Regularized Policy Gradient Algorithm for the Multi Armed Bandit

Gabriel Turinici
joint work with Ștefana-Lucia Anița

CEREMADE Université Paris Dauphine - PSL, Paris, France

Numerical Analysis, Numerical Modeling, Approximation Theory
(NA-NM-AT 2025, Nov. 3-6)

Introduction to Multi-Armed Bandits

Definition: The Multi-Armed Bandit (MAB) problem is a classic framework in **reinforcement learning**, where:

- An agent repeatedly chooses among multiple actions (e.g., pulling different slot machine arms), each giving some reward (random variable).
- The goal is to maximize the total reward over time.
- A balance must be struck between **exploration** (trying unknown actions) and **exploitation** (choosing the best-known action this far).

Multiple Applications:

- Optimizing online ad placements or A/B testing.
- Choosing the **best (happy hour) drink** in a bar without spending too much money.
- Allocating research funds to promising areas for maximum impact.

Softmax Policy Gradient MAB with $L2$ Regularization

Notations: The MAB has k arms (choices), each choice $a \leq k$ outcome is stochastic with a reward distribution of mean $q_*(a)$. At each time t, the agent selects an arm A_t and observes a reward R_t sampled from the distribution of the arm A_t. The objective is to maximize the cumulative reward.

Softmax Policy: The agent maintains a preference vector $H \in \mathbb{R}^k$, defining the probability of selecting arm A through: $\Pi_H(A) = \frac{e^{H(A)}}{\sum_{a=1}^k e^{H(a)}}$.

Our version : regularized 'loss': goal is to maximize: $\mathcal{L}_\gamma(H) = \mathbb{E}_{A \sim \Pi_H}\left[R(A) - \frac{\gamma}{2}\|\mathbf{H}\|^2\right]$, where $\gamma > 0$ is the $L2$ regularization coefficient.

Optimization algorithm : Policy gradient, a specific variant of Stochastic Gradient Ascent:

$$H_{t+1}(a) = H_t(a) + \rho_t\left[(R_t - \bar{R}_t)(\mathbb{1}_{a=A_t} - \Pi_{H_t}(a)) - \gamma H_t(a)\right],$$

with ρ_t = learning rate, $\bar{R}_t$ = average reward so far.

Theoretical results : fixed or variable ρ_t, large γ

Proposition (Convergence conditions, fixed or variable ρ_t, large γ)

Assume $\mu := \gamma - (\max_a q_*(a) - \min_a q_*(a)) > 0$. *Under appropriate hypotheses on distributions of arms A, there exists a unique optimum* H_*. *Moreover if*

$$\rho_t \to 0 \text{ and } \sum_{t \geq 1} \rho_t = \infty. \tag{1}$$

then

$$\lim_{t \to \infty} H_t \stackrel{\mathrm{L}^2}{=} H_*. \tag{2}$$

• Rq: many other results available in literature, cf. [2] and related works, but asymptotic & no regularization.

Theoretical results : "linear decay schedule" and small γ

Proposition (Convergence rate for "linear decay schedule")

Let $\beta_1, \beta_2 > 0$ two positive constants and take $\rho_t = \frac{\beta_1}{1+\beta_2 t}$.Under the same hypotheses as before for γ large enough : there is a unique solution H_* and the L2 regularized policy gradient MAB algorithm converges with the rate :

$$\mathbb{E}[\|H_t - H_*\|^2] = O\left(\frac{1}{t}\right) \text{ as } t \to \infty. \tag{3}$$

Question : optimum H^* will depend on γ, may pose problems, what about small γ ?

Lemma

Let $V(\gamma) := \max_{H\in\mathbb{R}^k} \mathcal{L}_\gamma(H)$. Then $\lim_{\gamma\to 0} V(\gamma) = V(0)$.

Numerical Simulations: Setup

Experiment Design (as in [3]):

- $M = 1000$ tests of 2000 steps each.
- $k = 10$ arms with rewards $R(A)$ normally distributed: $R(A) \sim \mathcal{N}(q_*(A), 1)$, where $q_*(A)$ has mean 4 and unit variance.
- Initial distributions tested:
  - **Uniform**: Π_{H_0} with $H_0 = (0, \dots, 0)$.
  - **Biased**: Π_{H_0} with $H_0 = (5, \dots, 0)$.

Plot Reward Normalization:

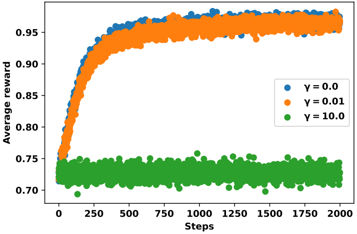
- Reward is scaled relative to the maximum possible reward:

$$R_{\text{scaled}} = \frac{R}{\max_a q_*(a)} \quad (\text{best reward} = 1).$$

Parameter Testing:

- Regularization coefficient: $\gamma \in \{0, 0.01, 10\}$.
- Learning rate: $\rho_t = 0.05$ or $\rho_t = \frac{1}{1+0.05t}$.

Results with constant learning step



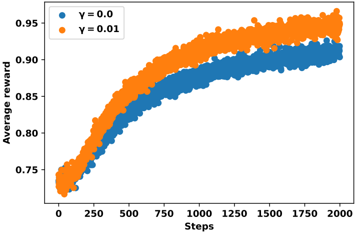


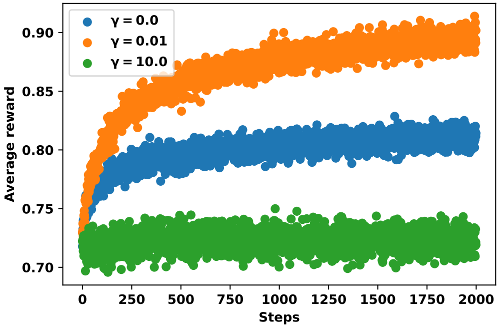
Figure: The average reward for $\rho_t = 0.05$ **(constant)**, γ is 0, 0.01 or 10 (see the legend). **Left :** start from a uniform distribution Π_{H_0} with $H_0 = (0, ..., 0)$. **Right :** start from a biased distribution Π_{H_0} with $H_0 = (5, ..., 0)$.

Uniform Start ($H_0 = (0, \ldots, 0)$):

- $\gamma = 0.01$ achieves comparable performance to $\gamma = 0$.
- $\gamma = 10$ biases the solution and reduces performance.

Biased Start ($H_0 = (5, \ldots, 0)$):

- $\gamma = 0.01$ improves performance over $\gamma = 0$.
- $\gamma = 10$ remains suboptimal.



- **Biased Start ($H_0 = (5, \ldots, 0)$):**
- Learning rate: $\rho_t = \frac{1}{1+0.05t}$ (linear decay schedule).
- Comparison of $\gamma = 0$ or 0.01 or 10.

Observations:

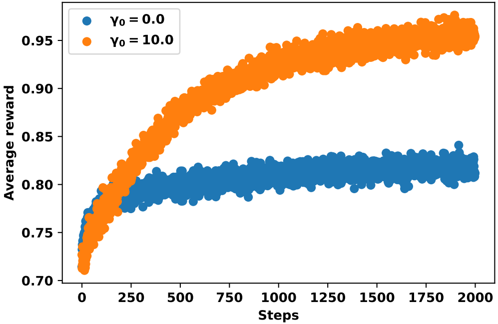
- $\gamma = 0.01$ achieves good performance, improving initial convergence.
- $\gamma = 10$ is too large, leading to suboptimal results.
- Decay of ρ_t helps transition from initial exploration to final convergence.

Empirical results so far:

- Convergence occurs as predicted by the theory.
- For non-uniform initial guesses H_0: regularization i.e., $\gamma > 0$ improves convergence significantly.
- large γ may lead to suboptimal limit points H^*.

Further idea: best of the two worlds : variable $\rho_t = \frac{1}{1+0.05*t}$ **and** $\gamma_t = \frac{\gamma_0}{1+0.2 \cdot t}$:

- Combines exploration and final convergence.
- $\gamma_0 > 0$ (orange in figure) outperforms non-regularized case $\gamma_0 = 0$ (blue).



Theoretical results : decreasing γ

Proposition (Convergence conditions, decreasing γ, SA & GT 2025)

Assume $\gamma_t \downarrow \bar{\gamma}$ and ρ_t such that it $\sum_t \rho_t = \infty$ and $\sum_t \rho_t^2 < \infty$. Under appropriate hypotheses on the distributions of arms A:

$$\sum_{t \geq 0} \rho_t \mathbb{E}[\|\nabla_H \mathcal{L}_{\gamma_t}(H_t)\|^2] < \infty \tag{4}$$

and therefore on a sub-sequence:

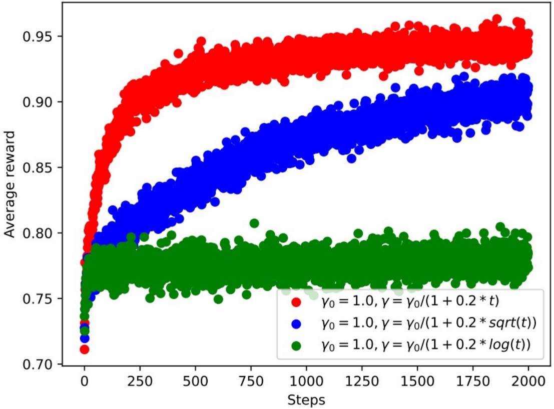
$$\lim_{\ell \to \infty} \nabla_H \mathcal{L}_{\gamma_{t_\ell}}(H_{t_\ell}) \stackrel{\mathrm{L}^2}{=} 0. \tag{5}$$

Remark

Work in progress: alternative results (asymptotic) for constant $\rho_t = \rho$.

Further numerical tests: γ regimes

Non-constant γ_t : we test several decay schedules: 'linear', logarithmic, square root... Not all behave the same, the 'linear' regime seems to be best for this parameter set.



Summary of theoretical and numerical results

Objective: we investigate a $L2$-regularized policy gradient algorithm for Multi-Armed Bandit (MAB).

Theoretical Results [1]:

- **Proposition 1:** Convergence established for both constant and variable step sizes (ρ_t).
- **Proposition 2:** Convergence rate proven to be $\mathcal{O}(1/t)$ for linear decay of ρ_t.
- **Lemma 1:** Regularization may shift the optimum but optimality is restored as $\gamma \to 0$.

Key Takeaway (theoretical and numerical): regularization helps improve convergence especially when starting far from optimality; it can be adjusted dynamically if needed.

Future Work: The convergence result for a decaying γ_t needs to be improved (optimality on the full sequence).

References I

Ștefana-Lucia Anița and Gabriel Turinici.
Convergence of a L2 regularized policy gradient algorithm for the multi armed bandit.
In *International Conference on Pattern Recognition*, pages 407–422. Springer, 2024.

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.

Richard S Sutton, Andrew G Barto, et al.
Reinforcement learning: An introduction, volume 1.
MIT press Cambridge, 1998.