Softmax Gradient Policy for Variance Minimization and Risk-Averse Multi-Armed Bandits

Gabriel Turinici

Dauphine | PSL CEREMADE cnrs
UMR CNRS 7534

CEREMADE Université Paris Dauphine - PSL, Paris, France

AIS 2026

12th International Artificial Intelligence Symposium, Sept. 21 – 24, 2026, Italy

Outline

Introduction: from mean maximization to risk-aware bandits

Classical MAB (maximize expected reward)
An agent chooses action A (can be random variable) among k arms repeatedly, gets reward (r.v.) $R(A)$; the goal is to maximize $\mathbb{E}[R(A)]$. This assumes **risk neutrality**.

Risk-aware setting: in many applications (finance, healthcare, safety-critical systems), stability matters more than raw return. We seek the arm with **lowest variance** σ_a^2, not highest mean.

Existing approaches: UCB/LCB-type bounds for mean-variance trade-off (Sani et al. 2012) or CVaR-based methods.

Our angle: Softmax policy gradient – no stored variance statistics; estimator computed on the fly via a single extra draw per step (cf. prior L2-regularized PG MAB in [1]).

Outline

Softmax parameterization and objective

k arms, each with unknown mean μ_a and variance σ_a^2.
Preference vector $H \in \mathbb{R}^k$, softmax policy:

$$\Pi_H(a) = \frac{e^{H(a)}}{\sum_{b=1}^{k} e^{H(b)}}.$$

Variance-minimizing objective:

$$\mathcal{L}(H) = \sum_{a=1}^{k} \Pi_H(a)\, \sigma_a^2 = \mathbb{E}_{A\sim\Pi_H}\big[\mathbb{V}(R(A) \mid A)\big].$$

Goal: $\min_H \mathcal{L}(H)$ so that the policy concentrates on the lowest-variance arm.

Difficulty: σ_a^2 cannot be read from a single reward – we need an unbiased estimator of the per-arm variance.

Key idea: unbiased variance from two independent draws

Idea: Sample $A_t \sim \Pi_{H_t}$, then draw two i.i.d. rewards $R_t, R'_t \sim R(A_t)$.
Define the composite reward:

$$\boxed{\mathscr{R}_t = \frac{1}{2}(R_t - R'_t)^2} \qquad \text{with } \mathbb{E}[\mathscr{R}_t \mid A_t = a] = \sigma_a^2. \tag{1}$$

Indeed, for i.i.d. X, Y: $\mathbb{E}[(X-Y)^2] = 2\sigma_a^2$.
Gradient update (Robbins–Monro):

$$H_{t+1}(a) = H_t(a) - \rho_t\,(\mathscr{R}_t - \bar{\mathscr{R}}_t)\,\big(\mathbb{1}_{a=A_t} - \Pi_{H_t}(a)\big),$$

with $\bar{\mathscr{R}}_t$ a running-average baseline for variance reduction.
No per-arm statistics stored – everything is on-the-fly.

Risk-aware extension: mean–variance trade-off

A general objective combining reward and variance is to minimize:

$$\mathcal{L}_r(H) = \sum_{a=1}^{k} \Pi_H(a) \left(\lambda_\sigma \sigma_a^2 + \lambda_\mu \mu_a\right), \quad \lambda_\mu \leq 0 \leq \lambda_\sigma.$$

Special cases:

- $\lambda_\mu = 0,\ \lambda_\sigma = 1$: pure variance minimization (previous slide).
- $\lambda_\mu < 0$, small λ_σ: prefer high-mean arms but penalize risk.

Algorithm: draw a mini-batch of ℓ rewards from the selected arm, compute empirical mean $\hat{\mu}_t$ and variance $\hat{\sigma}_t^2$, set composite reward $\mathscr{R}_t = \lambda_\sigma \hat{\sigma}_t^2 + \lambda_\mu \hat{\mu}_t$, then apply the same update.

Outline

Theory: almost-sure convergence

Proposition

Assume constant step-size $\rho_t = \rho$ and that all arms have distinct quality scores $q_a = \lambda_\sigma \sigma_a^2 + \lambda_\mu \mu_a$. If rewards are bounded, then:

$$\Pi_{H_t} \xrightarrow[t\to\infty]{} \Pi^\star = \delta_{a^\star} \quad \textit{almost surely},$$

where $a^\star$ uniquely minimizes q_a.

Proof: use general proofs for the softmax policy gradient cf. Mei et al. [2].
Remark: boundedness on rewards is strong but benign in practice (truncation preserves the optimum). Heavy-tailed distributions remain future work.

Outline

Numerical setup

Code: https://github.com/gabriel-turinici/min_variance_and_risk_averse_softmax_MAB

Variance minimizing version i.e. $\lambda_\mu = 0$, $\lambda_\sigma = 1$.

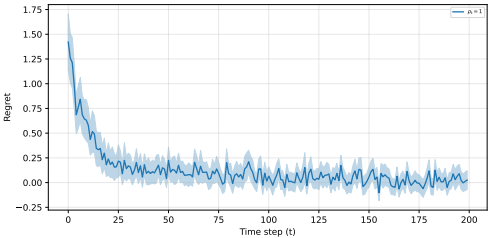
	Arms	Means	Std. dev.
Toy 2 arms	$k = 2$	$(0, 0)$	$(1, \sqrt{2})$
Toy 10 arms	$k = 10$	$(0, \ldots, 0)$	$\sigma_a = \begin{cases} \sqrt{4} & a < 10 \\ 1 & a = 10 \end{cases}$
Hard 10 arms	$k = 10$	$\mu_a \sim \mathcal{N}(4, 1)$	$\sigma_a^2 \sim U[1, 5]$ (Sutton-Barto)

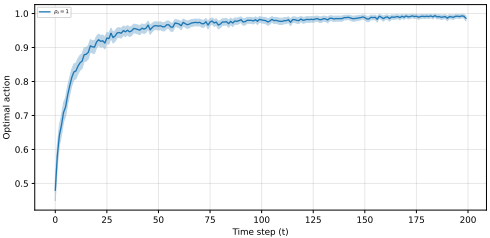
Detection metrics: regret and optimal-arm frequency, 95% CI over 1000 runs each.

All start from $H_0 = (0, \ldots, 0)$ (uniform).

Results: 2-arm toy problem

Toy 2 arms : $k = 2$; Means $(0, 0)$; Variances $(1, 2)$

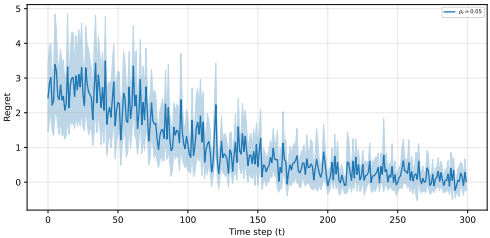


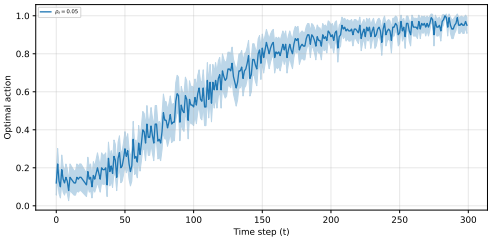


Regret $\to 0$ and optimal-arm frequency $\to 100\%$ in just 200 steps ($\rho = 0.5$, 1000 runs).
Clean separation ($\sigma^2 = 1$ vs 2) makes identification easy.

Results: 10-arm toy problem

Toy 10 arms : $k = 10$; means: $(0, \ldots, 0)$; $\sigma_a = \begin{cases} \sqrt{4} & a < 10 \\ 1 & a = 10 \end{cases}$





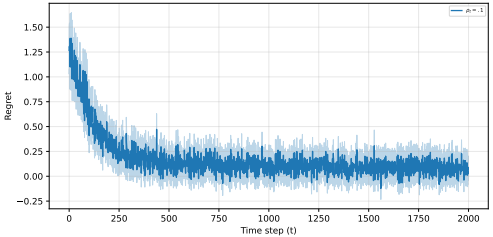
Again regret $\to 0$ and frequency $\to 100\%$ in 300 steps ($\rho = 0.05$, 1000 runs).

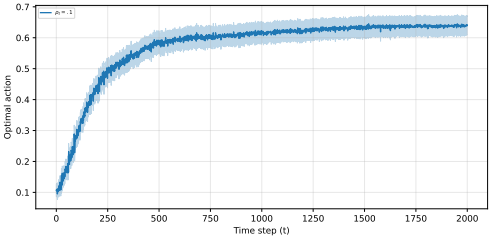
Note: the 9 sub-optimal arms $a = 1, ..., 9$ share same variance ($\sigma^2 = 4$); only the best arm $a = 10$ is separated, which is sufficient.

Results: hard 10-arm random separation

Sutton–Barto design: $\mu_a \sim \mathcal{N}(4, 1)$, $\sigma_a^2 \sim U[1, 5]$ (many near-ties).

$\rho = 0.1$, 2000 steps, 1000 runs.





Regret stays very low despite frequency plateauing around 70%. The near-ties cannot always be resolved, but the selected arm's variance is still close to optimal.

Outline

Summary and future work

What we did: risk-aware MAB where the target is the lowest-variance arm (or a mean–variance trade-off).

Key contribution: softmax policy-gradient algorithm with an unbiased variance estimator from 2 draws, no stored statistics, proven almost-sure convergence under mild assumptions.

Numerics: all three test cases confirm theory – toy examples converge to 100% optimal; the hard case shows that even when there are many near-ties limit regret remains low.

Future work:

- Contextual / time-dependent bandits and full RL extension.
- Variance-reduction techniques for sample efficiency.
- Convergence analysis for heavy-tailed distributions (relax boundedness).

References I

Ștefana-Lucia Aniță and Gabriel Turinici.
Convergence of a L2 Regularized Policy Gradient Algorithm for the Multi Armed Bandit.
In *Pattern Recognition*, pages 407–422, Cham, 2025. Springer Nature Switzerland.

Jincheng Mei, Zixin Zhong, Bo Dai, Alekh Agarwal, Csaba Szepesvari, and Dale Schuurmans.
Stochastic gradient succeeds for bandits.
In *International Conference on Machine Learning*, pages 24325–24360. PMLR, 2023.