A Proof of the strong duality (4)

In this section, we explain why the equalities (4) hold when the problem $(r, \boldsymbol{c}, \boldsymbol{B}', \nu)$ is feasible for some $\boldsymbol{B}' < \boldsymbol{B}$. We restate first these inequalities for the convenience of the reader:

$$
\begin{aligned}
\textsc{opt}(r, \boldsymbol{c}, \boldsymbol{B}) &= \sup_{\boldsymbol{\pi}:\mathcal{X}\to\mathcal{P}(\mathcal{A})} \inf_{\boldsymbol{\lambda}\geqslant\boldsymbol{0}} \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(\boldsymbol{X}, a)\,\pi_a(\boldsymbol{X}) + \left\langle \boldsymbol{\lambda},\, \boldsymbol{B} - \sum_{a\in\mathcal{A}} \boldsymbol{c}(\boldsymbol{X}, a)\,\pi_a(\boldsymbol{X}) \right\rangle\right] \\
&\overset{(\star)}{=} \min_{\boldsymbol{\lambda}\geqslant\boldsymbol{0}} \sup_{\boldsymbol{\pi}:\mathcal{X}\to\mathcal{P}(\mathcal{A})} \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \pi_a(\boldsymbol{X})\Big(r(\boldsymbol{X}, a) - \big\langle \boldsymbol{c}(\boldsymbol{X}, a) - \boldsymbol{B},\, \boldsymbol{\lambda}\big\rangle\Big)\right] \\
&= \min_{\boldsymbol{\lambda}\geqslant\boldsymbol{0}} \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\max_{a\in\mathcal{A}}\Big\{ r(\boldsymbol{X}, a) - \big\langle \boldsymbol{c}(\boldsymbol{X}, a) - \boldsymbol{B},\, \boldsymbol{\lambda}\big\rangle \Big\}\right].
\end{aligned}
$$

We deal here with a linear program, with primal variables in functional space $\mathcal{X} \to \mathbb{R}^{\mathcal{A}}$ and dual variables in the finite-dimensional space $\mathbb{R}^d$.

The first and third equalities are straightforward. The first equality holds because in $\textsc{opt}(r, \boldsymbol{c}, \boldsymbol{B})$, we want to consider only policies with expected cost smaller than or equal to $\boldsymbol{B}$; the infimum over $\lambda \geqslant \boldsymbol{0}$ equals $-\infty$ for policies with expected costs strictly larger than $\boldsymbol{B}$ and is achieved at $\boldsymbol{\lambda} = 0$ otherwise. The third inequality follows from identifying that for a given $\boldsymbol{\lambda}$, the best policy may be defined pointwise as the argument of the maximum written in the expectation. Thus, only the middle equality $(\star)$ deserves a proof. We obtain it by applying a general theorem of strong duality (which requires feasibility for slightly smaller cost constraints).

Statement of a general theorem of strong duality. We restate a result extracted from the monograph by Luenberger [1969]. It relies on the dual functional φ, whose expression we recall below.

Theorem 2 (stated as Theorem 1 in Section 8.6, page 224 in Luenberger, 1969)**.** *Let f be a real-valued convex functional defined on a convex subset Ω of a vector space X, and let G be a convex mapping of X into a normed space Z. Suppose there exists an x_1 such that $G(x_1) < \theta$ and that $\mu_0 = \inf\big\{f(x) :\ G(x) \leqslant \theta,\ x \in \Omega\big\}$ is finite. Then*

$$
\inf_{\substack{G(x)\leqslant\theta \\ x\in\Omega}} f(x) = \max_{z^\star\geqslant\theta} \varphi(z^\star)
$$

and the maximum on the right is achieved by some $z_0^\star \geqslant \theta$.

The dual function φ is defined as follows (Luenberger, 1969, pages 215, 223, and 224, which we quote). We denote by P a positive convex cone $P \subset Z$, and let $P^\star = \big\{z^\star \in Z^\star :\ \forall z \in P,\ \langle z,\, z^\star\rangle \geqslant 0\big\}$ be its corresponding positive cone in the dual space $Z^\star$. The dual functional is defined, for each element $z^\star \in P^\star$, as

$$
\varphi(z^\star) = \inf_{x\in\Omega}\big\{f(x) + \langle G(x),\, z^\star\rangle\big\}\,.
$$

Application. To prove the middle equality $(\star)$, we apply Theorem 2 with the vector space X of all functions $\mathcal{X} \to \mathbb{R}^{\mathcal{A}}$, its convex subset Ω formed by functions $\mathcal{X} \to \mathcal{P}(\mathcal{A})$, and note that the function f is actually linear in our situation:

$$
f(\pi) = -\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(\boldsymbol{X}, a)\,\pi_a(\boldsymbol{X})\right].
$$

(We take a $-$ sign to match the form of the minimization problem over π in Theorem 2.) We take $Z = \mathbb{R}^d = Z^\star$ with positive cones $P = P^\star = [0, +\infty)^d$. The mapping G is a linear mapping from X to $Z = \mathbb{R}^d$:

$$
G(\pi) = \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(\boldsymbol{X}, a)\,\pi_a(\boldsymbol{X})\right] - \boldsymbol{B}\,,
$$

and we take $\theta = \boldsymbol{0}$. We have that $\mu_0 = \textsc{opt}(r, \boldsymbol{c}, \boldsymbol{B})$ is indeed finite. We note that the condition "$(r, \boldsymbol{c}, \boldsymbol{B}', \nu)$ is feasible for some $\boldsymbol{B}' < \boldsymbol{B}$" is required to apply the theorem. The result follows from noting that we exactly have, for $z = \boldsymbol{\lambda} \in P^\star$,

$$
\varphi(\boldsymbol{\lambda}) = \inf_{\boldsymbol{\pi}:\mathcal{X}\to\mathcal{P}(\mathcal{A})} \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[-\sum_{a\in\mathcal{A}} r(\boldsymbol{X}, a)\,\pi_a(\boldsymbol{X}) + \sum_{a\in\mathcal{A}} \big\langle \boldsymbol{c}(\boldsymbol{X}, a) - \boldsymbol{B},\, \boldsymbol{\lambda}\big\rangle \pi_a(\boldsymbol{X})\right].
$$

B Proofs of Lemmas 1 and 2

In both proofs, we will use the following deviation inequalities, holding on an event $\mathcal{E}_{\text{H-Az}} \cap \mathcal{E}_\beta$ of probability at least $1-\delta/2$, where the event $\mathcal{E}_\beta$ is defined in Assumption 2 with a confidence level $1-\delta/4$ and the event $\mathcal{E}_{\text{H-Az}}$ is defined below: on $\mathcal{E}_{\text{H-Az}} \cap \mathcal{E}_\beta$,

$$\forall\, 1 \leqslant t \leqslant T, \qquad \sum_{\tau=1}^{t} \boldsymbol{c}_\tau \leqslant \big(\alpha_{t,\delta/4} + 2\beta_{t,\delta/4}\big)\mathbf{1} + \sum_{\tau=1}^{t} \widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) \tag{7}$$

$$\text{and} \qquad \sum_{\tau=1}^{t} r_\tau \geqslant -\big(\alpha_{t,\delta/4} + 2\beta_{t,\delta/4}\big) + \sum_{\tau=1}^{t} \widehat{r}^{\,\mathrm{ucb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau)\,, \tag{8}$$

where $\beta_{t,\delta/4}$ is also defined in Assumption 2 and

$$\alpha_{t,\delta/4} = \sqrt{2t \ln \frac{2(d+1)T}{\delta/4}}\,.$$

These inequalities come, on the one hand, by the Hoeffding-Azuma inequality applied $(d+1)T$ times on the range $[-1,1]$: it ensures that on an event $\mathcal{E}_{\text{H-Az}}$ with probability at least $1-\delta/4$, for all $1 \leqslant t \leqslant T$,

$$\left\| \sum_{\tau=1}^{t} \boldsymbol{c}_\tau - \sum_{\tau=1}^{t} \boldsymbol{c}(\boldsymbol{x}_\tau, a_\tau) \right\|_\infty \leqslant \alpha_{t,\delta/4}\,\mathbf{1} \qquad \text{and} \qquad \left| \sum_{\tau=1}^{t} r_\tau - \sum_{\tau=1}^{t} r(\boldsymbol{x}_\tau, a_\tau) \right| \leqslant \alpha_{t,\delta/4}$$

(where we denoted by $\|\cdot\|_\infty$ the supremum norm). On the other hand, Assumption 2 and the clipping (3) entail, in particular, that on an event $\mathcal{E}_\beta$ of probability at least $1-\delta/4$, for all $1 \leqslant t \leqslant T$,

$$\sum_{\tau=1}^{t} \boldsymbol{c}(\boldsymbol{x}_\tau, a_\tau) \leqslant \sum_{\tau=1}^{t} \big(\widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) + 2\varepsilon_{\tau-1}(\boldsymbol{x}_\tau, a_\tau, \delta/4)\,\mathbf{1}\big) = 2\beta_{t,\delta/4}\,\mathbf{1} + \sum_{\tau=1}^{t} \widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau)$$

$$\text{and (similarly)} \qquad \sum_{\tau=1}^{t} r(\boldsymbol{x}_\tau, a_\tau) \geqslant -2\beta_{t,\delta/4} + \sum_{\tau=1}^{t} \widehat{r}^{\,\mathrm{ucb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau)\,.$$

B.1 Proof of Lemma 1

For $\tau \geqslant 1$, by definition of $\boldsymbol{\lambda}_\tau$ in Box B,

$$\begin{aligned} \boldsymbol{\lambda}_\tau - \boldsymbol{\lambda}_{\tau-1} &= \Big(\boldsymbol{\lambda}_{\tau-1} + \gamma\big(\widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - (\boldsymbol{B} - b\mathbf{1})\big)\Big)_+ - \boldsymbol{\lambda}_{\tau-1} \\ &\geqslant \gamma\big(\widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - (\boldsymbol{B} - b\mathbf{1})\big)\,. \end{aligned}$$

For $t \geqslant 1$, as $\boldsymbol{\lambda}_0 = \mathbf{0}$ and $\boldsymbol{\lambda}_t \geqslant \mathbf{0}$, we get, after telescoping and taking non-negative parts,

$$\boldsymbol{\lambda}_t \geqslant \gamma \left(\sum_{\tau=1}^{t} \big(\widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - (\boldsymbol{B} - b\mathbf{1})\big) \right)_+,$$

$$\text{thus} \qquad \left\| \left(\sum_{\tau=1}^{t} \big(\widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - (\boldsymbol{B} - b\mathbf{1})\big) \right)_+ \right\| \leqslant \frac{\|\boldsymbol{\lambda}_t\|}{\gamma}\,. \tag{9}$$

Up to the deviation terms (7), $\|\boldsymbol{\lambda}_t\|/\gamma$ bounds how larger the cost constraints till round t are from $t(\boldsymbol{B} - b\mathbf{1})$. Most of the rest of the proof, namely, the Steps 1–4 below, thus focus on upper bounding the $\|\boldsymbol{\lambda}_t\|$, while Step 5 will collect all bounds together and conclude.

Step 1: Gradient-descent analysis. We introduce

$$\Delta\widehat{\boldsymbol{c}}_\tau = \widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - (\boldsymbol{B} - b\mathbf{1}) \tag{10}$$

and prove the following deterministic inequality: for all $1 \leqslant t \leqslant T$,

$$\forall\, \boldsymbol{\lambda} \geqslant \mathbf{0}, \qquad \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}\|^2 \leqslant \|\boldsymbol{\lambda}\|^2 + 4d\,\gamma^2 t + 2\gamma \sum_{\tau=1}^{t} \big\langle \Delta\widehat{\boldsymbol{c}}_\tau,\, \boldsymbol{\lambda}_{\tau-1} - \boldsymbol{\lambda} \big\rangle\,. \tag{11}$$

To do so, we proceed as is classical in (projected) gradient-descent analyses; see, e.g., Zinkevich [2003]. Namely, for all $1 \leqslant \tau \leqslant T$,

$$\begin{aligned}
2\bigl\langle -\Delta\widehat{\boldsymbol{c}}_\tau,\, \boldsymbol{\lambda}_{\tau-1} - \boldsymbol{\lambda} \bigr\rangle &= \frac{1}{\gamma}\Bigl(\|\boldsymbol{\lambda}_{\tau-1} - \boldsymbol{\lambda}\|^2 + \bigl\|\, \gamma\Delta\widehat{\boldsymbol{c}}_\tau \bigr\|^2 - \bigl\|\, \boldsymbol{\lambda}_{\tau-1} - \boldsymbol{\lambda} + \gamma\Delta\widehat{\boldsymbol{c}}_\tau \bigr\|^2 \Bigr) \\
&= \gamma \bigl\|\, \Delta\widehat{\boldsymbol{c}}_\tau \bigr\|^2 + \frac{1}{\gamma}\Bigl(\|\boldsymbol{\lambda}_{\tau-1} - \boldsymbol{\lambda}\|^2 - \bigl\|\, \boldsymbol{\lambda}_{\tau-1} + \gamma\Delta\widehat{\boldsymbol{c}}_\tau - \boldsymbol{\lambda} \bigr\|^2 \Bigr) \\
&\leqslant \gamma \bigl\|\, \Delta\widehat{\boldsymbol{c}}_\tau \bigr\|^2 + \frac{1}{\gamma}\Bigl(\|\boldsymbol{\lambda}_{\tau-1} - \boldsymbol{\lambda}\|^2 - \|\boldsymbol{\lambda}_{\tau} - \boldsymbol{\lambda}\|^2 \Bigr),
\end{aligned}$$

where the inequality comes from the definition $\boldsymbol{\lambda}_\tau = \bigl(\boldsymbol{\lambda}_{\tau-1} + \gamma\Delta\widehat{\boldsymbol{c}}_\tau\bigr)_+$ and the fact that

$$\forall\, x \in \mathbb{R}, \ \ \forall\, y \geqslant 0, \qquad \bigl|(x)_+ - y\bigr| \leqslant |x - y|$$

(which may be proved by distinguishing the cases $x \leqslant 0$ and $x \geqslant 0$).

We note that

$$\boldsymbol{B} - b\boldsymbol{1} \in [0,1]^d \quad \text{and} \quad \widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) \in [-1,1]^d\,, \qquad \text{so that} \qquad \bigl\|\, \Delta\widehat{\boldsymbol{c}}_\tau \bigr\|^2 \leqslant 4d\,.$$

Collecting all bounds above, we get, after summation and telescoping,

$$\sum_{\tau=1}^{t} \bigl\langle -\Delta\widehat{\boldsymbol{c}}_\tau,\, \boldsymbol{\lambda}_{\tau-1} - \boldsymbol{\lambda} \bigr\rangle \leqslant 2d\,\gamma t + \frac{1}{2\gamma}\Bigl(\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}\|^2 - \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}\|^2 \Bigr).$$

Rearranging and substituting $\boldsymbol{\lambda}_0 = \boldsymbol{0}$ yields the claimed inequality (11).

Step 2: Relating estimated costs (and rewards) to true conditional expectations. In this part, we upper bound the right-hand side of (11) by showing that on the event $\mathcal{E}_\beta$,

$$\begin{gathered}
\forall\, \boldsymbol{\lambda} \geqslant \boldsymbol{0}, \quad \forall\, 1 \leqslant t \leqslant T, \\
\sum_{\tau=1}^{t} \bigl\langle \Delta\widehat{\boldsymbol{c}}_\tau,\, \boldsymbol{\lambda}_{\tau-1} - \boldsymbol{\lambda} \bigr\rangle \leqslant 2\bigl(1 + \|\boldsymbol{\lambda}\|_1\bigr)\beta_{t,\delta/4} + \sum_{\tau=1}^{t} \bigl(g_\tau(\boldsymbol{\lambda}) - g_\tau(\boldsymbol{\lambda}_{\tau-1}) \bigr)\,, \qquad (12) \\
\text{where} \qquad g_\tau(\boldsymbol{\lambda}) = \max_{a \in \mathcal{A}} \Bigl\{ r(\boldsymbol{x}_\tau, a) - \bigl\langle \boldsymbol{c}(\boldsymbol{x}_\tau, a) - (\boldsymbol{B} - b\boldsymbol{1}),\, \boldsymbol{\lambda} \bigr\rangle \Bigr\}
\end{gathered}$$

and where we recall that $\|\cdot\|_1$ denotes the ℓ_1–norm.

Adding and subtracting $\widehat{r}^{\,\mathrm{ucb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau)$, here, we deal with

$$\begin{aligned}
\sum_{\tau=1}^{t} \bigl\langle \Delta\widehat{\boldsymbol{c}}_\tau,\, \boldsymbol{\lambda}_{\tau-1} - \boldsymbol{\lambda} \bigr\rangle = \sum_{\tau=1}^{t} &\Bigl(\widehat{r}^{\,\mathrm{ucb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - \bigl\langle \widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - (\boldsymbol{B} - b\boldsymbol{1}),\, \boldsymbol{\lambda} \bigr\rangle \Bigr) \\
- \sum_{\tau=1}^{t} &\Bigl(\widehat{r}^{\,\mathrm{ucb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - \bigl\langle \widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - (\boldsymbol{B} - b\boldsymbol{1}),\, \boldsymbol{\lambda}_{\tau-1} \bigr\rangle \Bigr).
\end{aligned}$$

Now, for each τ, by (3) and the fact that $\boldsymbol{\lambda} \geqslant \boldsymbol{0}$,

$$\begin{aligned}
&\widehat{r}^{\,\mathrm{ucb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - \bigl\langle \widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - (\boldsymbol{B} - b\boldsymbol{1}),\, \boldsymbol{\lambda} \bigr\rangle \\
&\leqslant r(\boldsymbol{x}_\tau, a_\tau) + 2\varepsilon_{\tau-1}(\boldsymbol{x}_\tau, a_\tau, \delta/4) - \bigl\langle \boldsymbol{c}(\boldsymbol{x}_\tau, a_\tau) - (\boldsymbol{B} - b\boldsymbol{1}),\, \boldsymbol{\lambda} \bigr\rangle + 2\varepsilon_{\tau-1}(\boldsymbol{x}_\tau, a_\tau, \delta/4)\, \|\boldsymbol{\lambda}\|_1 \\
&\leqslant g_\tau(\boldsymbol{\lambda}) + 2\bigl(1 + \|\boldsymbol{\lambda}\|_1\bigr)\varepsilon_{\tau-1}(\boldsymbol{x}_\tau, a_\tau, \delta/4)\,.
\end{aligned}$$

On the other hand, by definition of a_τ for the equality, and then by the other inequalities in (3), for each τ, and the fact that $\boldsymbol{\lambda}_{\tau-1} \geqslant \boldsymbol{0}$,

$$\begin{aligned}
&\widehat{r}^{\,\mathrm{ucb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - \bigl\langle \widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - (\boldsymbol{B} - b\boldsymbol{1}),\, \boldsymbol{\lambda}_{\tau-1} \bigr\rangle \qquad (13) \\
&= \max_{a \in \mathcal{A}} \Bigl\{ \widehat{r}^{\,\mathrm{ucb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a) - \bigl\langle \widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a) - (\boldsymbol{B} - b\boldsymbol{1}),\, \boldsymbol{\lambda}_{\tau-1} \bigr\rangle \Bigr\} \\
&\geqslant \max_{a \in \mathcal{A}} \Bigl\{ r(\boldsymbol{x}_\tau, a) - \bigl\langle \boldsymbol{c}(\boldsymbol{x}_\tau, a) - (\boldsymbol{B} - b\boldsymbol{1}),\, \boldsymbol{\lambda}_{\tau-1} \bigr\rangle \Bigr\} = g_\tau(\boldsymbol{\lambda}_{\tau-1})\,.
\end{aligned}$$

Collecting the two series of bounds concludes this part.

Step 3: Application of a Bernstein-Freedman inequality. We recall that we denoted by $\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}$ the optimal dual variable in (4) for $\mathrm{OPT}(r, \boldsymbol{c}, \boldsymbol{B} - b\mathbf{1})$; it exists because we assumed feasibility of a problem with average cost constraints $\boldsymbol{B}' < \boldsymbol{B} - b\mathbf{1}$.

We now upper bound the sum appearing in the right hand side of (12) at $\boldsymbol{\lambda} = \boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}$ by showing that on an event $\mathcal{E}_{\text{Bern-}\boldsymbol{c}}$ of probability at least $1 - \delta/4$, for all $1 \leqslant t \leqslant T$,

$$\sum_{\tau=1}^{t} \bigl(g_\tau(\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}) - g_\tau(\boldsymbol{\lambda}_{\tau-1})\bigr) \leqslant (1 + 2\Lambda_t)\sqrt{2t \ln \frac{T^2}{\delta/4}} + 2K_t \ln \frac{T^2}{\delta/4}\,, \tag{14}$$

$$\text{where} \qquad \Lambda_t = \max_{1 \leqslant \tau \leqslant t} \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}} - \boldsymbol{\lambda}_{\tau-1}\|_1 \qquad \text{and} \qquad K_t = 4\bigl(1 + \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|_1 + 2d\gamma t\bigr)\,.$$

We will do so by applying a version of the Bernstein-Freedman inequality for martingales stated in Cesa-Bianchi et al. [2005, Corollary 16] involving the sum of the conditional variances (and not only a deterministic bound thereon); it is obtained via peeling based on the classic version of Bernstein's inequality (Freedman, 1975). We restate it here for the convenience of the reader (after applying some simple boundings).

Lemma 4 (a version of the Bernstein-Freedman inequality by Cesa-Bianchi et al., 2005)**.** *Let $X_1, X_2, \ldots$ be a martingale difference with respect to the filtration $\mathcal{F} = (\mathcal{F}_s)_{s \geqslant 0}$ and with increments bounded in absolute values by K. For all $t \geqslant 1$, let*

$$\mathfrak{S}_t = \sum_{\tau=1}^{t} \mathbb{E}\bigl[X_\tau^2 \,\big|\, \mathcal{F}_{\tau-1}\bigr]$$

denote the sum of the conditional variances of the first t increments. Then, for all $\delta \in (0, 1)$ and all $t \geqslant 1$, with probability at least $1 - \delta$,

$$\sum_{\tau=1}^{t} X_\tau \leqslant \sqrt{2\mathfrak{S}_t \, \ln \frac{t}{\delta}} + 2K \ln \frac{t}{\delta}\,.$$

We introduce, for all $\boldsymbol{\lambda} \geqslant \mathbf{0}$, the common expectation of the $g_\tau(\boldsymbol{\lambda})$, namely,

$$G(\boldsymbol{\lambda}) = \mathbb{E}\bigl[g_\tau(\boldsymbol{\lambda})\bigr] = \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\max_{a\in\mathcal{A}}\Bigl\{r(\boldsymbol{X}, a) - \bigl\langle \boldsymbol{c}(\boldsymbol{X}, a) - (\boldsymbol{B} - b\mathbf{1}),\, \boldsymbol{\lambda}\bigr\rangle\Bigr\}\right],$$

and consider the martingale increments

$$X_\tau = \bigl(g_\tau(\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}) - g_\tau(\boldsymbol{\lambda}_{\tau-1})\bigr) - \bigl(G(\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}) - G(\boldsymbol{\lambda}_{\tau-1})\bigr)\,.$$

As $\boldsymbol{B} - b\mathbf{1} \in [0, 1]^d$ and $\boldsymbol{c}$ takes values in $[-1, 1]^d$, for all $\boldsymbol{x} \in \mathcal{X}$, all $a \in \mathcal{A}$, and all $\boldsymbol{v} \in \mathbb{R}^d$, the quantities $r(\boldsymbol{x}, a) - \bigl\langle \boldsymbol{c}(\boldsymbol{x}, a) - (\boldsymbol{B} - b\mathbf{1}),\, \boldsymbol{v}\bigr\rangle$ take absolute values smaller than $1 + 2\|\boldsymbol{v}\|_1$. Using that a difference of maxima is smaller than the maximum of the differences, we get, in particular,

$$\begin{aligned} &\bigl|g_\tau(\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}) - g_\tau(\boldsymbol{\lambda}_{\tau-1})\bigr| \leqslant 1 + 2\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}} - \boldsymbol{\lambda}_{\tau-1}\|_1 \quad \text{a.s.} && (15)\\ \text{and} \qquad &\bigl|G(\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}) - G(\boldsymbol{\lambda}_{\tau-1})\bigr| \leqslant 1 + 2\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}} - \boldsymbol{\lambda}_{\tau-1}\|_1\,. \end{aligned}$$

Now, given the update step in Step 3 of Box B, we have the deterministic bound $\|\boldsymbol{\lambda}_t\|_1 \leqslant 2d\gamma t$ for all $1 \leqslant t \leqslant T$. Therefore, by a triangle inequality, the martingale increments are bounded in absolute values by K_t, as

$$2\Bigl(1 + 2\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|_1 + 1 + 2\max_{\tau \leqslant t}\|\boldsymbol{\lambda}_{\tau-1}\|_1\Bigr) \leqslant 4\bigl(1 + \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|_1 + 2d\gamma t\bigr) = K_t\,.$$

The conditional variance of X_τ is smaller than the squared half-width of the conditional range (Popoviciu's inequality on variances); in particular, (15) thus entails

$$\mathfrak{S}_t \leqslant \sum_{\tau=1}^{t} \bigl(1 + 2\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}} - \boldsymbol{\lambda}_{\tau-1}\|_1\bigr)^2 \leqslant (1 + 2\Lambda_t)^2 t\,.$$

We now get the claimed inequalities (14) first by noting that by (4), for all $\tau \leqslant T$,

$$G(\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}) - G(\boldsymbol{\lambda}_{\tau-1}) \leqslant 0\,,$$

and second, by applying Lemma 4 for each $1 \leqslant t \leqslant T$, using a confidence level $\delta/(4T)$. By a union bound, this indeed defines an event $\mathcal{E}_{\text{Bern-}\boldsymbol{c}}$ of probability at least $1 - \delta/4$.

Step 4: Induction to bound the Λ_t. In this step, we show by induction that, with high probability, the norms $\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}} - \boldsymbol{\lambda}_t\|$ satisfy the bound (18) stated below.

To do so, we combine the outcomes of Steps 1–3 and obtain that on the event $\mathcal{E}_\beta \cap \mathcal{E}_{\text{Bern-}\boldsymbol{c}}$, for all $1 \leqslant t \leqslant T$,

$$\begin{aligned}
&\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}} - \boldsymbol{\lambda}_t\|^2 \\
&\leqslant \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|^2 + 4d\gamma^2 t + 2\gamma \left(2\big(1 + \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|_1\big)\beta_{t,\delta/4} + \sum_{\tau=1}^{t} \big(g_\tau(\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}) - g_\tau(\boldsymbol{\lambda}_{\tau-1}) \big) \right) \\
&\leqslant \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|^2 + 4d\gamma^2 t + 4\gamma\big(1 + \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|_1\big)\beta_{t,\delta/4} + 2\gamma(1 + 2\Lambda_t)\sqrt{2t \ln \frac{T^2}{\delta/4}} + 4\gamma K_t \ln \frac{T^2}{\delta/4} \,,
\end{aligned}$$

where we recall that norms not indexed by a subscript are Euclidean norms, and

$$\Lambda_t = \max_{1 \leqslant \tau \leqslant t} \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}} - \boldsymbol{\lambda}_{\tau-1}\|_1 \qquad \text{and} \qquad K_t = 4\big(1 + \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|_1 + 2d\gamma t\big) \,.$$

We upper bound $\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|_1$ and Λ_t in terms of Euclidean norms,

$$\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|_1 \leqslant \sqrt{d}\,\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\| \qquad \text{and} \qquad \Lambda_t \leqslant \sqrt{d} \max_{1 \leqslant \tau \leqslant t} \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}} - \boldsymbol{\lambda}_{\tau-1}\| \,,$$

perform some crude boundings like $\sqrt{t} \leqslant \sqrt{T}$ and $\beta_{t,\delta/4} \leqslant \beta_{T,\delta/4}$, and obtain the following induction relationship: on the event $\mathcal{E}_\beta \cap \mathcal{E}_{\text{Bern-}\boldsymbol{c}}$, for all $1 \leqslant t \leqslant T$,

$$\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}} - \boldsymbol{\lambda}_t\|^2 \leqslant A + B\,t + C \max_{1 \leqslant \tau \leqslant t} \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}} - \boldsymbol{\lambda}_{\tau-1}\| \,, \tag{16}$$

where

$$\begin{aligned}
&A = \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|^2 + \gamma \left(4\big(1 + \sqrt{d}\,\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|\big)\beta_{T,\delta/4} + 2\sqrt{2T \ln \frac{T^2}{\delta/4}} + 16\big(1 + \sqrt{d}\,\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|\big)\ln \frac{T^2}{\delta/4} \right), \\
&B = 4d\gamma^2 + 4 \times 4 \times 2d\gamma^2 \ln \frac{T^2}{\delta/4} = \left(4 + 32 \ln \frac{T^2}{\delta/4}\right) d\gamma^2 \leqslant 36\, d\gamma^2 \ln \frac{T^2}{\delta/4} \,, \\
&C = 4\gamma \sqrt{2dT \ln \frac{T^2}{\delta/4}} \,.
\end{aligned}$$

We now show that (16) implies that on $\mathcal{E}_\beta \cap \mathcal{E}_{\text{Bern-}\boldsymbol{c}}$,

$$\forall\, 0 \leqslant t \leqslant T, \qquad \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}} - \boldsymbol{\lambda}_t\| \leqslant M \overset{\text{def}}{=} \frac{C}{2} + \sqrt{A + BT + \frac{C^2}{4}} \,. \tag{17}$$

Indeed, for $t = 0$, given that $\boldsymbol{\lambda}_0 = \mathbf{0}$, we have $\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}} - \boldsymbol{\lambda}_t\| = \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\| \leqslant \sqrt{A}$. Now, if the bound (17) is satisfied for all $0 \leqslant \tau \leqslant t$, where $0 \leqslant t \leqslant T-1$, then (16) implies that

$$\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}} - \boldsymbol{\lambda}_{t+1}\| \leqslant A + B\,(t+1) + CM \leqslant A + B\,T + CM \leqslant M^2 \,,$$

where the final inequality follows from the fact that (by definition of M, and this explains how we picked M)

$$M^2 - CM = \left(M - \frac{C}{2}\right)^2 + \frac{C^2}{4} = A + BT \,.$$

Below, we will make repeated uses of $\sqrt{x+y} \leqslant \sqrt{x} + \sqrt{y}$, of $xy \leqslant 2(x^2 + y^2)$, and of $\sqrt{x} \leqslant 1 + x$, for $x, y \geqslant 0$. From (17), we conclude that on the event $\mathcal{E}_\beta \cap \mathcal{E}_{\text{Bern-}\boldsymbol{c}}$,

$$\begin{aligned}
\forall\, 0 \leqslant t \leqslant T, \qquad \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}} - \boldsymbol{\lambda}_t\| &\leqslant \sqrt{A} + \sqrt{BT} + C \\
&= \sqrt{A} + 6\gamma \sqrt{dT \ln \frac{T^2}{\delta/4}} + 4\gamma \sqrt{2dT \ln \frac{T^2}{\delta/4}} \leqslant \sqrt{A} + 6\gamma \beta'_{T,\delta/4} \,,
\end{aligned}$$

where we denoted

$$\beta'_{T,\delta/4} = \max\left\{ \beta_{T,\delta/4},\ 2\sqrt{dT \ln \frac{T^2}{\delta/4}} \right\} \geqslant \ln \frac{T^2}{\delta/4} \,.$$

For the sake of readability, we may further bound $\sqrt{A}$ as follows (in some crude way):

$$\begin{aligned}
\sqrt{A} &\leqslant \|\boldsymbol{\lambda}^{\star}_{\boldsymbol{B}-b\mathbf{1}}\| + \sqrt{\gamma}\sqrt{4\bigl(1+\sqrt{d}\,\|\boldsymbol{\lambda}^{\star}_{\boldsymbol{B}-b\mathbf{1}}\|\bigr)\beta_{T,\delta/4} + 2\sqrt{2T\ln\frac{T^2}{\delta/4}} + 16\bigl(1+\sqrt{d}\,\|\boldsymbol{\lambda}^{\star}_{\boldsymbol{B}-b\mathbf{1}}\|\bigr)\ln\frac{T^2}{\delta/4}} \\
&\leqslant \|\boldsymbol{\lambda}^{\star}_{\boldsymbol{B}-b\mathbf{1}}\| + \sqrt{\gamma}\sqrt{22\beta'_{T,\delta/4} + 20\sqrt{d}\,\|\boldsymbol{\lambda}^{\star}_{\boldsymbol{B}-b\mathbf{1}}\|\beta'_{T,\delta/4}} \\
&\leqslant \|\boldsymbol{\lambda}^{\star}_{\boldsymbol{B}-b\mathbf{1}}\| + 1 + 6\gamma\beta'_{T,\delta/4} + \|\boldsymbol{\lambda}^{\star}_{\boldsymbol{B}-b\mathbf{1}}\| + 5\gamma\sqrt{d}\,\beta'_{T,\delta/4}\,,
\end{aligned}$$

where we used the facts that $\sqrt{20xy} = 2\sqrt{5xy} \leqslant x + 5y$ and $\sqrt{22x} = 2\sqrt{22x/4} \leqslant 1 + 6x$.

All in all, we proved that on the event $\mathcal{E}_\beta \cap \mathcal{E}_{\text{Bern-}\boldsymbol{c}}$,

$$\forall\, 0 \leqslant t \leqslant T, \qquad \|\boldsymbol{\lambda}^{\star}_{\boldsymbol{B}-b\mathbf{1}} - \boldsymbol{\lambda}_t\| \leqslant 2\|\boldsymbol{\lambda}^{\star}_{\boldsymbol{B}-b\mathbf{1}}\| + 17\gamma\sqrt{d}\,\beta'_{T,\delta/4} + 1\,. \tag{18}$$

Step 5: Conclusion. We combine the bound (18) with the bound (9) of Step 1 and the bound (7): on the intersection of events $\mathcal{E}_\beta \cap \mathcal{E}_{\text{Bern-}\boldsymbol{c}} \cap \mathcal{E}_{\text{H-Az}}$, which has a probability at least $1 - 3\delta/4$, for all $1 \leqslant t \leqslant T$,

$$\begin{aligned}
\left\| \left(\sum_{\tau=1}^{t} \boldsymbol{c}_\tau - t(\boldsymbol{B} - b\mathbf{1}) \right)_{+} \right\| &\leqslant \sqrt{d}\bigl(\alpha_{T,\delta/4} + 2\beta_{T,\delta/4}\bigr) + \frac{\|\boldsymbol{\lambda}_t\|}{\gamma} \\
&\leqslant \sqrt{d}\bigl(\alpha_{T,\delta/4} + 2\beta_{T,\delta/4}\bigr) + \frac{\|\boldsymbol{\lambda}^{\star}_{\boldsymbol{B}-b\mathbf{1}}\|}{\gamma} + \frac{\|\boldsymbol{\lambda}^{\star}_{\boldsymbol{B}-b\mathbf{1}} - \boldsymbol{\lambda}_t\|}{\gamma} \\
&\leqslant \frac{3\|\boldsymbol{\lambda}^{\star}_{\boldsymbol{B}-b\mathbf{1}}\| + 1}{\gamma} + \sqrt{d}\bigl(\alpha_{T,\delta/4} + 19\beta'_{T,\delta/4}\bigr)\,.
\end{aligned}$$

This entails in particular the stated result, given the definition of $\Upsilon_{T,\delta}$ as $\max\bigl\{\beta'_{T,\delta/4},\, \alpha_{T,\delta/4}\bigr\}$.

B.2 Proof of Lemma 2

The proof is similar to (but much simpler and shorter than) the one of Lemma 1 and borrows some of its arguments. We use throughout this section the notation introduced therein; we also define a new event $\mathcal{E}_{\text{Bern-}r}$ of probability at least $1 - \delta/4$.

We start from (8) and introduce the same $\Delta\widehat{\boldsymbol{c}}_t$ quantity as in (10): on the event $\mathcal{E}_{\text{H-Az}} \cap \mathcal{E}_\beta$, for all $1 \leqslant t \leqslant T$,

$$\begin{aligned}
\sum_{\tau=1}^{t} r_\tau &\geqslant -\bigl(\alpha_{t,\delta/4} + 2\beta_{t,\delta/4}\bigr) + \sum_{\tau=1}^{t} \widehat{r}^{\,\text{ucb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) \\
&\geqslant -\bigl(\alpha_{t,\delta/4} + 2\beta_{t,\delta/4}\bigr) + \sum_{\tau=1}^{t} \Bigl(\widehat{r}^{\,\text{ucb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - \bigl\langle \Delta\widehat{\boldsymbol{c}}_\tau,\, \boldsymbol{\lambda}_{\tau-1} \bigr\rangle \Bigr) + \sum_{\tau=1}^{t} \bigl\langle \Delta\widehat{\boldsymbol{c}}_\tau,\, \boldsymbol{\lambda}_{\tau-1} \bigr\rangle\,.
\end{aligned}$$

On the one hand, the result (11) with $\boldsymbol{\lambda} = \mathbf{0}$ exactly states that

$$\sum_{\tau=1}^{t} \bigl\langle \Delta\widehat{\boldsymbol{c}}_\tau,\, \boldsymbol{\lambda}_{\tau-1} \bigr\rangle \geqslant \frac{\|\boldsymbol{\lambda}_t\|^2}{2\gamma} - 2d\,\gamma t \geqslant -2d\,\gamma t\,.$$

On the other hand, the result (13) states that on $\mathcal{E}_\beta$, for all $1 \leqslant \tau \leqslant T$,

$$\begin{aligned}
\widehat{r}^{\,\text{ucb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - \bigl\langle \Delta\widehat{\boldsymbol{c}}_\tau,\, \boldsymbol{\lambda}_{\tau-1} \bigr\rangle &= \widehat{r}^{\,\text{ucb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - \bigl\langle \widehat{\boldsymbol{c}}^{\,\text{lcb}}_{\delta/4,\tau-1}(\boldsymbol{x}_\tau, a_\tau) - (\boldsymbol{B} - b\mathbf{1}),\, \boldsymbol{\lambda}_{\tau-1} \bigr\rangle \\
&\geqslant g_\tau(\boldsymbol{\lambda}_{\tau-1})\,.
\end{aligned}$$

A similar application of Lemma 4 as in the proof of Lemma 1 shows that on a new event $\mathcal{E}_{\text{Bern-}r}$ of probability at least $1 - \delta/4$, for all $1 \leqslant t \leqslant T$,

$$\sum_{\tau=1}^{t} g_\tau(\boldsymbol{\lambda}_{\tau-1}) \geqslant \sum_{\tau=1}^{t} G(\boldsymbol{\lambda}_{\tau-1}) - \sqrt{2\sum_{\tau=1}^{t} \bigl(1 + 2\|\boldsymbol{\lambda}_{\tau-1}\|_1\bigr)^2 \ln\frac{T^2}{\delta/4}} - 8(1 + 2\gamma t)\ln\frac{T^2}{\delta/4}\,.$$

We relate ℓ^1–norms to Euclidean norms, resort to a triangle inequality, and substitute (18) to get that on $\mathcal{E}_{\text{Bern-}\boldsymbol{c}}$, for all $1 \leqslant t \leqslant T$,

$$\begin{aligned}
-\sqrt{2\sum_{\tau=1}^{t}\bigl(1+2\|\boldsymbol{\lambda}_{\tau-1}\|_1\bigr)^2 \ln\frac{T^2}{\delta/4}} &\geqslant -\biggl(1+2\sqrt{d}\max_{0\leqslant\tau\leqslant t-1}\|\boldsymbol{\lambda}_{\tau-1}\|\biggr)\sqrt{2t\ln\frac{T^2}{\delta/4}} \\
&\geqslant -\biggl(1+2\sqrt{d}\,\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\| + 2\sqrt{d}\max_{0\leqslant\tau\leqslant t-1}\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}-\boldsymbol{\lambda}_{\tau-1}\|\biggr)\sqrt{2t\ln\frac{T^2}{\delta/4}} \\
&\geqslant -\biggl(8\sqrt{d}\,\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\| + 34\gamma d\,\beta'_{T,\delta/4} + 2\sqrt{d}+1\biggr)\sqrt{2t\ln\frac{T^2}{\delta/4}} \\
&\geqslant -\biggl(8\sqrt{d}\,\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\| + 34\gamma d\,\beta'_{T,\delta/4} + 4\sqrt{d}\biggr)\sqrt{2t\ln\frac{T^2}{\delta/4}} \\
&\geqslant -6\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|\,\beta'_{T,\delta/4} - 25\gamma\sqrt{d}\,\bigl(\beta'_{T,\delta/4}\bigr)^2 - 2\sqrt{2}\,\beta'_{T,\delta/4} \\
&\geqslant -6\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|\,\beta'_{T,\delta/4} - 28\gamma\sqrt{d}\,\bigl(\beta'_{T,\delta/4}\bigr)^2,
\end{aligned}$$

where we performed some crude boundings using the definition of $1 \leqslant \beta'_{t,\delta/4} \leqslant \beta'_{T,\delta/4}$. We also note that

$$8(1+2\gamma t)\ln\frac{T^2}{\delta/4} \leqslant 8\ln\frac{T^2}{\delta/4} + 4\gamma\bigl(\beta'_{T,\delta/4}\bigr)^2 .$$

By (4), we have $\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}-b\mathbf{1}) = G(\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}) \leqslant G(\boldsymbol{\lambda})$, for all $\boldsymbol{\lambda} \geqslant \mathbf{0}$. Also,

$$\begin{aligned}
&G(\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}) + b\,\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|_1 \\
&= \mathbb{E}_{\boldsymbol{X}\sim\nu}\biggl[\max_{a\in\mathcal{A}}\Bigl\{r(\boldsymbol{X},a) - \bigl\langle \boldsymbol{c}(\boldsymbol{X},a)-(\boldsymbol{B}-b\mathbf{1}),\,\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\bigr\rangle\Bigr\}\biggr] + b\,\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|_1 \\
&= \mathbb{E}_{\boldsymbol{X}\sim\nu}\biggl[\max_{a\in\mathcal{A}}\Bigl\{r(\boldsymbol{X},a) - \bigl\langle \boldsymbol{c}(\boldsymbol{X},a)-\boldsymbol{B},\,\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\bigr\rangle\Bigr\}\biggr] \geqslant \mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}),
\end{aligned}$$

where we used again (4). In particular,

$$\sum_{\tau=1}^{t} G(\boldsymbol{\lambda}_{\tau-1}) \geqslant t\,\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}) - t\,b\,\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|_1 \geqslant t\,\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}) - t\,b\sqrt{d}\,\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|_1 .$$

Collecting all bounds above and using the definition of $\Upsilon_{T,\delta}$ as $\max\bigl\{\beta'_{T,\delta/4},\,\alpha_{T,\delta/4}\bigr\}$ and the fact that $dt \leqslant (\Upsilon_{T,\delta})^2$, we proved the following. On $\mathcal{E}_{\text{H-Az}} \cap \mathcal{E}_{\beta} \cap \mathcal{E}_{\text{Bern-}\boldsymbol{c}} \cap \mathcal{E}_{\text{Bern-}r}$, which is indeed an event with probability at least $1-\delta$, for all $1 \leqslant t \leqslant T$,

$$\begin{aligned}
&\sum_{\tau=1}^{t} r_\tau \\
&\geqslant -3\Upsilon_{T,\delta} - 2d\,\gamma t + \sum_{\tau=1}^{t} G(\boldsymbol{\lambda}_{\tau-1}) - \sqrt{2\sum_{t=1}^{T}\bigl(1+2\|\boldsymbol{\lambda}_{t-1}\|_1\bigr)^2\ln\frac{T^2}{\delta/4}} - 8\ln\frac{T^2}{\delta/4} - 4\gamma\bigl(\beta'_{T,\delta/4}\bigr)^2 \\
&\geqslant t\,\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}) - \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\|\Bigl(t\,b\sqrt{d}+6\Upsilon_{T,\delta}\Bigr) - 36\gamma\sqrt{d}\,\bigl(\Upsilon_{T,\delta}\bigr)^2 - 8\ln\frac{T^2}{\delta/4} .
\end{aligned}$$

This entails in particular the stated result.

C Proof of Theorem 1

The proof is divided into three steps: on a favorable event $\mathcal{E}_{\text{meta}}$ of probability at least $1-\delta$,
(i) we bound by $\operatorname{ilog} T$ the index of the last regime achieved in Box C;

Algorithm 1: Pseudo-code for the Box C strategy

```
Input: number of rounds $T$; confidence level $1-\delta$; margin $b$ on the average constraints;
       estimation procedure and error functions $\varepsilon_t$ of Assumption 2; optimistic estimates (2)
Initialization: $T_0 = 1$; sequence $\gamma_k = 2^k/\sqrt{T}$ of step sizes;
       sequence $M_{T,\delta,k} = 4\sqrt{T} + 20\sqrt{d}\,\Upsilon_{T,\delta/(k+2)^2}$ of cost deviations
for $k \geqslant 0$ do                                                          // Box C part
    $\boldsymbol{\lambda}_{T_k-1} = \mathbf{0}$;
    for $t \geqslant T_k$ do                                                    // Box B part
        if $t = T$ then
            Terminate algorithm;
        end
        Observe the context $\boldsymbol{x}_t$;
        Pick an action
            $a_t \in \mathop{\arg\max}_{a\in\mathcal{A}} \Bigl\{ \widehat{r}^{\,\mathrm{ucb}}_{\delta,t-1}(\boldsymbol{x}_t,a) - \bigl\langle \widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta,t-1}(\boldsymbol{x}_t,a) - (\boldsymbol{B}-b\mathbf{1}),\, \boldsymbol{\lambda}_{t-1} \bigr\rangle \Bigr\}$;
        Observe the payoff $r_t$ and the costs $\boldsymbol{c}_t$;
        Compute $\boldsymbol{\lambda}_t = \Bigl( \boldsymbol{\lambda}_{t-1} + \gamma_k \bigl( \widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta,t-1}(\boldsymbol{x}_t,a_t) - (\boldsymbol{B}-b\mathbf{1}) \bigr) \Bigr)_+$;      // make PGD update
        Compute the estimates $\widehat{r}^{\,\mathrm{ucb}}_{\delta,t}$ and $\widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta,t}$;
        if $\left\Vert \left( \sum_{\tau=T_k}^{t} \boldsymbol{c}_\tau - (t-T_k+1)\,(\boldsymbol{B}-b\mathbf{1}) \right)_+ \right\Vert > M_{T,\delta,k}$ then      // Box C part
            Break inner for loop;                       // finish phase $k$ and move to phase $k+1$
            $T_{k+1} = t+1$;                            // record beginning of phase $k+1$
        end
    end
end
```

(ii) we bound by $(1+\mathrm{ilog}\,T)\sqrt{T}$ the excess cumulative costs with respect to $T(\boldsymbol{B}-b_T\mathbf{1})$ and deduce that the cumulative costs are smaller than B;
(iii) we provide a regret bound, by summing the bounds guaranteed by Lemma 1 over each regime.

The favorable event $\mathcal{E}_{\text{meta}}$ is defined as follows. By construction, and thanks to the assumption of feasibility for $\boldsymbol{B}-2b_T\mathbf{1}$, the results of Lemmas 1 and 2 hold with probability at least $1-\delta/(k+2)^2$ at each round of each regime $k \geqslant 0$ that is actually achieved. Given that

$$\sum_{k\geqslant 0} \frac{1}{(k+2)^2} \leqslant \sum_{k\geqslant 0} \frac{1}{(k+1)(k+2)} = 1\,,$$

we may define $\mathcal{E}_{\text{meta}}$ as the event indicating that the (uniform-in-time) bounds of Lemmas 1 and 2 are satisfied within each of the regimes achieved.

C.1 Step 1: Bounding the number of regimes achieved

We show that, on $\mathcal{E}_{\text{meta}}$, if regime $K = \mathrm{ilog}\,\Vert\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b_T\mathbf{1}}\Vert$ is achieved, then this regime does not stop. Hence, on $\mathcal{E}_{\text{meta}}$, at most $K+1 \leqslant 1 + \mathrm{ilog}\,\Vert\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b_T\mathbf{1}}\Vert$ regimes take place.

Indeed, in regime $K = \mathrm{ilog}\,\Vert\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b_T\mathbf{1}}\Vert$, if it is achieved, the meta-strategy resorts to the Box B strategy with

$$\gamma_K = \frac{2^K}{\sqrt{T}} \geqslant \frac{\max\bigl\{\Vert\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b_T\mathbf{1}}\Vert,\, 1\bigr\}}{\sqrt{T}}\,.$$

Therefore, the bound of Lemma 1 entails that on $\mathcal{E}_{\text{meta}}$, for all $t \geqslant T_K$,

$$\left\| \left(\sum_{\tau=T_K}^{t} \boldsymbol{c}_\tau - (t - T_k + 1)\,(\boldsymbol{B} - b_T \mathbf{1}) \right)_+ \right\| \leqslant \frac{1 + 3\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b_T\mathbf{1}}\|}{\gamma_K} + 20\sqrt{d}\,\Upsilon_{T,\delta/(K+2)^2}$$
$$\leqslant 4\sqrt{T} + 20\sqrt{d}\,\Upsilon_{T,\delta/(K+2)^2} = M_{T,\delta,K}\,.$$

This is exactly the contrary of the stopping condition of regime K: the latter thus cannot be broken.

In some bounds, we will further bound $\operatorname{ilog}\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b_T\mathbf{1}}\|$ by $\operatorname{ilog} T$: this holds because the assumption of $(\boldsymbol{B} - 2b_T\mathbf{1})$–feasibility entails, by Lemma 3 and as OPT is always smaller than 1, the crude bound

$$\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b_T\mathbf{1}}\| \leqslant \frac{\mathrm{OPT}(r, \boldsymbol{c}, \boldsymbol{B} - b_T\mathbf{1}) - \mathrm{OPT}\big(r, \boldsymbol{c}, \boldsymbol{B} - (3/2) b_T \mathbf{1}\big)}{b_T/2} \leqslant \frac{2}{b_T} \leqslant \frac{\sqrt{T}}{7} \leqslant T\,,$$

where we used that $b_T \geqslant 14/\sqrt{T}$ given its definition. (Of course, sharper but more complex bounds could be obtained; however, they would only improve logarithmic terms in the bound, which we do not try to optimize anyway.)

C.2 Step 2: Bounding the cumulative costs

We still denote by K the index of the last regime and recall that $K \leqslant \operatorname{ilog}\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b_T\mathbf{1}}\| \leqslant \operatorname{ilog}(T)$, and that for $k \geqslant 0$, regime k starts at T_k and stops at $T_{k+1}-1$. By convention, $T_0 = 1$ and $T_{K+1} = T+1$.

By the very definition of the stopping condition of regime $k \geqslant 0$,

$$\left\| \left(\sum_{t=T_k}^{T_{k+1}-2} \boldsymbol{c}_t - (T_{k+1} - T_k - 1)\,(\boldsymbol{B} - b_T\mathbf{1}) \right)_+ \right\| \leqslant M_{T,\delta,k}\,.$$

For rounds of the form $t = T_{k+1} - 1$, we bound the Euclidean norm of $\boldsymbol{c}_t - (\boldsymbol{B} - b_T\mathbf{1})$ by $2\sqrt{d}$. Therefore, by a triangle inequality, satisfied both by the non-negative part $(\,\cdot\,)_+$ and the norm $\|\cdot\|$ functions, we have

$$\left\| \left(\sum_{t=1}^{T} \boldsymbol{c}_t - T\,(\boldsymbol{B} - b_T\mathbf{1}) \right)_+ \right\|$$
$$\leqslant \sum_{k=0}^{K} \left\| \left(\sum_{t=T_k}^{T_{k+1}-2} \boldsymbol{c}_t - (T_{k+1} - T_k - 1)\,(\boldsymbol{B} - b_T\mathbf{1}) \right)_+ \right\| + \sum_{k=0}^{K} \big\| \boldsymbol{c}_{T_{k+1}-1} - (\boldsymbol{B} - b_T\mathbf{1}) \big\|$$
$$\leqslant (K+1)\big(M_{T,\delta,\operatorname{ilog} T} + 2\sqrt{d}\big) \leqslant T\,b_T\,,$$

where we used the fact that $M_{T,\delta,k}$ increases with k, the bound $K \leqslant \operatorname{ilog} T$ proved in Step 1 and holding on the event $\mathcal{E}_{\text{meta}}$, as well as the definition of b_T. Therefore, on $\mathcal{E}_{\text{meta}}$, no component of

$$\left(\sum_{t=1}^{T} \boldsymbol{c}_t - T\,(\boldsymbol{B} - b_T\mathbf{1}) \right)_+$$

can be larger than $T\,b_T$, which yields the desired control $\displaystyle\sum_{t=1}^{T} \boldsymbol{c}_t \leqslant T\boldsymbol{B}$.

C.3 Step 3: Computing the associated regret bound

The total regret is the sum of the regrets suffered over each regime:

$$R_T = \sum_{k=0}^{K} \left((T_{k+1} - T_k)\mathrm{OPT}(r, \boldsymbol{c}, \boldsymbol{B}) - \sum_{t=T_k}^{T_{k+1}-1} r_t \right).$$

On the favorable event $\mathcal{E}_{\text{meta}}$, the bound of Lemma 2 holds in particular at the end of each regime; i.e., given the parameters $\gamma_k = 2^k/\sqrt{T}$ and $\delta/\bigl(4(k+2)^2\bigr)$ to run the Box B strategy in regime $k \in \{0, \ldots, K\}$, it holds on $\mathcal{E}_{\text{meta}}$ that

$$(T_{k+1} - T_k)\, \textsc{opt}(r, \boldsymbol{c}, \boldsymbol{B}) - \sum_{t=T_k}^{T_{k+1}-1} r_t \leqslant \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b_T\mathbf{1}}\| \Bigl((T_{k+1} - T_k)\, b_T \sqrt{d} + 6\Upsilon_{T,\delta/(k+2)^2} \Bigr) \\ + 36\sqrt{d}\, \bigl(\Upsilon_{T,\delta/(k+2)^2}\bigr)^2 \, \frac{2^k}{\sqrt{T}} + 8 \ln \frac{T^2}{\delta/\bigl(4(k+2)^2\bigr)} \,.$$

We now sum the above bounds and use the (in)equalities $\Upsilon_{T,\delta/(k+2)^2} \leqslant \Upsilon_{T,\delta/(K+2)^2}$,

$$\sum_{k=0}^{K} (T_{k+1} - T_k)\, b_T = T\, b_T \,, \qquad \text{and} \qquad \sum_{k=0}^{K} 2^k \leqslant 2^{K+1} \leqslant 2^{1+\operatorname{ilog} \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b_T\mathbf{1}}\|} \leqslant 4\, \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b_T\mathbf{1}}\|$$

to get

$$R_T \leqslant \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b_T\mathbf{1}}\| \Bigl(T\, b_T \sqrt{d} + 6K \Upsilon_{T,\delta/(K+2)^2} \Bigr) + 144\sqrt{d}\, \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b_T\mathbf{1}}\| \frac{\bigl(\Upsilon_{T,\delta/(K+2)^2}\bigr)^2}{\sqrt{T}} \\ + 8K \ln \frac{T^2}{\delta/\bigl(4(K+2)^2\bigr)} \,.$$

The final regret bound is achieved by substituting the inequality $K \leqslant \operatorname{ilog} T$ proved in Step 1:

$$R_T \leqslant \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b_T\mathbf{1}}\| \left(144\sqrt{d}\, \frac{\bigl(\Upsilon_{T,\delta/(2+\operatorname{ilog} T)^2}\bigr)^2}{\sqrt{T}} + T\, b_T \sqrt{d} + 6\Upsilon_{T,\delta/(2+\operatorname{ilog} T)^2} \operatorname{ilog} T \right) \\ + 8 \ln \frac{T^2}{\delta/\bigl(4(2+\operatorname{ilog} T)^2\bigr)} \operatorname{ilog} T \,. \qquad (19)$$

The order of magnitude is $\sqrt{T}$, up to poly-logarithmic terms, for the quantity $\Upsilon_{T,\delta/(2+\operatorname{ilog} T)^2}$, thus for $M_{T,\delta,\operatorname{ilog} T}$, thus for $T\, b_T$, therefore, the order of magnitude in $\sqrt{T}$ of the above bound is, up to poly-logarithmic terms,

$$\bigl(1 + \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b_T\mathbf{1}}\|\bigr)\sqrt{T},$$

as claimed.

D Proofs of Lemma 3 and Corollary 1

Proof of Lemma 3. For $\boldsymbol{\lambda} \geqslant \mathbf{0}$ and $\boldsymbol{C} \in [0,1]^d$, we denote

$$L(\boldsymbol{\lambda}, \boldsymbol{C}) = \mathbb{E}_{\boldsymbol{X}\sim\nu} \left[\max_{a\in\mathcal{A}} \Bigl\{ r(\boldsymbol{X}, a) - \bigl\langle \boldsymbol{c}(\boldsymbol{X}, a) - \boldsymbol{C},\, \boldsymbol{\lambda} \bigr\rangle \Bigr\} \right]$$

so that by (4) and the feasibility assumption, we have, at least for $\boldsymbol{C} = \widetilde{\boldsymbol{B}}$ and $\boldsymbol{C} = \boldsymbol{B} - b\mathbf{1}$:

$$\textsc{opt}(r, \boldsymbol{c}, \boldsymbol{C}) = \min_{\boldsymbol{\lambda}\geqslant\mathbf{0}} L(\boldsymbol{\lambda}, \boldsymbol{C}) = L(\boldsymbol{\lambda}^\star_{\boldsymbol{C}}, \boldsymbol{C}) \,.$$

The function L is linear in $\boldsymbol{C}$, so that

$$\textsc{opt}\bigl(r, \boldsymbol{c}, \widetilde{\boldsymbol{B}}\bigr) = L\bigl(\lambda^\star_{\widetilde{\boldsymbol{B}}}, \widetilde{\boldsymbol{B}}\bigr) \leqslant L\bigl(\lambda^\star_{\boldsymbol{B}-b\mathbf{1}}, \widetilde{\boldsymbol{B}}\bigr) = L\bigl(\lambda^\star_{\boldsymbol{B}-b\mathbf{1}}, \boldsymbol{B} - b\mathbf{1}\bigr) - \bigl\langle \lambda^\star_{\boldsymbol{B}-b\mathbf{1}},\, \boldsymbol{B} - b\mathbf{1} - \widetilde{\boldsymbol{B}} \bigr\rangle \\ = \textsc{opt}\bigl(r, \boldsymbol{c}, \boldsymbol{B} - b\mathbf{1}\bigr) - \bigl\langle \lambda^\star_{\boldsymbol{B}-b\mathbf{1}},\, \boldsymbol{B} - b\mathbf{1} - \widetilde{\boldsymbol{B}} \bigr\rangle \,.$$

The result follows from substituting

$$\bigl\langle \lambda^\star_{\boldsymbol{B}-b\mathbf{1}},\, \boldsymbol{B} - b\mathbf{1} - \widetilde{\boldsymbol{B}} \bigr\rangle \geqslant \|\lambda^\star_{\boldsymbol{B}-b\mathbf{1}}\|_1 \, \min\bigl(\boldsymbol{B} - b\mathbf{1} - \widetilde{\boldsymbol{B}}\bigr) \geqslant \|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\| \, \min\bigl(\boldsymbol{B} - b\mathbf{1} - \widetilde{\boldsymbol{B}}\bigr)$$

and from rearranging the inequality thus obtained. □

Proof of Corollary 1. We apply Lemma 3 with $\widetilde{\boldsymbol{B}} = \varepsilon\mathbf{1}$ for some $\varepsilon > 0$ sufficiently small and obtain

$$\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b\mathbf{1}}\| \leqslant \frac{\textsc{opt}(r, \boldsymbol{c}, \boldsymbol{B} - b\mathbf{1}) - \textsc{opt}\bigl(r, \boldsymbol{c}, \varepsilon\mathbf{1}\bigr)}{\min\bigl(\boldsymbol{B} - (b+\varepsilon)\mathbf{1}\bigr)} \,.$$

We conclude by substituting $\textsc{opt}(r, \boldsymbol{c}, \boldsymbol{B} - b\mathbf{1}) \leqslant \textsc{opt}(r, \boldsymbol{c}, \boldsymbol{B})$ and $\textsc{opt}\bigl(r, \boldsymbol{c}, \varepsilon\mathbf{1}\bigr) \geqslant \textsc{opt}\bigl(r, \boldsymbol{c}, \mathbf{0}\bigr)$ as well as $\min(\boldsymbol{B} - b\mathbf{1}) \geqslant \min \boldsymbol{B}/2$, and by letting $\varepsilon \to 0$. □

E Additional (sketch of) results concerning optimality

We detail here two series of claims made in Section 4 .

E.1 A proof scheme for problem-dependent lower bounds

We provide the proof scheme for proving the problem-dependent lower bound of order $\big(1+\|\lambda^\star_{\boldsymbol{B}}\|\big)\sqrt{T}$ announced in Section 4.

Step 0: Considering strict cost constraints. Our aim, as described in Box A, is to make sure that with high probability the cumulative costs are smaller than $T\,\boldsymbol{B}$. If we considered softer constraints, of the form $T\,\boldsymbol{B}+\widetilde{\mathcal{O}}\big(\sqrt{T}\big)$, then $\widetilde{\mathcal{O}}\big(\sqrt{T}\big)$ regret bounds would be possible (see Appendix F); i.e., the factor $\|\boldsymbol{\lambda}^\star_{\boldsymbol{B}-b_T\mathbf{1}}\|$ of Theorem 1 could be replaced by a constant. Thus, lower bounds are only interesting in the case of hard constraints stated in Box A.

Step 1: Necessity of a margin b_T of order $1/\sqrt{T}$. First, a classical lemma in CBwK (see, e.g., Agrawal and Devanur, 2016, Lemma 1) indicates that a sequence of adaptive cannot perform better than an optimal static policy. Denote by $\pi^\star_{\boldsymbol{B}'}$ a (quasi-)optimal policy for the average cost constraints $\boldsymbol{B}'$. Provided that costs are truly random (i.e., do not stem from Dirac distributions, which in particular, does not cover the cases where there is a null-cost action, see Limitation 2 in Section 4), then the law of iterated logarithm shows that when playing $\pi^\star_{\boldsymbol{B}'}$ at each round, the cumulative costs must (almost-surely, as $T\to+\infty$) be larger than $T\,\boldsymbol{B}'$ plus a positive term of the order of $\sqrt{T\ln\ln T}$. Therefore, to meet the hard constraints, one should pick $\boldsymbol{B}'$ of the form $\boldsymbol{B}-b_T\mathbf{1}$, where b_T is of order $1/\sqrt{T}$ up to logarithmic terms.

Step 2: Consequences in terms of regret. Therefore, the largest average reward a strategy may target is $\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}-b_T\mathbf{1})$. Deviations of the order $\sqrt{T}$ are also bound to happen. Therefore, up to logarithmic terms, the regret lower bound is approximatively larger than something of the order

$$T\big(\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B})-\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}-b_T\mathbf{1})\big)+\sqrt{T}\,.$$

Now, an argument similar to the one used in the proof of Lemma 3 shows that

$$\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B})-\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}-b_T\mathbf{1})\geqslant L(\lambda^\star_{\boldsymbol{B}},\boldsymbol{B})-L(\lambda^\star_{\boldsymbol{B}},\boldsymbol{B}-b_T\mathbf{1})=b_T\,\|\lambda^\star_{\boldsymbol{B}}\|\,.$$

All in all, the regret lower bound is thus approximatively larger than something of the order of

$$\big(1+\|\lambda^\star_{\boldsymbol{B}}\|\big)\sqrt{T}\,.$$

This matches the form of the bound of Theorem 1, but the dual vector $\lambda^\star_{\boldsymbol{B}-b_T\mathbf{1}}$ present in the upper bound is replaced by $\lambda^\star_{\boldsymbol{B}}$ in our lower bound.

E.2 Faster rates may be achievable in some specific cases

We explain here why, in some specific cases, faster rates would be achievable—with $\sqrt{T}$ in the regret bound of Theorem 1 replaced by $\sqrt{TB}$ for a problem with scalar costs and total-cost constraints B.

Indeed, consider a problem similar to Example 1: with scalar costs, total-cost constraint $B>0$, featuring a baseline action a_{null} with null cost but also null reward, and additional actions with larger rewards and expected costs $c(\boldsymbol{x},a)\geqslant\alpha>0$. Let N_T denotes the number of times non-null cost actions are played within the first T rounds. Deviation inequalities have it that

$$\sum_{t=1}^{T}c_t\geqslant\alpha N_T-\tilde{O}\Big(\sqrt{N_T}\Big)\,,$$

where we recall that $\tilde{O}$ is up to poly-logarithmic factors; as a consequence, the total-cost constraint enforces that

$$N_T\leqslant\frac{TB}{\alpha}+\tilde{O}\left(\sqrt{\frac{TB}{\alpha}}\right).$$

In particular, the margin b_T only needs to be of order $\sqrt{N_T}/T$, i.e., $\sqrt{B/(\alpha T)}$, instead of $1/\sqrt{T}$.

Similarly, since at most N_T non-null actions are played, the regret should be lower bounded by something of the order of

$$T\big(\mathrm{OPT}(r,c,B) - \mathrm{OPT}(r,c,B-b_T)\big) + \sqrt{N_T}\,,$$

where, as in Step 2 above,

$$\big(\mathrm{OPT}(r,c,B) - \mathrm{OPT}(r,c,B-b_T)\big) \geqslant b_T|\lambda^\star_B| = \tilde{O}\left(\sqrt{\frac{B}{\alpha T}}\right)\,|\lambda^\star_B|.$$

This suggests a lower bound on the regret of the order (up to poly-logarithmic factors) of

$$\big(|\lambda^\star_B|+1\big)\sqrt{TB/\alpha} \qquad \text{instead of} \qquad \big(1+|\lambda^\star_B|\big)\sqrt{T}\,.$$

The difference between the two bounds is significant when B is small, i.e., $B \ll 1$. While proving a matching upper bound with the Box B and Box C strategies looks a bit tricky, we feel that it must be possible to do so with the primal approach of Appendix F, at least when the context space $\mathcal{X}$ is finite. If our intuition holds, then it would be possible to get an upper bound

$$\begin{aligned} R_T &\leqslant \tilde{O}\Big(Tb_T\big(1+|\lambda^*_{B-b_T}|\big)\Big) \\ &\qquad\qquad = \tilde{O}\left(\sqrt{\frac{TB}{\alpha}}\left(1+\frac{\mathrm{OPT}(r,c,B)-\mathrm{OPT}(r,c,0)}{B}\right)\right) = \tilde{O}\Big(\sqrt{TB/\alpha^3}\Big)\,, \end{aligned}$$

where the last bound follows from $\mathrm{OPT}(r,c,B) - \mathrm{OPT}(r,c,0) \leqslant B/\alpha$, as explained in Example 1.

F Primal strategy

This section studies the primal strategy stated in Box D, which, at every round, solves an approximation of the primal optimization problem (1). The key issue in running such a primal approach is estimating ν, see comments after Notation 1; this primal approach is essentially worth for the case of finite context sets $\mathcal{X}$. The aim of this section is threefold:

1. In Appendix F.1, we provide a theory of "soft" constraints, when total-cost deviations from $T\boldsymbol{B}$ of order $\sqrt{T}$ up to logarithmic terms are allowed; at least when $\mathcal{X}$ is a finite set, the regret bound then becomes proportional to $\sqrt{T}$ up to logarithmic terms.
2. In Appendix F.2, we revisit and extend the results by Li and Stoltz [2022]. The extension consists of dealing with possibly signed constraints, and the revisited analysis (of the same strategy as in Li and Stoltz, 2022) consists in not directly dealing with KKT constraints (which, in addition, imposed the finiteness of $\mathcal{X}$) but in only relating optimization problems—defined with the true r and $\boldsymbol{c}$ or estimates thereof. We also offer a modular approach and separate the error terms coming from estimating ν and from estimating r and $\boldsymbol{c}$.
3. In Appendix F.3, we generalize the results of Appendix F.2, which rely on the existence of a null-cost action, and get guarantees that correspond quite exactly to the combination of Theorem 1 and the interpretation thereof offered by Lemma 3, at least in the case of a finite $\mathcal{X}$. Actually, in our research path, we had first obtained these primal results, before trying to obtain them in a more general case of a continuous $\mathcal{X}$, by resorting to a dual strategy. The proof technique in Appendix F.3 also inspired our approach to proving problem-dependent lower bounds presented in Appendix E.1.

Throughout this appendix, we will assume that the context distribution ν can be estimated in some way. We provide examples and pointers below.

Notation 1. *Fix $\delta \in (0,1)$. We denote by $\widehat{\nu}_{\delta,t}$ a sequence of estimators of ν, each constructed on the contexts $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_t$, and by $\xi_{t,\delta}$ a sequence of estimation errors such that, with probability at least $1-\delta$, for all bounded functions $f:\mathcal{X}\to[-1,1]$,*

$$\Big|\mathbb{E}_{\boldsymbol{X}\sim\nu}\big[f(\boldsymbol{X})\big] - \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta,t}}\big[f(\boldsymbol{X})\big]\Big| \leqslant \xi_{t,\delta}\,.$$

We also denote $\quad \Xi_{T,\delta} = \displaystyle\sum_{t=1}^{T}\xi_{t-1,\delta}\,.$

We consider the strategy described in Box D, whether $\mathcal{X}$ is finite (as in Li and Stoltz, 2022, where it was first considered) or not. Our simplified analyses do not rely on explicit KKT inequalities and therefore do not require anymore that $\mathcal{X}$ is finite. In Box D, by "when the empirical cost constraints are feasible", we mean that there exists some policy $\boldsymbol{\pi}$ such that

$$\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta,t-1}}\left[\sum_{a\in\mathcal{A}} \widehat{\boldsymbol{c}}^{\,\text{lcb}}_{\delta,t-1}(\boldsymbol{X},a)\,\pi_a(\boldsymbol{X})\right] \leqslant \boldsymbol{B} + b_t\boldsymbol{1}\,.$$

We need to guarantee the existence of a policy achieving the constrained maximum of Step 2 in Box D. This is immediate when $\mathcal{X}$ is finite, as the problem then corresponds to a finite-dimensional linear program. In general, we may note that when (4) holds, we read therein the optimal policy, as a pointwise maximum involving the optimal dual variables. The proofs reveal that up to slightly augmenting the $\xi_{t-1,\delta}$ of Notation 1, we may assume that the strict feasibility sufficient for (4) indeed holds.

We also note that even if the argument of the maximum $\boldsymbol{\pi}_t$ is guaranteed to exist, it may be difficult to compute: the linear program of Box D cannot be solved exactly if the $\widehat{\nu}_{\delta,t}$ are not finitely supported (which happens in general when $\mathcal{X}$ is not finite). We do not see this as a severe issues as the Box D strategy is an ideal strategy anyway, that we study for the sake of shedding lights on our results for the dual approach in the main body of the article.

A final remark on the Box D strategy is that the margins on the average cost constraints $\boldsymbol{B}$ can now be signed, which is why they will be referred to as signed slacks.

BOX D: PRIMAL CBWK STRATEGY, GENERALIZED FROM LI AND STOLTZ [2022]

Inputs: confidence level $1-\delta$; estimation procedure and error functions ε_t of Assumption 2; optimistic estimates (2); estimation procedure for ν as in Notation 1

Hyperparameter: signed slacks $b_1, b_2, \ldots, b_T \in \mathbb{R}$

Initialization: initial estimates $\widehat{r}^{\,\text{ucb}}_{\delta,0}(\boldsymbol{x},a)$ and $\widehat{\boldsymbol{c}}^{\,\text{lcb}}_{\delta,0}(\boldsymbol{x},a)$, as well as $\widehat{\nu}_{\delta,0}$

For rounds $t = 1, 2, 3, \ldots, T$:

1. Compute a policy $\boldsymbol{\pi}_t$ achieving

$$\begin{aligned} &\max_{\boldsymbol{\pi}:\mathcal{X}\to\mathcal{P}(\mathcal{A})} \quad \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta,t}}\left[\sum_{a\in\mathcal{A}} \widehat{r}^{\,\text{ucb}}_{\delta,t-1}(\boldsymbol{X},a)\,\pi_a(\boldsymbol{X})\right] \\ &\text{under} \qquad \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta,t-1}}\left[\sum_{a\in\mathcal{A}} \widehat{\boldsymbol{c}}^{\,\text{lcb}}_{\delta,t-1}(\boldsymbol{X},a)\,\pi_a(\boldsymbol{X})\right] \leqslant \boldsymbol{B} + b_t\boldsymbol{1} \end{aligned}$$

   when the empirical cost constraints are feasible, and pick an arbitrary policy $\boldsymbol{\pi}_t$ otherwise;
2. Observe $\boldsymbol{x}_t$ and draw an action $a_t \sim \boldsymbol{\pi}_t(\boldsymbol{x}_t)$;
3. Compute the estimate $\widehat{r}^{\,\text{ucb}}_{\delta,t}(\boldsymbol{x},a)$ and $\widehat{\boldsymbol{c}}^{\,\text{lcb}}_{\delta,t}(\boldsymbol{x},a)$, as well as $\widehat{\nu}_{\delta,t}$.

Discussion of the estimation of ν. The simplest case is when $\mathcal{X}$ is a finite set; in this case, we may take

$$\xi_{t,\delta} \qquad \text{of order} \qquad \sqrt{\frac{|\mathcal{X}|\ln(1/\delta)}{t}}\,,$$

where $\mathcal{X}$ denotes the cardinality of $\mathcal{X}$; see [Ai et al., 2022, Section 4.1], see also a less sharp bound based on Hoeffding's inequality in [Li and Stoltz, 2022, Section 5]. This leads to a total error term $\Xi_{T,\delta}$ of order $\sqrt{T}$ up to poly-logarithmic term.

When $\mathcal{X}$ is a continuous subset of $\mathbb{R}^n$, some regularity conditions are put on ν, which is typically assumed to have some smooth density with respect to the Lebesgue measure $\mathfrak{m}$. Estimates $\widehat{\nu}_{\delta,t}$ are obtained by estimating the density $\mathrm{d}\nu/\mathrm{d}\mathfrak{m}$: the criterion in Notation 1, which is proportional to the total-variation distance between ν and $\widehat{\nu}_{\delta,t}$, is then given by the $\mathbb{L}^1(\mathfrak{m})$ distance between the two densities. To control the latter with uniform convergence rates, so as to obtain deviation terms $\xi_{t,\delta}$ only depending on t and δ, heavy assumptions on the model to which ν belongs are in order, e.g., some Hölderian regularity, and uniform estimation rates obtained degrade with the ambient dimension n; they are generally much slower than $1/\sqrt{t}$. The total error terms $\Xi_{T,\delta}$ then prevent the bounds stated below in Proposition 1 and Theorems 3 and 4 from sharing the same orders of magnitude than the bounds proved with our dual approach in Theorem 1 and interpreted in Section 3.3. On this topic, see also [Ai et al., 2022, Section 4.1] for the estimation rates as well as a similar description in Han et al. [2022, end of Section 1] of the limitation of the primal approach in CBwK due to the estimation of densities. General references on density estimations are the monographs by Devroye and Györfi [1985] in $\mathbb{L}^1$ and Tsybakov [2008] in $\mathbb{L}^2$.

Note that in the dual approach, the knowledge of (the possibly complex) ν is replaced by the knowledge of the (finite-dimensional) optimal dual variables $\boldsymbol{\lambda}^\star_{\boldsymbol{B}} \in \mathbb{R}^d$, which is easier to learn. This explains the fundamental efficiency of the dual approach compared to the primal approach.

F.1 Analysis with "soft" constraints

We first provide an analysis for a version of the Box D strategy that may possibly breach the total-cost constraints $T\boldsymbol{B}$. More precisely, we allow deviations to $T\boldsymbol{B}$ of the order of $\sqrt{T}$ times poly-log factors: this is what we refer to as "soft" constraints. Our result is that the regret may then be bounded by a quantity of order $\sqrt{T}$ times poly-log factors, at least when $\mathcal{X}$ is finite.

We do so for two reasons: first, because we do not think that this is a well-known result, and second, for pedagogic reasons, as the proof for "hard" constraints follows from adapting the proof scheme for soft constraints (see Appendix F.2).

Proposition 1 (soft constraints)**.** *Fix $\delta \in (0,1)$. Under Assumption 2 and with Notation* (1)*, the strategy of Box D, run with $\delta/4$ and positive slacks $b_t = \xi_{t-1,\delta/4}$, ensures that with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} \boldsymbol{c}_t \leqslant T\boldsymbol{B} + \bigl(2\alpha_{T,\delta/4} + \beta_{T,\delta/4} + 2\Xi_{T,\delta/4}\bigr)\mathbf{1} \qquad \textit{and} \qquad R_T \leqslant 2\alpha_{T,\delta/4} + \beta_{T,\delta/4} + 2\Xi_{T,\delta/4}\,,$$

where $\alpha_{T,\delta/4} = \sqrt{2T\ln\bigl((d+1)/(\delta/4)\bigr)}$.

In particular, when $\mathcal{X}$ is finite, the deviation terms $2\alpha_{T,\delta/4} + \beta_{T,\delta/4} + 2\Xi_{T,\delta/4}$ are of order $\sqrt{T}$ up to poly-logarithmic terms, so that the bound of Proposition 1 reads: with high-probability,

$$\sum_{t=1}^{T} \boldsymbol{c}_t \leqslant T\boldsymbol{B} + \widetilde{\mathcal{O}}\bigl(\sqrt{T}\bigr) \qquad \text{and} \qquad R_T \leqslant \widetilde{\mathcal{O}}\bigl(\sqrt{T}\bigr)\,.$$

Put differently, soft-constraint satisfaction allows for $\widetilde{\mathcal{O}}\bigl(\sqrt{T}\bigr)$ regret bounds when $\mathcal{X}$ is finite.

Proof. As in the proofs of Lemmas 1 and 2 in Appendix B, we consider four events, each of probability at least $1-\delta/4$: two events $\mathcal{E}_{\text{H-Az1}}$ and $\mathcal{E}_{\text{H-Az2}}$, defined below, following from applications of the Hoeffding-Azuma inequality, the favorable event $\mathcal{E}_{\text{TVD}}$ of Notation 1 with $\delta/4$, and the favorable event $\mathcal{E}_\beta$ of Assumption 2 with $\delta/4$. Namely, given the value for $\alpha_{T,\delta/4}$ proposed in the statement (note this value is slightly different from the one considered in Appendix B), we have, on the one hand, on $\mathcal{E}_{\text{H-Az1}}$,

$$\sum_{t=1}^{T} \boldsymbol{c}_t \leqslant \alpha_{T,\delta/4}\mathbf{1} + \sum_{t=1}^{T} \boldsymbol{c}(\boldsymbol{x}_t, a_t) \qquad \text{and} \qquad \sum_{t=1}^{T} r_t \geqslant -\alpha_{T,\delta/4} + \sum_{t=1}^{T} r(\boldsymbol{x}_t, a_t)\,, \tag{20}$$

and on the other hand, on $\mathcal{E}_{\text{H-Az2}}$,

$$\sum_{t=1}^{T} \widehat{\boldsymbol{c}}^{\,\text{lcb}}_{\delta/4,t-1}(\boldsymbol{x}_t, a_t) \leqslant \alpha_{T,\delta/4}\mathbf{1} + \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \widehat{\boldsymbol{c}}^{\,\text{lcb}}_{\delta/4,t-1}(\boldsymbol{X}, a)\,\pi_{t,a}(\boldsymbol{X})\right]$$

$$\text{and} \qquad \sum_{t=1}^{T} \widehat{r}^{\,\text{ucb}}_{\delta/4,t-1}(\boldsymbol{x}_t, a_t) \geqslant -\alpha_{T,\delta/4} + \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \widehat{r}^{\,\text{ucb}}_{\delta/4,t-1}(\boldsymbol{X}, a)\,\pi_{t,a}(\boldsymbol{X})\right]. \tag{21}$$

The first two inequalities are obtained by considering conditional expectations with respect to the past, $\boldsymbol{x}_t$ and a_t, while the second two inequalities follow from taking conditional expectations with respect to the past and $\boldsymbol{x}_t$; we crucially use that, by definition, $\boldsymbol{x}_t \sim \nu$ is independent from $\boldsymbol{\pi}_t$ and the estimates $\widehat{r}^{\,\mathrm{ucb}}_{\delta/4,t-1}$ and $\widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,t-1}$.

We first note that by $\boldsymbol{B}$–feasibility of the problem, by (3), and by Notation (1), a policy $\boldsymbol{\pi}_t$ satisfying the constraints stated in Box D exists at each round $t \geqslant 1$ on the event $\mathcal{E}_\beta \cap \mathcal{E}_{\mathrm{TVD}}$. Indeed, denoting by $\boldsymbol{\pi}^{\mathrm{feas}}$ such a $\boldsymbol{B}$–feasible policy, this existence follows from the inequalities

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\!\left[\sum_{a\in\mathcal{A}} \widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\pi^{\mathrm{feas}}_a(\boldsymbol{X})\right] &\leqslant \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\!\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(\boldsymbol{X},a)\,\pi^{\mathrm{feas}}_a(\boldsymbol{X})\right] \\ &\leqslant \xi_{t-1,\delta/4} + \underbrace{\mathbb{E}_{\boldsymbol{X}\sim\nu}\!\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(\boldsymbol{X},a)\,\pi^{\mathrm{feas}}_a(\boldsymbol{X})\right]}_{\leqslant \boldsymbol{B}} \qquad (22)\end{aligned}$$

and from the choice $b_t = \xi_{t-1,\delta/4}$.

Cost-wise, we then successively have, by (20), then (3) and Assumption 2, and finally (21) and Notation (1), on the event $\mathcal{E}_{\text{H-Az1}} \cap \mathcal{E}_\beta \cap \mathcal{E}_{\text{H-Az2}} \cap \mathcal{E}_{\mathrm{TVD}}$ of probability at least $1-\delta$,

$$\begin{aligned}\sum_{t=1}^{T} \boldsymbol{c}_t &\leqslant \alpha_{T,\delta/4}\mathbf{1} + \sum_{t=1}^{T} \boldsymbol{c}(\boldsymbol{x}_t,a_t) \\ &\leqslant \big(\alpha_{T,\delta/4}+\beta_{T,\delta/4}\big)\mathbf{1} + \sum_{t=1}^{T} \widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,t-1}(\boldsymbol{x}_t,a_t) \\ &\leqslant \big(2\alpha_{T,\delta/4}+\beta_{T,\delta/4}\big)\mathbf{1} + \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{X}\sim\nu}\!\left[\sum_{a\in\mathcal{A}} \widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\pi_{t,a}(\boldsymbol{X})\right] \\ &\leqslant \big(2\alpha_{T,\delta/4}+\beta_{T,\delta/4}+\Xi_{T,\delta/4}\big)\mathbf{1} + \sum_{t=1}^{T} \underbrace{\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\!\left[\sum_{a\in\mathcal{A}} \widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\pi_{t,a}(\boldsymbol{X})\right]}_{\leqslant \boldsymbol{B}+b_t\mathbf{1}},\end{aligned}$$

where the inequalities $\leqslant \boldsymbol{B}+b_t\mathbf{1}$ follow from the definition of $\boldsymbol{\pi}_t$ in Box D. Substituting the choice $b_t = \xi_{t-1,\delta/4}$, we thus proved that on $\mathcal{E}_{\text{H-Az1}} \cap \mathcal{E}_\beta \cap \mathcal{E}_{\text{H-Az2}} \cap \mathcal{E}_{\mathrm{TVD}}$,

$$\sum_{t=1}^{T} \boldsymbol{c}_t \leqslant T\boldsymbol{B} + \big(2\alpha_{T,\delta/4}+\beta_{T,\delta/4}+2\Xi_{T,\delta/4}\big)\mathbf{1}\,,$$

as claimed.

The control for rewards mimics the steps above for costs (and then resorts to an additional argument). We obtain first that on $\mathcal{E}_{\text{H-Az1}} \cap \mathcal{E}_\beta \cap \mathcal{E}_{\text{H-Az2}} \cap \mathcal{E}_{\mathrm{TVD}}$,

$$\sum_{t=1}^{T} r_t \geqslant -\big(2\alpha_{T,\delta/4}+\beta_{T,\delta/4}+\Xi_{T,\delta/4}\big) + \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\!\left[\sum_{a\in\mathcal{A}} \widehat{r}^{\,\mathrm{ucb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\pi_{t,a}(\boldsymbol{X})\right].$$

We denote by $\pi^\star_{\boldsymbol{B}}$ an optimal static policy for the average cost constraints $\boldsymbol{B}$—when it exists, e.g., by (4), as soon as the problem is feasible for some $\boldsymbol{B}' < \boldsymbol{B}$; otherwise, we take a static policy achieving $\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B})$ up to some small $e>0$, which we let vanish in a final step of the proof. As in (22), we have, on $\mathcal{E}_\beta \cap \mathcal{E}_{\mathrm{TVD}}$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\!\left[\sum_{a\in\mathcal{A}} \widehat{\boldsymbol{c}}^{\,\mathrm{lcb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\pi^\star_{\boldsymbol{B},a}(\boldsymbol{X})\right] &\leqslant \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\!\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(\boldsymbol{X},a)\,\pi^\star_{\boldsymbol{B},a}(\boldsymbol{X})\right] \\ &\leqslant \xi_{t-1,\delta/4}\mathbf{1} + \underbrace{\mathbb{E}_{\boldsymbol{X}\sim\nu}\!\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(\boldsymbol{X},a)\,\pi^\star_{\boldsymbol{B},a}(\boldsymbol{X})\right]}_{\leqslant \boldsymbol{B}},\end{aligned}$$

where the $\leqslant \boldsymbol{B}$ inequality follows from the definition of $\boldsymbol{\pi}^\star_{\boldsymbol{B}}$. Thanks to the choice $b_t = \xi_{t-1,\delta/4}$, we have, by definition of $\boldsymbol{\pi}_t$ as some optimal policy in Box D and on $\mathcal{E}_\beta \cap \mathcal{E}_{\text{TVD}}$,

$$\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\left[\sum_{a\in\mathcal{A}} \widehat{r}^{\,\text{ucb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\pi_{t,a}(\boldsymbol{X})\right] \geqslant \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\left[\sum_{a\in\mathcal{A}} \widehat{r}^{\,\text{ucb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\pi^\star_{\boldsymbol{B},a}(\boldsymbol{X})\right].$$

Again by (3) and Notation (1), we have, on $\mathcal{E}_\beta \cap \mathcal{E}_{\text{TVD}}$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\left[\sum_{a\in\mathcal{A}} \widehat{r}^{\,\text{ucb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\pi^\star_{\boldsymbol{B},a}(\boldsymbol{X})\right] &\geqslant \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\left[\sum_{a\in\mathcal{A}} r(\boldsymbol{X},a)\,\pi^\star_{\boldsymbol{B},a}(\boldsymbol{X})\right] \\ &\geqslant -\xi_{t-1,\delta/4} + \underbrace{\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(\boldsymbol{X},a)\,\pi^\star_{\boldsymbol{B},a}(\boldsymbol{X})\right]}_{=\text{OPT}(r,\boldsymbol{c},\boldsymbol{B})}.\end{aligned}$$

Collecting all bounds, we proved that on $\mathcal{E}_{\text{H-Az1}} \cap \mathcal{E}_\beta \cap \mathcal{E}_{\text{H-Az2}} \cap \mathcal{E}_{\text{TVD}}$,

$$\sum_{t=1}^T r_t \geqslant T\,\text{OPT}(r,\boldsymbol{c},\boldsymbol{B}) - \bigl(2\alpha_{T,\delta/4} + \beta_{T,\delta/4} + 2\Xi_{T,\delta/4}\bigr),$$

which corresponds to the claimed regret bound. □

F.2 Analysis with "hard" constraints and a null-cost action

We now turn our attention to the main kind of result that we want to achieve: when constraints must be strictly satisfied—which we refer to as "hard" constraints. For the sake of simplicity, we do so for now in the presence of a null-cost action; Appendix F.3 explains how the analysis may be generalized to cases without such null-cost actions.

The following result corresponds to the combination of Theorem 1 with Corollary 1, and also, to Li and Stoltz [2022, main result: Theorem 2].

Theorem 3 (hard constraints)**.** *Fix $\delta \in (0,1)$. We consider the strategy of Box D, run with $\delta/4$ and negative slacks all equal to*

$$b_t \equiv -\overline{\Delta}_{T,\delta/4}, \qquad \textit{where} \qquad \overline{\Delta}_{T,\delta/4} \overset{\text{def}}{=} \frac{2\alpha_{T,\delta/4} + \beta_{T,\delta/4} + \Xi_{T,\delta/4}}{T}$$

and $\alpha_{T,\delta/4} = \sqrt{2T\ln\bigl((d+1)/(\delta/4)\bigr)}$. Assume that a null-cost action exists, that $\overline{\Delta}_{T,\delta/4} < \min \boldsymbol{B}$, that Assumption 2 holds, and use Notation (1)*. Then, with probability at least $1-\delta$,*

$$\sum_{t=1}^T \boldsymbol{c}_t \leqslant T\boldsymbol{B} \quad \textit{and} \quad R_T \leqslant \bigl(2\alpha_{T,\delta/4} + \beta_{T,\delta/4} + 2\Xi_{T,\delta/4}\bigr)\left(1 + \frac{\text{OPT}(r,\boldsymbol{c},\boldsymbol{B}) - \text{OPT}(r,\boldsymbol{c},\boldsymbol{0})}{\min \boldsymbol{B}}\right).$$

Proof. We explain how the proof of Proposition 1 may be adapted. We first justify the existence at each round $t \geqslant 1$ of a policy $\boldsymbol{\pi}_t$ satisfying the cost constraints stated in Box D: for the null-cost action a_{null}, we may impose that, for all $\boldsymbol{x} \in \mathcal{X}$, all $t \geqslant 0$, and all $\delta \in (0,1)$,

$$\widehat{\boldsymbol{c}}_t(\boldsymbol{x},a_{\text{null}}) = \boldsymbol{0} \quad \text{and} \quad \varepsilon_t(\boldsymbol{x},a_{\text{null}},\delta) = 0, \qquad \text{so that} \quad \widehat{\boldsymbol{c}}^{\,\text{lcb}}_{\delta,t}(\boldsymbol{x},a_{\text{null}}) = \boldsymbol{0} \ \text{ a.s.};$$

this shows that at least the static policy $\boldsymbol{\pi}^{\text{null}}$ always playing a_{null} satisfying the defining cost-constraints for $\boldsymbol{\pi}_t$. (Alternatively, we may note that the policy $\overline{\boldsymbol{\pi}}_t$ defined below also satisfies the cost constraints stated in Box D, on the high-probability event considered.)

We then handle total-cost constraints similarly as in the proof of Proposition 1 and obtain that on the event $\mathcal{E}_{\text{H-Az1}} \cap \mathcal{E}_\beta \cap \mathcal{E}_{\text{H-Az2}} \cap \mathcal{E}_{\text{TVD}}$ of probability at least $1-\delta$,

$$\sum_{t=1}^T \boldsymbol{c}_t \leqslant \bigl(2\alpha_{T,\delta/4} + \beta_{T,\delta/4} + \Xi_{T,\delta/4}\bigr)\boldsymbol{1} + \sum_{t=1}^T \underbrace{\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\left[\sum_{a\in\mathcal{A}} \widehat{\boldsymbol{c}}^{\,\text{lcb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\pi_{t,a}(\boldsymbol{X})\right]}_{\leqslant \boldsymbol{B} - \overline{\Delta}_{T,\delta/4}\boldsymbol{1}} \leqslant T\boldsymbol{B},$$

where this time, the slacks $b_t = -\overline{\Delta}_{T,\delta/4}$ are negative and were set to cancel out the positive deviation terms. Similarly, on $\mathcal{E}_{\text{H-Az1}} \cap \mathcal{E}_\beta \cap \mathcal{E}_{\text{H-Az2}} \cap \mathcal{E}_{\text{TVD}}$,

$$\sum_{t=1}^{T} r_t \geqslant -T\overline{\Delta}_{T,\delta/4} + \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\left[\sum_{a\in\mathcal{A}} \widehat{r}^{\,\text{ucb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\pi_{t,a}(\boldsymbol{X})\right]. \tag{23}$$

Now, the main modification to the proof of Proposition 1 is that (with its notation) we rather consider the policy

$$\overline{\boldsymbol{\pi}}_t = (1-w_t)\boldsymbol{\pi}^\star_{\boldsymbol{B}} + w_t\boldsymbol{\pi}^{\text{null}}, \qquad \text{where} \quad w_t = \min\left\{\frac{\overline{\Delta}_{T,\delta/4} + \xi_{t-1,\delta/4}}{\min \boldsymbol{B}},\ 1\right\}.$$

As in (22), we see that this policy satisfies, on $\mathcal{E}_\beta \cap \mathcal{E}_{\text{TVD}}$:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\left[\sum_{a\in\mathcal{A}} \widehat{\boldsymbol{c}}^{\,\text{lcb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\overline{\pi}_{t,a}(\boldsymbol{X})\right] &\leqslant \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(\boldsymbol{X},a)\,\overline{\pi}_{t,a}(\boldsymbol{X})\right] \\
&\leqslant \xi_{t-1,\delta/4}\mathbf{1} + \underbrace{\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(\boldsymbol{X},a)\,\overline{\pi}_{t,a}(\boldsymbol{X})\right]}_{\leqslant(1-w_t)\boldsymbol{B}}.
\end{aligned}$$

By using $\boldsymbol{B} \geqslant (\min \boldsymbol{B})\mathbf{1}$ and since we assumed that $\overline{\Delta}_{T,\delta/4} < \min \boldsymbol{B}$, we may continue the series of inequalities as follows:

$$\begin{aligned}
\xi_{t-1,\delta/4}\mathbf{1} + (1-w_t)\boldsymbol{B} &\leqslant \boldsymbol{B} + \xi_{t-1,\delta/4}\mathbf{1} - w_t(\min \boldsymbol{B})\mathbf{1} \\
&= \boldsymbol{B} - \min\bigl\{\overline{\Delta}_{T,\delta/4},\ \min \boldsymbol{B} - \xi_{t-1,\delta/4}\bigr\}\mathbf{1} \\
&\leqslant \boldsymbol{B} - \min\bigl\{\overline{\Delta}_{T,\delta/4},\ \overline{\Delta}_{T,\delta/4} - \xi_{t-1,\delta/4}\bigr\}\mathbf{1} = \boldsymbol{B} + b_t\mathbf{1},
\end{aligned}$$

meaning that $\overline{\boldsymbol{\pi}}_t$ is a policy satisfying the constraints stated in Step 2 of Box D on round $t \geqslant 1$. The consequence is that, first by definition of $\boldsymbol{\pi}_t$ as a maximizer and then by the optimistic estimates (3) and Notation (1), on $\mathcal{E}_\beta \cap \mathcal{E}_{\text{TVD}}$,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\left[\sum_{a\in\mathcal{A}} \widehat{r}^{\,\text{ucb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\pi_{t,a}(\boldsymbol{X})\right] \\
&\qquad \geqslant \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\left[\sum_{a\in\mathcal{A}} \widehat{r}^{\,\text{ucb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\overline{\pi}_{t,a}(\boldsymbol{X})\right] \\
&\qquad \geqslant -\xi_{t-1,\delta/4} + \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(\boldsymbol{X},a)\,\overline{\pi}_{t,a}(\boldsymbol{X})\right] \\
&\qquad = -\xi_{t-1,\delta/4} + (1-w_t)\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}) + w_t\mathrm{OPT}(r,\boldsymbol{c},\mathbf{0}) \\
&\qquad = \mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}) - \xi_{t-1,\delta/4} - \min\left\{\frac{\overline{\Delta}_{T,\delta/4} + \xi_{t-1,\delta/4}}{\min \boldsymbol{B}},\ 1\right\}\bigl(\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}) - \mathrm{OPT}(r,\boldsymbol{c},\mathbf{0})\bigr),
\end{aligned}$$

where $\mathrm{OPT}(r,\boldsymbol{c},\mathbf{0})$ refers to the expect reward achieved by the null-cost action a_{null}. After combination with (23) and summation, we get that on $\mathcal{E}_{\text{H-Az1}} \cap \mathcal{E}_\beta \cap \mathcal{E}_{\text{H-Az2}} \cap \mathcal{E}_{\text{TVD}}$:

$$\begin{aligned}
&\sum_{t=1}^{T} r_t - T\,\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}) \\
&\quad \geqslant -T\overline{\Delta}_{T,\delta/4} - \Xi_{T,\delta/4} - \sum_{t=1}^{T} \min\left\{\frac{\overline{\Delta}_{T,\delta/4} + \xi_{t-1,\delta/4}}{\min \boldsymbol{B}},\ 1\right\}\bigl(\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}) - \mathrm{OPT}(r,\boldsymbol{c},\mathbf{0})\bigr) \\
&\quad \geqslant -\bigl(T\overline{\Delta}_{T,\delta/4} + \Xi_{T,\delta/4}\bigr)\left(1 + \frac{\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}) - \mathrm{OPT}(r,\boldsymbol{c},\mathbf{0})}{\min \boldsymbol{B}}\right),
\end{aligned}$$

which corresponds to the stated regret bound. □

F.3 General analysis with "hard" constraints

We finally generalize Theorem 3 and get a result corresponding to Theorem 1 combined with Lemma 3.

Here, we "mix" the slacks $+\xi_{t-1,\delta/4}$ and $-\overline{\Delta}_{T,\delta/4}$ of Appendices F.1 and F.2, with a slight modification of $\overline{\Delta}_{T,\delta/4}$ to compensate for the $\xi_{t-1,\delta/4}$: we rather have a $2\Xi_{T,\delta/4}$ term in the numerator of $\overline{\Delta}'_{T,\delta/4}$, compared to the $\Xi_{T,\delta/4}$ term in the numerator of $\overline{\Delta}_{T,\delta/4}$.

Theorem 4 (hard constraints)**.** *Fix $\delta \in (0,1)$. We consider the strategy of Box D, run with $\delta/4$ and signed slacks (depending on $t \geqslant 1$)*

$$b_t = -\overline{\Delta}'_{T,\delta/4} + \xi_{t-1,\delta/4} \qquad \textit{where} \qquad \overline{\Delta}'_{T,\delta/4} = \frac{2\alpha_{T,\delta/4} + \beta_{T,\delta/4} + 2\Xi_{T,\delta/4}}{T}$$

and $\alpha_{T,\delta/4} = \sqrt{2T\ln\bigl((d+1)/(\delta/4)\bigr)}$. Assume that the problem is $(\boldsymbol{B}-\boldsymbol{m})$–feasible for some $\boldsymbol{m}$ that does not need to be known by the strategy, with $\boldsymbol{B}-\boldsymbol{m} \leqslant \boldsymbol{B} - \overline{\Delta}'_{T,\delta/4}\mathbf{1}$. Then, under Assumption 2 and with Notation (1)*, with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} \boldsymbol{c}_t \leqslant T\boldsymbol{B}$$

$$\textit{and} \qquad R_T \leqslant \bigl(2\alpha_{T,\delta/4} + \beta_{T,\delta/4} + 2\Xi_{T,\delta/4}\bigr)\left(1 + \frac{\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}) - \mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}-\boldsymbol{m})}{\min \boldsymbol{m}}\right).$$

Proof. We rather sketch the differences to the proofs of Theorem 3 and Proposition 1. We denote by $\mathcal{E}_{\text{all}}$ the event of probability at least $1-\delta$ obtained as the intersection of four convenient events of probability each at least $1-\delta/4$. All inequalities below hold on $\mathcal{E}_{\text{all}}$ but for the sake of brevity, we will not highlight this fact each time.

We denote by $\boldsymbol{\pi}^{\text{feas}}$ a (quasi-)optimal static policy among the ones achieving expected costs smaller than $\boldsymbol{B}-\boldsymbol{m}$; it therefore ensures that its expected costs are smaller than $\boldsymbol{B}-\boldsymbol{m}$ and achieves an expected reward of $\mathrm{OPT}(r,\boldsymbol{c},\boldsymbol{B}-\boldsymbol{m}) - e$, possibly up to a small factor $e \geqslant 0$ which we let vanish. We note that for each round $t \geqslant 1$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta,t-1}}\!\left[\sum_{a\in\mathcal{A}} \widehat{\boldsymbol{c}}^{\,\text{lcb}}_{\delta,t-1}(\boldsymbol{X},a)\,\pi^{\text{feas}}_a(\boldsymbol{X})\right] &\leqslant \xi_{t-1,\delta/T}\mathbf{1} + \mathbb{E}_{\boldsymbol{X}\sim\nu}\!\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(\boldsymbol{X},a)\,\pi^{\text{feas}}_a(\boldsymbol{X})\right]\\ &\leqslant \xi_{t-1,\delta/T}\mathbf{1} + \boldsymbol{B} - \boldsymbol{m}\mathbf{1} \leqslant \boldsymbol{B} + b_t\mathbf{1}\,,\end{aligned}$$

given the definition $b_t = -\overline{\Delta}'_{T,\delta/4} + \xi_{t-1,\delta/4}$ and the assumption $\boldsymbol{B}-\boldsymbol{m} \leqslant \boldsymbol{B} - \overline{\Delta}'_{T,\delta/4}\mathbf{1}$. Thus, on $\mathcal{E}_{\text{all}}$, the strategy $\boldsymbol{\pi}_t$ is indeed defined at each round $t \geqslant 1$ by the optimization problem stated in Step 2 of Box D (and not in an arbitrary manner).

Cost-wise, we thus have

$$\begin{aligned}\sum_{t=1}^{T} \boldsymbol{c}_t &\leqslant \bigl(2\alpha_{T,\delta/4} + \beta_{T,\delta/4} + \Xi_{T,\delta/4}\bigr)\mathbf{1} + \sum_{t=1}^{T} \overbrace{\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\!\left[\sum_{a\in\mathcal{A}} \widehat{\boldsymbol{c}}^{\,\text{lcb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\pi_{t,a}(\boldsymbol{X})\right]}^{\leqslant \boldsymbol{B}+b_t\mathbf{1}}\\ &\leqslant T\boldsymbol{B} + \bigl(2\alpha_{T,\delta/4} + \beta_{T,\delta/4} + \Xi_{T,\delta/4}\bigr)\mathbf{1} - T\overline{\Delta}'_{T,\delta/4} + \Xi_{T,\delta/4} = T\boldsymbol{B}\,,\end{aligned}$$

where the final equality follows from the definition of $\overline{\Delta}'_{T,\delta/4}$, which involves a $2\Xi_{T,\delta/4}$ term in its numerator.

Reward-wise, we introduce for each $t \geqslant 1$,

$$\overline{\boldsymbol{\pi}}_t = (1-w_t)\boldsymbol{\pi}^\star_{\boldsymbol{B}} + w_t\boldsymbol{\pi}^{\text{feas}}\,, \qquad \text{where} \quad w_t = \min\left\{\frac{\overline{\Delta}'_{T,\delta/4}}{\min \boldsymbol{m}},\ 1\right\},$$

whose empirical expected cost at round $t \geqslant 1$ is smaller than

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\!\left[\sum_{a\in\mathcal{A}} \widehat{\boldsymbol{c}}^{\,\text{lcb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\overline{\pi}_{t,a}(\boldsymbol{X})\right] &\leqslant \xi_{t-1,\delta/4}\mathbf{1} + \mathbb{E}_{\boldsymbol{X}\sim\nu}\!\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(\boldsymbol{X},a)\,\overline{\pi}_{t,a}(\boldsymbol{X})\right]\\ &\leqslant \xi_{t-1,\delta/4}\mathbf{1} + (1-w_t)\boldsymbol{B} + w_t(\boldsymbol{B}-\boldsymbol{m})\\ &= \boldsymbol{B} + \xi_{t-1,\delta/4}\mathbf{1} - w_t\boldsymbol{m}\,.\end{aligned}$$

By using $\boldsymbol{m} \geqslant (\min \boldsymbol{m})\mathbf{1}$ and thanks to the assumption $\boldsymbol{B} - \boldsymbol{m} \leqslant \boldsymbol{B} - \overline{\Delta}'_{T,\delta/4}\mathbf{1}$, which can be equivalently formulated as $\min \boldsymbol{m} \geqslant \overline{\Delta}'_{T,\delta/4}$, we may continue this series of inequalities by

$$\begin{aligned} \boldsymbol{B} + \xi_{t-1,\delta/4}\mathbf{1} - w_t \boldsymbol{m} &\leqslant \boldsymbol{B} - \min\Bigl\{\overline{\Delta}'_{T,\delta/4} - \xi_{t-1,\delta/4},\ \min \boldsymbol{m} - \xi_{t-1,\delta/4}\Bigr\}\,\mathbf{1} \\ &\leqslant \boldsymbol{B} - \min\Bigl\{\overline{\Delta}'_{T,\delta/4} - \xi_{t-1,\delta/4},\ \overline{\Delta}'_{T,\delta/4} - \xi_{t-1,\delta/4}\Bigr\}\,\mathbf{1} = \boldsymbol{B} + b_t \mathbf{1}\,, \end{aligned}$$

meaning that $\overline{\boldsymbol{\pi}}_t$ is a policy satisfying the constraints stated in Step 2 of Box D on round $t \geqslant 1$. In particular, by definition of $\boldsymbol{\pi}_t$ as a maximizer,

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\left[\sum_{a\in\mathcal{A}} \widehat{r}^{\,\mathrm{ucb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\pi_{t,a}(\boldsymbol{X})\right] \\ &\qquad \geqslant \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\left[\sum_{a\in\mathcal{A}} \widehat{r}^{\,\mathrm{ucb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\overline{\pi}_{t,a}(\boldsymbol{X})\right] \\ &\qquad \geqslant -\xi_{t-1,\delta/4} + \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(\boldsymbol{X},a)\,\overline{\pi}_{t,a}(\boldsymbol{X})\right] \\ &\qquad \geqslant -\xi_{t-1,\delta/4} + \bigl((1-w_t)\mathrm{OPT}(r,\boldsymbol{c}\,\boldsymbol{B}) + w_t \mathrm{OPT}(r,\boldsymbol{c}\,\boldsymbol{B}-\boldsymbol{m})\bigr) \\ &\qquad = \mathrm{OPT}(r,\boldsymbol{c}\,\boldsymbol{B}) - \min\left\{\frac{\overline{\Delta}'_{T,\delta/4}}{\min \boldsymbol{m}},\ 1\right\}\bigl(\mathrm{OPT}(r,\boldsymbol{c}\,\boldsymbol{B}) - \mathrm{OPT}(r,\boldsymbol{c}\,\boldsymbol{B}-\boldsymbol{m})\bigr) - \xi_{t-1,\delta/4}\,. \end{aligned} \tag{24}$$

Finally, by combining (23) and (24),

$$\begin{aligned} \sum_{t=1}^{T} r_t &\geqslant -\bigl(\overbrace{2\alpha_{T,\delta/4} + \beta_{T,\delta/4} + \Xi_{T,\delta/4}}^{=T\overline{\Delta}'_{T,\delta/4}-\Xi_{T,\delta/4}}\bigr)\mathbf{1} + \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_{\delta/4,t-1}}\left[\sum_{a\in\mathcal{A}} \widehat{r}^{\,\mathrm{ucb}}_{\delta/4,t-1}(\boldsymbol{X},a)\,\pi_{t,a}(\boldsymbol{X})\right] \\ &\qquad \geqslant T\,\mathrm{OPT}(r,\boldsymbol{c}\,\boldsymbol{B}) + \Xi_{T,\delta/4} - T\overline{\Delta}'_{T,\delta/4}\left(1 + \frac{\mathrm{OPT}(r,\boldsymbol{c}\,\boldsymbol{B}) - \mathrm{OPT}(r,\boldsymbol{c}\,\boldsymbol{B}-\boldsymbol{m})}{\min \boldsymbol{m}}\right) - \Xi_{T,\delta/4}, \end{aligned}$$

as claimed. □

G Numerical experiments: full description

This appendix reports numerical simulations performed on simulated data with the motivating example described in Chohlas-Wood et al. [2021] and alluded at in Section 2.1. These simulations are for the sake of illustration only.

A brief summary of the applicative background in AI for justice is the following. The learner wants to maximize the total number of appearances to court for people of concern. To achieve this goal, the learner is able to provide, or not, some transportation assistance: rideshare assistance (the highest level of help), or a transit voucher (a more modest level of help). There are of course budget limits on these assistance means, and the learner also wants to control how (un)fair the assistance policy is, in terms of subgroups of the population, while maximizing the total number of appearances. Some subgroups take a better advantage of assistance to appear in court, thus, without the fairness control, all assistance would go to these groups. The fairness costs described in Section 2.1 force the learner to perform some tradeoff between spending all assistance on most reactive subgroups and spending it equally among subgroups.

Outline of this appendix. We first recall the experimental setting of Chohlas-Wood et al. [2021]—in particular, how contexts, rewards, and costs are generated in their simulations (we cannot replicate their study on real data, which is not accessible). We then specify the strategies we implemented and how we tuned them—this includes describing the estimation procedure discussed in Section 2.2. We finally report the performance observed, in terms of (average) rewards and costs.

G.1 The experimental setting of Chohlas-Wood et al. [2021]

We follow strictly the experimental setting of Chohlas-Wood et al. [2021] as provided in the public repository `https://github.com/stanford-policylab/learning-to-be-fair`. This experimental setting, as reverse-engineered from the code, deals with purely simulated data and seems actually inconsistent with the description made in Chohlas-Wood et al. [2021, Section 5.4 and Appendix E], which would anyway rely on proprietary data that could not be made public.

Context generation. Each individual is described by four variables $\boldsymbol{x}$: age, proximity, poverty, and group, which are to be read in $\boldsymbol{x}$ in its components, referred to as x_{age}, x_{prox}, x_{pov} and $\operatorname{gr}(\boldsymbol{x}) = x_{\text{group}}$. The first three variables are assumed to be normalized and are simulated independently, according to uniform distributions on $[0,1]$. Groups are also simulated independently of these three variables: two groups are assumed, with respective probabilities of $1/2$. This defines the context distribution ν.

As we describe now, assistance has stronger impact on group 0 than in group 1.

Cost generation. We recall that three actions are available: offering some rideshare assistance (action a_{ride}), providing a transportation voucher (action a_{voucher}), or providing no help (which is a control situation: action a_{control}). The associated spending costs are deterministic and do not depend on $\boldsymbol{x}$, and there are two separate budgets for rideshares and vouchers. More precisely, at round t, upon taking action a_t, the following deterministic spending costs are suffered:

$$c_{\text{ride}}(\boldsymbol{x}_t, a_t) = \mathbb{1}_{\{a_t = a_{\text{ride}}\}} \qquad \text{and} \qquad c_{\text{voucher}}(\boldsymbol{x}_t, a_t) = \mathbb{1}_{\{a_t = a_{\text{voucher}}\}}\,,$$

where the corresponding average budgets are $B_{\text{ride}} = 0.05$ and $B_{\text{voucher}} = 0.20$.

We measure the extent of unfair allocation of spendings among groups with the following eight fairness costs: a first series of four fairness costs is given by

$$\begin{aligned}
2\,\mathbb{1}_{\{a_t=a_{\text{ride}}\}}\mathbb{1}_{\{\operatorname{gr}(\boldsymbol{x}_t)=0\}} - \mathbb{1}_{\{a_t=a_{\text{ride}}\}}\,, &\qquad 2\,\mathbb{1}_{\{a_t=a_{\text{ride}}\}}\mathbb{1}_{\{\operatorname{gr}(\boldsymbol{x}_t)=1\}} - \mathbb{1}_{\{a_t=a_{\text{ride}}\}}\,,\\
2\,\mathbb{1}_{\{a_t=a_{\text{voucher}}\}}\mathbb{1}_{\{\operatorname{gr}(\boldsymbol{x}_t)=0\}} - \mathbb{1}_{\{a_t=a_{\text{voucher}}\}}\,, &\qquad 2\,\mathbb{1}_{\{a_t=a_{\text{voucher}}\}}\mathbb{1}_{\{\operatorname{gr}(\boldsymbol{x}_t)=1\}} - \mathbb{1}_{\{a_t=a_{\text{voucher}}\}}\,,
\end{aligned}$$

and the second series is given by the opposite values of these costs. We denote by τ the corresponding average cost constraints τ, and set $\tau = 10^{-7}$ or $\tau = 0.025$ in our experiments.

All in all, the global (spending and fairness) vector costs $\boldsymbol{c}_t$ takes deterministic values in $\mathbb{R}^{10}$. The vector cost function $\boldsymbol{c}$ is fully known to the learner, and no estimation is needed.

Reward generation. The rewards are binary: $r_t = 1$ if the t–individual appeared in court, and $r_t = 0$ otherwise. That is, the expected reward equals the probability of appearance. A logistic regression model is assumed: denoting by $\Phi(u) = 1/(1+\mathrm{e}^u)$, we assume

$$\begin{aligned}
r(\boldsymbol{x}, a_{\text{control}}) &= \Phi(-x_{\text{age}}),\\
r(\boldsymbol{x}, a_{\text{voucher}}) &= \Phi(-x_{\text{age}} + 2x_{\text{prox}})\mathbb{1}_{\{\operatorname{gr}(\boldsymbol{x})=0\}} + \Phi(-x_{\text{age}} + x_{\text{prox}})\mathbb{1}_{\{\operatorname{gr}(\boldsymbol{x})=1\}}\,,\\
r(\boldsymbol{x}, a_{\text{ride}}) &= \Phi(-x_{\text{age}} + 4x_{\text{pov}})\mathbb{1}_{\{\operatorname{gr}(\boldsymbol{x})=0\}} + \Phi(-x_{\text{age}} + 2x_{\text{pov}})\mathbb{1}_{\{\operatorname{gr}(\boldsymbol{x})=1\}}\,.
\end{aligned}$$

We may write these six equalities in a compact format as:

$$r(\boldsymbol{x}, a) = \Phi\big(\varphi(\boldsymbol{x}, a)^{\mathrm{T}}\,\mu_\star\big)$$

$$\text{where} \qquad \boldsymbol{\varphi}(\boldsymbol{x}, a) = \begin{bmatrix} x_{\text{age}}\\ x_{\text{prox}}\,\mathbb{1}_{\{a=a_{\text{voucher}}\}}\\ x_{\text{prox}}\,\mathbb{1}_{\{a=a_{\text{voucher}}\}}\,\mathbb{1}_{\{\operatorname{gr}(\boldsymbol{x})=0\}}\\ x_{\text{pov}}\,\mathbb{1}_{\{a=a_{\text{ride}}\}}\\ x_{\text{pov}}\,\mathbb{1}_{\{a=a_{\text{ride}}\}}\,\mathbb{1}_{\{\operatorname{gr}(\boldsymbol{x})=0\}} \end{bmatrix} \qquad \text{and} \qquad \mu_\star = \begin{bmatrix} -1\\ 1\\ 1\\ 2\\ 2 \end{bmatrix}.$$

The learner ignores the coefficients $\mu_\star$ of this structure and only knows that expected rewards are of the form $\Phi\big(\varphi(\boldsymbol{x}, a)^{\mathrm{T}}\boldsymbol{\theta}\big)$ for some parameter $\boldsymbol{\theta} \in \mathbb{R}^5$. The learner will estimate the reward function r by estimating $\mu_\star$.

Reward estimation. We deal with a logistic model and follow the methodology described in Li and Stoltz [2022]; see Modeling 2 in Section 2.2. In particular, the parameters $\mu_\star$ are estimated, after each round $t \geqslant 1$, thanks to the maximum likelihood estimator

$$\widehat{\boldsymbol{\mu}}_t \in \operatorname*{arg\,max}_{\boldsymbol{\mu}\in\mathbb{R}^5} \sum_{s=1}^{t} r_s \Phi\big(\boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}}\boldsymbol{\mu}\big) + (1-r_s)\ln\Big(1 - \Phi\big(\boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}}\boldsymbol{\mu}\big)\Big) - \frac{\lambda^{\text{logistic}}}{2}\|\boldsymbol{\mu}\|^2\,,$$

where $\lambda^{\text{logistic}} \geqslant 0$ is a regularization factor. We define as $\widehat{r}_t(\boldsymbol{x}, a) = \Phi\big(\boldsymbol{\varphi}(\boldsymbol{x}, a)^{\mathrm{T}}\,\widehat{\boldsymbol{\mu}}_t\big)$ the estimated expected rewards, and the associated uniform estimation errors (see Assumption 2) are of the form

$$\varepsilon_t(\boldsymbol{x}, a, \delta) = C_\delta\big(1 + \ln t\big)\sqrt{\boldsymbol{\varphi}(a, \boldsymbol{x})^{\mathrm{T}}\, V_t^{-1} \boldsymbol{\varphi}(a, \boldsymbol{x})}$$

$$\text{where} \qquad V_t = \sum_{s=1}^{t} \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)\, \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}} + \lambda^{\text{logistic}} \mathrm{I}_5\,,$$

where I_5 is the 5×5 identity matrix.

In our simulations, we tested a range of values and picked (in hindsight) the well-performing values $C_\delta = 0.025$ and $\lambda^{\text{logistic}} = 0$.

G.2 Strategies implemented in this numerical study

We run all these strategies not with the total-cost constraints

$$\boldsymbol{B} = (0.05,\, 0.2,\, \tau,\, \tau,\, \tau,\, \tau)\,, \qquad \text{where} \qquad \tau \in \{10^{-7},\, 0.025\}\,,$$

but take a margin $b = 0.005$ and use $\boldsymbol{B}' = \boldsymbol{B} - (b, b, 0, 0, 0, 0)$ instead. This is a slightly different way of taking some margin on the average cost constraint: we do so because we are not aiming for a strict respect of the fairness constraints but rather want to report the level of violation on it. Again, we tried a range of values for b (between 0.001 to 0.01) and this value of 0.005 led to a good balance between (lack of) total-cost constraint violations and rewards.

Performance of optimal static policies. We use $\mathrm{OPT}(r, \boldsymbol{c}, \boldsymbol{B})$ as the benchmark in the definition of regret; our methodology also reveals that $\mathrm{OPT}(r, \boldsymbol{c}, \boldsymbol{B}')$ is another benchmark, see the discussion in Section 4. We report both values on our graphs. To compute them, we proceed as follows, e.g., for $\boldsymbol{B}$. As computing directly the minimum stated in (4) is difficult, even when fully knowing the distribution ν, we compute 100 estimates

$$\mathrm{OPT}^{(j)}(r, \boldsymbol{c}, \boldsymbol{B})\,, \qquad \text{which we average out into} \qquad \widehat{\mathrm{OPT}}(r, \boldsymbol{c}, \boldsymbol{B}) = \frac{1}{100}\sum_{j=1}^{100} \mathrm{OPT}^{(j)}(r, \boldsymbol{c}, \boldsymbol{B})\,.$$

For each j, we sample $S = 10{,}000$ contexts from the distribution ν, and denote by $\widehat{\nu}_S^{(j)}$ the associated empirical distribution; we then solve numerically the problem (4) with ν replaced by this empirical distribution:

$$\mathrm{OPT}^{(j)}(r, \boldsymbol{c}, \boldsymbol{B}) = \min_{\boldsymbol{\lambda} \geqslant 0}\ \mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_S^{(j)}}\left[\max_{a \in \mathcal{A}} \Big\{ r(\boldsymbol{X}, a) - \big\langle \boldsymbol{c}(\boldsymbol{X}, a) - \boldsymbol{B},\, \boldsymbol{\lambda} \big\rangle \Big\}\right].$$

Mixed policy knowing $\boldsymbol{\lambda}^\star_{\boldsymbol{B}'}$ but estimating r. For the sake of illustration, we report the performance of a policy that would have oracle knowledge of $\boldsymbol{\lambda}^\star_{\boldsymbol{B}'}$, which is a finite-dimensional parameter that summarizes ν, but would ignore r, i.e., the underlying logistic model. That is, this mixed policy would pick, at each round,

$$\max_{a \in \mathcal{A}} \Big\{ \widehat{r}^{\,\text{ucb}}_{\delta, t-1}(\boldsymbol{x}_t, a) - \big\langle \widehat{\boldsymbol{c}}^{\,\text{lcb}}_{\delta, t-1}(\boldsymbol{x}_t, a) - \boldsymbol{B}',\, \boldsymbol{\lambda}^\star_{\boldsymbol{B}'} \big\rangle \Big\}\,,$$

(and would omit the $\boldsymbol{\lambda}$ update in Box B).

To compute (an approximation of) $\boldsymbol{\lambda}^\star_{\boldsymbol{B}'}$, we proceed 100 times as described below to compute estimates

$$\boldsymbol{\lambda}^{\star,(j)}_{\boldsymbol{B}'}\,, \qquad \text{which we average out into} \qquad \widehat{\boldsymbol{\lambda}}^\star_{\boldsymbol{B}'} = \frac{1}{100}\sum_{j=1}^{100} \boldsymbol{\lambda}^{\star,(j)}_{\boldsymbol{B}'}\,,$$

where we noted that the numerical values obtained for the $\boldsymbol{\lambda}^{\star,(j)}_{\boldsymbol{B}'}$ are rather similar. With the notation above, for each j, we sample $S = 10{,}000$ contexts from the distribution ν, and solve

$$\boldsymbol{\lambda}^{\star,(j)}_{\boldsymbol{B}'} \in \operatorname*{arg\,min}_{\boldsymbol{\lambda} \geqslant \boldsymbol{0}} \mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_S^{(j)}}\left[\max_{a \in \mathcal{A}} \Big\{ r(\boldsymbol{X}, a) - \big\langle \boldsymbol{c}(\boldsymbol{X}, a) - \boldsymbol{B},\, \boldsymbol{\lambda} \big\rangle \Big\}\right].$$

(These estimations are independent from the estimations used to compute the OPT values, i.e., we use different seeds.)

PGD γ. We refer to the Box B strategy as PGD γ, for which we report the performance for values $\gamma \in \{0.01,\, 0.02,\, 0.04,\, 0.05,\, 0.1\}$.

We also implemented the Box C strategy, with an alternative value of $M_{T,\delta,k}$, given that the one exhibited based on the theoretical analysis was too conservative, so that no regime break occurred. We resort to a value of the same order of magnitude:

$$M'_{T,\delta,k} = c\, d \sqrt{T \ln\bigl(T(k+2)\bigr)}\,,$$

for a numerical constant c that we set to 0.01 in our simulations. We call this strategy PGD Adaptive in Figure 1 and Table 1.

G.3 Outcomes of the simulations

We take $T = 10{,}000$ individuals (instead of $T = 1{,}000$ as in the code by Chohlas-Wood et al., 2021) and set initial 50 rounds as a warm start for strategies (mostly because of the logistic estimation). We were limited by the computational power (see Appendix G.4) and could only perform $N = 100$ simulations for each (instance of each) strategy. We report averages (strong lines in the graphs) as well as ± 2 times standard errors (shaded areas in the graphs or values in parentheses in the table).

Graphs. In the first line of graphs in Figure 1, we report the average rewards (over the N runs) of the strategy under scrutiny as a function of the sample size. More precisely, with obvious notation, we plot

$$t \mapsto \frac{1}{N} \sum_{n=1}^{N} \left(\frac{1}{t} \sum_{s=1}^{t} r_s^{(n)} \right).$$

We report the values of $\textsc{opt}(r, \boldsymbol{c}, \boldsymbol{B})$ and $\textsc{opt}(r, \boldsymbol{c}, \boldsymbol{B'})$ as dashed horizontal lines.

In the second and third lines of graphs in Figure 1, we report the average costs suffered; again, with obvious notation, we plot

$$t \mapsto \frac{1}{N} \sum_{n=1}^{N} \left(\frac{1}{t} \sum_{s=1}^{t} \mathbb{1}_{\{a_s^{(n)} = a_{\text{ride}}\}} \right) \qquad \text{and} \qquad t \mapsto \frac{1}{N} \sum_{n=1}^{N} \left(\frac{1}{t} \sum_{s=1}^{t} \mathbb{1}_{\{a_s^{(n)} = a_{\text{voucher}}\}} \right).$$

We include the average budget constraints $B_{\text{ride}} = 0.05$ and $B_{\text{voucher}} = 0.20$ as dashed horizontal lines.

Finally, the fourth line of the graphs in Figure 1 reports the fairness costs; we average their absolute values and draw, again with obvious notation,

$$t \mapsto \frac{1}{N} \sum_{n=1}^{N} \left(\frac{1}{4} \sum_{a \in \{a_{\text{ride}}, a_{\text{voucher}}\}} \sum_{g \in \{0,1\}} \left| \frac{1}{t} \sum_{s=1}^{t} \Bigl(2\mathbb{1}_{\{a_s^{(n)} = a\}} \mathbb{1}_{\{\mathrm{gr}(\boldsymbol{x}_s^{(n)}) = g\}} - \mathbb{1}_{\{a_s^{(n)} = a\}} \Bigr) \right| \right).$$

We include the fairness tolerance τ as a dashed horizontal line.

Table. We also report the performance of the strategies at the final round T in following table, with ± 2 times standard errors in parentheses. We note that the performance of the mixed policy is poor, in particular in terms of fairness costs, which is why we omitted it on the graphs, to keep them readable.

Comments. When γ is well set, i.e., large enough (see the bound of Lemma 1), the PGD strategies control spending costs thanks to targeting $\boldsymbol{B'}$ instead of $\boldsymbol{B}$. We observe that for $\gamma = 0.01$, the PGD strategy does not control the rideshare costs, but for all larger values of γ, the associated strategies control the three costs considered. The average rewards achieved are coherent with the average spendings: the smaller the average spendings, the smaller the average rewards. There is some lag: the strategies tuned with γ parameters in $\{0.04,\, 0.05,\, 0.1\}$ could use costly actions more. We also observe that fairness costs remain under the target limits.

The mixed policy does not success in controlling the costs, in particular, the fairness costs. While the optimal dual variables $\boldsymbol{\lambda}^\star_{\boldsymbol{B'}}$ summarize the distribution ν, it seems that $\boldsymbol{\lambda}^\star_{\boldsymbol{B'}}$ has to be used with care: only with the exact values r and $\boldsymbol{c}$, and not with estimates. On the contrary, the PGD strategies of Box B are more stable, as the dual variables are learned based also on the estimated reward and cost functions.

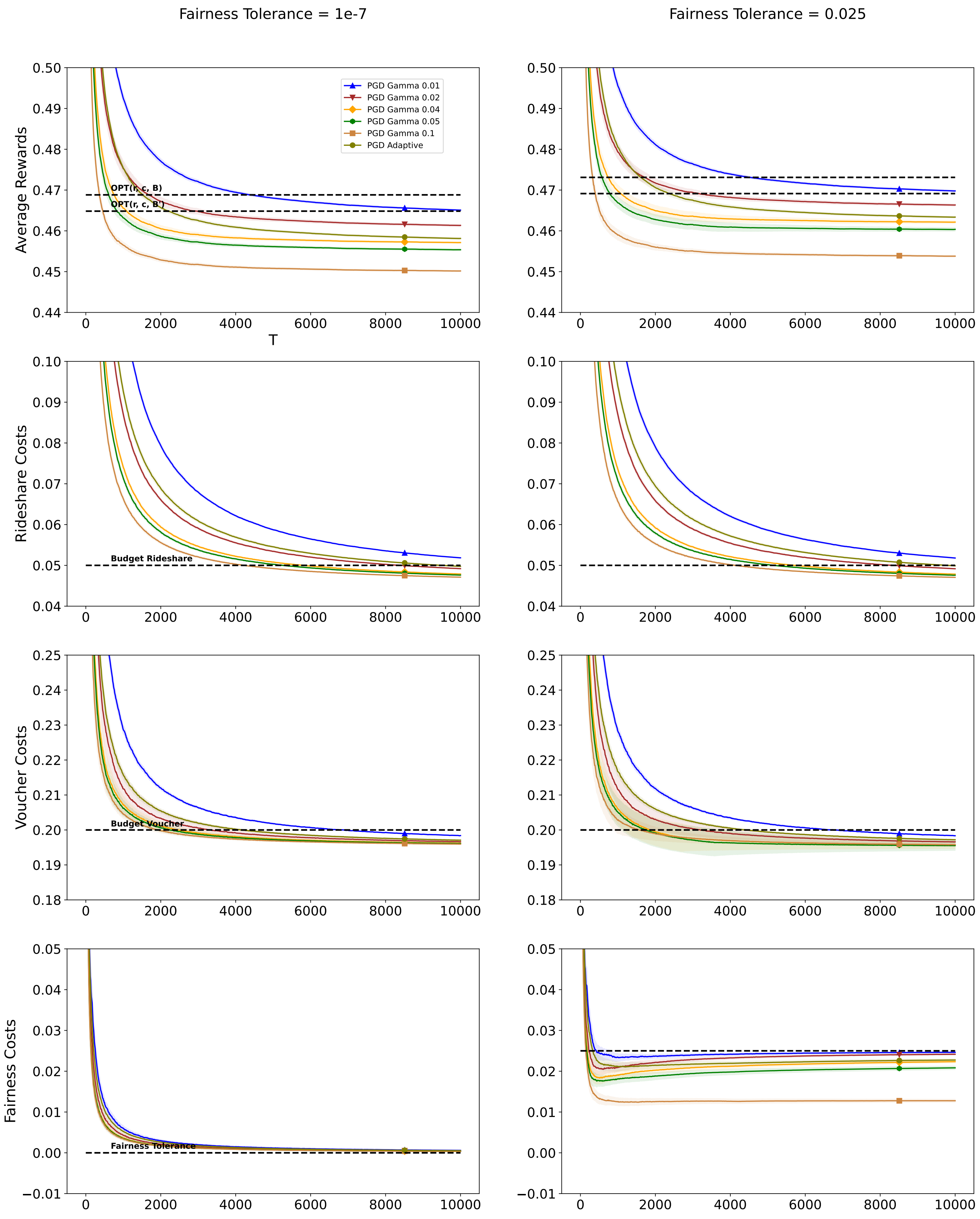

Figure 1: Performance of the PGD strategies.

Table 1: Performance of the strategies as computed at the final round T.

	Average rewards	Rideshare costs	Voucher costs	Fairness costs
	Fairness tolerance $\tau = 10^{-7}$			
$\textsc{opt}(r, \boldsymbol{c}, \boldsymbol{B})$	0.4688 (0.0002)			
$\textsc{opt}(r, \boldsymbol{c}, \boldsymbol{B}')$	0.4648 (0.0002)			
PGD $\gamma = 0.01$	0.4651 (0.0002)	0.0519 (<0.0001)	0.1984 (<0.0001)	0.0006 (<0.0001)
PGD $\gamma = 0.02$	0.4613 (0.0002)	0.0492 (<0.0001)	0.1967 (0.0004)	0.0004 (<0.0001)
PGD $\gamma = 0.04$	0.4571 (0.0002)	0.0479 (<0.0001)	0.1962 (0.0002)	0.0004 (<0.0001)
PGD $\gamma = 0.05$	0.4554 (0.0002)	0.0476 (<0.0001)	0.1961 (0.0002)	0.0003 (<0.0001)
PGD $\gamma = 0.1$	0.4502 (0.0002)	0.0471 (<0.0001)	0.196 (0.0002)	0.0003 (<0.0001)
PGD Adaptive	0.4581 (0.0002)	0.0498 (0.0002)	0.1971 (0.0002)	0.0005 (<0.0001)
Mixed Policy	0.4402 (0.0056)	0.0499 (0.0058)	0.1056 (0.017)	0.0411 (0.0052)
	Fairness tolerance $\tau = 0.025$			
$\textsc{opt}(r, \boldsymbol{c}, \boldsymbol{B})$	0.4731 (0.0002)			
$\textsc{opt}(r, \boldsymbol{c}, \boldsymbol{B}')$	0.4691 (0.0002)			
PGD $\gamma = 0.01$	0.4698 (0.0002)	0.0518 (0.0002)	0.1983 (<0.0001)	0.0246 (0.0002)
PGD $\gamma = 0.02$	0.4663 (0.0002)	0.0492 (<0.0001)	0.1966 (0.0006)	0.0242 (0.0002)
PGD $\gamma = 0.04$	0.4621 (0.0004)	0.0478 (<0.0001)	0.1958 (0.001)	0.0223 (0.0004)
PGD $\gamma = 0.05$	0.4604 (0.0004)	0.0476 (<0.0001)	0.1955 (0.0014)	0.0208 (0.0004)
PGD $\gamma = 0.1$	0.4538 (0.0002)	0.0471 (<0.0001)	0.1958 (0.0004)	0.0128 (0.0004)
PGD Adaptive	0.4634 (0.0002)	0.0499 (0.0002)	0.1972 (0.0002)	0.0228 (0.0002)
Mixed Policy	0.4466 (0.0054)	0.0566 (0.0052)	0.1053 (0.0164)	0.0473 (0.0054)

Finally, we note that in our experiments, the regimes in PGD Adaptive strategy typically covered range from $k = 0$ (corresponding to $\gamma_0 = 1/\sqrt{T} = 0.01$) to $k = 2$ (corresponding to $\gamma_2 = 2^2/\sqrt{T} = 0.04$). The PGD Adaptive strategy performs well and is (only) outperformed by the PDG strategy with a fixed $\gamma = 0.02$ (of course difficult to pick in advance). In particular, the PGD Adaptive strategy controls costs, and does so by switching to larger step sizes when needed.

G.4 Computation time and environment

As requested by the NeurIPS checklist, we provide details on the computation time and environment. Our experiments were ran on the following hardware environment: no GPU was required, CPU is 3.3 GHz 8 Cores with total of 16 threads, and RAM is 32 GB 4800 MHz DDR5. We ran 100 simulations with 10 different seeds on parallel each time. In the setting and for the data described above, the average time spend on each algorithm for a single run was inferior to 10 minutes.