  end if
end for

maintaining $|\mathcal{C}|$ separate distributions for all ϵ-close contexts $z \in \mathcal{Z}$, where $|\mathcal{C}|$ is the total number of L_1-balls created up to round T. For each L_1-ball (fixed $z \in \mathcal{C}$), any update rule that corresponds to the reformulated update rule of an expert problem can be used to obtain a high probability regret bound which then depends on the regret obtained by the expert algorithm.

Theorem 5. *Fix $\delta \in (0,1)$. Under the same assumptions as in Theorem 3, if a player plays according to Algorithm 1 with $p^t(z^t)$ computed according to Algorithm 5, then with probability at least $1-\delta$*

$$R^T = \mathcal{O}\left((L_r L_p)^{\frac{d}{d+2}} T^{\frac{d}{d+2}} \left(\sum_{z \in \mathcal{C}} (R^{T_z}(\mathcal{E}))^2 \right)^{\frac{1}{d+2}} + \sqrt{T \log(2/\delta)} + \beta_0^T \sqrt{T \gamma_0^T} \right)$$

$$\mathcal{V}_m^T = \mathcal{O}\left(\beta_m^T \sqrt{T \gamma_m^T} \right), \quad \forall m \in [M],$$

where $R^{T_z}(\mathcal{E})$ is the regret obtained by algorithm $\mathcal{E}$ after T_z rounds and $T_z = \sum_{t=1}^T 1_{\{\overline{z}^t = z\}}$ is the number of times the revealed context z^t belonged to a ball centered at $z \in \mathcal{Z}$.

For example, when using the update rule of the well-known no-regret algorithm Hedge Freund and Schapire (1995) to compute $p^t(z^t)$ in Algorithm 5, we obtain the following bound on the regret.

Corollary 3. *If algorithm $\mathcal{E}$ corresponds to the Hedge algorithm with step size $\eta^t = 2\sqrt{\log K / \sum_{\tau=1}^t 1_{\{\overline{z}^t = \overline{z}^\tau\}}}$ and prediction rule:*

$$p_{a_i}^{t+1}(z) = p_{a_i}^t(z) \exp(\eta^t UCB_0^t(a_i, a_{-i}^t, z^t) 1_{\{\overline{z}^t = z\}}, \quad \forall a_i \in \mathcal{A}_i, \ z \in \mathcal{Z},$$

then playing according to the so-called c.z.HedgeGP algorithm[8] yields the following high-probability regret bound:

$$R^T = \mathcal{O}((L_r L_p)^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}} \log(K)^{\frac{1}{d+2}} + \sqrt{T \log(2/\delta)} + \beta_0^T \sqrt{T \gamma_0^T}).$$

Proof. The proof follows from Theorem 5 and (Mourtada and Gaïffas, 2019, Proposition 1). □

Proof. (of Theorem 5) Lemma 2-4 in Appendix A.1.1 remain valid for infinite context space $\mathcal{Z}$ and context-independent constraints $g_m : \mathcal{A}_i \to \mathbb{R}$, where Assumption 1 implies $g_m(\overline{\pi}_i^*(z^t)) \leq 0$ for all $m \in [M]$. Thus, to

[8] Here, we consider the case where $p^t(z^t)$ is computed according to Algorithm 5.

prove Theorem 5 it is left to show that conditioned on equation (14) in Lemma 2 holding true, the regret is upper-bounded with probability at least $1-\frac{\delta}{2}$ by:

$$R^T \leq (L_r L_p)^{\frac{d}{d+2}} T^{\frac{d}{d+2}} \left(\sum_{z\in\mathcal{C}} (R^{T_z}(\mathcal{E}))^2 \right)^{\frac{1}{d+2}} + \sqrt{T/2\log(2/\delta)} + C_1 \beta_0^T \sqrt{T\gamma_0^T}. \tag{50}$$

The theorem then follows from the same standard probability arguments as in the proof of Theorem 1.

The main difference between Lemma 6 and the statement in equation (50) is that the update rule $p^t(z^t)$ is computed according to Algorithm 2 for the former and according to Algorithm 4 for the latter. Thus, to show the statement in equation (50), any steps and arguments from Lemma 6 that are independent of the computation rule of $p^t(z^t)$ can be repeated.

Analogously to c.z.AdaNormalGP (Algorithm 1 with $p^t(z^t)$ computed by Algorithm 3), c.z.$\mathcal{E}$GP (Algorithm 1 with $p^t(z^t)$ computed by Algorithm 5) builds an ϵ-net of the context space by creating new L_1-balls in a greedy fashion. Here, the variable $\overline{z}^t$ indicates the ball that context z^t belongs to. Then, the regret can be rewritten as:

$$\begin{aligned} R^T &= \sum_{z\in\mathcal{C}} \sum_{t:\overline{z}^t=z} r(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t) - r(a_i^t, a_{-i}^t, z^t) \\ &= \underbrace{\sum_{z\in\mathcal{C}} \sum_{t:\overline{z}^t=z} r(\overline{\pi}_i^*(z^t), a_{-i}^t, z^t) - r(\overline{\pi}_i^*(z), a_{-i}^t, z^t)}_{L-R^T} + \underbrace{\sum_{z\in\mathcal{C}} \sum_{t:\overline{z}^t=z} r(\overline{\pi}_i^*(z), a_{-i}^t, z^t) - r(a_i^t, a_{-i}^t, z^t)}_{R-R^T}, \end{aligned}$$

where in the last equality we add and subtract the term $r(\overline{\pi}_i^*(z), a_{-i}^t, z^t)$.

The first term $L-R^T$ can be bounded by making use of Assumption 4 and then applying the same arguments as in the proof of Lemma 6:

$$L - R^T \leq L_r L_p T \epsilon. \tag{51}$$

Next, we proceed with bounding $R-R^T$. It can be further upper-bounded by the same arguments and steps as in equations (24), (25) and (26). Then, with probability at least $1-\frac{\delta}{2}$ the following upper bound holds:

$$\begin{aligned} R - R^T &\leq \sum_{z\in\mathcal{C}} \sum_{t:\overline{z}^t=z} r(\overline{\pi}_i^*(z), a_{-i}^t, z^t) - r(a_i^t, a_{-i}^t, z^t) \\ &\leq \sum_{z\in\mathcal{C}} \sum_{t:\overline{z}^t=z} (\min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z), a_{-i}^t, z^t)\} - \min\{1, \mathrm{UCB}_0^t(a_i^t, a_{-i}^t, z^t)\}) \\ &+ C_1 \beta_0^T \sqrt{T\gamma_0^T} \\ &\sum_{z\in\mathcal{C}} \sum_{t:\overline{z}^t=z} \big(\min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z), a_{-i}^t, z^t)\} - \sum_{a_i\in\mathcal{A}_i} \overline{p}_{a_i}^t(z^t) \min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\} \big) \\ &+ \sqrt{T/2\log(2/\delta)} + C_1 \beta_0^T \sqrt{T\gamma_0^T}. \end{aligned} \tag{52}$$

where $C_1 = \frac{8}{\log(1+\sigma^{-2})}$.

Recall the function $\hat{\mathbf{r}}^t$ (line 7 of Algorithm 5) defined as:

$$\hat{r}^t(a_i) = \begin{cases} \min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\}, & \text{if } \mathrm{LCB}_m^t(a_i) \leq 0,\ \forall m \in [M], \\ \langle \overline{p}^t(z^t), \hat{r}^t \rangle, & \text{otherwise.} \end{cases} \tag{53}$$

For a fixed $z \in \mathcal{C}$, the first term in equation (52) can be rewritten as:

$$\begin{aligned}
&\sum_{t:z^t=z} \min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z), a_{-i}^t, z^t)\} - \sum_{a_i \in \mathcal{A}_i} \overline{p}_{a_i}^t(z^t) \min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\} \\
&= \sum_{t:z^t=z} \hat{r}^t(\overline{\pi}_i^*(z^t)) - \sum_{a_i \in \mathcal{A}_i} \overline{p}_{a_i}^t(z) \hat{r}^t(a_i^t) && (54) \\
&= \sum_{t:z^t=z} 1_{\overline{\pi}_i^*(z),t} \left(\hat{r}^t(\overline{\pi}_i^*(z)) - \sum_{a_i \in \mathcal{A}_i} \overline{p}_{a_i}^t(z^t) \hat{r}^t(a_i^t) \right), && (55) \\
&= \sum_{t:z^t=z} \left(\hat{r}^t(\overline{\pi}_i^*(z)) - \sum_{a_i \in \mathcal{A}_i} p_{a_i}^t(z^t) \hat{r}^t(a_i^t) \right), && (56)
\end{aligned}$$

Equation (54) follows by definition of $\hat{\mathbf{r}}$ (line 2 Algorithm 4). Since we conditioned on equation (14) in Lemma 2 holding true, $1_{\overline{\pi}_i^*(z),t}$ equals 1 for all $t \geq 1$ and therefore equation (54) and equation (55) are equivalent. Equivalence between equation (55) and equation (56) follows from the same arguments which are used in the proof of Theorem 4 to show equivalence between equation (46) and equation (47). Note furthermore, that since we condition on equation (14) in Lemma 2 holding true, $\mathrm{UCB}_0^t(\cdot) \geq 0$ since $r(\cdot) \geq 0$ and thus $\hat{r}^t(\cdot) \in [0,1]$. Now observe that equation (56) is precisely the regret with respect to expert $\overline{\pi}_i^*(z)$ which a player with reward function $\hat{r}^t(\cdot) \in [0,1]$ incurs in an expert problem after $T_z = \sum_{t=1}^T 1_{\{\overline{z}^t=z\}}$ repetitions of the game. Here T_z denotes the number of times the revealed context z^t belonged to the ball centered at $z \in \mathcal{C}$. Recall that for each context $z \in \mathcal{C}$ a distribution $p^t(z)$ is maintained which is updated whenever context z^t belongs to the ball with center z. The prediction $p^t(z^t)$ (Line 9 in Algorithm 5) is given by the no-regret algorithm $\mathcal{E}$ which receives full-information feedback $\hat{\mathbf{r}}^t = [\hat{r}^t(a_i)]_{a_i \in \mathcal{A}_i}$. Thus, equation (56) corresponds to the regret incurred by the algorithm $\mathcal{E}$ after T_z repetitions and is upper-bounded by the regret bound of $\mathcal{E}$:

$$\sum_{t:\overline{z}^t=z} \left(\hat{r}^t(\overline{\pi}_i^*(z)) - \sum_{a_i \in \mathcal{A}_i} p_{a_i}^t(z^t) \hat{r}^t(a_i^t) \right) \leq R^{T_z}(\mathcal{E}). \tag{57}$$

Summing over all the contexts $z \in \mathcal{C}$, we can further upper-bound equation (52) with probability at least $1 - \delta/2$:

$$\begin{aligned}
R - R^T \leq &\sum_{z \in \mathcal{C}} \sum_{t:\overline{z}^t=z} \min\{1, \mathrm{UCB}_0^t(\overline{\pi}_i^*(z), a_{-i}^t, z^t)\} - \sum_{a_i \in \mathcal{A}_i} \overline{p}_{a_i}^t(z^t) \min\{1, \mathrm{UCB}_0^t(a_i, a_{-i}^t, z^t)\} \\
&+ \sqrt{T/2 \log(2/\delta)} + C_1 \beta_0^T \sqrt{T \gamma_0^T} \\
&\leq \sum_{z \in \mathcal{C}} R^{T_z}(\mathcal{E}) + \sqrt{T/2 \log(2/\delta)} + C_1 \beta_0^T \sqrt{T \gamma_0^T} \\
&\leq \sqrt{|\mathcal{C}| \sum_{z \in \mathcal{C}} (R^{T_z}(\mathcal{E}))^2} + \sqrt{T/2 \log(2/\delta)} + C_1 \beta_0^T \sqrt{T \gamma_0^T} \\
&\leq \epsilon^{-d/2} \sqrt{\sum_{z \in \mathcal{C}} (R^{T_z}(\mathcal{E}))^2} + \sqrt{T/2 \log(2/\delta)} + C_1 \beta_0^T \sqrt{T \gamma_0^T}. && (58)
\end{aligned}$$

In the last equality we used $\mathcal{C} \leq (1/\epsilon)^d$, because the contexts space $\mathcal{Z} \subseteq [0,1]^d$ can be covered by at most $(1/\epsilon)^d$ balls of radius ϵ such that the distance between their centers is at least ϵ Clarkson (2005).

Then, with probability at least $1 - \frac{\delta}{2}$ the regret is bounded by:

$$\begin{aligned}
R^T &= L - R^T + R - R^T \\
&\leq L_r L_p T \epsilon + \epsilon^{-d/2} \sqrt{\sum_{z \in \mathcal{C}} (R^{T_z}(\mathcal{E}))^2} + \sqrt{T/2 \log(2/\delta)} + C_1 \beta_0^T \sqrt{T \gamma_0^T} \\
&= (L_r L_p)^{\frac{d}{d+2}} T^{\frac{d}{d+2}} \left(\sum_{z \in \mathcal{C}} (R^{T_z}(\mathcal{E}))^2 \right)^{\frac{1}{d+2}} + \sqrt{T/2 \log(2/\delta)} + C_1 \beta_0^T \sqrt{T \gamma_0^T},
\end{aligned}$$

where in the last equality ϵ is set as $(L_r L_p)^{-\frac{2}{d+2}} T^{-\frac{2}{d+2}} \left(\sum_{c \in \mathcal{C}} (R^{T_z}(\mathcal{E}))^2 \right)^{\frac{1}{d+2}}$. □

A.2.3 Proof of Proposition 1

In this section, we provide a proof for Proposition 1.

For any player i following a no-regret, no-violation algorithm the constrained contextual regret after T rounds of gameplay is upper bounded by:

$$R^T \geq \max_{\pi_i \in \Pi_i} \sum_{t=1}^{T} r(\pi_i(z^t), a_{-i}^t, z^t) - \sum_{t=1}^{T} r(a_i^t, a_{-i}^t, z^t) \tag{59}$$

for all $\pi_i \in \Pi_i$ such that $g_{i,m}(\pi_i(z^t)) \leq 0$ for all $m \in [M]$ and $t \in [T]$.

By definition of the empirical joint policy ρ^T at round T (see equation (12)) the following holds:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{a\sim\rho_{z^t}^T}\left[r_i(a_i, a_{-i}, z^t)\right] &= \frac{1}{T}\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} \rho_{z^t}^T(a) r_i(a_i, a_{-i}, z^t) \\ &\frac{1}{T}\sum_{t=1}^{T}\sum_{z\in\mathcal{Z}^T}\sum_{t:z^t=z} r_i(a_i^t, a_{-i}^t, z^t) \\ &= \frac{1}{T}\sum_{t=1}^{T} r_i(a_i^t, a_{-i}^t, z^t).\end{aligned} \tag{60}$$

Following the same argumentation it further holds that:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{a\sim\rho_{z^t}^T}\left[r_i(\pi_i(z^t), a_{-i}, z^t)\right] = \frac{1}{T}\sum_{t=1}^{T} r_i(\pi_i(z^t), a_{-i}^t, z^t). \tag{61}$$

By combining equation (59)-(61) and by definition of ϵ in Proposition 1 we obtain the following inequality:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{a\sim\rho_{z^t}^T}\left[r_i(a_i, a_{-i}, z^t)\right] &\geq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{a\sim\rho_{z^t}^T}\left[r_i(\pi_i(z^t), a_{-i}, z^t)\right] - \frac{R^T}{T} \\ &\geq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{a\sim\rho_{z^t}^T}\left[r_i(\pi_i(z^t), a_{-i}, z^t)\right] - \epsilon,\end{aligned}$$

for all $\pi_i \in \Pi_i$ such that $g_{i,m}(\pi_i(z^t)) \leq 0$ for all $m \in [M]$ and $t \in [T]$.

To prove that ρ^T is an ϵ-c.z.CCE it is left to show that the time-averaged expected constraint violations are bounded by ϵ. For any player i following a no-regret, no-violation algorithm the cumulative constraint violations after T rounds of game play are upper bounded by:

$$\mathcal{V}_{i,m} \geq \sum_{t=1}^{T} [g_{i,m}(a_i^t, z^t)]_+ \tag{62}$$

for all $m \in [M]$. Furthermore, it holds that:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{a\sim\rho_{z^t}^T}\left[[g_{i,m}(a_i, z^t)]_+\right] &= \frac{1}{T}\sum_{t=1}^{T} \rho_{z^t}^T(a)[g_{i,m}(a_i, z^t)]_+ \\ &= \frac{1}{T}\sum_{t=1}^{T}\sum_{z\in\mathcal{Z}^T}\sum_{t:z^t=z} [g_{i,m}(a_i^t, z^t)]_+ \\ &= \frac{1}{T}\sum_{t=1}^{T} [g_{i,m}(a_i^t, z^t)]_+\end{aligned} \tag{63}$$

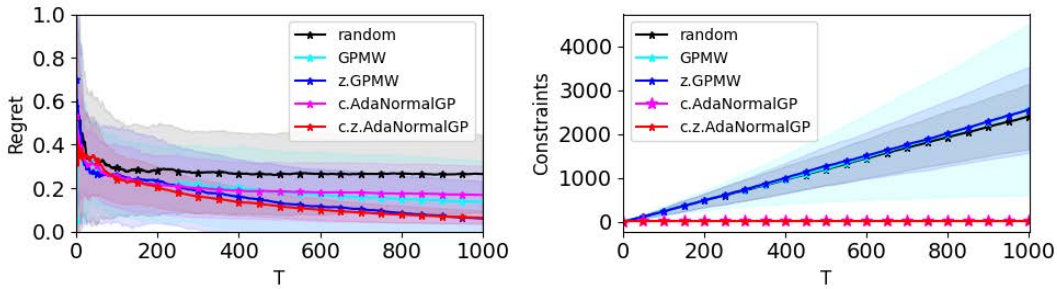

Figure 3: Regret and cumuluative constraint violations for players "random", "GPMW", "c.GPMW", "c.AdaNormalGP", and "c.z.AdaNormalGP". Shaded areas represent ± one standard deviation.

By combining equation (62) and (63) and by definition of ϵ in Proposition 1 we obtain the following inequality:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{a\sim\rho_{z^t}}\left[[g_{i,m}(a_i,z^t)]_+\right] \leq \frac{\mathcal{V}_{i,m}^T}{T} \leq \epsilon, \quad \forall m \in [M]$$

which completes the proof.

A.3 Supplementary Material for Section 3

In the following, we demonstrate c.z.AdaNormalGP on a random N-player game with unknown constraints. We furthermore, provide a complete description of our experimental setup for the application of controller design of robots presented in Section 3.1.

A.3.1 Random N-player games

We consider a repeated game between three players with action sets $\mathcal{A}_1 = \mathcal{A}_2 = \mathcal{A}_3 = \{0, 1, \ldots, K-1\}$ and contextual information of the form $\mathcal{Z} = \{1, \ldots, Z\}$. For each player, we simulate 10 different reward functions $r_i : \mathcal{A} \times \mathcal{Z} \to [0, 1]$ as the posterior mean computed from 10 samples of a GP, i.e., $r_i \sim \mathcal{GP}(0, k_r(\cdot,\cdot))$. Kernel $k_r = k_a * k_z$ is a composite kernel that can encode different dependences of r_i on $a \in \mathcal{A}$ and $z \in \mathcal{Z}$. Furthermore, the actions of each player are subject to a constraint $g_i(a_i) \leq 0$ for all $a_i \in \mathcal{A}_i$. We simulate 10 different constraint functions $g_i : \mathcal{A}_i \to \mathbb{R}$ as the posterior mean computed from 10 samples of a GP, i.e., $g_i \sim \mathcal{GP}(0, k_g(\cdot,\cdot))$. Concretely, we set $K = 7$, $Z = 5$, and $T = 1000$, the kernels are set as $k_a = k_{SE}$ with lengthscale $l = 2$ and $k_z = k_g = k_{SE}$ with lengthscale $l = 0.5$ and the noise is given by $\epsilon_i^t \sim \mathcal{N}(0, 1)$.

For every game, we consider player 1 to either play 1) random, i.e., selects actions uniformly at random or according to 2) GPMW Sessa et al. (2019), a no-regret algorithm, 3) z.GPMW Sessa et al. (2020), a contextual no-regret algorithm, 4) c.AdaNormalGP, a version of c.z.AdaNormalGP which ignores contextual information, and 5) c.z.AdaNormalGP. Player 2 and player 3 are random players. The regret obtained by "GPMW", "z.GPMW", "c.AdaNormalGP" and "c.z.AdaNormalGP" decreases over T, shown in Figure 3 (*left*). In particular, the two contextual no-regret algorithms "z.GPMW" and "c.z.AdaNormalGP" obtain smaller regret compared to their non-contextual counterparts since they exploit contextual information in their update rule. In addition, "c.AdaNormalGP" and "c.z.AdaNormalGP" learn the constraints on the action set and thus their cumulative constraint violations do not increase after a few rounds of gameplay, whereas "GPMW" and "z.GPMW" keep violating the constraints on the action set, shown in Figure 3 (*right*).