CS 748: Final Report
Risk-Aversion in Infinite-Armed Bandits

Kartik Gokhale
IIT Bombay
200100083@iitb.ac.in

Sarthak Mittal
IIT Bombay
200050129@iitb.ac.in

Harshvardhan Agarwal
IIT Bombay
200050050@iitb.ac.in

Abstract

Risk is an important factor in reinforcement learning, especially in tasks involving potential hazards. We seek to tackle risk in the multi-armed bandit problem model of the exploration-exploitation tradeoff of reinforcement learning. We propose a variant of the multi-armed bandit problem to consider risk in the short-term as well, which is the minimisation of risk in a small or insufficient horizon. We analyse the shortcomings of existing algorithms in our proposed setting and propose four novel algorithms under various assumptions which experimentally perform better than existing algorithms. Finally, we also prove a theoretical guarantee of a sub-linear quantile regret by our best algorithm, ExpExpSS.

1 Introduction

Risk plays a big role in decision-making and must be considered in high-stake applications such as driving, robotic surgery, and finance (1). The goal in risk-averse Reinforcement Learning is to avoid hazardous areas (2).

There have been several studies on the optimization of risk in reinforcement learning. According to Garcia et al., (3) and Gabriel et al.(4), enforcing safety is important for decision-making and therefore, certain guarantees need to be provided. An inherent risk associated with the exploration phase must be minimized. In the past, various criteria have been used, including distortion risk measures by Vijayan et al.(5), entropic risk measure by Borkar et al.(6),(7) and Fei et al.(8), which depends on the risk aversion of the user through the exponential utility function, and mean-variance by Sato et al.(9), Prashanth et al.(10),(11) and Xie et al.(12), which is the process of weighing risk expressed as variance against expected return.

This is particularly relevant during the exploration phase as there is always a tradeoff between risk and exploration, which is the decision between choosing a familiar option with a known reward value or choosing an unfamiliar option with an unknown or uncertain reward value, which has an inherent risk associated with it. We seek to understand and handle risk in the context of the multi-armed bandit problem.

The multi-armed bandits model the exploration-exploitation dilemma of reinforcement learning well. Amongst many variants of the bandit problem, there exist variants that incorporate risk-return trade-off(13), infinite arms(14), and quantile cumulative regret(15). We seek to formulate and solve one such variation of the multi-armed bandit problem, one that incorporates risk. Additionally, we do not seek to minimise risk only in the long-term, where the horizon tends to infinity but also in the short-term where the horizon is finite and comparable to the number of arms. Broadly speaking, we target solving the multi-armed bandit problem while striking a balance between risk and returns in the scenario of an insufficient horizon (comparable to the number of arms).

In this paper, we begin by stating the relevant related works and our exact problem statement mathematically. Following this, we propose four new algorithms, which were able to tackle the

Preprint. Under review.

various sub-problems we faced while solving this problem, each obtaining marginal improvements towards our end goal. We then show the performance of our algorithms against these baseline algorithms and attempt to obtain formal guarantees regarding the performance of the best of these algorithms.

2 Notation

This section introduces the notation for the expression of the proposed algorithms and the following theoretical analysis.

We consider a setting of infinite armed bandits with a finite horizon T. The arms are characterised by a distribution v_i bounded in the interval [0,1] with mean μ_i and variance σ_i^2. Following the approach by Sani et al.(13), we consider the mean-variance model of arms, which accommodates expected reward and risk modelled by variance, given as

$$MV_i = \sigma_i^2 - \gamma\mu_i \tag{2.1}$$

where γ is the coefficient of absolute risk tolerance.

For an algorithm A, Z_s denotes the s-th random sample obtained from the pulled distribution v_i of arm i pulled according to the algorithm. The empirical mean and variance of A upto intermediate horizon t will be given by

$$\hat{\mu}_t(A) = \frac{1}{t}\sum_{s=1}^{t} Z_s \tag{2.2}$$

$$\hat{\sigma}_t(A) = \frac{1}{t}\sum_{s=1}^{t}(Z_s - \hat{\mu}_t)^2 \tag{2.3}$$

We propose a similar definition of the mean-variance of the algorithm based on empirical statistics

$$\widehat{MV}_t(A) = \hat{\sigma}_t^2(A) - \gamma\hat{\mu}_t(A) \tag{2.4}$$

3 Problem Statement

We formulate our problem mathematically with quantile mean-variance regret given by

$$R_{\rho,T}(A) = \widehat{MV}_T(A) - MV_{i(\rho)} \tag{3.1}$$

where $i(\rho)$ represents the arm with $(1-\rho)$th highest mean variance value.

We want to find an algorithm A which is able to incur a reward $R_{\rho,T}(A)$ whose expected value decays to zero as T increases.

$$\lim_{T\to\infty} E[R_{\rho,T}(A)] = 0 \tag{3.2}$$

4 Proposed Algorithms

In this section, we detail the four new algorithms for the task of risk aversion in multi-armed bandits.

4.1 NewAlgo0

In this algorithm, we assume that the distribution from which the parameters θ_k, representing the distribution of each bandit arm, are sampled, is known. Thus, NewAlgo0 maintains a belief about the quality of each arm and then samples an arm if it lies in the top k-quantile of arms, which is calculated from the distribution from which the parameters are sampled. Since the distribution of θ_k is known, we define the *margin* to be the hyperplane in the parameter space separating the parameters resulting in the top k quantile of arms, in terms of minimising the mean-variance regret as defined by in Equation (2.1)

Algorithm 1: NewAlgo0

```
while $t \leq T$ do
    if min(MVlist) < margin then
        index ← argmin(MVlist)
    else
        index ← random(0,N)
    end
    reward ← pull(index); R ← R + reward
    $\mu$ ← (N + $\mu$ + reward)/(N + 1); $\nu$ ← N * ($\nu$ + ($\mu$ - reward)$^2$/(N + 1))/(N + 1)
    MVlist ← [MV($\mu_i$,$\nu_i$,$\gamma$) for $i$ in range(N)]
end
```

4.2 NewAlgo1

The fatal flaw that newAlgo0 faces lie in the update of its belief. We experimentally observed by running experiments that, in the cumulative regret scheme, wherein we consider only the mean, deviation of the empirical mean above the true mean can only be beneficial in the short run. This means that if the empirical mean of the arm is performing better than the expected value, the empirical reward that we have obtained is beneficial as it reduces the empirical regret. However, this is not true when we consider risk, as underestimating the variance can lead us to overestimate the empirical 'quality' of the arm, without it minimising regret. Thus, NewAlgo1 updates belief only after it has received a batch of samples. Thus, the deviation of these updates from the true value of the arm is lesser, on an expectation, than if we were to update the belief after every feedback.

Algorithm 2: NewAlgo1

```
while $t \leq T$ do
    if pulls[index] < 4 then
        pulls[index] ← pulls[index] + 1
    else
        if min(MVlist) < margin then
            index ← argmin(MVlist)
        else
            index ← random(0,N)
        end
        pulls[index] ← 1
    end
    reward ← pull(index); R ← R + reward
    $\mu$ ← (N + $\mu$ + reward)/(N + 1); $\nu$ ← N * ($\nu$ + ($\mu$ - reward)$^2$/(N + 1))/(N + 1)
    MVlist ← [MV($\mu_i$,$\nu_i$,$\gamma$) for $i$ in range(N) if pulls[$i$] $>$ 0 else margin]
end
```

4.3 NewAlgo2

In this algorithm, we assume that the distribution from which the parameters θ_k parameterise every arm is unknown. The goal of this algorithm is to estimate both the distribution from which the parameters of the arms are sampled from as well as the empirical mean-variance values of each of the arms. Estimating the distribution allows us to estimate the quantile at which each arm lies at. Thus, newAlgo2 uses exploration-exploitation on the 'margin' defined above in newAlgo1, separating 'good' arms from 'bad'. Then, with our empirical estimate of the margin, we use newAlgo1 as a black box for a fraction of the horizon operating on a margin, whose belief is updated regularly.

Algorithm 3: NewAlgo2

```
while $t \leq T$ do
    if $t$ % $(T/10)$ $is$ 0 then
        MVlist ← [MV($\mu_i$,$\nu_i$,$\gamma$) for $i$ in range(N) if pulls[$i$] $>$ 0]
        margin ← (1 - $q$)$^{th}$ quantile of MVlist
    end
    if pulls[index] < 4 then
        pulls[index] ← pulls[index] + 1
    else
        if min(MVlist) < margin then
            index ← argmin(MVlist)
        else
            index ← random(0,N)
        end
        pulls[index] ← 1
    end
    reward ← pull(index); R ← R + reward
    $\mu$ ← (N + $\mu$ + reward)/(N + 1); $\nu$ ← N * ($\nu$ + ($\mu$ - reward)$^2$/(N + 1))/(N + 1)
    MVlist ← [MV($\mu_i$,$\nu_i$,$\gamma$) for $i$ in range(N) if pulls[$i$] $>$ 0 else margin]
end
```

4.4 ExpExpSS

This algorithm subsamples a finite set of arms from the infinite distribution of arms and then employs ExpExp (13) to achieve sublinear regret on the selected arms. The sampling ensures that at least one good arm is picked up with high probability. In the exploration phase, the algorithm pulls a random arm from the selected set and updates the empirical MV value. Later in the exploitation phase, it pulls the arm with the minimum MV value.

Algorithm 4: ExpExpSS

```
$C \leftarrow 1$
$\alpha \leftarrow 0.3$
$K \leftarrow \frac{C}{\rho}T^{1/3-\alpha}$
$\tau \leftarrow K(T/14)^{2/3}$
while $t \leq T$ do
    if $t \leq \tau$ then
        index ← random(0,K)
        reward ← pull(index); R ← R + reward
        $\mu$ ← (N + $\mu$ + reward)/(N + 1); $\nu$ ← N * ($\nu$ + ($\mu$ - reward)$^2$/(N + 1))/(N + 1)
        MVlist ← [MV($\mu_i$,$\nu_i$,$\gamma$) for $i$ in range(N)]
    else
        index ← argmin(MVlist)
        reward ← pull(index)
    end
end
```

5 Theoretical Analysis

This section presents the main theorem providing the required guarantee on sub-linear regret for ExpExpSS.

Theorem 1 *For a given value of quantile factor ρ and sufficiently large values of T, the algorithm ExpExpSS achieves expected quantile regret bounded as follows where $\alpha < \frac{1}{3}$.*

$$E[R_{\rho,T}(A)] = O\left(\frac{1}{T^{\alpha}}\right) \tag{5.1}$$

The theorem directly implies that ExpExp achieves sublinear expected quantile regret. The proof can be found in Appendix A.

6 Experimental Analysis

In this section, we define our underlying bandit model, which we propose to solve along with the relevant assumptions.

Our bandit model comprises of Gaussian bandit arms, where each arm a_k is parameterized by $\theta_k = (\mu_k, \sigma_k^2)$ representing the mean and variance of the k-th arm respectively. Sampling from the Gaussian bandit arm a_k is equivalent to sampling a Gaussian distribution with parameters θ_k. The value obtained after sampling is treated as the reward. We start off by stating the hyperparameters used in the experiments.

- The means $\{\mu_k\}_{k=1}^K$ lie between some bounds, taken to be, for convenience, as 0 and 1, and the variances $\{\sigma_k^2\}_{k=1}^K$ are upper bounded by a known quantity, $M_{\sigma^2} = 1/16$.
- The underlying distribution from which the distribution parameters are sampled is known. This assumption is only for 'NewAlgo1'.
- We use the factor η to model the ratio between the number of arms and the horizon. We consider η to be comparable to 1, as we are considering an insufficient horizon.

Exact details about the running environment can be found in Appendix B.

7 Results

This section presents the results of newAlgo1, newAlgo2 and ExpExpSS algorithms. We consider the risk aversion setting for all the algorithms being compared.

We proceed with the analysis in three stages:

1. Comparison with baselines: Our baseline algorithms, for now, are (i) MVLCB: Modification of UCB for risk aversion, and (ii) ExpExp: An exploration-exploitation algorithm.
2. Performance with known distribution: Calculating the mean regret for different values of the horizon. We also vary the value of the parameter "eta" which is the ratio of the number of arms to the horizon.
3. Performance with unknown distribution: Here we use assume that the underlying distribution for the mean and variance of the Gaussian arms is unknown and run the algorithms.

The plots where $\rho = 0$ correspond to "variance minimization", and there is no contribution from the arm means. On the other hand, where $\rho = 1$, there is some weight given to variance along with the problem being "mean maximization", and the amount of impact the variance has is limited due to the assumed bound.

7.1 newAlgo1

As discussed, newAlgo0 seeks to obtain a **margin** separating 'good' arms from 'bad' arms. The margin can be treated as an underlying property of the underlying distribution. Though, it reduces into a mean maximisation problem; ignoring risk altogether. So we need to do better; only switch arms once the confidence in the empirical value crosses a threshold.

The results of newAlgo1 indicate that it is able to perform fairly well even when the number of arms is comparable to the horizon (infinite arms/finite horizon case). The comparison with baselines shows that newAlgo1 is able to consistently perform better than them. In the case where $\rho = 1$, ExpExp is able to incorporate variance to some extent because of the exploration and performs much better than MVLCB itself.

When the underlying distribution for the means and variances of the arms was assumed to be unknown (in this case Beta distributions with different parameters were used), newAlgo1 performed similarly to ExpExp when only variance was considered, but failed to beat ExpExp when both were considered.

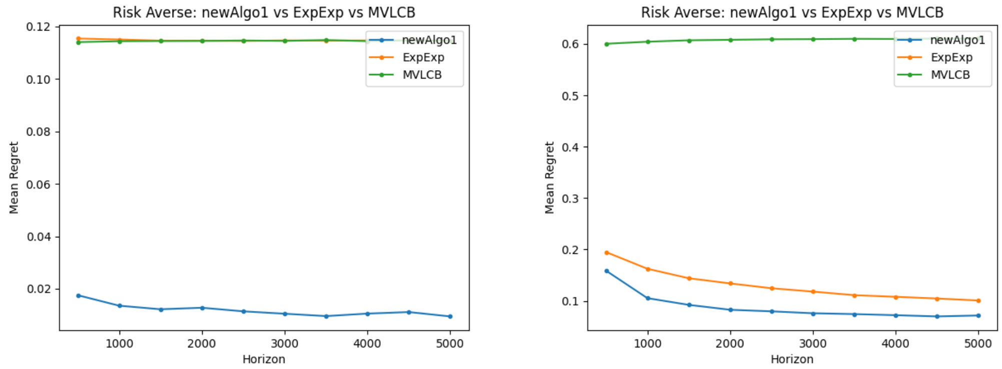

Figure 1: Comparison with $\rho = 0$ and $\rho = 1$

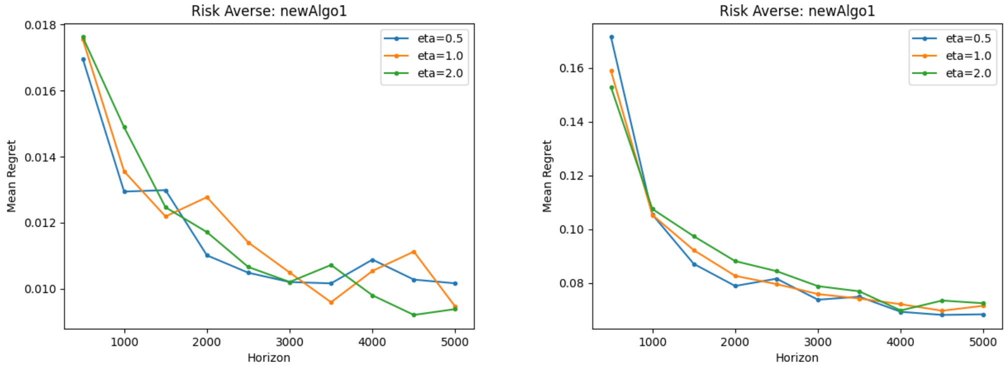

Figure 2: Performance with $\rho = 0$ and $\rho = 1$

7.2 newAlgo2

In newAlgo2, we use the preliminary algorithm as a module with an exploration-exploitation on the margin. We also introduce the idea of quantile regret as a metric of evaluation. Instead of calculating just the mean-variance regret, we use a modified regret formulation where the mean-variance reward is calculated, and then the quantile is used as the threshold for what can be considered a "good enough" arm for regret.

The results for newAlgo2 were quite similar to newAlgo1 in the case of comparison with baselines and variation with the ratio of the number of arms to the horizon. The differences we observed were that newAlgo2 was able to achieve sub-linear quantile regret along with risk aversion. However, both newAlgo1 and newAlgo2 were unable to beat ExpExp in the case where mean and variance were considered ($\rho = 1$).

We observed significant variations when only variance was considered, and the underlying distribution for mean and variance was assumed to be unknown. Though newAlgo2 could perform better than the baselines, the convergence was not clearly visible. When both mean and variance were considered, newAlgo2 and ExpExp achieved sub-linear quantile regret, but newAlgo2 could not beat ExpExp.

7.3 ExpExpSS

ExpExpSS also achieved sub-linear regret with variation in the number of arms to horizon ratio. With ExpExpSS, the results for the case where only the variance was considered were much more stable, and ExpExpSS was able to achieve sub-linear quantile regret while the other algorithms showed random fluctuations. When both mean and variance were considered, ExpExpSS performed comparable to newAlgo2 and ExpExp but could not beat them.

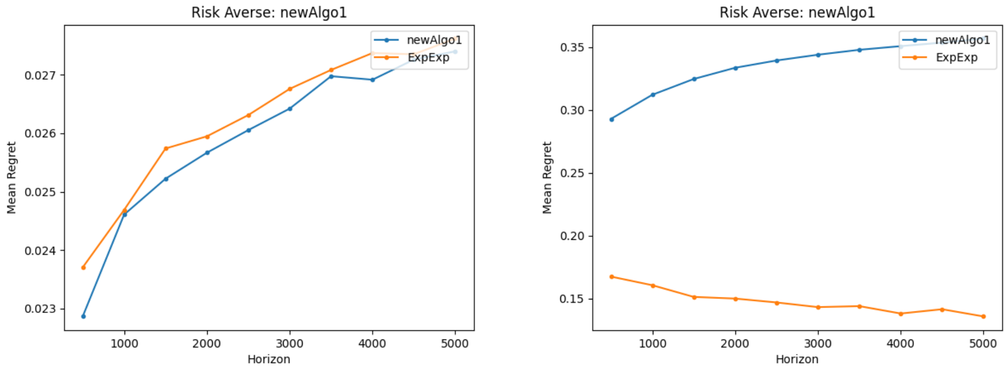

Figure 3: Performance of NewAlgo1 when distribution was unknown. Here, we consider the means to be distributed as $\beta(4, 20)$ and variances to be distributed as $\beta(20, 4)/16$ with $\rho = 0$ and $\rho = 1$

We are comparing ExpExpSS in situations where NewAlgo2 is unable to perform better than ExpExp. We observe that in the variance minimisation problem where $\rho = 0$, ExpExpSS performs better by a large margin, consistent with Theorem 1.

8 Related Works

In this section, we briefly discuss the related variants of the multi-armed bandit problem.

8.1 Risk-Aversion

This variant changes the objective from maximizing total reward to obtaining the best risk-return trade-off. The difficulty posed by this is that there is an exploration risk, which introduces a regret associated with the variability of an algorithm. This approach could be useful because the best method on average could have more variability in some applications. A less but more consistently effective method could be preferable in this case.

One such approach by Sani et al.(13) involves modifying the objective function to

$$R_T(A) = MV_A - MV^*, MV_i = \sigma_i - \rho\mu_i \tag{8.1}$$

where σ_i is the variance of arm I and μ_i is the mean of arm I. The objective is to design an algorithm whose regret decreases as the number of rounds increases.

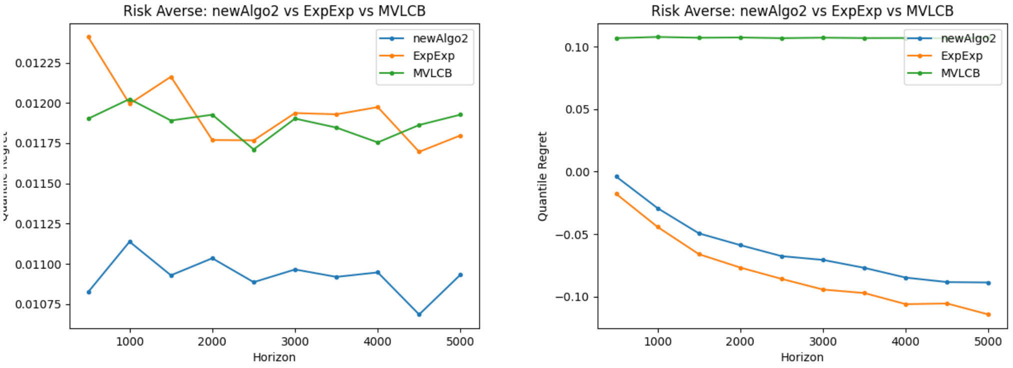

Figure 4: Performance on mean distribution $\beta(4, 20)$ and variance distribution $\beta(20, 4)/16$ with $\rho = 0$ and $\rho = 1$

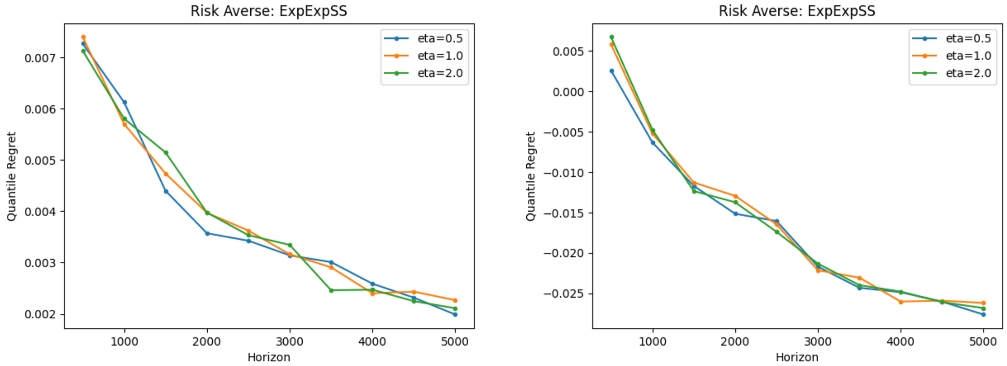

Figure 5: Performance on mean distribution $\beta(4, 20)$ and variance distribution $\beta(20, 4)/16$ with $\rho = 0$ and $\rho = 1$

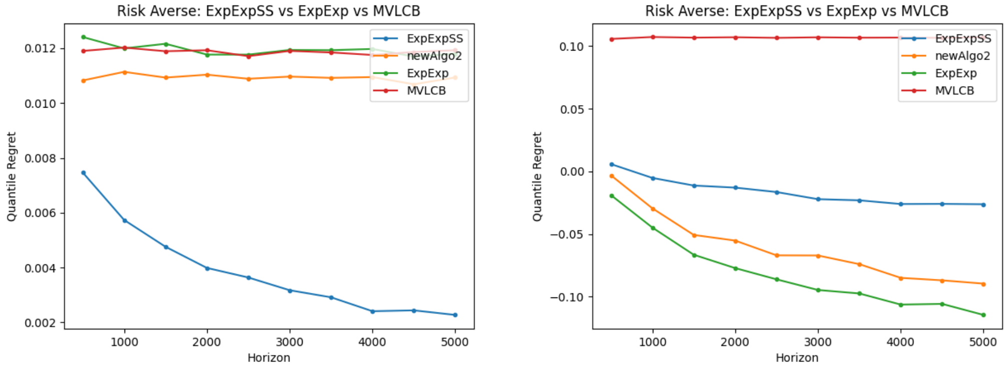

Figure 6: Performance on mean distribution $\beta(4, 20)$ and variance distribution $\beta(20, 4)/16$ with $\rho = 0$ and $\rho = 1$

For instance, consider two Gaussian arms with $\mu_1 = 0.5, \sigma_1^2 = 0.5$ and $\mu_2 = 0.4, \sigma_2^2 = 0$. In such a case, though arm one has a higher mean, the mean-variance of arm two is lesser and thus, by the above definition, becomes the optimal arm. This is because the first arm has a higher variance; thus, for a finite set of pulls, the guarantee (lower bound for regret with high probability) is higher for arm two than for arm one. This is how the mean-variance definition for regret incorporates risk aversion.

The algorithms MVLCB and ExpExp do not perform well when the arms are comparable to the horizon, as they cannot sample each arm adequately.

8.2 Infinite-Armed

This variant, introduced by Agrawal(14), relaxes the condition that the number of arms is finite. In the infinite armed case, the "arms" are a continuous variable in K-dimensional space. The arms are chosen from a subset of the real line, and the mean rewards are a continuous function of the arms. The difficulty this problem poses is that the learning task becomes infinite-dimensional. The original approach used a kernel estimator-based learning scheme for the mean rewards as a function of the arms.

An approach by Kalyanakrishnan et al.(15) introduces quantile cumulative regret. To focus on pulling "good arms", they define cumulative quantile regret as

$$R_T(\rho) = T\mu_\rho - \sum_i r_t \tag{8.2}$$

where μ_ρ is the infimum of the ρ-fraction of the arm means. The algorithm QRM1 incurs sub-linear regret with respect to horizon T given by

$$R_T(\rho) \in O(\rho^{-1} + \sqrt{(T/\rho)log(\rho T))}$$

Another version of the algorithm QRM2 achieves sub-linear regret for every ρ without requiring its value.

9 Conclusion

We feel this work will be relevant to situations in which the agent operates in an environment where the time horizon is comparable to the size of the action space. In such a setting, every 'bandit-arm' cannot be sampled an adequate number of times, as the horizon is inadequate. Additionally, if the rewards associated with the environment has high variance, the agent may need some guarantees on the reward obtained (managing "risk"). Our algorithm will help the agent to perform better in such an environment.

Our further plans include improving the explore-exploit tradeoff for newAlgo2 on the margin, such that we can obtain a sublinear regret guarantee on newAlgo2 as well. Furthermore, Reinforcement Learning based on ExpExpSS can be tested in the practical situations mentioned above, where risk must be considered and the horizon is insufficient.

References

[1] Vittori, E., M. Trapletti, M. Restelli. Option hedging with risk averse reinforcement learning, 2020.

[2] Greenberg, I., Y. Chow, M. Ghavamzadeh, et al. Efficient risk-averse reinforcement learning, 2022.

[3] García, J., F. Fernández. A comprehensive survey on safe reinforcement learning. *J. Mach. Learn. Res.*, 16:1437–1480, 2015.

[4] Dulac-Arnold, G., D. Mankowitz, T. Hester. Challenges of real-world reinforcement learning, 2019.

[5] Vijayan, N., P. L. A. Policy gradient methods for distortion risk measures, 2023.

[6] Borkar, V. S., S. P. Meyn. Risk-sensitive optimal control for markov decision processes with monotone cost. *Mathematics of Operations Research*, 27(1):192–209, 2002.

[7] Borkar, V., R. Jain. Risk-constrained markov decision processes. *IEEE Transactions on Automatic Control*, 59(9):2574–2579, 2014.

[8] Fei, Y., Z. Yang, Y. Chen, et al. Exponential bellman equation and improved regret bounds for risk-sensitive reinforcement learning, 2021.

[9] Sato, M., H. Kimura, S. Kobayashi. Td algorithm for the variance of return and mean-variance reinforcement learning. *Transactions of the Japanese Society for Artificial Intelligence*, 16(3):353–362, 2001. Copyright: Copyright 2011 Elsevier B.V., All rights reserved.

[10] L.A., P., M. Ghavamzadeh. Actor-critic algorithms for risk-sensitive mdps. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger, eds., *Advances in Neural Information Processing Systems*, vol. 26. Curran Associates, Inc., 2013.

[11] A., P. L., M. Ghavamzadeh. Variance-constrained actor-critic algorithms for discounted and average reward mdps, 2015.

[12] Liu, B., T. Xie, Y. Xu, et al. A block coordinate ascent algorithm for mean-variance optimization, 2018.

[13] Sani, A., A. Lazaric, R. Munos. Risk-aversion in multi-armed bandits. *NIPS*, 2013.

[14] Agrawal, R. The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 1995.

[15] Roy, C. A., K. Shivaram. Quantile-regret minimisation in infinitely many-armed bandits. 2018.

A Proof of Theorem 5.1

The objective of the problem is to minimize the following regret for algorithm A

$$R_{\rho,T}(A) = \widehat{MV}_T(A) - MV_{i(\rho)} \tag{A.1}$$

We select K arms from the distribution of infinite arms. With probability $(1-\rho)^K$, all the arms picked are worse that ρth best arm in which case the regret is bounded by some value B. In the other case, we can say $R_{\rho,T}(A) \leq R_T(A)$.

$$E[R_{\rho,T}(A)] \leq B(1-\rho)^K + E[R_T(A)] \tag{A.2}$$

The paper (14) introduces psuedo-regret which bounds true regret given by the inequality with probability $1-\delta$.

$$R_T(A) \leq \tilde{R_T}(A) + (5+\gamma)\sqrt{\frac{2Klog(6KT/\delta)}{T}} + 4\sqrt{2}\frac{Klog(6KT/\delta)}{T} \tag{A.3}$$

which can be modified to

$$E[R_T(A)] \leq E[\tilde{R_T}(A)] + (5+\gamma)\sqrt{\frac{2Klog(6KT/\delta)}{T}} + 4\sqrt{2}\frac{Klog(6KT/\delta)}{T} + B\delta \tag{A.4}$$

We can choose δ such that the last three terms are sub-linear. Select $\delta = \frac{Z}{T^\beta}$

We know that if ExpExp is employed with $\tau = K(T/14)^{2/3}$, then for any choice of distributions v_i, the expected psuedo regret is $E[\tilde{R_T}(A)] \leq 2\frac{K}{T^{1/3}}$. Select $K = \frac{C}{\rho}T^{1/3-\alpha}$. We know that τ should be less than equal to T which gives

$$\frac{C^3}{196\rho^3} \leq T^\alpha \tag{A.5}$$

Now, we have the following

$$\begin{aligned}
E[R_{\rho,T}(A)] &\leq B(1-\rho)^K + E[\tilde{R_T}(A)] + (5+\gamma)\sqrt{\frac{2Klog(6KT/\delta)}{T}} + 4\sqrt{2}\frac{Klog(6KT/\delta)}{T} + \frac{BZ}{T^\beta} \\
&\leq B(1-\rho)^K + 2\frac{K}{T^{1/3}} + (5+\gamma)\sqrt{\frac{2Klog(6KT/\delta)}{T}} + 4\sqrt{2}\frac{Klog(6KT/\delta)}{T} + \frac{BZ}{T^\beta} \\
&\leq Be^{-CT^{1/3-\alpha}} + \frac{2C}{T^\alpha} + (5+\gamma)\sqrt{\frac{2Clog(6KT^{1+\beta}/Z)}{\rho T^{2/3+\alpha}}} + 4\sqrt{2}\frac{Clog(6KT^{1+\beta}/Z)}{\rho T^{2/3+\alpha}} + \frac{BZ}{T^\beta}
\end{aligned}$$

If we take $\alpha < \frac{1}{3}$ and $\beta > \alpha$, then the inequality directly implies

$$E[R_{\rho,T}(A)] = O\left(\frac{1}{T^\alpha}\right)$$

This clearly shows that expected quantile regret is sub-linear in expectancy, given that T is sufficiently large.

B Experimental Analysis Details

Here, we present exact details related to our experimental analysis. For all the results detailed below, we operate our algorithm for 250 seeds and averaged out the results for consistent analysis. We vary our horizon to 5000 in intervals of 500 and plot out the corresponding regret achieved. In some of the plots, we vary the η ratio of the number of arms to the horizon.

The experimental setup considers Gaussian Arms, which subsample rewards from a Gaussian distribution with fixed mean and variance. The mean and variance of these arms are themselves subsampled from the two distributions, as detailed below.

- **Uniform Distribution**: Both mean and variance are sampled from a uniform distribution with mean ranging from 0 to 1 and variance ranging from 0 to $\frac{1}{16}$.
- **Beta Distribution**: Mean is sampled from $\beta(4, 20)$ and variance is sampled from $\frac{\beta(20,4)}{16}$

Since our aim is to develop an algorithm that performs "good" irrespective of the distribution of arms, we use a beta distribution which provides an alternative to the uniform distribution for testing purposes.

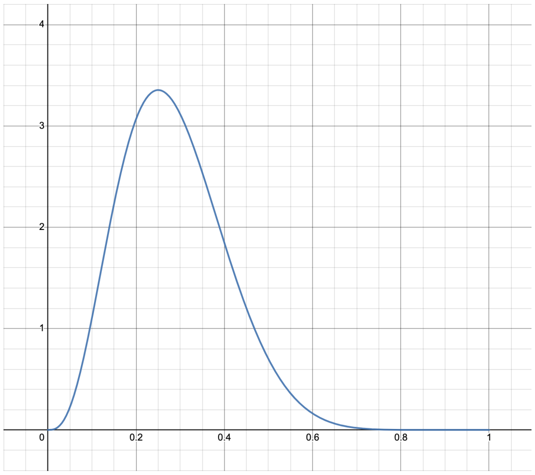

Figure 7: Distribution $\beta(4, 20)$

B.1 newAlgo0

For newAlgo0 we compared with the baseline implementations when the horizon was insufficient (we tried different values of η, which were close to 1). newAlgo0 relied on the fact that the arm means and variances were sampled from a uniform distribution. newAlgo0 was unable to achieve sub-linear regret with small horizons in the range of a few thousand when the underlying distribution was assumed to be unknown. newAlgo0 only used the risk-aversion parameter γ for simulation.

B.2 newAlgo1

In newAlgo1, we initialized the margin to a value resembling a hypothetical, optimal arm with a high mean and low variance. The primary change from the simulation point of view was that once an arm was pulled, it was pulled consecutively four times. This was done to ensure that variance calculation did not severely affect the margin updates. This change made newAlgo1 more stable when the underlying distribution was uniform. With just variance minimization, newAlgo1 achieved sub-linear regret with an insufficient horizon. Though the results with the unknown (taken as beta) distribution was still not good enough compared to ExpExp.

B.3 newAlgo2

Here we introduced two new parameters, ϵ and ρ, which were for deciding the exploration phase and the quantile fraction, respectively. For our experiments, we took both of them to be 0.1. We have also updated the margin on every one-tenth of the horizon and used the quantile fraction to decide the new margin (based on the MV values). For the starting ϵ fraction of the horizon, we randomly sampled arms to account for the loss in exploration due to pulling arms in sets of four consecutive pulls. The results of newAlgo2 were quite promising in all cases except when the distribution was unknown and mean-variance was considered.

B.4 ExpExpSS

For ExpExpSS, we introduce two hyperparameters - C and α, which can be varied to change the number of arms subsampled. We include a new variable τ which describes part of the horizon being used for exploration and is a function of the number of arms and horizon. The arms sampled (m) are given by $m = \lfloor \frac{C}{\rho} * n^{\frac{1}{3} - \alpha} + 1 \rfloor$ where n is the horizon. Also τ is set to $\tau = \lfloor \frac{n}{14}^{\frac{2}{3}} * m \rfloor$. The algorithm explores randomly in the exploration phase. At the end of exploration, it adheres to a particular arm and keeps sampling it for the rest of the exploitation phase.