在平均和折扣奖励的多目标（深度）强化学习中学习公平政策。

**Umer Siddique** 1 **Paul Weng** 1.2 **Matthieu Zimmer** 1.

# 摘要

由于自主系统的操作会同时影响多个用户，因此他们的设计必须考虑到公平性问题。与标准(深度)强化学习(RL)相比，我们研究的是学习一个公平对待用户的策略的概率。在本文中，我们提出了这个新的RL问题，在这个问题中，一个目标函数是优化的，它编码了一个公平性的概念，我们正式地去-ﬁne。对于这个问题，我们提供了一个理论讨论，我们研究了折现奖励和平均奖励的情况。在分析过程中，我们特别得出了一个在标准RL环境下的新结果，这个结果具有独立的意义：它指出了一个新的边界，即关于折现奖励的最优政策的op- timal平均奖励的近似误差。由于贴现奖励的学习通常比较容易，这个讨论进一步证明了通过学习贴现奖励的公平政策来确定平均奖励的公平政策。因此，我们描述了七种经典的深度RL算法如何适应我们的公平优化问题，并且我们通过三个不同领域的广泛实验来验证我们的方法。

# 1.1.简介

人工智能（AI）的进步及其在自主系统中的应用为人类社会创造了很多机会，也带来了很多挑战。事实上，一个训练有素的人工智能系统可以比人类更好地自动化或解决一些任务（[Pilarski等](#_bookmark54)，[2011](#_bookmark54)；[Silver等](#_bookmark61)，[2017](#_bookmark61)）。然而。

1上海交通大学UM-SJTU联合研究所2上海交通大学自动化系。 通讯作者：Paul Weng Paul Weng

[<paul.weng@sjtu.edu>。](mailto:paul.weng@sjtu.edu)

*Proceedings of the 37 th International Conference on Machine Learning*, Online, PMLR 119, 2020.Copyright 2020 by the au- thor(s).

目前的人工智能方法不能很好地处理影响众多用户的情况。

在这些人工智能系统中，通常的方法是最大化一个单一的整体效用（例如衡量效率、准确性或任务完成情况）。当一个人工智能系统影响到许多用户时，通常会采用一个*功利性的*目标，将所有用户的个人效用相加（或等价平均）。然而，这种方法并不能保证用户得到公平的对待。事实上，为了获得功利目标的有效的全局解决方案，一些用户的效用可能会被牺牲掉。因此，如果我们想让用户接受和信任自主系统，那么在设计自主系统时的公平性考量是至关重要的。

解决前述问题的一个典型方法是采用 *平等主义的*方法，其目标是使处境较差的用户的效用最大化。然而，这种maxmin方法的直接应用可能不会产生对所有用户严格有效的解决方案，因为只关注一个用户。在本文中，我们采用了一个更精确的公平性定义（[Moulin](#_bookmark44)，[2004](#_bookmark44)），它依赖于三个属性：效率、公正和公平（第[2.3](#_bookmark7)节）。为了对它们进行编码，我们使用广义的基尼评价函数（[Weymark](#_bookmark65)，[1981](#_bookmark65)）作为社会福利函数（即从所有用户效用中定义出整体效用的函数）。

在这项工作中，我们研究了在（深度）强化学习的背景下，考虑贴现奖励和平均奖励，对这个公平福利函数进行优化。

我们的贡献可以归纳如下：

1. 我们介绍这个新颖的问题，我们称之为"RL中的公平操作"（第[3](#_bookmark10)节）。
2. 我们研究了它的理论特性（第[3.1](#_bookmark13)节）。值得注意的是，(1)我们建立了固定马尔可夫策略对确定公平解的充分性，(2)我们讨论了公平最优性可能的状态依赖性，(3)我们提供了一个近似的误差边界，用于折现奖励的多线性最优，而不是平均奖励的最优。有趣的是，最后一个结果应用于单目标RL，导致了一个新颖的、简单的、且可互换的边界，这具有独立的意义。
3. 我们调整了三种深度RL算法，用于解决我们的公平的

优化问题（第[4](#_bookmark21)节）。

1. 我们提供了三个dif- ferent领域的广泛实验结果（第[5](#_bookmark22)节），这验证了我们的命题。

在下一节，我们在介绍我们的贡献之前，先介绍必要的背景。

# 背景资料

在本节中，我们首先回顾马尔科夫决策过程，以及它们在多目标环境下的扩展。然后，我们激励并回顾被称为广义基尼社会福利函数的福利函数，它编码公平性。

**符号。**矩阵用大写字母表示，向量用小写字母表示。两者都用粗体书写。矢量是列矢量，但表示概率分布的矢量除外，它们是行矢量。

初始分布d0，最优政策为：。

argmax ***d0v****π。*(2)

*π*

有趣的是，目标函数可以重写为 fol- lows: ***d0v****π* = ***d****πγ* ***r****π*，其中 ***d****πγ* 是*π*的状态上的*贴现 occu- pation 分布*[1](#_bookmark3)，它被定义为 ***d****π* = *o γ****td0P*** *t*，***P*** 0 = ***I***，***P*** *t* = ***P*** *t\_****1P****π*。

*γ*

*t*=0

*π*

*π*

*π*

*π*

←

值***d****πγ,s*表示总的贴现概率。

在政策*π*下从初始分布d0访问状态*s*。作为问题([2](#_bookmark0))的解的政策称为*γ最优*，表示为*πγ\** 。

利用平均报酬标准，政策*π*的价值函数通常称为*收益*，表示为***g****π*。对于初始状态*s*，它的定义为：

***g****π,s* = lim

1

k**户***π*

┌*Lh*

*rt | s*┐

*.*(3)

## 马尔科夫决策过程

*h→o h*

*t*=1

一个马尔科夫决策过程（MDP）（[Puterman](#_bookmark57), [1994](#_bookmark57)）是由以下元素组成的元组：一组一定的状态，一组一定的动作，每个*a*的过渡矩阵*Pa，*其中*Pa,s/* 表示概率。

*∈ a*

*纱*

在状态*s*中执行动作*a*后达到状态*s/*，每个*a*的奖励向量*ra*，其中*ra,s*是在*s*中执行*a后*获得的奖励，以及概率分布d0。

初始状态。在这个模型中，一个*政策π*确定了一个程序。

规定了如何在状态中选择行动。如果在每一个时间步骤中使用相同的程序，那么一个策略是*固定的*。如果它只根据当前状态选择行动，则是*马尔科夫策略*。在本文中，除非另有说明，否则策略是静止的和马尔科夫的。一个策略可以是*确定性的*（即 *s，π*(*s*) ）或*随机性的*（即， *s，a，π*(*a s*)去注意*s*中选择*a*的概率）。确定型政策是随机型政策的特殊情况。通过扩展。

←

*Ⅴ ∈ aⅤ|*

我们把***P****π*写成***P****π,ss/*= *π*(*a|s*)*Pa,ss/*，***r****π*写成

***r****π,s* =

*a π*(*a | s*)*ra,s*。一个政策诱导了一个*马尔科夫奖赏*

←*a*

给定初始状态d0上的分布，政策*π*所获得的预期平均报酬*μπ*被定义为*μπ*=***d0g****π*，也可以表示为μπ=dπrπ，其中dπ是政策π的固定分布，它被定义为d0P n的Cesa`ro-limit2。也可以表示为*μπ* = ***d****π****r****π*，其中***d****π*是政策*π*的*固定分布*，它被定义为***d0P*** *n*的Cesa`ro-[limit2](#_bookmark4)。分布***d****π*表示政策*π*在每个状态中花费的时间比例。对于

平均报酬标准，通过以下方式获得最优政策：

*π*

argmax *µπ。*(4)

*π*

作为这个问题的解的政策称为*平均最优*，表示为*π*1*\**。

在代理人和环境之间的互动持续时间较长的问题中，平均报酬标准通常是首选。然而，这两种标准是密切相关的（[Baxter& Bartlett，](#_bookmark39)2001[），](#_bookmark39)即对于任何一个问题，都可以采用平均报酬标准。

政策*π*，我们有***d****π****v****π* = *µπ* 。

*过程，*其转折和奖赏都是 resp.***P****π*和***r****π*。

在强化学习（RL）中，*Pa*'s和*ra*'s通常是未知的。在MDP或RL中，目标是找到一个能优化某些性能测量的政策，如*预期的折现总回报*或*预期的平均回报*。

利用贴现回报标准，价值函数为

从初始状态*s*出发的政策*π*的***v****π*的定义为：

1*\_γ*

在本文中，我们假设MDPs是*弱交流的*[3](#_bookmark6)。有趣的是，在这种MDPs中，最优增益***g****π\**是恒定的，即与初始状态无关。

1

当状态或动作空间变得太大或太多时，就需要进行函数逼近来进行泛化。通过参数化函数逼近（如神经网络或线性函数），函数*f*被逼近为

1从技术上讲，这不是一个概率分布，因为它

***v****π,s*

= k**户***π*

*o γt\_****1r***

*t*=1

┌L

*t*

(1)

不归一化。

2 序列的Cesa`ro限值 作为 是由

*u n - 关于*

*1imn→o* 1 *n\_*1 *ui*。 它是一个广义的极限概念，是指

*n*

*i*=0

其中，k **户***π*是相对于***P****π的*期望值，*γ*[0*，*1]是贴现因子，***rt***是代表在时间步骤*t*获得的报酬的随机变量。

*∈*

等于标准极限，如果后者存在的话；

3如果一个MDP的状态可以部分地分为两类：一类是在每个静止策略下所有的状态都是瞬时的，另一类是在某个静止策略下任何两个状态都可以互相到达，那么这个MDP就是*弱沟通的*。

由*f*ˆ(***θ***)表示，其中***θ***表示要学习的参数。在RL中，价值函数或策略都可以被近似。

标准的深度RL方法通常是为不计报酬而设计的。例如，深度Q网络(DQN)

其中***V****π*可以看作是一个*D*矩阵，*Rt*代表在时间步骤*t*获得的随机向量奖励。

在平均报酬的情况下，收益([3](#_bookmark1))变成。

*|s|×*

是Q-Learning的有效扩展（[Mnih等](#_bookmark40)，[2015](#_bookmark40)）。DQN结合了引导、非政策更新和func-。

***G****π,s* = lim

1

k**户***π*

┌*Lh*

*Rt | s*┐

,(6)

tion近似。为了提高学习的稳定性，它

*h→o h*

*t*=1

依靠经验重放（[Lin](#_bookmark36)，[1991](#_bookmark36)）和目标网工程。Q值函数的两个近似值

其中***G****π*可以看作是一个*|s|×D*维矩阵。

*Qπ*(*s，a*) = k**户***π*

┌←*ot*=1

*γt\_1rt | s，a*┐是学过的respec-。

MOMDPs中的多目标优化相当于解。

ing以下问题： argmax*π* ***J*** (*π*) 其中***J*** (*π*) 是

由***θ***和***θ***参数化。 目标网络作为

***l***

为了更新***θ***，回归目标是。 *Q*ˆ**e** (*s, a*) = *r* + *γ maxa/ eA Q*ˆ**e*←*** (*s/, a/*) 其中 (*s, a, s/, r*) 是一个从重放缓冲区中抽取的元组，分别由一个状态、一个...

行动、下一个状态和奖励。

政策也可以用***θ来***近似和参数化，政策梯度（[Sutton等人](#_bookmark63)，[2000](#_bookmark63)）给出了参数更新的方向。

*eJ*(*π***e**)=*ks~***a***π,a~π****9*** (*.|s*)[*Aπ*(*s,a*) elog *π***e**(*a s)*]。

*-*

*V V|*

其中，*Aπ*(*s，a*)=*Qπ*(*s，a*)*Vπ*(*s*)是优势func-。

tion。 由于这个功能是未知的，所以需要对其进行估算。

mated，可以用不同的方式进行。在Ad- vantage Actor-Critic(A2C)中，优势的估计方法是

←*-*

*A*ˆA2C(*st，at*)=*t*=1 *γt\_1rt V*ˆ(*st*)，其中*V*ˆ(*st*)是由批评网络近似的([Mnihetal.](#_bookmark42) , [2016](#_bookmark42))。A2C

行动者更新源自JA2C(*π***e**)=ks*~***a***π,a~π****9***(*.|s*)[*A*ˆA2C(*s,a*)]得到的策略梯度。在近端Pol- icy Optimization（PPO）中，优势*A*ˆPPO的估计值为

与*λ*-收益（[Schulman et al.](#_bookmark60) ，[2017](#_bookmark60)）。它也来自于政策梯度，但有一个额外的约束条件缓解-。

ing政策变化。由JPPO(*π***e**)=*ks* *~***a***π,a~π****9***(*.|s*)[min(*ρeA*ˆPPO(*s,a*)*,ρ*¯*eA*ˆPPO(*s,a*))]可得，其中*ρ*¯ =夹(*ρ ,* 1 *δ,*1+*δ*),*ρ* =*π****9***(*a|s*),*π*为极-。

**-e e** *θ b*

*δ*是一个超参数

*|*

*π* (*a s*)

*b*

产生过渡的冰块和

来控制约束。

## 多目标马尔科夫决策过程

多目标MDP(MOMDP)是一种MDP，其中奖励是向量(而不是标量)，其组成部分，称为*目标*，在多标准环境中被解释为不同的标准(例如，长度、成本、持续时间)，在多代理环境中被解释为单个的效用。从形式上看，MOMDP的报酬函数重新定义如下。*Ra,s* 驶*D*，其中*D*是目标数。因此，价值函数（现在表示为***V*** ，***Q***，***G***）也取值于驶*D*。

*∈*

前面所有关于MDPs的定义都自然延伸到MOMDPs。值得注意的是，在奖励打折的情况下，([1](#_bookmark5))变成：

([2](#_bookmark0))或([4](#_bookmark2))的多目标版本，而向量最大化是关于*帕累托优势*[4的](#_bookmark8)。由于不存在混淆的风险，帕累托优势简单来说就是

用*≥表示*其弱形式，用*>表示*其严格形式。

价值函数（或收益）不受帕累托支配的策略称为*帕累托最优*。MOMDPs中通常的方法是计算所有的帕累托最优解。然而，在某些问题中，这种解的数量可能非常大，这种方法在一般情况下可能是不可行的。事实上，在一些MOMDP实例中，帕累托最优策略的数量是MOMDP大小的指数([Perny et al.](#_bookmark50) , [2013](#_bookmark50))。

在实践中，特别是在自主代理的RL中，人们更希望关注一个解决方案，通常是在所有目标之间找到一个良好的平衡。一个只关注一个解决方案的天真方法是使用加权和来组合目标。然而，这种技术并不能控制目标值的平衡程度。更好的方法是使用一个非线性函数来组合目标。在我们的多用户环境中，找到平衡的解决方案相当于找到公平的解决方案。接下来我们将详细介绍我们的公平性方法。

## 笼统的基尼社会福利函数

在本文中，我们要求最优解满足三个属性，以质量为公平解。

**有效性** 一个公平的解决方案应该是帕累托最优的。

**公正性** 一个公平的解决方案应该满足*"平等对待"*原则，即具有相同特征的用户应该得到类似的待遇。

一个公平的解决方案应该满足*Pigou-Dalton原则*（[Moulin](#_bookmark44)，[2004年](#_bookmark44)）。直观地讲，这个原则指出，给定一个效用向量***v***驶*D*，从一个经济条件较好的用户转移到一个经济条件较差的用户身上，会产生一个新的向量，这个新的向量应该是优先的。形式上，对于任何指数*i*和*j*，如果*vi > vj，*那么对于任何c，使得vi *vj* ***>*** *c > 0，*新向量

*∈*

*-*

***v*** *cei* + *cej*优于***v***，其中***ei***表示第*i个*。

*--*

规范的基础向量[5](#_bookmark9)。

┌*罗* ┐

***V****π,s* = k**户**

*γt\_1Rt | s*

*π*

4 ***Vp****，****p****/ e*驶*D*，***p****弱帕累托支配****p****/ 兮Vi，pi＞pi/****。***

,(5)

此外，*帕累托主导****p****/ 兮Vi，****p。***

*> pi/ 和* aj, *p*

*> pj/ .*

*i*

*j*

*t*=1

5V向量***『****i』*为*i /*上*j*的***『****ij****』***上0，且***『****ii』*上í。

第一个属性是自然的，因为选择一个帕累托主导的解是非理性的。第二个属性在公平性方面是合理的。在我们的工作中，它是通过假设成立的：我们假设所有的目标都是平等的，因此应该以同样的方式对待。第三个是公平优化背景下的关键属性，因为它以一种自然的方式捕捉到了我们更倾向于对用户的效用分配是平衡的解决方案的想法。

为了具体落实这三个原则，我们采用了一种福利函数，称为*广义基尼**社会福利函数（*GGF）（[Weymark](#_bookmark65)，[1981](#_bookmark65)）。GGF的定义如下：

*D*

L

**GGFw**(***v***) = *wivi个。* (7)

*i*=1

其中，***v***驶*D*，***w***驶*D*是一个分量严格递减的ﬁxed正权重向量（即w1*>......>wD），*v个对应的向量的分量为

*∈∈*

的向量***v***按递增顺序排序（即v1*个≤......≤）。*

*vD个* )。此外，在不失一般性的前提下，我们假设

即GGF权重向量***w***经过归一化处理后，和为1（即***w****∈*[0*，*1]*D*，*D wi*=1）。

*i*=1

←

GGF满足所需的三个属性（[Weymark](#_bookmark65)，[1981](#_bookmark65)）。由于GGF权重为正值，GGF在帕累托优势方面是单调的。因此，它满足了效率属性（efﬁ- ciency property）。由于***v***的分量在([7](#_bookmark11))中被重新排序，GGF相对于它的分量是对称的。因此，它满足公正性属性。最后，由于GGF的权重是正的，而且是递减的，所以GGF是Schur-凹的（即它相对于Pigou-Dalton转移是单调的）。因此，它满足公平属性。

此外，GGF是一个片状线性凹函数。事实上，由于GGF的权重是正递减的，所以很容易检查出GGF可以改写成如下形式。

如果w1 1*,* w2 0*，...，wD* 0，GGF对应的是最大限度的平等主义的公平概念（[Rawls](#_bookmark58)，[1971](#_bookmark58)）。

如果w1 1*,* w2 *ε，...，wD ε*，GGF对应的是规则化的maxmin平等主义的公平概念。

* *→ →→*
* *→ →→*

如果w1 *1/D，...，wD 1/D*，GGF对应的是功利法。

* *→→*

如果*wk/wk*+1 + ，GGF对应的是公平性的利民概念（[Rawls](#_bookmark58)，[1971](#_bookmark58)；[Kurokawa等](#_bookmark32)，[2015](#_bookmark32)）。

* *→&*

# RL的公平政策

通过将GGF与MOMDPs整合在一起，我们现在可以正式制定本文所研究的*公平优化*问题，即确定一个能产生公平分配奖励给*D*ﬁxed用户的策略的问题。

argmax **GGFw**(***J***(*π*))*，*(9)

*π*

其中***J***(*π*)可以用折现或平均年龄报酬去定义。由于GGF是一个凹函数，([9](#_bookmark12))定义了一个凸优化问题。这个用折现报酬去定义的问题称为*GGF-γ问题*，而用平均报酬去定义的问题称为*GGF-平均问题。*它们的解分别称为*GGF-γ最优*政策和*GGF-平均最优*政策。

在本文中，我们旨在解决RL环境下的这个问题。由于GGF是一个非线性函数，公平优化是一个非线性凸优化问题。这带来了新的困难，我们接下来将讨论这个问题。

## 理论讨论

在这部分中，我们讨论了与MOMDPs中公平优化有关的三个重要点。(i)哪一个政策子集可以保证包含一个最优解，(ii)公平解可能取决于初始状态，(iii)GGF-*γ*-最优政策的平均向量报酬的GGF与GGF-*γ*-最优政策的GGF有多接近？

**GGFw**

(***v***) = min ***wTv****,* (8)

*σesD σ*

的最优平均向量奖励。我们理论结果的证明可以在附录A中找到。

其中$*D*是*D*度的对称群（即在1*，...，D*上的换元集合），*σ*是换元，***w****σ*=（***w****σ*(1)*，...，wσ(D)）。，****w****σ(D*)).等式([8](#_bookmark14))成立，因为将最大的权重赋给了

*}I*

***v***的最小分量，第二大分量对第二小分量，以此类推。

虽然GGF不是唯一的公平福利函数，但它享有不错的特性。(1) 简单性，因为它是洛伦兹空间的加权和([Chakravarty](#_bookmark45)，[1990年](#_bookmark45)；[Perny等人](#_bookmark50)，[2013年](#_bookmark50))。

(2)其广为人知的特性被[Weymark](#_bookmark65)([1981](#_bookmark65))公理化，(3)其通用性。

GGF可以通过适当设置其权重来涵盖各种特殊情况，如：。

**静态马尔科夫策略的有效性。**与问题([9](#_bookmark12))相关的第一个问题是，在所有(可能的非稳态)政策的集合中，哪种类型的政策是最优的。据我们所知，下面这个尚未被正式说明和证明的定理表明，对于问题（[9](#_bookmark12)），总是存在一个GGF-（*γ*或平均）最优的静止随机马尔可夫策略。

**Lemma 3.1.***对于任何MOMDP，问题*([9](#_bookmark12))的*解都是一个固定的随机马尔科夫策略。*

请注意，这个结果实际上对任何单调函数都是成立的，而不仅仅是问题（[9](#_bookmark12)）中的GGF。这意味着人们可以在较小的静止随机马尔可夫策略集合中寻找最优策略，而不是所有策略的集合。另外．

注意到，与单对象的情况相反，一个确定的

1 *γ*

*µ \* ≥ µ- κ* (***U*** )*rl,*

*\*-*

istic policy may not be optimal（[Busa-Fekete et al.](#_bookmark41)

*πγπ*12

1 *- γ|λ*2*|*

因为通过随机化可以得到更公平的解决方案。

**可能的状态依赖性最优性。**对于GGF-*γ*问题，已知最优性取决于初始状态，或者更一般地取决于初始状态的分布。(见例[3.2](#_bookmark17))。

(士*,* 0)(0*,* 7 )

*γ*



s1

s2

s3

(0*,* 0)( 5 *,* 5 )

*其中κ*2(***U*** )=***lU l2lU*** *\_*1l2*为****U****的条件数，****U*** 2=max**。**:*|*=1 ***Ux****，****x****为任意向量****x****的欧氏法线。*

据我们所知，这是对这种近似误差的唯一已知边界。然而，只有当***P****π\**有*n个*不同的特征值时，它才会成立，而且由于它涉及到相应特征向量矩阵的条件数，因此在实践中可能很难解释和评估。

1

*l l l l l l l*

这促使我们去证明另一种约束。使用基于Laurent数列扩展的ma- trix分解方法。

*γ γ* 和Drazin广义反，[Lamond & Puterman](#_bookmark33)

([1989](#_bookmark33))证明了贴现率之间的以下关系。

*图1.*MDP的例子，其中问题([9](#_bookmark12))的折现奖励的最优性取决于状态。

价值函数和静止政策的收益。

**定理3.5.***对于任何MDP，任何静止政策π，和*

**例3.2.***下面的例子改编自(*[*Ogryczak et al.*](#_bookmark48) *,* [*2013*](#_bookmark48)*)，表明GGF-γ-优选的Pol-*

*icy取决于初始状态。考虑图*[*1*](#_bookmark15)*中描绘的确定性双目标三状态MDP，其中的*

*σ*(***HP****π* ) *σ*(***HP****π* )+1

1

*任何γ∈*(

**户***π*

***v****π* =

***g****π* +

*γ*

1 *- γ*

*,* 1*),*

1 L

*o*

*n*=0

*γ*

╱ *γ -* 1 、*n*

*弧线代表行动，弧线权重对应于向量奖励，γ*[0*，*1)*是一个折现系数。每个状态都有两个动作(Up，Down)，除了状态*s3*，它是一个ab-*

*∈*

*Hn*+***1r****π，*(10)

*吸附状态。假设****w***=（5/9*，*4/9）*。那么，从*s1看*，最优策略π*1*在*s1*中选择Up，在*s2*中选择Up。然而，从*s2*看，π*1*不是最优的，因为在*s2*中行动Down是**首选。*

这一点在应用*γ*-最优政策时提出了一个潜在的困难：根据所访问的状态，人们可能有动力切换到另一个政策。在决策中

或，这个问题被称为*动态不一致的偏好。*

*其中****H*户***π是****I P****π的Drazin倒数，由*(***I P****π* +***P****π\**)*\_*1(***I P****π\**)*给出，****P****π\*是****P*** *n的Cesa`ro极限，对于 n ，σ*(***H*户***π* )*是矩阵****H*户***π的谱半径。*

利用Th.[3.5](#_bookmark16)，并假设*γ*足够接近1，我们可以证明GGF的误差边界（政策是GGF最优的）。

*π*

*→ &*

*--*

*-*

**定理3.6。***对于任何弱沟通的MOMDP：*

**GGFw(*µ****πγ\** ) *≥* **GGFw**(***µ****π*1*\** )

***- R***(1 *- γ*) ╱ρ(*γ，σ*(***H户π1\**** )) + *ρ*(*γ，σ*(***H户πγ\**** )*)、。*

*llππ* 1

([McClennen](#_bookmark37), [1990](#_bookmark37))。此外，这意味着Bellman最优性原则不再成立，因此动态编程不能直接用于计算公平的最优解。然而，我们可以证明在平均报酬的情况下，由于最优政策的收益是恒定的，所以偏好保持状态无关。

**Lemma 3.3.***对于任何弱沟通的MOMDP，GGF均值问题接纳的解是一个固定的随机马尔可夫策略，增益不变。*

**近似误差** RL中的一种常见做法是通过求解相关的折现报酬问题来确定平均报酬的近似最优政策。在单目标的情况下，([Kakade](#_bookmark56)，[2001](#_bookmark56))证明了这样做，一个*γ*-最优政策的收益和最优收益之间的差异可以被约束。

**定理3.4.***假设****P****π\*有n个不同的特征值。让*

1

***U***=(u1*，......，****u****|5|*)*为其右特征向量的矩阵，其对应的特征值为λ*1=1*>|λ2|≥......≥。*

*|λ|5||.那么。*

*其中****R***=max ***R****，ρ*(*γ，σ*)=*σ 。*

*γ\_*(1*\_γ*)*σ*

有趣的是，定理[3.6](#_bookmark18)应用于单目标的情况下（*D*=1），得到了另一个近似误差边界，它比定理[3.4](#_bookmark20)更普遍。

**推论3.7.***对于任何弱共通的MDP：μπγ\*≥μπ*1*\* -****r***(*1-γ*) *ρ*(*γ，σ*(***H*户***π*1*\** ))+*ρ*(*γ，σ*(***H*户***πγ\** )*，其中****r***=max*π* ***lr****πl。*

╱、

定理[3.6](#_bookmark18)和推论[3.7](#_bookmark19)中的边界清楚地表明，当*γ* 1时，近似误差 如预期的那样趋于零。与定理[3.4](#_bookmark20)中的边界一样，误差取决于实例具体常数。在这里，它主要取决于***I P****π*的Drazin逆的谱半径，其中*π*是GGF-*γ*-最优或GGF-平均最优策略。这个频谱半径可以直观地解释为政策在瞬态状态下可以花费多长时间的度量（因此，也是政策收敛到其平均报酬所需时间的度量）。 事实上，一个较大的

*→*

*-*

*σ*(***H*户***π* )意味着更大的*ρ*(*γ，σ*(***H*户***π* ))和更大的边界。

# 算法

在这一节中，我们将解释如何修改DQN和政策梯度算法，以便用折现奖励来解决问题（[9](#_bookmark12)）。如前文所讨论的，它可以提供该问题的近似解，且*γ*的平均报酬足够接近于1。

**DQN。**为了优化GGF，我们对深度Q网络（DQN）进行修改，使其取值为驶*|A|×D*而不是驶*|A|*。DQN被训练成预测多目标***Q***函数。请注意，直接预测GGF值会妨碍引导。因此，DQN的回归目标变成了。

***Q***ˆ*θ*（*s，a*）=***r***+*γ****Q***ˆ*θ/*（*s/，a\**）*。*

/、

*\**ˆ*/ /*

其中*a*

= *argmaxa/eA* **GGFw**

***r*** + *γ****Q****θ/* (*s , a* )。此

改编版的DQN称为*GGF-DQN*。

请注意，理想情况下，*a\**应选择为

↓!

/ˆ*/ /* 、

*argmaxa/eA* **GGFw**

*ks/*

***r*** + *γ****Q****θ/* (*s , a* )

.如何...

控制（TL），（三）数据中心控制（DC）。第一域(SC)对应于生态学中遇到的保护问题，其目标是维持几个相互影响的濒危物种的种群。我们对[Chade`s等人（2012](#_bookmark43)）提出的双物种模型进行了改编，专门考虑了两个物种的公平性，即一个濒危物种（海獭）和其猎物（北鲍）。一个状态编码了物种的种群数量。过渡函数基于两个物种的种群增长模型，考虑到偷猎（对鲍鱼）或石油泄漏（对海獭）等因素。为了保持两个种群的平衡，考虑了五种行动：什么都不做、引进海獭、实施反偷猎、控制海獭，以及一刀切。

一半反偷猎，一半控制海獭。矢量奖励对应于按比例排列的物种密度(单位：*m\_*2)。

第二个域（TL）对应于经典的交通灯控制问题，其中一个代理控制一个路口的交通灯，以优化交通流量。

通常解决这个问题的方法是最小化

曾经，即使期望值很高，也很难计算。

由样本平均值估计。因此，我们对DQN的修改实际上优化的是GGF的期望值，而不是期望值的GGF。根据Jensen不等式，这意味着我们实际上是在优化一个正确目标函数的下限。

**政策梯度法。**解决问题([9](#_bookmark12))的一个自然的替代方法是使用政策梯度法。与DQN的适应性相反，它直接对所需的目标函数进行定时操作。另一个优点是，它可以学习一个随机策略，它可能严格地支配一个确定性的GGF。

GGF的政策梯度制定如下。

**VeGGFw**(***J***(*π***e)**) =**VJ**(*π****9*** )**GGFw**(***J***(π**e**)) *．* ***VeJ***(*π***e**)

=***wTVeJ***（*π***e**）*。*

*σ*

其中，***eJ***（*π***e**）是一个*D N*矩阵，代表*D*目标上的经典政策梯度，***w****σ*按照***J***（*π***e**）排序，*N*是政策参数的数量。

*V×*

在实验中，我们将其应用于PPO和A2C：***eJ***（*π***e**）分别被 eJPPO（*π***e**）和 *θ*JA2C（*π***e**）代替。为了对***w进行***排序，将初始状态存储到经验的

*VV*

*V*

估计***J***（*π***e**）=*ks ~***a**[***V***ˆ（s0）]，其中***V***ˆ : *s →* 驶*D*为

所有车道的预期等待时间之和。相反，我们建议考虑到每条道路的公平性（在我们的实验中，*D* = 4）。换句话说，我们的目标是学习一个控制器，优化每条道路的预期等待时间。更具体地说，我们考虑了一个八车道。

交叉口，其中四个方向有2条车道。一个状态是由每个车道上停在十字路口的总等待时间和汽车密度（以[0*，*1]为单位）组成的，它还包括当前交通灯的阶段（即哪个车道和方向有绿灯/红灯）。它还包含了当前交通灯的相位（即哪些车道和方向有绿灯/红灯）。一个动作对应一个相位的变化。Trafﬁc随机生成的五次分布。在每一个时间步骤，车辆按照二项式分布，以给定的概率发射出ran- domly。二项式分布的fow近似于小概率的泊松分布，一些随机事件以一定的速度独立发生。奖励是一个向量，其成分是每个车道的负等待时间之和。

第三个领域（DC）是一个数据中心控制问题，其中一个集中式控制器管理由一定数量的主机共享的计算机网络，以优化每个主机的带宽（[Ruffyetal.](#_bookmark59)在这里，公平性是相对于主机而言的（在我们的实验中，*D* = 16）。例如，这种问题通常会发生在软件定义的网络（SDN）中。一个状态en-

0 0编码网络统计（即队列长度、导数过量、队列长度过量、队列长度过量、队列长度过量、队列长度过量、队列长度过量）。

与批判的近似。所得算法分别称为*GGF-PPO*和*GGF-A2C*。

# 实验结果

为了测试我们的三种算法，我们在三个不同的领域进行了实验（详细描述，见附录B）。(i) 物种保护 (SC), (ii) Trafﬁc light.

队列长度的时间、丢包次数、队列长度超过某个限制）和当前对主机的带宽分配。一个动作是一个带宽分配的向量。主机之间的流量是随机产生的。奖励是一个向量，它的成分是每个主机的带宽被队列长度之和惩罚（为了避免缓冲区膨胀）。

这三个领域大致按复杂程度的增加排序（目标数量也增加）。前两个域有离散的状态-动作空间，而第三个域有连续的状态-动作空间。由于它们都是偶发问题，所以它们都是交流(MO)MDPs。

在这些领域，我们通常运行DQN，PPO，A2C，和他们的适应GGF（即，GGF-DQN，GGF-PPO，GGF- A2C）。算法的超参数被opti- mized（附录C）和所有的实验进行了

1·4

1·2

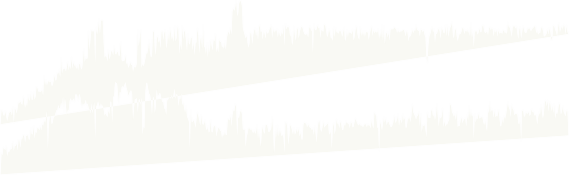
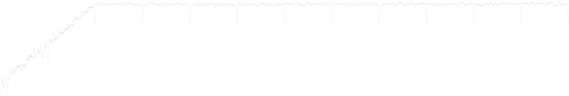
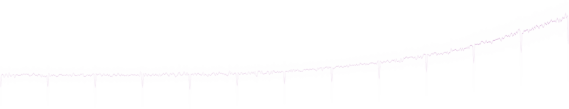
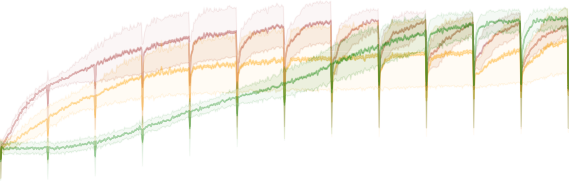
平均累积密度

1·0

0·8

0·6

0·4



PPO

A2C

DQN

随机

GGF-PPO GGF-A2C GGF-DQN

5000060000

步数

使用轻量级超参数优化器（LHPO），这是一个用于在集群上运行并行实验的开源库（[Zimmer](#_bookmark66)，[2018](#_bookmark66)）。已经使用了两台具有双CPU插座的计算机（英特尔Xeon CPU E5-2678 v3）。该

*图2.*DQN、A2C、PPO及其GGF版本在学习阶段与随机策略在SC域的平均累积密度。

未归一化的GGF系数被定义为*Wi。* 1

=

*2i*

从0到*D* 1。所有的实验结果(例如，小区)都是用SC和TL(resp.DC，因为它是一个复杂的环境)的不同种子的50(resp.20)运行的平均值。

*-*

现在介绍我们实验的主要结果（详见附录D）。它们是为了回答以下问题而设计的。**(A)** 优化GGF而不是目标的平均数有什么影响?**(B)**适应GGF的算法之间的比较如何？

0·9

0·8

GGF评分

0·7

0·6

DQN GGF-DQN A2C GGF-A2C PPO GGF-PPO

并与其标准版？**(C)**公平的确定性政策和随机性政策如何比较？**(D)** *γ*对GGF平均最优性的影响是什么？**(E)**这些算法在连续域中的表现如何？

**问题(A)为了**回答(A)，我们讨论了DQN、A2C、PPO及其GGF coun- terparts在SC域中的外周结果。我们可以首先比较它们在学习阶段（见图[2](#_bookmark23)）的两个积累密度的平均值。作为一个理智的检查，ﬁgure还包括均匀随机策略。由于密度在一个事件中积累，所有算法的曲线都会发生下降。正如预期的那样，随机策略表现最差，而标准RL算法是最好的，因为平均密度大致是这些算法中的优化值。在目标的平均数方面，一个为GGF而调整的算法通常会比它的原始版本表现更差，因为它在效率和公平之间进行了权衡。

更有趣的是，为了我们的目的，我们可以在GGF得分方面比较算法。训练后，将得到的策略在环境中应用50次。这个分数是生成轨迹的样本平均向量奖励的GGF。图[3](#_bookmark24)显示了DQN、A2C、PPO所学习的策略的该分数分布情况。

*图3.DQN、A2C、PPO及其GGF算法的GGF得分*DQN、A2C、PPO及其GGF算法在SC域测试阶段的GGF评分。

我们还提供了训练后估计的非聚集累积密度的图（图[4](#_bookmark25)），这对于SC域来说很容易做到，因为它是双目标的。我们可以再次观察到，标准RL算法获得的总累积密度比GGF对应的算法更高。然而，标准方法的两个物种的个体密度比我们的方法得到的密度更不相等，我们的方法提供了更公平的解决方案。

由于我们不能轻易地在所有领域中显示非聚集的对象，我们引入了额外的统计数据来评估公平性。值得注意的是，*变异系数*(CV)，它可以被理解为不平等性的简单测量。在图[5](#_bookmark27)中，每个优化GGF的算法都有一个较低的CV和较高的最小密度。

1·0

海员

阿巴隆斯

0·8

平均密度

0·6

0·4

0·2

及其GGF算法。训练和测试期间的步数为3650步（对应10年）。正如预期的那样，三种GGF算法都有较高的GGF

0·0

DQN GGF-DQN A2C GGF-A2C PPO GGF-PPO

比其原始算法的得分。

由于GGF的得分并不能直接清楚地反映出。

*图4.DQN、A2C、PPO及其GGF版本在SC域测试阶段的个体密度。*DQN、A2C、PPO及其GGF版本在SC域测试阶段的个体密度。

1*●*50

DQN

GGF-DQN

A2C

GGF-A2C

PPO

GGF-PPO

随机

1*●*25

1*●*00

0*●*75

0*●*50

0*●*25

0*●*00

CV 最小密度最大密度

GGF评分

*一*1-6 *一*1-8 *一*2-0 *一*2-2 *一*2-4

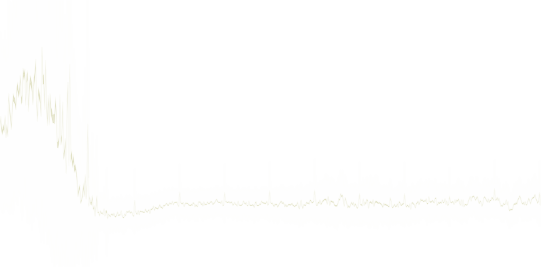
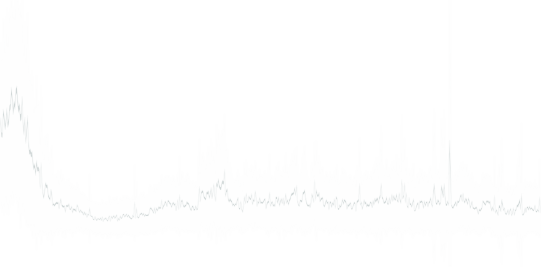
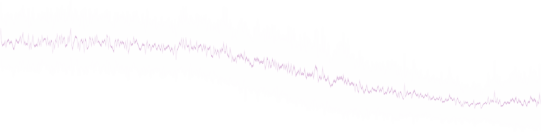
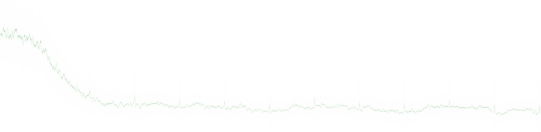
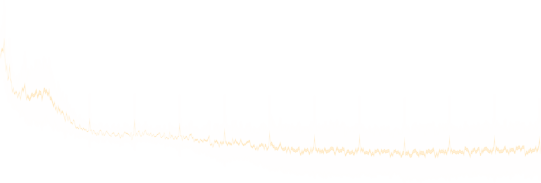
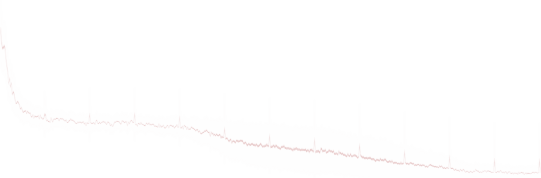
*×*107

DQN GGF-DQN A2C GGF-A2C PPO-0.99 GGF-PPO-0.99 PPO*-1-*GGF-PPO*-1-。*

*图5.*在SC域的测试阶段，DQN，A2C，PPO和其GGF对应的CV，最小和最大密度。

*图7.*在TL域的测试阶段，DQN、A2C、PPO及其GGF版本的GGF评分，当*γ*接近1时，PPO和GGF-PPO的GGF评分。

16000



PPO

A2C

DQN

随机

GGF-PPO GGF-A2C GGF-DQNFixed

平均累计等候时间

1·5

140001·0

12000

0·5

10000

8000

0·0

履历

最短等待时间

DQN

GGF-DQN

A2C

GGF-A2C

PPO

GGF-PPO

固定

随机

最大等待时间

6000

5000060000

步数

*图8.*在TL域的测试阶段，DQN、A2C、PPO及其GGF对应的CV、最小和最大等待时间。最小和最大等待时间有

*图6.*在学习阶段，DQN、A2C、PPO及其GGF对应的平均等待时间，以及在TL域的ﬁxed和随机策略的平均等待时间。

**问题(B)** 为了回答(B)，我们转向TL域，一个更复杂的环境。图 [6](#_bookmark29) 显示了我们考虑的六种不同 RL 算法在学习阶段获得的所有车道的平均等待时间。作为参考，我们添加了随机策略和ﬁxed策略的性能。后者的策略在所有相位之间以一个ﬁxed频率循环，这已经在许多模拟中得到优化。ﬁxed策略自然比随机策略表现更好。当训练结束时，它比所有RL算法都要差。正如预期的那样，所有的标准算法都比它们的GGF对应的算法表现得更好，因为av- erage累积的等待时间对应于原始算法优化的措施。

图[7](#_bookmark26)描述了训练后计算的GGF得分（超过减去等待时间）。它还包括PPO和GGF-PPO两种不同*γ*值（即0.99和接近1）的结果，我们在（D）的答案中讨论。虽然，由于空间的原因，没有包括ﬁxed策略，因为它的GGF-性能较差，但我们注意到GGF- DQN的性能比ﬁxed策略更好，这说明GGF-DQN确实对公平性进行了优化。三种GGF算法的GGF得分都比原来的同类算法要好。GGF-PPO实现了最好的得分。

除以3000，才能用CV显示。

为了证实那些高的GGF分数对应着更公平的解决方案，我们确实可以在图[8](#_bookmark28)中观察到，我们的pro- posed算法总是能实现比其原始算法更低的CV。在所有算法中，PPO per-形成最好的，因为它的等待时间最低。与PPO类似，GGF-PPO在所有道路上的等待时间也较低，但每条道路的等待时间分布比较均衡。

**问题(C)**优化随机策略的算法在目标的平均(或总)方面往往比DQN表现得更好(见图[2](#_bookmark23)、图[6](#_bookmark29)或附录中的图23)。然而，就GGF而言，对于像SC这样的简单做主，GGF-DQN实际上表现很好，而对于像TL这样更复杂的领域，结论则相反。这可能是由于do- main的部分可观察性和GGF的使用，这就需要随机策略来获得更公平的解决方案（正如我们的理论讨论所建议的）。图[6](#_bookmark29)表明，公平性的代价（即以优化GGF而非功利标准的目标平均值为代价的损失）是有限的。

**问题(D)**我们还运行了*γ*非常接近1的算法，即*γ*=0*.*99999。图[7](#_bookmark26)的最后两个boxpl图显示，GGF-PPO的结果与*γ*=0*.*99时的结果非常相似。这表明，发现政策

2*●*0

1*●*8

GGF评分

0*●*10

固定

随机

PPO

GGF-PPO

A2C

GGF-A2C

1*●*6

1*●*4

0*●*05

A2C GGF-A2C PPO GGF-PPOFixed

0*●*00

CV 最小bandwidth 最大带宽

*图9.*在DC域的测试阶段，A2C、PPO及其GGF版本的GGF得分，与已配置政策的GGF得分。

由GGF-PPO接近GGF-平均最优。同时，这也表明，在实践中，除了困难的MDP结构，使用*γ*=0*.*99是足够的。

**问题(E)** 虽然我们的理论讨论是针对有限MOMDPs的，但我们猜测在连续空间中通过增加一些通常的技术条件也可以得到类似的结果([Arapostathis等人](#_bookmark38)，[1993](#_bookmark38)).因此，我们也在DC域上尝试了我们的方法(用A2C和PPO)，其状态和动作是连续的。此前，我们还在DC域上尝试了我们的方法(用A2C和PPO)，其状态和动作是连续的。

图[9](#_bookmark30)显示了A2C、PPO、GGF-A2C和GGF-PPO在GGF评分方面的表现。作为参考，它还包括了ﬁxed策略的GGF得分。正如所料，ﬁxed保单的GGF得分最低。GGF-A2C和GGF-PPO的GGF得分比原来的同类政策要好。这清楚地表明，如果一个公平的政策很重要，那么基于加权总和来汇总目标（权重相等）的通常方法是不充分的。

图[10](#_bookmark31)说明了不同的RL al- gorithms及其GGF算法与ﬁxed和随机策略在CV、最小和最大带宽方面的性能。如同预期的那样，随机策略的表现较差，因为它具有最低的最小和最大带宽，也获得最低的累积带宽。ﬁxed策略比随机策略表现更好。GGF版本的A2C和PPO具有较低的CV，这表明它们比原来的算法更公平。

更多的实验细节和结果，见附录。

# 相关工作

应用数学、运筹学再搜索和理论计算机科学中的公平优化是一个活跃的再搜索方向([Ogryczak et al.](#_bookmark51) , [2014](#_bookmark51))。众多的类连续和组合优化问题（[Hurkala& Sliwinski，](#_bookmark52)2012[；](#_bookmark52)[Ogryczak等人，](#_bookmark48)2013[；](#_bookmark48)[Nguyen & Weng，2017；Lesca](#_bookmark46)等[人，2019）已经](#_bookmark34)被扩展到考虑 然而，这些作品作为-假设整个模型是已知的，因此只有

*图10.A2C、PPO和GGF对应的测试阶段的CV、最小和最大带宽。*在直流域的测试阶段，A2C、PPO及其对应的GGF的CV、最小和最大带宽。

着重于公平优化问题，因为不需要学习。其中一项值得注意的工作（[Ogryczak等人](#_bookmark48)，[2013](#_bookmark48)）使用线性编程方法解决了折现奖励的问题（[9](#_bookmark12)），这只有在过渡和奖励函数已知的情况下才有可能。相比之下，在我们的工作中，我们在RL环境下解决问题（[9](#_bookmark12)），既考虑了贴现奖励和平均奖励，也解决了大的甚至是连续状态空间的问题。

公平性考虑最近成为机器学习中的一个重要主题（[Busa-Fekete等人](#_bookmark41)，[2017](#_bookmark41)；[Spe-](#_bookmark62) [icher等人](#_bookmark62)，[2018](#_bookmark62)；[Agarwal等人](#_bookmark35)，[2018](#_bookmark35)；[Heidari等人](#_bookmark49)，[2018](#_bookmark49)；[Jiang & Lu，](#_bookmark55)2019[；](#_bookmark55)Weng[，](#_bookmark64)2019[）。](#_bookmark64)大多数工作集中在公平性的公正性方面。然而，少数显著的例外考虑了顺序决策问题中的公平优化。[Busa-Fekete等人](#_bookmark41)（[2017](#_bookmark41)）研究了一个与我们类似的问题，但在多臂匪徒集-ting中。[Jiang & Lu(](#_bookmark55)2019[)](#_bookmark55)考虑了多代理RL中学习公平策略的问题，其中公平性是在代理上deﬁned的，并且用不同的福利函数编码。此外，他们的重点是使用共识机制以分布式的方式学习分散的政策。最后，另一项最近的工作值得在此提及，尽管它并没有具体处理公平性。[Cheung](#_bookmark47)([2019](#_bookmark47))在UCRL环境下研究了一个比我们更普遍的问题([Jakschetal.](#_bookmark53) , [2010](#_bookmark53))，其重点是后悔最小化，以有效地解决探索-开发困境。那项工作还处理了表格式MDP中的非稳态威慑性策略，与我们的工作相反。

# 结论

在这项工作中，我们引入了RL中的新型公平运算问题，并进行了理论上的讨论。我们提出了三种深度RL算法的适应性来解决大规模问题，并提供了大量的经验验证。作为未来的工作，我们可能会考虑其他公平的wel- fare函数([Ogryczaketal.](#_bookmark51) , [2014](#_bookmark51))，扩展到分布式控制，或者直接适配平均报酬的RL算法。

# 鸣谢

这项工作得到了国家自然科学基金项目（编号：61872238）、上海市自然科学基金项目（编号：19ZR1426700）和雅虎FREP资助的部分支持。

# 参考文献

Agarwal, A., Beygelzimer, A., Dud´ık, M., Langford, J., and Wallach, H. A reductions approach to fair classiﬁcation.In *ICML*, 2018.

Arapostathis，A.，Borkar，V.，Fernandez-Gaucherand，E.，Ghosh，M.K.，和Marcus，S.I. Discrete-time controlled Markov processes with average cost criterion：a survey.*SIAM Journal on Control and Optimization*，31(2)：282-344，1993.ISSN 03630129.doi：10.1137/0331018.

Baxter, J. and Bartlett, P. L. Inﬁnite-horizon policy-gradient estimation.*JAIR*，15：319-350，2001。

Busa-Fekete, R., Szo¨ renyi, B., Weng, P., and Mannor, S. Multi-objective bandits:Optimizing the generalized Gini index. - In ICML, 2017.In *ICML*, 2017.

Chade`s，I.、Curtis，J.M.和Martin，T.G.为两个相互影响的濒危物种----海獭和北鲍----设定现实的恢复目标。《*养护生物学*》，26(6)：1016-1025，2012年。

Chakravarty, S. R. *Ethical Social Index Numbers*.Springer Verlag，1990年。

Cheung, W. C. Regret Minimization for Reinforcement Learning with Vectorial Feedback and Complex Objec- tives.In *NeurIPS*, 2019.

Heidari, H., Ferrari, C., Gummadi, K. P., and Krause, A. Fairness behind a veil of ignorance:A welfare analysis for automated decision making.In *NeurIPS*, 2018.

Hurkala, J. and Sliwinski, T. Fair ﬂow optimization with ad- vanced aggregation operators in wireless mesh networks.In *Federated Conference on Computer Science and Infor-* *mation Systems*, pp.

Jaksch, T., Ortner, R. and Auer, P. Near-optimal regret bounds for reinforcement learning.*Journal of Machine Learning Research*, 11:1563-1600, 2010.

Jiang, J. and Lu, Z. Learning Fairness in Multi-Agent Sys- tems.In *NeurIPS*, 2019.

Kakade, S. Optimizing average reward using discounted rewards.In *COLT*, 2001.

Kurokawa, D., Procaccia, A. D., and Shah, N. Leximin Allocations in the Real World.In *Proceedings of the Sixteenth ACM Conference on Economics and Compu- tation*，EC '15，pp.345-362，Portland，Oregon，USA，June 2015.Association for Computing Machinery.ISBN 978-1-4503-3410-5.doi：10.1145/2764468.2764490.

Lamond, B. F. and Puterman, M. L. Generalized Inverses in Discrete Time Markov Decision Processes.*SIAM Journal on Matrix Analysis and Applications*, 10(1):118-134, jan 1989.ISSN 0895-4798.doi:10.1137/0610009.SIAM Journal on Matrix Analysis and Applications, 10(1):118-134, jan 1989.

Lesca, J., Minoux, M., and Perny, P. The fair owa one-to-one assignment problem: Np-hardness and polynomial time special cases.*Algorithmica*，81(1)：98-123，2019.

Lin, L. Programming robots using reinforcement learning and teaching.In *AAAI*，pp.781-786，1991。

McClennen，E. *Rationality and dynamic choice:Founda- tional explorations*.Cambridge university press，1990.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidje- land, A.K., Ostrovski, G., Petersen, S。Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning.*Nature*, 518:529-533, 2015.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap,

T.P., Harley, T., Silver, D., and Kavukcuoglu, K. Asyn- chronous methods for deep reinforcement learning.In *ICML*, 2016.

Moulin，H. *Fair Division and Collective Welfare*.MIT Press，2004.

Nguyen, V. H. and Weng, P. An efﬁcient primal-dual algo- rithm for fair combinatorial optimization problems.In *COCOA*, 2017.

Ogryczak, W., Perny, P., and Weng, P. A compromise pro- gramming approach to multiobjective Markov decision processes.*International Journal of Information Technol- ogy & Decision Making，12*：1021-1053，2013.

Ogryczak, W., Luss, H., Pio´ro, M., Nace, D., and Tomaszewski, A. Fair optimization and networks:A survey.*Journal of Applied Mathematics*，2014，2014.

Perny, P., Weng, P., Goldsmith, J., and Hanna, J. Approx- imation of Lorenz-optimal solutions in multi-objective Markov decision processes.In *International Conference on Uncertainty in Artiﬁcial Intelligence（UAI）*，2013.

Pilarski, P. M., Dawson, M. R., Degris, T., Fahimi, F., Carey,

J.P., and Sutton, R. S. Online human training of a my- oelectric prosthesis controller via actor-critic reinforce- ment learning.In *IEEE International Conference on Rehabilitation Robotics*, 2011.

Puterman, M. *Markov decision processes: discrete stochas- tic dynamic programming*.Wiley，1994.

Rawls, J. *The Theory of Justice*.Havard university press, 1971.

Ruffy, F., Przystupa, M., and Beschastnikh, I. Iroko:A framework to prototype reinforcement learning for data center trafﬁc control. - In Workshop on ML for Systems at NeurIPS, 2019.In *Workshop on ML for Systems at NeurIPS*, 2019.URL [http://arxiv.org/abs/ 1812.09975](http://arxiv.org/abs/1812.09975)。

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms.*CoRR*，abs/1707.06347，2017。URL [http://arxiv. org/abs/1707.06347](http://arxiv.org/abs/1707.06347).

Silver，D.，Schrittwieser，J.，Simonyan，K.，Antonoglou，I.，Huang，A.，Guez，A.，Hubert，T.，Baker，L.，Lai，M.。

Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. Mastering game of go without human knowledge.*Nature*, 550:354-359, 2017.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. A uniﬁed approach to quantifying algorithmic unfairness:Mea- suring individual & group unfairness via inequality in- dices. - In KDD, pp.In *KDD*，pp.2239-2248，2018.ISBN 978-1-4503- 5552-0.doi：10.1145/3219819.3220046.URL [http:](http://doi.acm.org/10.1145/3219819.3220046)

[//doi.acm.org/10.1145/3219819.3220046](http://doi.acm.org/10.1145/3219819.3220046).

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation.在*NIPS*，2000年。

Weng, P. Fairness in reinforcement learning.In *AI for Social Good Workshop at IJCAI*, 2019.

Weymark, J. Generalized Gini inequality indices.*Mathe- matical Social Sciences*, 1:409-430, 1981.

Zimmer, . *Apprentissage par renforcement de´veloppemental*.PhD thesis, University of Lor- raine, January 2018.