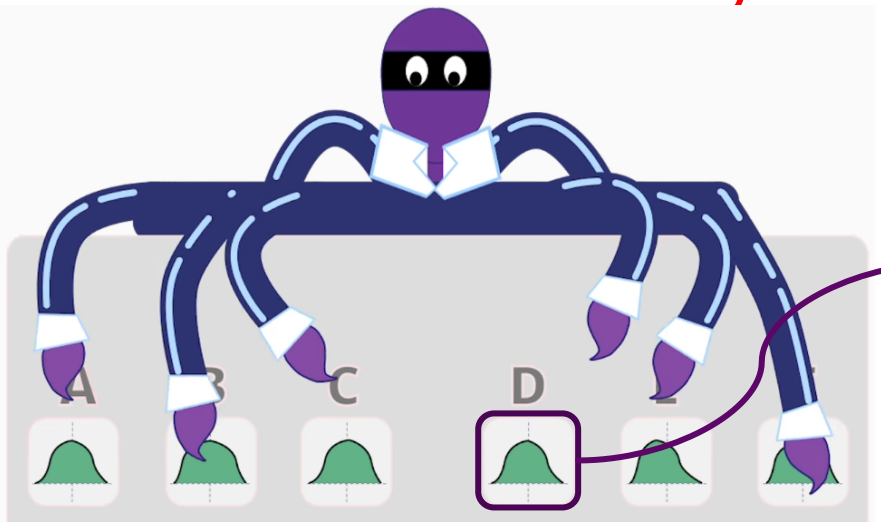
uniINF: Best-of-Both-Worlds Algorithm for Parameter-Free Heavy-Tailed MABs

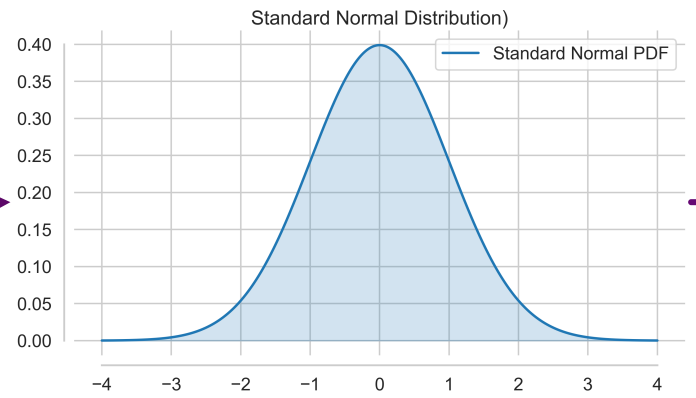
Published on ICLR2025 (Spotlight)

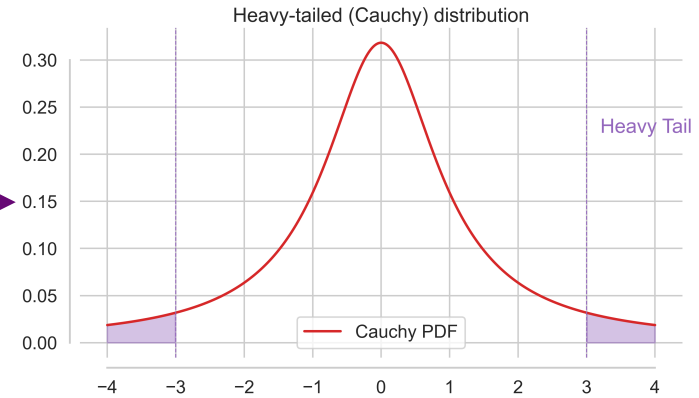
Yu Chen* Jiatai Huang* Yan Dai* Longbo Huang

Introduction

- Multi-Armed Bandits (MABs) is a classic theoretical model in online learning to address the exploration-exploitation trade-off problem.
- Many studies for MABs assume that the loss for each arm is sub-Gaussian, or even bounded.
- Challenges in the Real World: The loss distributions in many real-world scenarios are heavy-tailed.







Introduction

- Heavy-tailed distribution ν
  - Variance of loss $l \sim \nu$ may be infinite.
  - Only α-th moment is bounded ($\alpha \in (1, 2]$): $\mathbb{E}[|l|^\alpha] \leq \sigma^\alpha$ for some $\sigma \geq 0$.
- E.g., Pareto distribution, Student t distribution.
- Empirical evidence shows that heavy-tailed distributions frequently occur in many key real-world tasks.
  - Network Routing, e.g., [Liebeherr et al., 2012]
  - Algorithm Portfolio Selection, e.g., [Gagliolo & Schmidhuber, 2011]
  - Online Deep Learning, e.g., [Zhang et al.,2020]

Heavy-Tailed Multi-Armed Bandits (HTMABs)

- HTMAB problem is consisted with K arms and T rounds.
- In Round $t \in [T]$,

1. Loss vector $\boldsymbol{l}_t \in \mathbb{R}^K$ is generated by the environment from heavy tailed distribution $\boldsymbol{\nu}_t$ where $\mathbb{E}_{l\sim\nu_{t,i}}[|l|^\alpha] \leq \sigma^\alpha$ for each arm $i \in [K]$, which is hidden to the player.
2. Player choose an action $i_t \in [K]$ then observes and suffers a loss of l_{t,i_t}.

- The problem objective is minimizing the regret:

$$\Re_T = \max_{i\in[K]} \mathbb{E}\left[\sum_{t=1}^{T}\left(l_{t,i_t} - l_{t,i}\right)\right]$$

Heavy-Tailed Multi-Armed Bandits (HTMABs)

- **Unknown Parameters** (σ, α) : In real applications, we usually cannot predict the heavy-tailed parameters (σ, α) of the environment, which is called "parameter free" setting.
- **Unknown Environments:** The environment may be one of two types, and the algorithm cannot predict it:
  - Stochastic: The loss distribution is fixed and does not change over time ($\boldsymbol{\nu}_t = \boldsymbol{\nu}_1$).
  - Adversarial: The loss can be arbitrarily generated by an adversary in advance and changes over time.
- We aim to develop the Best of Both Worlds (BoBW) algorithm for parameter-free HTMABs

Overview of Related Works For HTMABs

Our Main Contribution

- **uniINF** : the **first** Parameter-Free HTMAB algorithm that achieves the BoBW guarantees.
  - Parameter-Free: without reliance on a prior knowledge of (σ, α).
  - BoBW: automatically achieves nearly-optimal regret bounds in both stochastic and adversarial cases.

Algorithm [a]	α-Free?	σ-Free?	Env.	Regret	Opt?
Lower Bound (Bubeck et al., 2013)	—	—	—	$\Omega\left(\sum_{i\neq i^*}(\frac{\sigma^\alpha}{\Delta_i})^{\frac{1}{\alpha-1}}\log T\right)$ $\Omega\left(\sigma K^{1-1/\alpha}T^{1/\alpha}\right)$	—
RobustUCB (Bubeck et al., 2013)	✗	✗	**Only Stoc.**	$\mathcal{O}\left(\sum_{i\neq i^*}(\frac{\sigma^\alpha}{\Delta_i})^{\frac{1}{\alpha-1}}\log T\right)$ $\widetilde{\mathcal{O}}\left(\sigma K^{1-1/\alpha}T^{1/\alpha}\right)$	✓ ✓
Robust MOSS (Wei & Srivastava, 2020)	✗	✗	**Only Stoc.**	$\mathcal{O}\left(\sum_{i\neq i^*}(\frac{\sigma^\alpha}{\Delta_i})^{\frac{1}{\alpha-1}}\log(\frac{T}{K}(\frac{\sigma^\alpha}{\Delta_i^\alpha})^{\frac{1}{\alpha-1}})\right)$ $\mathcal{O}\left(\sigma K^{1-1/\alpha}T^{1/\alpha}\right)$	✓ [b] ✓
APE^2 (Lee et al., 2020b)	✗	✓	**Only Stoc.**	$\mathcal{O}\left(e^\sigma+\sum_{i\neq i^*}(\frac{1}{\Delta_i})^{\frac{1}{\alpha-1}}(T\Delta_i^{\frac{\alpha}{\alpha-1}}\log K)^{\frac{\alpha}{(\alpha-1)\log K}}\right)$ $\widetilde{\mathcal{O}}\left(\exp(\sigma^{1/\alpha})K^{1-1/\alpha}T^{1/\alpha}\right)$	✗ ✗
HTINF (Huang et al., 2022)	✗	✗	**Stoc.** **Adv.**	$\mathcal{O}\left(\sum_{i\neq i^*}(\frac{\sigma^\alpha}{\Delta_i})^{\frac{1}{\alpha-1}}\log T\right)$ $\mathcal{O}\left(\sigma K^{1-1/\alpha}T^{1/\alpha}\right)$	✓ ✓
OptHTINF (Huang et al., 2022)	✓	✓	**Stoc.** **Adv.**	$\mathcal{O}\left(\sum_{i\neq i^*}(\frac{\sigma^{2\alpha}}{\Delta_i^{3-\alpha}})^{\frac{1}{\alpha-1}}\log T\right)$ $\mathcal{O}\left(\sigma^\alpha K^{\frac{\alpha-1}{2}}T^{\frac{3-\alpha}{2}}\right)$	✗ ✗
AdaTINF (Huang et al., 2022)	✓	✓	**Only Adv.**	$\mathcal{O}\left(\sigma K^{1-1/\alpha}T^{1/\alpha}\right)$	✓
AdaR-UCB (Genalti et al., 2024)	✓	✓	**Only Stoc.**	$\mathcal{O}\left(\sum_{i\neq i^*}(\frac{\sigma^\alpha}{\Delta_i})^{\frac{1}{\alpha-1}}\log T\right)$ $\widetilde{\mathcal{O}}\left(\sigma K^{1-1/\alpha}T^{1/\alpha}\right)$	✓ ✓
uniINF **(Ours)**	✓	✓	**Stoc.** **Adv.**	$\mathcal{O}\left(K(\frac{\sigma^\alpha}{\Delta_{\min}})^{\frac{1}{\alpha-1}}\log T\cdot\log\frac{\sigma^\alpha}{\Delta_{\min}}\right)$ $\widetilde{\mathcal{O}}\left(\sigma K^{1-1/\alpha}T^{1/\alpha}\right)$	✓ [c] ✓

[a] **α-Free?** and **σ-Free?** denotes whether the algorithm is parameter-free *w.r.t.* α and σ, respectively. **Env.** includes the environments that the algorithm can work; if one algorithm can work in ***both*** stochastic and adversarial environments, then we mark this column by green. **Regret** describes the algorithmic guarantees, usually (if applicable) instance-dependent ones above instance-independent ones. **Opt?** means whether the algorithm matches the instance-dependent lower bound by Bubeck et al. (2013) *up to constant factors*, or the instance-independent lower bound *up to logarithmic factors*.

[b] Up to $\log(\sigma^\alpha)$ and $\log(1/\Delta_i^\alpha)$ factors.

[c] Up to $\log(\sigma^\alpha)$ and $\log(1/\Delta_{\min})$ factors when all Δ_i's are similar to the dominant sub-optimal gap $\Delta_{\min}$.

Main Results

- Assumption 0 [Unique Best Arm, very common in MAB literatures]:

  There exists an optimal arm $i^* \in [K]$ such that $\mathbb{E}_{l\sim\nu_{t,i}}[l] - \mathbb{E}_{l\sim\nu_{t,i^*}}[l] > 0$ for any $i \neq i^*$.'

- Assumption 1 [Truncated Non-Negativity, Huang et al., 2022]:

  For any t and optimal loss $l \sim \nu_{t,i^*}$, $\mathbb{E}\left[l \cdot \mathbf{1}[|l| \geq M]\right] \geq 0$ holds for any $M \geq 0$.

With Ass 0 and 1, we have the following main theorem: **uniINF is BoBW algorithm**.

Main Theorem: `uniINF` guarantees near-optimal regret automatically under **both** stochastic and adversarial environments **without any prior knowledge** of the heavy-tailed parameters α and σ.

- **Adversarial** environments: `uniINF` achieves $\mathcal{R}_T = \widetilde{\mathcal{O}}\left(\sigma K^{1-1/\alpha} T^{1/\alpha}\right)$.
  - Matches stochastic-case instance-independent lower bound (Bubeck et al., 2013) up to logs.
- **Stochastic** environments: `uniINF` achieves $\mathcal{R}_T = \widetilde{\mathcal{O}}\left(K(\frac{\sigma^\alpha}{\Delta_{\min}})^{\frac{1}{\alpha-1}} \log T \cdot \log \frac{\sigma^\alpha}{\Delta_{\min}}\right)$.
  - Nearly matches stochastic-case instance-dependent lower bound up to polylog$(\sigma^\alpha, \Delta_{\min}^{-1})$.

Discussion on the Assumption 1

- We summarize the existing results in HTMABs for the relationship between the prior knowledge of heavy-tailed parameters (σ, α) and Assumption 1 in the following table.
  - [Cheng et al. (2024)]: when parameters (σ, α) are known, Assumption 1 is redundant.
  - [Genalti et al. (2024)]: achieving optimal regret guarantees is impossible without any assumptions.

Assumptions	**Known** (σ, α)	**Unknown** (σ, α)
With Assumption 1	`HTINF` achieves BoBW (Huang et al., 2022)	`uniINF` achieves BoBW **(Our Work)**
Weaker than Assumption 1?	Inferred by Cheng et al. (2024) (discussed below)	**Open Problem**
Without Any Assumption	`SAO-HT` achieves BoBW (Cheng et al., 2024)	No BoBW possible (Genalti et al., 2024, Theorems 2 & 3)

- **Our work**: Assumption 1 is sufficient for achieving BoBW under parameter-free settings.

uniINF: The Universal INF-type Algorithm For Parameter-Free HTMAB

- **Core framework**: "Follow the Regularized Leader" (FTRL).
- Three key-innovations components:

1. Refined Analysis for Log-Barrier Regularizer
2. Adaptive Skipping-Clipping Loss Tuning
3. Auto-Balancing Learning Rates

Algorithm `uniINF`: the universal INF-type algorithm for Parameter-Free HTMAB

Initialize the learning rate $S_1 \leftarrow 4$.
for $t = 1, 2, \ldots, T$ **do**
  Apply *Follow-the-Regularized-Leader* (FTRL) to calculate the action $\boldsymbol{x}_t \in \Delta^{[K]}$ with the log-barrier regularizer Ψ_t defined in Equation (1): ▷ FTRL with refined log-barrier regularizer.

$$\boldsymbol{x}_t \leftarrow \operatorname*{argmin}_{\boldsymbol{x} \in \Delta^{[K]}} \left(\sum_{s=1}^{t-1} \left\langle \widetilde{\boldsymbol{\ell}}_s, \boldsymbol{x} \right\rangle + \Psi_t(\boldsymbol{x}) \right), \quad \Psi_t(\boldsymbol{x}) := -S_t \sum_{i=1}^{K} \log x_i \tag{1}$$

  Sample action $i_t \sim \boldsymbol{x}_t$. Play i_t and observe feedback ℓ_{t,i_t}.
  for $i = 1, 2, \ldots, K$ **do**
    Calculate $\ell_{t,i}^{\texttt{skip}}$ and $\ell_{t,i}^{\texttt{clip}}$ with action-dependent threshold $C_{t,i}$ by Equation (2). ▷ Adaptive skipping-clipping loss tuning.

$$\ell_{t,i}^{\texttt{skip}} := \begin{cases} \ell_{t,i} & \text{if } |\ell_{t,i}| < C_{t,i} \\ 0 & \text{otherwise} \end{cases}, \quad \ell_{t,i}^{\texttt{clip}} := \begin{cases} C_{t,i} & \text{if } \ell_{t,i} \geq C_{t,i} \\ -C_{t,i} & \text{if } \ell_{t,i} \leq -C_{t,i} \\ \ell_{t,i} & \text{otherwise} \end{cases}, \quad \text{with } C_{t,i} := \frac{S_t}{4(1 - x_{t,i})} \tag{2}$$

  Calculate the importance sampling estimate of $\boldsymbol{\ell}_t^{\texttt{skip}}$, namely $\widetilde{\boldsymbol{\ell}}_t$, where

$$\widetilde{\ell}_{t,i} = \frac{\ell_{t,i}^{\texttt{skip}}}{x_{t,i}} \cdot \mathbb{1}[i = i_t], \quad \forall i \in [K].$$

  Update the learning rate S_{t+1} as ▷ Auto-balancing learning rates

$$S_{t+1}^2 = S_t^2 + (\ell_{t,i_t}^{\texttt{clip}})^2 \cdot (1 - x_{t,i_t})^2 \cdot (K \log T)^{-1}. \tag{3}$$

Analysis Preparation: FTRL Regret Decomposition

- By basic FTRL regret decomposition, we have

$$\mathcal{R}_T \leq \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \textsc{Div}_t\right]}_{\textsc{Bregman Divergence Terms}} + \underbrace{\mathbb{E}\left[\sum_{t=0}^{T-1} \textsc{Shift}_t\right]}_{\Psi\textsc{-Shifting Terms}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} |\textsc{SkipErr}_t| \cdot \mathbb{1}[i_t \neq i^*]\right]}_{\textsc{Sub-optimal Skipping Losses}} + \sigma K.$$

1. Bregman Divergence terms

$$\mathrm{Div}_t = \Psi_t(\boldsymbol{x}_t) - \Psi_t(\boldsymbol{z}_t) - \langle \nabla\Psi_t(\boldsymbol{x}_t), \boldsymbol{z}_t - \boldsymbol{x}_t \rangle,$$

$$\text{where } \boldsymbol{z}_t := \mathrm{argmin}_{\boldsymbol{z}\in\Delta^{[K]}} \left(\sum_{s=1}^{t} \langle \tilde{\boldsymbol{l}}_s, \boldsymbol{z} \rangle + \Psi_t(\boldsymbol{z}) \right),$$

2. The Ψ-Drift term Shift_t for benchmark $\tilde{y}$, $(\tilde{y}_{i^*} = \frac{1}{T}, \tilde{y}_i = 1 - \frac{K-1}{T}, \forall i \neq i^*)$

$$\mathrm{Shift}_t = (\Psi_{t+1} - \Psi_t)(\tilde{\boldsymbol{y}}) - (\Psi_{t+1} - \Psi_t)(\boldsymbol{x}_{t+1}),$$

3. Skipping Error terms $\mathrm{SkipErr}_t$:

$$\mathrm{SkipErr}_t = l_{t,i_t} - l_{t,i_t}^{\mathrm{skip}}.$$

Key Technique: Refined Analysis for Log-Barrier Regularizer

- **Log-barrier regularizer** $\Psi_t(\boldsymbol{x}) = -S_t \sum_{i=1}^{K} \log(x_i)$, where S_t^{-1} is learning rate in round t.
- It is known that log-barrier applied to a loss sequence $\{\boldsymbol{l}_t\}_{t=1}^{T}$ ensures the control of Div_t
  - [Foster et al., 2016, Lemma 16] $\mathrm{Div}_t \leq S_t^{-1} \sum_i x_{t,i}^2 l_{t,i}^2$ for non-negative $\boldsymbol{l}_t$.
  - [Dai et al., 2023, Lemma 3.1] $\mathrm{Div}_t \leq S_t^{-1} \sum_i x_{t,i} l_{t,i}^2$ for general $\boldsymbol{l}_t$.
- **Our work**: $\mathrm{Div}_t \leq S_t^{-1} \sum_i x_{t,i}^2 \left(1 - x_{t,i}\right)^2 l_{t,i}^2$ for general $\boldsymbol{l}_t$ when S_t is adequately large compared to $\|\boldsymbol{l}_t\|_\infty$.
- Extra $\left(1 - x_{t,i}\right)^2$ terms : contributes to exclude the optimal arm $i^* \in [K]$ in stochastic case analysis.

Key Technique: Adaptive Skipping-Clipping Loss Tuning

- We cannot straightforward control the Bregman divergence term

$$\mathrm{Div}_t \leq S_t^{-1} \sum_i x_{t,i}^2 \left(1 - x_{t,i}\right)^2 l_{t,i}^2$$

with original loss vector $\boldsymbol{l}_t$, since $\mathbb{E}\left[l_{t,i}^2\right]$ might be unbounded in HTMABs.

- **Previous Solution:** [Huang et al., 2022] developed a skipping technique that replaces the extremely large loss with 0 with dynamic threshold.
- **Challenge**: Skipping is not a free lunch, which will influence the update of learning rates and reduce the learning efficiency.

Key Technique: Adaptive Skipping-Clipping Loss Tuning

- **Our Solution**: we apply skipping-clipping techniques with adaptive thresholds $C_{t,i}$.
- We tune loss vector $\boldsymbol{l}_t$ via Eq.(2):

$$\ell_{t,i}^{\texttt{skip}} := \begin{cases} \ell_{t,i} & \text{if } |\ell_{t,i}| < C_{t,i} \\ 0 & \text{otherwise} \end{cases}, \quad \ell_{t,i}^{\texttt{clip}} := \begin{cases} C_{t,i} & \text{if } \ell_{t,i} \geq C_{t,i} \\ -C_{t,i} & \text{if } \ell_{t,i} \leq -C_{t,i} \\ \ell_{t,i} & \text{otherwise} \end{cases}, \quad \text{with } C_{t,i} := \frac{S_t}{4(1-x_{t,i})} \tag{2}$$

- This adaptive skipping-clipping technique ensure that every loss will influence the learning process.

Key Technique: Auto-Balancing Learning Rates

- **Challenge**: How to set optimal learning rates S_t when (σ, α) is unknown?
- **Our Solution:** Automatically balance Div_t and Shift_t by proper S_t

$$\underbrace{\mathtt{Div}_t \le \mathcal{O}\left(S_t^{-1}\left(\ell_{t,i_t}^{\mathtt{clip}}\right)^2 (1-x_{t,i_t})^2\right)}_{\text{Bregman Divergence}}, \quad \underbrace{\mathtt{Shift}_t \le \mathcal{O}\left((S_{t+1}-S_t)\cdot K\log T\right)}_{\Psi\text{-Shifting}}. \tag{4}$$

- Thus, to make Div_t roughly the same as Shift_t, it suffices to ensure

$$(S_{t+1}-S_t)S_t \approx \left(l_{t,i_t}^{\mathrm{clip}}\right)^2 \left(1-x_{t,i_t}\right)^2 \cdot (K\cdot \log T)^{-1}$$

- Moreover, notice that $(S_{t+1}-S_t)S_t \approx S_{t+1}^2 - S_t^2$ if S_t does not change too much. Therefore, our definition of S_t follows

$$S_{t+1}^2 = S_t^2 + (\ell_{t,i_t}^{\mathtt{clip}})^2 \cdot (1-x_{t,i_t})^2 \cdot (K\log T)^{-1}. \tag{3}$$

Conclusion & Summary

- **Our Contribution:** The First Parameter-Free BoBW Algorithm for HTMABs.
- We introduced **uniINF**, the **first** algorithm that solves the Parameter-Free Heavy-Tailed Multi-Armed Bandit (HTMAB) problem with the Best-of-Both-Worlds (BoBW) property.
  - uniINF performs nearly-optimally in both **stochastic** and **adversarial** environments without requiring any prior knowledge of the environment type or its heavy-tail parameters (σ, α).
- Our analytic tools and key innovation techniques could be independent interests in MABs community.

Thanks for Listening