Context Dependent Misalignment

Dor Fuchs

January 2026

1 Setup

1.1 Tasks (CMDPs)

We model the environment as a finite **sequence** of encountered constrained MDP tasks

$$\{M_k\}_{k=1}^{N}, \tag{1}$$

where tasks in the sequence **need not be distinct**. Each task is

$$M_k = (\mathcal{S}_k, \mathcal{A}_k, P_k, \hat{R}_k, \mathcal{C}_k, \mu_k), \tag{2}$$

where:

- $\mathcal{S}_k$ is the state space and $\mathcal{A}_k$ is the action space;
- $P_k(\cdot \mid s, a)$ is a transition kernel on $\mathcal{S}_k$;
- μ_k is an initial-state distribution on $\mathcal{S}_k$;
- $\hat{R}_k$ is the **proxy** reward specification used by the agent's training objective;
- $\mathcal{C}_k$ is a **finite** set of constraint functions

  $$g : \mathcal{S}_k \times \mathcal{A}_k \to \mathbb{R}, \tag{3}$$

  which encode alignment constraints.

We use the standard notation x^+ to denote the positive part of a scalar:

$$x^+ = \max\{x, 0\}. \tag{4}$$

We interpret $g(s, a)^+$ as the magnitude of violation of constraint g at (s, a).

1.2 Policies and induced trajectories (randomized history-dependent)

Fix a task M_k. Let $h_t = (s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t)$ denote the history up to decision epoch t. A **randomized history-dependent** decision rule at time t is a map

$$d_{k,t} : \{(\mathcal{S}_k \times \mathcal{A}_k)^{t-1} \times \mathcal{S}_k\} \to \mathcal{P}(\mathcal{A}_k), \tag{5}$$

where $\mathcal{P}(\mathcal{A}_k)$ denotes the set of probability measures on $\mathcal{A}_k$. A policy is a sequence $\pi = (d_{k,1}, d_{k,2}, \ldots)$, and we let Π_k denote the class of all such policies.

Given $\pi \in \Pi_k$, a trajectory $\tau = \{(s_t, a_t)\}_{t=1}^{\infty}$ is generated by

$$s_1 \sim \mu_k, \quad a_t \sim d_{k,t}(\cdot \mid h_t), \quad s_{t+1} \sim P_k(\cdot \mid s_t, a_t). \tag{6}$$

We write $\tau \sim (\pi, M_k)$ for the induced trajectory distribution.

1.3 Cumulative violation and ε-alignment

Fix a task M_k. For any policy $\pi \in \Pi_k$, define the **expected total cumulative constraint violation** as

$$\Delta_k(\pi) \;=\; \mathbb{E}_{\tau \sim (\pi, M_k)} \left[\sum_{t=1}^{\infty} \sum_{g \in \mathcal{C}_k} g(s_t, a_t)^{+} \right], \tag{7}$$

where $\sum_{t=1}^{\infty}$ is understood as the limit of partial sums (possibly $+\infty$).

Fix $\varepsilon > 0$. We define a policy π as ε-aligned for task M_k if

$$\Delta_k(\pi) < \varepsilon, \tag{8}$$

and ε-misaligned otherwise:

$$\Delta_k(\pi) \geq \varepsilon. \tag{9}$$

This induces two policy sets for each task M_k:

$$\Pi_k^{a,\varepsilon} = \{\pi \in \Pi_k : \Delta_k(\pi) < \varepsilon\}, \quad \Pi_k^{m,\varepsilon} = \{\pi \in \Pi_k : \Delta_k(\pi) \geq \varepsilon\}. \tag{10}$$

1.4 Temptation

Fix a task M_k. For concreteness, take the proxy-return criterion to be the infinite-horizon discounted value under the proxy reward:

$$V^{\hat{R}_k}(\pi) \;=\; \mathbb{E}_{\tau \sim (\pi, M_k)} \left[\sum_{t=1}^{\infty} \gamma_k^{t-1}\, \hat{r}_k(s_t, a_t) \right], \tag{11}$$

where $\hat{R}_k$ specifies $(\hat{r}_k, \gamma_k)$ with $\gamma_k \in (0, 1)$.

For any $\varepsilon > 0$, define the ε-temptation gap on task M_k as

$$d_k^{\varepsilon} = \sup_{\pi' \in \Pi_k^{m,\varepsilon}} V^{\hat{R}_k}(\pi') \;-\; \sup_{\pi \in \Pi_k^{a,\varepsilon}} V^{\hat{R}_k}(\pi). \tag{12}$$

We adopt the convention $\sup \emptyset = -\infty$ (extended reals), so d_k^ε is always defined (possibly $\pm\infty$).

We define the empirical history summary of temptations up to experience k as

$$D_0 = 0, \quad D_k = \frac{1}{k}\sum_{j=1}^{k} d_j^\varepsilon. \tag{13}$$

1.5 Experience indexing

We index by **experience** rather than within-task time. Experience k means: the agent has encountered the first k tasks in the sequence $\{M_j\}_{j=1}^{N}$.

1.6 Compliance state (analyst-side latent variable)

We introduce a compliance coefficient

$$c_k \in \mathbb{R}_{\geq 0} \tag{14}$$

as an **analyst-side latent state** indexed by experience. We interpret $c_k = 0$ as full compliance, and larger c_k as a greater latent tendency toward violating constraints.

We define c_k by the recursion

$$c_0 = 0, \quad c_k = (\alpha c_{k-1} + \beta D_{k-1} + \delta_k)^+. \tag{15}$$

1.7 Observable violations (measurement equation)

The quantity $\Delta_k(\pi_k)$ is a population expectation. What is observed in practice is a **noisy empirical statistic** (e.g., from finite rollouts / truncation), which we denote by $\widehat{\Delta}_k(\pi_k)$.

We connect the latent compliance state to observable behavior through a measurement model with contemporaneous temptation:

$$\widehat{\Delta}_k(\pi_k) = (\lambda c_k + \theta d_k^\varepsilon + \eta_k)^+, \tag{16}$$

where $\lambda \geq 0$ and θ are scaling coefficients and η_k is measurement noise.

1.8 Intuition for the pieces

- $\Delta_k(\pi)$ is the expected total cumulative constraint violation of policy π on task M_k.
- $\Pi_k^{a,\varepsilon}$ and $\Pi_k^{m,\varepsilon}$ split policies into ε-aligned versus ε-misaligned.
- d_k^ε measures how much better, in proxy-return terms, the best ε-misaligned behavior can be than the best ε-aligned behavior on the same task.

- c_k is an analyst-side latent compliance state whose dynamics depend on past temptation summaries D_{k-1} and an innovation term δ_k.
- The measurement equation treats observed violations $\widehat{\Delta}_k(\pi_k)$ as a noisy statistic depending on both latent tendency c_k and task-specific temptation d_k^{ε}.