Reinforcement Learning Lab Report

Assignment 2

Jayanth S (201081003)
Praveen Kumar N (201082001)
Rishabh Roy (201082002)

1 Bandit algorithms analysed

- ϵ-greedy
- Variable $\epsilon-$greedy
- Soft-max
- Upper confidence bound (UCB)
- Thompson sampling
- Reinforce without baseline
- Reinforce with baseline

2 Simulation settings

- Number of bandit arms considered is $K \in \{2, 5, 10\}$.
- Two reward distributions are analysed i.e,
  1. Bernoulli distribution
  2. Gaussian distribution
- Expected rewards of each arm is sampled from Uniform$(0, 1)$ distribution.
- Each algorithm is simulated for 20 times (runs) (for all values of K) and the performance metrics were averaged over 20 runs.
  1. For each run the expected rewards for each arm were uniformly sampled between 0 and 1.
  2. Each algorithm was run for 20000 time steps.

3 Bandits with Bernoulli reward distribution

Table 1: Performance of bandit algorithms with Bernoulli reward distribution

Algorithms	**Cumulative regret**		
	K = 2	K = 5	K = 10
ϵ-greedy ($\epsilon = 0.1$)	333.005	771.198	922.859
Variable ϵ greedy	77.520	678.512	1198.866
Soft-max ($\beta = 0.05$)	1156.521	1361.912	2233.716
UCB	86.359	247.374	540.835
Thompson sampling	7.526	49.618	45.003
Reinforce without baseline	31.879	162.975	393.057
Reinforce with baseline	36.959	247.870	279.704

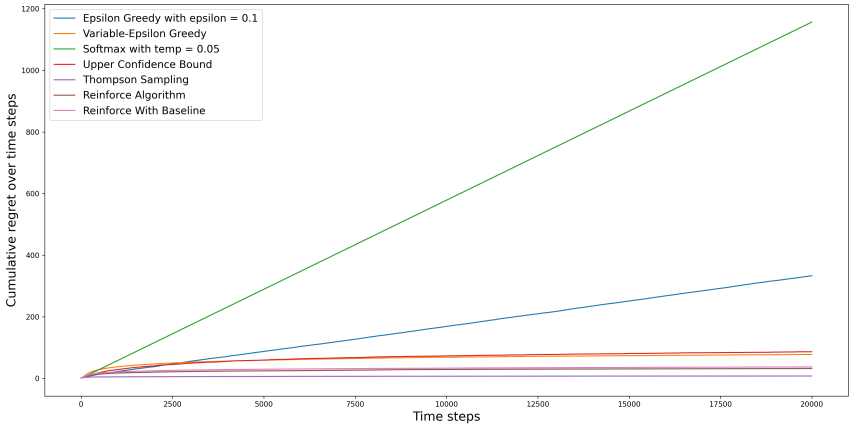

Figure 1: Variation of total regret with time steps for $K = 2$ with Bernoulli reward distribution

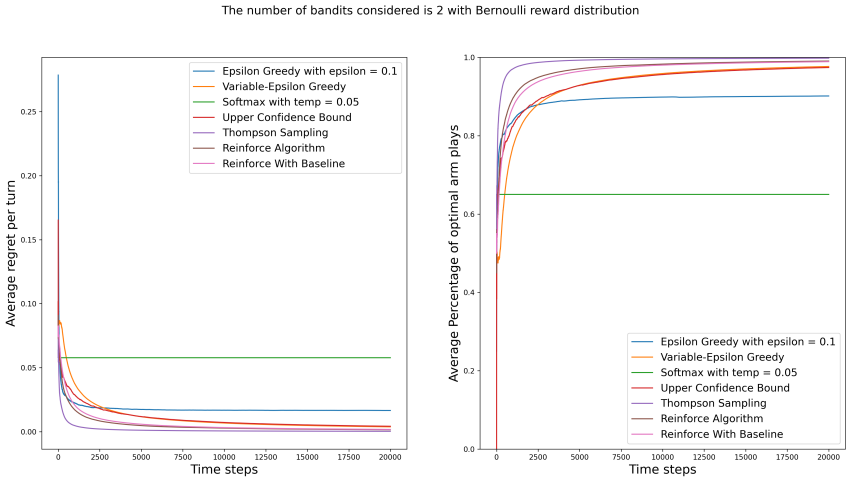

Figure 2: Variation of average regret per turn and average percent of optimal arm with time steps for $K = 2$ with Bernoulli reward distribution

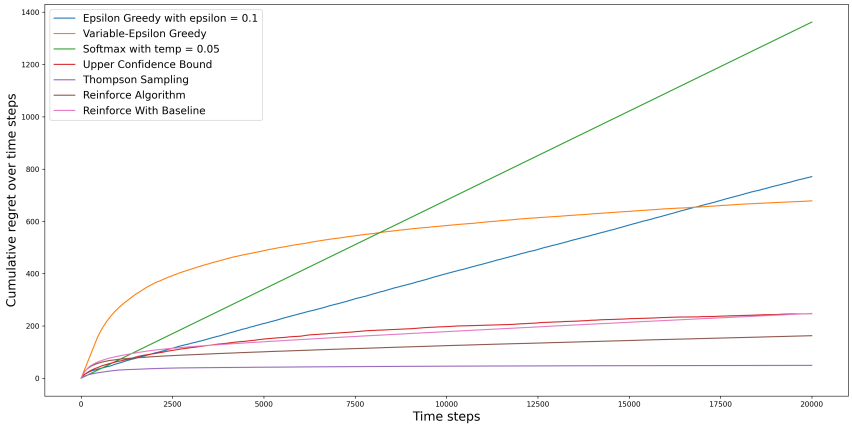

Figure 3: Variation of total regret with time steps for $K = 5$ with Bernoulli reward distribution

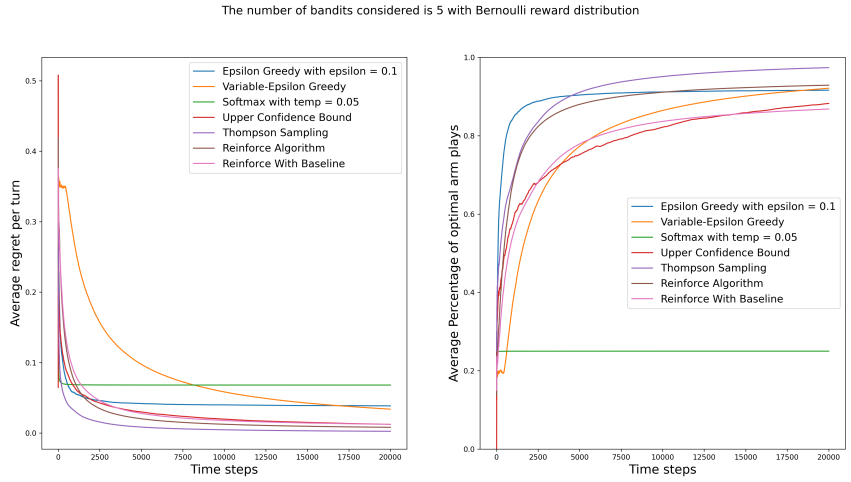

Figure 4: Variation of average regret per turn and average percent of optimal arm with time steps for $K = 5$ with Bernoulli reward distribution

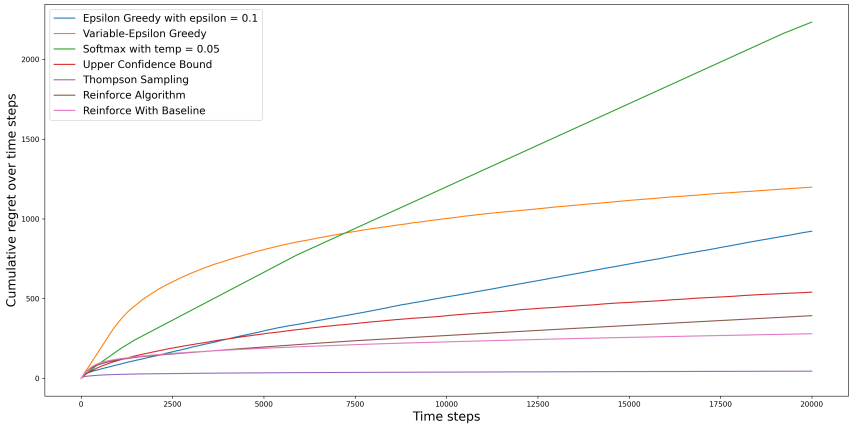

Figure 5: Variation of total regret with time steps for $K = 10$ with Bernoulli reward distribution

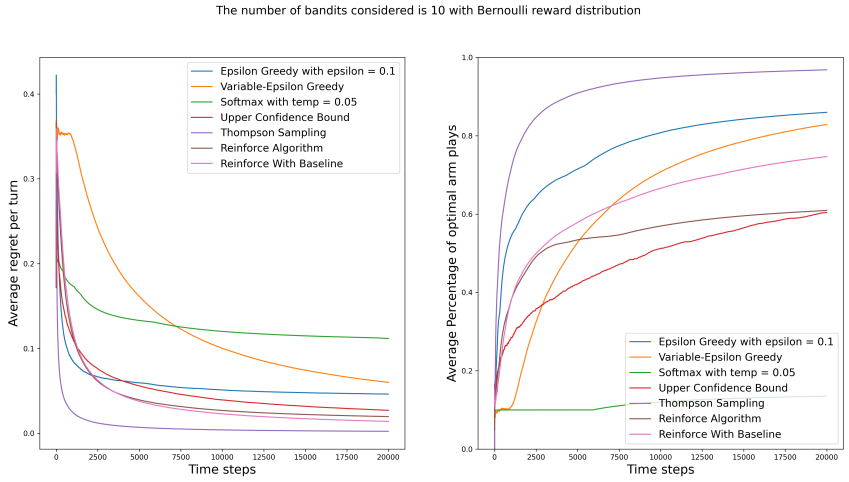

Figure 6: Variation of average regret per turn and average percent of optimal arm with time steps for $K = 10$ with Bernoulli reward distribution

3.1 Observations

- We observed that for bandits with Bernoulli reward distribution, irrespective of the number of arms **Thompson sampling** performed well.
- The rate of increase in average cumulative regret for $\epsilon-$greedy with $\epsilon = 0.1$ and Soft-max($\beta = 0.05$) algorithms was linear.
- Also for $\epsilon-$greedy and Soft-max the choice of hyper-parameter decides the performance of the algorithms.
- According to the obtained simulation results we rank the bandit algorithms in the following order : Thompson sampling >> Reinforce with and without baseline >> UCB $\approx$ Variable ϵ-greedy >> $\epsilon-$greedy $\approx$ Soft-max.
- We didn't observe a significant difference between reinforce algorithm with baseline and without baseline.
- We observed that, as the number of arms increases the average probability of picking the best arm will relatively decrease. A justification for

this is, since the average reward of each arm $\in [0, 1]$ with the increase in the number of arms the probability of the average rewards of different arms being close to each other increases. Hence the algorithms may not be able to differentiate the arms that are close to the optimal arm. However, the algorithms achieve sub-linear regret (or $O(log(T))$.

4 Bandits with Gaussian reward distribution

Here variance (σ^2) of all arms reward distribution is considered to be same. Algorithms are tested for three values of σ i.e, $0.01, 0.1$ and 1.

4.1 Bandits with Gaussian reward distribution with $\sigma = 0.01$

Table 2: Performance of algorithms with Gaussian reward distribution ($\sigma = 0.01$)

Algorithms	**Cumulative regret**		
	K = 2	K = 5	K = 10
ϵ-greedy ($\epsilon = 0.1$)	3332.102	6667.533	8190.580
Variable ϵ greedy	26.098	799.783	1013.742
Soft-max ($\beta = 0.05$)	733.375	2395.769	1472.530
UCB	138.481	200.309	628.3
Thompson sampling	3.479	14.445	31.685
Reinforce without baseline	37.031	182.564	660.266
Reinforce with baseline	34.981	104.587	247.971

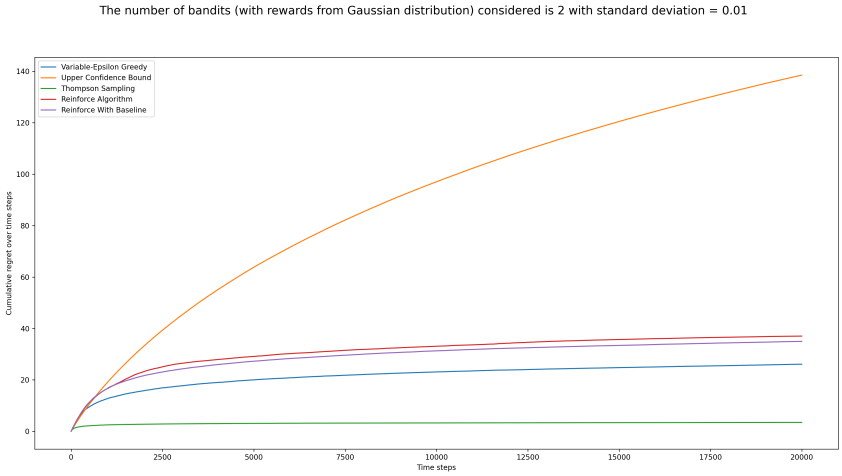

Figure 7: Variation of total regret with time steps for $K = 2$ with Gaussian reward distribution ($\sigma = 0.01$)

4.2 Bandits with Gaussian reward distribution with $\sigma = 0.1$

Table 3: Performance of algorithms for Gaussian reward distribution ($\sigma = 0.1$)

Algorithms	**Cumulative regret**		
	K = 2	K = 5	K = 10
ϵ-greedy ($\epsilon = 0.1$)	3326.135	6667.353	8179.818
Variable ϵ greedy	81.872	791.476	1340.028
Soft-max ($\beta = 0.05$)	727.078	3580.839	1389.425
UCB	81.741	156.033	517.477
Thompson sampling	3.797	14.506	32.526
Reinforce without baseline	39.880	109.630	447.888
Reinforce with baseline	38.885	104.269	389.414

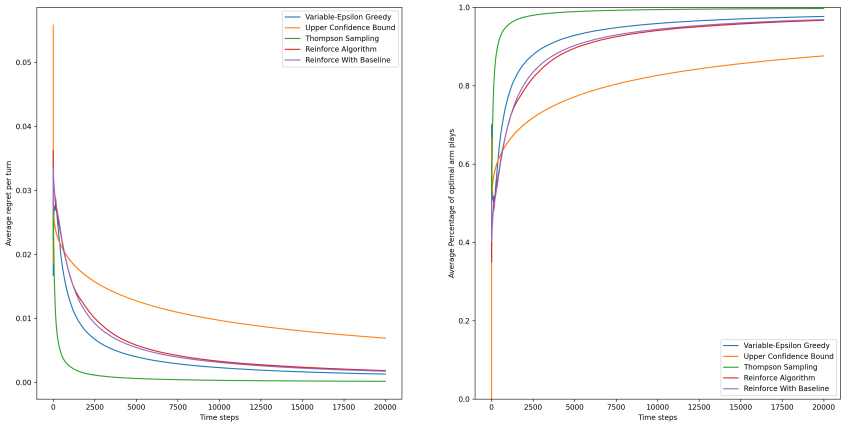

Figure 8: Variation of average regret per turn and average percent of optimal arm with time steps for $K = 2$ with Gaussian reward distribution ($\sigma = 0.01$)

4.3 Bandits with Gaussian reward distribution with $\sigma = 1$

Table 4: Performance of algorithms for Gaussian reward distribution ($\sigma = 1$)

Algorithms	**Cumulative regret**		
	K = 2	K = 5	K = 10
ϵ-greedy ($\epsilon = 0.1$)	3347.369	6667.150	8164.337
Variable ϵ greedy	100.056	918.101	1700.357
Soft-max ($\beta = 0.05$)	868.254	1587.601	3167.288
UCB	62.127	149.131	283.968
Thompson sampling	556.792	70.701	400.032
Reinforce without baseline	44.961	111.962	178.440
Reinforce with baseline	50.103	138.830	202.336

4.4 Observations

- We observed that for bandits with Gaussian reward distribution (fixed variance), irrespective of the number of arms **Thompson sampling** performed well for $\sigma = 0.01\&0.1$.

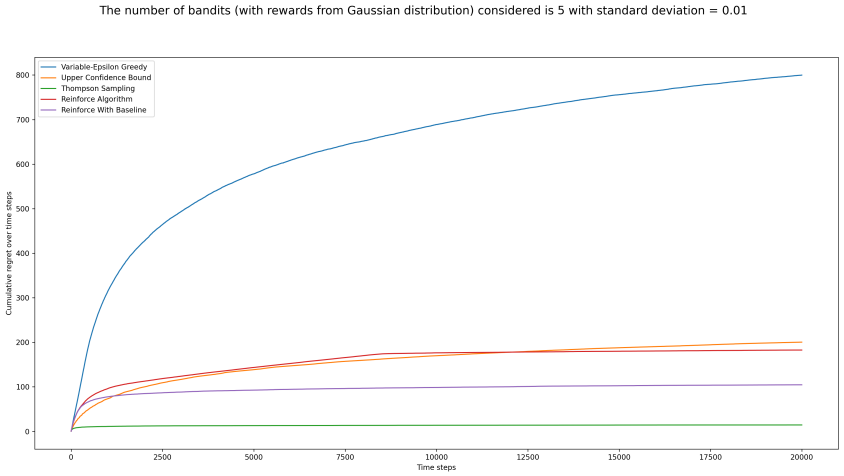

Figure 9: Variation of total regret with time steps for $K = 5$ with Gaussian reward distribution ($\sigma = 0.01$)

- According to the obtained simulation results we rank the bandit algorithms in the following order :
  For $\sigma = 0.01\&0.1$:
  Thompson sampling >> Reinforce with and without baseline >> UCB $\approx$ Variable ϵ-greedy >> $\epsilon-$greedy $\approx$ Soft-max. For $\sigma = 1$:
  Reinforce with and without baseline >> Thompson sampling >> UCB $\approx$ Variable ϵ-greedy >> $\epsilon-$greedy $\approx$ Soft-max.

- We observed that, when the $\sigma = 1$ and for $K = 10$ the performance of Thompson sampling was not better than the reinforce with and without baseline.

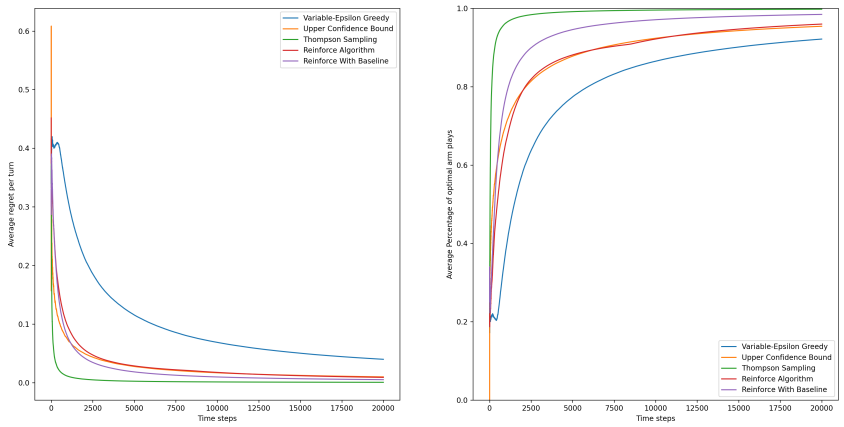

Figure 10: Variation of average regret per turn and average percent of optimal arm with time steps for $K = 5$ with Gaussian reward distribution ($\sigma = 0.01$)

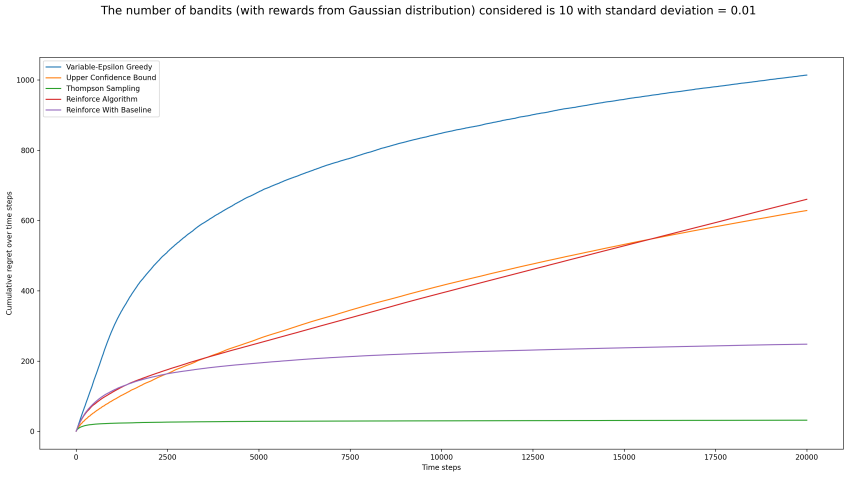

Figure 11: Variation of total regret with time steps for $K = 10$ with Gaussian reward distribution ($\sigma = 0.01$)

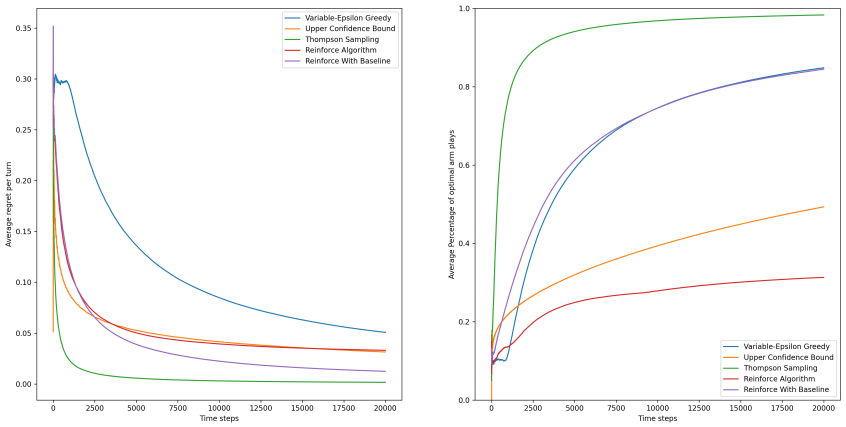

Figure 12: Variation of average regret per turn and average percent of optimal arm with time steps for $K = 10$ with Gaussian reward distribution ($\sigma = 0.01$)

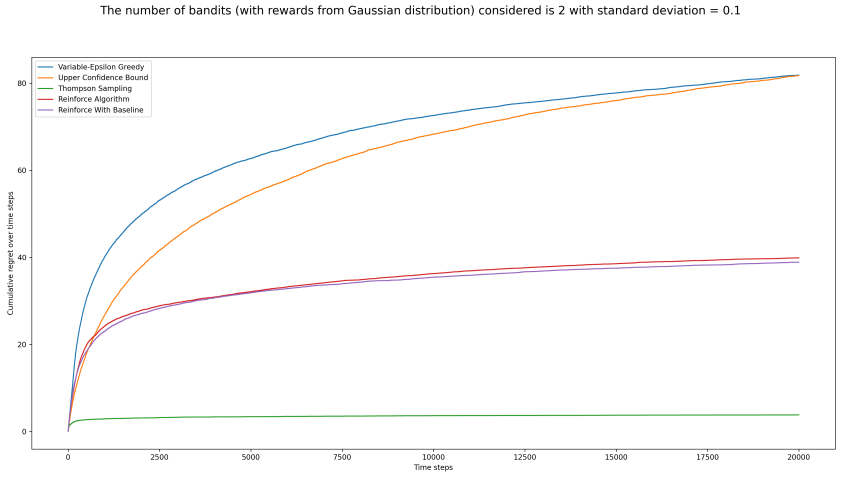

Figure 13: Variation of total regret with time steps for $K = 2$ with Gaussian reward distribution ($\sigma = 0.1$)

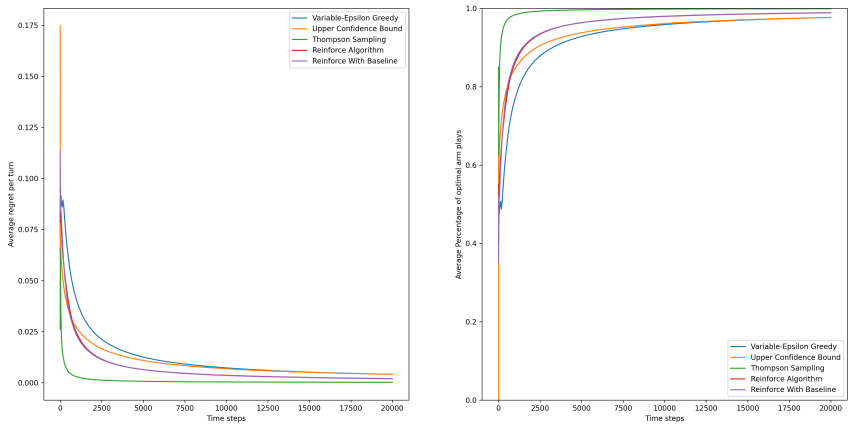

Figure 14: Variation of average regret per turn and average percent of optimal arm with time steps for $K = 2$ with Gaussian reward distribution ($\sigma = 0.1$)

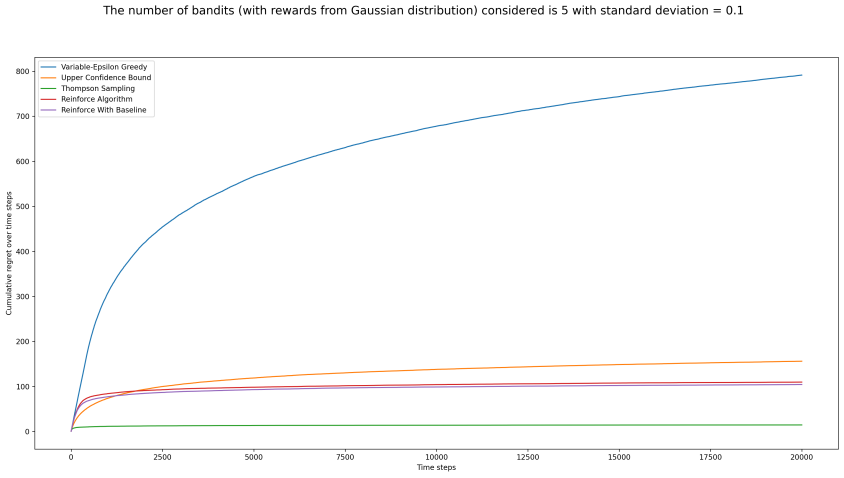

Figure 15: Variation of total regret with time steps for $K = 5$ with Gaussian reward distribution ($\sigma = 0.1$)

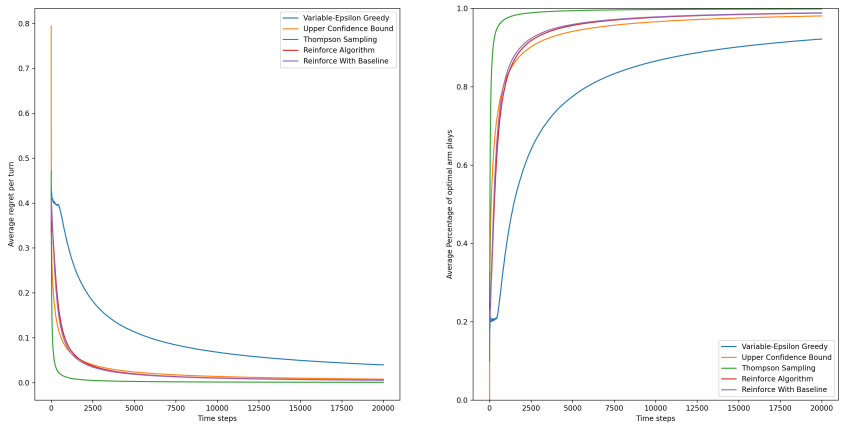

Figure 16: Variation of average regret per turn and average percent of optimal arm with time steps for $K = 5$ with Gaussian reward distribution ($\sigma = 0.1$)

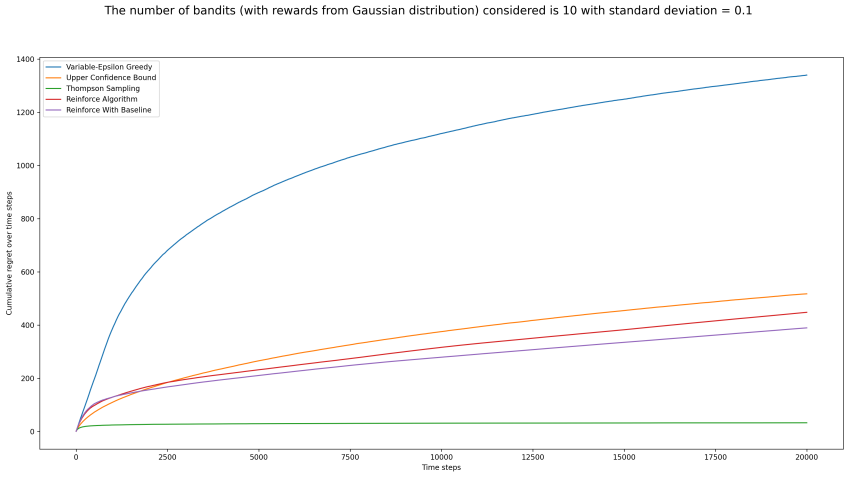

Figure 17: Variation of total regret with time steps for $K = 10$ with Gaussian reward distribution ($\sigma = 0.1$)

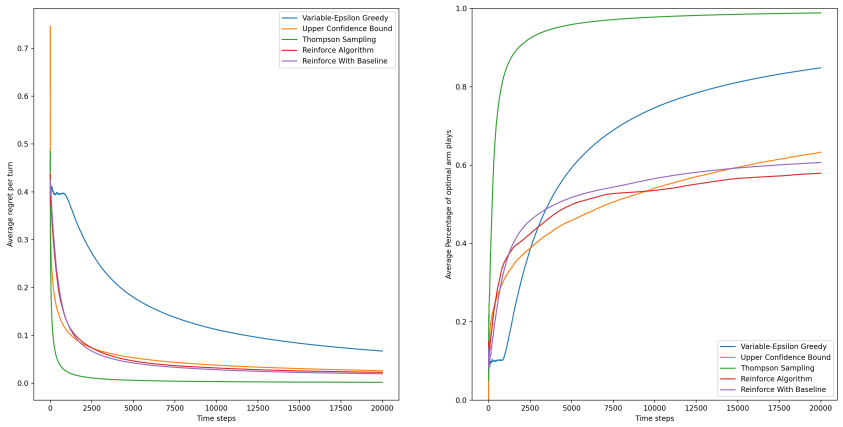

Figure 18: Variation of average regret per turn and average percent of optimal arm with time steps for $K = 10$ with Gaussian reward distribution ($\sigma = 0.1$)

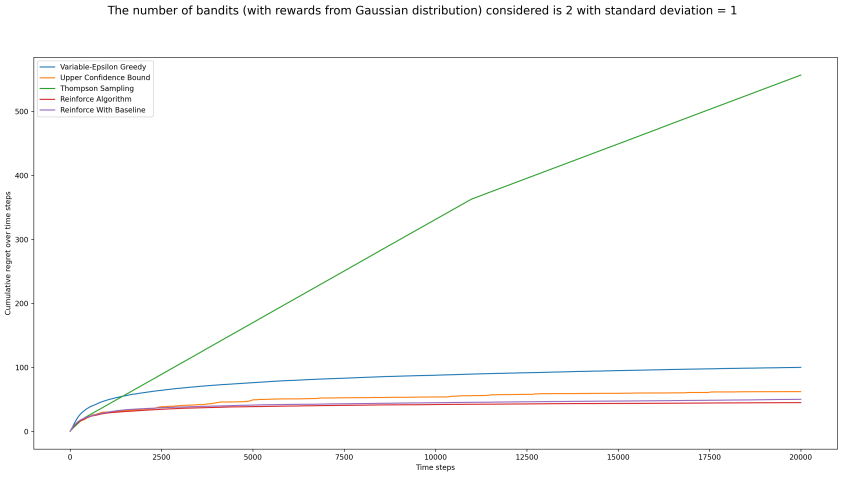

Figure 19: Variation of total regret with time steps for $K = 2$ with Gaussian reward distribution ($\sigma = 1$)

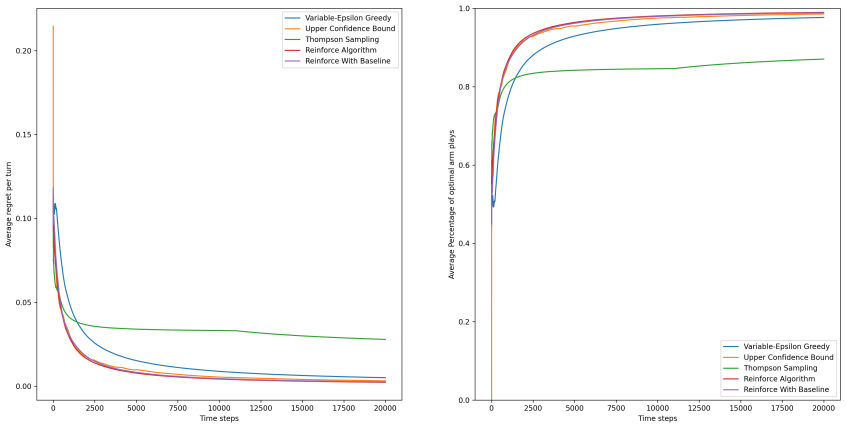

Figure 20: Variation of average regret per turn and average percent of optimal arm with time steps for $K = 2$ with Gaussian reward distribution ($\sigma = 1$)

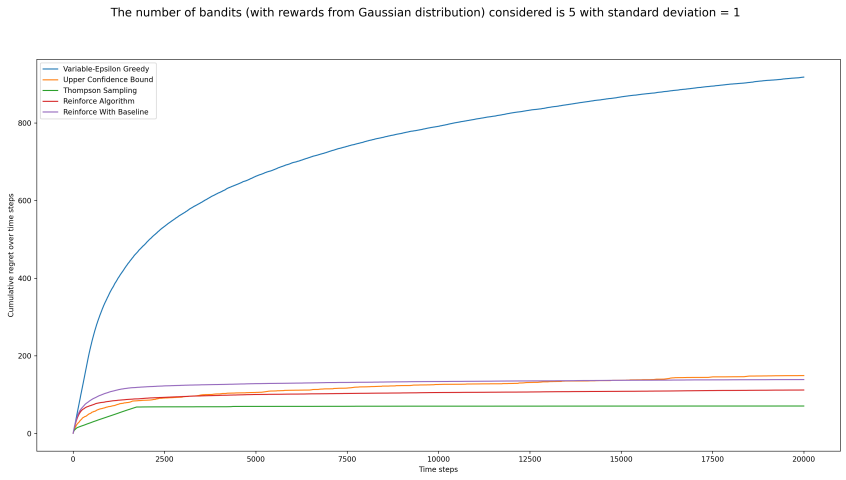

Figure 21: Variation of total regret with time steps for $K = 5$ with Gaussian reward distribution ($\sigma = 1$)

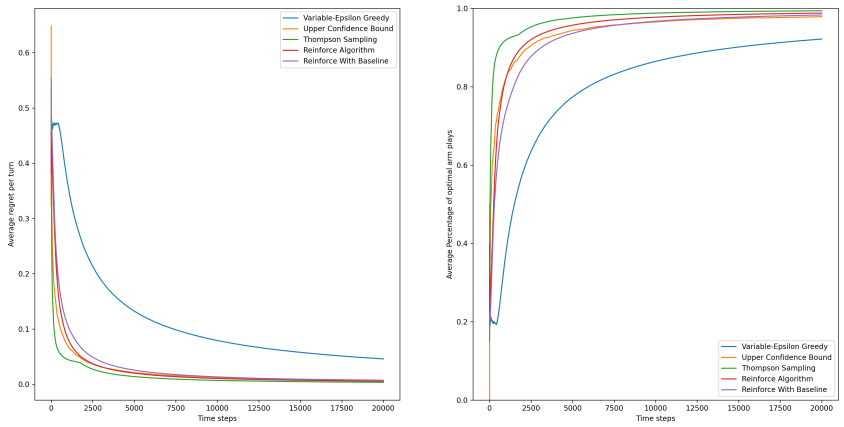

Figure 22: Variation of average regret per turn and average percent of optimal arm with time steps for $K = 5$ with Gaussian reward distribution ($\sigma = 1$)

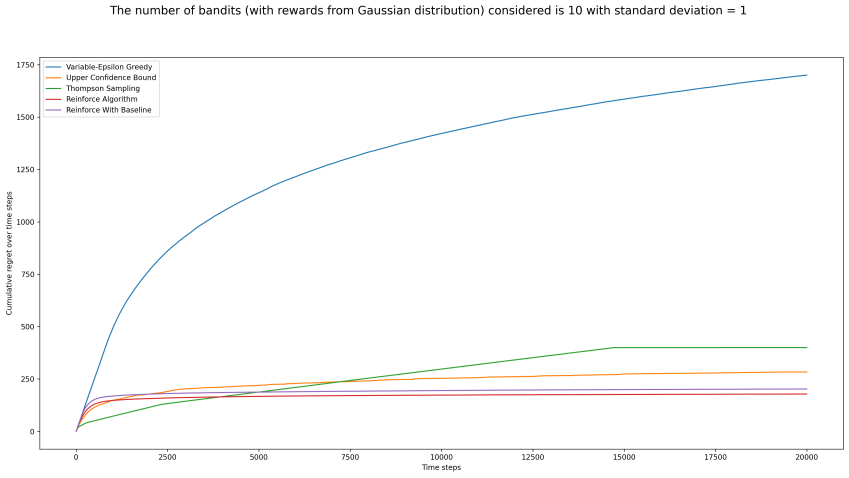

Figure 23: Variation of total regret with time steps for $K = 10$ with Gaussian reward distribution ($\sigma = 1$)

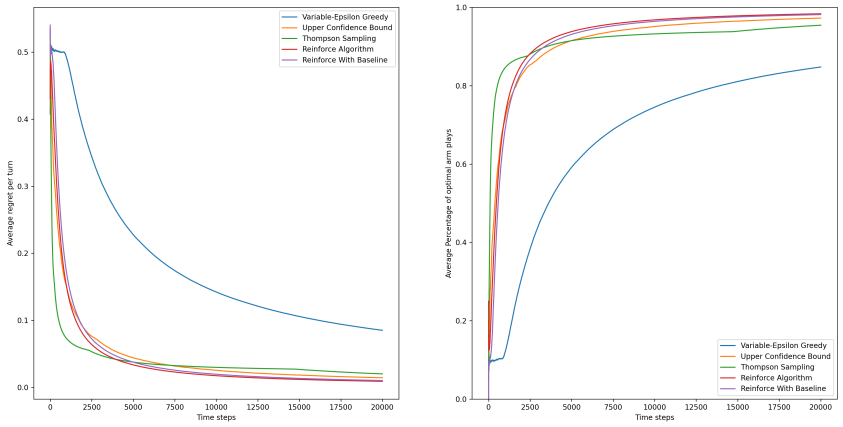

Figure 24: Variation of average regret per turn and average percent of optimal arm with time steps for $K = 10$ with Gaussian reward distribution ($\sigma = 1$)