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

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

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 $\mathcal{S}_k$.

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, stopping time, and induced trajectories (randomized history-dependent)

Fix a task M_k. Let

$$h_t = (s_1, a_1, \dots, s_{t-1}, a_{t-1}, s_t) \tag{4}$$

denote the history up to decision epoch t. Define the natural filtration $\mathcal{F}_t = \sigma(h_t)$, and let T be a stopping time with respect to $(\mathcal{F}_t)$.

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

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

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

Given $\pi \in \Pi_k$, a **random** trajectory

$$\tau = (s_1, a_1, s_2, a_2, \dots, s_T) \tag{6}$$

is generated by

$$s_1 \sim \mu_k, \qquad \forall t < T: \ a_t \sim d_{k,t}(\cdot \mid h_t), \qquad s_{t+1} \sim P_k(\cdot \mid s_t, a_t), \tag{7}$$

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)

Each task induces a (measurable) per-trajectory performance functional $R_k(\tau) \in \mathbb{R}$ ("return"). A standard choice is the discounted or undiscounted sum of rewards along τ, but we do not assume a specific form here.

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

$$\mathbb{E}[T] < \infty, \tag{8}$$

$$\sup_{\pi \in \Pi_k} \mathbb{E}_{\tau \sim (\pi, M_k)} \big[|R_k(\tau)| \big] < \infty, \tag{9}$$

$$\sup_{\pi \in \Pi_k} \mathbb{E}_{\tau \sim (\pi, M_k)} \Big[\sum_{t=1}^{T-1} \sum_{g \in \mathcal{C}_k} g(s_t, a_t)^+ \Big] < \infty. \tag{10}$$

(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** (as a subset of trajectories) by

$$\mathsf{Viol}_k(\tau) = \{\exists\, t \in \{1, \ldots, T-1\}\ \exists\, g \in \mathcal{C}_k \text{ such that } g(s_t, a_t) > 0\}\,. \tag{11}$$

Equivalently,

$$\mathsf{Viol}_k(\tau) = \left\{ \sum_{t=1}^{T-1} \sum_{g \in \mathcal{C}_k} g(s_t, a_t)^+ > 0 \right\}. \tag{12}$$

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

$$p_k(\pi) = \mathbb{P}_{\tau \sim (\pi, M_k)}\big(\mathsf{Viol}_k(\tau)\big). \tag{13}$$

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:

$$\Pi_k^0 = \{\pi \in \Pi_k : p_k(\pi) = 0\}, \qquad \Pi_k^{>0} = \{\pi \in \Pi_k : p_k(\pi) > 0\}. \tag{14}$$

1.6 Standing assumptions (alignability and nontriviality)

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

$$\forall k \in \{1, \ldots, N\} : \quad \Pi_k^0 \neq \emptyset \quad \text{and} \quad \Pi_k^{>0} \neq \emptyset. \tag{15}$$

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 and value

Fix a task M_k. The task induces a performance functional $R_k(\tau)$. Define the value of a policy $\pi \in \Pi_k$ by

$$V_k(\pi) = \mathbb{E}_{\tau \sim (\pi, M_k)}\big[R_k(\tau)\big]. \tag{16}$$

No additional structure (discounting, stationarity, etc.) is assumed unless stated later for a specific benchmark.

1.8 Stepwise temptation (action-level, protocol-defined return-to-go proxy)

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

$$\mathcal{A}_k^0(h_t) = \{a \in \mathcal{A}_k : \forall g \in \mathcal{C}_k,\ g(s_t, a) \le 0\}, \tag{17}$$

$$\mathcal{A}_k^{>0}(h_t) = \{a \in \mathcal{A}_k : \exists g \in \mathcal{C}_k,\ g(s_t, a) > 0\}. \tag{18}$$

For empirical work we fix a **reference evaluation protocol** that induces: (i) a trajectory-sampling policy π_k^{ref}, (ii) a compute budget, and (iii) an estimator for a return-to-go proxy. Let $Q_k^{\mathrm{ref}}(h_t, a)$ denote a chosen **return-to-go proxy** at (h_t, a), interpreted as the expected remaining 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

$$\delta_k(h_t) = \Big(\max_{a \in \mathcal{A}_k^{>0}(h_t)} Q_k^{\mathrm{ref}}(h_t, a) - \max_{a \in \mathcal{A}_k^0(h_t)} Q_k^{\mathrm{ref}}(h_t, a) \Big)^+. \tag{19}$$

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

1.9 Task-level temptation (maximum stepwise temptation along a trajectory)

Define the trajectory-level maximum temptation as

$$D_k^{\mathrm{ref}}(\tau) = \max_{t \in \{1,\dots,T-1\}} \delta_k(h_t). \tag{20}$$

Define the task-level temptation induced by the fixed reference protocol as

$$D_k^{\mathrm{ref}} = \mathbb{E}_{\tau \sim (\pi_k^{\mathrm{ref}}, M_k)} \big[D_k^{\mathrm{ref}}(\tau) \big]. \tag{21}$$

1.10 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). Define

$$\widehat{\delta}_k(h_t) = \Big(\max_{a \in \mathcal{A}_k^{>0}(h_t)} \widehat{Q}_k(h_t, a) - \max_{a \in \mathcal{A}_k^0(h_t)} \widehat{Q}_k(h_t, a) \Big)^+, \tag{22}$$

$$\widehat{D}_k^{\mathrm{ref}}(\tau) = \max_{t \in \{1,\dots,T-1\}} \widehat{\delta}_k(h_t). \tag{23}$$

The protocol-level quantity D_k^{ref} is approximated by repeated trajectory sampling under π_k^{ref} and averaging $\widehat{D}_k^{\mathrm{ref}}(\tau)$.

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

1.12 Compliance coefficient (mean violation magnitude)

Let $\pi_k \in \Pi_k$ denote the agent's deployed policy on task M_k. Define the per-step violation magnitude

$$v_{k,t} = \sum_{g \in \mathcal{C}_k} g(s_t, a_t)^+, \tag{24}$$

and define the compliance coefficient at experience k as the **expected cumulative violation magnitude**:

$$C_k = \mathbb{E}_{\tau \sim (\pi_k, M_k)} \Big[\sum_{t=1}^{T-1} v_{k,t} \Big]. \tag{25}$$

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

1.13 Modeling assumption (statistical ARX(1) hypothesis)

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

$$\forall k \in \{2, \dots, N\}: \quad C_k = \alpha C_{k-1} + \beta D_k^{\text{ref}} + \varepsilon_k. \tag{26}$$

For empirical estimation we replace D_k^{ref} by its computable approximation obtained from $\widehat{Q}_k$ and repeated trajectory sampling.

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

$$\mathbb{E}\big[\varepsilon_k \mid C_{k-1}, D_k^{\text{ref}}\big] = 0, \tag{27}$$

$$\text{Var}\big(\varepsilon_k \mid C_{k-1}, D_k^{\text{ref}}\big) = \sigma_k^2, \tag{28}$$

with no requirement that σ_k^2 is constant across tasks. Inference may use heteroscedasticity-robust or HAC standard errors.

1.14 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 (including π_k^{ref} and the estimator defining Q_k^{ref}) are fixed prior to observing the innovations $\{\varepsilon_k\}$. Equivalently, D_k^{ref} is treated as predetermined by the task specification and the fixed reference protocol.

1.15 Martingale structure of cumulative violations

Let $\mathcal{F}_t = \sigma(h_t)$ and define $v_{k,t}$ as above. Then the centered increments

$$m_{k,t} = v_{k,t} - \mathbb{E}[v_{k,t} \mid \mathcal{F}_{t-1}] \tag{29}$$

form a martingale difference sequence with respect to $(\mathcal{F}_t)$. Consequently,

$$M_{k,n} = \sum_{t=1}^{n} m_{k,t} \tag{30}$$

is a martingale. Under bounded-increment assumptions (e.g. bounded $v_{k,t}$ or bounded differences), one can use standard concentration tools (Azuma-Hoeffding / Freedman-type bounds) to control estimation error for cumulative violation statistics derived from finite rollouts. This is optional and only used when we build the inference layer.

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 ($p_k(\pi) > 0$).
- $\delta_k(h_t)$ is a stepwise temptation gap comparing the best misaligned vs best aligned action at history h_t under the protocol-defined proxy Q_k^{ref}.
- D_k^{ref} aggregates temptation as the expected maximum of $\delta_k(h_t)$ along trajectories sampled from the fixed reference protocol.
- 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.