تلاش‌هایی برای پیدا کردن کران بالایی بهتر برای مسئله بندیت پیوسته و لیپ‌شیتز

حامی عبادزاده مهدی دارابی

درس: آنالیز احتمالات ابعاد بالا
دانشگاه صنعتی شریف

۲ اسفند ۱۴۰۴

به نام خدا

فهرست مطالب

۱ مقدمه و شهود مسئله

۲ تعریف رسمی: بندیت گسسته

۳ UCB۱

۴ بندیت پیوسته و فرض لیپ‌شیتز

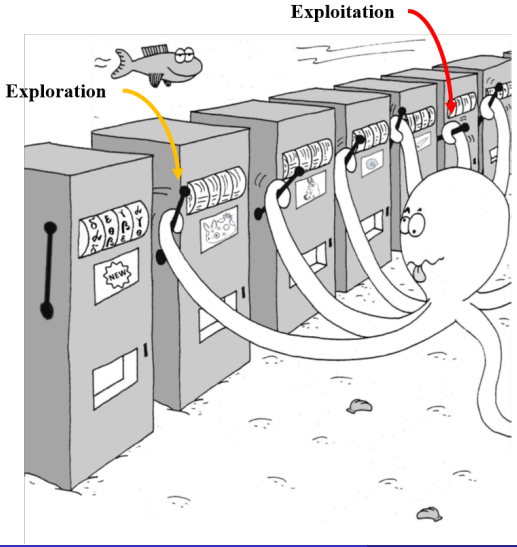
۵ گسسته‌سازی با εـ پوشش و سیاست پیشنهادی

۶ تلاش اول: کران بالا با گسسته‌سازی

۷ تلاش دوم: ایده ناموفقِ کران با سوپریمم یک فرایند

۸ نتیجه‌گیری و مسیرهای ادامه

مقدمه: موازنه اکتشاف/بهره‌برداری



- در تصمیم‌گیری تکرارشونده، کیفیت واقعی گزینه‌ها از پیش معلوم نیست.
- باید هم‌زمان:
  - **اکتشاف**: آزمون گزینه‌های کمترشناخته برای کسب اطلاعات،
  - **بهره‌برداری**: استفاده از گزینه‌هایی که تاکنون بهتر بوده‌اند.
- هدف: کمینه‌کردن **پشیمانی** در افق زمانی.

تعریف: بندیت چندبازویی

- K بازو؛ برای هر بازو i دنباله پاداش‌ها $Y_{i,۱}, Y_{i,۲}, \ldots$ مستقل و هم‌توزیع با توزیع ناشناخته P_i.
- $\mu_i := \mathbb{E}[Y_{i,۱}]$ و فرض می‌کنیم پاداش‌ها در بازه $[۰, ۱]$ کراندارند.
- $T_i(n)$ تعداد دفعات انتخاب بازوی i تا پایان گام n.

$$\mathbb{E}\left[\sum_{i=۱}^{K}\sum_{s=۱}^{T_i(n)} Y_{i,s}\right] = \sum_{i=۱}^{K} \mu_i \mathbb{E}[T_i(n)].$$

پشیمانی (Regret)

$$\mu^\star := \max_{1\le i\le K} \mu_i, \qquad R(n) := \mu^\star n - \sum_{i=1}^{K} \mu_i \mathbb{E}[T_i(n)].$$

هدف: طراحی سیاستی با $R(n)$ کوچک.

UCB۱: شاخص

$$\widehat{\mu}_i(t-۱) = \frac{۱}{T_i(t-۱)} \sum_{s=۱}^{T_i(t-۱)} Y_{i,s}, \qquad I_i(t) = \widehat{\mu}_i(t-۱) + \sqrt{\frac{۲\ln n}{T_i(t-۱)}}.$$

- جمله اول: بهره‌برداری.
- جمله دوم: اکتشاف (عدم‌قطعیت).

الگوریتم UCB1

UCB۱

1. **مقداردهی اولیه:** هر بازو $i = ۱, \dots, K$ را دقیقاً یک‌بار انتخاب کن و پاداش را ثبت کنید.
2. **برای هر گام** $t = K + ۱, \dots, T$:
   - برای هر بازو i شاخص زیر را محاسبه کنید:

$$I_i(t) = \widehat{\mu}_i(t-۱) + \sqrt{\frac{۲ \ln n}{T_i(t-۱)}}.$$

   - بازوی با بیشترین شاخص را انتخاب کنید و پاداش را ثبت کنید.

UCB1: کران پشیمانیِ وابسته به شکاف

$$\Delta_i := \mu^\star - \mu_i \quad \text{(برای بازوهای زیربهینه)}$$

نتیجه کلاسیک

$$\mathbb{E}[T_i(n)] \leq \frac{۸ \ln n}{\Delta_i^۲} + ۱ + \frac{\pi^۲}{۳}$$

$$R(n) \leq \sum_{i:\Delta_i > ۰} \left(\frac{۸ \ln n}{\Delta_i} + \left(۱ + \frac{\pi^۲}{۳}\right)\Delta_i \right)$$

تعریف: بندیت پیوسته

- فضای عمل: $\mathcal{X}$ (مثلاً زیرمجموعه‌ای فشرده از $\mathbb{R}^d$).
- در گام t عمل $X_t \in \mathcal{X}$ و پاداش $Y_t \in [۰, ۱]$.
- تابع میانگین: $f(x) = \mathbb{E}[Y_t \mid X_t = x]$ و $f^* = \sup_{x \in \mathcal{X}} f(x)$.

$$R(n) = \sum_{t=۱}^{n} \mathbb{E}\big[f^* - f(X_t)\big].$$

فرض لیپ‌شیتز

اگر برای ثابت $L > ۰$ داشته باشیم

$$|f(x) - f(x')| \leq L\, d(x, x') \qquad (\forall x, x' \in \mathcal{X}),$$

آن‌گاه اعمال نزدیک، میانگین پاداش نزدیک دارند و «ساختار هندسی» قابل استفاده می‌شود.

شهود لیپ‌شیتز

پیام شهودی

اگر ورودی‌ها (نقاط) نزدیک باشند، خروجی‌ها نمی‌توانند «پرش» بزرگی داشته باشند؛ شیب تغییرات با یک ثابت L کنترل می‌شود.

مثال قیمت‌گذاری و توجیه فرض لیپ‌شیتز

سناریو (کنش = قیمت)

فرض کنید کنش شما انتخاب یک قیمت باشد:

$$x \in [۰, ۱].$$

در هر مرحله، قیمت x را اعلام می‌کنید و درآمد تصادفی زیر را مشاهده می‌کنید:

$$Y = x \cdot ۱\{\text{مشتری با قیمت x خرید می‌کند}\}.$$

تابع میانگین پاداش

تابع پاداش میانگین (میانگین شرطی درآمد) برابر است با:

$$f(x) = \mathbb{E}[Y \mid X = x] = x\,p(x),$$

که در آن $p(x)$ احتمال خرید در قیمت x است.

در بسیاری از بازارها، تغییرات کوچک در قیمت معمولاً باعث «جهش ناگهانی» در $p(x)$ نمی‌شود؛ پس $p(x)$ و در نتیجه $f(x)$ به‌طور تدریجی تغییر می‌کنند. این دقیقاً همان شهودی است که فرض لیپ‌شیتز برای f بیان می‌کند.

گسسته‌سازی با εـ پوشش

اگر $(\mathcal{X}, d)$ فضای متریک باشد، $N_\varepsilon \subset \mathcal{X}$ یک εـ پوشش است اگر:

$$\forall x \in \mathcal{X}\ \exists \tilde{x} \in N_\varepsilon :\ d(x, \tilde{x}) \leq \varepsilon$$

اگر

$$N_\varepsilon = \{x_1, \ldots, x_K\}, \qquad K := |N_\varepsilon|$$

می‌توان نقاط شبکه را به عنوان بازوهای یک مسئله گسسته در نظر گرفت و UCB۱ را روی آن اجرا کرد.

مسئله پیشنهادی

پرسش اصلی

اگر در مسئله‌ی بندیتِ پیوسته با پاداش‌های کراندار، تابع میانگین پاداش f لیپ‌شیتز باشد، آیا می‌توان با گسسته‌سازی فضای عمل از طریق یک εـ پوشش و اجرای مستقیم الگوریتم UCB1 روی مجموعه‌ی گسسته‌ی N_ε، به کران پشیمانی‌ای دست یافت که (با انتخاب مناسب ε) از کران ارائه‌شده در حالت پیوسته (قضیه ۱) بهتر باشد؟

- فضای عمل: $\mathcal{X}$، تابع میانگین: $f : \mathcal{X} \to [0, 1]$.
- $N_\varepsilon \subset \mathcal{X}$ یک εـ پوشش و $|N_\varepsilon| < \infty$.
- سیاست پیشنهادی: محدود کردن انتخاب‌ها به N_ε و اجرای UCB1 روی آن.

تلاش‌ها

تلاش اول: گسسته‌سازی با εـ پوشش و اجرای UCB1

- **ایده:** فضای پیوسته‌ی عمل $\mathcal{X}$ را با یک εـ پوشش N_ε تقریب می‌زنیم و انتخاب‌ها را به N_ε محدود می‌کنیم.
- سپس مسئله به یک بندیت **گسسته** با $K := |N_\varepsilon|$ بازو تبدیل می‌شود و UCB1 را روی این K بازو اجرا می‌کنیم.
- پشیمانی کل به دو مؤلفه تفکیک می‌شود:

$$\text{(پشیمانی روی شبکه)} \; + \; \text{(خطای گسسته‌سازی)}.$$

- در ادامه، ε را طوری انتخاب می‌کنیم که این دو مؤلفه متعادل شوند.

سیاست «گسسته‌سازی + UCB۱»

ایده اصلی

فقط از مجموعه محدود N_ε عمل انتخاب کنید:

$$X_t \in N_\varepsilon$$

و UCB۱ را روی بازوهای متناظر با $\{x_1, \ldots, x_K\}$ اجرا کن، با میانگین بازوی i برابر:

$$\mu_i = f(x_i)$$

پشیمانی نسبت به بهینهٔ پیوسته:

$$R(n) = \sum_{t=1}^{n} \left(\mathbb{E}[f^* - f(X_t)]\right)$$

لم ۱: خطای تقریب شبکه (کنترل با لیپ‌شیتز)

بیان لم و اثبات کوتاه

فرض کنید N_ε یک εـ پوشش از χ و f با ثابت L لیپ‌شیتز باشد. تعریف کنید

$$f^\star := \sup_{x\in\chi} f(x), \qquad f^\star_\varepsilon := \max_{x\in N_\varepsilon} f(x).$$

آنگاه

$$0 \le f^\star - f^\star_\varepsilon \le L\varepsilon.$$

اثبات کوتاه. بگذارید $x^\star \in \arg\max_{x\in\chi} f(x)$. چون N_ε یک εـ پوشش است، نگاشتی $\pi : \chi \to N_\varepsilon$ وجود دارد به‌طوری‌که

$$d\big(x, \pi(x)\big) \le \varepsilon \qquad (\forall x \in \chi).$$

در نتیجه

$$f^\star = f(x^\star) \le f\big(\pi(x^\star)\big) + L\varepsilon \le f^\star_\varepsilon + L\varepsilon.$$

پس $f^\star - f^\star_\varepsilon \le L\varepsilon$ و از طرفی بدیهی است $f^\star_\varepsilon \le f^\star$.

نتیجه‌ی لم ۱: تجزیه و کران پشیمانی

اگر $X_t \in N_\varepsilon$ (یعنی سیاست فقط از نقاط شبکه انتخاب کند)، آنگاه

$$R(n) := \sum_{t=1}^{n}\big(\mathbb{E}[f^\star - f(X_t)]\big) = \sum_{t=1}^{n}\big(\mathbb{E}[f^\star - f^\star_\varepsilon]\big) + \sum_{t=1}^{n}\big(\mathbb{E}[f^\star_\varepsilon - f(X_t)]\big) \le L\varepsilon n + R^\varepsilon_n,$$

که در آن

$$R^\varepsilon_n := \sum_{t=1}^{n}\big(\mathbb{E}[f^\star_\varepsilon - f(X_t)]\big)$$

پشیمانی نسبت به بهترین نقطه روی شبکه است.

لم ۲: کران مستقل از فاصله برای UCB۱ روی K بازو

در مسئله گسسته با K بازو و پاداش‌های [۰ , ۱]، پشیمانی مورد انتظار UCB۱ پس از T گام:

$$R_n^{\varepsilon} \leq ۴\sqrt{۲K \times n\ln(n)} + \left(۱ + \frac{\pi^۲}{۳}\right)K$$

- ایده اثبات: بازوها را با آستانه α به «نزدیک‌بهینه» و «دور از بهینه» تقسیم می‌کنیم.
- برای نزدیک‌بهینه‌ها: $\leq \alpha n$.
- برای دورها: از کران کلاسیک $\mathbb{E}[T_i(n)]$ استفاده می‌کنیم.
- انتخاب بهینه $\alpha = \sqrt{(۸K\ln(n))/n}$ دو جمله اصلی را هم‌مرتبه می‌کند.

ترکیب لم‌ها: کران نهاییِ تلاش اول

از لم ۱ و لم ۲:

$$R(n) \leq L\varepsilon n + ۴\sqrt{۲K \times n\ln(n)} + \left(۱ + \frac{\pi^۲}{۳}\right)K$$

برای $\mathcal{X} = [۰, ۱]^d$ با متریک $\|\cdot\|_\infty$ می‌توان پوششی با

$$K = |N_\varepsilon| \lesssim \left(\frac{۱}{\varepsilon} + ۱\right)^d \approx \varepsilon^{-d}$$

در نظر گرفت.
با انتخاب

$$\varepsilon := n^{-۱/(d+۲)} \quad \Rightarrow \quad K = \varepsilon^{-d} = n^{d/(d+۲)}$$

به دست می‌آید:

$$R(n) \leq L\, n^{\frac{d+۱}{d+۲}} + ۴\sqrt{۲\ln(n)}\, n^{\frac{d+۱}{d+۲}} + \left(۱ + \frac{\pi^۲}{۳}\right) n^{\frac{d}{d+۲}}.$$

جمع‌بندی تلاش اول

- سیاست ساده: شبکه‌سازی ε و اجرای مستقیم UCB۱ روی شبکه.
- دو منبع پشیمانی:
  - خطای گسسته‌سازی: $L\varepsilon n$
  - پشیمانی گسسته UCB۱ روی K بازو
- بهینه‌سازی ε نرخ اصلی را از جنس $n^{\frac{d+1}{d+2}}$ (با عامل $\sqrt{\ln(n)}$) می‌دهد.

تلاش دوم (ناموفق): بازنویسی پشیمانی به‌صورت کران از بالا برای امید سوپریمم یک فرایند

انگیزه

ایده‌ی این تلاش استفاده از لم زیر بود که «امید سوپریمم» یک فرایند تصادفیِ لیپ‌شیتز و زیرگاوسی را با عدد پوشش کنترل می‌کند. به‌طور خلاصه، اگر $\{Z_t\}_{t\in T}$ یک فرایند لیپ‌شیتز و زیرگاوسی باشد، آنگاه کرانی از شکل زیر برقرار است:

$$\mathbb{E}\left[\sup_{t\in T} Z_t\right] \leq \varepsilon\,\mathbb{E}[C] + \sqrt{۲\sigma^۲ \log N(T,d,\varepsilon)}.$$

در این‌جا C ثابت لیپ‌شیتز، $\sigma^۲$ پارامتر زیرگاوسی است.

انگیزه تلاش دوم: لم سوپریمم فرایند زیرگاوسی

هدف: بازنویسی پشیمانی به شکل

$$R(n) \leq \mathbb{E}\left[\sup_{x\in\mathcal{X}} Z_x\right]$$

به گونه‌ای که فرایند $\{Z_x\}_{x\in\mathcal{X}}$:

- نسبت به x **لیپ‌شیتز** باشد،
- **زیرگاوسیِ مرکزدار** باشد،

تا بتوان امید سوپریمم را با عدد پوشش کنترل کرد.

یک انتخاب وسوسه‌انگیز و مشکل بنیادی

انتخاب صوری:

$$Z_x := n \times f(x) - \sum_{t=1}^{n} f(X_t)$$

- اگر $x^\star \in \arg\max f(x)$، آنگاه $Z_{x^\star}$ «شبیه» پشیمانیِ مبتنی بر f است.
- اگر f لیپ‌شیتز باشد:

$$|Z_x - Z_{x'}| = n \times |f(x) - f(x')| \leq n \times L\, d(x, x')$$

پس فرایند نسبت به x لیپ‌شیتز است.

- چون $f(\cdot) \in [0, 1]$، داریم $-n \leq Z_x \leq n$ و در نتیجه کران‌دار/زیرگاوسی است.

مشکل اصلی: Z_x مرکزی نیست:

$$\mathbb{E}[Z_x] = n \times f(x) - \sum_{t=1}^{n} \mathbb{E}[f(X_t)] \neq 0$$

مرکزی‌سازی و از بین رفتن وابستگی به x

اگر مرکزی کنیم:

$$\widetilde{Z}_x := Z_x - \mathbb{E}[Z_x]$$

آنگاه:

$$\widetilde{Z}_x = \left(n \times f(x) - \sum_{t=1}^{n} f(X_t)\right) - \left(n \times f(x) - \sum_{t=1}^{n} \mathbb{E}[f(X_t)]\right) = -\sum_{t=1}^{n} f(X_t) + \sum_{t=1}^{n} \mathbb{E}[f(X_t)]$$

که دیگر **به x وابسته نیست**. پس:

$$\sup_{x\in\mathcal{X}} \widetilde{Z}_x = \widetilde{Z}_x, \qquad \mathbb{E}\left[\sup_{x\in\mathcal{X}} \widetilde{Z}_x\right] = \mathbb{E}[\widetilde{Z}_x] = 0$$

بنابراین این مسیر در بهترین حالت «کران صفر» برای کمیت مرکزی می‌دهد و عملاً کران مفیدی برای پشیمانی تولید نمی‌کند.

کران بدیهی پشیمانی (بدون ساختار بیشتر)

چون در هر گام:

$$۰ \leq f^\star - f(X_t) \leq ۱$$

داریم:

$$۰ \leq R(n) = \sum_{t=۱}^{n} \mathbb{E}[f^\star - f(X_t)] \leq n \quad \Rightarrow \quad R(T) \leq n$$

و تلاش دوم در نهایت به چیزی بهتر از این نرخ عمومی نرسید.

جمع‌بندیِ تلاش دوم (ناموفق)

- هدف: کران‌گذاری پشیمانی با بازنویسی آن به شکل $\mathbb{E}[\sup_{x\in\mathcal{X}} Z_x]$ و استفاده از لم سوپریممِ فرایند لیپ‌شیتز—زیرگاوسی.
- مانع: نتوانستیم فرایندی بسازیم که هم وابستگی به x را حفظ کند و هم مرکزدار/زیرگاوسی باشد.
- نتیجه: بهترین کرانِ حاصل هم‌مرتبه با کران بدیهی $R(n) \leq n$ (یعنی $O(n)$) شد.

نتیجه‌گیری

- **تلاش اول (موفق):** گسسته‌سازی با εـ پوشش + UCB۱

$$R(n) \lesssim n^{\frac{d+1}{d+2}} \sqrt{\ln(n)} \quad (+\text{جملات پایین‌تر})$$

- **تلاش دوم (ناموفق):** ساخت فرایند Z_x برای کنترل $\mathbb{E}[\sup_x Z_x]$
  - چالشِ مرکزدار بودن و حفظ وابستگی به x
  - مرکزی‌سازی وابستگی به x را از بین برد

مراجع

Fischer، Paul and Cesa-Bianchi، Nicolò Auer، Peter
Finite-time analysis of the multiarmed bandit problem، *Machine Learning* ۴۷ (۲–۳):۲۳۵–۲۵۶، ۲۰۰۲.