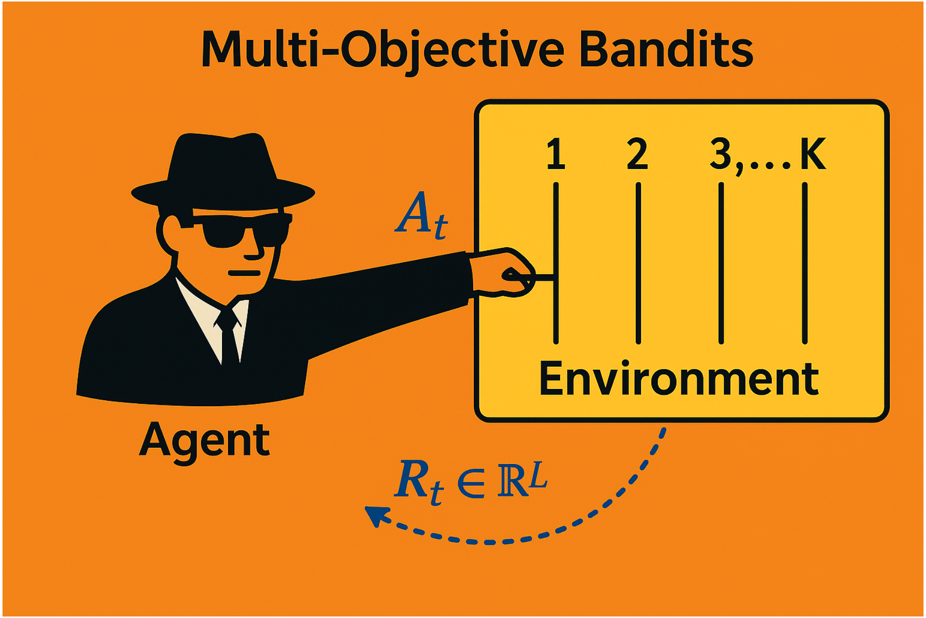
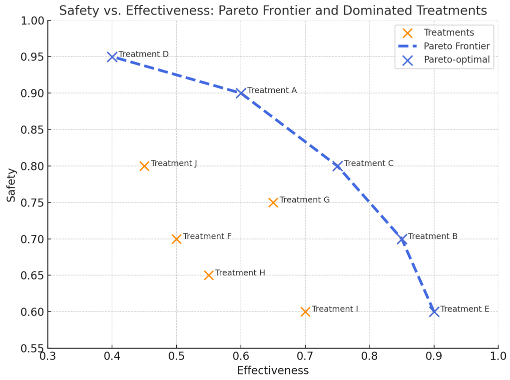
FraPPE: Fast and Efficient Preference-based Pure Exploration

Udvas Das, Apurv Shukla, Debabrota Basu

NeurIPS, 2025


Multi-Objective Bandits
1
2
3,...K
Environment
A_t
Agent
$R_t \in \mathbb{R}^L$

Pareto optimality: Conflicting objectives and need of preferences



Given a matrix of mean rewards of K-arms

$$M^{K\times L} \triangleq [M_1, M_2, ..., M_K] \in \mathcal{M}$$

and a preference cone

$$\mathcal{C} \triangleq \{\boldsymbol{x} \in \mathbb{R}^L \,|W\boldsymbol{x} \geq 0\},$$

the pareto optimal policy set is:

$$\boldsymbol{\pi}^\star \in \Pi^{\mathrm{P}} \triangleq \arg\max_{\boldsymbol{\pi}\in\Delta_K} \; M\boldsymbol{\pi} \text{ w.r.t } \mathcal{C}$$

Preference-based Pure Exploration (PrePEx)

Identify the set of Pareto optimal arms $\mathcal{P}^\star \subset \{1, 2, ..., K\}$ with probability at least $1-\delta$, while keeping **the expected number of interactions** $\mathbb{E}[\tau_\delta]$ **as low as possible**.

Lower bound on sample complexity: Characteristic time

Theorem (Lower bound (Shukla and Basu, 2024))

Given a bandit instance $M \in \mathcal{M}$, a preference cone $\mathcal{C}$, the expected stopping time of any $(1-\delta)$-correct PrePEx algorithm satisfies

$$\mathbb{E}[\tau_\delta] \geq \mathcal{T}_{M,\mathcal{C}} \log\left(\frac{1}{2.4\delta}\right)$$

$$(\mathcal{T}_{M,\mathcal{C}})^{-1} \triangleq \overbrace{\sup_{\boldsymbol{\omega}\in\Delta_K} \underbrace{\inf_{\substack{\boldsymbol{\pi}\in\Delta_K\setminus\{\boldsymbol{\pi}^\star\} \\ \boldsymbol{\pi}^\star\in\Pi^{\mathrm{P}}(M,\mathcal{C})}} \overbrace{\inf_{\tilde{M}\in\partial\Lambda(M)} \underbrace{\inf_{\boldsymbol{z}\in\mathcal{C}^+} \sum_{k=1}^{K} \boldsymbol{\omega}_k D_{\mathrm{KL}}\left(\boldsymbol{z}^\top M_k \,\middle\|\, \boldsymbol{z}^\top \tilde{M}_k\right)}_{\text{Preference that makes true and alt-instance most indistinguishable}}}^{\text{Most confusing instance}}}_{\text{Optimising over Pareto optimal policy space}}}^{\text{Optimal sampling ratio over arms}}$$

Efficiently solving the lower bound for efficient PrePEx

Step 1: Preference optimisation

Preference cone is closed, convex, and compact $\implies$ plug-in a cone programming solver

$$(\mathcal{T}_{M,\mathcal{C}})^{-1} \triangleq \overbrace{\sup_{\boldsymbol{\omega}\in\Delta_K} \underbrace{\inf_{\substack{\boldsymbol{\pi}\in\Delta_K\setminus\{\boldsymbol{\pi}^\star\} \\ \boldsymbol{\pi}^\star\in\Pi^{\mathrm{P}}(M,\mathcal{C})}} \overbrace{\inf_{\tilde{M}\in\partial\Lambda(M)} \underbrace{\min_{\boldsymbol{z}\in\mathcal{C}^+} \sum_{k=1}^{K} \boldsymbol{\omega}_k D_{\mathrm{KL}}\left(\boldsymbol{z}^\top M_k \,\middle\|\, \boldsymbol{z}^\top \tilde{M}_k\right)}_{\text{Preference that makes true and alt-instance most indistinguishable}}}^{\text{Most confusing instance}}}_{\text{Optimising over Pareto optimal policy space}}}^{\text{Optimal sampling ratio over arms}}$$

Efficiently solving the lower bound for efficient PrePEx

Step 2: Reducing policy set to $\mathcal{O}(\min\{K, L\})$ arms

– Π^{P} is spanned by p pure policies $\{\boldsymbol{\pi}_i^\star\}_{i=1}^p$ corresponding to p Pareto optimal arm.
– Number of neighbour of any Pareto optimal policy is $\mathcal{O}(\min\{K, L\})$.

$$(\mathcal{T}_{M,\mathcal{C}})^{-1} \triangleq \overbrace{\sup_{\boldsymbol{\omega}\in\Delta_K} \underbrace{\min_{\substack{\boldsymbol{\pi}_j\in\mathrm{nbd}(\boldsymbol{\pi}_i^\star)\\ \boldsymbol{\pi}_i^\star\in\{\boldsymbol{\pi}_i^\star\}_{i=1}^p}} \overbrace{\inf_{\tilde{M}\in\partial\Lambda(M)} \underbrace{\min_{\boldsymbol{z}\in\mathcal{C}^+} \sum_{k=1}^K \boldsymbol{\omega}_k D_{\mathrm{KL}}\left(\boldsymbol{z}^\top M_k \,\middle\|\, \boldsymbol{z}^\top \tilde{M}_k\right)}_{\text{Preference that makes true and alt-instance most indistinguishable}}}^{\text{Most confusing instance}}}_{\text{Optimising over Pareto optimal policy space}}}^{\text{Optimal sampling ratio over arms}}$$

Efficiently solving the lower bound for efficient PrePEx

Step 3: Reducing the Alt-set to union of lines

– The solution always lies at the boundary of the Alt-set.

– The boundary of the Alt-set is union of $\mathcal{O}(KL)$ lines

$$\overline{\Lambda}_i j(M) \triangleq \left\{ \tilde{M} \in \mathcal{M} \setminus \{M\} : \exists \boldsymbol{y} \in \mathrm{bd}(\mathcal{C}^\circ) \text{ such that } \tilde{M}(\boldsymbol{\pi}_j - \boldsymbol{\pi}_i^\star) = \boldsymbol{y} \right\}$$

$$(\mathcal{T}_{M,\mathcal{C}})^{-1} \triangleq \overbrace{\sup_{\boldsymbol{\omega} \in \Delta_K} \underbrace{\min_{\substack{\boldsymbol{\pi}_j \in \mathrm{nbd}(\boldsymbol{\pi}_i^\star) \\ \boldsymbol{\pi}_i^\star \in \{\boldsymbol{\pi}_i^\star\}_{i=1}^p}} \overbrace{\min_{\boldsymbol{y} \in \mathrm{bd}(\mathcal{C}^\circ)} \underbrace{\min_{\boldsymbol{z} \in \mathcal{C}^+} \sum_{k=1}^K \boldsymbol{\omega}_k D_{\mathrm{KL}} \left(\boldsymbol{z}^\top M_k \,\middle\|\, \boldsymbol{z}^\top \boldsymbol{y} f(\boldsymbol{\pi}_i^\star, \boldsymbol{\pi}_j) \right)}_{\text{Preference that makes true and alt-instance most indistinguishable}}}^{\text{Most confusing instance}}}_{\text{Optimising over Pareto optimal policy space}}}^{\text{Optimal sampling ratio over arms}}$$

Reduces the complexity from $\mathcal{O}\left(K^L\right)$ of (Crepon et al., 2024) to $\mathcal{O}\left(L\right)$.

Efficiently solving the lower bound for efficient PrePEx

Step 4: Frank-Wolfe for allocation optimisation in $\mathcal{O}(1)$

– Simplex is convex and closed. So, apply

$$\omega_{t+1} \leftarrow \textsc{FrankWolfe}(\omega_t, r_t, \hat{M}_t, \mathcal{C})$$

– The function to be optimised has bounded gradient and curvature.

$$(\mathcal{T}_{M,\mathcal{C}})^{-1} \triangleq \overbrace{\max_{\omega \in \Delta_K} \underbrace{\min_{\substack{\pi_j \in \mathrm{nbd}(\pi_i^\star) \\ \pi_i^\star \in \{\pi_i^\star\}_{i=1}^{p}}} \overbrace{\min_{y \in \mathrm{bd}(\mathcal{C}^\circ)} \underbrace{\min_{z \in \mathcal{C}^+} \sum_{k=1}^{K} \omega_k D_{\mathrm{KL}}\left(z^\top M_k \,\middle\|\, z^\top y f(\pi_i^\star, \pi_j)\right)}_{\text{Preference that makes true and alt-instance most indistinguishable}}}^{\textbf{Most confusing instance}}}_{\text{Optimising over Pareto optimal policy space}}}^{\text{Optimal sampling ratio over arms}}$$

We solve the lower bound with $\mathcal{O}(KL \min\{K, L\})$ complexity.

FraPPE: Frugal and Fast Preference-Based Pure Exploration

Algorithm FraPPE- **Fr**ugal and **F**ast **P**reference-Based **P**ure **E**xploration

```
Input: Confidence level δ, sequence {r_t}_{t≥1} = t^{-0.9}/K
Initialise: For t ∈ [K], sample each arm once s.t. ω_K = (1/K, ··· , 1/K), estimate M̂_K
for t > K do
    Estimate Pareto Indices: Compute P_t from M̂_t
    Compute the allocation policy with Frank-Wolfe: ω_{t+1} ← FrankWolfe(ω_t, r_t, M̂_t, C)
    C-tracking: Play A_t ∈ arg min_{a∈[K]} N_{a,t} − Σ_{s=1}^{t+1} ω_{a,s} (ties broken arbitrarily)
    Feedback: Get R_t ∈ ℝ^L and update M̂_t → M̂_{t+1}
    IF min_{π*_{i_t} ∈ {π*_i}_{i=1}^p} min_{π_j ∈ nbd(π*_{i_t})} inf_{y ∈ bd(C°)} min_{z ∈ C+} N_{t+1}^⊤ D_KL(z^⊤ M̂_{t+1} ‖ z^⊤ y f(π_j, π*_i)) > c(t+1, δ)
        break;
end for
Recommend: P_t as Pareto optimal set
```

Input: Confidence level δ, sequence $\{r_t\}_{t\geq 1} = t^{-0.9}/K$

Initialise: For $t \in [K]$, sample each arm once s.t. $\boldsymbol{\omega}_K = (1/K, \cdots, 1/K)$, estimate $\hat{M}_K$

for $t > K$ **do**

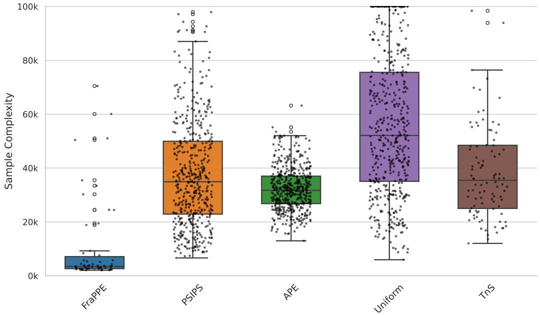
- **Estimate Pareto Indices:** Compute $\mathcal{P}_t$ from $\hat{M}_t$
- Compute the allocation policy with Frank-Wolfe: $\boldsymbol{\omega}_{t+1} \leftarrow \textsc{FrankWolfe}(\boldsymbol{\omega}_t, r_t, \hat{M}_t, \mathcal{C})$
- **C-tracking:** Play $A_t \in \arg\min_{a\in[K]} N_{a,t} - \sum_{s=1}^{t+1} \boldsymbol{\omega}_{a,s}$ (ties broken arbitrarily)
- **Feedback:** Get $R_t \in \mathbb{R}^L$ and update $\hat{M}_t \to \hat{M}_{t+1}$
- **IF** $\min_{\boldsymbol{\pi}^\star_{i_t} \in \{\boldsymbol{\pi}^\star_i\}_{i=1}^p} \min_{\boldsymbol{\pi}_j \in \mathrm{nbd}\left(\boldsymbol{\pi}^\star_{i_t}\right)} \inf_{\boldsymbol{y}\in \mathrm{bd}(\mathcal{C}^\circ)} \min_{\boldsymbol{z}\in\mathcal{C}^+} N_{t+1}^\top D_{\mathrm{KL}}\left(\boldsymbol{z}^\top \hat{M}_{t+1} \,\Big\|\, \boldsymbol{z}^\top \boldsymbol{y} f(\boldsymbol{\pi}_j, \boldsymbol{\pi}^\star_i)\right) > c(t+1, \delta)$
  - **break;**

end for

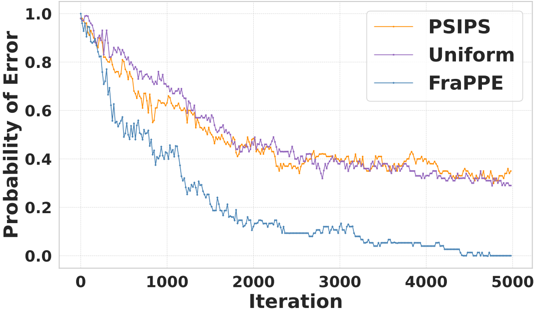
Recommend: $\mathcal{P}_t$ as Pareto optimal set

FraPPE achieves **asymptotically optimal sample complexity** and **a per-iteration computational complexity dominated by Pareto set computation**.

Empirical performance: Cov-Boost $(K = 20, L = 3, \delta = 0.01)$



$\sim$ 4 – 10X reduction



Uniformly lower probability of error

The Parting Message

We resolve an extension of the open problem (Crepon et al., 2024) for solving PrePEX

We design
- **a computationally efficient (polynomial in both K and L)** and
- **statistically optimal PrePEx algorithm**
- **beyond Gaussian rewards** and
- **for arbitrary preference cones**.

What's Ahead?

To scale FraPPE to practical applications of PrePEx, e.g. aligning LLMs with RL under Human Feedback (RLHF) (Ji et al., 2023).

References I

Crepon, E., Garivier, A., and M Koolen, W. (2024). Sequential learning of the Pareto front for multi-objective bandits. In Dasgupta, S., Mandt, S., and Li, Y., editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3583–3591. PMLR.

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Chen, B., Sun, R., Wang, Y., and Yang, Y. (2023). Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704.

Shukla, A. and Basu, D. (2024). Preference-based pure exploration. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.