Optimize your Bandit Problem using Thompson Sampling, e-greedy, UBC, and random sampling.

* Thompson Sampling (ThompsonSampling.py)
  + Without the initialized alpha and beta, the figure is:
    - 图表, 折线图

      描述已自动生成
  + With the initialized alpha and beta based on randomly generated probabilities, the figure is:
    - 图表, 折线图

      描述已自动生成
* E-Greedy (EGreedy.py)
  + When e equals 0.1, the figure is:
    - 图形用户界面

      描述已自动生成
* UCB (UpperConfidenceBound.py)
  + 图表, 折线图

    描述已自动生成
* Random Sampling (RandomSampling.py)
  + 图表, 折线图

    描述已自动生成

Which hyperparameters are important for Thompson Sampling, e-greedy, UBC, and random sampling? Show that they are important (15 Points)

* Thompson Sampling
  + K and t in the turn of generating the sample value from beta distribution, because it establishes the foundation when the trail and reward period starts to run.
    - 文本

      描述已自动生成
* E-Greedy
  + e, the probability of choosing exploration, determines the balance between exploration and exploitation, which affects the algorithm efficiency and the result accuracy, because the times of trials is finite.
* UCB
  + c constant in the formula of delta value, because delta value controls the speed of delta increasing, influencing the balance of exploration and exploitation.
    - 图片包含 图示

      描述已自动生成
* Random Sampling
  + The number of trials will affects the result accuracy, because it should be farther larger than the number of candidates to avoiding vague results.

How does the action space affect Thompson Sampling, e-greedy, UBC, and random sampling? Show why. (15 Points)

* Thompson Sampling
  + Thompson Sampling focuses on exporting those actions that not tried. When there are lots of actions waiting to try, and the result of each trail is zero or one, it requires a large amount of sources to try different actions many times to make the rewards close to zero, to weed them out. However, it decreases the rate of regrets.
* E-Greedy
  + Because greedy algorithm seeks for the locally optimal solution, the action space of E-Greedy is smaller than Thompson and UCB. It might not find the optimal solution, but there is high possibility for E-Greedy to find the action with reward probability close to that of the optimal action, and the possibility of regrets is low.
* UCB
  + Because the delta value calculation formula, which encourages UCB to try those actions with lower times of execution, it enlarges the action space when the number of trails goes bigger. However, it decreases the rate of regrets.
* Random Sampling
  + Random Sampling will not be affected by action space, because it does not receive and deal with the feedback of trail results, and decides which action to make merely dependent on the random seed.

How does stationary affect Thompson Sampling, e-greedy, UBC, and random sampling? Show why. (15 Points)

* Thompson Sampling
  + Thompson Sampling requires stationary to some degree. Although it could decrease the probability of regrets at the beginning of the sample changing, it will not perform as well as before over time. The ultimate regret increases.
* E-Greedy
  + When the parameters are resampled, the final performance of e-greedy goes worse over time because it focuses on locally optimal action.
* UCB
  + UCB could deal with non-stationary with good performance, and it decrease the probability of regrets. It will decrease regrets stably during resampling.
* Random Sampling
  + Random sampling will not be affected by the reward. Or, in other words, it does not deal with the reward of actions.

When do Thompson Sampling, e-greedy, UBC, and random sampling stop exploring? Explain why. Explain the exploration-exploitation tradeoff (15 Points)

* Thompson Sampling
  + When Thompson sampling repeats one action enough time, and meanwhile other actions are executed fixed times with low reward, it could stop.
* E-Greedy
  + When E-Greedy finds the locally optimal solution, it repeats making the same action for the fixed time, it will stop exploring.
* UCB
  + Because delta value depends on both the number of trials and the times of this action was made, when the estimated reward of one action is higher enough than other actions, and the times of other action execution is not too small, reaching the fixed number, it could stop.
* Random Sampling
  + Random sampling makes actions by randomly choosing, does not receive and reply to the feedback of state space. It will stop exploring as it has been set before the trails start.

How long do Thompson Sampling, e-greedy, UBC, and random sampling remember the past actions? Explain your answer. (10 Points)

* Thompson Sampling
  + It records the result of every actions by creating alpha and beta arrays for the whole period.
* E-Greedy
  + It only records the last and the current result.
* UCB
  + It only records the last and the current result.
* Random Sampling
  + It will stores the total rewards during the whole period.

Thompson Sampling with non-Beta distribution (5 Points) Modify the Thompson Sampling to run with a different distribution (e.g. Parteo, Normal, etc)

It should rewrite the code of calling pymc.rbeta() by using distribution method.