A Related Works

In the following, we survey some previous works that are tightly related to ours. In particular, we first describe works dealing with the online learning problem in MDPs, and, then, we discuss some works studying the constrained version of the classical online learning problem.

Online Learning in MDPs. There is a considerable literature on online learning problems (Cesa-Bianchi & Lugosi, 2006) in MDPs (see (Auer et al., 2008; Even-Dar et al., 2009; Neu et al., 2010) for some initial results on the topic). In such settings, two types of feedback are usually investigated: in the *full-information feedback* model, the entire loss function is observed after the learner's choice, while in the *bandit feedback* model, the learner only observes the loss due to the chosen action. Azar et al. (2017) study the problem of optimal exploration in episodic MDPs with unknown transitions and stochastic losses when the feedback is bandit. The authors present an algorithm whose regret upper bound is $\tilde{\mathcal{O}}(\sqrt{T})$, thus matching the lower bound for this class of MDPs and improving the previous result by Auer et al. (2008). Rosenberg & Mansour (2019b) study the online learning problem in episodic MDPs with adversarial losses and unknown transitions when the feedback is full information. The authors present an online algorithm exploiting entropic regularization and providing a regret upper bound of $\tilde{\mathcal{O}}(\sqrt{T})$. The same setting is investigated by Rosenberg & Mansour (2019a) when the feedback is bandit. In such a case, the authors provide a regret upper bound of the order of $\tilde{\mathcal{O}}(T^{3/4})$, which is improved by Jin et al. (2020) by providing an algorithm that achieves in the same setting a regret upper bound of $\tilde{\mathcal{O}}(\sqrt{T})$.

Online Learning in CMDPs with Long-term Constraints. All the previous works on the topic study settings in which constraints are selected stochastically. In particular, Zheng & Ratliff (2020) deal with episodic CMDPs with stochastic losses and constraints, where the transition probabilities are known and the feedback is bandit. The regret upper bound of their algorithm is of the order of $\tilde{\mathcal{O}}(T^{3/4})$, while the cumulative constraint violation is guaranteed to be below a threshold with a given probability. Wei et al. (2018) deal with adversarial losses and stochastic constraints, assuming the transition probabilities are known and the feedback is full information. The authors present an algorithm that guarantees an upper bound of the order of $\tilde{\mathcal{O}}(\sqrt{T})$ on both regret and constraint violation. Bai et al. (2020) provide the first algorithm that achieves sublinear regret when the transition probabilities are unknown, assuming that the rewards are deterministic and the constraints are stochastic with a particular structure. Efroni et al. (2020) propose two approaches to deal with the exploration-exploitation dilemma in episodic CMDPs. These approaches guarantee sublinear regret and constraint violation when transition probabilities, rewards, and constraints are unknown and stochastic, while the feedback is bandit. Qiu et al. (2020) provide a primal-dual approach based on *optimism in the face of uncertainty*. This work shows the effectiveness of such an approach when dealing with episodic CMDPs with adversarial losses and stochastic constraints, achieving both sublinear regret and constraint violation with full-information feedback. Wei et al. (2023) and Ding & Lavaei (2023) consider the case in which rewards and constraints are non-stationary, assuming that their variation is bounded. Thus, their results are *not* applicable to general adversarial settings.

Online Learning with Long-term Constraints. A central result is provided by Mannor et al. (2009), who show that it is impossible to suffer from sublinear regret and sublinear constraint violation when an adversary chooses losses and constraints. Liakopoulos et al. (2019) try to overcome such an impossibility result by defining a new notion of regret. They study a class of online learning problems with long-term budget constraints that can be chosen by an adversary. The learner's regret metric is modified by introducing the notion of a *K-benchmark*, *i.e.*, a comparator that meets the problem's allotted budget over any window of length K. Castiglioni et al. (2022a;b) deal with the problem of online learning with stochastic and adversarial losses, providing the first *best-of-both-worlds* algorithm for online learning problems with long-term constraints.

B EVENTS

Here we state the events that we use in the rest of the Appendix.

The following event states that the true occupancy measure space is always contained in the confidence set:

Event $E^{\Delta}(\delta)$**:** $\Delta(M) \subseteq \cap_i \Delta(\mathcal{P}_i)$.

In particular, under $E^{\Delta}(\delta)$, we have that $q^{\circ}, q^{*} \in \cap_i \Delta(\mathcal{P}_i)$. $E^{\Delta}(\delta)$ holds with probability at least $1-\delta$ (See Lemma 5).

The following event states that the cumulative error after T episodes due to the difference between q^{P,π_t} and $q^{P^{\widehat{q}_t},\pi_t}$ is small enough:

Event $E^{\widehat{q}}(\delta)$**:** $\sum_{t=1}^{T} ||q_t - \widehat{q}_t||_1 \leq \mathcal{E}^q_\delta$, *where* $\mathcal{E}^q_\delta := 4L|X|\sqrt{2T\ln\left(\frac{1}{\delta}\right)} + 6L|X|\sqrt{2T|A|\ln\left(\frac{T|X||A|}{\delta}\right)} \leq \tilde{\mathcal{O}}(\sqrt{T})$.

In the next sections we will often condition on the intersection of the previous events:

Event $E^{\Delta,\widehat{q}}(\delta)$**:** $E^{\widehat{q}}(\delta) \cap E^{\Delta}(\delta)$

$E^{\Delta,\widehat{q}}(\delta)$ holds with probability at least $1-2\delta$ (See Lemma 2).

The next event states that, in case the rewards are stochastic, the reward accumulated is not too far from the mean reward accumulated.

Event $E^r_{q^*}(\delta)$**:** $\left|\sum_{t=1}^{T}\left(r_t - \overline{r}\right)^{\top} q^*\right| \leq \mathcal{E}^r_\delta$, *where* $\mathcal{E}^r_\delta = \frac{L}{\sqrt{2}}\sqrt{T\ln\left(\frac{2}{\delta}\right)} \leq \tilde{\mathcal{O}}\left(\sqrt{T}\right)$

$E^r_{q^*}(\delta)$ holds with probability at least $1-\delta$ (See Lemma 3).

For the stochastic constraint setting, we define the quantity $\mathcal{E}^G_{t_1,t_2,\delta} := 2L\sqrt{2(t_2-t_1+1)\ln\left(\frac{T^2}{\delta}\right)}$ and then two events bounding the cumulative difference between the dual utility with the average constraints and that with the sampled constraints.

Event $E^G_{q^\circ}(\delta)$**:** *for all* $[t_1..t_2] \subseteq [1..T]$, $\left|\sum_{t=t_1}^{t_2} \lambda_t^{\top}(G_t^{\top} - \overline{G}^{\top})q^{\circ}\right| \leq \lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2,\delta}$

Event $E^G_{q^*}(\delta)$**:** *for all* $[t_1..t_2] \subseteq [1..T]$, $\left|\sum_{t=t_1}^{t_2} \lambda_t^{\top}(G_t^{\top} - \overline{G}^{\top})q^{*}\right| \leq \lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2,\delta}$

$E^G_{q^\circ}(\delta)$, $E^G_{q^*}(\delta)$ each hold with probability at least $1-\delta$ (See Lemma 4). We denote $\mathcal{E}^G_\delta := \mathcal{E}^G_{1,T,\delta}$

C Additional Details and Omitted Proof of Section 4

C.1 Algorithm

Algorithm 4 Upper Confidence Online Gradient Descent Policy Search (UC-O-GDPS)

Require: state space X, action space A, episode number T, and confidence parameter δ

Initialize epoch index $i = 1$ and confidence set $\mathcal{P}_1$ as the set of all transition functions. For all $k \in [0..L-1]$ and all $(x, a, x') \in X_k \times A \times X_{k+1}$, initialize counters $N_0(x,a) = N_1(x,a) = M_0(x' \mid x, a) = M_1(x' \mid x, a) = 0$ and occupancy measure

$$\widehat{q}_1(x, a, x') = \frac{1}{|X_k||A||X_{k+1}|}$$

Initialize policy $\pi_1 = \pi^{\widehat{q}_1}$

for $t \in [T]$ **do**

  Execute policy π_t for L steps and obtain trajectory x_k, a_k for $k \in [0..L-1]$ and loss ℓ_t

  for $k \in [0..L-1]$ **do**

    Update counters:

$$N_i(x_k, a_k) \leftarrow N_i(x_k, a_k) + 1,$$
$$M_i(x_{k+1} \mid x_k, a_k) \leftarrow M_i(x_{k+1} \mid x_k, a_k) + 1$$

  end for

  if $\exists k, N_i(x_k, a_k) \geq \max\{1, 2N_{i-1}(x_k, a_k)\}$ **then**

    Increase epoch index $i \leftarrow i + 1$

    Initialize new counters: for all (x, a, x'),

$$N_i(x,a) = N_{i-1}(x,a)$$
$$M_i(x' \mid x, a) = M_{i-1}(x' \mid x, a)$$

    Update confidence set $\mathcal{P}_i$ based on Equation (6)

  end if

  Update occupancy measure:

  $\eta_t = \frac{1}{\overline{\ell}_t C\sqrt{T}}$ with $\overline{\ell}_t = \max\{||\ell_t||_\infty\}_{t=1}^t$

$$\widehat{q}_{t+1} = \Pi_{\Delta(\mathcal{P}_i)}(\widehat{q}_t - \eta_t \ell_t)$$

  Update policy $\pi_{t+1} = \pi^{\widehat{q}_{t+1}}$

end for

Confidence Set. The description of how Confidence Set on the Transition Probability functions are built and used, follows precisely the description of Rosenberg & Mansour (2019b). We report the functioning for completeness.

UC-O-GDPS keeps counters of visits of each state-action pair (x, a) and each state-action-state triple (x, a, x'), in order to estimate the empirical transition function as:

$$\overline{P}_i(x' \mid x, a) = \frac{M_i(x' \mid x, a)}{\max\{1, N_i(x,a)\}}$$

where $N_i(x,a)$ and $M_i(x' \mid x, a)$ are the initial values of the counters, that is, the total number of visits of pair (x, a) and triple (x, a, x') respectively, before epoch i. Epochs are used to reduce the computational complexity; in particular, a new epoch starts whenever there exists a state-action whose counter is doubled compared to its initial value at the beginning of the epoch. Next, the confidence set for epoch i is defined as:

$$\mathcal{P}_i = \left\{\widehat{P} : \left\|\widehat{P}(\cdot|x,a) - \overline{P}_i(\cdot|x,a)\right\|_1 \leq \epsilon_i(x,a) \quad \forall (x,a) \in X \times A\right\} \tag{7}$$

with $\epsilon_i(x,a)$ defined as:

$$\epsilon_i(x,a) = \sqrt{\frac{2|X_{k(x)+1}| \ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1, N_i(x,a)\}}}$$

using $k(x)$ for the index of the layer that x belongs to and for some confidence parameter $\delta \in (0,1)$. We state the following Lemma by Rosenberg & Mansour (2019b), which provides the results related to the confidence set $\epsilon_i(x,a)$.

Lemma 5. *Rosenberg & Mansour (2019b) For any $\delta \in [0,1]$:*

$$\left\|P\left(\cdot|x,a\right) - \overline{P}_i\left(\cdot|x,a\right)\right\|_1 \leq \sqrt{\frac{2|X_{k(x)+1}|\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\left\{1, N_i(x,a)\right\}}}$$

holds with probability at least $1-\delta$ simultaneously for all $(x,a) \in X \times A$ and all epochs.

Lemma 5 implies that, with high probability, the occupancy measure space $\Delta(M)$ is included in the estimated one $\Delta(\mathcal{P}_i)\ \forall i$.

Occupancy Measure Update. The update of the occupancy measure is performed on the space $\Delta\left(\mathcal{P}_i\right)$, which is built on the estimated transition function set $\mathcal{P}_i$. More formally:

$$\widehat{q}_{t+1} = \Pi_{\Delta(\mathcal{P}_i)}\left(\widehat{q}_t - \eta_t \ell_t\right)$$

with $\eta_t = \frac{1}{\overline{\ell}_t C\sqrt{T}}$ with $\overline{\ell}_t = \max\{||\ell_t||_\infty\}_{t=1}^t$, and C constant. The employment of Online Gradient Descent has been necessary to achieve the interval regret results, while the adaptive learning rate was chosen to improve the performance in terms of Regret bounds.

C.2 Interval Regret

In the following subsections, we prove the theorem related to the interval regret of Algorithm 4. First, we will present the main theorem, then, all the necessary lemmas.

Theorem 3. *With probability at least $1-2\delta$, when $\eta_t = \left(\overline{\ell}_t C\sqrt{T}\right)^{-1}$, UC-O-GDPS satisfies for any $q \in \cap_i \Delta\left(\mathcal{P}_i\right)$:*

$$R^{\mathsf{P}}_{t_1,t_2}(q) \leq \overline{\ell}_{t_1,t_2}\mathcal{E}^q_\delta + \overline{\ell}_{t_2} LC\sqrt{T} + \overline{\ell}_{t_1,t_2}\frac{|X||A|}{2}\frac{(t_2-t_1+1)}{C\sqrt{T}},$$

where $\overline{\ell}_{t_1,t_2} := \max\{||\ell_t||_\infty\}_{t=t_1}^{t_2}$, $\overline{\ell}_t := \overline{\ell}_{1,t}$ and $\delta \in [0,1]$.

Proof. Assume Event $E^{\Delta,\widehat{q}}(\delta)$ holds. By definition 2:

$$\begin{aligned}
R_{t_1,t_2}(q) &= \sum_{t=t_1}^{t_2} \ell_t^\top (q_t - q) \\
&= \underbrace{\sum_{t=t_1}^{t_2} \ell_t^\top (q_t - \widehat{q}_t)}_{①} + \underbrace{\sum_{t=t_1}^{t_2} \ell_t^\top (\widehat{q}_t - q)}_{②} \\
&\leq \overline{\ell}_{t_1,t_2}\mathcal{E}^q_\delta + \overline{\ell}_{t_2} LC\sqrt{T} + \overline{\ell}_{t_1,t_2}\frac{|X||A|}{2}\frac{(t_2-t_1+1)}{C\sqrt{T}}
\end{aligned}$$

where the Inequality holds by Lemmas 9 and 10. We focus on bounding the first term ① and the second term ②. □

C.2.1 Bound on the First Term

In order to bound the first term of the Interval Regret, we state some useful Lemmas by Rosenberg & Mansour (2019b).

Lemma 6. *Rosenberg & Mansour (2019b) Let $\{\pi_t\}_{t=1}^T$ be policies and let $\{P_t\}_{t=1}^T$ be transition functions. Then,*

$$\sum_{t=1}^T ||q^{P_t,\pi_t} - q^{P,\pi_t}||_1 \leq \sum_{t=1}^T \sum_{x\in X}\sum_{a\in A} |q^{P_t,\pi_t}(x,a) - q^{P,\pi_t}(x,a)| + \sum_{t=1}^T \sum_{x\in X}\sum_{a\in A} q^{P,\pi_t}(x,a)||P_t(\cdot|x,a) - P(\cdot|x,a)||_1 \tag{8}$$

where $P_t = P^{\widehat{q}_t}$.

The following Lemma, shows how to bound the first term in Equation (8) with the second one.

Lemma 7. *Rosenberg & Mansour (2019b) Let $\{\pi_t\}_{t=1}^T$ be policies and let $\{P_t\}_{t=1}^T$ be transition functions. Then, for every $k \in [1..L-1]$ and every $t = 1, ..., T$ it holds that:*

$$\sum_{x_k \in X_k} \sum_{a_k \in A} |q^{P_t,\pi_t}(x_k, a_k) - q^{P,\pi_t}(x_k, a_k)| \leq \sum_{s=0}^{k-1} \sum_{x_s \in X_s} \sum_{a_s \in A} q^{P,\pi_t}(x_s, a_s) ||P_t(\cdot|x_s, a_s) - P(\cdot|x_s, a_s)||_1$$

where $P_t = P^{\widehat{q}_t}$.

and finally, Equation (8) is upper bounded given:

Lemma 8. *Rosenberg & Mansour (2019b) Let $\{\pi_t\}_{t=1}^T$ be policies and let $\{P_t\}_{t=1}^T$ be transition functions such that $q^{P_t,\pi_t} \in \Delta(\mathcal{P}_i)$ for every t. Then, with probability at least $1 - 2\delta$ Event $E^{\Delta}(\delta)$ holds and:*

$$\sum_{t=1}^{T} \sum_{k=0}^{L-1} \sum_{s=0}^{k-1} \sum_{x_s \in X_s} \sum_{a_s \in A} q^{P,\pi_t}(x_s, a_s) ||P_t(\cdot|x_s, a_s) - P(\cdot|x_s, a_s)||_1 \leq 2L|X| \sqrt{2T \ln\left(\frac{1}{\delta}\right)} + 3L|X| \sqrt{2T|A| \ln\left(\frac{T|X||A|}{\delta}\right)}$$

where $P_t = P^{\widehat{q}_t}$.

From the previous Lemmas, it easy to show that:

Lemma 2. *If the confidence set $\mathcal{P}$ is updated as in Equation (6), with probability at least $1 - 2\delta$ $\sum_{t=1}^T ||q_t - \widehat{q}_t||_1 \leq \mathcal{E}_\delta^q$, where $\mathcal{E}_\delta^q \leq \tilde{\mathcal{O}}(\sqrt{T})$.*

Proof. Following Rosenberg & Mansour (2019b), by Lemmas 6, 7 and 8 we obtain that with probability at least $1 - 2\delta$ Event $E^{\Delta}(\delta)$ holds and: $\sum_{t=1}^T ||q^{P_t,\pi_t} - q^{P,\pi_t}||_1 \leq 4L|X| \sqrt{2T \ln\left(\frac{1}{\delta}\right)} + 6L|X| \sqrt{2T|A| \ln\left(\frac{T|X||A|}{\delta}\right)}$ □

Now, we are ready to bound ①.

Lemma 9. *Under Event $E^{\Delta,\widehat{q}}(\delta)$ it holds:*

$$\sum_{t=t_1}^{t_2} \ell_t^\top (q_t - \widehat{q}_t) \leq \overline{\ell}_{t_1,t_2} \mathcal{E}_\delta^q$$

with $\overline{\ell}_{t_1,t_2} := \max\{||\ell_t||_\infty\}_{t=t_1}^{t_2}$

Proof.

$$\begin{aligned}
\sum_{t=t_1}^{t_2} \ell_t^\top (q_t - \widehat{q}_t) &\leq \sum_{t=t_1}^{t_2} ||\ell_t||_\infty ||q_t - \widehat{q}_t||_1 \\
&\leq \overline{\ell}_{t_1,t_2} \sum_{t=t_1}^{t_2} ||q_t - \widehat{q}_t||_1 \\
&\leq \overline{\ell}_{t_1,t_2} \sum_{t=1}^{T} ||q_t - \widehat{q}_t||_1 \\
&\leq \overline{\ell}_{t_1,t_2} \mathcal{E}_\delta^q \qquad (9)
\end{aligned}$$

with $\overline{\ell}_{t_1,t_2} := \max\{||\ell_t||_\infty\}_{t=t_1}^{t_2}$ and where Inequality (9) holds under the event $E^{\widehat{q}}(\delta)$. □

C.2.2 BOUND ON THE SECOND TERM

Lemma 10. *For any $q \in \cap_i \Delta(\mathcal{P}_i)$, the Projected OGD update:*

$$\widehat{q}_{t+1} = \Pi_{\Delta(\mathcal{P}_i)}(\widehat{q}_t - \eta_t \ell_t)$$

with $\eta_t = \frac{1}{\overline{\ell}_t C\sqrt{T}}$ and $\overline{\ell}_t = \max\{||\ell_t||_\infty\}_{t=1}^t$ ensures:

$$\sum_{t=t_1}^{t_2} \ell_t^\top (\widehat{q}_t - q) \leq U_1 \frac{\overline{\ell}_{t_2}}{2} C\sqrt{T} + U_2 \frac{\overline{\ell}_{t_1,t_2}}{2} \frac{(t_2 - t_1 + 1)}{C\sqrt{T}}$$

where $U_1 = 2L$, $U_2 = |X||A|$, $\overline{\ell}_{t_1,t_2} = \max\{||\ell_t||_\infty\}_{t=t_1}^{t_2}$.

Proof. By the standard analysis of Projected Online Gradient Descent [Lemma 2.12 Orabona (2019)] we have:

$$\ell_t^\top (\widehat{q}_t - q) \leq \frac{1}{2\eta_t} ||\widehat{q}_t - q||_2^2 - \frac{1}{2\eta_t} ||\widehat{q}_{t+1} - q||_2^2 + \frac{\eta_t}{2} ||\ell_t||_2^2.$$

Observe that for any two occupancy measures q_1, q_2 it holds:

$$\begin{aligned} ||q_1 - q_2||_2^2 &\leq ||q_1||_2^2 + ||q_2||_2^2 \\ &\leq ||q_1||_1 + ||q_2||_1 \\ &\leq 2L \end{aligned}$$

where the second Inequality follows from $q(x,a) \in [0,1] \ \ \forall x, a$. Then, summing over the interval $[t_1 .. \ t_2]$ we get:

$$\begin{aligned} \sum_{t=t_1}^{t_2} \ell_t^\top (\widehat{q}_t - q) \leq & \frac{1}{2\eta_{t_1}} ||\widehat{q}_{t_1} - q||_2^2 \underbrace{- \frac{1}{2\eta_{t_2}} ||\widehat{q}_{t_2+1} - q||_2^2}_{\leq 0} \\ & + \frac{1}{2} \sum_{t=t_1}^{t_2-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) ||\widehat{q}_{t+1} - q||_2^2 + \frac{1}{2} \sum_{t=t_1}^{t_2} \eta_t ||\ell_t||_2^2 \\ \leq & \frac{L}{\eta_{t_1}} + L \sum_{t=t_1}^{t_2-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) + \frac{1}{2C\sqrt{T}} \sum_{t=t_1}^{t_2} \frac{1}{\overline{\ell}_t} \sum_{x,a} \ell_t(x,a)^2 \qquad (10) \\ \leq & \frac{L}{\eta_{t_1}} + L \underbrace{\sum_{t=t_1}^{t_2-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right)}_{= \frac{1}{\eta_{t_2}} - \frac{1}{\eta_{t_1}}} + \frac{1}{2C\sqrt{T}} \sum_{t=t_1}^{t_2} \underbrace{\frac{||\ell_t||_\infty}{\max\{||\ell_\tau||_\infty\}_{\tau=1}^t}}_{\leq 1} ||\ell_t||_\infty \sum_{x,a} 1 \\ \leq & L \overline{\ell}_{t_2} C\sqrt{T} + \frac{|X||A|}{2} \overline{\ell}_{t_1,t_2} \frac{(t_2 - t_1 + 1)}{C\sqrt{T}} \qquad (11) \end{aligned}$$

where Inequality (10) follows from the definition of η_t, and from $\eta_t > \eta_{t+1}$, while Inequality (11) comes from the telescopic sum over $[t_1 .. t_2]$ and from the definition of η_{t_2}.

□

D OMITTED PROOF OF SECTION 5

D.1 INTERVAL REGRETS

In this section, we show the Interval Regrets, attained by both primal and dual player, in our specific framework.

D.1.1 INTERVAL REGRET OF THE DUAL

In this subsection, we show the Interval Regret obtained by dual player. Recall that the dual variables are updated with Projected Online Gradient Descent as shown in (5) or equivalently:

$$\lambda_{t+1,i} = \min\left\{\max\left\{0, \lambda_{t,i} + \eta[G_t^\top]_i \widehat{q}_t\right\}, T^{1/4}\right\} \tag{12}$$

with $\eta = \left[K\sqrt{T\ln\left(\frac{T^2}{\delta}\right)}\right]^{-1}$.

Let

$$R^{\mathsf{D}}_{t_1,t_2}(\lambda) := \sum_{t=t_1}^{t_2} \left(\lambda - \lambda_t\right)^\top G_t^\top \widehat{q}_t$$

denote the regret accumulated by OGD from episode t_1 to episode t_2 with respect to the constant multiplier λ. By standard analysis of OGD Orabona (2019) we have that:

$$R^{\mathsf{D}}_{t_1,t_2}(\lambda) \leq \frac{||\lambda_{t_1} - \lambda||_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=t_1}^{t_2} ||G_t^\top \widehat{q}_t||_2^2$$

We can upper-bound the quantity $||G_t^\top \widehat{q}_t||_2^2$ as:

$$||G_t^\top \widehat{q}_t||_2^2 = \sum_{i=1}^{m}\left(\sum_{x,a} g_{t,i}(x,a)\widehat{q}_t(x,a)\right)^2 \leq \sum_{i=1}^{m}\left(\sum_{x,a}\widehat{q}_t(x,a)\right)^2 \leq mL^2$$

obtaining:

$$R^{\mathsf{D}}_{t_1,t_2}(\lambda) \leq D_1\frac{||\lambda_{t_1} - \lambda||_2^2}{\eta} + D_2\eta(t_2 - t_1 + 1)$$

with $D_1 = \frac{1}{2}$, $D_2 = \frac{mL^2}{2}$.

We bound the distance between lagrange multipliers for consecutive episodes.

Lemma 11. *If the dual player employs Projected Online Gradient Descent as in Update (12), it holds:*

$$||\lambda_{t+1}||_1 - ||\lambda_t||_1 \leq m\eta L$$

Proof. Since the dual minimizer is performing projected gradient descent with learning rate η, and the gradient of the Lagrangian at time t with respect to λ is equal to $\widehat{q}_t^\top G_t^\top$, element-wise it holds that:

$$\begin{aligned}
\lambda_{t+1,i} &= \min\left\{\max\left\{0, \lambda_{t,i} + \eta[G_t^\top]_i\widehat{q}_t\right\}, T^{\frac{1}{4}}\right\} \\
&\leq \max\left\{0, \lambda_{t,i} + \eta[G_t^\top]_i\widehat{q}_t\right\} \\
&\leq \max\left\{0, \lambda_{t,i} + \eta||[G_t^\top]_i||_\infty||\widehat{q}_t||_1\right\} \\
&\leq \max\left\{0,\ \lambda_{t,i} + \eta L\right\} \\
&= \lambda_{t,i} + \eta L
\end{aligned}$$

Thus,

$$||\lambda_{t+1}||_1 - ||\lambda_t||_1 = \sum_{i=1}^{m}\lambda_{t+1,i} - \sum_{i=1}^{m}\lambda_{t,i} \leq \sum_{i=1}^{m}\lambda_{t,i} + \sum_{i=1}^{m}\eta L - \sum_{i=1}^{m}\lambda_{t,i} = m\eta L$$

□

D.1.2 INTERVAL REGRET OF THE PRIMAL

We restate Lemma 10:

Lemma 10. *For any $q \in \cap_i \Delta(\mathcal{P}_i)$, the Projected OGD update:*

$$\widehat{q}_{t+1} = \Pi_{\Delta(\mathcal{P}_i)}\left(\widehat{q}_t - \eta_t \ell_t\right)$$

with $\eta_t = \frac{1}{\overline{\ell}_t C\sqrt{T}}$ and $\overline{\ell}_t = \max\{||\ell_t||_\infty\}_{t=1}^{t}$ ensures:

$$\sum_{t=t_1}^{t_2} \ell_t^\top (\widehat{q}_t - q) \leq U_1 \frac{\overline{\ell}_{t_2}}{2} C\sqrt{T} + U_2 \frac{\overline{\ell}_{t_1,t_2}}{2} \frac{(t_2 - t_1 + 1)}{C\sqrt{T}}$$

where $U_1 = 2L$, $U_2 = |X||A|$, $\overline{\ell}_{t_1,t_2} = \max\{||\ell_t||_\infty\}_{t=t_1}^{t_2}$.

Let

$$\lambda_{t_1,t_2} := \max\{||\lambda_t||_1\}_{t=t_1}^{t_2}.$$

Then it holds $\overline{\ell}_{t_1,t_2} \leq 1 + \lambda_{t_1,t_2}$ and we can restate the interval regret of the primal in terms of the 1-norm of the Lagrange multipliers as:

$$\sum_{t=t_1}^{t_2} r_t^{\mathcal{L}\top}(q - \widehat{q}_t) \leq U_1 \frac{(1 + \lambda_{1,t_2})}{2} C\sqrt{T} + U_2 \frac{(1 + \lambda_{t_1,t_2})}{2} \frac{(t_2 - t_1 + 1)}{C\sqrt{T}}. \tag{13}$$

D.2 Bound on the Lagrange multipliers

We prove Theorem 4, which we restate for convenience.

Theorem 4. *If Condition 2 holds and PDGD-OPS is used, then, when $\zeta := \frac{20mL^2}{\rho^2}$, it holds*

$$||\lambda_t||_1 \leq \zeta \qquad \forall t \in [T+1]$$

with probability at least $1 - 2\delta$ in the stochastic constraint setting and with probability at least $1 - \delta$ in the adversarial constraint setting.

Proof. Suppose event $E^{\Delta}(\delta)$ holds. If the constraints are stochastic, suppose event $E^{G}_{q^\circ}(\delta)$ holds too. Let $M > 1$ be a constant. We prove the statement by absurd. Suppose by absurd that there exists $t_2 \in [T]$ such that:

$$\forall t \leq t_2 \quad ||\lambda_t||_1 \leq \frac{2LM}{\rho^2} \qquad \wedge \qquad ||\lambda_{t_2+1}||_1 > \frac{2LM}{\rho^2}$$

and let $t_1 < t_2$ be such that:

$$||\lambda_{t_1-1}||_1 \leq \frac{2L}{\rho} \qquad \wedge \qquad \forall t : t_1 \leq t \leq t_2 \quad ||\lambda_t||_1 \geq \frac{2L}{\rho}.$$

By construction it holds that $1 < \frac{2L}{\rho} \leq ||\lambda_t||_1 \leq \frac{2LM}{\rho^2}$ for all $t_1 \leq t \leq t_2$. Also notice that by Lemma 11, for $\eta \leq \frac{1}{mL}$ it holds that:

$$||\lambda_{t_1}||_1 \leq ||\lambda_{t_1-1}||_1 + m\eta L \leq \frac{2L}{\rho} + m\eta L \leq \frac{4L}{\rho}$$

Focus on the quantity $\sum_{t=t_1}^{t_2} -\lambda_t^\top G_t^\top q^\circ$: in the stochastic constraint setting we have, under the event $E_{q^\circ}^G(\delta)$:

$$\begin{aligned}
\sum_{t=t_1}^{t_2} -\lambda_t^\top G_t^\top q^\circ &\geq \sum_{t=t_1}^{t_2} -\lambda_t^\top \overline{G}^\top q^\circ - \lambda_{t_1,t_2}\mathcal{E}_{t_1,t_2}^G \\
&\geq \sum_{t=t_1}^{t_2}\sum_{i=1}^{m} -\lambda_{t,i}\left[\overline{G}^\top q^\circ\right]_i - \lambda_{t_1,t_2}\mathcal{E}_{t_1,t_2}^G \\
&\geq \rho\sum_{t=t_1}^{t_2}\sum_{i=1}^{m}\lambda_{t,i} - \lambda_{t_1,t_2}\mathcal{E}_{t_1,t_2}^G \\
&= \rho\sum_{t=t_1}^{t_2}||\lambda_t||_1 - \lambda_{t_1,t_2}\mathcal{E}_{t_1,t_2}^G \\
&\geq \rho\frac{2L}{\rho}(t_2-t_1+1) - \lambda_{t_1,t_2}\mathcal{E}_{t_1,t_2}^G \\
&= 2L(t_2-t_1+1) - \lambda_{t_1,t_2}\mathcal{E}_{t_1,t_2}^G
\end{aligned}$$

While in the adversarial setting it holds:

$$\begin{aligned}
\sum_{t=t_1}^{t_2} -\lambda_t^\top G_t^\top q^\circ &\geq \sum_{t=t_1}^{t_2}\sum_{i=1}^{m} -\lambda_{t,i}\left[G_t^\top q^\circ\right]_i \\
&\geq \rho\sum_{t=t_1}^{t_2}\sum_{i=1}^{m}\lambda_{t,i} \\
&= \rho\sum_{t=t_1}^{t_2}||\lambda_t||_1 \\
&\geq \rho\frac{2L}{\rho}(t_2-t_1+1) \\
&= 2L(t_2-t_1+1)
\end{aligned}$$

In particular, we have that:

$$\sum_{t=t_1}^{t_2} -\lambda_t^\top G_t^\top q^\circ \geq 2L(t_2-t_1+1) - \lambda_{t_1,t_2}\mathcal{E}_{t_1,t_2}^G$$

is true in both settings under the required events.

We can lower bound the cumulative value of the Lagrangian function, namely ${r_t^\mathcal{L}}^\top \widehat{q}_t$, from t_1 to t_2 by that achievable by the primal minimizer by always playing the feasible occupancy measure q°:

$$\begin{aligned}
\sum_{t=t_1}^{t_2} {r_t^\mathcal{L}}^\top \widehat{q}_t &= \sum_{t=t_1}^{t_2} {r_t^\mathcal{L}}^\top q^\circ - \sum_{t=t_1}^{t_2} {r_t^\mathcal{L}}^\top (q^\circ - \widehat{q}_t) \\
&= \underbrace{\sum_{t=t_1}^{t_2} r_t^\top q^\circ}_{\geq 0} + \sum_{t=t_1}^{t_2} -\lambda_t^\top G_t^\top q^\circ - \sum_{t=t_1}^{t_2} {r_t^\mathcal{L}}^\top (q^\circ - \widehat{q}_t) \\
&\geq 2L(t_2-t_1+1) - \lambda_{t_1,t_2}\mathcal{E}_{t_1,t_2,\delta}^G - \sum_{t=t_1}^{t_2} {r_t^\mathcal{L}}^\top (q^\circ - \widehat{q}_t)
\end{aligned}$$

Applying Lemma 10 and observing that by construction $1 \leq \lambda_{t_1,t_2} \leq \frac{2LM}{\rho^2}$, we can bound $1+\lambda_{t_1,t_2} \leq \frac{4LM}{\rho^2}$ and obtain:

$$\sum_{t=t_1}^{t_2} r_t^{\mathcal{L}\top} \widehat{q}_t \geq 2L(t_2 - t_1 + 1) - \frac{2LM}{\rho^2}\mathcal{E}^G_{t_1,t_2,\delta} - U_1\frac{2LM}{\rho^2}C\sqrt{T} - U_2\frac{2LM}{\rho^2}\frac{(t_2 - t_1 + 1)}{C\sqrt{T}}$$

since under $E^{\Delta}(\delta)$ we have that $q^\circ \in \cap_i \Delta(\mathcal{P}_i)$.

We can upper-bound the same quantity with the value achievable by the dual by always playing a vector of zeroes.

$$\begin{aligned}
\sum_{t=t_1}^{t_2} r_t^{\mathcal{L}\top} \widehat{q}_t &= \sum_{t=t_1}^{t_2} r_t^\top \widehat{q}_t - \sum_{t=t_1}^{t_2} \lambda_t^\top G_t^\top \widehat{q}_t \\
&\leq \sum_{t=t_1}^{t_2} r_t^\top \widehat{q}_t - \sum_{t=t_1}^{t_2} \underline{0}^\top G_t^\top \widehat{q}_t + R^{\mathsf{D}}_{t_1,t_2}(\underline{0}) \\
&\leq \sum_{t=t_1}^{t_2} L + D_1\frac{||\lambda_{t_1}||_2^2}{\eta} + D_2\eta(t_2 - t_1 + 1) \\
&\leq \sum_{t=t_1}^{t_2} L + D_1\frac{||\lambda_{t_1}||_1^2}{\eta} + D_2\eta(t_2 - t_1 + 1) \\
&\leq L(t_2 - t_1 + 1) + D_3\frac{L^2}{\rho^2\eta} + D_2\eta(t_2 - t_1 + 1)
\end{aligned}$$

With $D_3 = 4D_1$.

Combining the bounds on the cumulative value of the Lagrangian, we have:

$$\begin{gathered}
2L(t_2 - t_1 + 1) - \frac{2LM}{\rho^2}\mathcal{E}^G_{t_1,t_2,\delta} - U_1\frac{2LM}{\rho^2}C\sqrt{T} - U_2\frac{2LM}{\rho^2}\frac{(t_2 - t_1 + 1)}{C\sqrt{T}} \\
\leq \\
L(t_2 - t_1 + 1) + D_3\frac{L^2}{\rho^2\eta} + D_2\eta(t_2 - t_1 + 1)
\end{gathered}$$

Observing that $\mathcal{E}^G_{t_1,t_2,\delta} = 2L\sqrt{2(t_2 - t_1 + 1)\ln\left(\frac{T^2}{\delta}\right)} \leq U_3 l_1\sqrt{t_2 - t_1 + 1}$ with $l_1 = \sqrt{\ln\left(\frac{T^2}{\delta}\right)}$ and $U_3 = 2L\sqrt{2}$ and rearranging the terms we obtain:

$$\begin{aligned}
L(t_2 - t_1 + 1) \leq\ & U_3\frac{2LM}{\rho^2}l_1\sqrt{t_2 - t_1 + 1} + \\
&+ U_1\frac{2LM}{\rho^2}C\sqrt{T} + \\
&+ U_2\frac{2LM}{\rho^2}\frac{(t_2 - t_1 + 1)}{C\sqrt{T}} + \\
&+ D_2\eta(t_2 - t_1 + 1) + \\
&+ D_3\frac{1}{\eta}\frac{L^2}{\rho^2}
\end{aligned}$$

We will make use of the following lemma:

Lemma 12. *For $\eta \leq \frac{1}{mL}$ and $\frac{M}{\rho} > 4$ it holds:*

$$(t_2 - t_1 + 1) > \frac{M}{\rho^2 m\eta}$$

Proof. By Lemma 11 we have:

$$\sum_{t=t_1}^{t_2} (||\lambda_{t+1}||_1 - ||\lambda_t||_1) \leq \sum_{t=t_1}^{t_2} m\eta L$$

which, since the sum in the LHS is telescopic, implies:

$$||\lambda_{t_2+1}||_1 - ||\lambda_{t_1}||_1 \leq (t_2 - t_1 + 1)m\eta L.$$

Also note that:

$$\frac{2LM}{\rho^2} - \frac{4L}{\rho} \leq ||\lambda_{t_2+1}||_1 - ||\lambda_{t_1}||_1.$$

Rearranging the terms, we obtain, for $\frac{M}{\rho} > 4$:

$$\frac{M}{\rho^2 m\eta} < \frac{2L(\frac{M}{\rho} - 2)}{\rho m\eta L} \leq (t_2 - t_1 + 1)$$

□

Applying Lemma 12 we show that the above leads to a contradiction for some choices of C, M and η, namely, we show that:

$$\begin{aligned} L(t_2 - t_1 + 1) > \; & U_3 \frac{2LM}{\rho^2} l_1 \sqrt{t_2 - t_1 + 1} + && (1) \\ & + U_1 \frac{2LM}{\rho^2} C\sqrt{T} + && (2) \\ & + U_2 \frac{2LM}{\rho^2} \frac{(t_2 - t_1 + 1)}{C\sqrt{T}} + && (3) \\ & + D_2 \eta (t_2 - t_1 + 1) + && (4) \\ & + D_3 \frac{1}{\eta} \frac{L^2}{\rho^2} && (5) \end{aligned}$$

In the followings, we prove that each of the terms on the RHS is upper bounded by $\frac{1}{5}L(t_2 - t_1 + 1)$:

1. By trivial computations and applying Lemma 12:

$$\begin{aligned} \frac{1}{5}L(t_2 - t_1 + 1) &> U_3 \frac{2LM}{\rho^2} l_1 \sqrt{T} \geq U_3 \frac{2LM}{\rho^2} l_1 \sqrt{t_2 - t_1 + 1} \\ (t_2 - t_1 + 1) &> U_3 \frac{10M}{\rho^2} l_1 \sqrt{T} \\ (t_2 - t_1 + 1) &> \frac{M}{\rho^2 m\eta} \geq U_3 \frac{10M}{\rho^2} l_1 \sqrt{T} \\ \frac{1}{m\eta} &\geq 10 U_3 l_1 \sqrt{T} \end{aligned}$$

   which is ensured by:

$$\boxed{\eta \leq \frac{1}{10 m U_3 l_1 \sqrt{T}}}$$

2. Then applying again Lemma 12:

$$\begin{aligned} \frac{1}{5}L(t_2 - t_1 + 1) &> U_1 \frac{2LM}{\rho^2} C\sqrt{T} \\ (t_2 - t_1 + 1) &> \frac{M}{\rho^2 m\eta} \geq 10 U_1 \frac{M}{\rho^2} C\sqrt{T} \end{aligned}$$

   which is true for:

$$\boxed{\eta \leq \frac{1}{10 m U_1 C\sqrt{T}}}$$

3. We solve the third term with respect to C.

$$\frac{1}{5}L(t_2 - t_1 + 1) \geq U_2 \frac{2LM}{\rho^2} \frac{(t_2 - t_1 + 1)}{C\sqrt{T}}$$

which is ensured by:

$$\boxed{C \geq 10U_2 \frac{M}{\rho^2} \frac{1}{\sqrt{T}}}$$

4.

$$\frac{1}{5}L(t_2 - t_1 + 1) > D_2\eta(t_2 - t_1 + 1)$$
$$\frac{1}{5}L > D_2\eta$$

Which is ensured by

$$\boxed{\eta < \frac{L}{5D_2}}$$

5. Applying Lemma 12, we solve the Inequality with respect to M:

$$\frac{1}{5}L(t_2 - t_1 + 1) > D_3 \frac{1}{\eta} \frac{L^2}{\rho^2}$$
$$(t_2 - t_1 + 1) > \frac{M}{\rho^2 m\eta} \geq 5D_3 \frac{1}{\eta} \frac{L}{\rho^2}$$
$$\frac{M}{m} \geq 5D_3 L$$

from which:

$$\boxed{M \geq 5mD_3L}$$

We recall all the constants: $D_2 = \frac{mL^2}{2}$, $D_3 = 2$, $U_1 = 2L$, $U_2 = |X||A|$, $U_3 = 2L\sqrt{2}$. We choose $M = 10mL$ and recall Condition 2:

$$\rho \geq T^{-\frac{1}{8}} L\sqrt{20m} \;\; \Rightarrow \;\; \frac{20mL^2}{\rho^2} \leq T^{\frac{1}{4}} \leq \sqrt{T}$$

We now focus on the condition on C:

$$C \geq 10U_2 \frac{10mL}{\rho^2} \frac{1}{\sqrt{T}}$$
$$= 5\frac{U_2}{L} \frac{20mL^2}{\rho^2} \frac{1}{\sqrt{T}}$$

is thus always ensured by $C = 5\frac{U_2}{L}$. The conditions on η are satisfied if:

$$\eta \leq \min\left\{ \frac{L}{5D_2},\ \frac{1}{10mU_1C\sqrt{T}},\ \frac{1}{10mU_3l_1\sqrt{T}} \right\}.$$

Observe that:

$$\min\left\{ \frac{L}{5D_2},\ \frac{1}{10mU_1C\sqrt{T}},\ \frac{1}{10mU_3l_1\sqrt{T}} \right\}$$
$$= \min\left\{ \frac{1}{2.5mL},\ \frac{1}{10mU_1\left(\frac{5U_2}{L}\right)\sqrt{T}},\ \frac{1}{20\sqrt{2}mLl_1\sqrt{T}} \right\}$$

which, if we plug in the value of l_1, leads to the choice:

$$\eta = \frac{1}{50m \max\left\{\frac{U_1U_2}{L},\ L\right\} \sqrt{T \ln\left(\frac{T^2}{\delta}\right)}}$$

The remaining conditions $\frac{M}{\rho} > 4$, $\eta \leq \frac{1}{mL}$ are trivially satisfied. Summing the conditions $(1-5)$ proves the contradiction.

If we plug the values of U_1 and U_2 corresponding to UC-O-GDPS, we have $\max\left\{\frac{U_1 U_2}{L},\ L\right\} = \max\left\{2|X||A|,\ L\right\} = 2|X||A|$ and thus obtain:

$$\eta = \frac{1}{100m|X||A|\sqrt{T\ln\left(\frac{T^2}{\delta}\right)}}$$

□

D.3 Analysis with Stochastic constraints

D.3.1 Lower bound on the dual cumulative utility

We start proving a useful Lemma in which we lower bound the dual cumulative utility. This Lemma holds both for the stochastic constraints and the adversarial constraint setting.

Lemma 13. *Under the event $E^{\widehat{q}}(\delta)$, the cumulative dual utility $\sum_{t=1}^{T} \lambda_t^\top G_t^\top q_t$ is lower bounded as:*

$$\sum_{t=1}^{T} \lambda_t^\top G_t^\top q_t \geq -\lambda_{1,T}\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0})$$

where $\lambda_{t_1,t_2} := \max\{\|\lambda_t\|_1\}_{t=t_1}^{t_2}$.

Proof. We exploit the fact that the dual is no-regret with respect to the $\underline{0}$ vector:

$$\begin{aligned}
\sum_{t=1}^{T} \lambda_t^\top G_t^\top q_t &= \sum_{t=1}^{T} \lambda_t^\top G_t^\top (q_t - \widehat{q}_t) + \sum_{t=1}^{T} \lambda_t^\top G_t^\top \widehat{q}_t \\
&\geq \sum_{t=1}^{T} \lambda_t^\top G_t^\top (q_t - \widehat{q}_t) + \sum_{t=1}^{T} \underline{0}^\top G_t^\top \widehat{q}_t - R_T^{\mathsf{D}}(\underline{0}) \\
&\geq \sum_{t=1}^{T} -\underbrace{\|\lambda_t\|_1}_{\leq \lambda_{1,T}} \underbrace{\left\|G_t^\top\right\|_\infty}_{\leq 1} \|q_t - \widehat{q}_t\|_1 - R_T^{\mathsf{D}}(\underline{0}) \\
&\geq -\lambda_{1,T} \sum_{t=1}^{T} \|q_t - \widehat{q}_t\|_1 - R_T^{\mathsf{D}}(\underline{0}) \\
&\geq -\lambda_{1,T}\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0})
\end{aligned}$$

where the last Inequality holds under $E^{\widehat{q}}(\delta)$. □

D.3.2 Analysis when Condition 2 holds

We start by introducing the notation $\widehat{v}_{t,i} := [G_t^\top]_i \widehat{q}_t$, that is the violation of the i-th constraint incurred by $\widehat{q}_t$. We further denote $\widehat{V}_{t,i} := \sum_{\tau=1}^{t} \widehat{v}_{\tau,i}$. Observe that, when Condition 2 holds, thanks to Theorem 4 we have $||\lambda_t||_1 \leq T^{\frac{1}{4}}$ for all t and thus $\lambda_{t,i} \leq T^{\frac{1}{4}}$. This means that $\lambda_{t,i}$ never gets past the upper extreme and the update of the dual is effectively equivalent to that of OGD working on the set $\mathrm{R}_{\geq 0}^m$:

$$\lambda_{t,i} = \max\{\lambda_{t,i} + \eta\widehat{v}_{t,i},\ 0\}$$

Lemma 14. *If Condition 2 holds, then for each episode $t \in [T]$ and each constraint i it holds:*

$$\lambda_{t,i} \geq \eta\widehat{V}_{t-1,i}$$

Proof. We prove the result by induction. Suppose that the statement holds for episode t. Then

$$\begin{aligned}\lambda_{t+1,i} &= \max\{\lambda_{t,i} + \eta\widehat{v}_{t,i},\ 0\} \\ &\geq \lambda_{t,i} + \eta\widehat{v}_{t,i} \\ &\geq \eta\widehat{V}_{t-1,i} + \eta\widehat{v}_{t,i} \\ &= \eta\widehat{V}_{t,i}\end{aligned}$$

Observe that for $t = 1$ the statement holds as the sum on the RHS evaluates to 0. □

Lemma 15. *If Condition 2 holds, under the events $E^{\Delta}(\delta)$, $E^{\widehat{q}}(\delta)$ and $E^{G}_{q^\circ}(\delta)$ for the stochastic constraint setting and under the events $E^{\Delta}(\delta)$ and $E^{\widehat{q}}(\delta)$ for the adversarial constraints one, it holds:*

$$V_T \leq \widehat{V}_{T,i^*} + \mathcal{E}^q_\delta$$

Proof. Let i^* denote the most violated constraint, e.g. $i^* = \arg\max_i \sum_{t=1}^T [G_t^\top q_t]_i$. Then we have:

$$\begin{aligned}V_T &= \sum_{t=1}^T [G_t^\top q_t]_{i^*} \\ &= \sum_{t=1}^T [G_t^\top \widehat{q}_t]_{i^*} + \sum_{t=1}^T [G_t^\top (q_t - \widehat{q}_t)]_{i^*} \\ &= \widehat{V}_{T,i^*} + \sum_{t=1}^T [G_t^\top]_{i^*}(q_t - \widehat{q}_t) \\ &\leq \widehat{V}_{T,i^*} + \sum_{t=1}^T |||[G_t^\top]_{i^*}||_\infty ||q_t - \widehat{q}_t||_1 \\ &\leq \widehat{V}_{T,i^*} + \mathcal{E}^q_\delta\end{aligned}$$

Where the last step holds under $E^{\widehat{q}}(\delta)$ since $|||[G_t^\top]_{i^*}||_\infty \leq 1$. □

We are now ready to prove the regret and violation bounds for the stochastic constraint setting.

Theorem 5. *In the stochastic constraint setting, when Condition 2 holds, the cumulative regret and constraint violation incurred by PDGD-OPS are upper bounded as follows. If the rewards are adversarial, then with probability at least $1 - 4\delta$ Algorithm 2 provides $R_T \leq \zeta\mathcal{E}^G_\delta + \zeta\mathcal{E}^q_\delta + R^{\mathsf{D}}_T(\underline{0}) + R^{\mathsf{P}}_T(q^*)$ and $V_T \leq \frac{1}{\eta}\zeta + \mathcal{E}^q_\delta$. If the rewards are stochastic, then with probability at least $1 - 5\delta$ Algorithm 2 provides $R_T \leq \mathcal{E}^r_\delta + \zeta\mathcal{E}^G_\delta + \zeta\mathcal{E}^q_\delta + R^{\mathsf{D}}_T(\underline{0}) + R^{\mathsf{P}}_T(q^*)$, and $V_T \leq \frac{1}{\eta}\zeta + \mathcal{E}^q_\delta$. In both cases:*

$$R_T \leq \tilde{\mathcal{O}}\left(\zeta\sqrt{T}\right),\ \ V_T \leq \tilde{\mathcal{O}}\left(\zeta\sqrt{T}\right).$$

Proof. Assume events $E^G_{q^\circ}(\delta)$, $E^G_{q^*}(\delta)$, $E^{\Delta}(\delta)$ and $E^{\widehat{q}}(\delta)$ hold.

Recall that $\lambda_{1,T} \leq \zeta$ under the events $E^{\Delta}(\delta)$ and $E^G_{q^\circ}(\delta)$ since Condition 2 holds (see proof of Theorem 4).
By Lemma 15 we have:

$$\begin{aligned}V_T &\leq \widehat{V}_{T,i^*} + \mathcal{E}^q_\delta \\ &\leq \frac{1}{\eta}\lambda_{T+1,i^*} + \mathcal{E}^q_\delta \\ &\leq \frac{1}{\eta}||\lambda_{T+1}||_1 + \mathcal{E}^q_\delta \\ &\leq \frac{1}{\eta}\zeta + \mathcal{E}^q_\delta\end{aligned}$$

Where the third Inequality holds for Lemma 14. By the definition of regret of the primal:

$$\begin{aligned}
\sum_{t=1}^{T} r_t^\top q_t &\geq \sum_{t=1}^{T} r_t^\top q^* - \sum_{t=1}^{T} \lambda_t^\top G_t^\top q^* + \sum_{t=1}^{T} \lambda_t^\top G_t^\top q_t - R_T^{\mathsf{P}}(q^*) \\
&\geq \sum_{t=1}^{T} r_t^\top q^* - \sum_{t=1}^{T} \lambda_t^\top G_t^\top q^* - \lambda_{1,T}\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*) \qquad (14) \\
&\geq \sum_{t=1}^{T} r_t^\top q^* - \sum_{t=1}^{T} \lambda_t^\top \overline{G}^\top q^* - \lambda_{1,T}\mathcal{E}_\delta^G - \lambda_{1,T}\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*) \qquad (15) \\
&\geq \sum_{t=1}^{T} r_t^\top q^* - \sum_{t=1}^{T} \sum_{i} \lambda_{t,i} \underbrace{(\overline{G})_i q^*}_{\leq 0} - \lambda_{1,T}\mathcal{E}_\delta^G - \lambda_{1,T}\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*) \qquad (16) \\
&\geq \sum_{t=1}^{T} r_t^\top q^* - \zeta\mathcal{E}_\delta^G - \zeta\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*)
\end{aligned}$$

where Inequality (14) holds for Lemma 13, and Inequality (15) holds under Event $E_{q^*}^G(\delta)$. We now focus on the case in which the rewards are adversarial. We have:

$$\sum_{t=1}^{T} r_t^\top q^* = T \cdot \overline{r}^\top q^* = T \cdot \text{OPT}_{\overline{r},\overline{G}}$$

and thus we obtain the stated bound:

$$\sum_{t=1}^{T} r_t^\top q_t \geq T \cdot \text{OPT}_{\overline{r},\overline{G}} - \zeta\mathcal{E}_\delta^G - \zeta\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*)$$

By union bound on $E_{q^\circ}^G(\delta)$, $E_{q^*}^G(\delta)$ and $E^{\Delta,\widehat{q}}(\delta)$, the result holds with probability at least $1 - 4\delta$.

For the stochastic rewards case, we require also event $E_{q^*}^r(\delta)$ to hold. Thus,

$$\sum_{t=1}^{T} r_t^\top q^* \geq \sum_{t=1}^{T} \overline{r}^\top q^* - \mathcal{E}_\delta^r = T \cdot \text{OPT}_{\overline{r},\overline{G}} - \mathcal{E}_\delta^r$$

and thus we obtain the stated bound:

$$\sum_{t=1}^{T} r_t^\top q_t \geq T \cdot \text{OPT}_{\overline{r},\overline{G}} - \mathcal{E}_\delta^r - \zeta\mathcal{E}_\delta^G - \zeta\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*)$$

By union bound on $E_{q^\circ}^G(\delta)$, $E_{q^*}^G(\delta)$, $E^{\Delta,\widehat{q}}(\delta)$ and $E_{q^*}^r(\delta)$, the result holds with probability at least $1 - 5\delta$.

Observe that under $E^{\Delta,\widehat{q}}(\delta)$ it holds:

$$R_T^{\mathsf{P}}(q^*) \leq \tilde{\mathcal{O}}\left((1 + \lambda_{1,T})\sqrt{T}\right) = \tilde{\mathcal{O}}\left(\zeta\sqrt{T}\right)$$

and

$$R_T^{\mathsf{D}}(\underline{0}) \leq \frac{mL^2}{2} \frac{1}{100m|X||A|\sqrt{\ln\left(\frac{T^2}{\delta}\right)}}\sqrt{T} \leq \mathcal{O}\left(\sqrt{T}\right)$$

□

D.3.3 ANALYSIS WHEN CONDITION 2 DOES NOT HOLD

Lemma 16. *If Condition 2 does not hold, then*

$$\widehat{V}_{T,i} \leq (2+2L)\frac{1}{\eta}T^{\frac{1}{4}} \qquad \forall T, i$$

holds under the event $E^{\Delta}(\delta)$ in the adversarial constraint setting and under the events $E^{\Delta}(\delta)$, $E^{G}_{q^\circ}(\delta)$, in the stochastic constraint setting.

Proof. Assume events $E^{\Delta}(\delta)$, $E^{G}_{q^\circ}(\delta)$ hold and suppose by absurd that $\widehat{V}_{T,i} = (2+2L+\epsilon)\frac{1}{\eta}T^{\frac{1}{4}}$, with $\epsilon > 0$, for some T and i.

We can lower bound the quantity $\sum_{t=1}^{T} {r_t^{\mathcal{L}}}^\top \widehat{q}_t$:

$$\begin{aligned}
\sum_{t=1}^{T} {r_t^{\mathcal{L}}}^\top \widehat{q}_t &= \underbrace{\sum_{t=1}^{T} r_t^\top q^\circ}_{\geq 0} - \sum_{t=1}^{T} \lambda_t^\top G_t^\top q^\circ - \sum_{t=1}^{T} {r_t^{\mathcal{L}}}^\top (q^\circ - \widehat{q}_t) \\
&\geq \underbrace{-\sum_{t=1}^{T} \lambda_t^\top \overline{G}^\top q^\circ}_{\geq 0} - \lambda_{1,T}\mathcal{E}^G_\delta - \sum_{t=1}^{T} {r_t^{\mathcal{L}}}^\top (q^\circ - \widehat{q}_t) \\
&\geq -mT^{\frac{1}{4}}\mathcal{E}^G_\delta - \sum_{t=1}^{T} {r_t^{\mathcal{L}}}^\top (q^\circ - \widehat{q}_t) \qquad (17)
\end{aligned}$$

Where Inequality (17) holds since $||\lambda_t||_1 \leq mV^{\frac{1}{4}}$ by construction of the dual space. Observe that, if we are in the Adversarial setting, then from the (stronger) definition of ρ and q° it holds $-\sum_{t=1}^{T} \lambda_t^\top G_t^\top q^\circ \geq 0$ and we obtain the tighter bound

$$\sum_{t=1}^{T} {r_t^{\mathcal{L}}}^\top \widehat{q}_t \geq -\sum_{t=1}^{T} {r_t^{\mathcal{L}}}^\top (q^\circ - \widehat{q}_t)$$

The dual is no regret with respect to the vector $\tilde{\lambda}$, whose elements are 0 for $j \neq i$ and $T^{\frac{1}{4}}$ in position $j = i$:

$$\begin{aligned}
\sum_{t=1}^{T} {r_t^{\mathcal{L}}}^\top \widehat{q}_t &= \sum_{t=1}^{T} r_t^\top \widehat{q}_t - \sum_{t=1}^{T} \lambda_t^\top G_t^\top \widehat{q}_t \\
&\leq \sum_{t=1}^{T} r_t^\top \widehat{q}_t - \sum_{t=1}^{T} \tilde{\lambda}^\top G_t^\top \widehat{q}_t + R_T^{\mathsf{D}}(\tilde{\lambda}) \\
&= \sum_{t=1}^{T} r_t^\top \widehat{q}_t - T^{\frac{1}{4}} \sum_{t=1}^{T} [G_t^\top \widehat{q}_t]_i + R_T^{\mathsf{D}}(\tilde{\lambda}) \\
&\leq LT - T^{\frac{1}{4}}\widehat{V}_{T,i} + R_T^{\mathsf{D}}(\tilde{\lambda})
\end{aligned}$$

Combining the bounds we have:

$$\begin{aligned}
-mT^{\frac{1}{4}}\mathcal{E}^G_\delta - \sum_{t=1}^{T} {r_t^{\mathcal{L}}}^\top (q^\circ - \widehat{q}_t) &\leq LT - T^{\frac{1}{4}}\widehat{V}_{T,i} + R_T^{\mathsf{D}}(\tilde{\lambda}) \\
T^{\frac{1}{4}}\widehat{V}_{T,i} &\leq LT + mT^{\frac{1}{4}}\mathcal{E}^G_\delta + \sum_{t=1}^{T} {r_t^{\mathcal{L}}}^\top (q^\circ - \widehat{q}_t) + R_T^{\mathsf{D}}(\tilde{\lambda}) \\
\frac{\sqrt{T}}{\eta}(2+2L+\epsilon) &\leq LT + mT^{\frac{1}{4}}\mathcal{E}^G_\delta + \sum_{t=1}^{T} {r_t^{\mathcal{L}}}^\top (q^\circ - \widehat{q}_t) + R_T^{\mathsf{D}}(\tilde{\lambda}) \qquad (18)
\end{aligned}$$

Observe that:

$$R_T^{\mathsf{D}}(\tilde{\lambda}) \leq \frac{1}{2}\frac{||\tilde{\lambda}||_2^2}{\eta} + \frac{mL^2}{2}\eta T = \frac{\sqrt{T}}{2\eta} + \frac{mL^2}{2}\frac{1}{100m|X||A|\sqrt{T\ln\left(\frac{T^2}{\delta}\right)}}T \leq L\frac{\sqrt{T}}{\eta}$$

Since $|X| \geq L$.

For the primal it holds by Lemma 10:

$$\begin{aligned}
\sum_{t=1}^{T} {r_t^{\mathcal{L}}}^{\top}(q^{\circ} - \widehat{q}_t) &= \sum_{t=1}^{T} {\ell_t}^{\top}(\widehat{q}_t - q^{\circ}) \\
&\leq \lambda_{1,T}U_1 C\sqrt{T} + \lambda_{1,T}U_2\frac{\sqrt{T}}{C} \\
&\leq mT^{\frac{1}{4}}\sqrt{T}\left(U_1 C + \frac{U_2}{C}\right) \\
&= m\left(U_1\frac{U_2}{5} + 5\right)\sqrt{T}\,T^{\frac{1}{4}} \\
&= m\left(2L\frac{|X||A|}{5} + 5\right)\sqrt{T}\,T^{\frac{1}{4}} \\
&\leq 6mL|X||A|\sqrt{T}\,T^{\frac{1}{4}} \\
&\leq \frac{L}{\eta}T^{\frac{1}{4}} \leq L\frac{\sqrt{T}}{\eta}
\end{aligned}$$

And for the Azuma-Hoeffding term it holds:

$$mT^{\frac{1}{4}}\mathcal{E}_{\delta}^{G} = mT^{\frac{1}{4}}2L\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)} \leq \frac{1}{\eta}T^{\frac{1}{4}} = \frac{\sqrt{T}}{\eta}$$

Observe that $LT \leq \frac{\sqrt{T}}{\eta}$ holds trivially.
Dividing both the terms in Equation (18) by $\frac{\sqrt{T}}{\eta}$, we obtain

$$2 + 2L + \epsilon \leq 2 + 2L$$

which is absurd. □

We are now ready to prove the Regret and Violation bounds when Assumption 2 does not hold:

Theorem 6. *In the stochastic constraint setting, when Condition 2 does not hold, the cumulative regret and constraint violations incurred by PDGD-OPS are upper bounded as follows. If the rewards are adversarial, then with probability at least* $1 - 4\delta$ *Algorithm 2 provides* $R_T \leq mT^{\frac{1}{4}}\mathcal{E}_{\delta}^{G} + mT^{\frac{1}{4}}\mathcal{E}_{\delta}^{q} + R_T^{\mathsf{D}}(\underline{0}) + R_T^{\mathsf{P}}(q^*)$ *and* $V_T \leq (2+2L)\frac{1}{\eta}T^{\frac{1}{4}} + \mathcal{E}_{\delta}^{q}$. *If the rewards are stochastic, then with probability at least* $1 - 5\delta$ *Algorithm 2 provides* $R_T \leq \mathcal{E}_{\delta}^{r} + mT^{\frac{1}{4}}\mathcal{E}_{\delta}^{G} + mT^{\frac{1}{4}}\mathcal{E}_{\delta}^{q} + R_T^{\mathsf{D}}(\underline{0}) + R_T^{\mathsf{P}}(q^*)$ *and* $V_T \leq (2+2L)\frac{1}{\eta}T^{\frac{1}{4}} + \mathcal{E}_{\delta}^{q}$. *In both cases, it holds:*

$$R_T \leq \tilde{\mathcal{O}}\left(T^{\frac{3}{4}}\right), \;\; V_T \leq \tilde{\mathcal{O}}\left(T^{\frac{3}{4}}\right).$$

Proof. Assume events $E^{\Delta}(\delta)$, $E^{\widehat{q}}(\delta)$, $E_{q^*}^{G}(\delta)$, $E_{q^{\circ}}^{G}(\delta)$ hold. We avoid the computations and restart from (16), since the previous part of the proofs are identical:

$$\begin{aligned}
\sum_{t=1}^{T} r_t^{\top} q_t &\geq \sum_{t=1}^{T} r_t^{\top} q^* - \sum_{t=1}^{T}\sum_{i}\lambda_{t,i}\underbrace{(\overline{G})_i q^*}_{\leq 0} - \lambda_{1,T}\mathcal{E}_{\delta}^{G} - \lambda_{1,T}\mathcal{E}_{\delta}^{q} - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*) \\
&\geq \sum_{t=1}^{T} r_t^{\top} q^* - mT^{\frac{1}{4}}\mathcal{E}_{\delta}^{G} - mT^{\frac{1}{4}}\mathcal{E}_{\delta}^{q} - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*)
\end{aligned}$$

By the same reasoning as in the proof of Theorem 5, we obtain that if the rewards are adversarial then

$$\sum_{t=1}^{T} r_t^\top q_t \geq T \cdot \text{OPT}_{\overline{r},\overline{G}} - mT^{\frac{1}{4}}\mathcal{E}_\delta^G - mT^{\frac{1}{4}}\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*)$$

with probability at least $1-4\delta$ by union bound on $E^{\Delta,\widehat{q}}(\delta)$, $E_{q^*}^G(\delta)$ and $E_{q^\circ}^G(\delta)$, while if the rewards are stochastic, under the event $E_{q^*}^r(\delta)$ we have that:

$$\sum_{t=1}^{T} r_t^\top q_t \geq T \cdot \text{OPT}_{\overline{r},\overline{G}} - \mathcal{E}_\delta^r - mT^{\frac{1}{4}}\mathcal{E}_\delta^G - mT^{\frac{1}{4}}\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*)$$

with probability at least $1-5\delta$ by union bound on $E^{\Delta,\widehat{q}}(\delta)$, $E_{q^*}^G(\delta)$, $E_{q^\circ}^G(\delta)$ and $E_{q^*}^r(\delta)$.

Observe that:

$$R_T^{\mathsf{P}}(q^*) \leq \tilde{\mathcal{O}}\left(T^{\frac{3}{4}}\right)$$

and

$$R_T^{\mathsf{D}}(\underline{0}) = \frac{mL^2}{2}\eta T \leq \tilde{\mathcal{O}}\left(\sqrt{T}\right).$$

In order to bound the violation, we apply Lemma 16:

$$V_T \leq \widehat{V}_{T,i^*} + \mathcal{E}_\delta^q \leq (2+2L)\frac{1}{\eta}T^{\frac{1}{4}} + \mathcal{E}_\delta^q$$

□

D.4 Analysis with Adversarial Constraints

D.4.1 Analysis when Condition 2 holds

Theorem 7. *In the adversarial constraint setting, when Condition 2 holds, the cumulative regret and constraint violations incurred by PDGD-OPS are upper bounded as follows. If the rewards are adversarial, then with probability at least $1-2\delta$ Algorithm 2 provides $R_T \leq \frac{1}{1+\rho}T \cdot \text{OPT}_{\overline{r},\overline{G}} + \zeta\mathcal{E}_\delta^q + R_T^{\mathsf{D}}(\underline{0}) + R_T^{\mathsf{P}}(\tilde{q})$ and $V_T \leq \frac{1}{\eta}\zeta + \mathcal{E}_\delta^q$. If the rewards are stochastic, then with probability at least $1-3\delta$ Algorithm 2 provides $R_T \leq \frac{1}{1+\rho}T \cdot \text{OPT}_{\overline{r},\overline{G}} + \mathcal{E}_\delta^r + \zeta\mathcal{E}_\delta^q + R_T^{\mathsf{D}}(\underline{0}) + R_T^{\mathsf{P}}(\tilde{q})$ and $V_T \leq \frac{1}{\eta}\zeta + \mathcal{E}_\delta^q$. In both cases, it holds:*

$$\sum_{t=1}^{T} r_t^\top q_t \geq \Omega\left(\frac{\rho}{1+\rho}T \cdot \text{OPT}_{\overline{r},\overline{G}}\right), \;\; V_T \leq \tilde{\mathcal{O}}\left(\zeta\sqrt{T}\right).$$

Proof. Assume events $E^\Delta(\delta)$ and $E^{\widehat{q}}(\delta)$ hold.

Recall that $\lambda_{1,T} \leq \zeta$ under the event $E^\Delta(\delta)$ since Condition 2 holds (see the proof of Theorem 4). Following the same steps of the proof of Theorem 5, we obtain:

$$V_T \leq \frac{1}{\eta}\zeta + \mathcal{E}_\delta^q$$

Let $\tilde{q} = \frac{\rho}{1+\rho}q^* + \frac{1}{1+\rho}q^\circ$, observe that it holds for all t and for all i:

$$[G_t^\top \tilde{q}]_i = \frac{\rho}{1+\rho}\underbrace{[G_t^\top q^*]_i}_{\leq 1} + \frac{1}{1+\rho}\underbrace{[G_t^\top q^\circ]_i}_{\leq -\rho} \leq 0$$

$$r_t^\top \tilde{q} = \frac{\rho}{1+\rho}r_t^\top q^* + \frac{1}{1+\rho}r_t^\top q^\circ \geq \frac{\rho}{1+\rho}r_t^\top q^*$$

By the definition of regret of the primal:

$$\begin{aligned}\sum_{t=1}^{T} r_t^\top q_t &\geq \sum_{t=1}^{T} r_t^\top \tilde{q} - \sum_{t=1}^{T} \lambda_t^\top G_t^\top \tilde{q} + \sum_{t=1}^{T} \lambda_t^\top G_t^\top q_t - R_T^{\mathsf{P}}(\tilde{q}) \\ &\geq \frac{\rho}{1+\rho} \sum_{t=1}^{T} r_t^\top q^* - \sum_{t=1}^{T} \sum_i \lambda_{t,i} \underbrace{[G_t^\top \tilde{q}]_i}_{\leq 0} + \sum_{t=1}^{T} \lambda_t^\top G_t^\top q_t - R_T^{\mathsf{P}}(\tilde{q}) \\ &\geq \frac{\rho}{1+\rho} \sum_{t=1}^{T} r_t^\top q^* - \lambda_{1,T} \mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(\tilde{q}) \\ &\geq \frac{\rho}{1+\rho} \sum_{t=1}^{T} r_t^\top q^* - \zeta \mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(\tilde{q})\end{aligned}$$

where the third Inequality holds for Lemma 13.
By the same reasoning as in the proof of Theorem 5, we obtain that if the rewards are adversarial it holds:

$$\begin{aligned}\sum_{t=1}^{T} r_t^\top q_t &\geq \frac{\rho}{1+\rho} T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - \zeta \mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(\tilde{q}) \\ &= T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - \frac{1}{1+\rho} T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - \zeta \mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(\tilde{q})\end{aligned}$$

with probability at least $1 - 2\delta$, since we are conditioning on $E^{\Delta,\widehat{q}}(\delta)$.
If the rewards are stochastic, requiring also event $E_{q^*}^r(\delta)$ to hold we obtain:

$$\frac{\rho}{1+\rho} \sum_{t=1}^{T} r_t^\top q^* \geq \frac{\rho}{1+\rho} \sum_{t=1}^{T} \overline{r}^\top q^* - \frac{\rho}{1+\rho} \mathcal{E}_\delta^r \geq \frac{\rho}{1+\rho} T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - \mathcal{E}_\delta^r$$

And thus,

$$\sum_{t=1}^{T} r_t^\top q_t \geq T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - \frac{1}{1+\rho} T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - \mathcal{E}_\delta^r - \zeta \mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(\tilde{q})$$

with probability at least $1 - 3\delta$. Finally observe that, under Assumption 2 and event $E^{\Delta,\widehat{q}}(\delta)$, it holds:

$$R_T^{\mathsf{P}}(\tilde{q}) \leq \tilde{\mathcal{O}}\left((1 + \lambda_{1,T})\sqrt{T}\right) \leq \tilde{\mathcal{O}}\left(\zeta\sqrt{T}\right)$$

and

$$R_T^{\mathsf{D}}(\underline{0}) \leq \frac{mL^2}{2} \frac{1}{100m|X||A|\sqrt{\ln\left(\frac{T^2}{\delta}\right)}} \sqrt{T} \leq \mathcal{O}\left(\sqrt{T}\right)$$

□

D.5 Azuma-Hoeffding Bounds and Proofs

In this subsection we prove that events $E_{q^*}^r(\delta)$, $E_{q^\circ}^G(\delta)$, $E_{q^*}^G(\delta)$ each hold with probability at least $1 - \delta$.

Lemma 3. *If the rewards are stochastic, then, with probability at least* $1 - \delta$*, it holds:*

$$\left| \sum_{t=1}^{T} \left(r_t - \overline{r}\right)^\top q^* \right| \leq \mathcal{E}_\delta^r,$$

where $\mathcal{E}_\delta^r := \frac{L}{\sqrt{2}} \sqrt{T \ln\left(\frac{2}{\delta}\right)}$.

Proof. Observe that:

$$\begin{aligned}\max_{t\in[t_1..t_2]}\left|(r_t-\overline{r})^\top q^*\right| &\leq \max_{t\in[t_1..t_2]}\underbrace{\|r_t-\overline{r}\|_\infty}_{\leq 1}\|q^*\|_1\\ &\leq L\end{aligned}$$

where the second Inequality holds since since $q^*(x,a)\geq 0$. By the Azuma-Hoeffding inequality for martingales we have that:

$$\mathbb{P}\left[\left|\sum_{t=t_1}^{t_2}(r_t-\overline{r})^\top q^*\right|\geq \frac{L}{\sqrt{2}}\sqrt{T\ln\left(\frac{2}{\delta}\right)}\right]\leq\delta.$$

□

We perform the same analysis for the constraints, obtaining:

Lemma 4. *If the constraints are stochastic, given a sequence of occupancy measures $(q_t)_{t=1}^T$, then with probability at least $1-\delta$, for all $[t_1..t_2]\subseteq[1..T]$, it holds:*

$$\left|\sum_{t=t_1}^{t_2}\lambda_t^\top\left(G_t^\top-\overline{G}^\top\right)q_t\right|\leq\lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2,\delta},$$

where $\mathcal{E}^G_{t_1,t_2,\delta}:=2L\sqrt{2(t_2-t_1+1)\ln\left(\frac{T^2}{\delta}\right)}$ and $\lambda_{t_1,t_2}:=\max\{\|\lambda_t\|_1\}_{t=t_1}^{t_2}$.

Proof. Observe that:

$$\begin{aligned}\max_{t\in[t_1..t_2]}\left|\lambda_t^\top(G_t^\top-\overline{G}^\top)q_t\right| &\leq \max_{t\in[t_1..t_2]}\|\lambda_t\|_1\underbrace{\left\|G_t^\top-\overline{G}^\top\right\|_\infty}_{\leq 2}\|q_t\|_1\\ &\leq \max_{t\in[t_1..t_2]}2||\lambda_t||_1L\\ &=2\lambda_{t_1,t_2}L\end{aligned}$$

where the second Inequality holds since $q_t(x,a)\geq 0$ and $\lambda_{t,i}\geq 0$. By the Azuma-Hoeffding inequality for martingales we have that:

$$\mathbb{P}\left[\left|\sum_{t=t_1}^{t_2}\lambda_t^\top(G_t^\top-\overline{G}^\top)q_t\right|\geq 2\lambda_{t_1,t_2}L\sqrt{2(t_2-t_1+1)\ln\left(\frac{2}{\delta}\frac{T^2}{2}\right)}\right]\leq 2\delta/T^2.$$

A union bound over all the t_1,t_2 such that $[t_1..t_2]\subseteq[1..T]$ concludes the proof. □