تحلیل نمودار termporal difference error

الهه رضاپناه ، زهرا رستمی

9912762858–9912762789

با توجه به نمودار خطای تفاوت زمانی (TD Error) برای الگوریتم‌های Q-learning و SARSA:

بر اساس مشاهدات از نمودار متوجه می‌شویم که:

– خطای TD برای Q-learning (خط آبی) که در عکس به دلیل هم‌پوشانی با خطای TD SARSA (منطقه نارنجی) قابل مشاهده نیست .

– خطای TD SARSA (منطقه نارنجی) نشان می‌دهد که این مقادیر در ابتدا دارای نوسانات زیادی هستند و به تدریج با افزایش تعداد گام‌های زمانی تثبیت می‌شوند.

– با گذشت زمان، خطای TD برای SARSA به مقدار مرکزی نزدیک به صفر می‌رسد که نشان‌دهنده کاهش و تثبیت خطا با پیشرفت یادگیری است.

مورد دیگیری که باید به آن اشاره کرد همگرایی خطای TD است:

– مقدار مرکزی که خطای TD به آن نزدیک می‌شود، برای هر دو الگوریتم Q-learning و SARSA نزدیک به صفر است. این همگرایی به سمت صفر یک ویژگی مطلوب در یادگیری تقویتی است، زیرا نشان می‌دهد که الگوریتم‌ها در حال یادگیری هستند و مقادیر Q تخمینی آنها به مقادیر واقعی Q نزدیک می‌شود.

– نوسانات اولیه بالا در خطای TD می‌تواند به فاز اکتشاف نسبت داده شود، جایی که عامل هنوز در حال یادگیری و تصمیم‌گیری‌های تصادفی بیشتری است که منجر به خطاهای TD بالاتر و متغیرتر می‌شود.

از دلایل همگرایی می‌توان به موارد زیر اشاره کرد:

– کاهش خطای TD در طول زمان نتیجه فرایند یادگیری است. وقتی عامل تجربیات بیشتری کسب می‌کند و مقادیر Q خود را بر اساس قانون یادگیری TD به‌روزرسانی می‌کند، تخمین‌ها دقیق‌تر می‌شوند و خطای TD کاهش می‌یابد.

– پارامترها، مانند نرخ یادگیری (ALPHA) و ضریب تخفیف (GAMMA)، بر سرعت همگرایی خطای TD تأثیر می‌گذارند. تعادل مناسب این پارامترها اطمینان می‌دهد که مقادیر Q به‌طور مؤثر همگرا می‌شوند بدون نوسان یا واگرایی.

به طور خلاصه، خطای TD برای هر دو الگوریتم Q-learning و SARSA به مرور زمان به مقادیری نزدیک به صفر همگرا می‌شود که نشان‌دهنده یادگیری موفقیت‌آمیز است. نوسانات اولیه بالا در خطای TD با کسب تجربیات بیشتر توسط الگوریتم‌ها کاهش می‌یابد و منجر به تخمین‌های دقیق‌تر مقادیر Q می‌شود. همگرایی خطای TD به صفر نشان‌دهنده این است که الگوریتم‌ها به درستی پاداش‌های آینده را تخمین می‌زنند که هدف اصلی یادگیری تقویتی است.