NEURAL INFORMATION PROCESSING SYSTEMS

A Single Deep Preference-Conditioned Policy for Learning Pareto Coverage Sets

Akihiro Kubo[1,4], Kosuke Nakanishi[4,2], Shin Ishii[4,3,1]

[1] International Research Center for Neurointelligence, The University of Tokyo

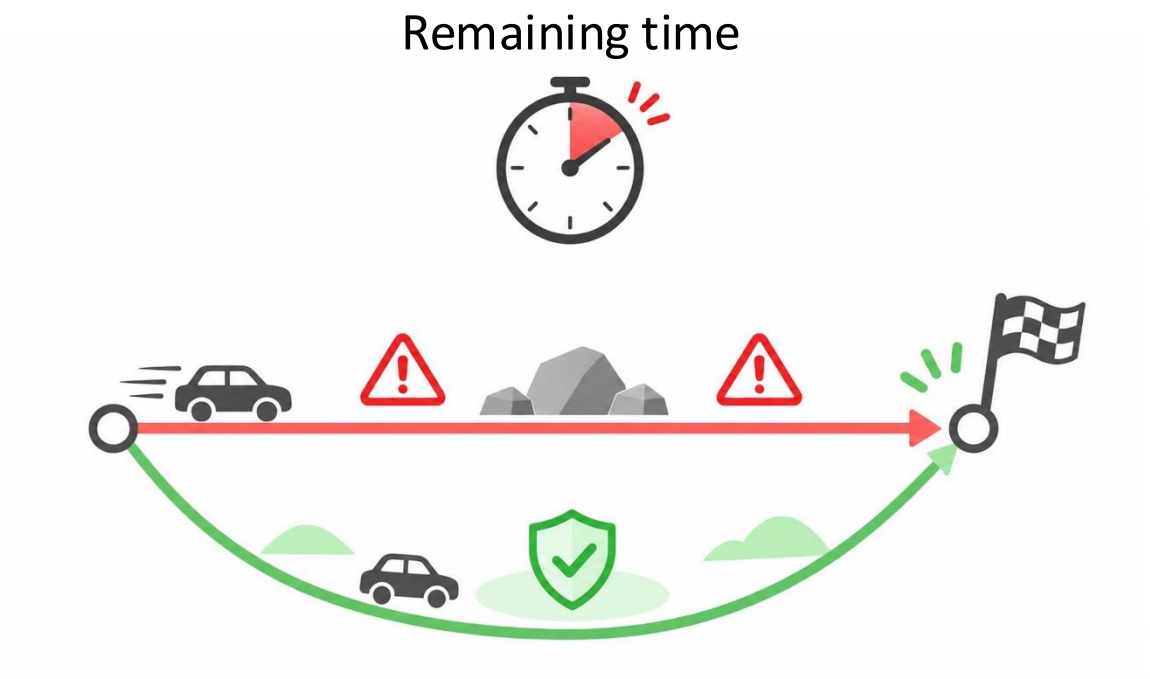
[2] Honda R&D Co., Ltd.

[3] Advanced Telecommunications Research Institute International

[4] Department of Information Science, Kyoto University

What counts as “best” when two objectives compete?

Two goals: Higher safety and more time to spare at arrival



Route A: Shorter

A higher accident risk

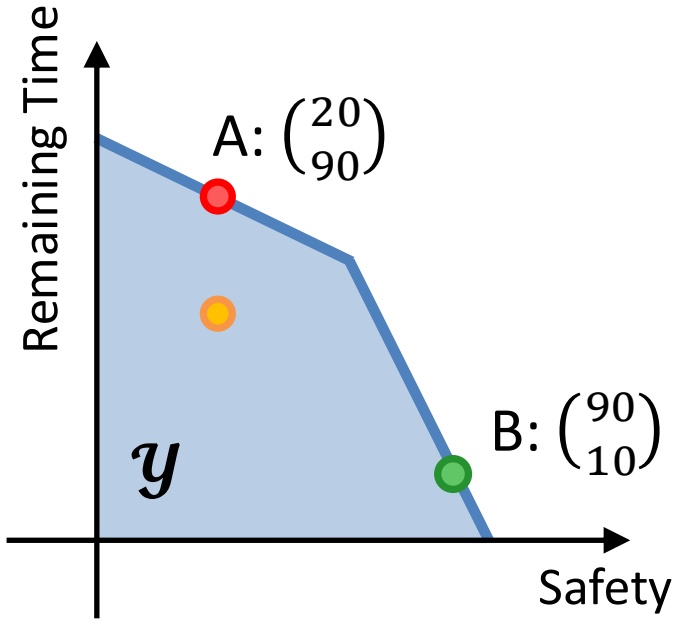
Route B: Safer

A longer detour

The Pareto front

- **Outcome (plotted point)**: The pair of scores a route gets on those objectives, plotted as one point.
- **Dominance**: An outcome $\boldsymbol{y}$ dominates $\boldsymbol{x}$ if $y_i \geq x_i$ for all i and $y_j \geq x_j$ for some j.
- **Pareto front:** The set of points for all undominated outcomes $\boldsymbol{\mathcal{Y}}$. Each is Pareto-optimal.



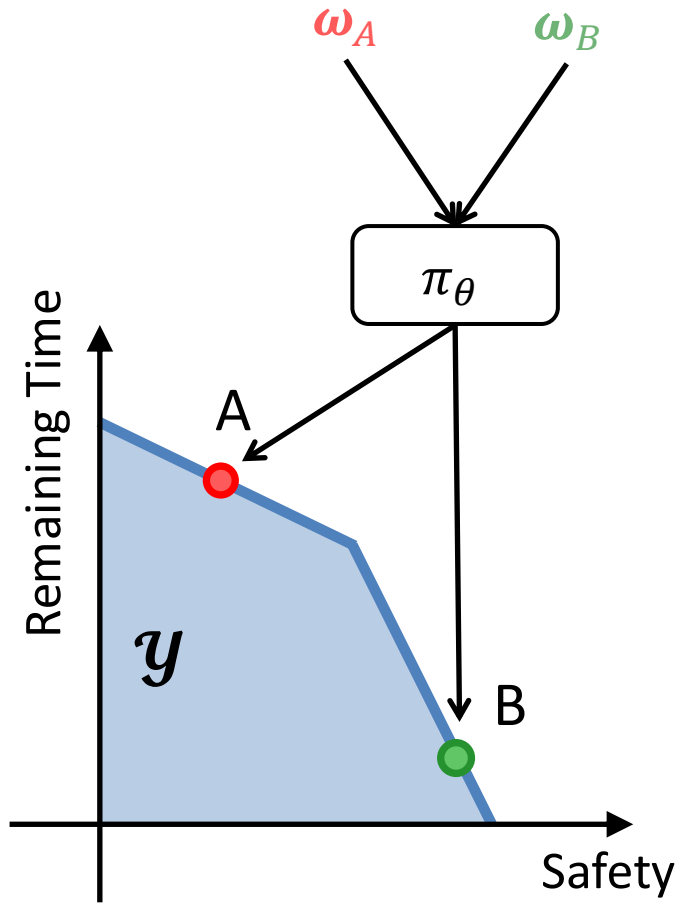


One model can offer many trade-offs

- **Preference vector ω**
  - One positive number per objective, with $\sum_i \omega_i = 1$. The numbers ω_i express relative priorities.
  - $\omega = (0.7, 0.3)$ assigns 0.7 to safety and 0.3 to time to spare.
- **Return vector in multi-objective reinforcement learning (MORL)**

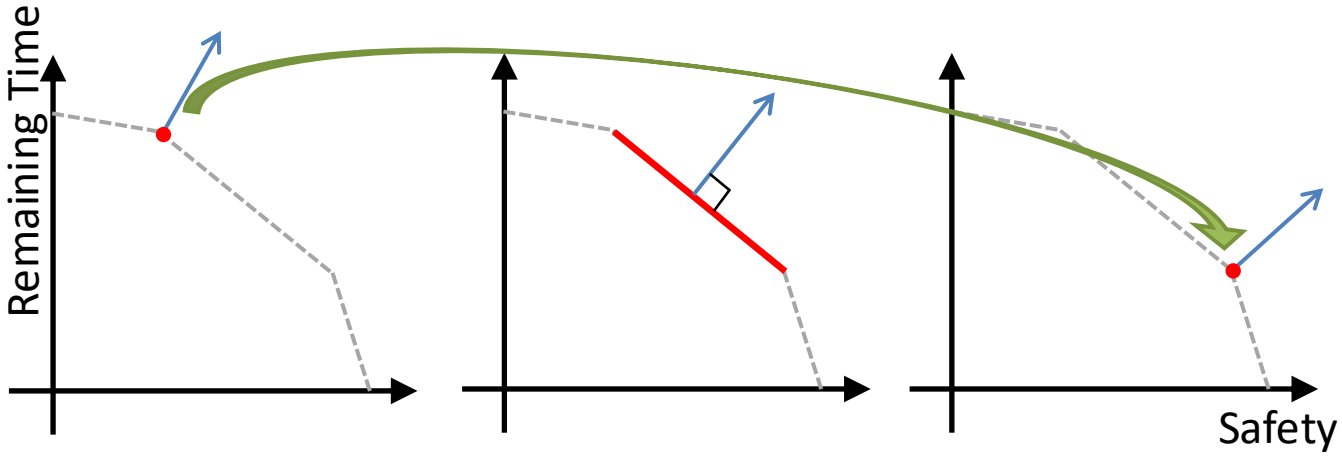
  We summarize a policy's outcome by $\mathbf{J}(\pi) = \mathbb{E}[\sum_t \gamma^t \mathbf{r}(s_t, a_t)]$ with one expected discounted sum of rewards per objective.
- **Preference-conditioned policy**

  $\pi_\theta(a \mid s, \omega)$ use shared parameters θ. Vary ω to aim to get every point on the Pareto front.



A small preference change can cause a big jump

- **Linear scalarization**
  - A weighted sum of the scores $\sum_i \omega_i J_i$
  - At a tie, a whole flat segment can be optimal. A tiny change in preference can switch the chosen point from one end to the other. [3, 14, 15]



- **Nonlinear scalarization**
  - A nonlinear utility of the return vector (scalarized expected return, SER)
  - General concave utilities require different optimization methods. The choice of utility determines whether the target return vector is unique and varies continuously. [25, 26], Paper §5

Research question

- Can we select target return vectors that vary continuously with preference and cover the Pareto front?
- Can we design policy updates that make the return vectors converge to those targets?

STCH gives us targets we can trace across the front

Smooth Tchebycheff scalarization (STCH) [19]

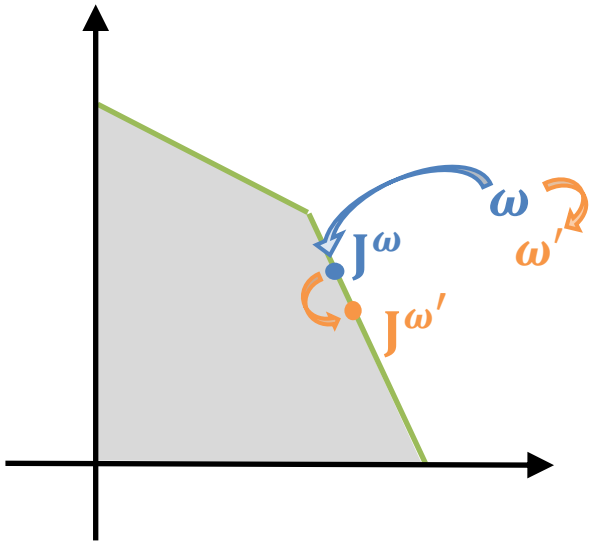
A smooth utility that emphasizes the largest weighted shortfalls from a utopia point $\boldsymbol{I}$.

$$u(\mathbf{f}, \omega) := -\tau \log \sum_{k=1}^{m} \exp\left(\frac{\omega_k (I_k - f_k)}{\tau}\right)$$

τ: smoothing parameter

The **target** $\mathbf{J}^{\omega}$ is the attainable return vector that maximizes the utility at preference ω

1. Unique and Pareto optimal [Prop. 2.1]
   A single-valued mapping to Pareto-optimal vectors
2. Full Pareto-front coverage [Prop. 2.2]
   Image set for all preference $\{\mathbf{J}^{\omega} | \forall_{\omega}\}$ corresponds to Pareto front
3. Lipschitz continuous in preference [**Thm. 2.3**]



CMDPI implements mirror descent through policy iteration

Concave Mirror Descent Policy Iteration (CMDPI)

1. Occupancy space

μ: discounted state–action visitation frequencies. $\bar{\mathbf{J}}(\mu) = \mathbf{J}$ is linear in μ and the STCH is concave in μ, so minimizing its negative over $\mathcal{M}$ is a convex problem. Apply relative-smooth mirror descent (MD). [Eqs. (9)–(10), Lemma 3.1]

$$\min_{\mu \in \mathcal{M}} f(\mu), \qquad f(\mu) = -u\big(\bar{\mathbf{J}}(\mu), \omega\big)$$

2. Lemma 2.4.

The Bregman divergence induced by negative conditional entropy equals the occupancy-weighted policy KL. This makes each MD step equivalent to solving a KL-regularized MDP with reference policy π_k. [Eqs. (5), (10)–(11)]

$$D_{\psi_\gamma}(\mu \mid \mu_k) = \Gamma_\gamma(\pi, \pi_k) \qquad \pi_{k+1} = \arg\max_{\pi \in \Pi} \mathbb{E}_\pi\left[\sum_{t \geq 0} \gamma^t r_k(s_t, a_t)\right] - L_{\text{rel}} \Gamma_\gamma(\pi, \pi_k)$$

3. Solve a KL-regularized MDP.

Form a scalar reward $r_k = \langle \mathbf{g}_k, \mathbf{r} \rangle$ with $\mathbf{g}_k = \nabla_{\mathbf{J}} \boldsymbol{u}(\mathbf{J}(\pi_k), \omega)$, solve a soft Bellman fixed point, and improve the policy. [Eqs. (12)–(14), Alg. 3]

$$\pi_{k+1}(a|s) \propto \pi_k(a|s) \exp\left(\frac{Q_{k+1}(s,a)}{\alpha_{\text{KL}}}\right)$$

Finite-step policies are continuous, and returns approach the target

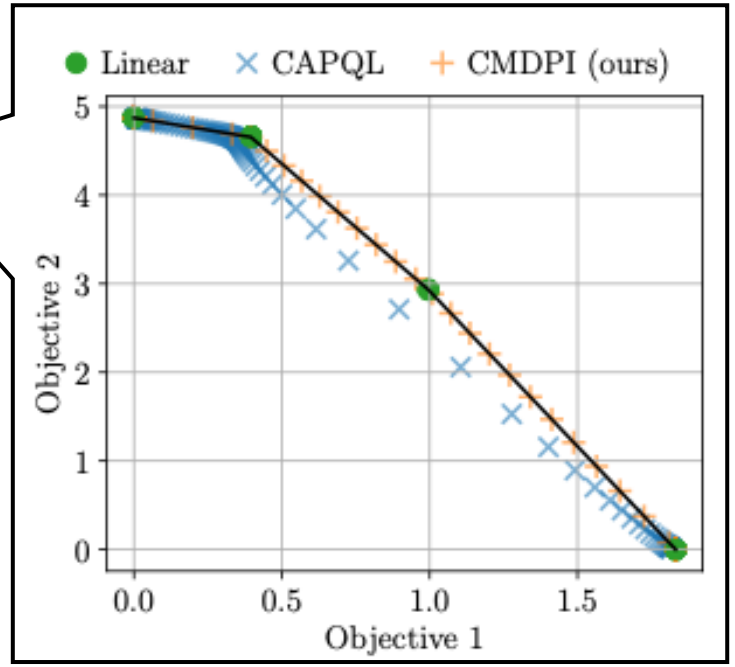
What we vary	What exact tabular CMDPI guarantees
Preference $\boldsymbol{\omega}$ at fixed finite iteration k	$\boldsymbol{\omega} \mapsto \pi_k^{\boldsymbol{\omega}}$ is continuous. Start from a preference-continuous policy with full action support. [Prop. G.1]
Iteration k at fixed preference $\boldsymbol{\omega}$	$\|\mathbf{J}(\pi_k^{\omega}) - \mathbf{J}^{\omega}\|_2 = O(k^{-1/2})$. Use full-support initialization, positive initial-state mass $p_0(s) > 0$ for every state, and a valid relative-smoothness constant L_{rel}. [**Thm. 3.2**]

Proof roadmap

Relative-smooth mirror descent gives an $O(1/k)$ utility gap. STCH quadratic growth converts it into an $O(k^{-1/2})$ distance in return space.

Planning shows denser coverage, and deep results vary by task

Evaluation	Main observation
Exact tabular planning: 4 states, 2 actions, 100 preferences.	CMDPI covers the Pareto front more densely and uniformly.
Discrete deep RL: 8 MO-Gymnasium tasks, 8 seeds, 2M steps per task.	Deep CMDPI uses previous-policy KL in the critic target and has the best mean HV rank.
Continuous deep RL: 8 MuJoCo tasks, model-based comparison at 500K real-environment steps.	PCSAC-MB, an STCH-based actor–critic surrogate, has higher mean HV than CAPQL-MB on 6 of 8 tasks.



HV= hypervolume

	LS	CAPQL	PreCo	C-MORL	PCSAC	CMDPI
HV	4.25 ± 0.36	4.12 ± 0.33	3.75 ± 0.54	4.12 ± 0.69	3.00 ± 0.72	$\mathbf{1.75 \pm 0.33}$
EUM	3.88 ± 0.53	3.50 ± 0.49	3.38 ± 0.57	3.62 ± 0.67	3.50 ± 0.74	$\mathbf{3.12 \pm 0.54}$
SP	$\mathbf{2.25 \pm 0.35}$	3.75 ± 0.44	$\mathbf{2.25 \pm 0.46}$	4.62 ± 0.69	4.38 ± 0.62	3.75 ± 0.47

Summary

- **STCH gives us targets**

  One Pareto-optimal return vector per positive preference, Lipschitz preference sensitivity, and full-front coverage under the stated assumptions. [Props. 2.1–2.2, **Thm. 2.3**]

- **CMDPI gives us updates**

  With exact tabular updates, policies vary continuously with preference at each finite step, and their return vectors converge to the STCH targets. [Prop. G.1, **Thm. 3.2**]

A theoretical step toward justifying a single preference-conditioned policy for Pareto-front coverage.

Extending the guarantees to deep models remains open.