**Stochastic Search and Optimisation:**

**Analysis of Multi-armed Bandits**

**1 The Problem**

This report will indicate how to maximize profits in Multi-armed Bandits problem.

The introduction of Multi-armed Bandits problem as follows:

There are  machines, and one of them is selected to pull the lever. The machine provides a random reward, and the reward of each machine is subject to a specific probability distribution. A gambler has  chances of lever pulls. His goal is to maximize the reward. For each round of selection, the main problem is Exploration-Use. That is, the gambler should pull the arm with Highest expected payoff or play other machines to get more information about the expected payoffs of them.

This report will analyze several different machine selection methods. The first is the naive probability based on Bayesian theory. On this basis, the ε-greedy method and UCB method are considered.

**2 Solution Method**

Multi-armed Bandits的基本实现由example program “mab\_sim.r” 给出。它被进行了一些改进使得可以使用choose function自动选择machine而不是手动输入。

为了进行比较，多种choose functions被实现。

最基本的选择基于朴素的贝叶斯思想，在该问题中，由于每台Bandits的pull结果服从伯努利分布，后验概率等于reward expectations。

在此基础上容易实现ε-greedy method. 它的选择过程是，对于（为所有machine的reward expectations），有的概率选择第台machine作为本次pull的对象，有的概率随机选择一台machine作为本次pull的对象。

最后是UCB方法。它使用公式：



对每台machine得到一个平衡Exploration-Use的均衡化评分。公式中前项代表目前对该machine的reward expectations，后项代表“exploration”该machine有多大价值。可以看出，公式的两项本身并无一个可比较量纲约束其相对值大小，这样会在各个machine概率方差不同时造成结果的不稳定。因此需要对两项添加系数进行约束。容易考虑，随着试验过程的深入，“Use”的权重应当被增加，而“exploration”的权重应当被降低。因此最终公式为：



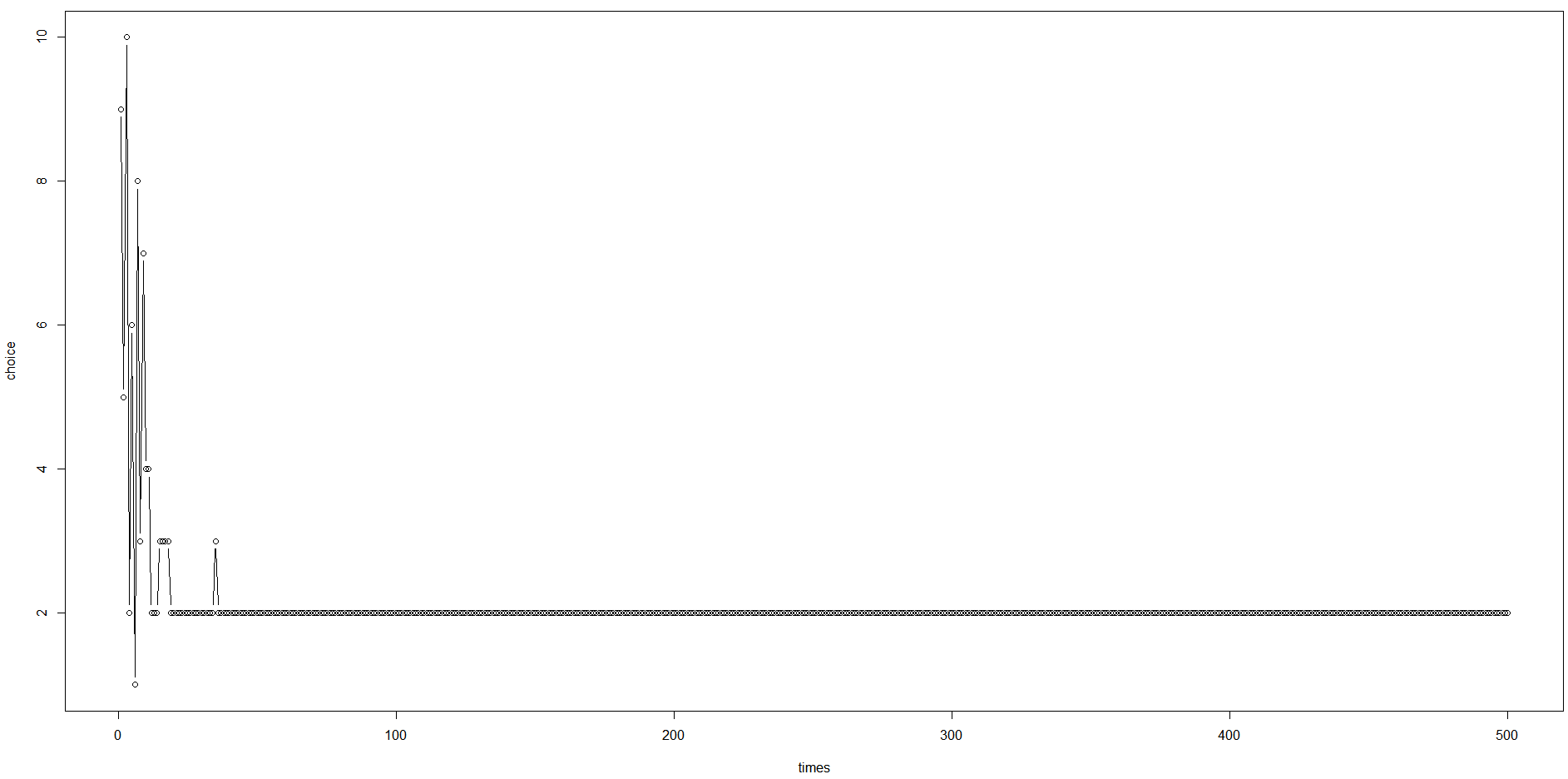
UCB评分最高的machine将被选择。可以看出，这种对于公式的变形与退火算法起到的作用是相同的。

另外还有基于贝叶斯决策思想的Thompson sampling。它使用每个machine的两参数生成一个beta分布。从所有machine的beta分布中进行采样，值最大的machine被选择。

**3 Experiment & Discussion**

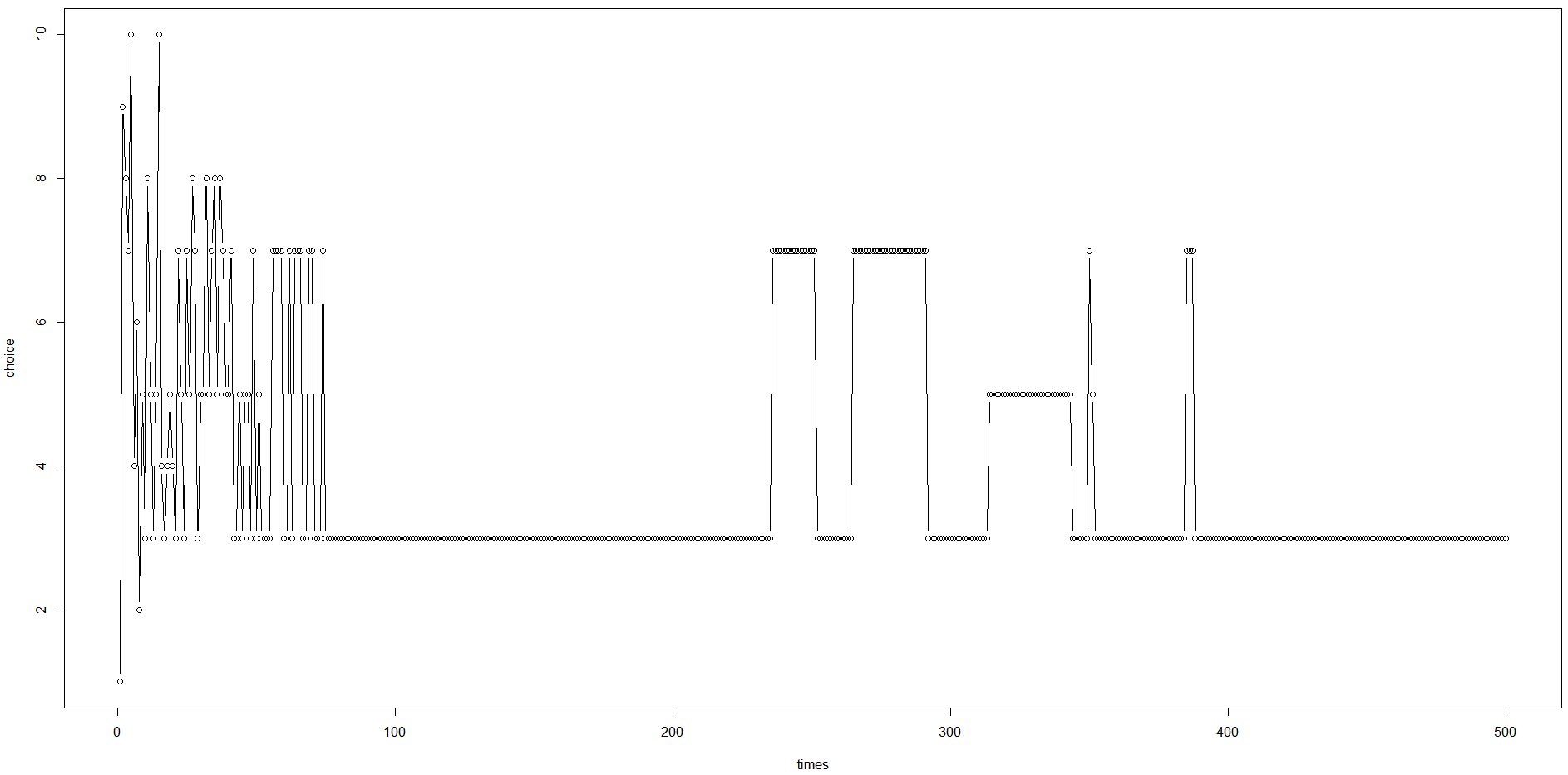
以10个machines，500次选择机会为参数进行实验。

对于直接选择最大reward expectations的方法（Greedy method），存在两种典型情况：



*当各机器reward概率分别为*

*0.6698579 0.7717658 0.8094393 0.8043435 0.1346799 0.06263159 0.06610808 0.5332526 0.4820622 0.3480183*

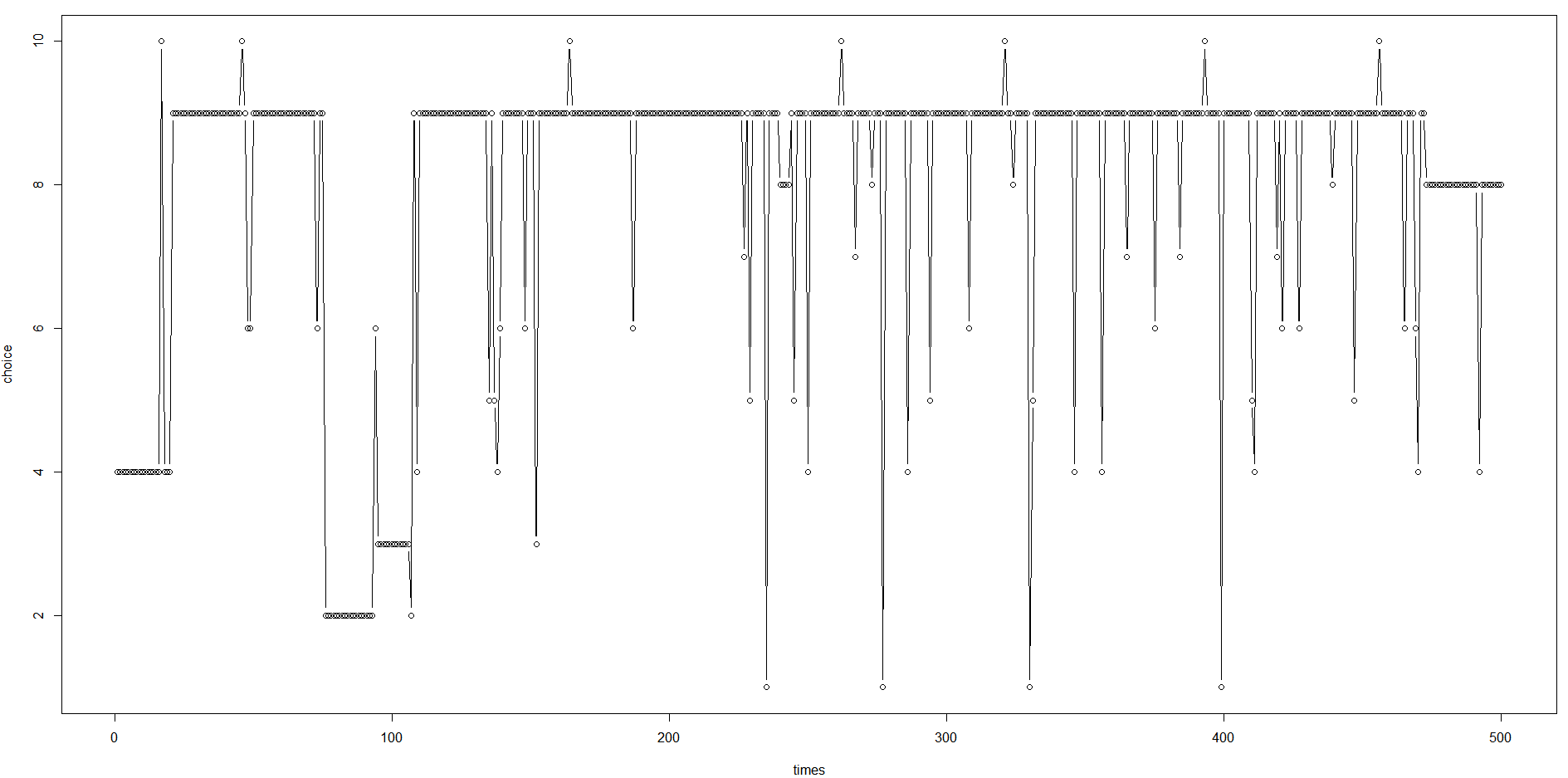


*当各机器reward概率分别为*

0.05696483 0.2703675 0.9262651 0.5431344 0.8945411 0.2681076 0.8948032 0.9364386 0.2179528 0.3285078

可以看出，当各个machine概率方差较大时，此方法可以很快收敛并得到较好结果。而当方差较小时，则会出现选择震荡情况。

对于ε-greedy方法，以上分析仍然适用。除此之外，一种新的有代表性的情况如下：



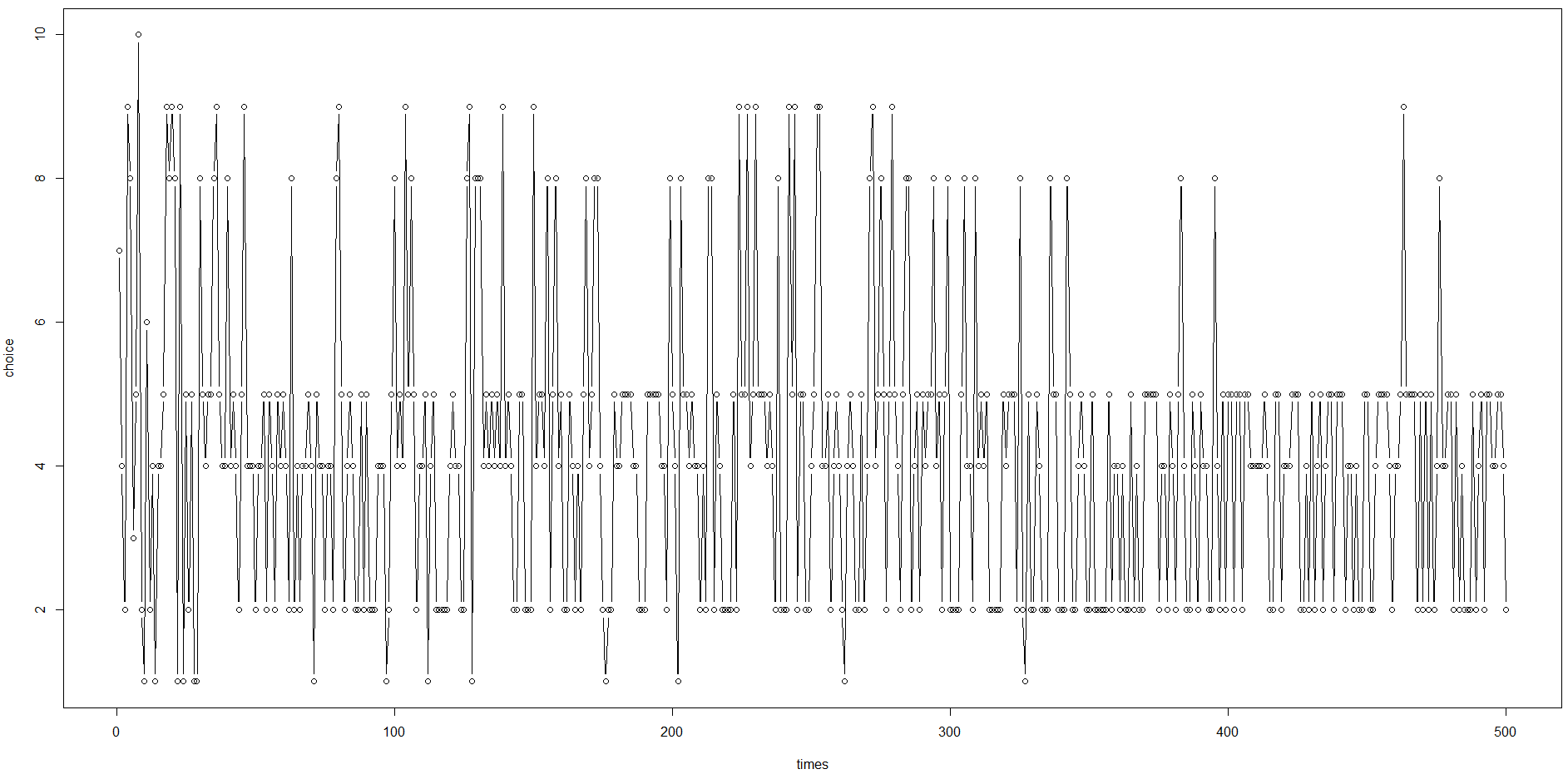
*当各机器reward概率分别为*

*0.1443913 0.75339 0.6888177 0.6754279 0.03809189 0.3998524 0.3650092 0.7453877 0.7152574 0.2146367*

此次实验中，不仅各个machine概率方差较小，而且最大值也比较小。这使得收敛到的值不够大。因此即使已经得到了相对准确的reward expectations，ε-greedy依然会不断尝试“exploration”，造成了选择的震荡（选择的震荡会直接导致更大的cumulative regret）。使用退火方法约束试验后期的exploration选择可以一定程度缓解这种情况。

但对于machine概率方差较小而最大值较大的情况，ε-greedy可以相对greedy方法获得更好结果。

Thompson sampling的震荡情况相比ε-greedy更加严重。普遍性结果如下：



*当各机器reward概率分别为*

*0.2264612 0.9853143 0.3358236 0.9316129 0.97296 0.5930124 0.3902846 0.6714077 0.5611102 0.2146527*

造成这种结果的原因是采样带来的不稳定性。与ε-greedy可以通过退火方法约束exploration不同，对固定的后验分布进行采样无法进行恰当约束。或许在更多的machine带来更大的探索空间下，该方法会相对更有优势。

对于UCB method，以greedy method作为baseline（因为ε-greedy稳定性较差），各实验十次（各机器概率在0-1高斯随机采样）结果如下：

*422 448 428 446 339 382 440 434 371 456*

*UCB方法*

*488 426 452 408 306 440 486 477 488 348*

*greedy method*

可以看出，结果并无很大区别，甚至greedy method的结果更好一些。但如果将采样范围调整至0.6-0.8，400次选择机会，结果如下：

*315 303 308 283 282 278 262 311 294 307*

*UCB方法*

*295 296 320 307 289 303 283 299 281 313*

*greedy method*

可以看出，UCB与ε-greedy一样在各个machine概率方差较小时会有更好的表现。另外，选择次数较少时，UCB策略可以比其它策略更快的获得较高概率的最优解。另外，适当降低UCB公式中第二项的权重衰减速度，可以使其更接近greedy方法，从而在一些情况下达到更好的效果。