بسم الله الرحمن الرحیم

یادگیری و تعادل

کیانا هوشانفر
مریم هارونی

رشته برق ـ کنترل
تیرماه ۱۴۰۳

مقاله

تمرکز مقاله:

- مرور ادبیات اخیر در زمینه یادگیری غیرتعادلی در بازی‌ها.
- محدودیت‌های موجود به دلیل محدودیت فضا.
- تأکید بر مدل‌های یادگیری فردی.
- تمرکز کمتر بر مدل‌های تکاملی و مدل‌های تنظیم کوتاه‌مدت.

مقدمه

عقلانیت و تعادل

- عقلانیت ⟵ تعادل نش
- آگاهی مشترک از عقلانیت بین بازیکنان ⟵ تعادل نش
- هماهنگی در یک تعادل واحد

یادگیری از تجربه

- بازی مکرر ⟵ نتایجی نزدیک به پیش‌بینی‌های تعادلی
- ایده ایجاد تعادل از طریق یادگیری

نظریه یادگیری در بازی‌ها

- فرآیند ایجاد تعادل از طریق یادگیری، تطبیق و/یا تقلید بلندمدت غیرتعادلی را رسمی می‌کند.

مقدمه

اهداف عاملان

تعادل نش ✗

حداکثر کردن سود خود ✓

سوال اصلی

چه زمانی یادگیری به رفتار تعادلی منجر می‌شود؟

قابلیت قبول قوانین یادگیری

- نبود جایگزین‌های ساده و واضح بهتر.
- شرایط واقع‌گرایانه برای یادگیری.

مقدمه

مدل‌های خاص مورد بحث

تنظیمات ساده یادگیری

مشاهده استراتژی‌های عاملان در انتهای هر دور

مواجه تصادفی عاملان با حریفان ناشناس

یادگیری در بازی‌های فرم گسترده

مشاهده استراتژی‌های حریفان ✗ فقط توالی اقدامات ✓

بررسی محدودیت‌های یادگیری

مقدمه

بخش 2A بازی خیالی و بازی خیالی تصادفی

مدل‌ها نسبتا ساده + تفکرعوامل بیزی + تفکرمحیطی ثابت

محیط ثابت یا تقریبا ثابت ← عملکرد خوب

سادگی این مدل‌ها ← تحلیل با تکنیک‌های تقریب تصادفی

بخش 2B عملکرد نامتقارن

همسانی هانان ← عملکرد خوب در محیط‌های ثابت

شرایط قوی‌تر ← تضمین عملکرد خوب در محیط‌های عمومی‌تر

تحت کالیبراسیون ← همگرایی بازی به مجموعه‌ای از تعادل‌های همبسته **مدل‌های**

بخش 2C یادگیری پیچیده‌تر

مدل‌هایی که بازیکنان به‌طور فرضی ماتریس پرداخت را نمی‌دانند، شامل مدل‌های یادگیری تقویتی و مدل‌های تقلید.

تفسیر بازی خیالی تصادفی به عنوان یادگیری تقویتی.

بازی خیالی

عاملان به‌طور مکرر یک بازی استراتژیک ثابت را بازی می‌کنند.

فرضیات

عاملان فضای استراتژی و تابع سود خود را می‌دانند.
عاملان استراتژی حریف خود را در هر دوره مشاهده می‌کنند.
عاملان فرض می‌کنند که توزیع استراتژی‌های حریفان ثابت اما ناشناخته است.

نزدیک‌بینی استراتژیک

فودنبرگ و کرپس (۱۹۹۳) ⬅ مدل "جمعیت بزرگ"

- هر دوره همه عاملان برای بازی روبهرو می‌شوند و فقط بازی مقابل خود را می‌بینند.
- احتمال مواجهه عوامل با همان حریفدر دوره های مکرر کم است.

این تنظیم به عوامل اجازه می دهد تا بیشتر بر بازده های فوری تمرکز کنند تا بر تعاملات آینده با حریفان خاص.

بازی خیالی

FP ⟸ رویکرد بیزی به روز رسانی باورهای خود درباره استراتژی های حریفان

$$k_i^t(s^{-i}) = k_i^t(s^{-i}) + \begin{cases} 1 & if \quad s^{-i} = s_{t-1}^{-i} \\ 0 & if \quad s^{-i} \neq s_{t-1}^{-i} \end{cases}$$

$$\gamma_i^t(s^{-i}) = \frac{k_i^t(s^{-i})}{\sum_{\tilde{s}^{-i} \in s^{-i}} k_i^t(\tilde{s}^{-i})}$$

براساس این باورهای به روز شده بهترین اقدامات را برای به حداکثر رساندن بازده های مورد انتظار

بازی خیالی

مثالی از فودنبرگ و کرپس برای نشان دادن محدودیت های احتمالی FP

i. وزن اولیه $(1,\sqrt{2})$

ii. تخمین هردو از بازی رقیب B

iii. هردو A بازی میکنند

iv. وزن به روز شده $(2,\sqrt{2})$

v. هردو B بازی میکنند

vi. .....

	A	B
A	0,0	1,1
B	1,1	0,0

همگرایی فرکانس های تجربی به تعادل نش مختلط ⬅ بازده صفر به دلیل وجود چرخه

بازی خیالی تصادفی

مقدمه

بخش ۲

بخش ۳

نتیجه‌گیری

FP

بازیکنان به طور مکرر بازی را انجام می دهند و استراتژی های خود را بر اساس رفتار مشاهده شده حریفان تنظیم می کنند.

SFP

بازیکنان اقدامات را بر اساس یک تابع بهترین پاسخ تصادفی انتخاب می‌کنند.

تصادفی بودن : نوسانات غیرقابل پیش بینی در پاداش‌ها را منعکس می‌کند.

تصادفی بودن در تصمیم گیری ⬅ SFP مدلی واقع بینانه‌تر از رفتار استراتژیک در بازی ها

⬅ همگرایی نرم تر به تعادل ها و ارائه سازگاری در پرداخت های بلندمدت

بازی خیالی تصادفی

توزیع بهترین پاسخ

هر بازیکن i شوک های تصادفی η^i

$$\overline{BR}^i(\sigma^{-i})(s^i) = Prob[\eta^i \; s.t. \; s^i \; is \; a \; best \; response \; to \; \sigma^{-i}]$$

یک مثال خاص از بهترین پاسخ تصادفی = بهترین پاسخ تصادفی لجستیک

احتمال انتخاب s^i

$$\overline{BR}^i(\sigma^{-i})(s^i) = \frac{\exp(\beta u(s^i,\sigma^{-i}))}{\sum_{\acute{s}^i} \exp(\beta u(\acute{s}^i,\sigma^{-i}))}$$

تقاطع توابع رو توزیع نش نامیدند.

با افزایش β به سمت بی‌نهایت، توزیع‌های نش به تعادل نش بازی با اطلاعات کامل همگرا می‌شوند.

بازی خیالی تصادفی

مزایای SFP

- همگرایی بهتر
- مقاومت در برابر نویز
- سازگاری جهانی

تابع بهترین پاسخ صاف

یک بهترین پاسخ صاف عمومی می تواند از حداکثرسازی یک تابع پرداخت تعیینی مختل شده که اقدامات خالص را جریمه می کند باشد:

$$\arg\max_{\sigma^i} u^i(\sigma^i, \sigma^{-i}) + \beta^{-1} v^i(\sigma^i)$$

$$v^i(\sigma^i) = \sum_{s^i} -\sigma^i(s^i) \log\sigma^i(s^i)$$

بازی خیالی تصادفی

بررسی سیستم‌هایی از عوامل که همگی از بازی ساختگی تصادفی یا SFP استفاده می‌کنند.

استفاده از روش‌های "تقریب تصادفی" اعمال می‌شود.

نسخه Benaim برای SFP :

$$x_{n+1} - x_n = \frac{1}{n+1}(F(x_n) + U_n + b_n)$$
$$F(x(t)) = \mathrm{dx(t)/dt}$$

تقریب تصادفی برای بازی دو نفره SFP :

$$\theta_{1,n+1} - \theta_{1,n} = (1/(n+1))[\overline{B}\overline{R}_1(\theta_{2,n}) - \theta_{1,n} + U_{1,n} + b_{1,n}]$$
$$\theta_{2,n+1} - \theta_{2,n} = (1/(n+1))[\overline{B}\overline{R}_2(\theta_{1,n}) - \theta_{2,n} + U_{2,n} + b_{2,n}]$$

بازی خیالی تصادفی

آن‌ها سپس از تقریب تصادفی برای ارتباط دادن رفتار حدی سیستم با رفتار سیستم قطعی زیر استفاده کردند:

$$\dot{\theta}_1 = BR_1(\theta_2) - \theta_1, \dot{\theta}_2 = BR_2(\theta_1) - \theta_2$$

توجه داشته باشید که نقاط تعادل این سیستم دقیقاً توزیع‌های تعادلی هستند.
بنابراین تقریب تصادفی به طور تقریبی می‌گوید که SFP نمی‌تواند به یک توزیع نش ناپایدار خطی همگرا شود و باید به یکی از مجموعه‌های داخلی زنجیروار متعدی سیستم همگرا شود.

بازی خیالی تصادفی

Benaim
تکاملی از FP تا "FP تصادفی وزن‌دار" (نسبت هندسی وزن کمتری به مشاهدات قدیمی‌تر)

به طور خلاصه، FPصاف وزن‌دار با وزن‌هایی که به ۱ همگرا می‌شوند، مسیرها و مجموعه‌های حدی مشابه FP تصادفی دارد؛ اما تفاوت در سرعت حرکت است و از این روی در نحوه همگرایی توزیع تجربی متفاوت است.

عملکرد مجانبی و همگرایی جهانی

چگونه SFP می‌تواند در محیط‌های قابل تعویض عملکرد خوبی داشته باشد ← نیاز به قوانین یادگیری پیچیده‌تر ← توسعه قوانین یادگیری بهتر و کارآمدتر

ضمانت

SFP تضمین می کند که بازیکنان حداقل به اندازه بیشینه سازی در برابر میانگین طول زمان بازی خوب عمل می کنند. (به شرط محیط ثابت)

محدودیت

SFP دارای محدودیت در بازی های دارای الگوها، روند یا چرخه

عملکرد مجانبی و همگرایی جهانی

مقدمه

بخش ۲

بخش ۳

نتیجه‌گیری

یادگیری بیزی

به روزرسانی باورهای بازیکنان بر اساس اطلاعات جدید

محدودیت

رویدادهایی با احتمال پیشین صفر

مسئله تعامل

تعامل پیچیده ← استراتژی هایی با پیش فرض غیرممکن ← شرایط بهینه غیر بیزی

سازگاری جهانی

سازگاری جهانی
یک قانون یادگیری جهانی سازگار است اگر در طول زمان به اندازه ای که انگار فرکانس های اقدامات حریفان را از قبل می دانست، عمل کند.

تئوری: هانان و بلکول وجود چنین قوانینی را اثبات کردند. این قوانین تضمین می کنند که میانگین بازده طولانی مدت حداقل به اندازه مقدار مینیمکس است.

مثال: بازی تطابق سکه ها

سازگاری جهانی

قابل دستیابی

مفهوم

بلکول ایده قابل دستیابی را معرفی کرد که شامل ایجاد قوانینی است که دستیابی به مجموعه هدف را در طول زمان تضمین می کنند.

کاربرد

ساخت قوانین یادگیری جهانی ← اطمینان از رسیدن استراتژی‌های بازیکنان به نتایج مطلوب در طول بازی‌های تکراری

سازگاری جهانی

Lambson و Probst

دو بازیکن + طول الگوهای مساوی + همگرایی بازی دقیق یا FP
⇐ همگرایی توزیع تجمعی تجربی بازی به پوش محدب مجموعه تعادل‌های نش

ما انتظار داریم که تشخیص الگوهای طولانی‌تر یک مزیت باشد!

یک مثال جالب!

کالیبراسیون

مقدمه

بخش ۲

بخش ۳

نتیجه‌گیری

Foster & Vohra

قواعد یادگیری مشتق شده از پیش‌بینی‌های کالیبره‌شده

پیش بینی به خوبی کالیبره شده یعنی چه؟!

کالیبراسیون به نظر می‌رسد که ویژگی مطلوبی برای یک پیش‌بین باشد.

کالیبراسیون

سوال مربوط به کالیبراسیون : در تمام مواقعی که یک بازیکن اقدام خاصی را انجام داد، پاسخ آن چقدر خوب بود؟

هارت و ماس-کولل (۲۰۰۰)
سنجش پشیمانی
اگر بدون توجه به بازی حریفان، بازیکن به طور نامحدود کالیبره شده باشد به این معنا که میانگین زمانی پشیمانی برای هر اقدام به صفر برسد، می‌گوییم که بازیکن به طور جهانی کالیبره شده است.
فاستر و ووهرا (۱۹۹۷)
وجود روش‌های یادگیری دارای این ویژگی ⬅ اگر همه بازیکنان از چنین قواعدی پیروی کنند، میانگین زمانی فرکانس بازی باید به مجموعه‌ای از تعادل‌های همبسته بازی همگرا شود.

کالیبراسیون

الگوریتم اولیه فاستر و ووهرا شامل روشی پیچیده برای یافتن ماتریس‌های تصادفی و بردارهای ویژه

سوال:

- آیا فرض اینکه بازیکنان از قواعد جهانی کالیبره شده پیروی می‌کنند، تقریب خوبی است یا نه؟!
- از نظر مفهومی تشخیص اینکه آیا قواعد یادگیری در مسیر بازی کالیبره شده‌اند آسان است؟! اگر چنین باشد، میانگین زمانی بازی مشترک به مجموعه‌ای از تعادل‌های همبسته همگرا می‌شود.

کالیبراسیون

تحقیقات بعدی

- گسترش قابل توجه مجموعه قواعدی به طور جهانی کالیبره شده
- ساده سازی بسیار الگوریتم‌ها و روش‌های

قواعد یادگیری به طور جهانی سازگار ⬅ ساختن قواعد یادگیری به طور جهانی کالیبره شده ⬅ حل یک مسئله نقطه ثابت

این مسئله نقطه ثابت یک مسئله خطی است که با وارونه‌سازی ماتریس حل می‌شود، همانطور که در فودنبرگ و لوین (۱۹۹۸) نشان داده شده است.

کالیبراسیون

آیا وارونه‌سازی یک ماتریس ساده است؟!

تفاوت ظریف بین کالیبره بودن و کالیبره بودن جهانی!

یادگیری کالیبره ⬅ الگوریتم‌های ساده Hart و Mas-Colell (2001)

حل مسئله نقطه ثابت برای کالیبراسیون جهانی $Rq = R^T q$

μ یک عدد بزرگ

$$q(b) = \begin{cases} \left(\frac{1}{\mu}\right) R(a,b) & for\ b \neq a \\ 1 - \sum_{c \neq a} q(c) & for\ b = a \end{cases}$$

کالیبراسیون

Hart & Mas-Colell (2000)

که اگر همه افراد از قوانین مشابهی استفاده کنند، کالیبراسیون دارد.

Cahn (2001)

این قاعده کالیبره است اگر همه افراد از قوانینی استفاده کنند که نرخ تغییر اقداماتشان مشابه باشد. به طور شهودی، اگر بازیکنان دیگر اقدامات خود را به سرعت زیادی تغییر ندهند، روش بالا به طور ضمنی ماتریس لازم برای حل معادله $Rq = RTq$ را معکوس می‌کند.

آزمون

یک تفسیر از مفهوم کالیبراسیون در روند یادگیری، آن است که "یادگیرنده" بتواند به درستی پاسخ‌های خود را با تکرار زمانی بهبود ببخشد، حتی اگر در ابتدا دانش زیادی نداشته باشد.

ساندرونی (۲۰۰۳)
دو ویژگی برای آزمون‌های یادگیری ارائه داده است:

I. آزمون باید پس از یک تعداد محدود دوره‌ها اعلام کند که یادگیرنده از آزمون گذرانده است یا نه.
II. آزمون باید با احتمال بالا واقعیت یادگیرنده را تأیید کند.

آزمون

دکل و فینبرگ (۲۰۰۶) و اولسیفسکی و ساندرونی (۲۰۰۶)
شرطی را که آزمون باید در زمان محدود نتیجه قطعی دهد، رفع کردند.

فورتنو و وهرا (۲۰۰۸)
یک الگوریتم نادانه که برخی آزمون‌ها را پشت سر گذاشته باشد، باید به ضرورت پیچیده‌ای محاسباتی باشد.

آل-نجار و وینشتاین (۲۰۰۷)
تمایز دادن بین دو یادگیرنده اطلاعاتی آسان‌تر است نسبت به ارزیابی یک یادگیرنده در تنهایی.

فینبرگ و استوارت (۲۰۰۷)
احتمال مقایسه بین چندین متخصص مختلف، برخی واقعی و برخی نادرست را در نظر گرفته و نشان دادند که تنها متخصصان واقعی تضمین می‌کنند که آزمون را با چه شرایط دیگری انجام دهند.

همگرایی به نش

مقدمه

بخش ۲

بخش ۳

نتیجه‌گیری

چرا همگرایی به تعادل نش؟
استفاده گسترده‌ای از تعادل نش در نظریه بازی ها
تعادل نش = توصیف شرایطی که در آن‌ها دیگر یادگیری امکان‌پذیر نیست

سوال:
یک کلاس از قوانین ⬅ استفاده همه بازیکنان از قوانین یادگیری این کلاس استفاده
⬅ آیا به تعادل نش همگرا می‌شویم.

مثال
همه بازیکنان قوانین یادگیری کالیبره یکنواخت ⬅ همگرایی بازی به مجموعه‌ی تعادل مشترک همگرا

همگرایی به نش

مقدمه

بخش ۲

بخش ۳

نتیجه‌گیری

آیا کلاس‌هایی از قوانین یادگیری وجود دارند که زمانی که همه بازیکنان از آن‌ها استفاده می‌کنند، به همگرایی جهانی به یک تعادل نش منتهی می‌شوند؟
نتایج اولیه در این راستا منفی بود.

مفهوم دیکوپله بودن
هارت و ماس-کلل (۲۰۰۳)
در سیستم‌های پیوسته و زمان حقیقی و بدون استفاده از اطلاعات گذشته، نمی‌توان به همگرایی به تعادل در بازی اطمینان داشت.
در ادامه، نشان دادند که در روش‌های گسسته و زمان تصادفی، همگرایی به تعادل نیز تضمین نمی‌شود در شرایط "۱-یادآوری" که در آن وضعیت سیستم بر اساس آخرین پروفایل بازی تعیین می‌شود ⬅ بهبود نتایج همگرایی گذشته فوستر و یانگ تضمین همگرایی در شرایط استفاده از دو دوره گذشته

همگرایی به نش

مقدمه

بخش ۲

بخش ۳

نتیجه‌گیری

مدل یادگیری تصادفی فوستر و یانگ

فرایند یادگیری از یک اقدام "وضع موجود" ← بازبینی دوره‌ای به صورت تصادفی

بازبینی‌ها در زمان‌های تصادفی :

وضع موجود ← زمان‌های بازبینی تصادفی ← مقایسه عادلانه‌ای بین نتایج ← بررسی

عملکرد ← ... ← یک اقدام وضع موجود جدید به طور تصادفی ← ...

همگرایی به نش

مقدمه

بخش ۲

بخش ۳

نتیجه‌گیری

برخی روش‌های یادگیری مانند بازی خیالی ⬅ همگرا در بسیاری از محیط

نیاز به قوانین یادگیری پیچیده‌تر!

مدل لوین (۱۹۹۱)

بازیکنانی که فقط به حرکات فوری خود نگاه نمی‌کنند و کمی صبورتر هستند، از تعادل نش دور می‌شوند زیرا ارزش تعهدات بلندمدت خود را درک می‌کنند.

شاما و ارسلان (۲۰۰۵)

دینامیک‌های پیش‌بینی حرکت بعدی حریف بدون دید بلندمدت

چرا و چگونه به قوانین یادگیری شامل پیش‌بینی حرکات آینده حریفان نیاز داریم؟!

همگرایی به نش

مدل شاما و ارسلان ⬅ محیط بازی خیالی نرم با وزن‌دهی نمایی به مشاهدات

$$\sigma_i(t) = (1-\lambda)z_i(t-1) + \lambda\sigma_i(t-1)$$

ثابت بین ۰ و ۱
(نشان‌دهنده نرخ کاهش اهمیت مشاهدات)

یک بردار با یک درایه ۱ و ما بقی ۰
(اقدام بازی شده در زمان t توسط بازیکن i)

فرکانس وزنی تجربی بازیکن i تا زمان t (میانگین وزنی اقدامات بازیکن i در طول زمان)

فرکانس وزنی تجربی همیشه بین مشاهدات جدید و گذشته تعادلی برقرار کند و به مرور زمان تغییرات را به طور نمایی در نظر بگیرد.

همگرایی به نش

$\sigma_{-i}(t-1)$ نشان دهنده عمل بازیکن i- در گذشته نه حال حاضر

⬅ پیش‌بینی بازی فعلی بازیکن i- برای تخمین بهتری از بازی او در حال حاضر

شما و ارسلان ⬅ کنترل مشتق تناسبی ⬅ معرفی r_{-i} برای استفاده در پیش‌بینی
(r_{-i} ، σ_{-i} را دنبال می‌کند.)

$$r_{-i}(t) = r_{-i}(t-1) + \lambda(\sigma_{-i}(t-1) - r_{-i}(t-1))$$

پیش‌بینی بازی بازیکن i-

$$\sigma_{-i}(t-1) + \gamma(r_{-i}(t) - r_{-i}(t-1))$$

γ پارامتر تعیین کننده وزن به تخمین$r(t)$

λ پارامتر تعیین کننده سرعت تنظیم متغیر کمکی

همگرایی به نش

مدل شاما و ارسلان ⬅ استفاده از روش‌های تقریب تصادفی ⬅ بررسی همگرایی سیستم به نش

$$\dot{\sigma}_i = \emptyset[\beta_i(\sigma_{-i} + \gamma\dot{r}_{-i}) - \sigma_i]$$

$$\dot{r}_{-i} = \lambda(\sigma_{-i} - r_{-i})$$

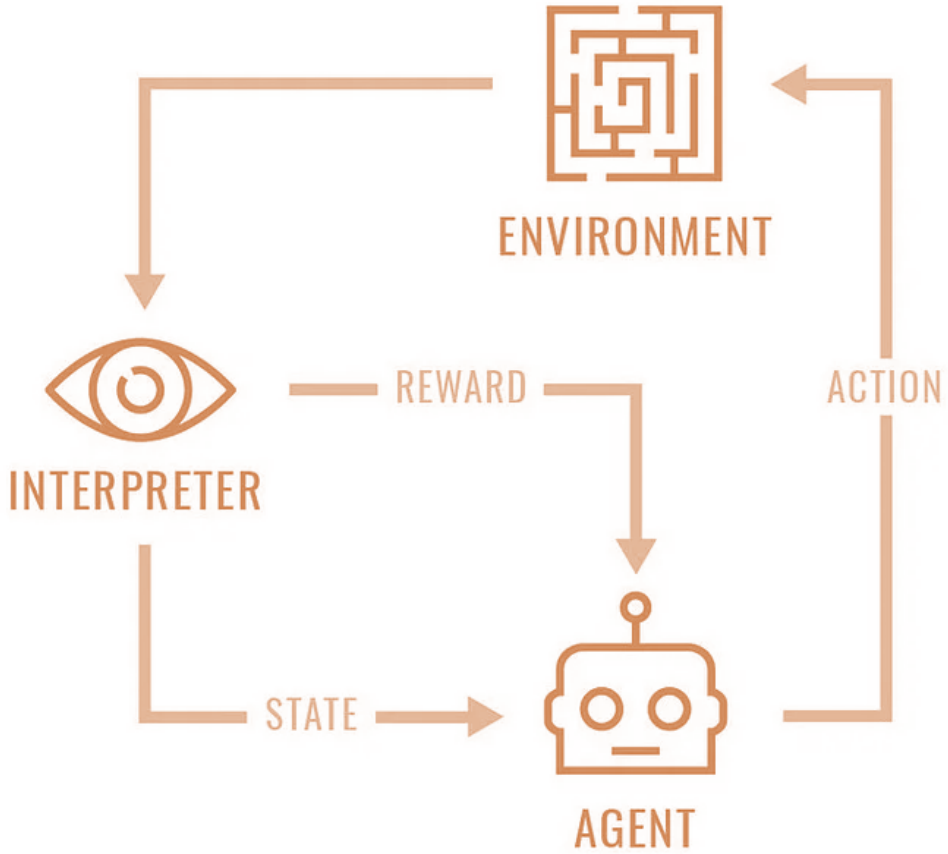
$$\ddot{r}_{-i} = \lambda(\dot{\sigma}_{-i} - \dot{r}_{-i})$$

این مدل نشان می‌دهد که تحت شرایطی خاص، سیستم می‌تواند به تعادل نش همگرا شود.

اگر $\ddot{r}_{-i}$ کوچک باشد:

- $\dot{r}_{-i}$ تقریب خوبی از $\dot{\sigma}_{-i}$ است
- سیستم به طور جهانی به یک توزیع تعادل نش همگرا می‌شود

یادگیری تقویتی



پاداش تقویتی تجمعی (CPR)

پاداش تقویتی تجمعی CPR : یک مدل بنیادی از یادگیری تقویتی که توسط لاسلیه و همکاران مطالعه شده است.

تنظیمات اولیه:

- عوامل با تخصیص وزن‌های اولیه به اقدامات مختلف شروع می‌شوند.

مکانیسم به‌روزرسانی وزن:

- وزن‌ها بر اساس پاداش‌های دریافتی پس از هر اقدام به‌روزرسانی می‌شوند.

احتمال انتخاب اقدام:

- احتمال انتخاب یک اقدام خاص متناسب با پاداش تجمعی آن است.
- این احتمال با پاداش‌های کلی همه اقدامات مقایسه می‌شود.

معادله احتمال عمل

احتمال یک عمل در CPR :

- بدست آمده توسط معادله:

$$P_k(t) = \frac{CU_k(t)}{\sum_j CU_j(t)}$$

جایی که CUk(t) امتیاز تجمعی عمل k در زمان t است.

ویژگی های کلیدی:

- اطمینان حاصل می کند که هر عملی بی نهایت بارها امتحان شده است.
- به عوامل اجازه می دهد تا ارزش همه اقدامات را بیاموزند.

رویکرد احتمالی:

- اکتشاف و بهره برداری را متعادل می کند.
- به مرور زمان منجر به یادگیری موثرتر می شود.

همگرایی و ثبات در CPR

یافته های Laslier و همکاران:

- در سناریوهایی که یک بازیکن از یک استراتژی ثابت استفاده می کند، فرآیند یادگیری به بهترین اقدامات همگرا می شود.
- هر عمل بی‌نهایت و اغلب انجام می‌شود، تا اطمینان حاصل شود که عوامل ارزش همه اقدامات را یاد می‌گیرند.
- تجزیه و تحلیل با استفاده از تقریب تصادفی:

مدل سازی شده با یک معادله دیفرانسیل زمان پیوست

$$\dot{x} = -x + r(x)$$

x نشان دهنده کسری از زمان انتخاب هر عمل است.
r(x) احتمال هر عمل بر اساس وضعیت فعلی است.

فرآیندهای تصادفی تقریبی

مقدمه

بخش ۲

بخش ۳

نتیجه‌گیری

استفاده از فرآیندهای تصادفی تقریبی در RL :

- فرآیند RL زمان گسسته را با یک معادله دیفرانسیل زمان پیوسته تقریب می‌کند.

هدف:

- به محققان کمک می کند تا رفتار بلند مدت و ویژگی های همگرایی فرآیند یادگیری را درک کنند.

بینش های به دست آمده:

- درک چگونگی تنظیم استراتژی های خود توسط عوامل در طول زمان را فراهم می کند.

یادگیری تقویتی آشفته

مطالعه هاپکینز در مورد CPR:

- بررسی نسخه های CPRبا اغتشاشات تصادفی کوچک.

دینامیک تقویت آشفته:

آشفتگی ها تغییرات کوچک و تصادفی را در فرآیند یادگیری ایجاد می کنند.

یافته های کلیدی:

- نشان دهید که چگونه استراتژی ها بر اساس موفقیت نسبی تکامل می یابند.
- استحکام مدل های RL را در برابر نوسانات و عدم قطعیت های جزئی توضیح دهید.

مدل های تقویتی ثابت

تحلیل بورگرز و سارین:

- یک مدل تقویتی با نرخ به روز رسانی ثابت مورد بررسی قرار گرفت.

یافته های کلیدی:

- نشان داد که مدل ممکن است به اقدامات نادرست همگرا شود.
- نرخ به روز رسانی ثابت مانع از یادگیری مداوم بهترین اقدامات توسط بازیکنان می شود.

اهمیت یافته ها:

- ضرورت مکانیسم‌های به‌روزرسانی تطبیقی را در مدل‌های RL برجسته می‌کند.

$$u_k(t+1) = u_k(t) + \gamma(R_k(t) - u_k(t))$$

مدل های تقلیدی

مقدمه

بخش ۲

بخش ۳

نتیجه‌گیری

مدل های تقلیدی

عواملی را درگیر کنید که بر اساس موفقیت مشاهده شده از اقدامات دیگران تقلید می کنند.

بیورنستاد و وایبول

عوامل مطالعه با اطلاعات پر سر و صدا در مورد پاداش دیگران.
نشان داد که چگونه نمایندگان به استراتژی‌های با بهترین عملکرد تغییر می‌کنند.

بینمور و ساموئلسون

تقلید کاوش شده با آرزوهای ثابت.

نواک و همکاران

تقلید ترکیبی از استراتژی های موفق و محبوب.
پویایی یادگیری متنوع را نشان داد.

مدل های مبتنی بر آرزوها

مدل های مبتنی بر آرزو

اگر پاداش های دریافتی کمتر از سطح هدف باشد، نمایندگان رفتار خود را تنظیم می کنند.

آثار چو و ماتسویی

نشان داد که چگونه عوامل می توانند همکاری یا مخالفت را بر اساس رضایت خود از پاداش های فعلی یاد بگیرند.

بینش های کلیدی

این مدل‌ها نشان می‌دهند که چگونه آرمان‌های کارگزاران بر فرآیندهای یادگیری و تصمیم‌گیری آنها تأثیر می‌گذارد.

تعادل خود تأیید کننده SCE

مقدمه

بخش ۲

بخش ۳

نتیجه‌گیری

تعادل خود تأیید کننده SCE:

- مفهومی که در آن بازیکنان اعتقاداتی در مورد استراتژی های حریف دارند که بر اساس مشاهدات آنها صحیح است.

عقاید و اعمال:

- این باورها اقدامات بازیکنان را هدایت می کنند، حتی اگر کاملاً دقیق نباشند.

اهمیت SCE:

- توضیح می دهد که چگونه نتایج پایدار می تواند بدون آگاهی کامل رخ دهد.
- بازیکنان بر اساس واقعیت درک شده خود بهینه سازی می کنند.

آزمایش و تعادل

تعیین آزمایش کافی:

- برای دستیابی به تعادل robust ضروری است.

یافته های فودنبرگ و کرپس:

- نتایج غیر نش نمی توانند پایدار باشند اگر هر اقدامی بی نهایت مرتب انجام شود.

بینش کلیدی:

- اهمیت آزمایش کافی را برجسته می کند.
- اطمینان حاصل می کند که تمام استراتژی های بالقوه بررسی شده و تصمیمات بهینه گرفته شده است.

مدل های مبتنی بر باور

مقدمه

بخش ۲

بخش ۳

نتیجه‌گیری

مدل های مبتنی بر اعتقاد:

- نیاز به باورهای بازیکنان برای همسویی با داده های مشاهده شده.
- آگاهی از عملکردهای بازده دیگران را فرض نکنید.

مثال ها:

- تعادل حدسی منطقی روبینشتاین و وولینسکی.
- تعادل منطقی خود تأیید کننده دکل و همکاران.

بینش های کلیدی:

- به درک اینکه چگونه باورها یادگیری و سازگاری استراتژی را شکل می دهند کمک کنید.

یادگیری بیزی

یادگیری بیزی:

- شامل بازیکنانی می شود که باورهای خود را بر اساس داده های مشاهده شده به روز می کنند.

یافته های Acemoglu و همکاران:

- نشان داد که چگونه عوامل بیزی منطقی می توانند با وجود مواجهه با داده های مشابه، عقاید متفاوتی را حفظ کنند.
- پیچیدگی های یادگیری در زمینه های اقتصادی را برجسته می کند.

اهمیت مدل های بیزی:

- چارچوبی برای درک به‌روزرسانی‌های باورها فراهم کنید.
- کمک به تصمیم گیری در شرایط عدم اطمینان.

یادگیری استقرا معکوس

استقرا معکوس:

- تکنیکی که در آن بازیکنان از پایان بازی به عقب کار می کنند تا استراتژی های بهینه را تعیین کنند.

همگرایی به تعادل زیربازی کامل:

- با آزمایش کافی، بازیکنان می توانند به این همگرایی دست یابند.
- تصمیم گیری بهینه را در هر مرحله تضمین می کند.

کاربرد:

- به ویژه در بازی های فرم گسترده با تصمیم گیری های متوالی مفید است.

یادگیری غیرتعادلی در اقتصاد کلان

نقش یادگیری در نظریه های اقتصاد کلان:

- برای درک پویایی های اقتصادی و سیاست گذاری بسیار مهم است.

یادگیری غیر تعادلی:

- به خصوص تعادل خود تأیید کننده SCE

برنامه های کاربردی:

- مطالعه چگونگی سازگاری اقتصادها با تغییرات
- درک چگونگی تثبیت اقتصادها در طول زمان.

کاربردهای عملی

مقدمه

بخش ۲

بخش ۳

نتیجه‌گیری

کاربردهای عملی مدل های یادگیری:

- در زمینه های مختلف از جمله تئوری حراج، کالاهای عمومی و پویایی بازار استفاده می شود.

فواید:

- به سیاستگذاران کمک می کند تا مکانیسم های بهتری طراحی کنند.
- به پیش بینی رفتار در سیستم های اقتصادی و اجتماعی کمک می کند.

مثال ها:

- طراحی حراج های کارآمد
- درک واکنش بازار به تغییرات سیاست

نتیجه گیری

مقدمه

بخش ۲

بخش ۳

نتیجه‌گیری

Score-Based Equilibrium Learning in Multi-Player Finite Games with Imperfect Information

Exploiting Hidden Structures in Non-Convex Games for Convergence to Nash Equilibrium

Scalable and Independent Learning of Nash Equilibrium Policies in Stochastic Games with Unknown Independent Chains

Paths to Equilibrium in Normal-Form Games

The Complexity of Approximate (Coarse) Correlated Equilibrium for Incomplete Information Games

Thanks