We start by addressing (5.1). KPI stands for Key Performance Index. For Target Set Objectives, specifying \(U = \{u : w_k \geq \rho_k \forall 1 \leq k \leq K\}\) is sufficient for ensuring \(\bar{V}_{1:T,k} \geq \rho_k\) whenever possible, thanks to the \(\min_{u \in U}\) operator in (1). To see this, consider setting \(L_1 = \ldots = L_K = 0, L_0 = 1\). We claim that \(g_{\text{MO}}(\bar{V}_{1:T}) = -(1/2K)\sum_{k=1}^{K} \max\{\rho_k - \bar{V}_{1:T,k}, 0\}^2\). Indeed,

\[
g_{\text{MO}}(\bar{V}_{1:T}) = \frac{1}{2K} \min_{u \in \Pi_{k=1}^{K} (\rho_k, \infty)} \left\{ \sum_{k=1}^{K} (\bar{V}_{1:T,k} - u_k)^2 \right\} = \frac{1}{2K} \sum_{k=1}^{K} \min_{u_k \in (\rho_k, \infty)} \left\{ (\bar{V}_{1:T,k} - u_k)^2 \right\}.
\]

For the \(k\)th summand, if \(\bar{V}_{1:T,k} \geq \rho_k\), the argmin is \(\bar{V}_{1:T,k}\) and the summand = 0. Else, we have \(\bar{V}_{1:T,k} < \rho_k\), the argmin is \(\rho_k\) and the summand = \((\bar{V}_{1:T,k} - \rho_k)^2\). Thus, the claim is proved.

Maximizing \(g_{\text{MO}}(\bar{V}_{1:T})\) is equivalent to minimizing \((1/2K)\sum_{k=1}^{K} \max(\rho_k - \bar{V}_{1:T,k}, 0)^2\). If the KPI \(\rho\) is achievable, then the optimal policy would generate \(\bar{V}_{1:T}\) for which \(\bar{V}_{1:T,k} \geq \rho_k\) for all \(k\), yielding objective value \(g_{\text{MO}}(\bar{V}_{1:T}) = 0\).

Otherwise, the shortfall of \(\bar{V}_{1:T}\) compared to \(\rho\) is minimized in the mean squared error sense.

(2.2): Any maximizer \(\bar{V}_{1:T}\) of \(g_{\text{MO}}(\bar{V}_{1:T})\) is Pareto-optimal. To see this, first observe that the \(\bar{V}_{1:T}\) generated by any policy satisfies \(\bar{V}_{1:T} \in [0,1]^K\), since \(V(s,a) \in [0,1]\) always. Suppose the contrary that there is a \(\bar{V}_{1:T}\), where \(\bar{V}_{1:T,k} \geq \bar{V}_{1:T,k} \forall k, \text{ and } \bar{V}_{1:T,1} > \bar{V}_{1:T,1}\). These mean that \(0 \leq 1 - \bar{V}_{1:T,k} \leq 1 - \bar{V}_{1:T,1} \forall k\), and \(0 \leq 1 - \bar{V}_{1:T,1} < 1 - \bar{V}_{1:T,1}\). Consequently, \(g_{\text{MO}}(\bar{V}_{1:T}) > g_{\text{MO}}(\bar{V}_{1:T})\), contradicting the maximality of \(\bar{V}_{1:T}\) on \(g_{\text{MO}}\). Thus, \(\bar{V}_{1:T}\) is Pareto-optimal. Altogether, \(g_{\text{MO}}\) with suitably chosen \(\rho, U\) captures Pareto-optimality.

Moreover, \(g_{\text{MO}}\) captures the State Space Exploration problem, which goes beyond Pareto-optimality.

(2.3): Capturing Pareto-optimality allows us to model many real world problems. Our framework allows any smooth concave \(g\) and not just \(g_{\text{MO}}\) (App. D), which captures other applications such as Maximum Entropy Exploration [23].

The design and analysis of GTP: (5.2), (2.1), (2.4). We start by addressing (5.2). For instance (1b), we claim that \(\text{opt}(P_{\mathcal{M}}) = 0\). In addition, the solution \(x^*\), defined as \(x^*(s^1, r_1) = x^*(s^2, r_1) = 1/2\) and \(x^*(s, a) = 0\) for all other \(s, a\), is optimal to \(P_{\mathcal{M}}\). Indeed, \(x^*\) is feasible to \(P_{\mathcal{M}}\) (recall \(p(s^1 | s^1, r_1) = p(s^2 | s^2, r_1) = 1\)), and that \(\sum_{s,a} v(s,a) x^*(s,a) = \frac{1}{2} + \frac{1}{2} = 1\), and that \(g_{\text{MO}}(\sum_{s,a} v(s,a) x^*(s,a)) = 0\).

The bad policy in Line 139, which causes \(\bar{V}_{1:T} \approx (1/6, 1/6)^T\), incurs \(\text{Reg}(T) = 0 - (-1/6 - 1/2)^2 = \Omega(1)\). The \(\Omega(1)\) regret is caused by the \(\Theta(T)\) implicit switching cost, where the agent switches between \(s^1, s^2\) (hence visits \(s^1\) for \(\Theta(T)\) times in \(T\) time steps. In an MO-OMDP instance, the implicit switching cost occurs when the agent switches from a prior class to another, and visits a state that does not contribute to the objective (like \(s^1\)) during the switch.

(2.1): GTP (see Lines 175-177) consists of the maintenance of distance measure \(\Psi\) in Line 13 in Alg 1, and the first criterion \(\Psi < Q\) in Line 9 in Alg 1. GTP keeps the implicit switching cost bounded, while balances the contributions by \(\{\bar{V}_{1:T,k}\}_{k=1}^{K}\). As said in Lines 188-193 for Fig 1b, GTP ensures the agent only switches between \(s^1, s^2\) for \(O(\sqrt{T})\) times in \(T\) steps, and \(\bar{V}_{1:T,k} - 0.5 = O(1/\sqrt{T})\) for \(k = 1, 2\), thus \(\text{Reg}(T) = O(1/\sqrt{T})\). GTP reduces the implicit switching cost from \(\Theta(T)\) to \(O(\sqrt{T})\), by looping at each \(s^1, s^2\) for \(\Theta(\sqrt{T})\) times before switching (cf. Lines 190-191).

(2.4): Lemma 4.1 follows from concentration inequalities, which are not our contributions. GTP is new, and its design and analysis are our novel contributions. Compared to UCRL2 for SO-OMDPs, analysing TFW-UCRL2 for MO-OMDPs requires crucial effort on bounding two costs: (i) the implicit switching cost (see Lemma 4.3) due to GTP. (ii) there is a delay cost caused by GTP on the gradient updates. In Fig 1b, the delay cost is the \(O(1/\sqrt{T})\) error on \(\bar{V}_{1:T,k} - 0.5\). The delay cost, included by eqn. (11), is discussed in Lines 244-248 and bounded by Proposition 4.2. These switch and delay costs are not present in UCRL2, and their analyses certainly do not follow from the literature.

Reviewer # 8: In fact, the regret under \(Q = L/\sqrt{K}\) (where \(L = L_0 + \max_k |L_k|\)) is quite close to the optimal regret by tuning \(Q\). We chose \(Q = L/\sqrt{K}\) to optimize the dependence on \(L, K\) in the regret order bound in Theorem 3.1. The regret could be improved by tuning \(Q\) online, or by optimizing \(Q\) in the actual regret bound. \(\rho, L_0, \ldots, L_K\) parameterize the objective function \(g_{\text{MO}}\), which is assumed to be fixed, while \(Q\) parameterizes the algo, so we only consider tuning \(Q\). If accepted, we will conduct the suggested empirical comparisons with [26, 28, 34] in their settings.