ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

**Факультет информатики, математики и компьютерных наук**

Программа подготовки магистров по направлению

01.04.02 Прикладная математика и информатика

Богатова Маргарита Дмитриевна

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА МАГИСТРА**

на тему:

**«Исследование построения агентов на основе метода Active Inference»**

**Научный руководитель:**

профессор кафедры прикладной математики и информатики

Крылов Владимир Владимирович

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Подпись

**Рецензент:**

д.ф.м.н, профессор кафедры прикладной математики и информатики

Калягин Валерий Александрович

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Подпись

Нижний Новгород  
2022

АННОТАЦИЯ

Определения, обозначения, сокращения

AIF – active inference – активный вывод

dAIF – deep active inference – глубокий активный вывод.

EFE

VFE

PCM

MDP

POMDP

MCTS

ОГЛАВЛЕНИЕ

[АННОТАЦИЯ 2](#_Toc104157536)

[Определения, обозначения, сокращения 3](#_Toc104157537)

[ОГЛАВЛЕНИЕ 4](#_Toc104157538)

[Введение 5](#_Toc104157539)

[Глава 1. Обзор литературы и постановка задачи 8](#_Toc104157540)

[1.1. Вариационная свободная энергия 9](#_Toc104157541)

[1.2. Ожидаемая свободная энергия 10](#_Toc104157542)

[1.3. Связь с обучением c подкреплением 12](#_Toc104157543)

[1.4. Активный вывод в задачах непрерывного управления 12](#_Toc104157544)

[1.5. Глубокий активный вывод (dAIF) 13](#_Toc104157545)

[1.6. Анализ существующих работ 14](#_Toc104157546)

[Глава 2. Предлагаемое решение задачи 17](#_Toc104157547)

[Глава 3. Реализация и эксперименты 18](#_Toc104157548)

[Заключение 19](#_Toc104157549)

[Список литературы 20](#_Toc104157550)

Введение

Активный вывод (AIF) — это математическая структура, возникшая в вычислительной нейронауке как теория того, как мозг реализует действие, восприятие и обучение [19]. В основе активного вывода лежит принцип минимизации свободной энергии, предполагающий в что агенты действуют, противодействуя силам окружающей среды, которые мешают им оставаться в ограниченном наборе желаемых состояний мира, и минимизируют неожиданность сенсорных наблюдений за счет минимизации аппроксимации этой неожиданности, которая называется свободной энергией. Согласно этому предположению, биологические агенты развивают различные навыки, такие как восприятие, действие, планирование и обучение, которые агенты постоянно адаптируют на протяжении всей своей жизни.

Активный вывод и принцип свободной энергии использовались для объяснения и моделирования нескольких сложных процессов в различных дисциплинах. Например, в психологии они использовались для распознавания эмоций [1]. В экономике принцип свободной энергии использовался для формулирования процесса оптимизации агентов с точки зрения их когнитивных искажений [1].

Недавние разработки в области глубокого обучения открыли новые горизонты для изучения и экспериментирования с различными теориями восприятия и поведения. Одним из популярных примеров в этом отношении является обучение с подкреплением (RL) [20], теория, которая связывает сигналы дофамина в мозге с сигналами вознаграждения, которые можно использовать для подкрепления правильного поведения [21], и описывает, как поведение агента может быть оптимизировано через максимизацию вознаграждения. Сочетание RL с моделями глубокого обучения для оценки функций позволило достичь успехов в построении искусственных агентов. Такие агенты нашли широкое применение в видеоиграх [22], настольных игры [23, 24] или в управлении роботами [25, 26]. Методы глубокого обучения также начинают возникать в контексте активного вывода [5,6,7]. Интегрированный с машинным обучением активный вывод называют deep active inference (dAIF).

Обучение с подкреплением и активный вывод призваны решать одни и те же задачи, которые обычно формулируются как MDP или POMDP. Ввиду этой образующейся связи, активный вывод сейчас часто применятся для решения задач именно в такой формулировке, что позволяет внедрять передовые разработки в области RL в AIF. Как следствие и использовании глубокого обучения, dAIF достигают сравнимых с RL результатов.

Современные задачи управления роботами поставленные как POMDP\MDP успешно решаются с применением методов обучения с подкреплением. В свою очередь в подавляющем большинстве работ по активному выводу рассматриваются игрушечные дискретные задачи. Вектор исследования в области активного вывода сейчас направлен в сторону расширения области применения на задачи с непрерывным пространством действий и состояний [8, 10, 12].

На данный момент есть ряд исследований, в которых активный вывод применятся для задач с пространством состояний большой размерности [5,6,7], однако исследований для многомерного пространства и состояний, и действий еще очень мало и многие из них не используют все возможности глубокого обучения. Таким образом **задача** адаптации активного вывода для непрерывного пространства действий, решающая **проблему** масштабирования этого метода для комплексных задач таких как управление роботом является **актуальной**.

**Объектом** исследования являются алгоритмы активного вывода, а также методы глубокого обучения для управления роботом. **Предметом** работы является исследование применения теории активного вывода для реальных задач большой размерности. **Цель** исследования состоит в изучении существующих подходов к построению AIF агентов и разработка на их основе алгоритма для решения задачи управления роботом.

Для достижения поставленной цели требуется решить следующие **задачи**:

1. Изучение теории активного вывода.
2. Исследование научных источников, посвященных проблеме масштабирования активного вывода.
3. Разработка алгоритма на основе активного вывода для решения задачи управления роботом.
4. Реализация этого алогизма и проведение экспериментов для оценки качества работы агента, а также определения лучшей его конфигурации.

В процессе работы использовались следующие **методы**: анализ литературных источников, описывающих передовые достижения в области глубокого активного вывода, и обобщение полученной информации; проведение экспериментов по решению различных задач управления роботом dAIF агентом и сравнение результатов, визуализация и анализ данных результатов проведенных экспериментов.

**Новизна** работы заключается в том, что в ней:

* Предложен алгоритм для обучения dAIF агента для окружений с многомерными разнотипными данными и непрерывными действиями.
* Обученный dAIF агент успешно применен для решения задачи управления роботом.

**Практическая значимость**: построенный алгоритм может быть использован в других прикладных задачах, а также может помочь исследователям в будущем разрабатывать новые подходы.

Работа имеет следующую **структуру**: в первой главе содержится описание основных принципов активного вывода, а также обзор и анализ существующих AIF работ для его масштабирования. Во второй главе приводится ход разработки алгоритма для решения задачи. Третья глава содержит описание окружений, конкретных задач, а также проведенных экспериментов. В заключении содержаться выводы о проделанной работе.

Глава 1. Обзор литературы и постановка задачи

Теория активного обучения была предложена К.Дж. Фристоном как единая теория мозга Главная концепция активного обучения состоит в минимизации функционала свободной энергии.

В активном выводе предполагается, что окружение, в котором действует агент описывается как частично наблюдаемый марковский процесс принятия решений (POMDP), который обозначается в виде набора M = {S, A, T, Ω, O, γ}, где S — множество ненаблюдаемых состояний, A — множество действий, T — функция перехода между состояниями, также называемая динамики окружающей среды, Ω — множество наблюдений, O — множество условных вероятностей наблюдения, γ — коэффициент температуры. Такая формальность обусловлена тем, что агент не имеет «доступа» к истинному состоянию мира. Используя прошлые наблюдения, он строит **убеждение** об истинном состоянии.

Агент получает наблюдения из среды, имеющей скрытые состояния, которые агент пытается вывести. Он также может предпринимать действия, которые изменяют состояние среды и, следовательно, будущие наблюдения.

Поведение агента в среде определятся его генеративной моделью своих наблюдений, состояний и действий, которые можно факторизовать следующим образом:

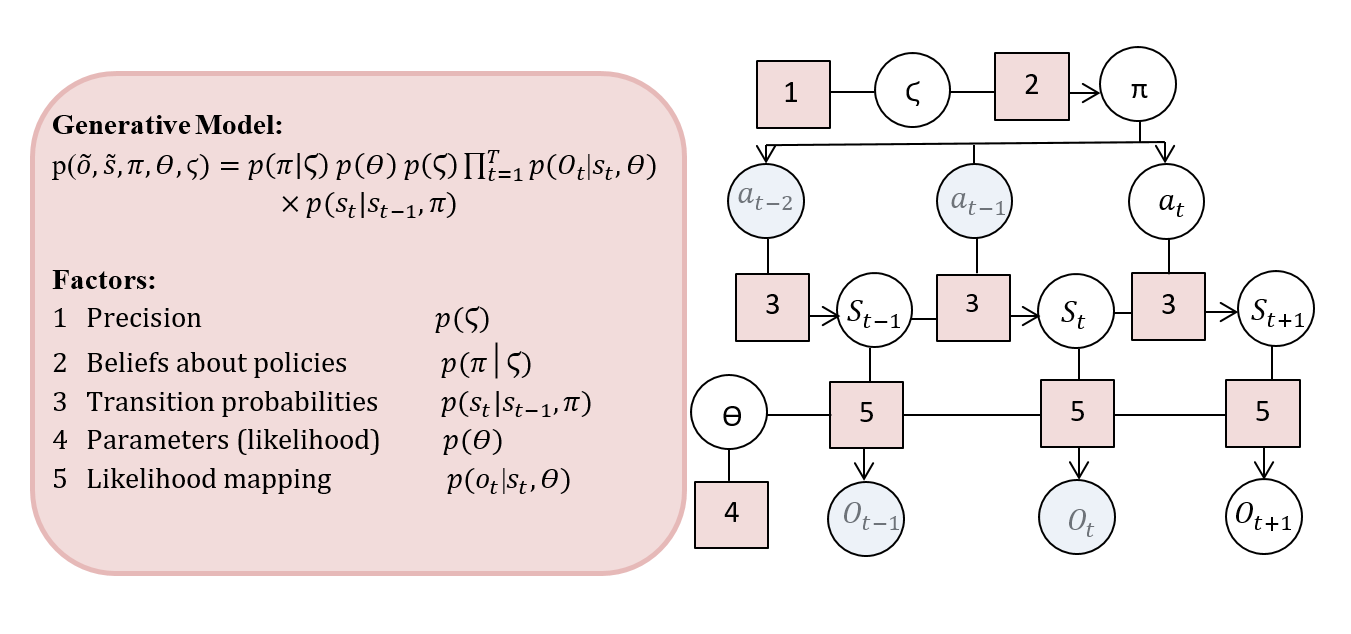


Рисунок 1. Генеративная модель в виде фактор-графа.

Путем оптимизации генеративной модели агент улучшает свою способность выяснить, какое действие привело к наблюдению. В каждый момент времени агент вычисляет ожидаемую свободную энергию (EFE) каждой политики (траектории, последовательности дейсвтий), которая количественно определяет вариационную свободную энергию (VFE) и изменяет среду, чтобы свести к минимуму VFE.

* 1. Вариационная свободная энергия

Ключевой смысл принципа свободной энергии состоит в том, что агент стремится найти равновесие между своей внутренней моделью и внешним миром. В терминах Баесовского вывода, агент максимизирует обоснованность (evidence) модели, приближая внутреннее состояние, описываемое плотностью q(s), к апостериорному миру p(s|o). Помимо интуиции, введение вариационного распределения необходимо по причине того, что p(s|o) в большинстве реальных случаев нельзя посчитать точно.

Для того, чтобы подобрать лучшую аппроксимацию q(s) минимизируют расстояние Кульпака-Лейблера (KL) между распределениями q(s) и p(s|o). Для простоты записи, рассмотрим упрощенную генеративную модель .

Из этого определения вытекает формула вариационной свободной энергии (VFE):

Эта формула представляет собой функцию внутренних убеждений агента об окружающей среде, и означает верхнюю границу неожиданности наблюдений. VFE измеряет меру расхождения между вариационной плотностью q(s) и генеративной моделью .

Далее, формула свободной энергии может быть переписана следующим образом:

С одной стороны, минимизация свободной энергии подразумевает максимизацию точности (второе уравнение, второе слагаемое) при одновременном снижении сложности апостериорного распределения (второе уравнение, первое слагаемое). С другой стороны, это подразумевает оптимизацию обоснованности модели (первое уравнение).

Очередной раз преобразуем формулу: используем условие марковского процесса о том, что распределение вероятностей будущих состояний процесса зависит только от нынешнего состояния и заменим на .

Это наиболее часто используемый в литературе вид формулы свободной энергии для урезанной генеративной модели. В статье [5] была выведена формула для генеративной модели, изображенной на схеме 1, которая берет во внимание распределения действий агента. Выглядит она следующим образом:

*Воспользуемся формулой расстояния Кульпака-Лейблера и :*

* 1. Ожидаемая свободная энергия

«Активной» составляющей агента активного вывода является взаимодействие с окружением. Теперь необходимо взять во внимание то, что агент совершает действия. Для того, чтобы действовать в среде наиболее оптимально, в подходе предлагается помимо минимизации вариационной свободной энергии, также минимизировать ожидаемую свободную энергию (EFE), которая представляет собой комбинацию VFE всех ожидаемых траекторий в будущем с учетом текущего состояния и действия.

В общем виде формула выглядит так:

Стоит отметить, что в стандартной литературе по активному выводу рассматривается генеративная модель , где называется политикой, однако далее будет выведена упрощенная формула, на основе предположения что .

Для вывода формулы рассмотрим генеративную модель . Наиболее критичным для определения распределением является которое означает предпочтения агента, или другими словами наблюдения, которые агент ожидает получить.

*=*

*Допустим, что*

*Пусть действия распределены равномерно:*

*Перегруппируем части уравнения:*

Интуитивно, выведенная формула означает, что агент будет выбирать такие действия, для которых наблюдения содержат больше информации о состояниях, которые будут вести к получению предпочтительных результатов и в то же время сохраняя действия настолько вариативными, насколько это возможно.

Таким образом, мы определили приближенные формулы, которые обычно являются основой для описания различных подходов. Теперь определим основные для активного вывода понятия:

* – модель переходов (*transition model*).
* – предпочтения агента *(agent preferences)*.
* - условное распределение действий, обычно в литературе называет моделью действий.
* – модель наблюдений *(observation model)*. Также в литературе называется репрезентацией. Предполагается, что генеративная модель агента способна генерировать воображаемые наблюдения.
  1. **Связь с обучением c подкреплением**

Активный вывод был предложен как альтернатива обучению с подкреплением. Эти подходы призваны решать одни и те же задачи, и за иногда даже оперируют одними и теми же формулами, однако идеи, лежащие в основе, у них разные. Например, концепция «вознаграждения» в активном выводе сильно отличается от вознаграждения в RL, поскольку вознаграждение — это не сигналы, используемые для привлечения агента, а скорее сенсорные состояния, которые агенты стремятся часто посещать, чтобы минимизировать свою свободную энергию. Авторы активного обучения утверждают, что агенты здесь функция награждения совсем не обязательна и агенты смогут изучать свою собственную.

* 1. Активный вывод в задачах непрерывного управления

Фокус этой работы направлен на применение активного вывода для решения задачи непрерывного управления. Особенности этой задачи состоит в том, что при ее рассмотрении как POMDP, внутреннее состояние системы и действия задаются следующим образом: и . Например, **задача управления роботизированной** рукой – наблюдения и действия в такой модели — это вектора, значения которого варьируются от -1 до 1, которые внутри симуляции трансформируются в углы между подвижными частями робота.

Реализация агента активного вывода, который бы действовал в окружении с непрерывным пространством состояний и действий на текущий момент открытая проблема. Потому что стандартные [13] формулы для расчета и обучения AIF агента применимы только для дискретных случаев. Так, в [13, 14] плотности в уравнении свободной энергии задаются матрицами. Например, кодируется набором матриц для каждого возможного действия, строки и столбцы которых задают плотность для каждой возможной пары состояний.

Еще более сложная ситуация с непрерывным пространством действий. Чтобы сделать вывод о своей политике, агент должен вычислить ожидаемую свободную энергию (EFE) для каждой траектории. На практике это означает, что для *каждой* *возможной* *политики* агенту необходимо на определенный горизонт времени вперед сгенерировать состояния и наблюдения, в которых он будет, если будет следовать этой политике. И далее вычисляется свободная энергия этих состояний и оценивается ожидаемая свободная энергия для каждой политики. Когда доступных политик и действий неопределенно много, такой способ подсчета ожидаемой свободной энергии также становится экспоненциально сложным.

Такой подход называют **табличным активным выводом**, и он был довольно хорошо изучен [15,16,17]. Однако эти методы, хотя и дают представление об идее активного вывода и важности различных параметров, не масштабируются из-за их экспоненциальной сложности, и они не применялись ни к чему, кроме простых игрушечных задач.

* 1. Глубокий активный вывод (dAIF)

В некоторых недавних научных работах были предприняты попытки масштабирования активного вывода для более комплексных и реальных задач. В последние годы для построения MDP/POMDP агентов большую популярность имеют методы глубокого обучения с подкреплением. Именно применение нейросетей позволило достигнуть значительных успехов в этой области. Эта тенденция не обошла стороной и теорию активного вывода - использования нейронных сетей для аппроксимации плотностей положило начало развитию активного обучения в этом направлении.

Насколько это известно, [4] является первой работой, в которой предлагается аппроксимировать модели наблюдений и переходов с помощью глубоких нейронных сетей. Авторы представляют все плотности в виде диагонального гаусовского распределения, а для аппроксимации параметров этих распределений используют однослойные сети с шестнадцатью нейронами. Также именно они ввели понятие *deep active inference* или *dAIF.*

Большое влияние на область dAIF также оказала работа [5]. Авторы использовали последние на то время разработки в области обучения с подкреплением и создали смешанный алгоритм. Основываясь на этом подходе были разработаны такие алгоритмы как [6] и [7]. Вкратце, авторы аппроксимировали с помощью нейронных сетей модели наблюдений и переходов, но что самое интересное, позволили агенту самому предсказывать значение ожидаемой свободной энергии для сразу всех доступных действий, и сделано это тем же способом, как это сделано в обучении с подкреплением при аппроксимации Q-функции.

Возвращаясь к проблеме масштабирования активного вывода, далее статьи будут рассматриваться с точки зрения того, какие именно методы предлагаются и предлагаются ли вообще для работы с непрерывными пространствами состояний и действий. Так, в упомянутых ранее [5,6,7] реализуют масштабирование модели наблюдений с помощью вариационного автоэнкодера, однако VFE считают для всех возможных действий.

Наиболее интересной в этом плане является работа [8], в которой авторы преследуют такие же цели что и в текущем исследовании. Все плотности они рассматривают как нормальные распределения, а их параметры аппроксимируют с помощью нейронных сетей. Модель наблюдений строится на основе вариационного автоэнкодера, что позволяет применять метод в случае высокоразмерных наблюдений (например, изображений). Ключевой особенностью здесь является то, как реализован выбор политики и планирование. В каждый момент времени они выбирают такое действие, которое будет математическим ожиданием модели действий, которая также представлена как нормальное распределение. Для это из нее семплируются конечная последовательность действий, для каждой такой последовательности производится оценка свободной энергии и далее модель действий оптимизируется с помощью метода кросс-энтропии (CEM) [9]. Авторы показывают, что такой вид dAIF является конкурентным с передовыми методами. Некоторым продолжением и в своем роде упрощением предыдущего подхода стала работа [10], в которой авторы попробовали усилить алгоритмы обучения с подкреплением с помощью идей активного обучения. На основе принципа ожидаемой свободной энергии была разработана целевая функция, минимизируя которую мы могли бы получить оптимальную модель действий. Благодаря такому упрощению, этот подход можно использовать и для задач с большим входом и непрерывными действиями.

В исследовании [11] попытались оптимизировать подход к выбору оптимального действия (то есть минимизации ожидаемой свободной энергии в пределах некоторого горизонта времени), чтобы избежать полного перебора всех возможных траекторий. Авторы предлагают использовать MCTS для выбора оптимального действия в дереве решений, которое образуется, если рассматривать задачу минимизации ожидаемой свободной энергии для конечного числа возможных действий.

Наконец, из наиболее релевантных последних исследований выделяется [12], авторы которого адаптировали стандартный алгоритм обучения с подкреплением - актер-критик - для предсказания действий, и тем самым создали модель действий, которая применима для непрерывного пространства. Помимо этого, было предложено использовать так называемое контрастное обучение, а в качестве репрезентации использовали обычную нейронную сеть. Как результат, алгоритм, который работает с непрерывным пространством состояний и действий, превосходящий другие dAIF методы на дискретных задачах.

* 1. Анализ существующих работ

Резюмируя обзор статей из прошлого пункта, выделим основные подходы к масштабированию моделей наблюдений и действий.

Так, **для репрезентации** наблюдений большой размерности самым популярным подходом оказался вариационный автоэнкодер, и это оправдано – так как этот метод идеально подходит для аппроксимации с теоретической точки зрения, и при этом является не требовательным к виду входных данных – наблюдения могут быть как изображениями, так и, например, текстом или любым числовым вектором. Однако эта модель довольно тяжела для обучения, и поэтому некоторые исследования [12, 10] стремятся найти другие архитектуры сетей.

С вычислительной точки зрения наиболее сложный аспект минимизации ожидаемой свободной энергии состоит в том, как выбрать действия, которые в будущем приведут к более правильному убеждению агента. На практике оптимизация VFE трансформируется в поиск по дереву среди наиболее вероятных политик. Тем не менее, поиск может быть дорогостоящим, особенно в многомерных пространствах и с непрерывными действиями.

В обзоре [1] выделяют три пути для организации планирования и выбора действия:

1. Наивная реализация активного вывода, где политика представляет собой некоторый план действий . Каждому такому плану ставится в соответствие ожидаемая свободная энергия, и следующее действие выбирается из лучшего плана в соответствии с уравнением общим уравнением, описанным в пункте 2.

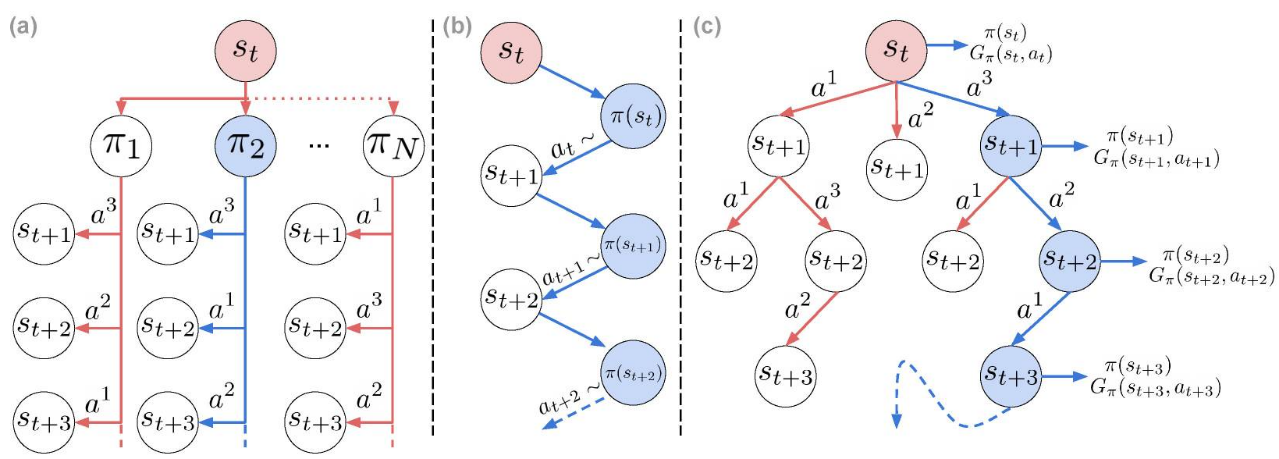


Рисунок 2. Различные подходы для выбора оптимального действия. Источник [1].

1. Построение отображения состояние-действие (), которое также называется политикой привычек. Работает по принципу нахождения оптимальной политики, которая бы обеспечивала ожидаемое наилучшее действие для каждого состояния. Эта формулировка была вдохновлена методами в обучении с подкреплением и применялась в работах [5,6,7,12].
2. Объединение первых двух вариантов – сначала для конкретного состояния и действия оценивается ожидаемая свободная энергия, а затем производится поиск по дереву с помощью эвристических алгоритмов.

Оценим три описанных подхода с точки зрения выбора оптимального действия для непрерывного управления. Первый, наивный, вариант уже был обсужден – перебор всех возможных траекторий неэффективен. Третий вариант - это также в своей наивной реализации требует перечисления всех возможных действий, что также невозможно в контексте рассматриваемой задачи. Обычно исследования, которые применяют этот подход, используют MCTS [11]. Однако, недавно были разработаны расширения MCTS для комплексных пространств действий [18]. А это значит, что *третий вариант* также *потенциально может быть применен* для задачи управления роботом. Однако, наиболее подходящим вариантом кажется второй – потому что в текущих передовых RL подходах наиболее успешными являются алгоритмы, где непрерывное вещественное действие так или иначе предсказывается нейронной сетью. То есть перебор действий и как следствие долгосрочное планирование отсутствуют.

Таким образом, за основу решения задачи управления роботом в текущей работе возьмем вариационный автоэнкодер как модель наблюдений и политику привычек как модель действий. Такой выбор основывается на том, ***во-первых***, автоэнкодеры широко используются и позволяют работать с любыми типами наблюдений, и ***во-вторых***, именно такой подход к выбору оптимального действия агентом доказал свою эффективность в области обучения с подкреплением.

Также важно отметить, что среди рассмотренных алгоритмов глубокого активного вывода лишь два [10, 12] были разработаны с перспективой применения их для комплексных задач и по факту применили свою разработку к таким задачам и показали результаты. Однако более сложные задачи, такие как управление роботом, еще не рассматривались в данном контексте. Из этого и вытекает цель этой работы – разработка алгоритма глубокого активного вывода для непрерывных пространств действий и решение с помощью него задачи управления роботом.

Глава 2. Предлагаемое решение задачи

In the current paper, we take an alternative approach and employ amortized inference [24], which utilizes function approximators (i.e., neural networks) to parameterize distributions. Free energy is then minimized with respect to the parameters of the function approximators, and not the variational parameters themselves

In our model we do not use this approach, but instead use a learned amortized inference distribution Q(a|s) and minimize this using a variational approach on the divergence with the approximated true posterior of the value function p(a|s),which is learned through a bootstrapping estimation procedure. Due to this our method is end-to-end differentiable and all networks can be trained through gradient descent on the variational free-energy.

Примеры существующих решений этой проблемы в аи, ее результаты и недостатки.

Существующие решения в целом в реинфорсмент лернинг

Одно из отличий pomdp и scaling active inference – это то что они исключают часть связанную с q(a|s) из уравнения свободной энергии. И его оптимизируют уже через метод кросс энтропии или msts

Глава 3. Реализация и эксперименты

In Reinforcement Learning, you really shouldn't typically be paying attention to the precise values of your loss values. They are not informative in the same sense that they would be in, for example, supervised learning. The loss values should only be used to compute the correct updates for your RL approach, but they do not actually give you any real indication of how well or poorly you are doing.

**This is because in RL, your learning targets are often non-stationary; they are often a function of the policy that you are modifying (hopefully improving!)**. It's very well possible that, as the performance of your RL agent improves, your loss actually increases. Due to its improvement, it may discover new parts of its search space which lead to new target values that your agent was previously completely oblivious to.

Your only really reliable metric for how well your agent is doing is the returns it collects in evaluation runs.

Заключение

Список литературы

1. Pietro Mazzaglia, Tim Verbelen , Ozan Çatal and Bart Dhoedt, The Free Energy Principle for Perception and Action: A Deep Learning Perspective. 2022.
2. Pablo Lanillos, Cristian Meo, Corrado Pezzato, Ajith Anil Meera, Mohamed Baioumy, Wataru Ohata, Alexander Tschantz, Beren Millidge, Martijn Wisse, Christopher L. Buckley, and Jun Tani. Active Inference in Robotics and Artificial Agents: Survey and Challenges. 2021.
3. How Active Inference Could Help Revolutionise Robotics 2022
4. Ueltzhöffer K. Deep active inference. Biol. Cybern. 2018;112:547–573. doi: 10.1007/s00422-018-0785-7
5. Millidge B. Deep Active Inference as Variational Policy Gradients. arXiv. 2019 doi: 10.1016
6. Aleksey Zelenov, Vladimir Krylov, Deep active inference in control tasks, 2021.
7. Otto van der Himst, Pablo Lanillos, Deep Active Inference for Partially Observable MDP
8. Alexander Tschantz, Manuel Baltieri, Anil. K. Seth, Christopher L. Buckley. Scaling active inference. 2019
9. Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. arXiv:1811.04551 [cs, stat], 2018
10. Alexander Tschantz, Beren Millidge, Anil K. Seth, Christopher L. Buckley. Reinforcement Learning through Active Inference 2020
11. Zafeirios Fountas, Noor Sajid, Pedro A.M. Mediano, Karl Friston. Deep active inference agents using Monte-Carlo methods, 2020
12. Pietro Mazzaglia, Tim Verbelen, Bart Dhoedt, Contrastive Active Inference. 2022
13. Ryan Smith, K. Friston, Christopher Whyte. A Step-by-Step Tutorial on Active Inference and its Application to Empirical Data. 2021.
14. Noor Sajid, Philip J. Ball, Thomas Parr, Karl J. Friston. Active inference: demystified and compared. 2020.
15. Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. Cognitive neuroscience, 6(4):187–214.
16. Friston, K., Samothrakis, S., and Montague, R. (2012). Active inference and agency: optimal control without cost functions. Biological cybernetics, 106(8-9):523–541
17. Parr, T. and Friston, K. J. (2018b). The anatomy of inference: Generative models and brain structure. Frontiers in computational neuroscience
18. Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Mohammadamin Barekatain, Simon Schmitt, David Silver. Learning and Planning in Complex Action Spaces. 2021
19. Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, O Doherty J, Pezzulo G. Active inference and learning. Neurosci Biobehav Rev. 2016 Sep;68:862-879. doi: 10.1016
20. Sutton R.S., Barto A.G. Reinforcement Learning: An Introduction. MIT Press; Cambridge, MA, USA: 2018
21. Glimcher P.W. Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. Proc. Natl. Acad. Sci. USA. 2011
22. Mnih V., Kavukcuoglu K., Silver D., Graves A., Antonoglou I., Wierstra D., Riedmiller M. Playing Atari with Deep Reinforcement Learning
23. Silver D., Hubert T., Schrittwieser J., Antonoglou I., Lai M., Guez A., Lanctot M., Sifre L., Kumaran D., Graepel T., et al. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm
24. Schrittwieser J., Antonoglou I., Hubert T., Simonyan K., Sifre L., Schmitt S., Guez A., Lockhart E., Hassabis D., Graepel T., et al. Mastering Atari, Go, chess and shogi by planning with a learned model.
25. Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, Sergey Levine. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor 2018
26. Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, Daan Wierstra. Continuous control with deep reinforcement learning. 2015.