Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Smart Process Manufacturing—Article

碳配额市场下以乙醇胺溶液进行碳捕集的电厂的优化竞标和运行：基于强化学习的 Sarsa 时间差分算法的解决方案

李子昂 [a]，丁正桃 [a,*]，王美宏 [b]

[a] School of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, UK
[b] Department of Chemical and Biological Engineering, The University of Sheffield, Sheffield S1 3JD, UK

ARTICLE INFO

Article history:
Received 17 January 2017
Revised 2 March 2017
Accepted 10 March 2017
Available online 24 March 2017

关键词
电厂
燃烧后碳捕集
化学吸收
碳配额市场
决策优化
强化学习

摘要

对于处在碳配额市场条件下以乙醇胺 (MEA) 进行碳捕集的燃煤电厂，本文应用了基于强化学习的 Sarsa 时间差分算法为其自行搜寻一种统一的竞标和运行策略。电厂的决策者的目的被定义为最大化电厂寿命下的贴现累计利润。其中，我们引入以下两个限制条件：一是碳捕集的高能耗和电力生产之间的权衡；二是碳排放交易市场中竞得的碳配额数量与电力生产导致的实际碳排放量的近似相等。本文给出了三个案例方便研究。第一个案例中，我们展示了 Sarsa 算法将收敛到一个确定且优化的竞标和运行策略。第二个案例中，相互独立设计的运行和竞标策略与统一设计的运行和竞标策略相互比较，以表明加入了随时间变化、市场导向的碳捕集水平后，Sarsa 算法将有助于电厂决策者获得更高的贴现累计利润。第三个案例则引入了处在同一碳配额市场的另一电厂作为原电厂的竞争对手。两家电厂设置了相同的发电和二氧化碳捕集设备，但新电厂采用不同的策略获得利润。比较两家电厂的贴现累计利润，结果表明：采用 Sarsa 学习算法、找到统一的竞标和运行策略的原电厂会更具竞争力。

1. 引言

二氧化碳(CO_2)是电厂排放的主要温室气体。基于胺溶液的燃烧后碳捕集由于可以通过对传统火电厂的简单改造而实现，是一种前景广阔的大范围碳捕集技术[1]。而类属伯胺的乙醇胺(MEA)则是采用溶液类碳捕集法最可行的策略，因为相比仲胺和叔胺，它与二氧化碳的反应速度更快[2]。之前的研究聚焦在特定碳捕集水平下，碳捕集过程的优化运行[1,3–7]。然而，乙醇胺溶液的再生能耗高、价格贵，而且碳配额市场下每季度拍卖的结算价格又会变化。固定碳捕集过程中碳捕集水平的运行方式并不经济。文献[8,9]已经对不同二氧化碳排放价格条件下需对碳捕集水平进行优化有所阐释。不过这些二氧化碳排放的定价机制类似碳税[10]。对于灵活的市场导向的碳配额贸易机制[11–13]，决策者需确定向碳排放市场提出竞标的碳配额数量。因此，我们有必要为配有碳捕集设备的电厂设计一种统一的竞标和运行策略来最大化电厂整个寿命周期中的利润。我们运用了Sarsa时间差分算法来为装配碳捕集设备的特定电厂的决策者寻找竞标和运行策略。竞标与运行的关系是通过随时间变化的碳配额市场下决策者的持有账户建立的[14]。对于寻找到的策略的优劣，可由获得的贴现累计利润来评

* Corresponding author.
E-mail address: zhengtao.ding@manchester.ac.uk

英文原文: Engineering 2017, 3(2): 257–265
引用本文: Ziang Li , Zhengtao Ding, Meihong Wang. Optimal Bidding and Operation of a Power Plant with Solvent-Based Carbon Capture under a CO2 Allowance Market: A Solution with a Reinforcement Learning-Based Sarsa Temporal-Difference Algorithm. *Engineering*,http://dx.doi.org/10.1016/J.ENG.2017.02.014

估，即电厂的贴现现金流[9]。本文将按如下顺序介绍。在第2节，在考虑到各季度碳配额拍卖市场的前提下，我们会建立一个应用了碳捕集技术的电厂的利润模型，并依此讨论、提出基于溶液进行碳捕集的燃煤电厂的竞标和运行问题。在第3节，Sarsa时间差分算法会被引入并用于寻找前述系统的最优解。在第4节，通过案例，我们会解释Sarsa时间差分算法可找到一个统一的竞标和运行策略，且该策略可以最大化特定电厂的利润。本文末尾会给出相应结论。

2. 问题形成

在本节，我们将建立一个配有碳捕集过程的燃煤电厂利润模型以及一个简化后的温室气体排放贸易系统。此后，在碳排放贸易系统下，我们会给出电厂寿命周期中基于贴现累计利润的目标函数。

2.1. 基于乙醇胺的碳捕集模型的建立

基于乙醇胺的碳捕集过程模型可由Aspen Plus®构建[15]。模型的物理特性由电解质非随机双液法(electrolyte non-random two liquid，eNRTL)计算。模型本身由来自碳捕集试验场的实验数据验证[16]。验证后的模型将按比例放大，匹配额定容量650 MW的亚临界燃煤电厂排放的废气。图1是基于乙醇胺的碳捕集过程的流程图[6,17]。图中共有两座吸收塔(Absorber)用于废气中二氧化碳的吸收[18]。表1列出了吸收塔和汽提塔(Stripper)的参数。贫碳MEA溶液由分流器Splitter 1分为相等的两个部分，随后分别送入两座吸收塔的顶部。同时，来自电厂的废气由分流器Splitter 2等分并注入吸收塔底部。在吸收塔中，废气中的二氧化碳将自发地与乙醇胺溶液反应。吸收了大部分二氧化碳的气相会排入空气而富含碳的液相则被抽送至换热器，之后输送至汽提塔内。在汽提塔中，二氧化碳从富碳MEA溶液中分解出来，同时重新生成贫碳MEA溶液，离开汽提塔底部。为了使重新生成的贫碳MEA溶液的温度满足两座吸收塔进液口特定的温度目标，新生成的溶液需由换热器及下游的冷却器冷却。此外，在循环前，流失的MEA和水分由混合器Mixer 3补充。最后，贫碳MEA溶液会被送回至吸收塔，用于持续的二氧化碳吸收。通过汽提塔的冷凝器(Condenser)，分解出的高浓度二氧化碳产品可被压缩和运输。

图1中，基于乙醇胺的燃烧后碳捕集过程受控于四个控制回路。类似的控制方案在文献[3,17]中已有讨论。相应地，在Aspen Plus®建立的稳态模型中，我们提出以下设定：①通过改变冷凝器热负荷，汽提塔最上层的温度设定在35℃；②通过改变冷却器热负荷，吸收塔顶层

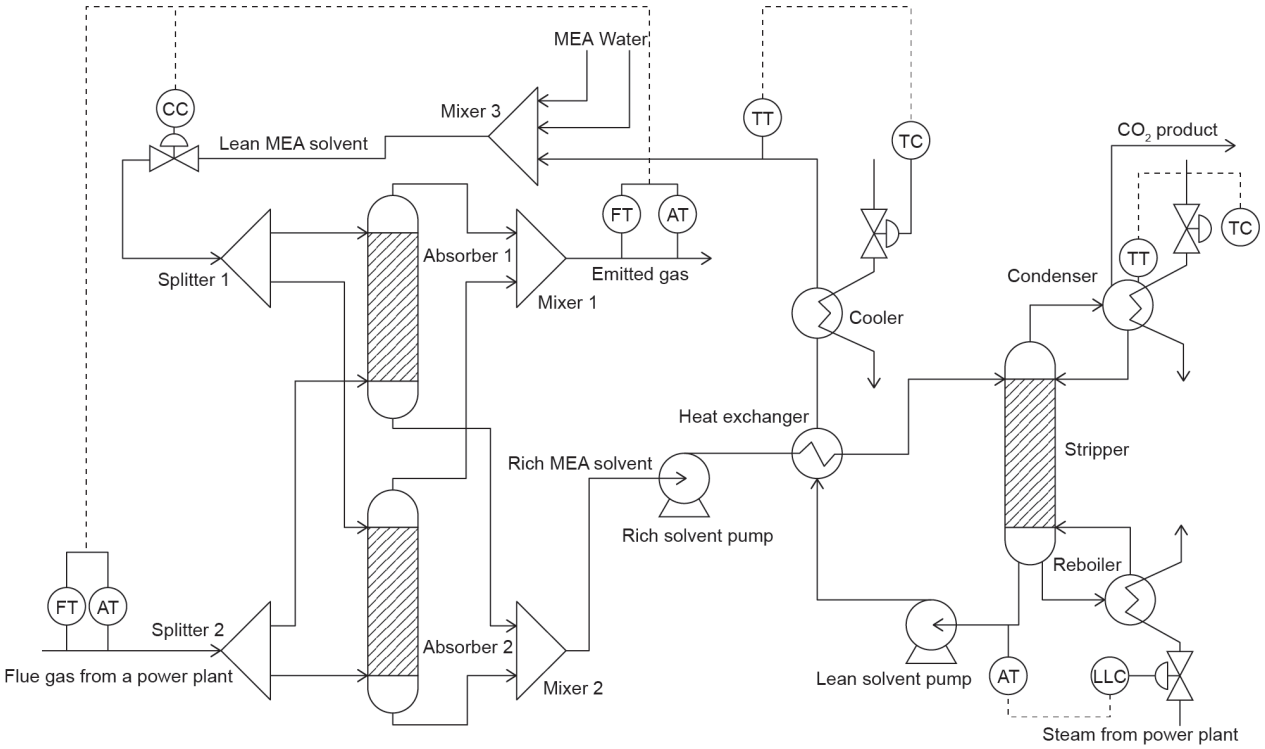

图1. 燃烧后碳捕集过程[6,17]。

贫碳MEA溶液的温度设定在40℃；③贫碳MEA溶液的贫碳胺荷载比(lean loading)设置在0.2 $mol_{CO_2}\cdot mol^{-1}_{MEA}$左右[18,19]；④通过改变贫碳MEA溶液的流速，碳捕集水平可以设置为来自离散值集合{50%，60%，70%，80%，90%}中的某一特定数值。值得注意的是，实际中的lean loading测量困难，重沸器(reboiler)的温度会由其热负荷决定并以此体现lean loading的大致情况。

在给出了吸收塔、汽提塔的规格(表1)及废气、贫碳MEA溶液的基本输入设定(表2)后，我们进一步改变贫碳MEA溶液的摩尔流速和lean loading，意图在最低重沸器热负荷下，实现特定的碳捕集水平。表3摘录了各个设定下的运行表现，从中我们可以确定在第t季度不同碳捕集水平c_t对应的最优的重沸器热负荷$Q_{reb}(c_t)$。

2.2. 亚临界燃煤电厂的利润模型

燃煤电厂消耗了全世界范围内绝大部分能源并且在所有电力生产系统中排放了最多的二氧化碳，因此十分重要[20]。这一节中，我们为配有碳捕集的亚临界燃煤电厂构建了利润模型。我们也会同时建立电厂运行设定与电力输出之间的关系。依据美国能源情报署(EIA)的报告[21]，电厂第t季度的成本C_t(USD·qtr^{-1})可由如下公式计算：

$$C_t = FOM + VOM_t + F_t + B_t \tag{1}$$

其中，FOM为每一季度固定的运行和维修(OM)成本；VOM_t为每一季度变化的运行和维修成本；F_t为每一季度的燃料成本；B_t为每一季度碳配额竞标成本。根据加利福尼亚州碳配额拍卖机制[11]，竞标成本可定义为

$$B_t = v_t \cdot w_t \tag{2}$$

其中，v_t为第t季度拍卖的碳配额结算价格(USD·$allowance^{-1}$)；w_t为决策者第t季度通过拍卖赢取的碳配额吨数(t·$allowance^{-1}$)。配额1 t将允许电厂排放1 t的二氧化碳。式(1)的其他项则定义如下：

$$FOM = 0.25\cdot\beta\cdot P_n \tag{3}$$

$$VOM_t = \delta\cdot E_t/1000 \tag{4}$$

$$F_t = f\cdot H_t \tag{5}$$

其中，β和δ分别为固定和变化的运行维修成本系数；f为单位燃料成本；P_n为电厂额定容量。这些变量及其单位都列在表4中。此外，E_t是电力输出(kW·h·qtr^{-1})，H_t是燃料消耗(GJ·qtr^{-1})。燃煤电厂电力生产的收益可以表示为

$$R_t = \lambda\cdot E_t \tag{6}$$

其中，λ为电价[USD·$(kW\cdot h)^{-1}$]并在表4中给出。之后，第t季度的利润可作如下推导:

$$\begin{aligned} P_t &= R_t - C_t \\ &= \lambda\cdot E_t - 0.25\cdot\beta\cdot P_n - \delta\cdot E_t/1000 - f\cdot H_t - v_t\cdot w_t \\ &\triangleq P(E_t, H_t, w_t, v_t) \end{aligned} \tag{7}$$

式中，利润P_t是随变量E_t，H_t，w_t和v_t变化的，所以我们将其记作P_t=$P(E_t, H_t, w_t, v_t)$。

表1 吸收塔和汽提塔参数

Parameters	Absorber	Stripper
Packing type	Mellapak	Mellapak
Dimension	250Y	250Y
Number of columns	2	1
Diameter (m)	16.9	16.9
Packing height (m)	23.5	23.5
Top stage pressure (Pa)	101 325	170 273

表2 燃烧后碳捕集物质流

Parameters		Flue gas	Lean MEA solvent
Mole flow rate (kmol·s^{-1})		25	—
Temperature (°C)		40	40
Pressure (Pa)		105 117	170 273
Mass fraction	MEA	0	0.3098
	H_2O	0.0964	0.6434
	CO_2	0.2068	0.0468
	N_2	0.6703	0
	O_2	0.0265	0

表3 不同设定下燃烧后碳捕集过程的运行性能

Capture level	Lean MEA flow rate (kg·s^{-1})	Lean loading ($mol_{CO_2}\cdot mol^{-1}_{MEA}$)	Q_{reb} (MW_{th})
50%	948.8	0.20	293.3
60%	1148.2	0.20	354.5
70%	1350.3	0.20	416.9
80%	1557.3	0.20	480.9
90%	1837.5	0.21	547.3

表4 配有碳捕集的电厂的参数[21]

Parameters	Symbol	Value	Unit
Nominal capacity	P_n	650 000	kW
Capacity factor	ζ	0.55	Unitless
Efficiency	η	38.78	%
Fixed OM coefficient	β	80.53	USD·$(kW\cdot a)^{-1}$
Variable OM coefficient	δ	9.51	USD·$(MW\cdot h)^{-1}$
Electricity price	λ	0.102	USD·$(kW\cdot h)^{-1}$
Fuel price	f	1.545	USD·GJ^{-1}

在本文中，电力生产和碳捕集的总燃料消耗H_t假定为常量。据此，主电厂的电力输出和碳捕集设施的捕集能耗是要相互权衡的。对于如表4所设定的额定容量为P_n=650 000 kW的燃煤电厂，1个季度的燃料消耗可按下式计算：

$$H_t = P_n / \eta \cdot 2190 \cdot 3600 / 10^6 \cdot \zeta = 7267999 \ \mathrm{GJ \cdot qtr^{-1}} \tag{8}$$

式中的“2190”代表1个季度的小时数。每一季度的碳捕集能耗U_t(GJ·qtr^{-1})需由下式约束：

$$H_t = U_t + E_t / \eta \cdot 3600 / 10^6 \tag{9}$$

U_t应是第2.1节中讨论的重沸器热负荷的函数，即

$$U_t = Q_{\mathrm{reb}}(c_t) \cdot 3600 / 1000 \cdot 2190 \cdot \zeta \tag{10}$$

结合式(9)和(10)，第t季度的电力输出E_t可以推导为

$$E_t = 10^6 / 3600 \cdot [H_t - 7884 \cdot Q_{\mathrm{reb}}(c_t) \cdot \zeta] \cdot \eta \triangleq E(c_t) \tag{11}$$

以上算式表明碳捕集水平c_t可以唯一确定电力输出E_t。电厂每一季度的利润[公式(7)]可以简化为

$$P_t = P(H_t, E(c_t), w_t, v_t) \triangleq P(c_t, w_t, v_t) \tag{12}$$

总体说来，对于某一配有碳捕集技术的特定电厂，假定燃料消耗H固定以及燃料和电力价格(f和l)持续不变，电厂第t季度的利润P_t可以由碳捕集水平c_t、碳拍卖市场中获得的碳配额数量w_t和单位碳配额的结算价格v_t唯一确定。

2.3. 碳配额市场

在第2.2节，虽然一个季度的利润P_t已经被完全定义，但我们只讨论了E_t和H_t两个自由度。另外两个自由度的变量w_t和v_t会受到碳配额市场条件的影响。只有当市场包含的或选择加入的所有实体(如发电公司)向拍卖操作员上交了竞标选项(包括竞标数量q和竞标价格p)后，每一季度的市场条件才能被完全定义。决策者相关的实体的竞标数量和竞标价格分别记为$q_{0,t}$和$p_{0,t}$；所有其他实体的竞标数量和竞标价格分别记为$q_{i,t}$和$p_{i,t}$。其中，$i \in \mathbb{I}$，$\mathbb{I}=\{1, 2, 3, \cdots, I\}$，是除去决策者实体外的、包含在碳配额市场下的实体的集合。操作员之后会实行如下的密封竞标拍卖机制[14]。

在一个季度中，若某一实体或竞标人的竞标违反了购买限制、持有限制或竞标担保的相关规定，拍卖操作员会因其不合格而拒绝该竞标。接下来，会依据竞标价格按降序考虑所有竞标人的合格的竞标方案。从出价最高的竞标开始，各个竞标价格下呈递的竞标方案都会依竞标人的竞标数量出售等量的碳配额，直到满足如下任意一条件：一是所有碳交易市场下的碳配额A(单位：allowance)售完；二是下一个竞标人的竞标价格低于拍卖的竞标底价g_t(单位：USD·allowance^{-1})[11]。如果拍卖的碳配额售完，则结算价格是出售了碳配额的最后一个竞标方案的竞标价格；如果结算价格等于拍卖底价，则出售的碳配额应该等于所有在拍卖底价之上的各竞标方案共同累计的竞标数量。拍卖操作员之后可以计算出每一个竞标人或实体赢得的碳配额(如决策者赢得的碳配额在本文中记为w_t)、总共售出的碳配额u_t和对应于所有实体的一个统一结算价格v_t，即

$$w_t = w(q_{0,t}, p_{0,t}, q_{1,t}, p_{1,t}, \ldots, q_{I,t}, p_{I,t}) \tag{13}$$

$$v_t = v(q_{0,t}, p_{0,t}, q_{1,t}, p_{1,t}, \ldots, p_{I,t}, q_{I,t}) \tag{14}$$

$$u_t = u(q_{0,t}, p_{0,t}, q_{1,t}, p_{1,t}, \ldots, q_{I,t}, p_{I,t}) \tag{15}$$

在式(13)、式(14)或式(15)中，决策者只能决定他自身的竞标数量$q_{0,t}$和竞标价格$p_{0,t}$。但利用其他实体的历史竞标数据，决策者可以对其他公司的竞标选项进行估计。例如，在第2.4节，其他实体的各个竞标选项是由概率来表示的。

在本文中，为了判断一个竞标方案是否合格，我们只考虑电厂决策者的持有限制h_1(单位：allowance)。为了简单，不再考虑购买限制和竞标担保。持有限制是包含在碳配额市场下某一实体账户的持有上限。如果任意实体的竞标数量会潜在地导致其持有账户里的碳配额超过持有限制，那么这一呈递的竞标方案将会因不合格而受到拍卖操作员的拒绝。本文中的持有账户只拥有加利福尼亚州出台的法规中多个账户的类似功能[11]。我们假设持有账户的碳配额有以下约束：

$$h_{t+1} \le h_t + q_{0,t} \le h_1 \tag{16}$$

$$h_{t+1} = h_t + (w_t - e_t) \ge 0 \tag{17}$$

其中，h_{t+1}为第t+1季度拍卖开始时持有账户的碳配额。值得注意的是，如果赢得的碳配额w_t低于电厂碳排放e_t，额外的碳配额将从持有账户中上缴；如果赢得的碳配额高于碳排放，则剩余的赢取的碳配额将保留至持有账户。在第t季度前的所有季度中，累积在持有账户的总碳配额被记为h_t。式(17)的不等式表明持有账户的碳配额不能耗尽，否则，没有缴纳碳配额的过度排放将产

生额外的罚金。根据加利福尼亚州法规[11]，对于某实体产生二氧化碳排放并未及时上缴碳配额的那部分，作为惩罚，新的遵守义务将是上缴更多的碳配额，其数量四倍于本应上缴的碳配额数量。因未及时上缴而导致的该义务的履行需要额外的竞标和贸易机制。为了简单，不同于加利福尼亚州法规中规定的惩罚，我们假定未及时上缴时，每吨过度排放二氧化碳需交的罚金为320 USD。因此，式(17)可以认为是一个软边界。另外，式(16)表明对于任意t，决策者只能呈递不会潜在地导致h_{t+1}超过之前提到的h_l的竞标数量$q_{0,t}$。变量e_t代表电厂第t季度的碳排放(单位：$\mathrm{t\cdot qtr^{-1}}$)，由下式决定：

$$e_t = 148.6\cdot(1-c_t)\cdot 3600\cdot 2190/1000\cdot\zeta \triangleq e(c_t) \tag{18}$$

c_t碳捕集水平是基于溶液的碳捕集过程的运行参数，而w_t和$q_{0,t}$是在碳配额市场下的竞标参数。式(16)和式(17)中的不等式表明了竞标和运行的潜在关系。

2.4. 目标构建

在第2.2节，利润[式(12)]可由碳捕集水平c_t、决策者赢得的碳配额数量w_t和结算价格v_t表示。碳捕集水平c_t可以由决策者任意确定，而赢得的碳配额数量w_t和结算价格v_t必须由式(13)和式(14)中出现的所有实体的竞标选项决定。如果所有其他实体已经呈递了他们的竞标选项(如任意$i\in\mathbb{I}$下的$p_{i,t}$和$q_{i,t}$)，配有碳捕集电厂的决策者只需决定自己运行方案c_t和竞标方案$(q_{0,t}, p_{0,t})$，就能估计式(12)中相应的利润。可以定义如下统一的决策者的行为是

$$a_t = (c_t, q_{0,t}, p_{0,t})^{\mathrm{T}} \in \mathbb{A}(s_t)=\mathbb{A} \tag{19}$$

其中，$\mathbb{A}(s_t)$是在状态s_t下的一个离散的行为集合，并且进一步认为是不随状态变化的集合$\mathbb{A}$。决策者的电厂只知道自己的竞标数量$q_{0,t}$和价格$p_{0,t}$；任意$i\in\mathbb{I}$下的其余竞标人的$q_{i,t}$和$p_{i,t}$必须由决策者通过先验知识获得。在本文中，其他实体的竞标数量和价格被提前假定为受到上一季度拍卖市场结算价格、出售碳配额的影响。一个相似的状态选择方案在文献[22]中有所讨论。因此，第t季度的状态s_t可按下式记为

$$s_t = (v_{t-1}, u_{t-1}, h_t, t)^{\mathrm{T}} \in \mathbb{S} \tag{20}$$

其中，因为持有账户的碳配额h_t必须充足[式(17)]但不能潜在地超过持有上线h_l[式(16)]，所以被认为是一个状态变量。此外，我们趋向于最大化所关心的电厂寿命下的贴现累计利润。因此，时间t也作为式(20)中s_t的一个状态。不同时间状态表示的是电厂寿命的不同阶段，在各个阶段决策者可以设定不同的行为方案。

若实体的竞标数量集合和竞标价格集合分为$\mathbb{Q}_i$和$\mathbb{P}_i$，则决策者可估计任意竞标人选择某竞标选项的概率

$$\kappa(s, p_i, q_i) = \Pr(q_{i,t}=q_i, p_{i,t}=p_i \mid s_t=s, q_i\in\mathbb{Q}_i, p_i\in\mathbb{P}_i) \tag{21}$$

需对任意$i\in\mathbb{I}$成立。虽然每个季度中实体会选择自己的竞标选项，对于数量和价格的竞标选项的集合(即$\mathbb{Q}_i$和$\mathbb{P}_i$)在本文中假定为时不变且不随状态而变化。之后，我们建立以下马尔可夫决策过程。在特定的状态s_t=s下，决策者会执行行为a_t=a。在定义了联合概率

$$\mathbb{P}^a_{ss'} = \prod_{i\in\mathbb{I}} \kappa(s, p_i, q_i) \tag{22}$$

后，所有其他竞标人将如式(22)所设定的那样选择他们自己的竞标选项，使得下一季度的状态$s_{t+1}=(v_t, u_t, h_{t+1}, t+1)=s'$在执行行为$a_t=a$后可以被唯一确定。进一步地，奖励值函数$r_{t+1}$是由状态$s_t$过渡至$s_{t+1}$时基于式(12)导出，即

$$r_{t+1} \triangleq P_t = P(c_t, w_t, v_t) \tag{23}$$

注意到“奖励值”是在强化学习(RL)框架下定义的专业术语。从物理意义上来说，第t+1季度的奖励值r_{t+1}就是电厂在第t季度的利润P_t[式(12)]。因为t可以是任意时刻，决策者可以递归地获得有限时间区间内的奖励值序列$s_t, a_t, s_{t+1}, r_{t+1}, a_{t+1}, s_{t+2}, r_{t+2}, \cdots, a_{t+N-1}, s_{t+N}, r_{t+N}$，称为一段竞标和运行事件。变量$N$表示配有基于乙醇胺溶液碳捕集过程的电厂的寿命。对于$\forall k\in\{0, 1, \cdots, N-1\}$，目标函数可以构建为

$$\max_{\pi} V^{\pi}(s) = \max_{\pi} \mathbb{E}_{\pi}\left\{\sum_{k=0}^{N-1}\gamma^k r_{t+k+1} \mid s_t=s\right\} \tag{24}$$

并服从

$$r_{t+k+1} = P(c_{t+k}, w_{t+k}, v_{t+k}) \tag{25}$$

$$h_{t+k} + q_{0,t+k} \le h_l \tag{26}$$

$$h_{t+k+1} = h_{t+k} + (w_{t+k} - e_{t+k}) \ge 0 \tag{27}$$

其中，$V^p(s_t)$为在方针p下关于s_t的状态值函数；r_{t+k+1}为由状态s_{t+k}过渡至s_{t+k+1}的奖励值；γ为贴现率；$\mathbb{E}_{\pi}\{\cdot\}$为贴现奖励值序列在策略$p$下的期望。对于决策者，一个随机方针的概率可以写为$p(s_{t+k}, a_{t+k})$，其中，在各个状态下任意行为出现的概率应该事先由决策者确定，目的是最大化电厂寿命下的贴现累计利润。我们先考虑随机方

针是因为最优方针是由基于强化学习的Sarsa算法搜索得到的。最后，随机方针将逐渐演变为一个可应用的确定优化方针。此外，式(26)和式(27)可由式(16)和式(17)分别获得。

3. Sarsa 算法：介绍与应用

基于强化学习的Sarsa时间差分算法可在Maltlab®中实现，找到在第2节中所定义问题的最优方针或策略。由于该算法的自适应及无模型特性，它可使电厂利润最大化。一个最初的优化方针可以根据第2节中的仿真环境自行找到；进一步的策略调整可以在agent与真实环境互动后实现。Sarsa时间差分算法比起动态规划法计算时间更少，而比起另一种基本的强化学习法(Q-学习法)[23]，收敛性要更好。然而，如果调节参数值(如ε)的计划给定序列不恰当，Sarsa时间差分算法经常会找到一些较差的方针。参数ε是在行为集合$\mathbb{A}$内的等可能寻找行为的概率。在ε-贪心方针下，各个状态下的应对方针都包含两种模式。一种是直接选择该状态下的贪心行为。该贪心行为是当前状态下决策者认为最好的行为，且无穷迭代后将收敛至最优行为。选择该行为模式的概率是1–ε。另一种是为了探索当前状态下各个行为的优劣，即是否存在更好的行为。选择该模式的概率为ε。若决策者随机地进入了探索模式，行为集合$\mathbb{A}$内各个行为都有被采用的可能，各行为被采用的概率都是$\varepsilon/n_{\mathbb{A}}$。$n_{\mathbb{A}}$是集合$\mathbb{A}$包括的行为的数量，将在式(32)给出。

为了设计Sarsa时间差分算法，我们根据式(24)定义了一个最优行为值函数，即

$$Q^*(s,a) \triangleq \max_{\pi} Q^{\pi}(s,a) = \max_{\pi} \mathbb{E}_{\pi}\left\{\sum_{k=0}^{N-1} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a\right\} \quad (28)$$

该式对于所有$s \in \mathbb{S}$和$a \in \mathbb{A}$成立。其中，Q^{π}为在方针π下的行为值函数。以此定义的最优方针为

$$a^* = \arg\max_{a} Q^*(s,a) \quad (29)$$

然而，在最优方针并未获取的前提下，最优的行为值Q^*是未知的。根据文献[23,24]，行为值函数的迭代方案[式(30)]可以确保：行为值函数$Q_{t+k+1}(s, a)$在$k \to \infty$且无穷次访问所有状态$s \in \mathbb{S}$和所有行为$a \in \mathbb{A}$后，将收敛至$Q^{\pi}(s, a)$。该迭代方案记为

$$Q_{t+k+1}(s,a) = Q_{t+k}(s,a) + \alpha[r + \gamma Q_{t+k}(s',a') - Q_{t+k}(s,a)] \quad (30)$$

其中，a为当前方针π下的行为。如果根据式(30)得到的$Q^{\pi}(s, a)$的估计值为$\hat{Q}^{\pi}(s, a)$，改进方针可通过下式实现：

$$\tilde{a} = \arg\max_{a} \hat{Q}^{\pi}(s,a) \text{ for } \forall s \in \mathbb{S} \quad (31)$$

$$\pi'(s,a) = \begin{cases} 1-\varepsilon+\varepsilon/n_{\mathbb{A}}, \text{ if } a = \tilde{a} \\ \varepsilon/n_{\mathbb{A}}, \text{ if } a \neq \tilde{a} \end{cases} \quad (32)$$

其中，$\tilde{a}$为使方针$\pi'(s, a)$比$\pi(s, a)$有所提升的贪心行为；$n_{\mathbb{A}}$为行为集合$\mathbb{A}$中包含的行为的数量；ε为从行为集合$\mathbb{A}$中均匀选取任意行为的概率。除了贪心行为，其他所有行为称为探索行为。探索行为可以保证找到每个状态对应的全局最大状态值V^{π}[式(24)]而非局部最大值。通过设定新的方针$\pi \leftarrow \pi'$，式(30)至式(32)构成了一组可重复迭代无穷多次的行为值迭代算法。这一算法可以用来获取最优方针π^*。但使用前，a和ε都需设置计划值序列。学习率a起初要很大来实现对$Q(s, a)$[式(30)]的快速初始化，但最终减小以确保行为值的收敛。虽然a的计划值序列存在理论条件，但它们很少被实际应用[23]。探索行为集合的概率ε在探索初期等于1，但会逐渐降低至0，因为我们最终将推出一个确定方针。表5给出应用了ε-贪心方针的Sarsa时间差分算法的伪代码。

表5 应用ε-贪心方针的Sarsa时间差分算法

Input	Discount coefficient γ; scheduled ε and α; arbitrary policy π
Initialization	Initialize $Q(s, a)$ for all $s \in \mathbb{S}$, all $a \in \mathbb{A}$
For each **policy improvement**	
For every episode μ	
Initialize s, choose a for state s with the ε-greedy policy π	
For each step of an episode	
Take action a and observe r, s'	
Choose a' for state s' with the ε-greedy policy π	
$Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma Q(s', a') - Q(s, a)]$	
$s \leftarrow s'$, $a \leftarrow a'$	
End for	
End for	
$\hat{Q}^{\pi}(s, a) \leftarrow Q(s, a)$ for all $s \in \mathbb{S}$ and all $a \in \mathbb{A}$	
Policy improvement:	
Apply Eqs. (31) and (32); $\pi \leftarrow \pi'$ for all $s \in \mathbb{S}$ and all $a \in \mathbb{A}$	
Scheduled parameter update: ε, α	
End for	

Sarsa时间差分算法是无模型的在线算法，因此可以通过与环境(即真实碳配额市场)的交互来实现。然而，在本文中，对于其他实体的竞标选项及相应概率的估计是必要的，这些可以用来形成一个模拟碳配额市场。这样的先验知识可由其他电厂的历史竞标数据

来获得。如果历史竞标数据无从获取，通过统计分析，历史的市场条件仍可以用来识别状态之间相互过渡的概率[22]。在此基础上，我们可以得到一个基于强化学习的Sarsa时间差分算法(表5)的初始方针。这样的好处是与真实碳拍卖市场的交互次数减少。文献[23]给出了一个统一的规划和学习视角，同时考虑了仿真模型和真实环境。

4. 结果与讨论

在案例研究中，共有8个包含在市场下的实体，标记为0，1，2，3，4，5，6，7。实体0是运行燃煤电厂的决策者，其设定如表4所示。这一实体被假定为我们自己的公司，并且想要最大化该公司下一家电厂的贴现累计利润。决策者将利用Sarsa时间差分算法来寻找各个状态下合适的竞标和运行行为。实体1运行着一家与实体0在设定上完全相同的电厂，但采用了不同的竞标和运行策略。该策略将会在第4.3节介绍。其他实体(即实体2~7)的竞标策略是事先定义且被假定由决策者的仿真环境来预先估计。对于式(24)中的目标函数，初始时刻是t=0，而相应的时间范围是N=100季度(即电厂的寿命是25年)，因此k的取值范围是$k\in\{0, 1, 2, \cdots, 99\}$。由以上定义推知，任意随时间变化的量可由$k$来索引，如$s_k$，$a_k$，$r_{k+1}$。拥有25年寿命的电厂的年度贴现率设置为8%[9]，所以季度制的贴现率为$\gamma = 1/(1 + 8\%)^{0.25} \approx 0.98$。持有限制$h_l$的制定需基于年度碳配额预算[11]。然而，年度碳配额预算是计划制定的，且可能每年各不相同。为了简单起见，本文认为h_l是一个常数，即6×10^6 allowances。在表5中，μ是8段事件，a由1/20逐渐变至1/200，ε由1逐渐变至0.1。变量a和ε都在执行方针改进后改变。

式(20)的状态变量需要聚合为离散的子集来缓解状态空间的维数灾难；该方法又被称为状态聚合[22,23]。状态聚合可由如下方式实现：结算价格和出售碳配额将一起被考虑，因为当一个处于某个特定的区间时，另一个也应被约束在某个特定值。例如，当碳配额拍卖中的出售碳配额u_{k-1}小于总体拍卖的碳配额A = 1 500 000 t，则结算价格必须等于竞标底价g，即表6中i_s=1，2，3对应行的状态。同样地，时刻k和持有账户的碳配额h_k分别聚合并记录在表6、表7中。根据这两个表格，原始的状态空间$\mathbb{S}$被离散为8 × 5 × 14 = 560个聚合状态。行为变量[式(19)]则归为两部分。其一是运行部分，包括5种决策者燃煤电厂可设定的碳捕集水平，即表3中的$\mathbb{C}=\{50\%, 60\%, 70\%, 80\%, 90\%\}$；其二是16种可行的竞标选项，包括竞标数量和价格，即$(q_0, p_0)\in\mathbb{B}_0$。类似于其他实体的竞标数量和竞标价格集合(即$\mathbb{Q}_t$和$\mathbb{P}_t$)，我们只考虑时不变的、不随状态变化的竞标选项集$\mathbb{B}_0$。由此，对于决策者的各个聚合状态，共有5 × 16 = 80种不同的行为方式。接下来的内容里，我们将在决策者实际采用了某个特定行为时提及该行为的具体值。为简便起见，由$\mathbb{C}$和$\mathbb{B}_0$集合给出的80中不同行为并未在本文列出。

4.1. Sarsa 时间差分算法的收敛性

在这一节，我们提供了特定状态某个行为的行为值的收敛特性。由于状态都已聚合，我们只考虑在表6和表7中各个状态聚合后的子集，而非实际任意k下的状态值s_k。图2展示了某一状态行为对(s, a)的行为值$Q(s, a)$的收敛情况。其中，状态s被归为元组(i_s, j_s, v_s) = (7, 7, 4)，而行为a的索引为i_a=61。索引i_a=61对应的行为是a = (300 000, 14.5, 27)。该行为事先设定在离散行为集合$\mathbb{A}$中。这一状态行为对应的$Q(s, a)$通过式(30)的迭代将收敛至式(29)中的最优行为值$Q^*(s, a)$。由于表5中的算法可以访问到所有的状态行为对(s, a)，因此对于任意$s \in \mathbb{S}$和$a \in \mathbb{A}$，表5的迭代会得到对应的$Q^*(s, a)$的值，从而各个状态的最优行为可由式(29)确定。如前所述，每个

表6 结算价格和售出碳配额对(v_{k-1}, u_{k-1})的子集及时刻k的子集分布

Level i_s	v_{k-1} domain	u_{k-1} domain	Level v_s	k domain
1	$v_{k-1} = g$	$[0, 0.5A)$	1	{0, 2, ..., 24}
2	$v_{k-1} = g$	$[0.5A, 0.8A)$	2	{25, 26, ..., 49}
3	$v_{k-1} = g$	$[0.8A, 1.0A)$	3	{50, 51, ..., 74}
4	$(1.0g, 1.1g)$	$u_t = A$	4	{75, 76, ..., 99}
5	$[1.1g, 1.2g)$	$u_t = A$	5	k = 100
6	$[1.2g, 1.3g)$	$u_t = A$		
7	$[1.3g, 1.4g)$	$u_t = A$		
8	$[1.4g, \infty)$	$u_t = A$		

表7 持有账户碳配额的子集分布

Level j_s	h_k domain (× 1000)	Level j_s	h_k domain (× 1000)
1	[0, 64]	8	(2050, 3050]
2	(64, 129]	9	(3050, 4050]
3	(129, 193]	10	(4050, 5050]
4	(193, 258]	11	(5050, 5700]
5	(258, 322]	12	(5700, 5750]
6	(322, 1050]	13	(5750, 5850]
7	(1050, 2050]	14	(5850, 6000]

状态都有如图3所示的80种行为值。基于这些行为值，我们可以得出行为索引是$i_a=61$的行为有最大的Q值，所以它给出了在该状态下最好的行为方式。最优方针因此可以通过寻找每个聚合状态下的最大行为值找到。

4.2. Sarsa 时间差分算法的性能

在这一节，通过与大多数相关文献中设定固定碳捕集水平的运行方式对比，我们展示了考虑到随时间变化碳捕集水平的Sarsa时间差分算法可以获取更多的贴现累计利润。k=0时刻的初始竞标底价为12.73 USD·allowance^{-1}[11]。另外，年度竞标底价的增长率τ被引入，用来提高每年的竞标底价。年度竞标底价增长率可以刺激新兴碳捕集和埋存技术的进步。某一结算价格的案例已显示在图4中。从中不仅可以看到结算价格在整个时间范围内(即100个季度)的波动，而且由于5%的年度增长率，结算价格出现了计划增长。该年度增长率同样设置于加利福尼亚和魁北克的联合温室气体拍卖机制中[11,14]。

为了展现Sarsa时间差分算法的自适应性，计划的碳配额竞标底价的年度增长率τ被分别假定为0%、5%、10%和15%。如果特定的年度增长率固定在如图5所示的$\tau=0\%$，则除了竞争对手(competitor，即Entity1)所示曲线，四组不同的奖励值序列表明了实体0的决策者采用四组不同竞标和运行策略的获利情况。其中一组的竞标运行策略是用考虑到随时间变化碳捕集水平的Sarsa时间差分算法找到的。对于其余策略：运行方面，固定了碳捕集水平(即碳捕集水平在相关的一段竞标和运行事件中被设置为50%、70%或90%)；竞标方面，事先定义了各聚合状态下所有行为中各竞标选项出现的概率。对于固定了碳捕集水平的策略，可能的竞标选项来自于竞标选项集合$\mathbb{B}_0$。这一集合也是基于Sarsa的统一的竞标和运行策略集合。之前的奖励值序列显示了以季度为基础的电厂寿命下的利润。通过计算特定奖励值序列的贴现和，我们可以获得某个特定竞标和运行策略的贴现累计利润。例如，根据图5，我们得出在年度竞标底价

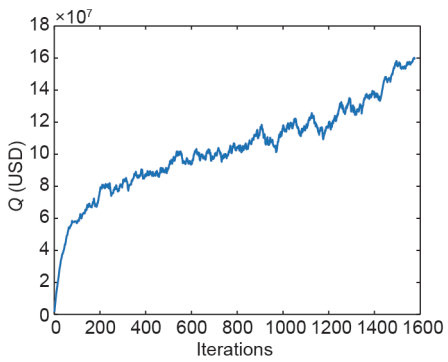

图2. 典型状态行为对在$i_s=5$，$j_s=7$，$v_s=4$，$i_a=6$处的收敛性。

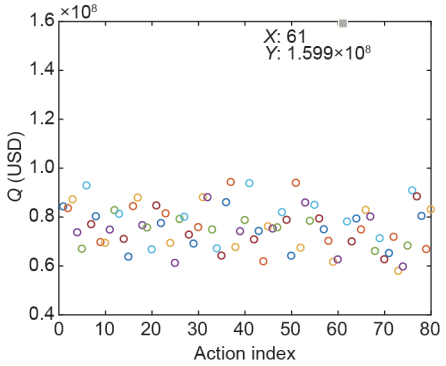

图3. 在特定状态$i_s=5$，$j_s=7$，$v_s=4$处所有行为的行为值。

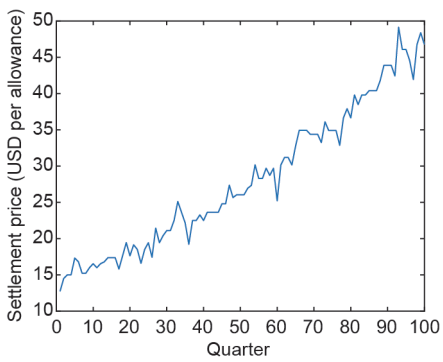

图4. 在年增长率$\tau=5\%$和初始竞标底价$g=12.73$ USD·allowance^{-1}下的结算价格。

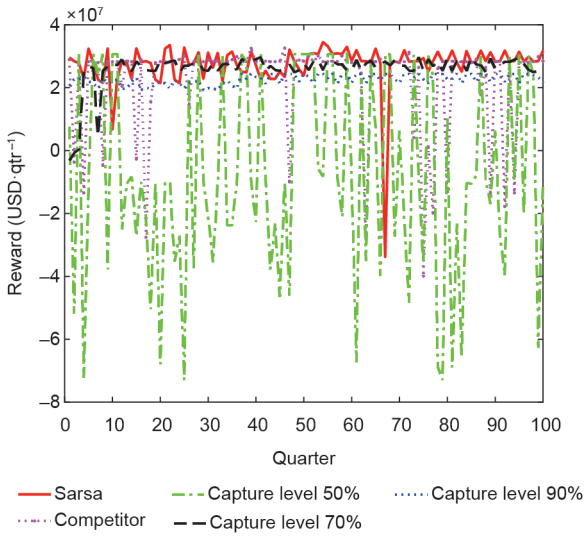

图5. 在年增长率$\tau=0\%$和持有账户初始碳配额$h_0=0.05\times10^6$ allowances下不同竞标和运行策略的奖励值。

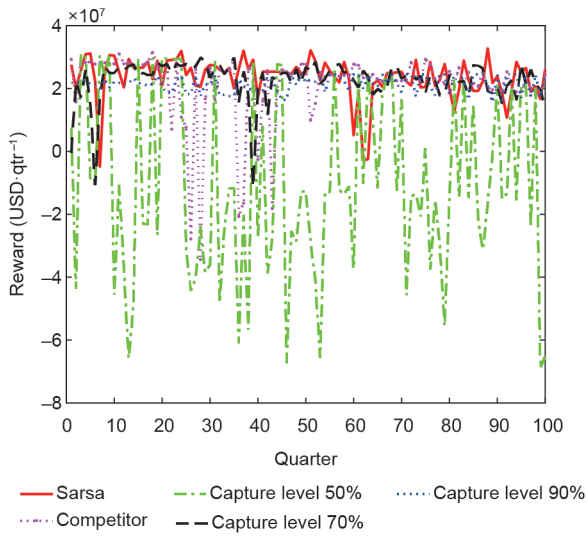

图6. 在年增长率$\tau=5\%$和持有账户初始碳配额$h_0=0.05\times10^6$ allowances下不同竞标和运行策略的奖励值。

增长率τ为0时各个策略对应的贴现累计利润。

类似地，如图6至图8所示，我们可以得到在持有账户初始碳配额为$h_0 = 0.05\times10^6$和τ由5%变化至15%时的贴现累计利润。在持有账户特定初始碳配额为h_0=0.05 $\times10^6$ t的前提下，不同竞标底价增长率对应的贴现累计利润如图9所示。

对于以其他持有账户初始碳配额数量为前提的贴现累计利润展示在图10和图11中。可以推知，无论采用基于固定碳捕集水平方案的固定碳捕集水平设置为何值，Sarsa时间差分法找到的统一、灵活的运行和竞标策略都表现得更为出色。

4.3. 与碳配额市场下另一实体的比较

在同一碳配额市场下，我们将比较采用了Sarsa时间差分算法的决策者和其竞争对手(实体1)的获利情况。对于该竞争对手，其电厂的所有设置都假定与实体0的决策者的相同。而在关于运行和竞标策略的问题上，实体1将其碳捕集水平固定在60%，而它的竞标选项则从与实体0选用的相同集合$\mathbb{B}_0$中独立选取。我们假设实体1在竞标选项的选择行为上可由实体0的决策者根据玻尔兹曼分布近似，即

$$\Pr(y)=\exp[\omega(y)/\xi]/\sum_{z=1}^{n_b}\exp[\omega(z)/\xi] \tag{33}$$

其中，y和z为可选的竞标选项的索引；n_b为总的竞标选项的数量(对于竞标选项集$\mathbb{B}_0$，n_b=16)；$\Pr(y)$记为选择以y为索引的竞标选项的概率；ξ为分布的温度常数。从式(33)中可以看出，ξ很大时，所有可能的竞标选项的选择接近于等概率。为了简单，本例中的温度常数为ξ=1。变量ω代表由y或z索引的各个选项的权重。在本文有关仿真中，所有权重中最大的被定义为常值，即$\omega_{max}=n_b$。假定第13个竞标选项被赋予最大的权重，即$\omega(y=13)=\omega_{max}=16$，则以索引$y=13$为中心，所有竞标选项的权

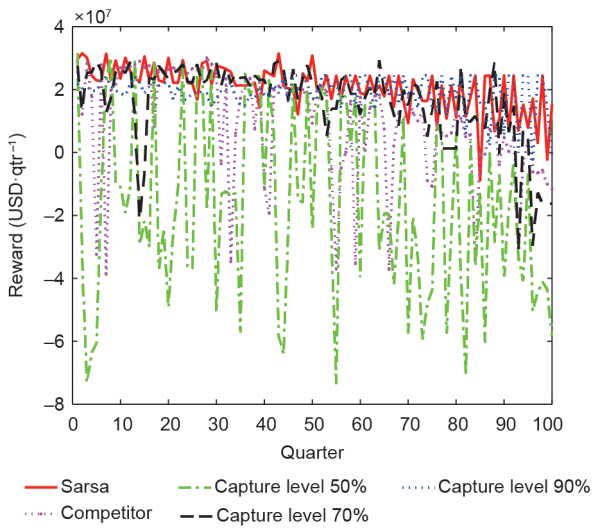

图7. 在年增长率τ = 10%和持有账户初始碳配额h_0 = 0.05 × 10^6 allowances下不同竞标和运行策略的奖励值。

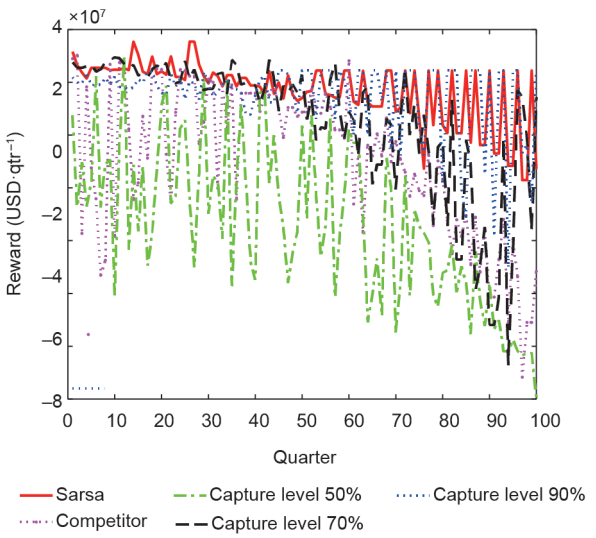

图8. 在年增长率τ = 15%和持有账户初始碳配额h_0 = 0.05 × 10^6 allowances下不同竞标和运行策略的奖励值。

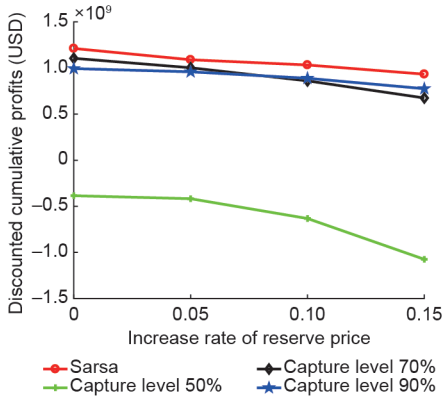

图9. 在持有账户初始碳配额h_0 = 0.05 × 10^6 allowances下的贴现累计利润。

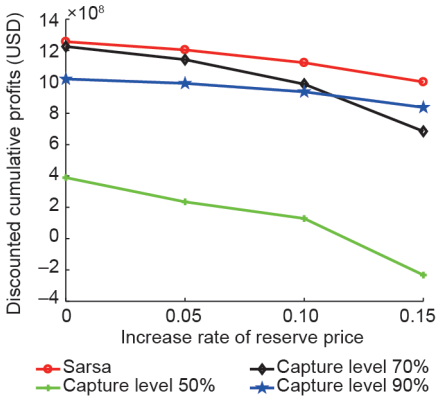

图10. 在持有账户初始碳配额h_0 = 3 × 10^6 allowances下的贴现累计利润。

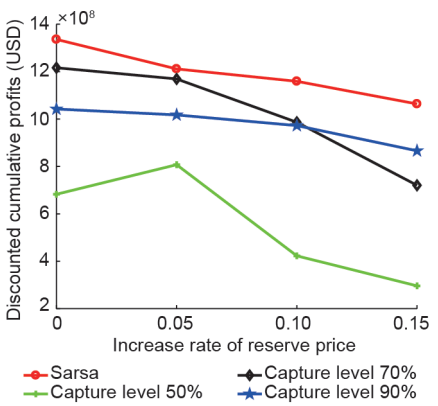

图11. 在持有账户初始碳配额h_0 = 5 × 10^6 allowances下的贴现累计利润。

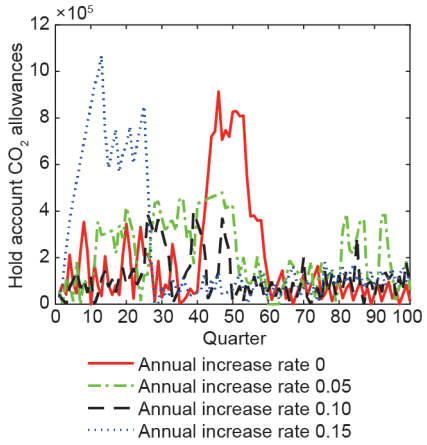

图12. 不同竞标底价下采用Sarsa时间差分策略的决策者持有账户的碳配额数量。

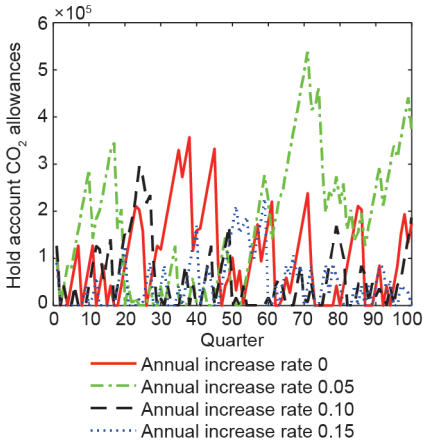

图13. 不同竞标底价下竞争对手持有账户的碳配额数量。

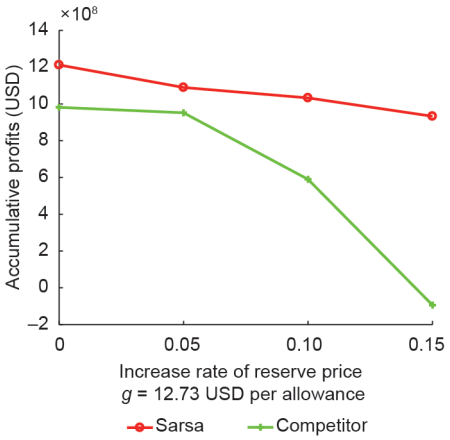

图14. 在持有账户初始碳配额h_0= 0.05 × 10^6 allowances下实体0的决策者和实体1的贴现累计利润。

重依据离该中心的远近而指定为：4，5，6，7，8，9，10，11，12，13，14，15，16，15，14，13。根据这些权重，选择某个可能竞标选项的概率可以由式(33)事先定义。实际中，决策者需获得该竞争对手的历史竞标数据或历史市场条件来识别这些权重。

决策者持有账户的碳配额h_k和竞争对手的碳配额数量分别如图12和图13所示。两个实体的贴现累计利润则可根据图5和图6得到并记录在图14中。图中，决策者在不同竞标底价的年度增长率下，都获得了更多的贴现累计利润。这些表明，决策者通过Sarsa时间差分法得到的竞标和运行策略要比处在同一碳配额市场下的竞争对手的策略更好。

5. 结论

本文为燃煤电厂设计了由Sarsa时间差分算法得到的统一的竞标和运行策略。当考虑了来自碳捕集水平集合的随时间变化的灵活捕集水平和一个共同设计的竞标选项集合后，提出的策略要优于运行上固定碳捕集水平、竞标上独立设计的策略。Sarsa时间差分算法能够最大化在不同碳配额市场条件下电厂的贴现累计利润，如竞标底价采用不同的年增长率或是持有账户拥有不同的初始碳配额。此外，通过与另一个固定其碳捕集水平并以玻尔兹曼随机分布独立设计竞标策略的电厂比较，运用了Sarsa时间差分算法的决策者在碳配额市场下更具竞争力。

Compliance with ethics guidelines

Ziang Li, Zhengtao Ding, and Meihong Wang declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Lawal A, Wang M, Stephenson P, Yeung H. Dynamic modelling of CO_2 absorption for post combustion capture in coal-fired power plants. Fuel 2009;88(12):2455–62.
[2] Wang M, Lawal A, Stephenson P, Sidders J, Ramshaw C. Post-combustion CO_2 capture with chemical absorption: A state-of-the-art review. Chem Eng Res Des 2011;89(9):1609–24.
[3] Lin YJ, Pan TH, Wong DSH, Jang SS, Chi YW, Yeh CH. Plantwide control of CO_2 capture by absorption and stripping using monoethanolamine solution. Ind Eng Chem Res 2011;50(3):1338–45.
[4] Lin YJ, Wong DSH, Jang SS, Ou JJ. Control strategies for flexible operation of power plant with CO_2 capture plant. AIChE J 2012;58(9):2697–704.
[5] Luu MT, Manaf NA, Abbas A. Dynamic modelling and control strategies for flexible operation of amine-based post-combustion CO_2 capture systems. Int J Greenh Gas Control 2015;39:377–89.
[6] Nittaya T, Douglas PL, Croiset E, Ricardez-Sandoval LA. Dynamic modelling and control of MEA absorption processes for CO_2 capture from power plants. Fuel 2014;116:672–91.
[7] Sahraei MH, Ricardez-Sandoval L. Controllability and optimal scheduling of a CO_2 capture plant using model predictive control. Int J Greenh Gas Control 2014;30:58–71.
[8] Luo X, Wang M. Optimal operation of MEA-based post-combustion carbon capture for natural gas combined cycle power plants under different market conditions. Int J Greenh Gas Control 2016;48(2):312–20.
[9] Mac Dowell N, Shah N. Identification of the cost-optimal degree of CO_2 capture: An optimisation study using dynamic process models. Int J Greenh Gas Control 2013;13:44–58.
[10] Luckow P, Stanton EA, Fields S, Biewald B, Jackson S, Fisher J, et al. 2015 carbon dioxide price forecast. Cambridge (MA): Synapse Energy Economics, Inc; 2015 Mar.
[11] California Environmental Protection Agency. California cap on greenhouse gas emissions and market-based compliance mechanisms [Internet]. Eagan: Thomson Reuters; c2017 [cited 2016 Nov 5]. Available from: https://govt.westlaw.com/calregs/Browse/Home/California/CaliforniaCodeofRegulations?guid=I47A831C02EBC11E194EACEFFB46E37D1&originationContext=documenttoc&transitionType=Default&contextData=(sc.Default)&bhcp=1.
[12] Chen Y, Wang L. A power market model with renewable portfolio standards, green pricing and GHG emissions trading programs. In: Proceedings of the Energy 2030 Conference; 2008 Nov 17–18; Atlanta, USA. Piscataway: IEEE; 2008. p. 1–7.
[13] Nanduri V. Application of reinforcement learning-based algorithms in CO_2 allowance and electricity markets. In: Proceedings of the 2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL); 2011 Apr 11–15; Paris: France. Piscataway: IEEE; 2011. p. 164–9.
[14] Air Resources Board. 2016 detailed auction requirements and instructions, California cap-and-trade program and Québec cap-and-trade system joint auction of greenhouse gas allowances [Internet]. [cited 2016 Oct 24]. Available from: https://www.arb.ca.gov/cc/capandtrade/auction/auction.htm.
[15] AspenTech. Rate-based model of the CO_2 capture process by MEA using Aspen Plus. Burlington: Aspen Technology, Inc; 2008. 23p.

[16] Dugas RE. Pilot plant study of carbon dioxide capture by aqueous monoethanolamine [dissertation]. Austin: The University of Texas at Austin; 2006.
[17] Lawal A, Wang M, Stephenson P, Obi O. Demonstrating full-scale post-combustion CO_2 capture for coal-fired power plants through dynamic modelling and simulation. Fuel 2012;101:115–28.
[18] Agbonghae EO, Hughes KJ, Ingham DB, Ma L, Pourkashanian M. Optimal process design of commercial-scale amine-based CO_2 capture plants. Ind Eng Chem Res 2014;53(38):14815–29.
[19] Aroonwilas A, Veawab A. Integration of CO_2 capture unit using single- and blended-amines into supercritical coal-fired power plants: Implications for emission and energy management. Int J Greenh Gas Control 2007;1(2):143–50.
[20] Oko E, Wang M. Dynamic modelling, validation and analysis of coal-fired subcritical power plant. Fuel 2014;135:292–300.
[21] US Energy Information Administration. Updated capital cost estimates for utility scale electricity generating plants. Final report. Washington DC: US Energy Information Administration; 2013 Apr.
[22] Song H, Liu CC, Lawarrée J, Dahlgren RW. Optimal electricity supply bidding by Markov decision process. IEEE Trans Power Syst 2000;15(2):618–24.
[23] Sutton RS, Barto AG. Reinforcement learning: An introduction. Cambridge: MIT press; 1998.
[24] Busoniu L, Babuska R, De Schutter B, Ernst D. Reinforcement learning and dynamic programming using function approximators. Boca Raton: CRC press; 2010.