Сравнение стратегий в многоруких бандитах для распределения Стьюдента с разными степенями свободы

Михаил Давыдов

13 марта 2024 г.

Аннотация

Я рассмотрел эффективность алгоритмов для решения задачи о многоруких бандитах для распределений, отличных от нормального. В результате получил снижение эффективности стандартных алгоритмов для распрделений Стьюдента с малым числом степеней свободы и низкую эффективность известных стратегий для распределения Коши.

1 Введение

Инвесторы вкладывают свои деньги в акции, чтобы сохранить или приумножить свои богатства. Они хотят использовать для распределения денег эффективные алгоритмы. Для этого они могут использовать модели для приближенного вычисления изменения стоимости акций [BP03]. В качестве одной из таких моделей можно использовать модель многоруких бандитов.

Задача о многоруких бандитах звучит так: есть k рычагов, каждый ход можно выбрать один из рычагов, и в ответ на выбор i-ого рычага игрок получает награду из распределения, связанного с i-ым рычагом [SB18]. Цель – максимизировать среднюю награду за определенное число, например, 1000 шагов. Оптимальной стратегией является нажатие на лучшие рычаги, то есть рычаги с наибольшим матожиданием (в случае, если нам неинтересны риски) или, если матожидания нет, с наибольшим значением по какой-то единой для всех рычагов метрике, например, по медиане. Проблема заключается в том, что изначально распределения наград для рычагов неизвестны, и их надо приблизить нажатиями на рычаги и получениями выборки наград.

Можно считать, что изменение стоимости акции тоже приходит из какого-то распределения. Изначально эти изменения акций нам неизвестны, и задача инвестора – последовательными изменениями своего экономического портфеля найти такое распределение денег по акциям, которое в среднем дает наибольший прирост стоимости портфеля. Для нахождения распределения денег можно использовать стандартные алгоритмы из задачи о многоруких бандитах. Однако эти алгоритмы были опробованы на гауссовских распределениях, в то время как в реальной жизни гораздо чаще наблюдаются степенные распределения с более тяжелыми хвостами. Поэтому я решил проверить эффективность алгоритмов на распределениях Стьюдента с разными степенями свободы.

2 Связанные работы

Главной работой, на которую я опирался, является книга Ричарда Саттона и Эндрю Барто "Reinforcement Learning: An Introduction". Во второй главе авторы обсуждают задачу о многоруких бандитах и способы ее решения. Однако в качестве распределений для рычагов используются нормальные распределения и не рассматриваются никакие другие распределения, что является серьезным упущением.

3 Методология

Для рассмотрения алгоритмов я использовал 10 рычагов, распределения всех рычагов одинаковы с точностью до смещения. Я рассмотрел 4 распределения: стандартное нормальное распределение

($\mathcal{N}(0,1)$ или t_∞), которое является предельным случаем распределения Стьюдента при числе степеней свободы, стремящемся к бесконечности; распределение Стьюдента с 3 степенями свободы (t_3), домноженное на $\frac{1}{\sqrt{3}}$, чтобы дисперсия была равна 1; распределение Стьюдента с 2 степенями свободы (t_2) без нормировки, поскольку у t_2 нет дисперсии; распределение Стьюдента с 1 степенью свободы, или же стандартное распределение Коши (t_1). Каждое из этих распределений было смещено на значение, выбранное случайно из $\mathcal{N}(0,1)$.

Введем некоторые обозначения:

- R_t – награда, полученная на t-ом шагу (то есть нажали на рычаг в t-ый раз).
- A_t – номер рычага, выбранный на t-ом шагу.
- $N_t(a) := \sum_{i=1}^{t-1} I(A_i = a)$ – количество нажатий на рычаг a на t-ом шагу.
- $Q_t(a) := \frac{\sum_{i=1}^{t-1} R_i \cdot I(A_i=a)}{N_t(a)}$ – средняя награда при нажатии рычага с номером a. Можно считать, что обновление происходит так: если на шаге t было выбрано действие a, то для всех остальных действий b значение $Q_t(b)$ не меняется, а для действия a $Q_t(a) = Q_t(a) + \frac{1}{N_t(a)+1}(R_t - Q_t(a))$.
- $\bar{R}_t := \frac{\sum_{i=1}^{t-1} R_i}{max(t-1,1)}$ – средняя награда за все предыдущие шаги, или, как ее называют по-другому, baseline.

Стратегии и порядок проверки распределений полностью совпадают с таковыми в описанной выше книге Саттона и Барто, а именно:

1. Рассматриваются ϵ-greedy стратегии, то есть стратегии, действия в которых выбираются по формуле
$$A_t = \begin{cases} \arg\max\limits_a Q_t(a), & \text{with probability } 1-\epsilon, \\ \text{a random action}, & \text{with probability } \epsilon. \end{cases}$$
(среди равных значений выбор происходит случайно). Далее выбираются ϵ-greedy стратегии с $\epsilon = 0$ (то есть просто жадная стратегия), $0.01, 0.1$. Изначально $\forall a\ Q_t(a) = 0$.

2. Далее берутся стратегии с оптимистичной инициализацией, то есть стратегии, для которых изначально $\forall a\ Q_t(a) = d,\ d > 0$. Я использовал $d = 5$ (как в книге). Обновление $Q_t(a)$ происходит с константынм шагом, то есть если на шаге t было выбрано действие a, то для всех остальных действий b значение $Q_t(b)$ не меняется, а для действия a $Q_t(a) = Q_t(a) + \alpha(R_t - Q_t(a)),\ \alpha = const$. α называется step-size. Так же, как и в книге, я выбрал $\alpha = 0.1$. Я сравнил greedy-стратегию с оптимистичной инициализацией с 0.1-greedy стратегиями с обычной и оптимистичной инициализациями.

3. Upper-Confidence-Bound Action Selection – выбор действия происходит по формуле
$$A_t = \arg\max_a \left[Q_t(a) + c\sqrt{\frac{ln\ t}{N_t(a)}} \right],\ c > 0$$
В случае, когда $N_t(a) = 0$, значение функции справа считается равным бесконечности. Сравнил UCB с $c = 2$ с 0.1-greedy стратегией.

4. Gradient bandits: для каждого действия a вводится значение $H_t(a)$. На t-ом шаге выбор происходит среди действий соответственно вероятностям $\Pr\{A_t = a\} = \frac{e^{H_t(a)}}{\sum_{i=1}^k e^{H_t(i)}} := \pi_t(a)$. После выбора действия A_t и получения награды R_t обновления H_t происходят по формулам:
$$H_{t+1}(A_t) = H_t(A_t) + \alpha(R_t - \bar{R}_t)(1 - \pi_t(A_t)), \text{ and}$$
$$H_{t+1}(a) = H_t(a) - \alpha(R_t - \bar{R}_t)\pi_t(a),\ a \neq A_t.$$
Я сравнил результаты gradient bandits для $\alpha = 0.1,\ 0.4$, и в случаях, когда baseline есть и когда его нет. При этом моды распределений специально смещены на 4, чтобы показать, что, в отличие от других стратегий gradient bandit невосприимчив к смещению распределений.

Каждую из стратегий я запустил для каждого из распределений на 2000 независимых тестах (для каждого теста свое смещение распределений) длиной 1000 шагов каждый, после чего посчитал среднюю награду и процент выбранных оптимальных действий на каждом шагу (во всех распределениях оптимальным считался рычаг с наибольшей медианой, что в случае нормального и Стьюдента с > 1 степеней свободы равносильно наибольшему матожиданию). Для каждого распределения и каждой метрики результаты для одной группы стратегий визуализированы на графике. Кроме того, я решил дополнительно изобразить для каждой стратегии и для каждой метрики на одном графике результаты по всем распределениям.

В конце я провел обзор всех стратегий для различных значений гиперпараметров. Для каждой стратегии варьировался один ключевой гиперпараметр, варьирование происходило по значениям

$$\frac{1}{128}, \frac{1}{64}, \frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4$$

(для ϵ-greedy значения 2 и 4 не рассматривались). Взяты стратегии:

- ϵ-greedy, варьирование по ϵ
- greedy с оптимистичной инициализацией и $\alpha = 0.1$, варьирование по $Q_0(a)$
- UCB, варьирование по c
- gradient bandits, варьирование по α

Ввиду долгого выполнения процент оптимальных действий и средняя награда брались по 1000 тестам, при этом количество шагов осталось равным 1000. В случае средней награды бралась средняя награда на 1000-ом шагу, в случае оптимального действий брался наилучший процент за 1000 шагов. Все эти вычисления были выполнены для всех распределений, после чего для каждого распределения и для каждой метрики значения были визуализированы на графике.

Сам код можно найти в этом репозитории в папке "gradient bandit".

4 Результаты

На всех графиках график средней награды для распределения Коши не имеет смысла ввиду отсутствия у распределения матожидания.

1. ϵ-greedy: Как можно видеть, для t_2, t_3, t_∞ при увеличении ϵ до значения $\frac{1}{k} = \frac{1}{10}$ средняя награда и процент оптимальных действий увеличиваются. Для t_2 характерны резкие прыжки в средней награде ввиду отсутствия дисперсии. Для t_1 процент оптимальных действий значительно ниже других распределений и составляет около 30%, более того, оптимальный процент действий для каждой из стратегий примерно одинаков, изменение ϵ не ведет к изменению процента оптимальных действий для t_1. Предположу, что при $\nu \to 0$ процент оптимальных действий t_ν стремится к $\frac{1}{k}$. При положительном ϵ средняя награда и процент оптимальных действий для t_3 и t_∞ примерно одинаковы, а для t_2 эти значения несколько ниже. Для greedy стратегии наблюдается другая картина: метрики для t_∞ выше, чем для t_3, что можно объяснить тем, что после сжатия t_3 его пик стал более острым, и более резких изменений стало меньше, чем у t_∞, из-за чего исправлений неправильных выборов рычагов стало меньше. Для t_2 наоборот, метрики выше, чем у t_∞, что объясняется отсутствием дисперсии и более частыми “далекими” значениями, что повышает вероятность исправления неправильных выборов рычагов.

2. Оптимистичная инициализация: здесь ситуация несколько отличается. Оптимистичная жадная стратегия лучше оптимистичной ϵ-greedy и реалистичной ϵ-greedy только для t_∞ и t_3, причем для t_3 процент оптимальных действий для всех трех стратегий выравнивается к 1000-му шагу. Для t_2 и t_1 средняя награда и процент оптимальных действий для оптимистичной жадной стратегии хуже, чем для остальных стратегий. То есть жадность стратегии сильнее влияет на результат, чем инициализация. При этом для всех распределений для ϵ-greedy оптимистичной и реалистичной стратегий средняя награда и процент оптимальных действий выравниваются, что объясняется маленьким вкладом начальной инициализации

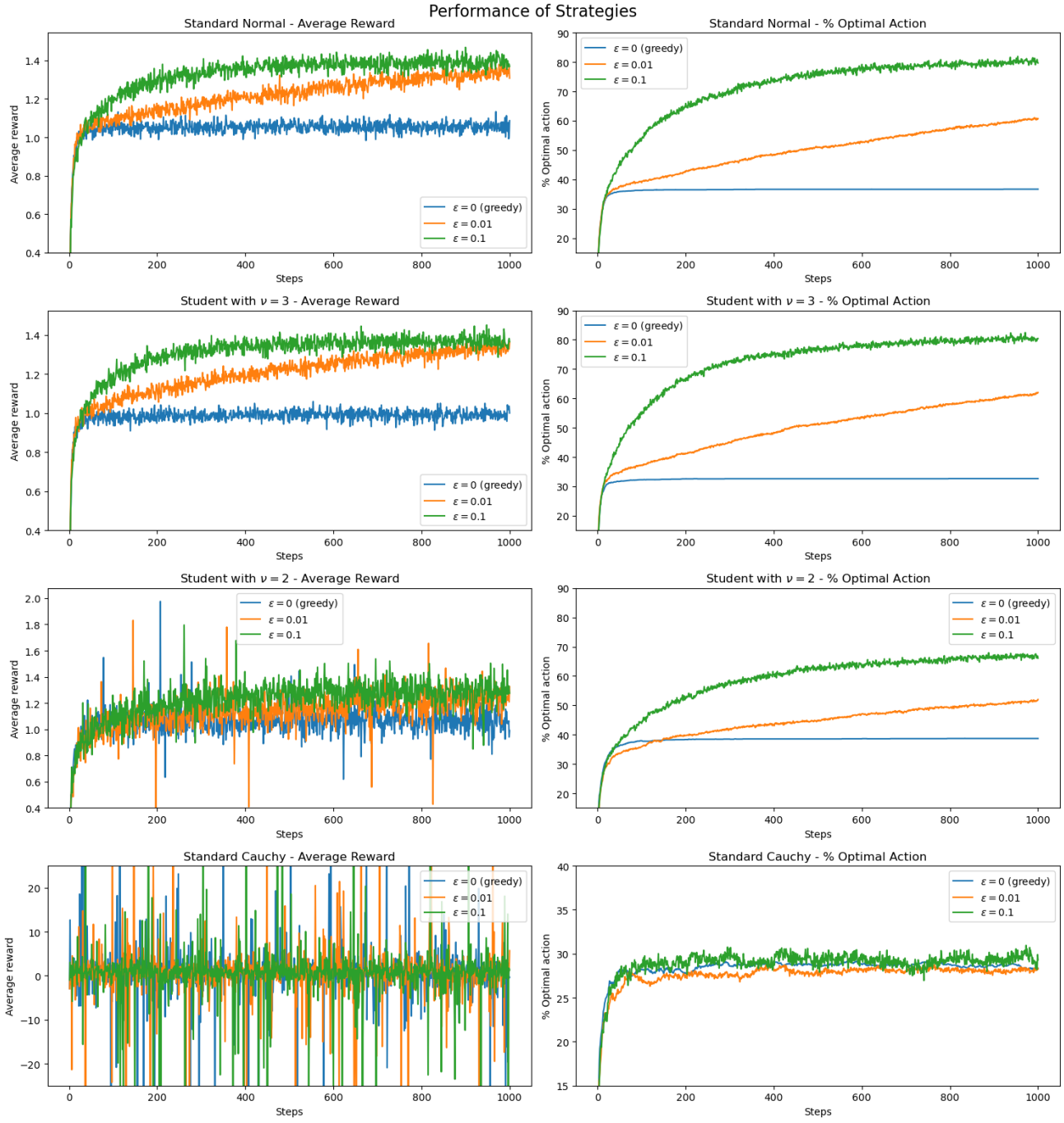

Рис. 1: Значения средней награды и процента оптимального выбора для ϵ-greedy стратегий, сгруппировано по распределениям

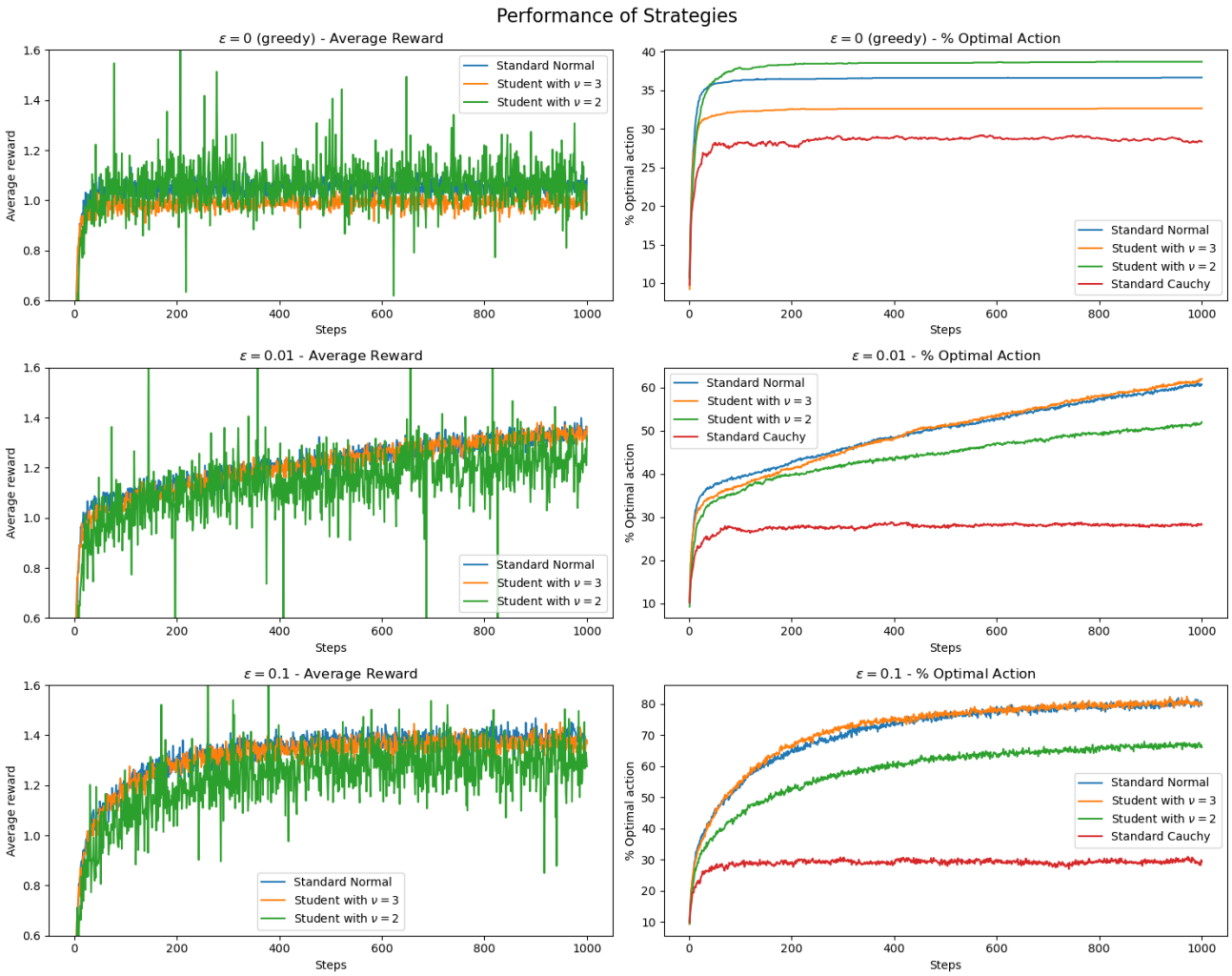

Рис. 2: Значения средней награды и процента оптимального выбора для ϵ-greedy стратегий, сгруппировано по стратегиям

на 1000-ом шагу. Кроме того, для t_3, t_2, t_1 и константного step-size наблюдается переобучение: в какой-то момент средняя награда и процент оптимальных действий начинают падать. Такая проблема возникает в силу того, что для константного step-size вклад выбросов не уменьшается со временем, и потому для распределений с тяжелыми хвостами возникает ситуация, когда приход “выброса” резко изменяет среднюю награду оптимального действия в отрицательную сторону, и затем это действие выбирается гораздо реже. Аналогично предыдущему эксперименту, для t_2 виден разброс в средней награде из-за отсутствия дисперсии. Если сравнивать распределения, то с уменьшением степеней свободы обе метрики падают. При этом для t_2 высота пика в проценте оптимальных действий на 10-ом шагу выше, чем для t_∞, что опять же, можно объяснить большей остротой пика.

3. UCB: на всех распределениях и всех метриках видно, что UCB лучше ϵ-greedy стратегии. Это можно объяснить тем, что UCB с увеличением уверенности реже производит "исследование"действий. При этом с уменьшением числа степеней свободы средняя награда незначительно падает, а процент оптимальных действий падает заметно. Для t_∞ и t_3 метрики падают на очень малую величину, позволяющую сказать, что UCB одинаково применимо для t_∞ и t_3.

4. Gradient bandits: для всех распределений присутствие baseline повышает значения метрик. Кроме того, значения метрик при $\alpha = 0.1$ лучше, чем при $\alpha = 0.4$ при заданных распределении и baseline, поскольку изменение $H_t(a)$ и, значит, $\pi_t(a)$, происходит менее резко, и

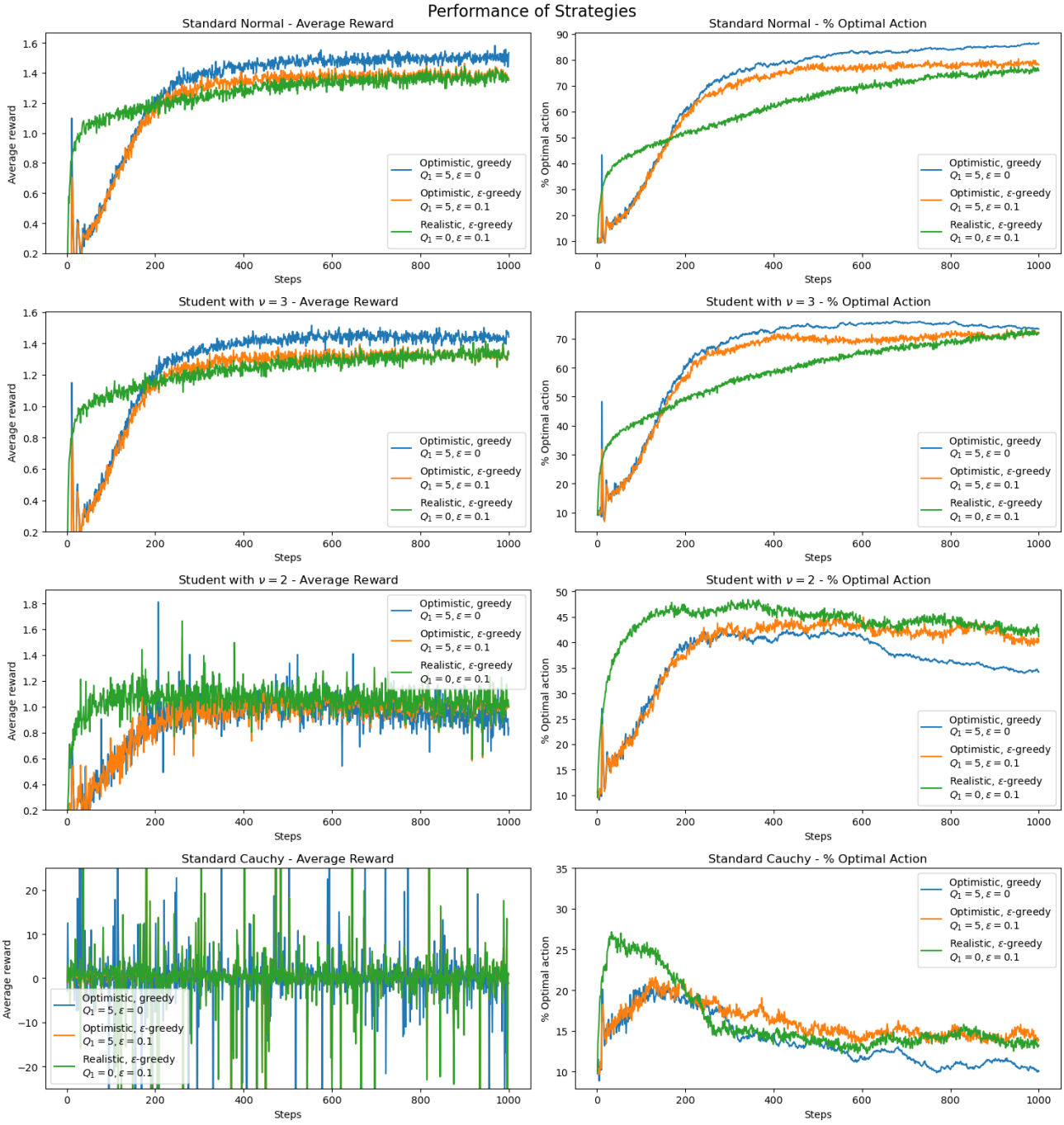

Рис. 3: Значения средней награды и процента оптимального выбора для оптимистичной стратегии в сравнении с ϵ-greedy стратегий для константного step-size, сгруппировано по распределениям. Для t_3, t_2, t_1 видно переобучение

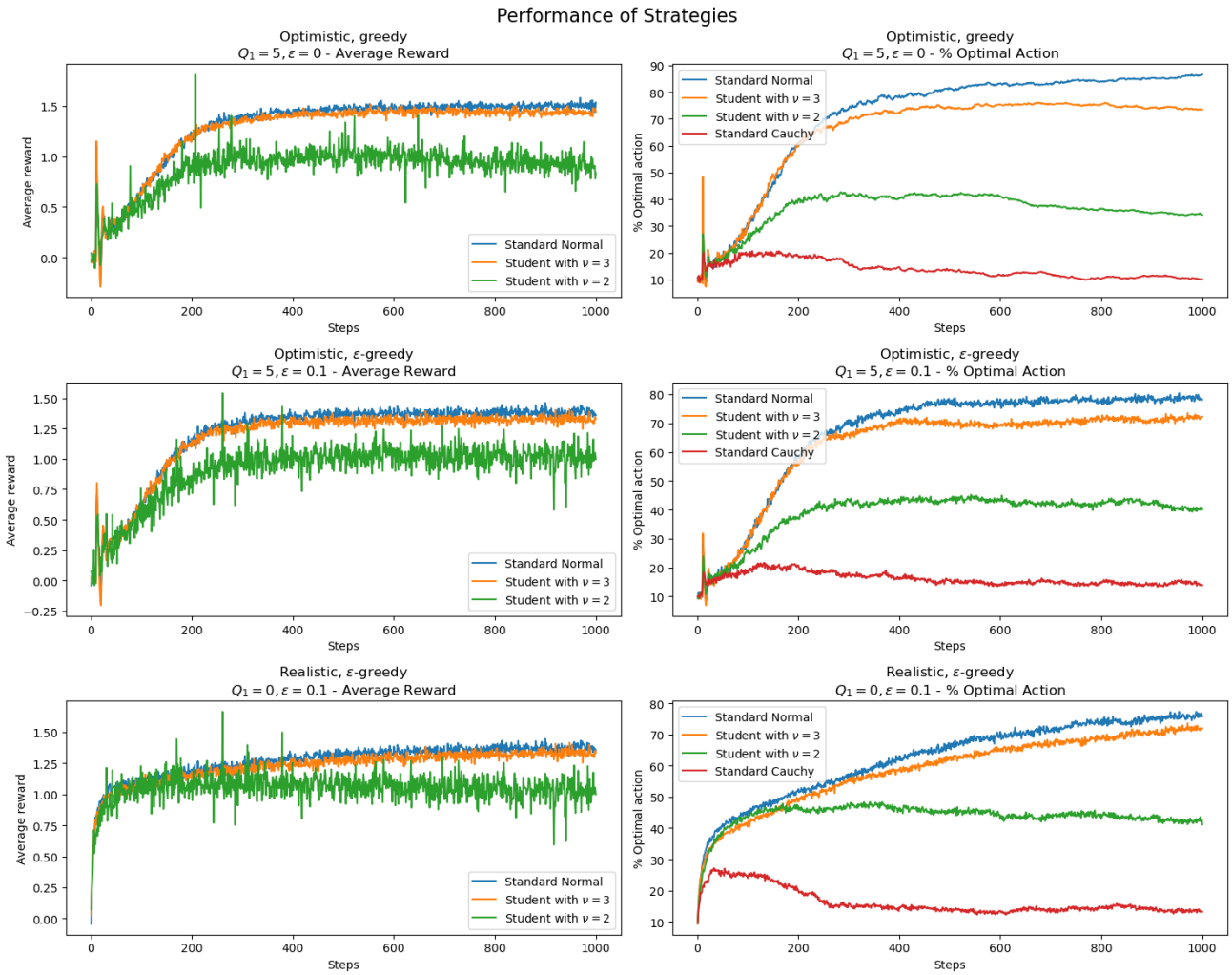

Рис. 4: Значения средней награды и процента оптимального выбора для оптимистичной и ϵ-greedy стратегий для константного step-size, сгруппировано по стратегиям.

выбросы меньше влияют на результат. С уменьшением ν разрыв в проценте оптимальных действий для для $\alpha = 0.4$ с baseline и $\alpha = 0.1$ без baseline сокращается, а для t_1 и вовсе почти совпадают, что дает говорить о том, что с уменьшением числа степеней свободы влияние baseline уменьшается. Впрочем, при $\nu < 1$ само понятие выборочного среднего как приближение матожидания перестает иметь смысл. Аналогично предыдущим пунктам, для t_2 и t_1 значение метрик ниже, чем для t_3 и t_∞. Причем при увеличении α в ситуации с присутствием baseline значения метрик для t_∞ становятся ниже, чем для t_3, при отсутствии baseline ситуация ровно наоборот.

В целом, характерны следующие тенденции:

1. Везде, кроме greedy стратегий и стратегий с постоянным step-size, при увеличении числа степеней свободы значения метрик увеличиваются (или незначительно уменьшаются).

2. У t_2 имеются резкие перепады в средней награде с сохранением тенденции к увеличению (или уменьшению для постоянного step-size). Это объясняется отсутсвием у t_2 дисперсии.

3. Везде, кроме greedy стратегий и стратегий с постоянным step-size, занчения метрик для t_∞ и t_3 почти совпадают, то есть стратегии одинаково применимы для этих двух распределений.

Наконец, проанализируем графики со сравнениями всех стратегий. Для t_∞ средняя награда достигает наибольшего значения для оптимистичной инициализации, gradient bandits и UCB, а

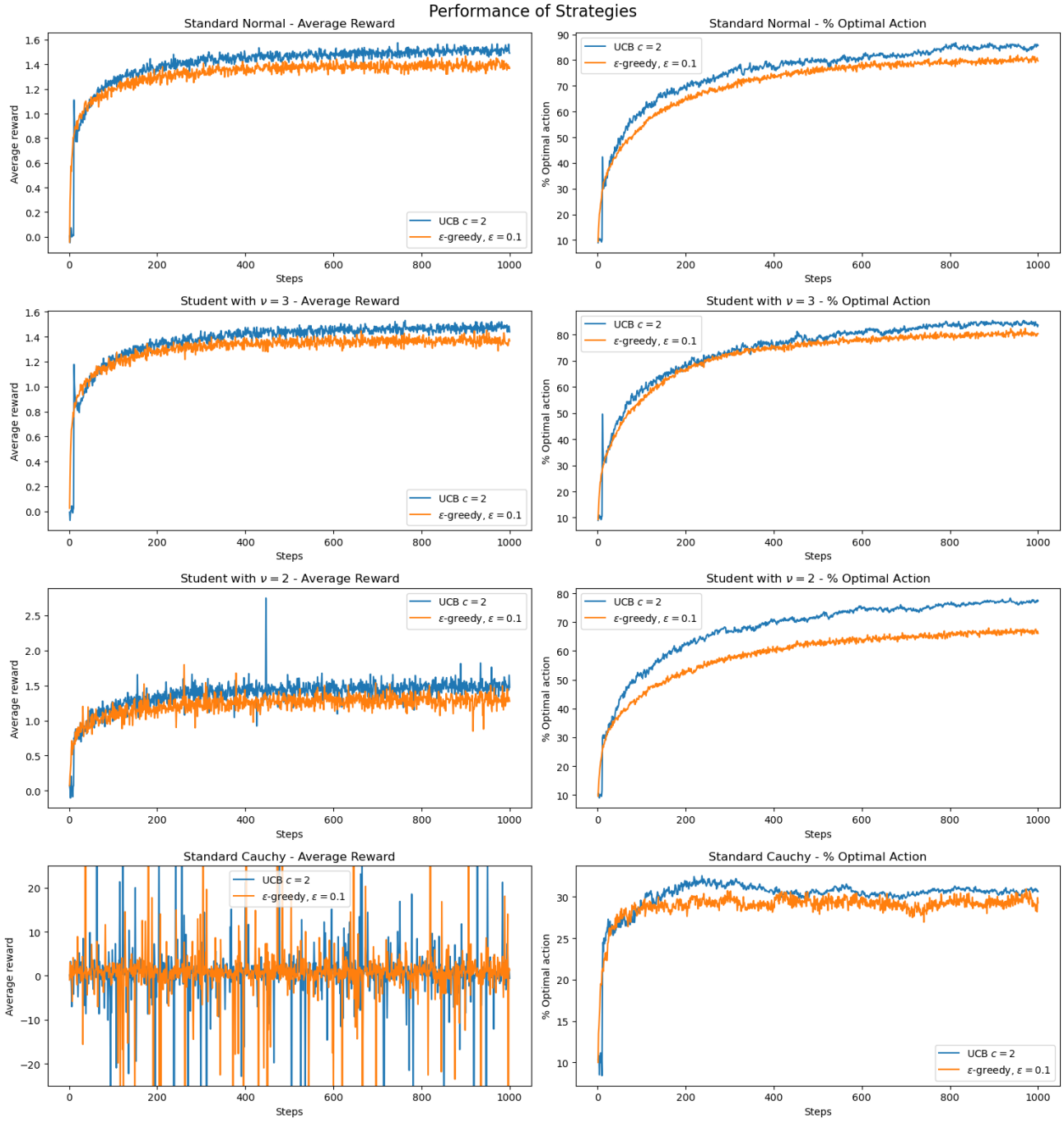

Рис. 5: Значения средней награды и процента оптимального выбора для UCB в сравнении с ϵ-greedy, сгруппировано по распределениям. UCB лучше ϵ-greedy на всех распределениях

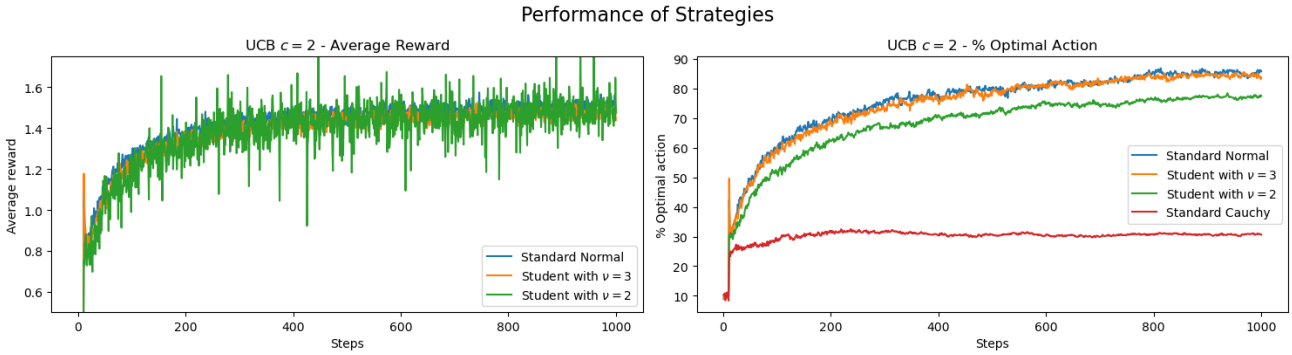

Рис. 6: Значения средней награды и процента оптимального выбора для UCB, сгруппировано по стратегиям.

процент оптимальных действий – для UCB. Для t_3 ситуация такая же, за исключением того, что средняя награда для ϵ-greedy достигает примерно таких же значений, что и для других стратегий. Для t_2 средняя награда и процент оптимальных действий для постоянного step-size значительно падают относительно других стратегий, что можно объяснить приданием одинакового значения выбросам выборки для любого шага. Средняя награда лучшая для ϵ-greedy и UCB, процент оптимальных действий лучший для UCB и gradient bandits. Как я уже говорил, график средней награды для распределения Коши не имеет смысла, для процента оптимальных действий лучшие значения достигаются для ϵ-greedy и UCB, при этом значения метрики для gradient bandits значительно падают относительно других стратегий и даже хуже, чем для стратегии с постоянным step-size. Так происходит, поскольку baseline есть среднее по всем предыдущим шагам, что не сходится ни к какому значению при любом распределении выбора действий.

В целом UCB на всех распределениях показывает лучший или один из лучших результатов, что можно объяснить тем, что эта стратегия дает любому рычагу воможность быть выбранным, при этом эта вероятность быть выбранным уменьшается с увеличением числа шагов и при этом в меньшей степени привязана к матожиданию, как, например, gradient bandits. Так как так же дается ненулевая вероятность быть выбранным каждому рычагу, ϵ-greedy и gradient bandits тоже дают хорошие результаты на t_2, t_3, t_∞.

5 Выводы

Проделанные эксперименты позволяют судить о том, что Gradient bandits, ϵ-greedy и UCB – стратегии показывают высокую эффективность на степенных распределениях. Так как UCB – единственная из стратегий, показывающая высокую эффективность на всех метриках и всех распределений, то эта стратегия – лучший из кандидатов для применения в оптимизации портфолио в модели прироста стоимости акций как многоруких бандитов.

Список литературы

[BP03] Jean-Philippe Bouchaud and Mark Potters. Theory of financial risk and derivative pricing. pages 202–203, 2003.

[SB18] Richard S. Sutton and Andrew G. Barto. Reinforcemnt Learning: An Introduction, second edition. pages 47–68, 2018.

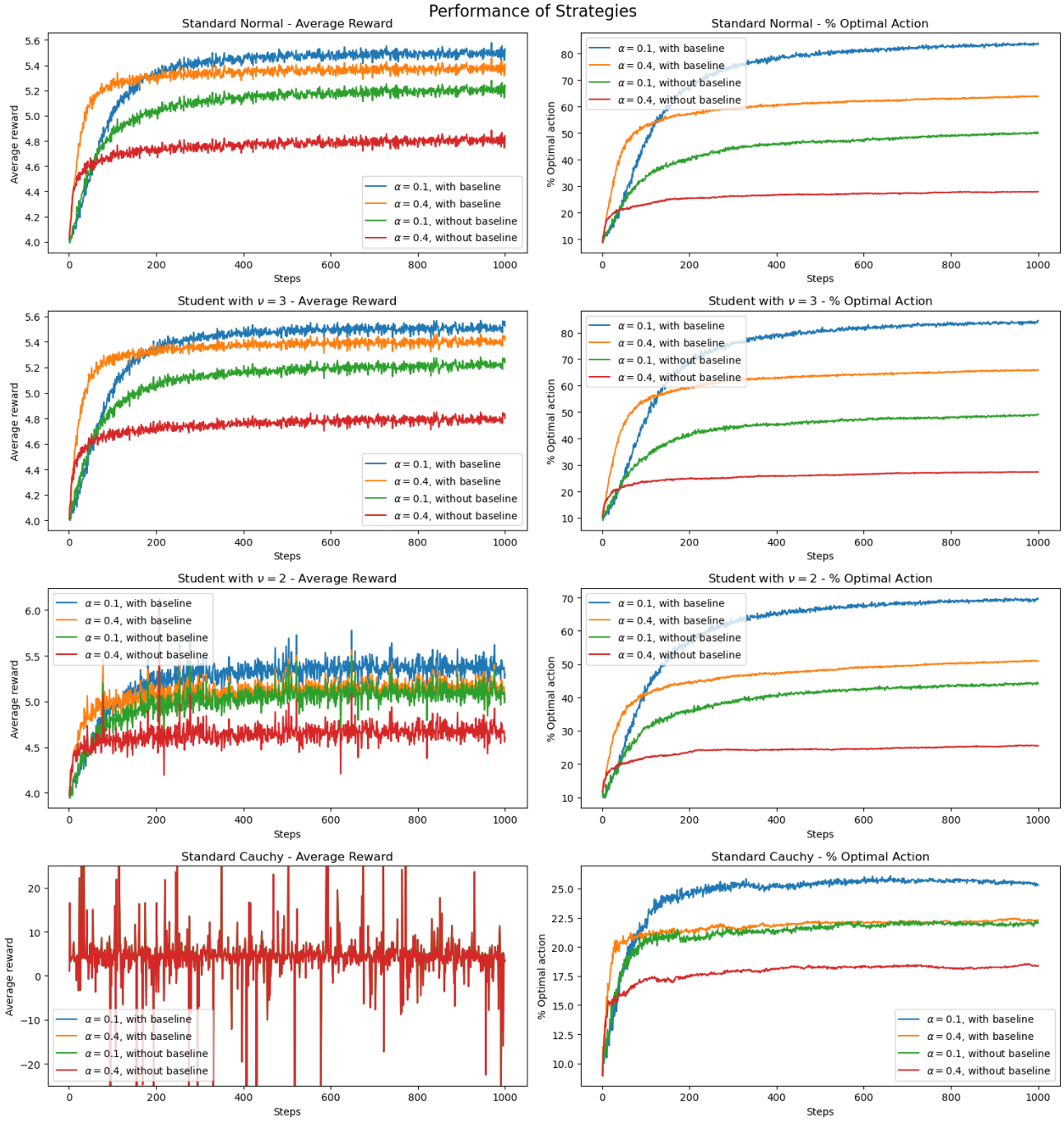

Рис. 7: Значения средней награды и процента оптимального выбора для gradient bandits, сгруппировано по распределениям. Меньшее α и присутствие baseline приводят к лучшим результатам

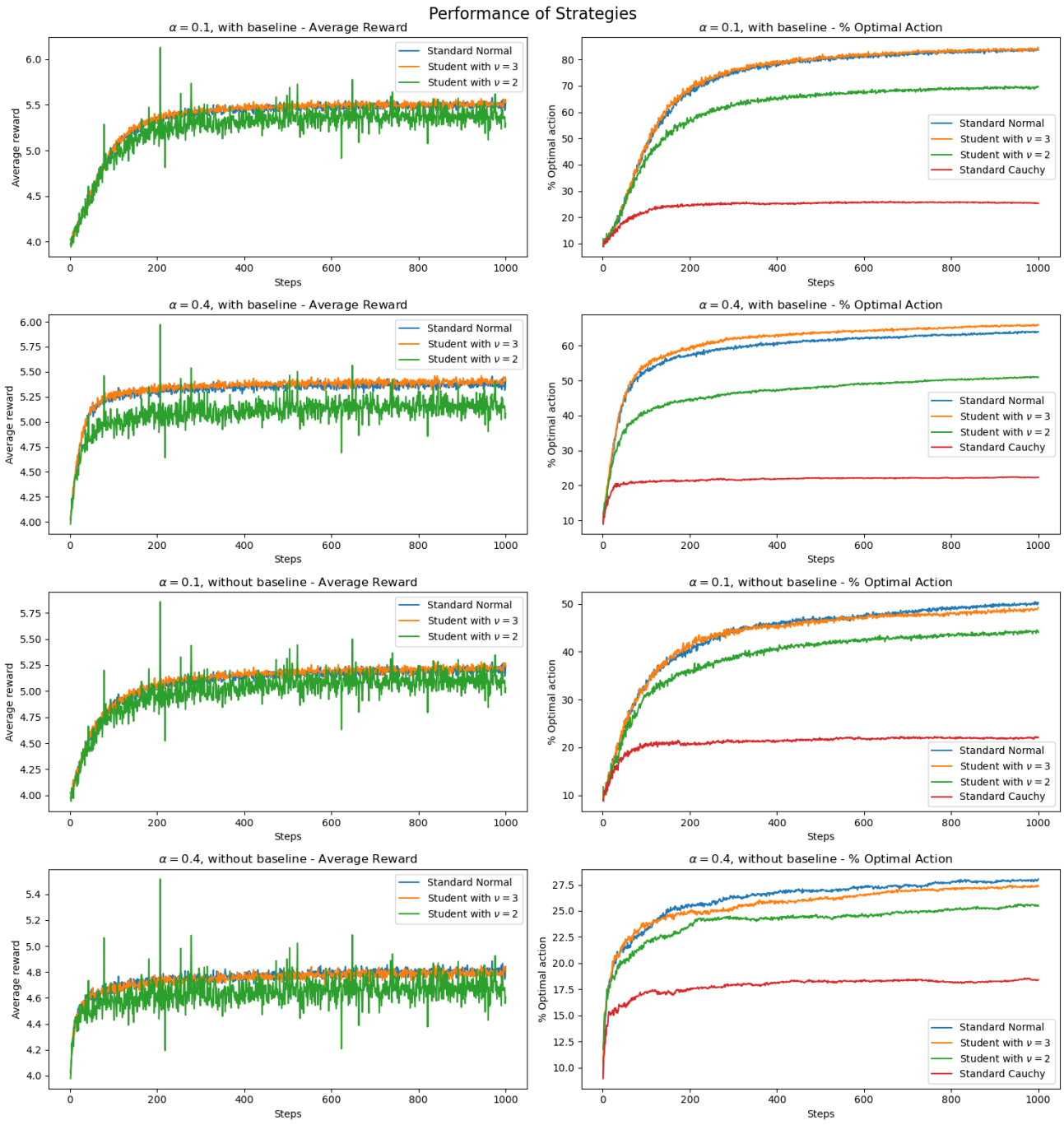

Рис. 8: Значения средней награды и процента оптимального выбора для gradient bandits, сгруппировано по стратегиям. Заметно странное поведение метрик: когда есть baseline, значения для t_3 лучше, чем для t_∞

.

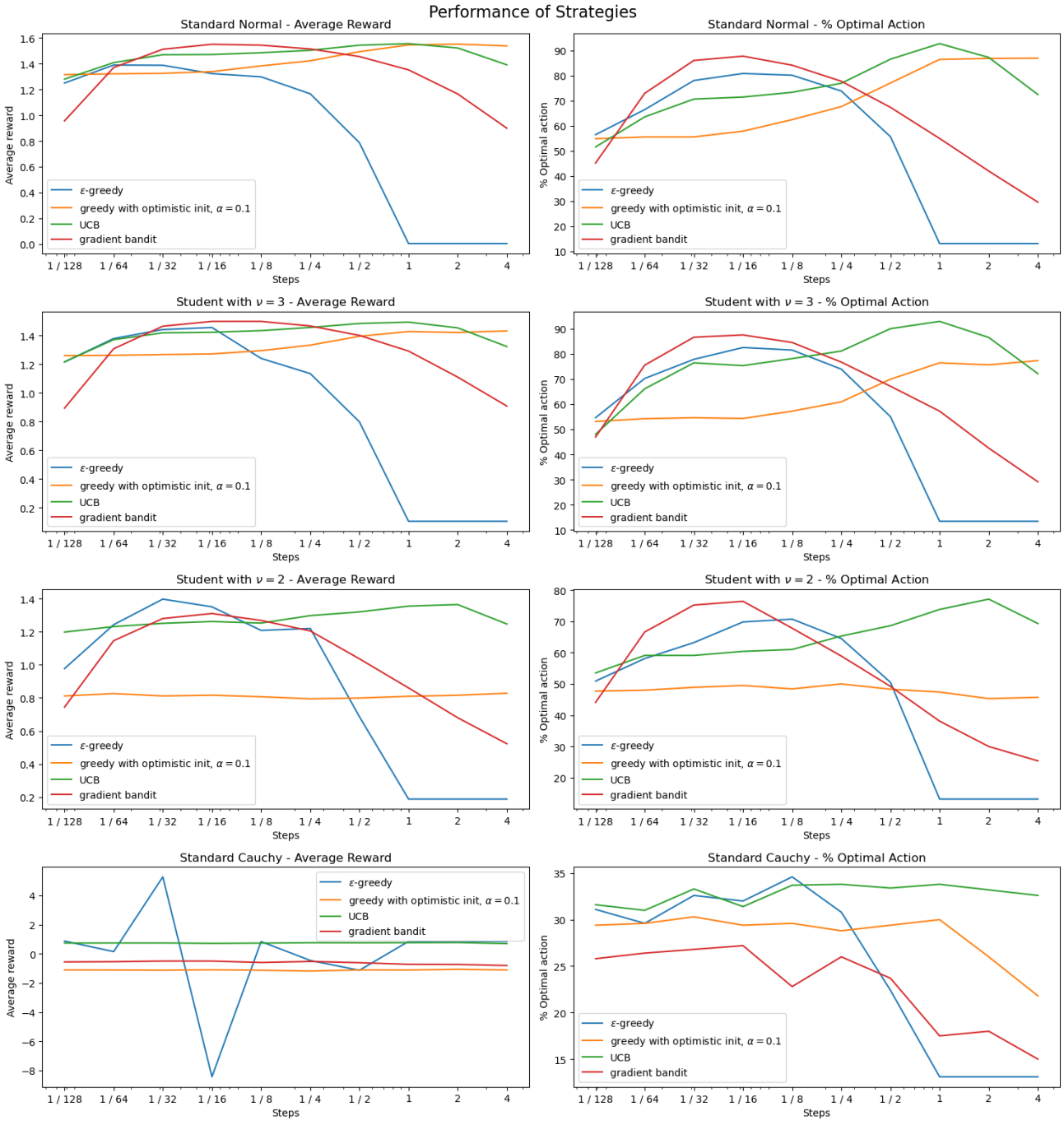

Рис. 9: Сравнение всех стратегий при варьировании гиперпараметров