Towards Painless Policy Optimization for Constrained MDPs: Supplementary material

Arushi Jain[1*] **Sharan Vaswani**[2*] **Reza Babanezhad**[3] **Csaba Szepesvári**[4,5] **Doina Precup**[1,5]

[1]Mila, McGill University
[2]Simon Fraser University
[3]SAIT AI Lab, Montreal
[4]Amii, University of Alberta
[5]DeepMind

ORGANIZATION OF THE APPENDIX

A THEORETICAL GUARANTEES IN THE TABULAR SETTING

In the tabular setting, we use m independent trajectories for *each* (s, a) pair. By Hoeffding's inequality and union bound across all states and actions, the sampling error can be bounded by $\frac{1}{1-\gamma}\sqrt{\frac{\log(2SA/\delta)}{2m}}$ (similar to the proof of Lemma B.5). Since all action-value functions can be represented in the tabular setting, the bias error term $\varepsilon_{\text{b}} = 0$, and hence $\tilde{\varepsilon} = \frac{1}{1-\gamma}\sqrt{\frac{\log(2SA/\delta)}{2m}}$. Compared to the linear function approximation setting in Section 5 that has a computational complexity proportional to $O(d^2)$, the computational cost in the tabular setting is $O(SA)$. However, the approximation error is smaller than that in Lemma B.5.

Now that we have bounded the approximation errors in the tabular setting, we instantiate Theorem 3.1 for GDA in Section 4.2.1. Plugging in the value of U, the primal and dual regret from Equation (5) and $\tilde{\varepsilon}$, we obtain the following corollary.

Corollary A.1. *For the gradient descent ascent updates in Equations (6) and (7) with the specified step-sizes,* $U = \frac{2}{\zeta\,(1-\gamma)}$*, using* m *trajectories, the average optimality gap (*`OG`*) and constraint violation (*`CV`*) can be bounded as:*

$$\texttt{OG} \leq \frac{\left(\frac{(1+U)\sqrt{2\log|A|}}{1-\gamma} + U\right)}{(1-\gamma)\sqrt{T}} + \frac{\varepsilon_s(1+2U)}{1-\gamma},$$

$$\texttt{CV} \leq \frac{\zeta\left(\frac{(1+U)\sqrt{2\log|A|}}{1-\gamma} + U\right)}{\sqrt{T}} + \varepsilon_s(1+2U),$$

where $\varepsilon_s = \frac{1}{1-\gamma}\sqrt{\frac{\log(2SA/\delta)}{2m}}$.

*The first two authors contributed equally. Email: arushi.jain@mail.mcgill.ca, vaswani.sharan@gmail.com.

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

Proof. To get the result we replace the regrets for primal and dual of GDA (Orabona, 2019, Theorem 6.8) in Theorem 3.1 and get the required results. Specifically we set

$$\mathcal{R}^p(\pi^*, T) \leq \frac{1+U}{1-\gamma}\sqrt{2\log|A|}\sqrt{T},$$

and

$$\mathcal{R}^d(0,T), \mathcal{R}^d(U,T) \leq \frac{U}{1-\gamma}\sqrt{T}.$$

□

Hence, the average optimality gap for GDA is $O\left(\frac{1}{(1-\gamma)^3\sqrt{T}} + \frac{\varepsilon_s}{(1-\gamma)^2}\right)$, while the average constraint violation scales as $O\left(\frac{1}{(1-\gamma)^2\sqrt{T}} + \frac{\varepsilon_s}{1-\gamma}\right)$. Compared to the tabular result in Ding et al. (2020), the above bound on the optimality gap is worse by a factor of $O(1/1-\gamma)$ and matches their bound on the constraint violation. On the other hand, in the tabular setting without sampling error (when $\varepsilon_s = 0$), Xu et al. (2021, Theorem 3) obtain an $O\left(\frac{1}{(1-\gamma)^{1.5}\sqrt{T}}\right)$ bound on both the optimality gap and constraint violation. However, in order to set this bound, they require the knowledge of $\mathrm{KL}(\pi^*||\pi_0)$ to set the algorithm hyper-parameters. This information is not available, making it difficult to implement their algorithm.

Now, we instantiate Theorem 3.1 for the coin-betting algorithms in Section 4.2.2. Plugging in the value of U, the primal and dual regret and $\tilde{\varepsilon}$, we obtain the following corollary.

Corollary A.2. *Using the primal updates in Equation (8), and the dual updates in Equation (9), with $U = \frac{2}{\zeta(1-\gamma)}$, using m trajectories, the average optimality gap (OG) and constraint violation (CV) for CBP can be bounded as:*

$$\texttt{OG} \leq \frac{\left(\frac{3(1+U)\sqrt{1+KL(\pi_0||\pi^*)}}{1-\gamma} + \Psi\right)}{(1-\gamma)\sqrt{T}} + \frac{\varepsilon_s(1+2U)}{1-\gamma},$$

$$\texttt{CV} \leq \frac{\zeta\left(\frac{3(1+U)\sqrt{1+KL(\pi_0||\pi^*)}}{1-\gamma} + \Psi\right)}{\sqrt{T}} + \zeta\,\varepsilon_s(1+2U),$$

where $\varepsilon_s = \frac{1}{1-\gamma}\sqrt{\frac{\log(2SA/\delta)}{2m}}$ and $\Psi = 4U\sqrt{\log((T+1)U)} + 1$.

Proof. To get the result we replace the regrets for primal and dual of CB in Theorem 3.1 and get the required results. Specifically from (Orabona and Pal, 2016, Corollary 6) and (Orabona and Tommasi, 2017, Theorem 8), we get the upper-bound for primal regret and the dual regret:

$$\mathcal{R}^p(\pi^*, T) \leq \frac{3(1+U)}{1-\gamma}\sqrt{T}\sqrt{1+\mathrm{KL}(\pi_0||\pi^*)},$$

and

$$\mathcal{R}^d(\lambda, T) \leq \frac{1}{1-\gamma} + \left\|\lambda - \lambda^0\right\|\sqrt{\left(\frac{1}{(1-\gamma)^2} + \frac{G_T}{1-\gamma}\right)\Gamma_T}$$

where $\Gamma_T = \log\left(1 + (G_T(1-\gamma)+1)^2\left\|\lambda-\lambda^0\right\|^2\right)$ and $G_T = \sum_{i=0}^{T}|\hat{V}_c^{\pi_i}(\rho) - b|$. Since $|\hat{V}_c^{\pi_i}(\rho) - b| \leq \frac{1}{1-\gamma}$ we have $G_T \leq T/1-\gamma$ and $\left\|\lambda-\lambda^0\right\| \leq 2U$ for all λ. Using these upperbound and replace in $\mathcal{R}^d(\lambda, T)$ we get:

$$\mathcal{R}^d(\lambda, T) \leq \frac{4U\sqrt{(T+1)\log((T+1)U)}+1}{1-\gamma}$$

□

B MAIN PROOFS

The following well known result will be useful:

Lemma B.1 (Value difference lemma)**.** *For any value function V^π (reward or cost), and any two memoryless policies π and π',*

$$V^{\pi'} - V^\pi = (I - \gamma P_{\pi'})^{-1} [T_{\pi'} V^\pi - V^\pi]$$

where $T_{\pi'} V^\pi = [r_{\pi'} + \gamma P_{\pi'} V^\pi]$ is the Bellman operator for policy π'.

Proof. As is well known, $V^{\pi'} = (I - \gamma P_{\pi'})^{-1} r_{\pi'}$. Hence,

$$\begin{aligned} V^{\pi'} - V^\pi &= (I - \gamma P_{\pi'})^{-1} r_{\pi'} - V^\pi \\ &= (I - \gamma P_{\pi'})^{-1} (r_{\pi'} - (I - \gamma P_{\pi'}) V^\pi) \\ &= (I - \gamma P_{\pi'})^{-1} (r_{\pi'} + \gamma P_{\pi'} V^\pi - V^\pi) \\ &= (I - \gamma P_{\pi'})^{-1} [T_{\pi'} V^\pi - V^\pi] \end{aligned}$$

□

Let us now turn to the proof of Theorem 3.1:

Theorem B.2. *Assuming that $\left\| Q_r^t - \hat{Q}_r^t \right\|_\infty \le \tilde{\varepsilon}$ and $\left\| Q_c^t - \hat{Q}_c^t \right\|_\infty \le \tilde{\varepsilon}$, for a generic algorithm producing a sequence of polices $\{\pi_0, \pi_1, \ldots, \pi_{T-1}\}$ and dual variables $\{\lambda_0, \lambda_1, \ldots, \lambda_{T-1}\}$ such that for all t, λ_t is constrained to lie in the $[0, U]$ where $U > \lambda^*$, `OG` and `CV` can be bounded as:*

$$\begin{aligned} \texttt{OG} &\le \frac{\mathcal{R}^p(\pi^*, T) + (1-\gamma)\mathcal{R}^d(0, T)}{(1-\gamma)T} + \tilde{\varepsilon}\, g(U), \\ \texttt{CV} &\le \frac{\mathcal{R}^p(\pi^*, T) + (1-\gamma)\mathcal{R}^d(U, T)}{(U-\lambda^*)(1-\gamma)T} + \frac{\tilde{\varepsilon}\, g(U)}{(U-\lambda^*)}, \end{aligned}$$

where $g(U) := \left[\frac{1+U}{1-\gamma} + U\right]$.

Proof. We will begin with bounding the value differences in the Lagrangian using Lemma B.1. Let $T^r_{\pi^*}$ and $T^c_{\pi^*}$ be the Bellman operators of the optimal policy for the reward and cost respectively. Then,

$$[V_r^{\pi^*} - V_r^{\pi_t}] + \lambda_t [V_c^{\pi^*} - V_c^{\pi_t}] = (I - \gamma P_{\pi^*})^{-1} \left[[T^r_{\pi^*} V_r^{\pi_t} - V_r^{\pi_t}] + \lambda_t [T^c_{\pi^*} V_c^{\pi_t} - V_c^{\pi_t}] \right]$$

Let M_π be the state-action operator applied Q functions such that $M_\pi(Q)(s) = \sum_a \pi(a|s) Q(s,a)$. Observe that $T^r_{\pi^*} V_r^{\pi_t} = M_{\pi^*} Q_r^{\pi_t}$ and $V_r^{\pi_t} = M_{\pi_t} Q_r^{\pi_t}$. The expressions for the constraint rewards are analogous. Rewriting the above expression,

$$\begin{aligned} [V_r^{\pi^*} - V_r^{\pi_t}] + \lambda_t [V_c^{\pi^*} - V_c^{\pi_t}] &= (I - \gamma P_{\pi^*})^{-1} \left[[M_{\pi^*} Q_r^{\pi_t} - M_{\pi_t} Q_r^{\pi_t}] + \lambda_t [M_{\pi^*} Q_c^{\pi_t} - M_{\pi_t} Q_c^{\pi_t}] \right] \\ &= (I - \gamma P_{\pi^*})^{-1} \left[[M_{\pi^*} - M_{\pi_t}] [Q_r^{\pi_t} + \lambda_t Q_c^{\pi_t}] \right] \\ &= (I - \gamma P_{\pi^*})^{-1} \left[[M_{\pi^*} - M_{\pi_t}] [\hat{Q}_r^{\pi_t} + \lambda_t \hat{Q}_c^{\pi_t}] \right] \\ &\quad + \underbrace{(I - \gamma P_{\pi^*})^{-1} \left[[M_{\pi^*} - M_{\pi_t}] [Q_r^{\pi_t} - \hat{Q}_r^{\pi_t} + \lambda_t (Q_c^{\pi_t} - \hat{Q}_c^{\pi_t})] \right]}_{\text{Error}} \end{aligned}$$

Let us first bound the maximum norm of the "Error" term,

$$\begin{aligned}\|\text{Error}\|_\infty &= \left\| (I-\gamma P_{\pi^*})^{-1} \left[[M_{\pi^*} - M_{\pi_t}] \left[Q_r^{\pi_t} - \hat{Q}_r^{\pi_t} + \lambda_t (Q_c^{\pi_t} - \hat{Q}_c^{\pi_t}) \right] \right] \right\|_\infty \\ &\leq \frac{1}{1-\gamma} \left\| [Q_r^{\pi_t} - \hat{Q}_r^{\pi_t} + \lambda_t (Q_c^{\pi_t} - \hat{Q}_c^{\pi_t})] \right\|_\infty \\ &\leq \frac{1}{1-\gamma} \left\| Q_r^{\pi_t} - \hat{Q}_r^{\pi_t} \right\|_\infty + \lambda_t \left\| Q_c^{\pi_t} - \hat{Q}_c^{\pi_t}] \right\|_\infty\end{aligned}$$

By assumption, $\left\| Q_r^{\pi_t} - \hat{Q}_r^{\pi_t} \right\|_\infty, \left\| Q_r^{\pi_t} - \hat{Q}_r^{\pi_t} \right\|_\infty \leq \varepsilon$.

$$\implies \|\text{Error}\|_\infty \leq \frac{\varepsilon}{1-\gamma}(1+\lambda_t)$$

Since the dual variables are projected onto the $[0, U]$ interval, $\lambda_t \leq U$, implying that

$$\|\text{Error}\|_\infty \leq \frac{\varepsilon}{1-\gamma}(1+U)$$

Substituting in this bound on the error, using the convention that left-multiplication by a measure means integration with respect to it,

$$\begin{aligned}[V_r^{\pi^*}(\rho) - V_r^{\pi_t}(\rho)] + \lambda_t [V_c^{\pi^*}(\rho) - V_c^{\pi_t}(\rho)] &\leq \rho (I-\gamma P_{\pi^*})^{-1} \left[[M_{\pi^*} - M_{\pi_t}] [\hat{Q}_r^{\pi_t} + \lambda_t \hat{Q}_c^{\pi_t}] \right] + \frac{\varepsilon}{1-\gamma}(1+U) \\ &\leq \frac{1}{1-\gamma} \nu_{\rho,\pi^*} \left[[M_{\pi^*} - M_{\pi_t}] [\hat{Q}_r^{\pi_t} + \lambda_t \hat{Q}_c^{\pi_t}] \right] + \frac{\varepsilon}{1-\gamma}(1+U)\,,\end{aligned}$$

where $\nu_{\rho,\pi^*} = (1-\gamma)\rho (I-\gamma P_{\pi^*})^{-1}$ is the discounted probability measure over the states obtained when starting from ρ and following π^*. Summing from $t=0$ to $T-1$ and dividing by T.

$$\frac{1}{T}\nu_{\rho,\pi^*} \sum_{t=0}^{T-1} \left[[V_r^{\pi^*}(\rho) - V_r^{\pi_t}(\rho)] + \lambda_t [V_c^{\pi^*}(\rho) - V_c^{\pi_t}(\rho)] \right] \leq \frac{\nu_{\rho,\pi^*}}{(1-\gamma)T} \sum_{t=0}^{T-1} \left[[\mathcal{M}_{\pi^*} - \mathcal{M}_{\pi_t}][\hat{Q}_r^{\pi_t} + \lambda_t \hat{Q}_c^{\pi_t}] \right] + \frac{\varepsilon}{1-\gamma}(1+U)$$

Now, observe that

$$\nu_{\rho,\pi^*} \sum_{t=0}^{T-1} \left[[\mathcal{M}_{\pi^*} - \mathcal{M}_{\pi_t}][\hat{Q}_r^{\pi_t} + \lambda_t \hat{Q}_c^{\pi_t}] \right] = \sum_{t=0}^{T-1} \langle \pi^*(\cdot|s) - \pi_t(\cdot|s), \hat{Q}_r^{\pi_t}(s,\cdot) + \lambda_t \hat{Q}_c^{\pi_t}(s,\cdot)] \rangle_{s\sim\nu_{\rho,\pi^*}} = \mathcal{R}^p(\pi^*, T)$$

Putting everything together,

$$\frac{1}{T}\sum_{t=0}^{T-1} [V_r^{\pi^*}(\rho) - V_r^{\pi_t}(\rho)] + \frac{1}{T}\sum_{t=0}^{T-1} \lambda_t [V_c^{\pi^*}(\rho) - V_c^{\pi_t}(\rho)] \leq \frac{\mathcal{R}^p(\pi^*, T)}{(1-\gamma)T} + \frac{\varepsilon}{1-\gamma}(1+U)\,. \tag{B.1}$$

The above result bounds the sub-optimality in the Lagrangian. Next, we will see how this result implies a bound on the sub-optimality in the objective and the constraint violation. To bound the reward sub-optimality, we will upper bound the negative of the second term on the left-hand side in the above equation, i.e., we upper bound $\frac{1}{T}\sum_{t=0}^{T-1} \lambda_t [V_c^{\pi_t}(\rho) - V_c^{\pi^*}(\rho)]$. We have,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \lambda_t [V_c^{\pi_t}(\rho) - V_c^{\pi^*}(\rho)] &\leq \frac{1}{T}\sum_{t=0}^{T-1} \lambda_t [V_c^{\pi_t}(\rho) - b] && \text{(since } V_c^{\pi^*}(\rho) \geq b\text{)} \\ &= \frac{1}{T}\sum_{t=0}^{T-1} \lambda_t [V_c^{\pi_t}(\rho) - \hat{V}_c^{\pi_t}(\rho)] + \frac{1}{T}\sum_{t=0}^{T-1} \lambda_t [\hat{V}_c^{\pi_t}(\rho) - b] \\ &= \frac{1}{T}\sum_{t=0}^{T-1} \lambda_t [V_c^{\pi_t}(\rho) - \hat{V}_c^{\pi_t}(\rho)] + \frac{\mathcal{R}^d(0,T)}{T} \\ &\leq U\varepsilon + \frac{\mathcal{R}^d(0,T)}{T}\,. && \text{(B.2)}\end{aligned}$$

Using Equations (B.1) and (B.2),

$$\texttt{OG} = \frac{1}{T}\sum_{t=0}^{T-1}[V_r^{\pi^*}(\rho) - V_r^{\pi_t}(\rho)] \leq \frac{\mathcal{R}^p(\pi^*, T) + (1-\gamma)\mathcal{R}^d(0,T)}{(1-\gamma)T} + \frac{\varepsilon}{1-\gamma}(1+U) + U\varepsilon \tag{B.3}$$

This proves the first part of the theorem. We now bound the constraint violation. For an arbitrary λ,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}[(\lambda_t - \lambda)(V_c^{\pi_t}(\rho) - b)] &= \frac{1}{T}\sum_{t=0}^{T-1}\left[(\lambda_t - \lambda)(V_c^{\pi_t}(\rho) - \hat{V}_c^{\pi_t}(\rho))\right] + \frac{1}{T}\sum_{t=0}^{T-1}\left[(\lambda_t - \lambda)(\hat{V}_c^{\pi_t}(\rho) - b)\right] \\ &= \frac{1}{T}\sum_{t=0}^{T-1}\left[(\lambda_t - \lambda)(V_c^{\pi_t}(\rho) - \hat{V}_c^{\pi_t}(\rho))\right] + \frac{\mathcal{R}^d(\lambda, T)}{T},\end{aligned}$$

implying

$$\frac{1}{T}\sum_{t=0}^{T-1}[(\lambda_t - \lambda)(V_c^{\pi_t}(\rho) - b)] \leq U\varepsilon + \frac{\mathcal{R}^d(\lambda, T)}{T}. \tag{B.4}$$

Adding Equation (B.4) and Equation (B.1) and reordering the terms gives

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}(V_r^{\pi^*}(\rho) - V_r^{\pi_t}(\rho)) + \frac{\lambda}{T}\sum_{t=0}^{T-1}(b - V_c^{\pi_t}(\rho)) \\ &\leq \frac{1}{T}\sum_{t=0}^{T-1}\underbrace{\lambda_t(b - V_c^{\pi^*}(\rho))}_{\leq 0 \text{ since } V_c^{\pi^*}(\rho) \geq b.} + \underbrace{\frac{\mathcal{R}^p(\pi^*, T) + (1-\gamma)\mathcal{R}^d(\lambda, T)}{(1-\gamma)T} + \frac{\varepsilon}{1-\gamma}(1+U) + U\varepsilon}_{h(\lambda)} \\ \implies &\frac{1}{T}\sum_{t=0}^{T-1}(V_r^{\pi^*}(\rho) - V_r^{\pi_t}(\rho)) + \frac{\lambda}{T}\sum_{t=0}^{T-1}(b - V_c^{\pi_t}(\rho)) \leq h(\lambda)\end{aligned}$$

We consider two cases: (i) if $\sum_{t=0}^{T-1}(b - V_c^{\pi_t}(\rho)) \geq 0$, we set $\lambda = U$, else, if (ii) $\sum_{t=0}^{T-1}(b - V_c^{\pi_t}(\rho)) < 0$, we set $\lambda = 0$. Using these choices, and since $\mathcal{R}^d(\lambda, T)$ is linearly increasing in λ,

$$\frac{1}{T}\sum_{t=0}^{T-1}(V_r^{\pi^*}(\rho) - V_r^{\pi_t}(\rho)) + \frac{U}{T}\left[\sum_{t=0}^{T-1}(b - V_c^{\pi_t}(\rho))\right]_+ \leq h(U)$$

Now take the policy π' such that $V_r^{\pi^*}(\rho) - V_r^{\pi'}(\rho) = \frac{1}{T}\sum_{t=0}^{T-1}(V_r^{\pi^*}(\rho) - V_r^{\pi_t}(\rho))$ and $V_c^{\pi^*}(\rho) - V_c^{\pi'}(\rho) = \frac{1}{T}\sum_{t=0}^{T-1}(b - V_c^{\pi_t}(\rho))$. Then,

$$[V_r^{\pi^*}(\rho) - V_r^{\pi'}(\rho)] + U\left[b - V_c^{\pi'}(\rho)\right]_+ \leq h(U).$$

Using Lemma B.3 with $C = U > \lambda^*$ and $\beta = h(U)$, we get

$$\begin{aligned}\texttt{CV} &= \frac{1}{T}\left[\sum_{t=0}^{T-1} b - V_c^{\pi_t}(\rho)\right]_+ = \left[b - V_c^{\pi'}(\rho)\right]_+ \\ &\leq \frac{h(U)}{U - \lambda^*} = \frac{\mathcal{R}^p(\pi^*, T) + (1-\gamma)\mathcal{R}^d(U, T)}{(U-\lambda^*)(1-\gamma)T} + \frac{1}{(U-\lambda^*)}\left[\frac{\varepsilon}{(1-\gamma)}(1+U) + U\varepsilon\right],\end{aligned}$$

which completes the proof subject to proving Lemma B.3. □

Lemma B.3 (Constraint violation bound)**.** *For any $C > \lambda^*$ and any $\tilde{\pi}$ s.t. $V_r^{\pi^*}(\rho) - V_r^{\tilde{\pi}}(\rho) + C[b - V_c^{\tilde{\pi}}(\rho)]_+ \leq \beta$, we have $[b - V_c^{\tilde{\pi}}(\rho)]_+ \leq \frac{\beta}{C - \lambda^*}$.*

Proof. Define $\nu(\tau) = \max_\pi \{V_r^\pi(\rho) \mid V_c^\pi(\rho) \geq b + \tau\}$ and note that by definition, $\nu(0) = V_r^{\pi^*}(\rho)$ and that ν is a decreasing function for its argument.

Let $V_l^{\pi,\lambda}(\rho) = V_r^\pi(\rho) + \lambda(V_c^\pi(\rho) - b)$. Then, for any policy π s.t. $V_c^\pi(\rho) \geq b + \tau$, we have

$$\begin{aligned}
V_l^{\pi,\lambda^*}(\rho) &\leq \max_{\pi'} V_l^{\pi',\lambda^*}(\rho) \\
&= V_r^{\pi^*}(\rho) && \text{(by strong duality)} \\
&= \nu(0) && \text{(from above relation)} \\
\implies \nu(0) - \tau\lambda^* &\geq V_l^{\pi,\lambda^*}(\rho) - \tau\lambda^* = V_r^\pi(\rho) + \lambda^* \underbrace{(V_c^\pi(\rho) - b - \tau)}_{\text{Positive}} \\
\implies \nu(0) - \tau\lambda^* &\geq \max_\pi \{V_r^\pi(\rho) \mid V_c^\pi(\rho) \geq b + \tau\} = \nu(\tau)\,. \\
\implies \tau\lambda^* &\leq \nu(0) - \nu(\tau)\,. && \text{(B.5)}
\end{aligned}$$

Now we choose $\tilde{\tau} = -(b - V_c^{\tilde{\pi}}(\rho))_+$.

$$\begin{aligned}
(C - \lambda^*)|\tilde{\tau}| &= \lambda^*\tilde{\tau} + C|\tilde{\tau}| && \text{(since } \tilde{\tau} \leq 0\text{)} \\
&\leq \nu(0) - \nu(\tilde{\tau}) + C|\tilde{\tau}| && \text{(Equation (B.5))} \\
&= V_r^{\pi^*}(\rho) - V_r^{\tilde{\pi}}(\rho) + C|\tilde{\tau}| + V_r^{\tilde{\pi}}(\rho) - \nu(\tilde{\tau}) && \text{(definition of } \nu(0)\text{)} \\
&= V_r^{\pi^*}(\rho) - V_r^{\tilde{\pi}}(\rho) + C(b - V_c^{\tilde{\pi}}(\rho))_+ + V_r^{\tilde{\pi}}(\rho) - \nu(\tilde{\tau}) \\
&\leq \beta + V_r^{\tilde{\pi}}(\rho) - \nu(\tilde{\tau})\,.
\end{aligned}$$

Now let us bound $\nu(\tilde{\tau})$:

$$\begin{aligned}
\nu(\tilde{\tau}) &= \max_\pi \{V_r^\pi(\rho) \mid V_c^\pi(\rho) \geq b - (b - V_c^{\tilde{\pi}}(\rho))_+\} \\
&\geq \max_\pi \{V_r^\pi(\rho) \mid V_c^\pi(\rho) \geq V_c^{\tilde{\pi}}(\rho)\} && \text{(tightening the constraint)} \\
\nu(\tilde{\tau}) \geq V_r^{\tilde{\pi}}(\rho) &\implies (C - \lambda^*)|\tilde{\tau}| \leq \beta \implies (b - V_c^{\tilde{\pi}}(\rho))_+ \leq \frac{\beta}{C - \lambda^*}
\end{aligned}$$

□

B.1 PROOF OF LEMMA 4.1

Lemma B.4. *The objective Equation (1) satisfies strong duality, and the optimal dual variables are bounded as* $\lambda^* \leq \frac{1}{\zeta(1-\gamma)}$*, where* $\zeta := \max_\pi V_c^\pi(\rho) - b > 0$.

Proof. Starting from the Lagrangian form in Equation (3),

$$V_r^*(\rho) := \max_\pi \min_{\lambda \geq 0} V_r^\pi(\rho) + \lambda[V_c^\pi(\rho) - b]$$

Using the linear programming formulation of CMDPs in terms of the state-occupancy measures μ, we know that both the objective and the constraint are linear functions of μ, and strong duality holds w.r.t μ. Since μ and π have a one-one mapping, we can switch the min and the max (Paternain et al., 2019), implying,

$$\begin{aligned}
V_r^*(\rho) &= \min_{\lambda \geq 0} \max_\pi V_r^\pi(\rho) + \lambda[V_c^\pi(\rho) - b] \\
&= \max_\pi V_r^\pi(\rho) + \lambda^*[V_c^\pi(\rho) - b]\,.
\end{aligned}$$

Define $\tilde{\pi} := \arg\max_\pi V_c^\pi(\rho)$. Then,

$$\begin{aligned}
V_r^*(\rho) &\geq V_r^{\tilde{\pi}}(\rho) + \lambda^*[V_c^{\tilde{\pi}}(\rho) - b] \\
\implies \lambda^* &\leq \frac{V_r^*(\rho) - V_r^{\tilde{\pi}}(\rho)}{[V_c^{\tilde{\pi}}(\rho) - b]} \leq \frac{1}{(1-\gamma)\zeta}\,.
\end{aligned}$$

□

B.2 PROOFS FOR SECTION 5

Lemma B.5. *For policy π, any distribution ω and subset $\mathcal{C}$, if we use m trajectories to estimate the action-value function for each $(s,a) \in \mathcal{C}$, and solve Equation (10) to compute $\hat{\theta}_r^\pi$, then for any $(s,a) \in (\mathcal{S} \times \mathcal{A})$ pair, the error $|\langle \phi(s,a), \hat{\theta}_r^\pi \rangle - Q_r^\pi|$ can be upper-bounded by*

$$\varepsilon_b(1 + \|\phi(s,a)\|_{G_\omega^\dagger}) + \frac{\|\phi(s,a)\|_{G_\omega^\dagger}}{1-\gamma} \sqrt{\frac{\log(2|\mathcal{C}|/\delta)}{2m}},$$

where $G_\omega = \sum_{(s,a)\in\mathcal{C}} \omega(s,a)\phi(s,a)\phi(s,a)^\intercal$ and $A^\dagger$ is pseudoinverse of A.

Proof. By solving $\hat{\theta}_r^\pi = \arg\min_\theta \sum_{z\in\mathcal{C}} \omega(z) \left[\langle \theta, \phi(z)\rangle - q_r(z)\right]^2$ (Equation (10)), we get that

$$\theta_r^\pi = G_\omega^\dagger \sum_{(s,a)\in\mathcal{C}} \omega(s,a)\phi(s,a)\, q_r^\pi(s,a)$$

and lets denote $z = (s,a)$, $\phi(z) = \phi(s,a)$ and $\varepsilon(z) = q_r(z) - Q_r^\pi(z) + Q_r^\pi(z) - \langle \phi(z), \theta_r^* \rangle$ and $\theta_r^* := \arg\min_\theta \max_{(s,a)} \|Q_r^\pi(s,a) - \langle \theta, \phi(s,a)\rangle\|$ is the optimal parameter for the given policy π.

Therefore we can write $q_r(z) = \varepsilon(z) + \langle \phi(z), \theta_r^* \rangle$

$$\begin{aligned}
|\langle \phi(z), \theta_r^\pi \rangle - Q_r^\pi| &= |\langle \phi(z), \theta_r^\pi \rangle - \langle \phi(z), \theta_r^* \rangle + \langle \phi(z), \theta_r^* \rangle - Q_r^\pi| \\
&\le |\langle \phi(z), \theta_r^\pi \rangle - \langle \phi(z), \theta_r^* \rangle| + \varepsilon_b \quad [\varepsilon_b \text{ from } \textit{Assumption } 5.1]
\end{aligned}$$

Now we need to bound the first term of above inequality. Based on the definition of $\varepsilon(z)$, we can write $\theta_r^\pi = G_\omega^\dagger \sum_{z'\in\mathcal{C}} (\langle \phi(z'), \theta_r^* \rangle + \varepsilon(z'))\, \omega(z')\phi(z')$. Using this equality we can get easily that:

$$\begin{aligned}
|\langle \phi(z), \theta_r^\pi \rangle - \langle \phi(z), \theta_r^* \rangle| &= |\sum_{z'\in\mathcal{C}} \varepsilon(z')\omega(z')\phi(z)^T G_\omega^\dagger \phi(z')| \\
&\le \sum_{z'\in\mathcal{C}} |\varepsilon(z')||\omega(z')\phi(z)^T G_\omega^\dagger \phi(z')| \\
&\le |\max_{z'\in\mathcal{C}} \varepsilon(z')| \sum_{z'\in\mathcal{C}} |\omega(z')\phi(z)^T G_\omega^\dagger \phi(z')|
\end{aligned}$$

To bound the sum term we can

$$\begin{aligned}
\left(\sum_{z'\in\mathcal{C}} \omega(z')|\phi(z)^T G_\omega^\dagger \phi(z')|\right)^2 &\le \sum_{z'\in\mathcal{C}} \omega(z') \left(|\phi(z)^T G_\omega^\dagger \phi(z')|\right)^2 \qquad \text{(Jensen's inequality)} \\
&= \phi(z)^T G_\omega^\dagger \left(\sum_{z'\in\mathcal{C}} \left[\omega(z')\phi(z')\phi(z')^T\right]\right) G_\omega^\dagger \phi(z) \\
&= \|\phi(z)\|^2_{G_\omega^\dagger}
\end{aligned}$$

To finish the proof we need to bound $|\max_{z'\in\mathcal{C}} \varepsilon(z')|$. Based on the definition of $\varepsilon(z')$ we have

$$\begin{aligned}
|\varepsilon(z')| &\le |q_r(z') - Q_r^\pi(z')| + |Q_r^\pi(z') - \langle \phi(z'), \theta_r^* \rangle| \\
&\le |q_r(z) - Q_r^\pi(z)| + \varepsilon_b \\
&\le \frac{1}{1-\gamma} \sqrt{\frac{\log(2/\delta)}{2m}} + \varepsilon_b
\end{aligned}$$

where the second inequality is due to function approximation error (Assumption 5.1) and the last inequality comes from Hoeffding's inequality. Specifically, since the m trajectories are independent, and the action-value functions lie in the $[0, 1/1-\gamma]$ range, we use Hoeffding's inequality to conclude that the sampling error for each

$z \in \mathcal{C}$ can be upper-bounded by $\frac{1}{1-\gamma}\sqrt{\frac{\log(2/\delta)}{2m}}$. Since we desire uniform control over all states and actions in $\mathcal{C}$, by union bound, with probability $1-\delta$, $|q_r(z) - Q_r^\pi(z)| \leq \frac{1}{1-\gamma}\sqrt{\frac{\log(2|\mathcal{C}|/\delta)}{2m}}$. Putting everything together we get the result. □

Corollary B.6. *Under Assumption 5.1, `OG` and `CV` of CBP can be bounded as:*

$$\texttt{OG} \leq \frac{\left(\frac{3(1+U)\sqrt{1+KL(\pi_0||\pi^*)}}{1-\gamma} + \Psi\right)}{(1-\gamma)\sqrt{T}} + \frac{\tilde{\varepsilon}(1+2U)}{1-\gamma},$$
$$\texttt{CV} \leq \frac{\zeta\left(\frac{3(1+U)\sqrt{1+KL(\pi_0||\pi^*)}}{1-\gamma} + \Psi\right)}{\sqrt{T}} + \zeta\,\tilde{\varepsilon}(1+2U),$$

where $U = \frac{2}{\zeta(1-\gamma)}$, $\tilde{\varepsilon} = \varepsilon_b(1+\sqrt{d}) + \frac{\sqrt{d}}{1-\gamma}\sqrt{\frac{\log(2d(d+1)/\delta)}{2m}}$ *and* $\Psi = 4U\sqrt{\log((T+1)U)} + 1$.

Proof. The proof is similar to the proof of Corollary A.2 but with a different $\tilde{\varepsilon}$. □

C ADDITIONAL IMPLEMENTATION DETAILS

In Appendix C.1, we describe a more practical variant of CBP and in Appendix C.2, we describe the offline G-experimental design procedure required to form the coreset $\mathcal{C}$ for CBP. Details about the synthetic tabular environment are presented in Appendix C.3 whereas Appendix C.4 details the hyperparameters used across the different experiments.

C.1 PRACTICAL COIN-BETTING POLITEX ALGORITHM

We present the practical version of CBP which uses a parameter α_λ in Algorithm 1.

Algorithm 1: Practical Coin-Betting Politex

Input: $\alpha_\lambda > 0$ (parameter), π_0 (policy initialization), λ_0 (dual variable initialization), m (Number of trajectories), T (Number of iterations), Feature map Φ.

Initialize: $L_0 = 0$

Compute coreset $\mathcal{C}$ and distribution ω

Solve the unconstrained problem $\max_\pi \hat{V}_c^\pi(\rho)$ to estimate ζ in Lemma 4.1 and set $U = \frac{2}{\zeta(1-\gamma)}$.

for $t \leftarrow 0$ **to** $T-1$ **do**

For every $(s,a) \in \mathcal{C}$, use m trajectories starting from (s,a) using policy π_t and estimate the action-value functions $q_r(s,a)$ and $q_c(s,a)$.

Compute and store $\hat{\theta}_r^{\pi_t}$ and $\hat{\theta}_c^{\pi_t}$ using Equation (10).

for *every s encountered in the trajectory generated by π_t, and for every a* **do**

Compute $\hat{Q}_r^t(s,a) = \langle \hat{\theta}_r^{\pi_t}, \phi(s,a) \rangle$; $\hat{Q}_c^t(s,a) = \langle \hat{\theta}_c^{\pi_t}, \phi(s,a) \rangle$ and $\hat{Q}_l^t(s,a) = \hat{Q}_r^t(s,a) + \lambda_t \hat{Q}_c^t(s,a)$.

Update policy,

$$\hat{A}_l^t(s,a) = \frac{1-\gamma}{1+U}\left[\hat{Q}_l^t(s,a) - \left\langle \hat{Q}_l^t(s,\cdot), \pi_t(\cdot|s) \right\rangle\right]$$

$$\tilde{A}_l^t(s,a) = \hat{A}_l^t(s,a)\,\mathcal{I}\{w_t(s,a) > 0\} + [\hat{A}_l^t(s,a)]_+\,\mathcal{I}\{w_t(s,a) \le 0\}$$

$$w_{t+1}(s,a) = \frac{\sum_{i=0}^t \tilde{A}_l^i(s,a)}{(t+1) + T/2}\left(1 + \sum_{i=0}^t \tilde{A}_l^i(s,a)\, w_i(s,a)\right)$$

$$\pi_{t+1}(a|s) = \begin{cases} \pi_0(a|s), & \text{if } \sum_a \pi_0(a|s)\,[w_{t+1}(s,a)]_+ = 0 \\ \frac{\pi_0(a|s)\,[w_{t+1}(s,a)]_+}{\sum_a \pi_0(a|s)\,[w_{t+1}(s,a)]_+}, & \text{otherwise.} \end{cases}$$

end

Update dual variable,

$$\hat{V}_c^t(\rho) = \langle \rho(\cdot), \langle \hat{Q}_c^t(s,\cdot), \pi_t(\cdot|s) \rangle \rangle$$

$$g_t = b - \hat{V}_c^t(\rho)$$

$$L_t = \max(L_{t-1}, |g_t|)$$

$$\lambda_{t+1} = \lambda_0 + \frac{\sum_{i=0}^t g_i}{L_t \max(\sum_{i=0}^t |g_i| + L_t, \alpha_\lambda L_t)}\left(L_t + \sum_{i=0}^t [(\lambda_i - \lambda_0) g_i]_+\right)$$

end

C.2 OFFLINE G-EXPERIMENTAL DESIGN TO BUILD CORESET $\mathcal{C}$

We use offline G-experimental design to form the coreset in Line 2 of Algorithm 1. In particular, we use the greedy iterative algorithm in Algorithm 2 to build $\mathcal{C}$: in iteration τ, go through all the states and actions adding the (s, a) pair (to $\mathcal{C}$) with the highest marginal gain computed as $g_\tau(s, a) := \|\phi(s, a)\|_{G_\tau^\dagger}$. Here G_τ is the Gram matrix formed by the features of the (s, a) pairs present in $\mathcal{C}$ at iteration τ. For a specified input $\varepsilon' > 0$, the algorithm terminates at iteration T when $\max_{(s,a)} g_{\mathrm{T}}(s, a) \leq \varepsilon'$. Hence, the algorithm directly controls $\sup_{(s,a)} \|\phi(s, a)\|_{G_\omega^\dagger} \leq \varepsilon'$ in Lemma B.5, and hence controls ε_s in practice. However, this procedure does not have a guarantee on how large $|\mathcal{C}|$ can be. In practice, we set ε' such that $|\mathcal{C}| = O(d)$. Although we only consider forming the coreset in an offline manner that involves iterating through all SA state-action pairs, efficient online variants forming the coreset while running the algorithm have been developed recently (Li et al., 2021). Such techniques are beyond the scope of this paper and we plan to explore them in future work.

Algorithm 2: Coreset $\mathcal{C}$ formation using G-experimental design

Input: Φ (Feature map), $\varepsilon' > 0$ (tolerance parameter), $\nu = 1$ (default value).
Initialize: $G^\dagger = \frac{1}{\nu}\mathcal{I}_d$, $\mathcal{C} = \emptyset$, $g_{max} = \infty$ (maximum marginal gain).
while $g_{max} \leq \varepsilon'$ **do**
  $g_{max} = 0$
  for $\forall (s, a) \in (\mathcal{S} \times \mathcal{A})$ **do**
    Compute $g(s, a) = \sqrt{\phi(s, a)^\intercal G^\dagger \phi(s, a)}$ [marginal gain]
    if $g_{max} < g(s, a)$ **then**
      $(s^*, a^*) = (s, a)$
      $g_{max} = g(s, a)$
    end
  end
  $\mathcal{C} = \mathcal{C} \cup \{(s^*, a^*)\}$
  $G^\dagger = G^\dagger - \frac{G^\dagger \phi(s^*,a^*) \phi(s^*,a^*)^\intercal G^\dagger}{1+\phi(s^*,a^*)^\intercal G^\dagger \phi(s^*,a^*)}$ [Sherman-Morrison to compute $\left(G + \phi(s^*, a^*)\, \phi(s^*, a^*)^\intercal\right)^\dagger$]
end

C.3 SYNTHETIC TABULAR ENVIRONMENT

In Figure 6, we show the synthetic tabular environment which is modified from Example 3.5 (Sutton and Barto, 2018) to add the constraint rewards.

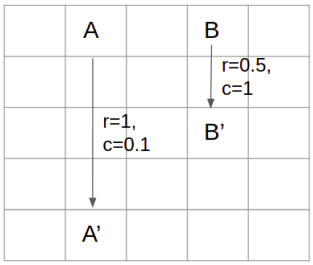


Figure 6: **Tabular environment:** A $5X5$ gridworld environment where all actions results in reward(r) and constraint reward(c) as 0, except special states denoted by A and B. All four actions in states (A, B) transitions the agent to states (A', B') and results in reward as $(1, 0.5)$ and constraint rewards as $(0.1, 1)$ respectively. The remaining transitions incur zero reward and zero constrain reward.

C.4 HYPER-PARAMETERS

Table 1: **Hyperparameters for tabular setting:** Shows the hyperparameters for different algorithms CBP, GDA and CRPO for experiments in Appendix D.1.

Experiments	**CBP**	**GDA**	**CRPO**
Model-based	$\alpha_\lambda = 8$	$\alpha_\pi = 1.0$, $\alpha_\lambda = 0.1$	$\alpha_\pi = 0.75, \eta = 0.0$
Model-free	$\alpha_\lambda = 8$	$\alpha_\pi = 1.0$, $\alpha_\lambda = 0.1$	$\alpha_\pi = 0.75, \eta = 0.0$

Table 2: **Hyperparameters for LFA setting with sampling :** Shows the hyperparamters used for different d dimension features for CBP, GDA, CRPO with fixed number of samples for $\hat{Q}$ approximations (for gridworld experiments in Appendix D.2).

Algorithms	$d = 40$	$d = 56$	$d = 80$
CBP	$\alpha_\lambda = 0.25$	$\alpha_\lambda = 0.25$	$\alpha_\lambda = 0.1$
GDA	$\alpha_\pi = 1.0$, $\alpha_\lambda = 0.1$	$\alpha_\pi = 1.0$, $\alpha_\lambda = 1.0$	$\alpha_\pi = 1.0$, $\alpha_\lambda = 0.1$
CRPO	$\alpha_\pi = 0.75$	$\alpha_\pi = 0.75$	$\alpha_\pi = 0.75$

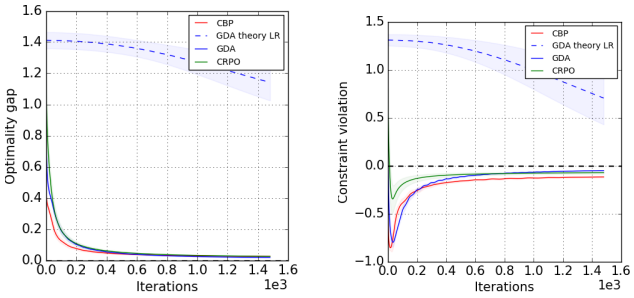
Table 3: **Hyperparameters for Cartpole environment**: Shows the best hyperparameter for different values of entropy regularization coefficient ν and different algorithms namely CBP, GDA and CRPO (for experiments in Appendix D.2).

Algorithms	$\boldsymbol{\nu = 0}$	$\boldsymbol{\nu = 0.1}$	$\boldsymbol{\nu = 0.01}$	$\boldsymbol{\nu = 0.001}$
CBP	$\alpha_\lambda = 0.1$	$\alpha_\lambda = 0.1$	$\alpha_\lambda = 5.0$	$\alpha_\lambda = 0.5$
GDA	$\alpha_\pi = 0.01$, $\alpha_\lambda = 0.001$	$\alpha_\pi = 0.001$, $\alpha_\lambda = 0.001$	$\alpha_\pi = 0.1$, $\alpha_\lambda = 0.1$	$\alpha_\pi = 0.01$, $\alpha_\lambda = 0.0001$
CRPO	$\alpha_\pi = 0.1$, $\eta = 10$	$\alpha_\pi = 0.1$, $\eta = 0$	$\alpha_\pi = 0.5$, $\eta = 10$	$\alpha_\pi = 0.5$, $\eta = 0$

D ADDITIONAL EXPERIMENTAL RESULTS

D.1 TABULAR SETTING

Model-based setting: In Figure 7, we demonstrate performance – optimality gap (OG) and constraint violation (CV) – with best hyperparameters for three algorithms namely, CBP, GDA and CRPO. In addition, we show the performance of GDA with theoretical learning rates of π and λ to focus on the importance of tuning GDA's hyperparameter for practical purpose. We observe OG converges to zero quickly for our CBP as compared to GDA and CRPO with constraint satisfaction (when $CV \leq 0$). The ideal performance metric is when both OG and CV converges to 0 value. Refer Table 1 for best values of hyperparameter. We used the following ranges of hyperparameters. For CBP, $\alpha_\lambda = \{1, 2, 5, 8, 15, 50, 100, 300, 500\}$. The hyperparameter of GDA varied as $\alpha_\pi = \{0.001, 0.01, 0.1, 1.0\}$ (learning rate policy) and $\alpha_\lambda = \{0.0001, 0.001, 0.01, 0.1, 1.0\}$ (learning rate dual variable). For CRPO, the learning rate of policy varied as $\alpha_\pi = \{0.001, 0.01, 0.05, 0.1, 0.5, 0.75\}$ and tolerance parameter $\eta = \{0, 0.25\}$.



(a) OG (b) CV

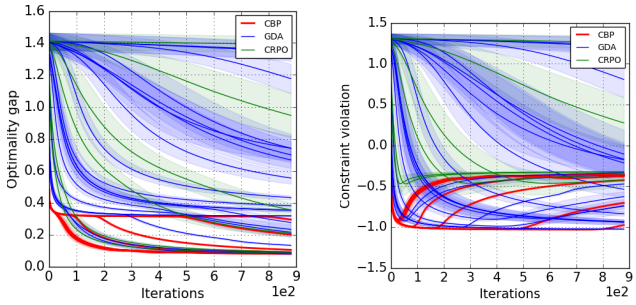
Figure 7: **Model-based in tabular case:** OG and CV for CB, GDA and CRPO with best hyperparameters. The results are averaged over 5 runs with 95% confidence interval. We assume that we have access to true CMDP. The best hyperparameter has the least OG and satisfies the condition $CV \in [-0.25, 0]$. We also show the performance of baseline GDA with theoretical $\alpha_\pi = \sqrt{\frac{2\log|A|}{T}}\frac{1-\gamma}{1+U}$ and $\alpha_\lambda = \frac{U(1-\gamma)}{\sqrt{T}}$. Here, $U = \frac{2}{\zeta(1-\gamma)}$. This is shown in blue dashed line.

Model-free setting: Here, we test the performance of algorithms in the model-free setting (don't have access to true CMDP model). We use TD(0) based sampling approach (Sutton, 1988) to estimate the Q action-value function. We sample data for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. In Figure 3, we observe the effect on performance by varying the number of samples for Q action-value estimation. Here, we consider one-hot encoded features (no overlapping of features). We observe that CBP consistently converges faster than its counterpart in the sampling-based approach. Further, it also matches the expectation that the performance improves with the increase in number of samples. In Figures 8 and 9, we show the robustness of CBP with hyperparameter sensitivity and environment misspecification respectively. The hyperparameters used for these experiments are presented in Table 1 in Appendix C.4.

D.2 LINEAR SETTING

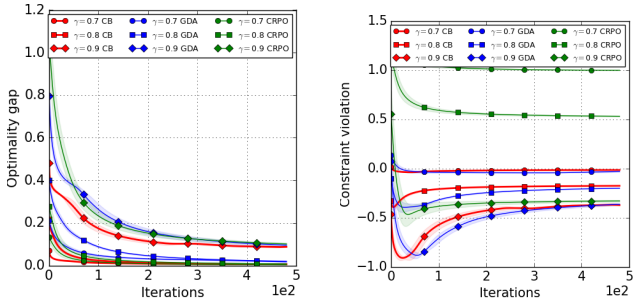
Gridworld environment: We use 5×5 gridworld environment as show in Figure 6. Tile coding is used to learn the feature representation for every (s, a) pair in the environment. Number of tilings used are 1 and we vary the tiling size to change the dimension of the features (feature overlap for multiple (s, a) pairs). In Figure 10 we show hyperparameter sensitivity on performance for all the three algorithms with different d dimension of features. The values of all other parameters were kept fixed. Similar observation holds here, CBP is robust to varying values of hyperparameters. The range of hyperparameter is similar to one in Figure 8.

G-experimental design for gridworld environment: In Figure 11 we show the performance with G-experimental design (Appendix C.2). Here subset of $(s, a) \in \mathcal{C}$ pairs are chosen from a *coreset*.



(a) OG (b) CV

Figure 8: **Sensitivity to hyperparameters in model-free gridworld environment:** Performance with different hyperparameters for CB, GDA and CRPO. The results are averaged over 5 runs with 95% confidence interval. We used 2000 samples for $\forall(s,a) \in \mathcal{S} \times \mathcal{A}$ to estimate Q values for both reward and cost. The results are demonstrated on one-hot features (with no feature overlap). The hyperparameters for CB are $\alpha_\lambda = \{1, 2, 5, 8, 15, 50, 100, 300, 500\}$. The hyperparameter for GDA are $\alpha_\pi = \{0.001, 0.01, 0.1, 1.0\}$ (learning rates for policy) and $\alpha_\lambda = \{0.0001, 0.001, 0.01, 0.1, 1.0\}$ (learning rate for dual variable). The hyperparameter for CRPO are $\alpha_\pi = \{0.001, 0.01, 0.05, 0.1, 0.5, 0.75\}$. We use $\eta = 0.0$ for CRPO. The key observation is that CBP is robust against the variations in hyperparameters with a smaller variance in performance against multiple runs.



(a) OG (b) CV

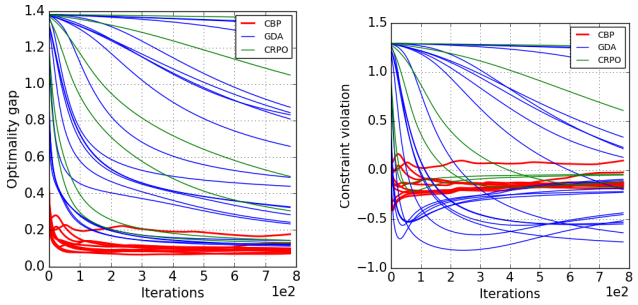
Figure 9: **Environment misspecification in model-free gridworld by varying discount factor γ:** We only introduce sampling error by estimating Q function with 2000 samples for all (s,a) pair for all three algorithms (no feature overlap). We vary the discount factor $\gamma = \{0.7, 0.8\}$ to observe the effects of environment misspecification on CBP, GDA and CRPO. We keep the hyperparameters fixed for all the algorithms, similar to the one for original CMDP with $\gamma = 0.9$. The results are averaged over 5 runs with 95% confidence interval.The hyperparameters used are reported in Table 1. We observe that CRPO does not even satisfy constraint for case when $\gamma = 0.8$ ($CV > 0$). Further, our CBP converges consistently faster than the baselines.

Exploration in continuous state-spaces: We used G-experimental design for the discrete state-action environment in the previous section. However, such a procedure is difficult to implement for the continuous state-action spaces we consider in this section. In order to achieve enough exploration in practice, similar to Xu et al. (2021), we use entropy regularization (Geist et al., 2019; Cen et al., 2021) for the policy updates. Specifically, for a specified regularization parameter ν, our task is to find a sequence of policies $\{\pi_0, \pi_1, \ldots, \pi_{T-1}\}$ that minimize the regularized primal regret,

$$\mathcal{R}_\nu^p(\pi^*, T) := \sum_{t=0}^{T-1} \sum_{s=0}^{\mathcal{S}-1} \left[\langle \pi^*(\cdot|s) - \pi_t(\cdot|s), \hat{Q}_r^t + \lambda_t \hat{Q}_c^t \rangle + \nu\, d^{\pi_t}(s) \sum_{a \in \mathcal{A}} \pi(a|s) \log(\pi(a|s))\right].$$

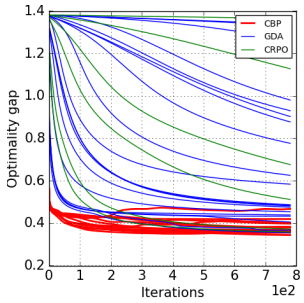
It can be easily seen (Geist et al., 2019) that the form of the algorithm updates remain the same, but the action-value functions for policy π need to be redefined to depend on the "effective reward" equal to $r(s,a) - \nu \log \pi(a|s)$. Therefore, the new $\hat{Q}_l^t$ with exploration is equal to $\hat{Q}_l^t(s,a) = \hat{Q}_r^t(s,a) + \lambda_t \hat{Q}_c^t(s,a) - \nu \log \pi_t(a|s)$.

Cartpole environment : We added two constraint rewards (c_1, c_2) to the classic OpenAI gym Cartpole environment. (1) Cart receives a $c_1 = 0$ constraint reward value when enters the area

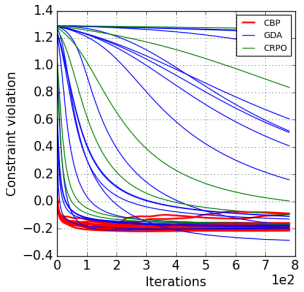


(a) OG ($d = 40$)

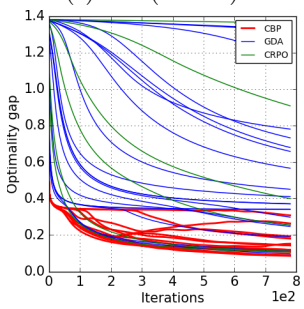
(b) CV ($d = 40$)



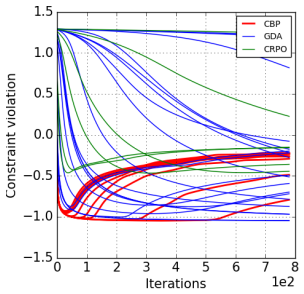
(c) OG ($d = 56$)



(d) CV ($d = 56$)

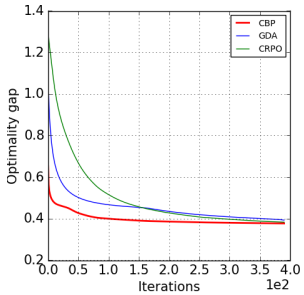


(e) OG ($d = 80$)

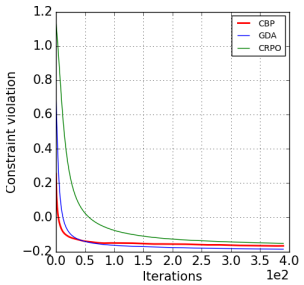


(f) CV ($d = 80$)

Figure 10: **Linear function approximation in gridworld environment:** We approximate the Q value using LSTDQ with 300 samples for each (s, a) pair. The dimension of the features (denoted by d) are varied to observe the sensitivity to a range of hyperparameter values. We kept all the other parameters fixed. We use (a,b) $d = 40$, (b,c) $d = 56$, (e,f) $d = 80$ dimension features respectively. CBP is consistently robust to variation in the hyperparameters for different dimensions as compared to baselines GDA and CRPO.



(a) OG



(b) CV

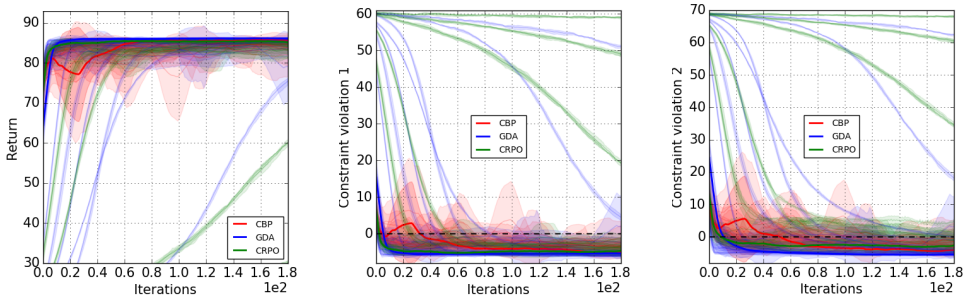
Figure 11: **G-experimental design:** We show the performance with G-experimental design, where subset of $(s, a) \in \mathcal{C}$ are chosen to learn the weight vectors of Q function. Here, we are in model-free setting with $d = 56$ dimensional features in LFA. We used 300 samples for $c \in \mathcal{C}$ to learn the estimate of Q for all three algorithms.

$[-2.4, -2.2], [-1.3, -1.1], [1.1, 1.3], [2.2, 2.4]$, else receive $c_1 = +1$. (2) When the angle of the cart is less than 4

degrees receive $c_2 = +1$, else everywhere $c_2 = 0$. Each episode length is no longer than 200.

We used tile coding (Sutton and Barto, 2018) to discretize the continuous state space of the environment. The dimension of the features is 2^{12}. We used 8 number of tilings with each grid size 4×4. For experimenting the effect of adding exploration on the performance, we incorporated the entropy coefficient (Haarnoja et al., 2018; Geist et al., 2019). We varied the entropy regularizer $\nu = \{0, 0.1, 0.01, 0.001\}$. Refer to Figure 12 for the experiment with ν coefficient.

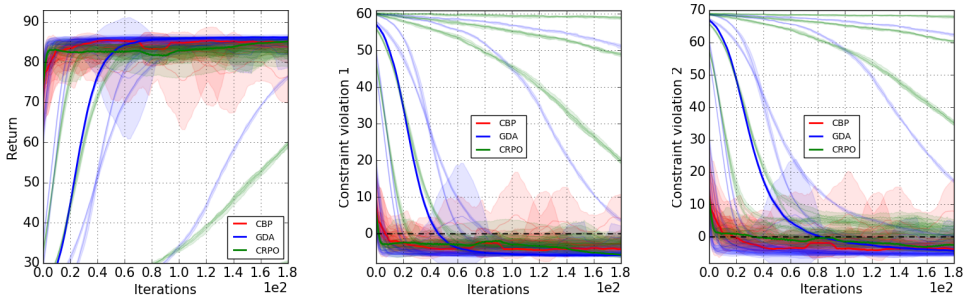
We conducted the experiments with following α_λ parameter value of CBP $\{0.1, 0.5, 5, 50, 250, 500, 750, 1000\}$. For GDA, we varied the learning rate of policy $\alpha_\pi = \{0.1, 0.01, 0.001, 0.0001\}$ and learning rate of dual variable $\alpha_\lambda = \{0.1, 0.01, 0.001, 0.0001\}$. For CRPO baseline, the following values of learning rate of policy $\alpha_\pi = \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$ are experimented with. We kept the tolerance parameter η of CRPO as $\{0, 10\}$. The best hyperparameters are summarized in Table 3 for the different values of entropy regularizer ν.



(a) Return ($\nu = 0$)

(b) CV 1 ($\nu = 0$)

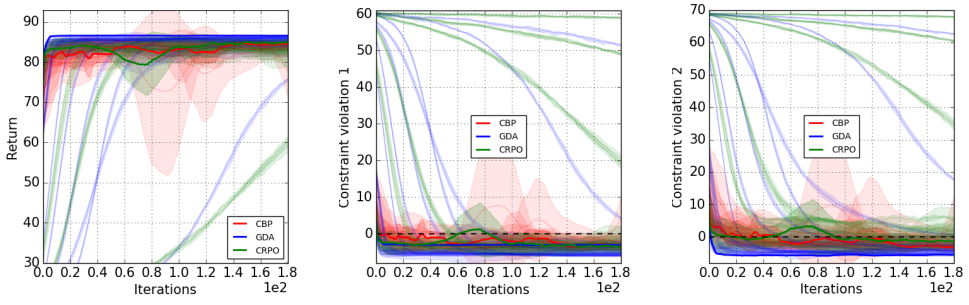
(c) CV 2 ($\nu = 0$)



(d) Return ($\nu = 0.1$)

(e) CV 1 ($\nu = 0.1$)

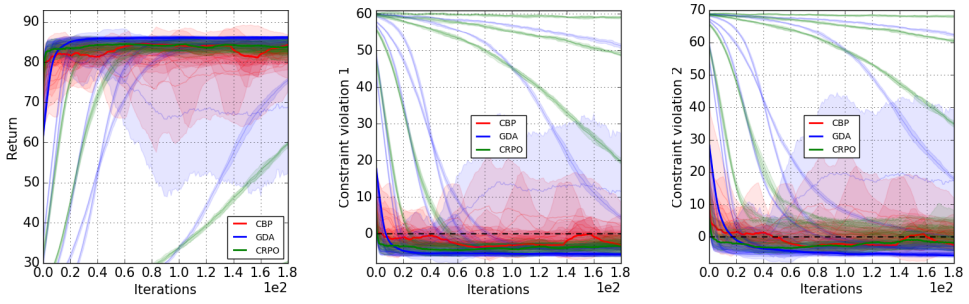
(f) CV 2 ($\nu = 0.1$)



(g) Return ($\nu = 0.01$)

(h) CV 1 ($\nu = 0.01$)

(i) CV 2 ($\nu = 0.01$)



(j) Return ($\nu = 0.001$)

(k) CV 1 ($\nu = 0.001$)

(l) CV 2 ($\nu = 0.001$)

Figure 12: **Cartpole environment:** We show the sensitivity to entropy regularization ν for all three algorithms CBP, GDA and CRPO. The performance is averaged over 5 runs with 95% confidence interval. Different rows corresponds to different value of $\nu = \{0, 0.1, 0.01, 0.001\}$. Darker lines show the performance with the best hyperparameters. Lighter shade lines show performance with other values of hyperparameters. The range of hyperparameter for CBP is $\alpha_\lambda = \{0.1, 0.5, 5, 50, 250, 500, 750, 1000\}$. For GDA, we vary learning rates of both policy and dual variable as $\{0.1, 0.01, 0.001, 0.0001\}$. For CRPO, we vary $\alpha_\pi = \{0.1, 0.5, 0.01, 0.001, 0.005\}$ and tolerance hyperparameter as $\eta = \{0, 10\}$.

References

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 2021.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained Markov decision processes. *Advances in Neural Information Processing Systems*, 33, 2020.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR, 2019.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

Gen Li, Yuxin Chen, Yuejie Chi, Yuantao Gu, and Yuting Wei. Sample-efficient reinforcement learning is feasible for linearly realizable MDPs with limited revisiting. *Advances in Neural Information Processing Systems*, 34, 2021.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Francesco Orabona and David Pal. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29:577–585, 2016.

Francesco Orabona and Tatiana Tommasi. Training deep networks without learning rates through coin betting. *Advances in Neural Information Processing Systems*, 30, 2017.

Santiago Paternain, Luiz FO Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 7555–7565, 2019.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. second. *A Bradford Book*, 2018.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. CRPO: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pages 11480–11491. PMLR, 2021.