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

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

where:

- $\mathcal{S}_k$ is a **finite** state space,
- $\mathcal{A}_k$ is a **finite** action space,
- $P_k(\cdot \mid s, a)$ is a transition kernel on $\mathcal{S}_k$,
- $\mathcal{R}_k \subseteq [0, 1]$ is the set of reward values,
- r_k is the reward function, $r_k : \mathcal{S}_k \times \mathcal{A}_k \times \mathcal{S}_k \to \mathcal{R}_k$,
- $\mathcal{C}_k$ is a **finite** set of constraint functions $g : \mathcal{S}_k \times \mathcal{A}_k \to \mathbb{R}$,
- μ_k is an initial-state distribution on $\mathcal{S}_k$,
- T is to be explained in the next section.

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

$$x^+ = \max\{x, 0\}. \tag{3}$$

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. We define the natural filtration $\mathcal{F}_t = \sigma(h_t)$, and assume T is a stopping time w.r.t. $(\mathcal{F}_t)$.

A **randomized history-dependent** decision rule at time t is a map

$$d_{k,t} : \{(\mathcal{S}_k \times \mathcal{A}_k)^{t-1} \times \mathcal{S}_k\} \to \mathcal{P}(\mathcal{A}_k), \tag{4}$$

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 **random** trajectory $\tau = (s_1, a_1, s_2, a_2, \ldots, s_T)$ is generated by

$$s_1 \sim \mu_k, \quad \forall t < T : \; a_t \sim d_{k,t}(\cdot \mid h_t), \quad s_{t+1} \sim P_k(\cdot \mid s_t, a_t), \tag{5}$$

where no action is taken at time T. We write $\tau \sim (\pi, M_k)$ for the induced trajectory distribution.

1.3 Well-posedness (finite return and finite violation magnitude)

We assume each task is well-posed in the sense that the discounted return and cumulative violation magnitude are integrable under any deployed policy class of interest. Concretely, we assume for each k:

$$\mathbb{E}[T] < \infty \qquad \text{and} \qquad \sup_{\pi \in \Pi_k} \mathbb{E}_{\tau \sim (\pi, M_k)} \left[\sum_{t=1}^{T-1} \sum_{g \in \mathcal{C}_k} g(s_t, a_t)^+ \right] < \infty. \tag{6}$$

(For empirical benchmarks one may enforce a hard horizon $T \leq H$ almost surely.)

1.4 Pathwise-perfect alignment and violation probability

Fix a task M_k and a policy $\pi \in \Pi_k$. Define the **event of any constraint violation** by

$$\mathsf{Viol}_k(\pi) \;=\; \{\exists\, t \in \{1, \ldots, T-1\} \; \exists\, g \in \mathcal{C}_k \text{ such that } g(s_t, a_t) > 0\}\,. \tag{7}$$

Equivalently, $\mathsf{Viol}_k(\pi) = \left\{ \sum_{t=1}^{T-1} \sum_{g \in \mathcal{C}_k} g(s_t, a_t)^+ > 0 \right\}$.

We define the **violation probability** of π on task M_k as

$$p_k(\pi) \;=\; \mathbb{P}_{\tau \sim (\pi, M_k)}\big(\mathsf{Viol}_k(\pi)\big). \tag{8}$$

1.5 Aligned and misaligned policy classes (binary, pathwise)

We adopt the stance that **aligned means pathwise-perfect**: a policy is aligned iff it violates no constraint along the trajectory, almost surely.

Accordingly, define the **aligned** and **misaligned** policy classes:

$$\Pi_k^0 \;=\; \{\pi \in \Pi_k : p_k(\pi) = 0\}, \qquad \Pi_k^{>0} \;=\; \{\pi \in \Pi_k : p_k(\pi) > 0\}. \tag{9}$$

1.6 Standing assumptions (alignability and nontriviality)

We impose the following assumptions for all tasks in the sequence:

$$\forall k \in \{1, \ldots, N\} : \qquad \Pi_k^0 \neq \emptyset \quad \text{and} \quad \Pi_k^{>0} \neq \emptyset. \tag{10}$$

The first condition states that the alignment constraints are **attainable** (there exists at least one pathwise-perfect aligned policy), so the aligned baseline is well-defined. The second excludes degenerate tasks in which no policy can ever violate constraints.

1.7 Return

Fix a task M_k. Define the discounted return for any policy $\pi \in \Pi_k$ as

$$V_k(\pi) \;=\; \mathbb{E}_{\tau \sim (\pi, M_k)}\left[\sum_{t=1}^{T-1} \gamma_k^{t-1}\, r_k(s_t, a_t, s_{t+1})\right], \tag{11}$$

where $\gamma_k \in (0, 1)$ is the task discount factor.

1.8 Global temptation (policy-level)

Define the (nonnegative) **global temptation gap** on task M_k as

$$D_k \;=\; \left(\sup_{\pi' \in \Pi_k^{>0}} V_k(\pi') \;-\; \sup_{\pi \in \Pi_k^0} V_k(\pi)\right)^+. \tag{12}$$

In many benchmark-style tasks with a saturated score ceiling, it is possible for both aligned and misaligned strategies to achieve the maximal return, in which case D_k may be 0 despite the presence of locally tempting misaligned actions.

1.9 Stepwise temptation (action-level, return-to-go)

Fix a task M_k and a history h_t with current state s_t. Define the aligned and misaligned action sets at h_t by

$$\mathcal{A}_k^0(h_t) = \{a \in \mathcal{A}_k : \forall g \in \mathcal{C}_k,\ g(s_t, a) \leq 0\}, \qquad \mathcal{A}_k^{>0}(h_t) = \{a \in \mathcal{A}_k : \exists g \in \mathcal{C}_k,\ g(s_t, a) > 0\}. \tag{13}$$

For empirical work we fix a **reference evaluation protocol** that induces a value proxy (e.g. a reference policy class, a compute budget, and an estimator). Let $Q_k^{\mathrm{ref}}(h_t, a)$ denote a chosen **return-to-go proxy** at (h_t, a), interpreted as the expected discounted future return obtained by taking action a at history h_t and then following the reference protocol thereafter.

Define the **stepwise temptation gap** at history h_t by

$$\Delta_k(h_t) \;=\; \left(\max_{a \in \mathcal{A}_k^{>0}(h_t)} Q_k^{\mathrm{ref}}(h_t, a) \;-\; \max_{a \in \mathcal{A}_k^{0}(h_t)} Q_k^{\mathrm{ref}}(h_t, a) \right)^{+}. \tag{14}$$

(If $\mathcal{A}_k^0(h_t) = \emptyset$ at a history reached under the chosen protocol, the task is not alignable along that trajectory; in empirical benchmarks one may enforce that $\mathcal{A}_k^0(h_t) \neq \emptyset$ for all encountered histories.)

1.10 Prefix-normalized temptation (task-level)

Given a trajectory $\tau = (s_1, a_1, \ldots, s_T)$ with histories $\{h_t\}_{t=1}^{T-1}$, define the **prefix-average temptation** up to step m by

$$\overline{\Delta}_{k,m}(\tau) \;=\; \frac{1}{m} \sum_{t=1}^{m} \Delta_k(h_t), \qquad m \in \{1, \ldots, T-1\}, \tag{15}$$

and the **prefix-normalized temptation** along τ by

$$D_k^{\mathrm{pref}}(\tau) \;=\; \max_{m \in \{1,\ldots,T-1\}} \overline{\Delta}_{k,m}(\tau). \tag{16}$$

Finally define the **task-level prefix temptation** as the expectation under the fixed reference protocol:

$$D_k^{\mathrm{pref}} \;=\; \mathbb{E}_{\tau \sim (\pi_k^{\mathrm{ref}}, M_k)} \Big[D_k^{\mathrm{pref}}(\tau) \Big], \tag{17}$$

where π_k^{ref} denotes the (fixed) trajectory-sampling policy induced by the reference protocol.

1.11 Operational temptation (computable estimator)

Empirically we estimate $Q_k^{\mathrm{ref}}(h_t, a)$ with an approximation $\widehat{Q}_k(h_t, a)$ (e.g. Monte Carlo rollouts, fitted Q evaluation, or a learned critic under a fixed compute budget), yielding an estimator $\widehat{\Delta}_k(h_t)$ and hence

$$\widehat{D}_k^{\mathrm{pref}}(\tau) \;=\; \max_{m \in \{1,\ldots,T-1\}} \frac{1}{m} \sum_{t=1}^{m} \widehat{\Delta}_k(h_t), \qquad \widehat{D}_k^{\mathrm{pref}} \;=\; \mathbb{E}\Big[\widehat{D}_k^{\mathrm{pref}}(\tau) \Big], \tag{18}$$

where the outer expectation is approximated via repeated trajectory sampling under the reference protocol.

1.12 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 $\{M_j\}_{j=1}^{N}$.

1.13 Compliance coefficient (mean violation magnitude)

Let $\pi_k \in \Pi_k$ denote the agent's deployed policy on task M_k. We define the compliance coefficient at experience k as the **expected cumulative violation magnitude**:

$$C_k \;=\; \mathbb{E}_{\tau\sim(\pi_k,M_k)}\left[\sum_{t=1}^{T-1}\sum_{g\in\mathcal{C}_k} g(s_t,a_t)^+\right]. \tag{19}$$

Note: the alignment classification is *binary* via $p_k(\pi)$, while C_k measures *how much* violation occurs under the deployed policy.

1.14 Modeling assumption (statistical ARX(1) hypothesis)

We impose the modeling assumption that violation magnitude evolves as an auto-regressive process with exogenous regressors of order 1 ($ARX(1)$):

$$\forall k \in \{2,\dots,N\}.\quad C_k \;=\; \alpha\, C_{k-1} \;+\; \beta\, D_k^{\text{pref}} \;+\; \varepsilon_k. \tag{20}$$

For empirical estimation we replace D_k^{pref} by its computable approximation $\widehat{D}_k^{\text{pref}}$.

Innovation assumptions. We model $\{\varepsilon_k\}$ as heteroscedastic innovations satisfying

$$\mathbb{E}\Big[\varepsilon_k \,\Big|\, C_{k-1}, D_k^{\text{pref}}\Big] = 0, \qquad \mathrm{Var}\Big(\varepsilon_k \,\Big|\, C_{k-1}, D_k^{\text{pref}}\Big) = \sigma_k^2, \tag{21}$$

with no requirement that σ_k^2 is constant across tasks. Optionally, one may additionally assume

$$\mathbb{E}\Big[\varepsilon_k\varepsilon_{k-\ell} \,\Big|\, C_{k-1}, D_k^{\text{pref}}\Big] = 0 \quad \text{for all } \ell \geq 1, \tag{22}$$

and test this assumption empirically via residual autocorrelation diagnostics. Inference may use heteroscedasticity-robust or HAC standard errors.

1.15 Predetermined protocol (exogeneity of temptation)

For empirical tests that interpret β as the effect of temptation on compliance, we assume the task sequence $\{M_k\}_{k=1}^{N}$ and the reference protocol used to define Q_k^{ref} (and hence D_k^{pref}) are fixed **prior** to observing the innovations $\{\varepsilon_k\}$. Equivalently, D_k^{pref} is treated as predetermined by the task specification and the fixed evaluation protocol.

1.16 Intuition for the pieces

- $p_k(\pi)$ is the probability that policy π violates *any* alignment constraint on task M_k (binary notion of misalignment).
- Π_k^0 contains pathwise-perfect aligned policies ($p_k(\pi) = 0$); $\Pi_k^{>0}$ contains policies that violate with positive probability ($p_k(\pi) > 0$).
- D_k is a global (policy-level) temptation gap, which may be 0 in tasks with a saturated score ceiling.
- $\Delta_k(h_t)$ is a stepwise temptation gap comparing the best misaligned vs best aligned action at history h_t under a return-to-go proxy Q_k^{ref}.
- D_k^{pref} aggregates stepwise temptation as the maximum over prefix-average temptation, normalized by step count.
- C_k is the expected cumulative violation magnitude of the deployed policy π_k on task M_k, modeled as $ARX(1)$ with temptation input.