دانشگاه تربیت مدرّس
دانشکده فنی و مهندسی

پایان‌نامه دوره دکتری مهندسی برق–کنترل

طراحی کنترلگر عاملگرای هوشمند برای ناوبری ربات

ولی درهمی

استاد راهنما:

دکتر وحید جوهری مجد

اساتید مشاور:

دکتر مجید نیلی احمد آبادی

دکتر حمید رضا مومنی

پاییز ۱۳۸۶

بسمه تعالی

تاییدیه اعضای هیات داوران حاضر در جلسه دفاع از رساله‌دکتری

آقای ولی درهمی رساله‌۳۴ واحدی خود را **با عنوان طراحی کنترلگر عامل گرای هوشمند برای ناوبری ربات** در تاریخ ۱۳۸۶/۸/۶ ارائه کردند.

اعضای هیات داوران نسخه نهایی این رساله را از نظر فرم و محتوا تایید کرده و پذیرش آنرا برای تکمیل درجه دکتری مهندسی برق - کنترل پیشنهاد می کنند.

عضو هیات داوران	نام و نام خانوادگی	رتبه علمی	امضا
استاد راهنما	دکتر وحید جوهری مجد	دانشیار	
استاد مشاور	دکتر مجید نیلی احمدآبادی	دانشیار	
استاد مشاور	دکتر حمیدرضا مومنی	دانشیار	
استاد ناظر	دکتر سعید جلیلی	استادیار	
استاد ناظر	دکتر محمد رضا امین ناصری	دانشیار	
استاد ناظر	دکتر نجار اعرابی	دانشیار	
استاد ناظر	دکتر محمد فرخی	دانشیار	
نماینده شورای تحصیلات تکمیلی	دکتر سعید جلیلی	استادیار	

دستور العمل حق مالکیت مادی و معنوی در مورد نتایج پژوهشهای علمی دانشگاه تربیت مدرس

مقدمه:

با عنایت به سیاست های پژوهشی دانشگاه در راستای تحقق عدالت وکرامت انسانها که لازمه شکوفایی علمی و فنی است و رعایت حقوق مادی ومعنوی دانشگاه و پژوهشگران لازم است اعضای هیأت علمی دانشجویان دانش آموختگان و دیگر همکاران طرح درمورد نتایج پژوهشهای علمی که تحت عناوین پایان نامه وساله و طرحهای تحقیقاتی با هماهنگی دانشگاه انجام شده است موارد ذیل را رعایت نمایید:

ماده۱: حقوق مادی و معنوی پایان نامه‌ها / رساله های مصوب دانشگاه متعلق به دانشگاه است و هر گونه بهره‌برداری ازآن باید با ذکر نام دانشگاه و رعایت آیین‌نامه‌ها و دستورالعمل‌های مصوب دانشگاه باشد.

ماده ۲- انتشار مقاله یا مقالات مستخرج از پایان نامه/ رساله به صورت چاپ در نشریات علمی و یا ارائه در مجامع علمی می‌باید به نام دانشگاه بوده و استاد راهنما نویسنده مسئول مقاله باشند.

تبصره: در مقالاتی که پس از دانش آموختگی بصورت ترکیبی از اطلاعات جدید و نتایج حاصل از پایان‌نامه/ رساله نیز منتشر می‌شود نیز باید نام دانشگاه درج شود.

ماده۳- انتشار کتاب حاصل از نتایج پایان‌نامه/ رساله و تمامی طرحهای تحقیقاتی دانشگاه باید با مجوز کتبی صادره از طریق حوزه پژوهشی دانشگاه و بر اساس آیین‌نامه‌های مصوب انجام می‌شود.

ماده ۴- ثبت اختراع و تدوین دانش فنی و یا ارائه در جشنواره های ملی، منطقه‌ای و بین المللی که حاصل نتایج مستخرج از پایان‌نامه/ رساله و تمامی طرح های تحقیقاتی دانشگاه باید با هماهنگی استاد راهنما یا مجری طرح از طریق حوزه پژوهشی دانشگاه انجام گیرد.

ماده ۵- این دستورالعمل در ۵ ماده و یک تبصره در تاریخ ۱۳۸۴/۴/۲۵ در شورای پژوهشی دانشگاه به تصویب رسید ه و از تاریخ تصویب لازم الاجرا است و هر گونه تخلف از مفاد این دستورالعمل، از طریق مراجع قانونی قابل پیگیری خواهد بود.

نام و نام خانوادگی: ولی درهمی

تاریخ و امضاء: ۸۶/۱۰/۱۵

آیین نامه چاپ پایان نامه‌های دانشجویان دانشگاه تربیت مدرس

نظر به اینکه چاپ و انتشار پایان نامه‌های تحصیلی دانشجویان دانشگاه تربیت مدرس، مبین بخشی از فعالیتهای علمی- پژوهشی دانشگاه است بنابر این بمنظور آگاهی و رعایت حقوق دانشگاه، دانش آموختگان این دانشگاه نسبت به رعایت موارد ذیل متعهد می‌شوند:

- **ماده ۱:** در صورت اقدام به چاپ پایان نامه‌های خود ، مراتب را قبلا بطور کتبی به «دفتر نشرآثار علمی» دانشگاه اطلاع دهد.
- **ماده ۲:** در صفحه سوم کتاب (پس از برگ شناسنامه)، عبارت ذیل را چاپ کند:

« کتاب حاضر، حاصل رساله دکتری نگارنده در رشته **مهندسی برق - کنترل** است که در سال ۱۳۸٦ در دانشکده فنی و مهندسی دانشگاه تربیت مدرس به راهنمایی **جناب آقای دکتر وحید جوهر مجد** و مشاوره **جناب آقایان دکتر مجید نیلی احمد آبادی و دکتر حمید رضا مومنی** از آن دفاع شده است. »

- **ماده ۳:** به منظور جبران بخشی از هزینه های انتشارات دانشگاه، تعداد یک درصد شمارگان کتاب (در هر نوبت چاپ) را به « دفتر نشر آثار علمی » دانشگاه اهدا کند. دانشگاه می تواند مازاد نیاز خود را به نفع مرکز نشر در معرض فروش قرار دهد.
- **ماده ٤:** در صورت عدم رعایت ماده ۳، ۵۰٪ بهای شمارگان چاپ شده را به عنوان خسارت به دانشگاه تربیت مدرس، تادیه کند.
- **ماده ۵:** دانشجو تعهد و قبول می کند در صورت خودداری از پرداخت بهای خسارت، دانشگاه می‌تواند خسارت مذکور را از طریق مراجع قضایی مطالبه و وصول کند؛ به علاوه به دانشگاه حق می دهد به منظور استیفای حقوق خود، از طریق دادگاه، معادل وجه مذکور در ماده ٤ را از محل توقیف کتابهای عرضه شده نگارنده برای فروش، تامین نماید.
- **ماده ٦:** اینجانب **ولی درهمی** دانشجوی رشته **مهندسی برق-کنترل** مقطع **دکتری** تعهدات فوق و ضمانت اجرایی آن را قبول کرده، به آن ملتزم می شوم.

نام و نام خانوادگی: ولی درهمی

تاریخ و امضا: ۸۶/۱۰/۱۶

این رساله را تقدیم می‌کنم به

همسر مهربان و فداکارم

و دو فرزند عزیزم

سارا و صدار

تقدیر و تشکّر

خداوند را سپاس می‌گویم بخاطر لطف و مهربانی هایش و اینکه به من لذت توفیق کسب علم را عطا کرده است و امیدوارم بتوانم تکلیف خود را در برابر آنچه به من ارزانی داشته است بخوبی بانجام برسانم.

از استاد راهنمای محترم جناب آقای دکتر وحید مجد، بخاطر راهنماییهای مفیدی که نموده اند، و همچنین صبر و حوصله ای که در اصلاح مقالات و رساله کشیده اند تشکر و سپاسگزاری می‌نمایم.

از استاد مشاور محترم جناب آقای دکتر مجید نیلی احمدآبادی، بخاطر دلسوزی ها و راهنماییهای ارزنده ای که نموده اند و وقت زیادی که بسیار بیشتر از آنچه از استاد مشاور انتظار می‌رود برای مشاوره اختصاص دادند، تشکر و قدردانی می‌نمایم. همچنین از جناب آقای دکتر حمید رضا مومنی، استاد مشاور دوم این رساله سپاسگزار هستم.

از زحمات دوست عزیزم آقای دکتر امیر رضا عطارها که در تهیه مقالات انگلیسی کمک بسیار نمودند، قدردانی فراوان می‌نمایم.

در پایان ممنون از صبر، فداکاری، مهربانی، و دلسوزی همسرم هستم که اگر حمایت های او نبود هرگز موفق به پایان رساندن این مسیر که همراه با فراز و نشیب های بسیاری برای من و خانواده ام بود، نمی شدم.

چکیده

عدم وجود تحلیل ریاضی و عدم کاوش مناسب دو نقطه ضعف عمده در اکثر الگوریتم‌های آموزش تقویتی پیوسته هستند. تحقیقات این رساله بر دو موضوع فوق متمرکز است. یک الگوریتم جدید "آموزش تقویتی فازی" بر مبنای معماری نقاد-تنها، که "آموزش سارسای فازی" (FSL) نامیده می‌شود، ارائه می‌گردد. الگوریتم FSL پارامترهای تالی قواعد سیستم فازی را بصورت روی خط تنظیم می‌کند. انتخاب عمل در هر قاعده بر طبق فرمول جدید ارائه شده در این رساله بنام "بیشینه نَرم بهبود یافته" انجام می‌شود. این نحوه انتخاب عمل موجب می‌گردد تا توزیع احتمال انتخاب عمل نهایی در الگوریتم FSL از توزیع بولتزمن که یک توزیع پیوسته است، تبعیت کند. وجود نقاط ایستای منطبق بر نقاط ثابت الگوریتم "تکرار تقریب ارزش عمل" (ارائه شده در این رساله) برای FSL اثبات می‌شود. همچنین همگرایی بردار وزن FSL به مقدار یکتا وقتی که از سیاست انتخاب عمل ایستا استفاده شود اثبات می‌گردد. نتایج شبیه سازی در بکارگیری الگوریتم FSL در مقابل الگوریتم "آموزش Q فازی" (FQL)، حاکی از کیفیت و سرعت بالاتر آموزش FSL نسبت به FQL است. همچنین همگرایی FSL در برابر واگرایی FQL در یک مسأله چند هدفه جاییکه هر دو الگوریتم از سیاست ایستا استفاده می‌نمایند، نشان داده می‌شود. سپس موضوع تعادل میان "کاوش برای کسب تجربه جدید" و "بهره‌برداری از تجربیّات گذشته" در آموزش تقویتی فازی پیوسته مورد بررسی قرار می‌گیرد. کیفیت برقراری تعادل وابستگی شدیدی به دقت تابع ارزشِ عمل تقریب زده شده دارد. بنابراین، در ابتدا تفاوت‌های بین تخمین تابع ارزش در الگوریتم‌های آموزش تقویتی گسسته با تقریب آن در آموزش تقویتی پیوسته‌ایِ که از تقریب زننده استفاده می‌کند شرح داده شده و مشکل "فوق برازش" معرفی می‌گردد. برای رفع این مشکل یک روش جدید محاسبه نرخ آموزش ارائه می‌شود. مقدار نرخ آموزش پیشنهادی رابطه معکوس با مقدار ملاقاتِ فازی حالت جاری دارد و باعث تاثیر تقریباً یکسان داده‌های آموزشی روی پارامترهای وزن تقریب زننده و جلوگیری از فوق برازش می‌گردد. برای ایجاد تعادل بین کاوش و بهره‌برداری، یک سیستم فازی T-S مرتبه صفر که مقدار ضریب دما را برای فرمول بیشینه نَرم تولید می‌کند، ارائه می‌گردد. با اضافه نمودن دو تکنیک "نرخ آموزش تطبیقی" و "تعادل دهنده فازیِ کاوش و بهره‌برداری" به الگوریتم FSL، الگوریتم حاصل را FSL بهبود یافته یا EFSL می‌نامیم. نتایج شبیه‌سازی حاکی از برتری قابل توجه در کیفیت و زمان آموزشEFSL در مقابل FSL است. در ادامه راهکارهایی برای چالشهای موجود در بکارگیری آموزش تقویتی در مسأله ناوبری ربات ارائه می‌شود و ساختار یک کنترلگر فازی برای آن تعیین می‌گردد. در انتها عملکرد الگوریتم‌های ارائه شده در تنظیم پارامترهای کنترلگر در مسأله مذکور با یکدیگر مقایسه می‌شود.

کلید واژه ها: آموزش تقویتی، کاوش، بهره‌برداری، سیستم‌های فازی، ناوبری ربات

فهرست مطالب

عنوان — صفحه

فهرست شکل‌ها

عنوان **صفحه**

فهرست جدول‌ها

صفحه

عنوان

فصل ۱– مقدمه

۱-۱- پیشگفتار

طراحی رویِ خط[1] کنترلگر بهینه برای سیستم‌های پیچیده در محیط‌هایی که دارای عدم قطعیت[2] و نایقینی[3] هستند، اگر غیر ممکن نباشد، کار پیچیده‌ای است. بخصوص برای سیستمهایی (مانند ناوبری ربات[4] ها) که می‌خواهند بطور هوشمند با محیط تعامل داشته باشند. با توجه به اینکه عموماً اطلاعات زیادی در خصوص خروجی مطلوب کنترلگر در این نوع سیستم‌ها وجود ندارد، استفاده از از روشهای آموزش بدون ناظر برای طراحی و تنظیم پارامترهای کنترلگر ارجحیت دارد [۱-۳].

آموزش تقویتی (RL)[5]، یک روش قوی مدرن برای آموزش رویِ خط استراتژی‌های کنترل از طریق تعامل با محیط است. این روش تنها با استفاده از یک معیار اسکالر راندمان، که سیگنال تقویت یا پاداش[6] نامیده می‌شود، بدون نیاز به سرپرست قادر به آموزش عاملها در محیطهای پیچیده، ناقطعی و تصادفی می‌باشد [۴-۶]. قابلیتهای مذکور باضافه قدرت کاوش بالای آموزش تقویتی در جهت یافتن پاسخ بهینه، منجر به استفاده روز افزون آن در وظایف کنترلی پیچیده همچون کنترل حرکت ربات‌ها، و ناوبری ربات شده است.

سازماندهی الگوریتم‌های آموزش تقویتی بر مبنای تخمین ارزشِ[7] حالت (یا جفت حالت- عمل) می‌باشد [۶، ۷]. در آموزش تقویتی گسسته مقدار ارزشِ حالت (یا جفت حالت- عمل) در جدول ارزش ذخیره شده و در هر قدم که آن حالت (یا جفت حالت- عمل) ملاقات شوند به روز رسانی انجام می‌گیرد [۶]. از آنجا که تعداد پارامترهای قابل تنظیم در آموزش تقویتی گسسته، رابطه مستقیمی با عدد اصلی[8] فضای متغیرهای حالت و عمل مسأله دارد، در مسائل کنترل با فضای حالت- عمل بزرگ و یا پیوسته مانند ناوبری ربات، مشکل تنگنای ابعاد[9] وجود دارد. بنابراین استفاده از تقریب زننده‌های تابع برای تقریب تابع ارزش[10] در اینگونه مسائل ضروری است [۸، ۹].

بر این اساس محققین با ترکیب الگوریتم‌های آموزش تقویتی گسسته با تقریب زننده‌هایی همچون شبکه‌های عصبی و منطق فازی، الگوریتم‌های آموزش تقویتی پیوسته را ارائه داده‌اند. کاربردهای گسترده و عملکرد مطلوب سیستم استنتاج فازی در کنترل و مسائل پیچیده و نیز مزایایی چون امکان گنجاندن دانش بشری، ارائه دانش بصورت قواعد اگر- آنگاه، و قابلیت مدلسازی و کنترل سیستمهای غیر خطی با

---

[1] Online
[2] Nondeterministic
[3] Uncertainty
[4] Robot navigation
[5] Reinforcement Learning
[6] Reward
[7] Value
[8] Cardinality
[9] Cures of dimentsionality
[10] Value function

دقت دلخواه [۱۰-۱۲]، باعث شده است تا محققین [۱۳-۲۶] با ترکیب سسیتم‌های فازی بعنوان تقریب زننده با روشهای آموزش تقویتی، الگوریتمهای آموزش تقویتی فازی (FRL)[1] را ارائه دهند. تمرکز ما نیز در این رساله بر روی الگوریتمهای FRL خواهد بود.

توجه شود که در آموزش تقویتی گسسته مقادیر ارزش حالت (یا جفت حالت- عمل) از هم مستقل می‌باشند، در حالیکه در حالت پیوسته به خاطر استفاده از تقریب زننده، مقدار ارزش در هر قدم زمانی بر مبنای پارامترهای تقریب زننده حاصل می‌شود، لذا به روز رسانی پارامترها در هر قدم زمانی بر روی مقادیر ارزش در کل فضا تاثیر می‌گذارد.

۱-۲- دو موضوع مهم در آموزش تقویتی پیوسته

دو موضوع در بکارگیری الگوریتم‌های FRL (یا بطور کلی الگوریتمهای آموزش تقویتی پیوسته) در مسائل کنترلی مورد توجه ویژه هستند .[۲۷]:

۱-۲-۱- تحلیل ریاضی الگوریتم

بر خلاف نتایج و قضایایی که در خصوص همگرایی مقادیر ارزش برای الگوریتم‌های آموزش تقویتی گسسته ارائه شده است [۶،۲۸-۳۰]، اکثر روشهای آموزش تقویتی پیوسته موجود، فاقد تحلیل ریاضی می‌باشند. بنابراین تحلیل ریاضی الگوریتمهای آموزش تقویتی پیوسته، یک موضوع مورد علاقه برای محققین می‌باشد.

۱-۲-۲- تعادل بین کاوش و بهره‌برداری از تجربیّات

در حین آموزش تقویتی دو هدف مخالف باید با هم ترکیب شوند: از یک طرف کلیه جفت‌های حالت– عمل مسأله باید به حد کافی برای بدست آوردن دانش مورد کاوش[2] قرار گیرند و از طرف دیگر تجربه‌های بدست آمده باید در انتخاب عمل بکار گرفته شوند. از نقطه نظر کاوش، مهمترین سؤال این است که چگونه می‌توان زمان یادگیری را حداقل نمود. متناظراً از نقطه نظر بهره‌برداری از تجربیّات[3] گذشته سؤال این است که چگونه می‌توان جریمه‌ها را کاهش داد [۳۱]. برای آموزش مؤثّر، عملها باید بگونه‌ای انتخاب شوند که محیط بطور مناسب آزموده شده و از جریمه‌ها نیز حتی الامکان اجتناب گردد. انجام کامل این دو کار بطور همزمان ممکن نیست. در ادامه این رساله برای مختصر نویسی بجای عبارت "تعادل[4] بین کاوش و بهره‌برداری از تجربیّات" از عبارت "تعادل" استفاده می‌کنیم. برقراری تعادل مناسب، ضمن کاهش زمان آموزش، امکان گریز عامل از نقاط اکسترمم محلی را فراهم نموده و منجر به یافتن جواب مطلوب می‌گردد. تعدادی از محققین راهکارهایی برای انتخاب عمل و برقراری تعادل مناسب در آمورش تقویتی گسسته ارائه داده اند، لیکن در خصوص تعادل در آموزش تقویتی پیوسته، نتایج تحقیق جامعی ارائه

[1] Fuzzy Reinforcement Learning
[2] Exploration
[3] Exploitation
[4] Balance

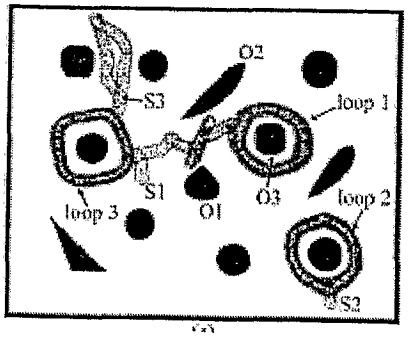
نشده است. بطور کلی مقوله تعادل بین کاوش و بهره‌برداری از تجربیّات یک موضوع باز و جدید در آموزش تقویتی پیوسته است که دومین موضوعی است که در این رساله بر روی آن تحقیق می‌شود.

۱-۳- مروری بر مطالعات انجام شده

دو معماری معروف استفاده شده در FRL ها، روش عملگر- نقاد[1] [۳۲،۱۰] و نقاد– تنها[2] [۳۳،۶] هستند. در اینجا، ابتدا دو معماری مذکور شرح داده شده و سپس مطالعات انجام شده در خصوص تحلیل ریاضی و تعادل مرور می‌گردد.

۱-۳-۱- معماری عملگر– نقاد

آموزش عملگر- نقاد فازی(FACL)[3] دارای دو بخش نقاد و عملگر می‌باشد. بخش نقاد برای تقریب تابع ارزشِ حالت و بخش عملگر برای تولید عمل استفاده می‌گردد. در واقع از بخش نقاد برای تخمین خطای ارزش (خطای تفاضل موقتی (TD)[4]) و بکارگیری این خطا برای تنظیم پارامترهای بخش عملگر استفاده میشود [۱۴]. علیرغم بکارگیری الگوریتمهای FACL در بسیاری از مسائل، هیچ تحلیل ریاضی و یا اثبات همگرایی برای الگوریتمهای FACL موجود که تنها از دو بخش عملگر و نقاد استفاده می‌کنند، موجود نیست. ضعف دیگر ساختار فوق فقدان کاوش مناسب است. در واقع انتخاب عمل در الگوریتم مذکور بر مبنای سیاست حریصانه[5] می‌باشد و امکان ارزیابی عملهای دیگر در هر حالت را نمی‌دهد. شکل ۱-۱ نمونه‌ای از به دام افتادن ربات را در حلقه‌ی بین موانع حین عبور از آنها در هنگام آموزش با الگوریتم FACL نشان می‌دهد [۲۰].



شکل ۱-۱: نمونه ای از به دام افتادن ربات در بین موانع [۲۰].

[1] Actor-critic
[2] Critic-only
[3] Fuzzy Actor Critic Learning
[4] Temporal Diffference
[5] Greedy

دلیل رخداد مشکل مذکور، فقدان کاوش در الگوریتم‌های FACL می‌باشد. تعدادی از محققین [۳۴،۱۹] برای رفع مشکل فوق، خروجی نهایی را با یک مقدار تصادفی برمبنای تابع توزیع چگالی احتمال گوسی با میانگین صفر جمع نموده‌اند. علیرغم بهبودهای صورت گرفته به دلیل لحاظ کردن کاوش در کنترلگر، تعادل مناسبی بین کاوش و بهره‌برداری از تجربیّات حاصل نشده و نقاط ضعفی چون توزیع احتمال یکنواخت خروجی حول عمل با بالاترین ارزش، و همچنین فقدان ارتباط میان احتمال انتخاب عمل‌ها با مقدارهای ارزش عمل[1] همچنان پا برجاست.

در [۵] یک الگوریتم FRL جدید با ساختاری شبیه عملگر- نقاد ارائه گردیده و همگرایی روش اثبات شده است. در این الگوریتم بخش نقاد به تقریب مقدار ارزش عمل بجای مقدار ارزش حالت می‌پردازد و خروجی بخش عملگر احتمال انتخاب هر عمل می‌باشد. همچنین بخش دیگری نیز برای تقریب مقدار ارزش در نظر گرفته شده است. گرچه این الگوریتم نسبت به روشهای دیگر FACL دارای تحلیل ریاضی و درجه کاوش بالاتری نسبت به دیگر FACL ها است، لیکن بالاسری محاسباتی بالای آن منجر به عدم گرایش محققین به استفاده از آن شده است. ضمناً به مقوله تعادل بین کاوش و استفاده از تجربیّات پرداخته نشده است.

۱-۳-۲- معماری نقاد-تنها

در مقابلِ معماری عملگر- نقاد، روش نقاد- تنها فقط دارای، یک بخش نقاد است که برای تقریب تابع ارزش عمل استفاده می‌شود و عمل نهایی با توجه به مقادیر ارزش تقریب زده شده تولید می‌گردد. این نحوه انتخاب عمل ضمن فراهم نمودن امکان درجه کاوش بالا، پتانسیل لازم برای برقراری تعادل بین کاوش و بهره‌برداری از تجربیّات را دارا می‌باشد. ضمناً ساختار فوق امکان گنجاندن دانش بشری را براحتی ممکن می‌سازد.

در [۲۲-۲۶]، معماری نقاد-تنها با ترکیب یک سیستم فازی با روش آموزش Q[2] (که یک روش "مستقل از سیاست"[3] است) پیاده سازی شده است. یک نسخه کامل از الگوریتم مذکور با نام آموزش Q فازی (FQL)[4] در [۲۳] معرفی گردید. در الگوریتم FQL تالی قواعد یک سیستم فازی T-S[5] مرتبه صفر، بصورت روی خط تنظیم می‌شوند. این الگوریتم در مسائل زیادی از جمله رباتها بکار گرفته شده است[۲۲-۲۶]، اما روش مذکور بر مبنای تجربه بوده و دارای دو ضعف عمده‌ی فقدان تحلیل ریاضی و امکان واگرایی است. مثالها و تحلیل‌هایی از واگراییِ الگوریتم FQL و نیز واگرایی الگوریتم‌هایی که از ترکیب آموزش Q با "تقریب زننده‌های تابعی خطی"[6] (تقریب زننده‌های تابع بصورت ترکیبهای خطی از

[1] Action value
[2] Q-Learning
[3] Off-policy
[4] Fuzzy Q-Learning
[5] Takagi-Sugeno
[6] Linear function approximators

یک سری توابع پایه[۹]) استفاده می‌کنند در [۳۵،۸] آمده است. در واقع گرچه قابلیتِ مستقل از سیاست آموزش Q خوشایند است، لیکن تازه سازی مقدار ارزش جفت حالت- عمل بر طبق یک توزیع متفاوت از دینامیک زنجیره مارکف باعث ناپایداری الگوریتم در حالت پیوسته می‌گردد [۳۶]. بعلاوه الگوریتم‌های آموزش Q ترکیب شده با تقریب زننده‌های تابع، ویژگی مستقل از سیاست را از آموزش Q گسسته، به ارث نبرده و مقدار نهایی ارزش جفت حالت- عمل تقریب زده شده به سیاست استفاده شده وابستگی شدید دارد.

.بر خلاف نتایج تحلیلی منفی و مثالهایی از واگرایی که برای الگوریتم‌های آموزش تقویتی پیوسته که از تعمیم روش آموزش Q حاصل شده‌اند، موجود است، تعدادی از نتایج تحلیلی مثبت در خصوص ترکیب روشهای آموزش تقویتی "وابسته به سیاست"[1] با تقریب زننده های تابع وجود دارد [۲۷،۳۷-۳۹]. وجود نقاط ایستا[2] در روش تفاضل موقتی خطی که از سیاست انتخاب عمل بیشینه نَرم[3] بهره می‌برد، در [۳۷] اثبات شد. البته این اثبات تنها مربوط به تقریب تابع ارزش حالت است نه تابع ارزش عمل. بعلاوه، با توجه به آنکه در فرمول بیشینه نَرم مقدار ارزش عمل مورد نیاز است، در مقاله مذکور مقدار ارزش عمل در هر قدم زمانی، از مقدار تابع ارزش با فرض دانستن مدل محیط محاسبه شده است، در حالیکه در مسائل آموزش تقویتی معمولاً مدل محیط، ناشناخته فرض می‌گردد.

در مقالات [۳۸،۳۹] از ترکیب روش سارسا[4] (که یک روش وابسته به سیاست است) با تقریب زننده های تابعی خطی (ما این ترکیب را سارسای خطی می‌نامیم) برای تقریب تابع ارزش عمل استفاده شده است. در [۳۸] همگرایی پارامترهای وزن سارسای خطی به یک ناحیه در صورت استفاده از یک سیاست ایستا[5] در همه رویدادها[6]، اثبات شده است. در [۳۹] همگرایی الگوریتم ارائه شده به پاسخ یکتا به شرط ایستا بودن سیاست ها در هر اپیسود اثبات گردیده است. الگوریتم مذکور دارای دو ضعف عمده می‌باشد: ۱- نتایج قضیه، تضمینی در خصوص کیفیت سیاست نهایی که الگوریتم به آن همگرا می‌شود ارائه نمی‌دهد، ۲- با توجه به آنکه پس از تعیین هر سیاست جدید باید آموزش تا همگرایی وزنها ادامه یابد، سرعت آموزش کند می‌باشد. در حالیکه در مسائل کنترل با آموزش رویِ خط، مطلوب آن است که در هر قدم زمانی سیاست انتخاب عمل به روز رسانی گردد. توجه شود که هیچگونه پیاده سازی و شبیه سازی برای سه الگوریتم اشاره شده در بالا [۳۷-۳۹] ارائه نشده و در واقع دو چالش عمده برای پیاده سازی این الگوریتم‌ها وجود دارد: انتخاب تابع تقریب زننده تابعی خطی و نحوه انتخاب عمل، به نحوی که این دو بتوانند شرایط بیان شده در لم‌ها وقضایای بیان شده در مقالات مذکور را ارضاء نمایند.

در مجموع، بر خلاف روشهای آموزش تقویتی گسسته، هیچ قضیه‌ای در خصوص همگرایی آموزش تقویتی (فازی) بر مبنای معماری نقاد-تنها وقتی که سیاست انتخاب عمل ایستا نباشد و یا هنگامی که

---

[1] On-policy
[2] Stationary points
[3] Softamx
[4] Sarsa
[5] Stationary
[6] Episodes

سیاست بر مبنای مقدار ارزش عمل در هر قدم زمانی به روز رسانی شود، وجود ندارد. لذا این موضوع یکی از مواردی است که در این رساله بر روی آن تحقیق می‌گردد.

در ادامه به روشهای انتخاب عمل بکار رفته در الگوریتم‌های اشاره شده در بالا پرداخته می‌شود. الگوریتم‌های FQL ارائه شده در[۲۲-۲۵] از ایده‌های موجود انتخاب عمل و تعادل در حالت گسسته [۶، ۳۱] برای انتخاب یک عمل در هر قاعده فازی استفاده کرده‌اند. همچنین الگوریتم FQL ارائه شده در [۲۳]، برای انتخاب عملها در هر قاعده از ترکیب فرمول بیشینه نَرم با روش "نرخ انجام عمل" [۳۱] بهره برده است. در واقع روشهای انتخاب عمل در مقالات مذکور [۲۲-۲۶] یک گسترش روشهای انتخاب عمل در حالت گسسته بدون توجه به ویژگی‌های تابع ارزش عمل در حالت پیوسته می‌باشد.

در آموزش تقویتی گسسته، تعداد حالت‌ها و عمل‌ها قابل شمارش و محدود می‌باشند. مقادیر حالت ها و (یا جفت‌های حالت-عمل) در یک جدول ارزش ذخیره شده و این مقادیر بطور مستقل به روز رسانی می‌گردند. در حالیکه در آموزش تقویتی گسسته، تعداد حالت‌ها و عمل‌ها نا محدود و غیر قابل شمارش می‌باشند. با توجه به استفاده از تقریب زننده‌ها برای تقریب تابع ارزش، به روز رسانی پارامترهای تقریب زننده در هر قدم زمانی می‌تواند باعث تغییر مقدار ارزش در کل فضای مسأله گردد. این تفاوتها باعث می‌شود که ایده‌های ارائه شده برای حالت گسسته چندان در حالت پیوسته کارآ نباشند.

نکته دیگر آنکه نحوه انتخاب عمل در مقالات فوق [۲۲-۲۶] هیچ تضمینی برای اینکه تابع توزیع احتمال انتخاب عمل نهایی از یک تابع پیوسته تبعیت کند را ارائه نمی‌دهد. در حالیکه یک نکته مهم در آموزش تقویتی پیوسته، تولید عمل نهایی بر طبق یک توزیع احتمال پیوسته است به گونه‌ای که تغییرات کوچک در مقدار ارزش عمل باعث تغییرات شدید در رفتار عامل نشود [۳۹].

یک مشکل دیگر در RL پیوسته، خطای تقریب تابع ارزش است، بعضاً مقدار این خطا بقدری بزرگ می‌شود که انتخاب عمل بر مبنای آن عملکرد بسیار ضعیفی را باعث می‌گردد. دلیل این خطای زیاد، غالب شدن اثر داده های آموزشی ناحیه و یا نواحیی می‌باشد که عامل نسبت به نواحی دیگر بسیار بیشتر ملاقات نموده است. این مشکل را ما "فوق برازش"[1] می‌نامیم و برای اولین بار در این رساله در فصلهای بعد بطور کامل معرفی می‌گردد. برای رفع این معضل ما یک نرخ آموزش تطبیقی را معرفی می‌نماییم. لذا در اینجا اشاره ای هم به نرخ های آموزش استفاده شده در FRL ها می‌شود.

بعنوان یک نکته عملی در اکثر مسائل RL، نرخ آموزش با تابعی از زمان کاهش می‌یابد [۶، ۲۲، ۲۴، ۲۵]. از آنجا که نرخ آموزش مذکور تابعی از زمان است، مقدار آن در هنگام به روز رسانی برای پارامترهای قابل تنظیم قواعدی که در شروع آموزش تحریک نمی گردند، بقدری کوچک خواهد بود که تغییر این پارامترها محسوس نمی‌باشد. برای رفع مشکل مذکور، در [۴۰] یک نرخ جدید آموزش پیشنهاد گردیده است. در این روش برای پارامترهای قابل تنظیم هر قاعده مستقل از قواعد دیگر یک نرخ آموزش تعیین می‌شود. مقدار نرخ آموزش برای هر قاعده با توجه به مجموع شدت آتش های گذشته آن قاعده محاسبه

[1] Overfitting