Catoni Contextual Bandits are Robust to Heavy-tailed Rewards

Chenlu Ye[1] Yujia Jin[2] Alekh Agarwal[2] Tong Zhang[1]

[1]University of Illinois Urbana-Champaign [2]Google Research

Highlights

- **The Problem: Heavy-Tailed Rewards in Contextual Bandits**:
  - Standard Assumption: rewards are bounded within a fixed range $[0, R]$
  - Previous Work: regret scales polynomially with R
  - Limitation: heavy-tailed rewards or rewards where the worst-case range can be substantially larger than the variance.
  - E.g. Financial markets (stock prices), online advertising (value returns), communication networks (waiting times)
- **The Goal**: obtain performance guarantees depending on the reward variance, not the worst-case range

Table: Comparison between different algorithms for stochastic contextual bandits

Algorithm	Function Type	Known Variances	Regret Bound
Weighted OFUL+ (Zhou and Gu, 2022)	Linear	✓	$\widetilde{O}(d\sqrt{\sum_{t\in[T]}\sigma_t^2} + dR)$
Heavy-OFUL (Huang et al., 2024) [1] AdaOFUL (Li and Sun, 2024)	Linear	✓	$\widetilde{O}(d\sqrt{\sum_{t\in[T]}\sigma_t^2})$
OLS (Pacchiano, 2024)	Non-linear	✓	$\widetilde{O}(\sigma\sqrt{d_{\mathcal{F}}\ln N_{\mathcal{F}}} + Rd_{\mathcal{F}}\ln N_{\mathcal{F}})$
Catoni-OFUL (Theorem 2)	Non-linear	✓	$\widetilde{O}(\sqrt{\sum_{t\in[T]}\sigma_t^2\cdot d_{\mathcal{F}}\ln N_{\mathcal{F}}} + d_{\mathcal{F}}\ln N_{\mathcal{F}})$
SAVE (Zhao et al., 2023b)	Linear	✗	$\widetilde{O}(d\sqrt{\sum_{t\in[T]}\sigma_t^2} + dR)$
DistUCB (Wang et al., 2024b) [2]	Non-linear	✗	$\widetilde{O}(\sqrt{\sum_{t\in[T]}\sigma_t^2\cdot\tilde{d}_{\mathcal{F}}\ln N_{\mathcal{F}}} + R\tilde{d}_{\mathcal{F}}\ln N_{\mathcal{F}})$
Unknown-Variance OLS (Pacchiano, 2024)	Non-linear	✗	$\widetilde{O}(d_{\mathcal{F}}\sqrt{\sum_{t\in[T]}\sigma_t^2\cdot\ln N_{\mathcal{F}}} + Rd_{\mathcal{F}}\ln N_{\mathcal{F}})$
VACB (Theorem 3)	Non-linear	✗	$\widetilde{O}(d_{\mathcal{F}}\sqrt{\sum_{t\in[T]}\sigma_t^2\cdot\ln N_{\mathcal{F}}} + d_{\mathcal{F}}(\ln N_{\mathcal{F}})^{3/4})$

Formulation

- Catoni Estimator: unique zero of the increasing function

$$f(x;\{Z_i\}_{i\in[t]},\theta) := \sum_{i\in[t]}\Psi(\theta(Z_i - x)),\quad \Psi(x) = \begin{cases}\log(1+x+x^2/2) & \text{if } x\ge 0,\\ -\log(1-x+x^2/2) & \text{if } x<0.\end{cases}$$

- How it Works: Catoni Regression with weights $\bar{\sigma}_i$

$$\min_f\max_{f'}\underbrace{R(f)-R(f')}_{\text{estimate robustly using Catoni mean}},\quad R(f)=\sum_i\mathbb{E}_i\frac{1}{\bar{\sigma}_i^2}[(f(x_i)-y_i)^2]\qquad(1)$$

Algorithm for Known Variance

- combines the OFUL framework with a variance-weighted regression approach
- Uses the Catoni estimator to construct a robust confidence set for the true reward function.

Algorithm 1: Catoni-OFUL

Input: Parameter $\alpha > 0$, δ and $\hat{\beta}_t$ for each $t \in [T]$.
for t=1,2,. . . ,T **do**
Pick action $x_t = \mathrm{argmax}_{x\in\mathcal{X}_t}\max_{f\in\mathcal{F}_{t-1}} f(x)$;
Observe the reward y_t;
Let $\bar{\sigma}_t = \max\left(\alpha, \sigma_t, \sqrt{4\iota(\delta)L_f D_{\mathcal{F}_{t-1}}(x_t; x_{[t-1]}, \bar{\sigma}_{[t-1]})}\right)$;
Estimate $\hat{f}_t$ in (3);
Construct confidence set

$$\mathcal{F}_t := \left\{f\in\mathcal{F}_{t-1} : \sum_{i\in[t]}\frac{1}{\bar{\sigma}_i^2}\left(f(x_i)-\hat{f}_t(x_i)\right)^2 \le \hat{\beta}_t^2\right\};$$

end for

- Result:
  - Upper bound:

$$\tilde{O}\Big(\big(\sum_{t\in[T]}\sigma_t^2\cdot d_F\log N_F\big)^{1/2} + d_F\log N_F\Big),$$

where σ_t is the reward variance, d_F, N_F are the eluder dimension and log-covering number of function space $\mathcal{F}$

  - Match lower bound interms of $\Omega\Big(\mathbb{E}\sum_{t\in[T]}\sigma_t^2\Big)$.

Algorithms for Unkown Variance

- Employs a "peeling" technique: samples are grouped into levels based on their uncertainty.
- Instead of estimating variance for each round, it robustly estimates an aggregate variance for each level using the Catoni estimator.
- Avoiding the need for a separate function class to predict variances.

Algorithm 2: Variance-Agnostic Catoni Bandit

Input: Parameter $\gamma > 0$, $L = \lceil\log_2(1/\gamma)\rceil$, $l_\star = \lceil\log_2(1076\iota'(\delta))\rceil$.
Initialize the estimators for all layers: $\lambda^l \leftarrow 2^{-2l}$, $\hat{\beta}_0^l \leftarrow 2^{-l+1}$, $\Psi_0^l \leftarrow \emptyset$ for all $l \in [l_\star, L]$.
for t=1,. . . ,T **do**
Observe $\mathcal{X}_t$, and initialize $\mathcal{X}_t^1 \leftarrow \mathcal{X}_t$, $l \leftarrow l_\star$.
while x_t is not specified **do**
if $D_t^l(x) \le \gamma$ for all $x \in \mathcal{X}_t^l$ **then**
Choose $x_t, f_{t-1}^l \leftarrow \mathrm{argmax}_{x\in\mathcal{X}_t^l, f\in\mathcal{F}_{t-1}^l} f(x)$
Observe y_t.
Break.
else if $D_t^l(x) \le 2^{-l}$ for all $x \in \mathcal{X}_t^l$ **then**
Update $\mathcal{X}_t^{l+1} \leftarrow \{x \in \mathcal{X}_t^l \mid \hat{f}_{t-1}^l(x) \ge \max_{x\in\mathcal{X}_t^l}\hat{f}_{t-1}^l(x) - 2^{-l+1}\hat{\beta}_{t-1}^l\}$.
else
Choose $x_t \in \mathcal{X}_t^l$ such that $D_t^l(x_t) > 2^{-l}$ and observe y_t.
Update $w_t \leftarrow 2^l D_t^l(x_t)$.
Update the index sets: $\Psi_t^l \leftarrow \Psi_{t-1}^l \cup \{t\}$ and $\Psi_t^{l'} \leftarrow \Psi_{t-1}^{l'}$ for $l' \ne l$.
Optimize $\hat{f}_t^l$ as in (7), and choose the confidence set $\mathcal{F}_t^l$ defined in (9).
end if
Update $l \leftarrow l + 1$.
end while
For $l \in [L]$ s.t. $\Psi_t^l = \Psi_{t-1}^l$, $\hat{f}_t^l \leftarrow \hat{f}_{t-1}^l$, $\mathcal{F}_t^l \leftarrow \mathcal{F}_{t-1}^l$.
end for

Algorithms for Unkown Variance

- Result: Still achieves a variance-dependent regret bound with only a logarithmic dependence on the reward range R.
- Key takeout: by using **peeling**, instead of **uniforming variance** (variance weighting)

$$\begin{aligned}S_t &:= \sum_{i\in\Psi_t}\mathrm{Var}[Z_i(f,f')]\\ &= \sum_{i\in\Psi_t}\mathbb{E}\Big[\frac{1}{w_i^2}(f(x_i)-f'(x_i))^2(f^\star(x_i)-y_i)^2\,\Big|\,x_i\Big]\\ &\le \sum_{i\in\Psi_t}\frac{(f(x_i)-f'(x_i))^2}{w_i^2}\cdot\frac{\sigma_i^2}{w_i^2}.\end{aligned}$$

We **uniform uncertainty**

$$\begin{aligned}S_t &\le \max_{i\in\Psi_t}\frac{(f(x_i)-f'(x_i))^2}{w_i^2}\cdot\sum_{i\in\Psi_t}\frac{\sigma_i^2}{w_i^2}\\ &\le \underbrace{\max_{i\in\Psi_t}\frac{D_i^2}{w_i^2}\cdot\Big(\sum_{\tau\in[i-1]}\frac{(f(x_\tau)-f'(x_\tau))^2}{w_\tau^2}+\lambda\Big)}_{\text{Uniform bound }\le 2^{-2l}\cdot 4\hat{\beta}_{t-1}^2}\cdot\sum_{i\in\Psi_t}\frac{\sigma_i^2}{w_i^2},\end{aligned}$$

Reference

[1] Zhao, H., He, J., Zhou, D., Zhang, T., and Gu, Q. (2023b). Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. In The Thirty Sixth Annual Conference on Learning Theory, pages 4977–5020. PMLR.

[2] Huang, J., Zhong, H., Wang, L., and Yang, L. (2024). Tackling heavy-tailed rewards in reinforcement learning with function approximation: Minimax optimal and instance-dependent regret bounds. Advances in Neural Information Processing Systems, 36.

[3] Li, X. and Sun, Q. (2024). Variance-aware decision making with linear function approximation under heavytailed rewards. Transactions on Machine Learning Research.