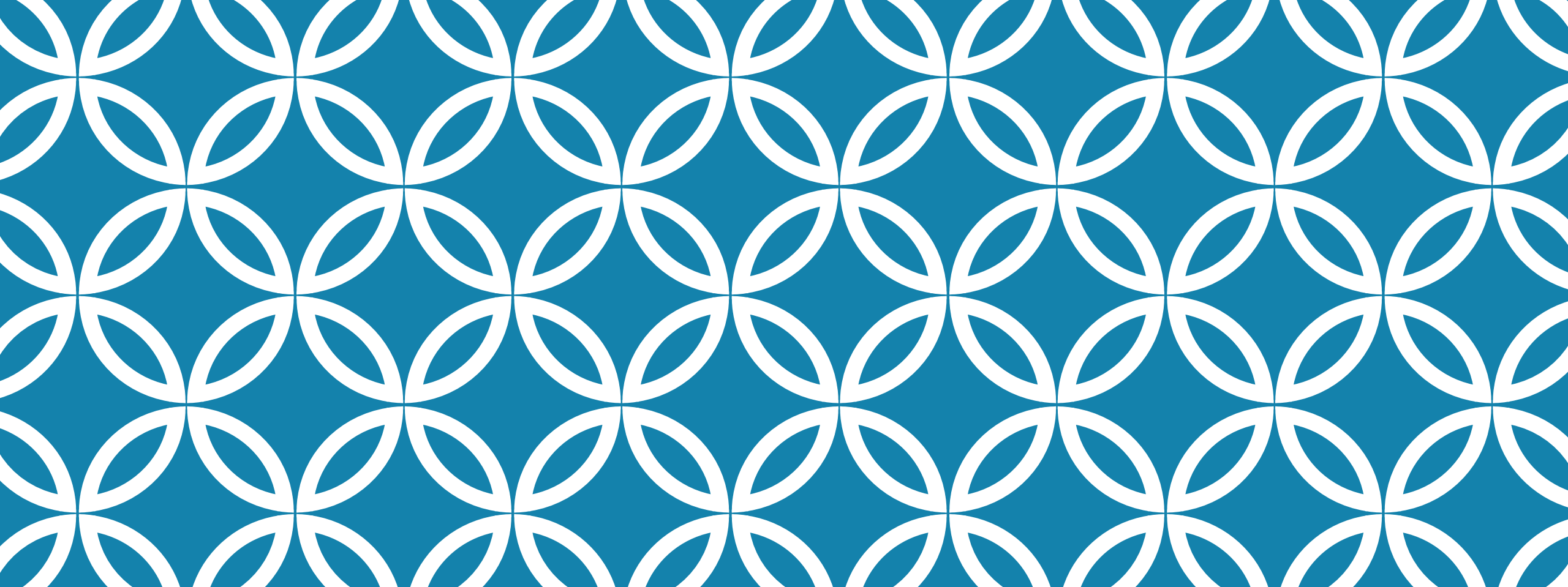




یادگیری تقویتی
کاوش در برابر بهره‌برداری

محسن هوشمند
دانشکده تکنولوژی اطلاعات و علم رایانه
دانشگاه تحصیلات تکمیلی علوم پایه زنجان

اجزای یادگیری تقویتی

محیط

عامل

حالت

کنش (عمل و حرکت) سیاست

پاداش

مسیر $s_0, a_0, r_1, s_1, a_1, r_2, \dots$

با هدف یافتن بهترین دنبالهٔ پیشنهاد ساز مجموعه پاداش‌ها

اصول یادگیری تقویتی

- i. مشاهده محیط (آگهی عامل از حالت فعلی)
- ii. تصمیم به چگونگی اقدام مبنی بر استراتژی (سیاستی) و اقدام
- iii. تغییر محیط (انتقال به حالت جدید) و دریافت پاداش یا جریمه
- iv. پردازش و تحلیل بازخورد (پاداش دریافتی از محیط) [یادگیری از تجارب و اصلاح سیاست] ارزیابی سیاست
- v. تکرار تا یافتن سیاست بهینه

کاوش و بهره‌برداری

کاوش - جستجو در فضای ناشناخته جهت یافتن سیاست بهتر
بهره‌برداری - استفاده از دانش موجود برای تعیین سیاست
یادگیری همزمان و تعامل با محیط
استفاده محض از بهره‌برداری منجر به سیاست زیربهبوده
نیاز به کاوش جهت جستجوی تمامی فضای جستجو
▪ آزمایش تمام مسیرها و حالت‌ها

یادگیری تقویتی

وجه ممیزه‌ی ت از دیگر انواع یادگیری

▪ استفاده از اطلاعات آموزشی جهت «ارزیابی» کنش

▪ محاسبه پاداش دریافتی با اجرای کنشی خاص

▪ دیگر روش‌ها «دیکته» یا «دستور» کنش مناسب

▪ موجب ایجاد نیاز به کاوش فعالانه و پویا

▪ در راستای جستجوی آگاهانه و مشخص رفتار مناسب

▪ موجب دو نوع روش $\Leftarrow$ بازخورد ارزیابی در مقابل بازخورد دستوری

▪ بازخوردگیری ارزیابی صرف در مقابل بازخوردگیری دستوری صرف

▪ مورد متقدم

▪ نمایشگر میزان خوبی کنش انجام پذیرفته

▪ ارزیابی اعمال ارزش به کنش انجام یافته

▪ عدم نمایش بهترین یا بدترین کنش

▪ کاملاً وابسته به انجام کنش

▪ مورد متاخر

▪ در بازخورد دستوری، اعمال ارزش به کنش بهینه‌ای که باید انجام می‌پذیرفت

▪ نمایشگر بهترین کنش ممکن

▪ کامل مستقل از وقوع کنش

تمرین: انواع روش‌های یادگیری در امور انسانی تحقیق شود
و با مرجع گزارش شود.

ماشین سکه‌ای چند اهرم (راهزن)

مثال مسئله راهزن

- جهت بررسی جنبه‌های ارزیابانه
- صرفاً در یک موقعیت
- تک حالت
- Non-associative

مسئله راهزن (دیگر نام‌ها)

- ماشین سکه
- Multi arm bandit (Ame)
- Single arm bandit
- Jackpot
- Slot machine(Bri)
- Fruitmachine (Bri)

ماشین سکه‌ای چند اهرم

انتخاب مکرر از بین چند گزینه

- هر انتخاب منجر به دریافت مقداری پاداش
- کدام اهرم را پشت سر هم بکشم تا در آینده مجموعه پاداش‌ها به بیشترین حد خود برسد.
- عدم اطلاع کاربر از هر ماشین
- امتحان چند بارهٔ دستگاه و ایجاد تدریجی درکی از امتیازات هر اهرم
- $\Leftarrow$ بیشینه‌سازی امید کل پاداش در طول زمان
- در قالب گام‌های زمانی

تمرین: صورت کلی مسئله راهزن را تحقیق کنید و چند مورد استفاده آن را در رایانه و دیگر شاخه‌های علوم گزارش کنید.

ماشین سکه‌ای چند اهرم

مسئله

فرض پاداش‌ها دارای توزیع نرمال

هر اهرم دارای یک میانگین و یک انحراف معیار

▪ بیش از ۹۹ درصد داده در بازه 6σ

⇐ در هر بار اجرا دریافت پاداشی با توزیع احتمال مشخص

به دنبال اجرای کنشی در طولانی مدت با بیشترین میانگین پاداش

نحوه کلی کار

▪ تخصیص ارزشی به هر کنش انجام شده

▪ ارزیابی آنها و انتخاب کنش

ماشین سکه‌ای چند اهرم

ارزش واقعی کار یا ارزش بهینه کنش a از k کنش
 $q_*(a) = E[R_t | A_t = a]$

عدم امکان محاسبه بالا در عمل

چرا؟

- در عمل شناختی از پاداش‌ها نداریم.
- استفاده از تقریب عددی

بازخورد ارزیابی

- فرایند تعیین تقریبی ارزش کنش‌ها از روی پاداش دریافتی

با تعیین ارزش کنش‌ها، انتخاب کنش با بیشترین ارزش

ماشین سکه‌ای چند اهرم

$Q_t(a)$

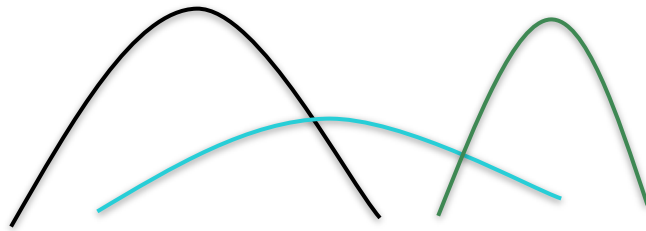
- نمایشگر تخمینی از ارزش کنش a در زمان t
- توجه به تفاوت آن با $q_*(a)$
- مورد اخیر میزان واقعی که کاربر اطلاعی از آن ندارد.
- انتخاب کنش با بیشترین ارزش

$$A_t = \arg \max_a Q_t(a)$$

- j_a^t : جمع پاداش‌های انتخاب کنش a تا پیش از زمان t
- N_a^t : تعداد دفعات انتخاب کنش a تا پیش از زمان t

آن‌گاه:

$$\forall a, a \in \{1:k\}: Q_t(a) = \frac{j_a^t}{N_a^t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}} a$$



روش میانگین نمونه

- برابر با بهره‌برداری از اطلاعات تاکنون
- امکان بیشتر بودن پاداش کنشی در کوتاه مدت
- در ی‌ت محیط غالباً تصادفی $\Leftarrow$ مقایسه‌ها آماری
- وجود روش‌های دیگر
- لزوماً بهترین روش نیست و وجود روش‌های دیگر برای تخمین مقدار تخمین ارزش - تمرین

سیاست حریرانه در مقابل حریرانه-اپیلون

سیاست حریرانه:

- در صورت آزمایش کنش‌ها در هر زمان کنشی دارای بیشترین ارزش تخمینی
- انتخاب چنین عملی $\Leftarrow$ سیاست حریرانه
- انتخاب کنش با بیشترین ارزش (بهره‌برداری)

$$A_t = \arg \max_a Q_t(a)$$

- بهره‌برداری از دانش فعلی ارزش کنش‌ها
- عدم انتخاب بیشترین مقدار
- غیرحریرانه
- در وادی کاوش
- ایجاد امکان بهبود تقریب از ارزش کنش‌ها

سیاست حریرانه در مقابل حریرانه-اپیلون

سیاست حریرانه:

- در صورت آزمایش کنش‌ها در هر زمان کنشی دارای بیشترین ارزش تخمینی
- انتخاب چنین عملی $\Leftarrow$ سیاست حریرانه
- انتخاب کنش با بیشترین ارزش (بهره‌برداری)

$$A_t = \arg \max_a Q_t(a)$$

- بهره‌برداری از دانش فعلی ارزش کنش‌ها
- عدم انتخاب بیشترین مقدار
- غیرحریرانه
- در وادی کاوش
- ایجاد امکان بهبود تقریب از ارزش کنش‌ها

بهره‌برداری قدمی مناسب برای بیشینه‌سازی ارزش در تک قدم

کاوش امکان تولید پاداش کل بزرگتر در بلندمدت

- پاداش کمتر کاوش در کوتاه‌مدت ولی پاداش بیشتر در بلندمدت (هنگام یافتن کنش‌های با ارزش بالاتر) (بار معنای -)

همیشه یک پای ماجرا کم است!؟

- عدم امکان کاوش و بهره‌برداری با اجرای کنشی واحد و منفرد
- $\Leftarrow$ تعارض بین کاوش و بهره‌بردن

سیاست حریصانه در مقابل حریصانه-اپسیلون

انتخاب بین کاوش و بهره‌برداری

- پیچیده
- بسته به
- ارزش دقیق تخمین‌ها
- عدم قطعیت‌ها
- تعداد اقدام باقی‌مانده

وجود روش‌های فراوان، و پرچم و خم جهت تعادل بین بهره و کاوش در راهزن k -اهرمی

- اما راهی خونی
- بیشتر روش‌ها دارای فرضیات قوی دربارهٔ مانایی و دانش قبلی
- هزینه محدود یا بهینگی دارای کمترین معنی در مواجهه با تغییر فرضیات

سیاست حریصانه در مقابل حریصانه-اپسیلون

سیاست حریصانه:

▪ انتخاب کنش با بیشترین ارزش (بهره‌برداری)

$$A_t = \arg \max_a Q_t(a)$$

با میل t به بی‌نهایت، میل $Q_t(a)$ به $q_*(a)$

سیاست حریصانه با اپسیلون:

▪ انتخاب کنش با بیشترین ارزش با احتمال $1 - \epsilon$ (بهره‌برداری) و انتخاب تصادفی با احتمال ϵ (کاوش)

$$A_t = \begin{cases} \arg \max_a Q_t(a), & p \in [0, 1 - \epsilon) \\ a \sim U(\{a_1, a_2, \dots, a_k\}), & p \in [1 - \epsilon, 1] \end{cases}$$

احتمال انتخاب کنش عمل حریصانه در روش مذکور

$$P = 1 - \epsilon + \epsilon \times \frac{1}{k}$$

سیاست حریرانه در مقابل حریرانه-اوسیلون

روش‌های ارزش‌کنش یا ارزش-کنش

$V(s)$ تابع ارزش حالت

$Q(s,a)$ تابع ارزش حالت-کنش

مسئله ماشین سکه

- صرفاً یک حالت و چندین کنش

- پس $Q(s,a) \rightarrow Q(a)$

- کاربردهای عملی چون مدل‌سازی رفتار انسانی، سیستم‌های پیشنهاد دهنده، و آزمایشات بالینی

سیاست حریصانه در مقابل حریصانه-اپسیلون

الگوریتم حریصانه حالت خاصی از الگوریتم حریصانه با اپسیلون
▪ اپسیلون برابر صفر

محیط مانا

▪ در ابتدا انتخاب اپسیلون بزرگ و سپس میل تدریجی مقدار اپسیلون به صفر

محیط پویا (غیرمانا)

▪ اپسیلون غیر صفر

امکان محاسبه ارزش‌ها با حافظه کمتر

محاسبه بازگشتی ارزش‌ها

Q_{n+1} ارزش کنش a پس از n -بار انتخاب در t تکرار آزمایش

$$Q_{n+1} = \frac{1}{n} \sum_{i=1}^n R_i$$

محاسبه بازگشتی ارزش‌ها

Q_{n+1} ارزش کنش a پس از n -بار انتخاب در t تکرار آزمایش

$$Q_{n+1} = \frac{1}{n} \sum_{i=1}^n R_i$$

نیاز به حافظه زیاد- چاره‌ای برای بهینگی

محاسبه بازگشتی ارزش‌ها

Q_{n+1} ارزش کنش a پس از n -بار انتخاب در t تکرار آزمایش

$$Q_{n+1} = \frac{1}{n} \sum_{i=1}^n R_i$$

نیاز به حافظه زیاد- چاره‌ای برای بهینگی
▪ استفاده از رابطه بازگشتی برای کاهش پیچیدگی حافظه

محاسبه بازگشتی ارزش‌ها

Q_{n+1} ارزش کنش a پس از n -بار انتخاب در t تکرار آزمایش

نیاز به حافظه زیاد- چاره‌ای برای بهیمنگی

▪ استفاده از رابطه بازگشتی برای کاهش پیچیدگی حافظه

$$Q_{n+1} = \frac{1}{n} \sum_{i=1}^n R_i$$

محاسبه بازگشتی ارزش‌ها

Q_{n+1} ارزش کنش a پس از n -بار انتخاب در t تکرار آزمایش

نیاز به حافظه زیاد- چاره‌ای برای بهیمنگی

▪ استفاده از رابطه بازگشتی برای کاهش پیچیدگی حافظه

$$Q_{n+1} = \frac{1}{n} \sum_{i=1}^n R_i = \frac{1}{n} \left(R_n + \sum_{i=1}^{n-1} R_i \right)$$

محاسبه بازگشتی ارزش‌ها

Q_{n+1} ارزش کنش a پس از n -بار انتخاب در t تکرار آزمایش

نیاز به حافظه زیاد- چاره‌ای برای بهیمنگی

▪ استفاده از رابطه بازگشتی برای کاهش پیچیدگی حافظه

$$Q_{n+1} = \frac{1}{n} \sum_{i=1}^n R_i = \frac{1}{n} \left(R_n + \sum_{i=1}^{n-1} R_i \right) = \frac{1}{n} \left(R_n + (n-1) \frac{1}{n-1} \sum_{i=1}^{n-1} R_i \right)$$

محاسبه بازگشتی ارزش‌ها

Q_{n+1} ارزش کنش a پس از n -بار انتخاب در t تکرار آزمایش

نیاز به حافظه زیاد- چاره‌ای برای بهیمنگی

▪ استفاده از رابطه بازگشتی برای کاهش پیچیدگی حافظه

$$\begin{aligned} Q_{n+1} &= \frac{1}{n} \sum_{i=1}^n R_i = \frac{1}{n} \left(R_n + \sum_{i=1}^{n-1} R_i \right) = \frac{1}{n} \left(R_n + (n-1) \frac{1}{n-1} \sum_{i=1}^{n-1} R_i \right) \\ &= \frac{1}{n} \left(R_n + (n-1) \underbrace{\frac{1}{n-1} \sum_{i=1}^{n-1} R_i}_{Q_n} \right) \end{aligned}$$

محاسبه بازگشتی ارزش‌ها

Q_{n+1} ارزش کنش a پس از n -بار انتخاب در t تکرار آزمایش

نیاز به حافظه زیاد- چاره‌ای برای بهیمنگی

▪ استفاده از رابطه بازگشتی برای کاهش پیچیدگی حافظه

$$\begin{aligned} Q_{n+1} &= \frac{1}{n} \sum_{i=1}^n R_i = \frac{1}{n} \left(R_n + \sum_{i=1}^{n-1} R_i \right) = \frac{1}{n} \left(R_n + (n-1) \frac{1}{n-1} \sum_{i=1}^{n-1} R_i \right) \\ &= \frac{1}{n} \left(R_n + (n-1) \underbrace{\frac{1}{n-1} \sum_{i=1}^{n-1} R_i}_{Q_n} \right) = \frac{1}{n} (R_n + (n-1)Q_n) \end{aligned}$$

محاسبه بازگشتی ارزش‌ها

Q_{n+1} ارزش کنش a پس از n -بار انتخاب در t تکرار آزمایش

نیاز به حافظه زیاد- چاره‌ای برای بهیمنگی

▪ استفاده از رابطه بازگشتی برای کاهش پیچیدگی حافظه

$$\begin{aligned} Q_{n+1} &= \frac{1}{n} \sum_{i=1}^n R_i = \frac{1}{n} \left(R_n + \sum_{i=1}^{n-1} R_i \right) = \frac{1}{n} \left(R_n + (n-1) \frac{1}{n-1} \sum_{i=1}^{n-1} R_i \right) \\ &= \frac{1}{n} \left(R_n + (n-1) \underbrace{\frac{1}{n-1} \sum_{i=1}^{n-1} R_i}_{Q_n} \right) = \frac{1}{n} (R_n + (n-1)Q_n) = \frac{1}{n} (R_n + nQ_n - Q_n) \end{aligned}$$

محاسبه بازگشتی ارزش‌ها

Q_{n+1} ارزش کنش a پس از n -بار انتخاب در t تکرار آزمایش

نیاز به حافظه زیاد- چاره‌ای برای بهینگی

▪ استفاده از رابطه بازگشتی برای کاهش پیچیدگی حافظه

$$\begin{aligned} Q_{n+1} &= \frac{1}{n} \sum_{i=1}^n R_i = \frac{1}{n} \left(R_n + \sum_{i=1}^{n-1} R_i \right) = \frac{1}{n} \left(R_n + (n-1) \frac{1}{n-1} \sum_{i=1}^{n-1} R_i \right) \\ &= \frac{1}{n} \left(R_n + (n-1) \underbrace{\frac{1}{n-1} \sum_{i=1}^{n-1} R_i}_{Q_n} \right) = \frac{1}{n} (R_n + (n-1)Q_n) = \frac{1}{n} (R_n + nQ_n - Q_n) \end{aligned}$$

$$Q_{n+1} = Q_n + \frac{1}{n} [R_n - Q_n]$$

محاسبه بازگشتی ارزش‌ها

Q_{n+1} ارزش کنش a پس از n -بار انتخاب در t تکرار آزمایش

نیاز به حافظه زیاد- چاره‌ای برای بهیمنگی

▪ استفاده از رابطه بازگشتی برای کاهش پیچیدگی حافظه

$$\begin{aligned} Q_{n+1} &= \frac{1}{n} \sum_{i=1}^n R_i = \frac{1}{n} \left(R_n + \sum_{i=1}^{n-1} R_i \right) = \frac{1}{n} \left(R_n + (n-1) \frac{1}{n-1} \sum_{i=1}^{n-1} R_i \right) \\ &= \frac{1}{n} \left(R_n + (n-1) \underbrace{\frac{1}{n-1} \sum_{i=1}^{n-1} R_i}_{Q_n} \right) = \frac{1}{n} (R_n + (n-1)Q_n) = \frac{1}{n} (R_n + nQ_n - Q_n) \end{aligned}$$

$$Q_{n+1} = Q_n + \frac{1}{n} [R_n - Q_n]$$

[تقریب، قدیم - هدف] اندازه، قدم + تقریب، قدیم = تقریب، جدید

خطا $R_n - Q_n$

به دنبال یافتن تخمینی از هدف

محاسبه بازگشتی یا افزایشی ارزش‌ها

Q_{n+1} ارزش کنش a پس از n -بار انتخاب در t تکرار آزمایش

نیاز به حافظه زیاد- چاره‌ای برای بهینگی
▪ استفاده از رابطه بازگشتی برای کاهش پیچیدگی حافظه

$$\begin{aligned} Q_{n+1} &= \frac{1}{n} \sum_{i=1}^n R_i = \frac{1}{n} (R_n + \sum_{i=1}^{n-1} R_i) = \frac{1}{n} \left(R_n + (n-1) \frac{1}{n-1} \sum_{i=1}^{n-1} R_i \right) \\ &= \frac{1}{n} \left(R_n + (n-1) \underbrace{\frac{1}{n-1} \sum_{i=1}^{n-1} R_i}_{Q_n} \right) = \frac{1}{n} (R_n + (n-1)Q_n) = \frac{1}{n} (R_n + nQ_n - Q_n) \end{aligned}$$

$$Q_{n+1} = Q_n + \frac{1}{n} [R_n - Q_n]$$

[تقریب، قدیم - هدف] اندازه، قدم + تقریب، قدیم = تقریب، جدید

- خطا $R_n - Q_n$
- به دنبال یافتن تخمینی از هدف
- روش بازگشتی یا افزایشی «میانگین نمونه»

سیاست حریرانه

شباهت با روش‌های بهینه‌سازی هموار و گردایانی و دیگر انواع یادگیری

نیاز به شرط اولیه

▪ در بیشتر اوقات $Q_1 = 0$

▪ در صورت وجود شناخت امکان تعریف مقدار خاصی برای بعضی از کنش‌ها

سیاست حریم

دلیل میانگین گیری

▪ فرض: پاداش ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

سیاست حریصانه

دلیل میانگین گیری

▪ فرض: پاداش ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$

سیاست حریرانه

دلیل میانگین گیری

▪ فرض: پاداش ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$
$$E(\hat{\mu})$$

سیاست حریرانه

دلیل میانگین گیری

▪ فرض: پاداش ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$
$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right)$$

سیاست حریرانه

دلیل میانگین گیری

▪ فرض: پاداش ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$
$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n} \sum_{i=1}^n E(r_i)$$

سیاست حریرانه

دلیل میانگین گیری

▪ فرض: پاداش ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$
$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n} \sum_{i=1}^n E(r_i) = \frac{1}{n} \sum_{i=1}^n E(r_i)$$

سیاست حریصانه

دلیل میانگین گیری

▪ فرض: پاداش ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$
$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n} \sum_{i=1}^n E(r_i) = \frac{1}{n} \sum_{i=1}^n \underbrace{E(r_i)}_{\mu}$$

سیاست حریصانه

دلیل میانگین گیری

▪ فرض: پاداش ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$
$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n} \sum_{i=1}^n E(r_i) = \frac{1}{n} \sum_{i=1}^n \underbrace{E(r_i)}_{\mu} = \frac{1}{n} \sum_{i=1}^n \mu$$

سیاست حریرانه

دلیل میانگین گیری

▪ فرض: پاداش ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$
$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n} \sum_{i=1}^n E(r_i) = \frac{1}{n} \sum_{i=1}^n \underbrace{E(r_i)}_{\mu} = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \times n\mu$$

سیاست حریصانه

دلیل میانگین گیری

▪ فرض: پاداش ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$
$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n} \sum_{i=1}^n E(r_i) = \frac{1}{n} \sum_{i=1}^n \underbrace{E(r_i)}_{\mu} = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \times n\mu = \mu$$

سیاست حریصانه

دلیل میانگین گیری

▪ فرض: پاداش ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$
$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n} \sum_{i=1}^n E(r_i) = \frac{1}{n} \sum_{i=1}^n \underbrace{E(r_i)}_{\mu} = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \times n\mu = \mu$$
$$Var(\hat{\mu})$$

سیاست حریصانه

دلیل میانگین گیری

▪ فرض: پاداش ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$
$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n} \sum_{i=1}^n E(r_i) = \frac{1}{n} \sum_{i=1}^n \underbrace{E(r_i)}_{\mu} = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \times n\mu = \mu$$
$$Var(\hat{\mu}) = Var\left(\frac{1}{n} \sum_{i=1}^n r_i\right)$$

سیاست حریصانه

دلیل میانگین گیری

▪ فرض: پاداش ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$
$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n} \sum_{i=1}^n E(r_i) = \frac{1}{n} \sum_{i=1}^n \underbrace{E(r_i)}_{\mu} = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \times n\mu = \mu$$
$$Var(\hat{\mu}) = Var\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n^2} \sum_{i=1}^n Var(r_i)$$

سیاست حریصانه

دلیل میانگین گیری

▪ فرض: پاداش ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$
$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n} \sum_{i=1}^n E(r_i) = \frac{1}{n} \sum_{i=1}^n \underbrace{E(r_i)}_{\mu} = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \times n\mu = \mu$$
$$Var(\hat{\mu}) = Var\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n^2} \sum_{i=1}^n Var(r_i) = \frac{1}{n^2} \sum_{i=1}^n \underbrace{Var(r_i)}_{\sigma^2}$$

سیاست حریصانه

دلیل میانگین گیری

▪ فرض: پاداش ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$

$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n} \sum_{i=1}^n E(r_i) = \frac{1}{n} \sum_{i=1}^n \underbrace{E(r_i)}_{\mu} = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \times n\mu = \mu$$

$$Var(\hat{\mu}) = Var\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n^2} \sum_{i=1}^n Var(r_i) = \frac{1}{n^2} \sum_{i=1}^n \underbrace{Var(r_i)}_{\sigma^2} = \frac{1}{n^2} \sum_{i=1}^n \sigma^2$$

سیاست حریصانه

دلیل میانگین گیری

▪ فرض: پاداش‌ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$

$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n} \sum_{i=1}^n E(r_i) = \frac{1}{n} \sum_{i=1}^n \underbrace{E(r_i)}_{\mu} = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \times n\mu = \mu$$

$$Var(\hat{\mu}) = Var\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n^2} \sum_{i=1}^n Var(r_i) = \frac{1}{n^2} \sum_{i=1}^n \underbrace{Var(r_i)}_{\sigma^2} = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{n\sigma^2}{n^2}$$

سیاست حریصانه

دلیل میانگین گیری

▪ فرض: پاداش‌ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$
$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n} \sum_{i=1}^n E(r_i) = \frac{1}{n} \sum_{i=1}^n \underbrace{E(r_i)}_{\mu} = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \times n\mu = \mu$$
$$Var(\hat{\mu}) = Var\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n^2} \sum_{i=1}^n Var(r_i) = \frac{1}{n^2} \sum_{i=1}^n \underbrace{Var(r_i)}_{\sigma^2} = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

سیاست حریصانه

دلیل میانگین گیری

▪ فرض: پاداش‌ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$
$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n} \sum_{i=1}^n E(r_i) = \frac{1}{n} \sum_{i=1}^n \underbrace{E(r_i)}_{\mu} = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \times n\mu = \mu$$
$$Var(\hat{\mu}) = Var\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n^2} \sum_{i=1}^n Var(r_i) = \frac{1}{n^2} \sum_{i=1}^n \underbrace{Var(r_i)}_{\sigma^2} = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$\xrightarrow[n \rightarrow \infty]{} Var(\hat{\mu})$

سیاست حریصانه

دلیل میانگین گیری

▪ فرض: پاداش‌ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$
$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n} \sum_{i=1}^n E(r_i) = \frac{1}{n} \sum_{i=1}^n \underbrace{E(r_i)}_{\mu} = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \times n\mu = \mu$$
$$Var(\hat{\mu}) = Var\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n^2} \sum_{i=1}^n Var(r_i) = \frac{1}{n^2} \sum_{i=1}^n \underbrace{Var(r_i)}_{\sigma^2} = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$
$$\xrightarrow{n \rightarrow \infty} Var(\hat{\mu}) = 0$$

سیاست حریصانه

دلیل میانگین گیری

▪ فرض: پاداش‌ها دارای مقادیر توزیع نرمال $R \sim N(\mu, \sigma^2)$

n بار نمونه برداری مستقل

$$\hat{\mu} = E(R) \approx \frac{1}{n} \sum_{i=1}^n r_i$$
$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n} \sum_{i=1}^n E(r_i) = \frac{1}{n} \sum_{i=1}^n \underbrace{E(r_i)}_{\mu} = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \times n\mu = \mu$$
$$Var(\hat{\mu}) = Var\left(\frac{1}{n} \sum_{i=1}^n r_i\right) = \frac{1}{n^2} \sum_{i=1}^n Var(r_i) = \frac{1}{n^2} \sum_{i=1}^n \underbrace{Var(r_i)}_{\sigma^2} = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$\xrightarrow{n \rightarrow \infty} Var(\hat{\mu}) = 0$

▪ نتیجه نمونه‌گیری زیاد وردائی را صفر می‌کند و در نتیجه عدم قطعیت کاهش می‌یابد.

الگوریتم حل مسئله ماشین سکه چند بازو با سیاست حریصانه با اپسیلون

مقداردهی اولیه برای $a = 1:k$

$$Q(a) \leftarrow 0 \quad \blacksquare$$

$$N(a) \leftarrow 0 \quad \blacksquare$$

حلقه برای ابد

$$A = \begin{cases} \arg \max_a Q_t(a), p \in [0, 1 - \epsilon] \\ a_{\text{تصادفی}}, p \in [1 - \epsilon, 1] \end{cases} \quad \blacksquare$$

$$R \leftarrow \text{rahzan}(A) \quad \blacksquare$$

$$N(A) \leftarrow N(A) + 1 \quad \blacksquare$$

$$Q(A) \leftarrow Q(A) + \frac{1}{N(A)} [R - Q(A)] \quad \blacksquare$$

انتهای الگوریتم

شبیه سازی

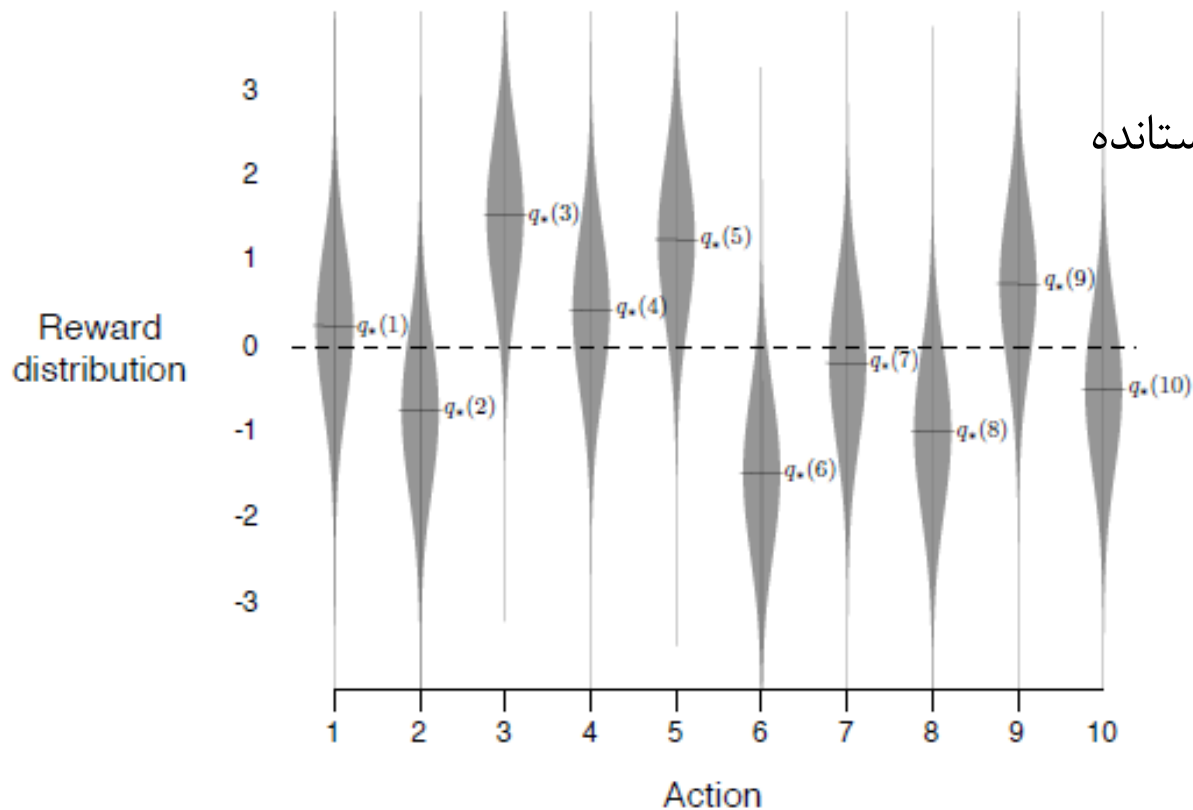
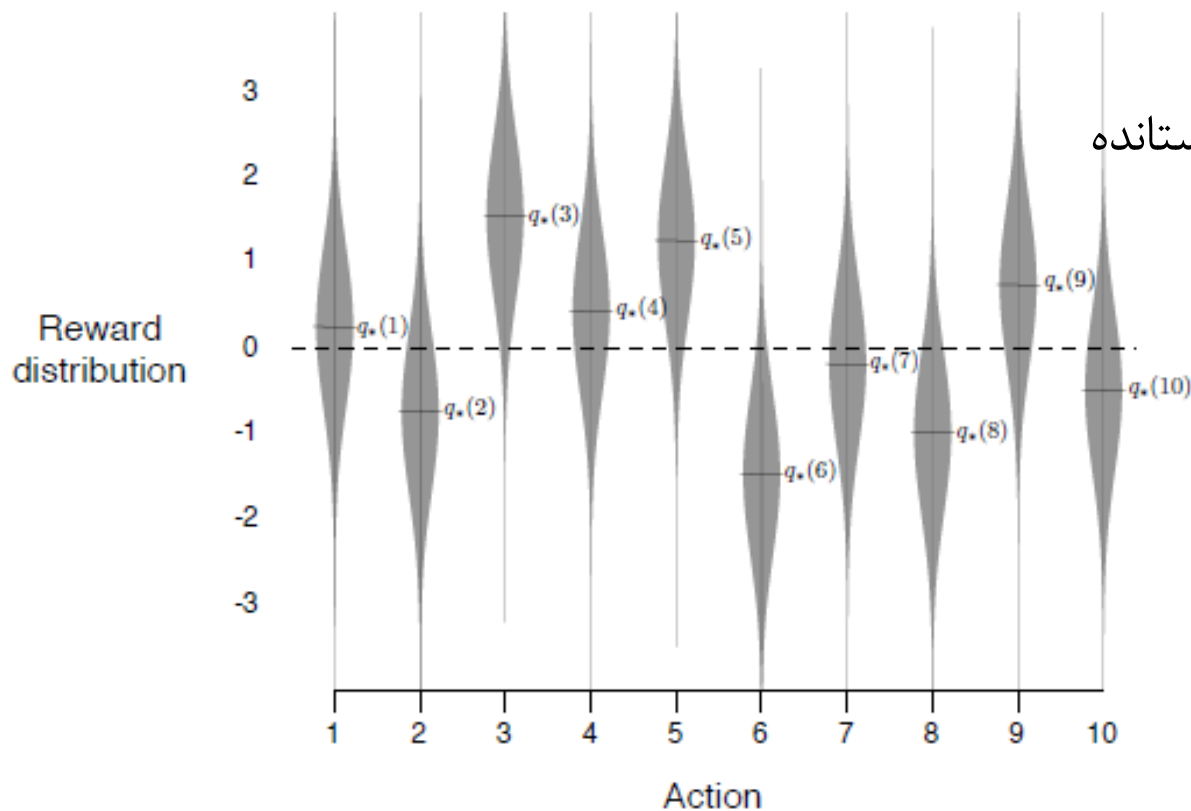


Figure 2.1: An example bandit problem from the 10-armed testbed. The true value $q_*(a)$ of each of the ten actions was selected according to a normal distribution with mean zero and unit variance, and then the actual rewards were selected according to a mean $q_*(a)$ unit variance normal distribution, as suggested by these gray distributions.

شبیه سازی



ده کنش روی کارهایی با ارزش های دارای توزیع نرمال استاندارد

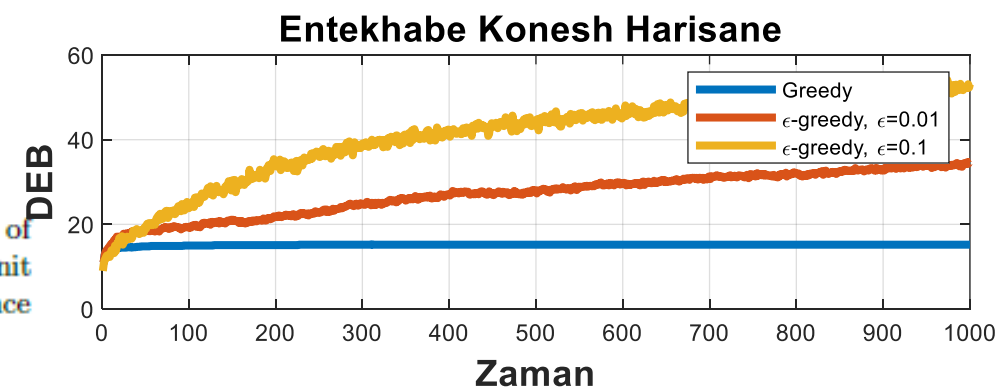
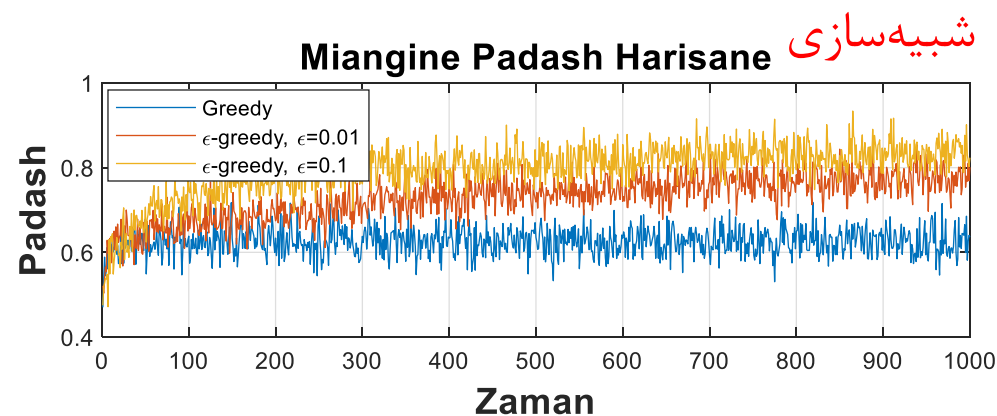


Figure 2.1: An example bandit problem from the 10-armed testbed. The true value $q_*(a)$ of each of the ten actions was selected according to a normal distribution with mean zero and unit variance, and then the actual rewards were selected according to a mean $q_*(a)$ unit variance normal distribution, as suggested by these gray distributions.

شبیه‌سازی

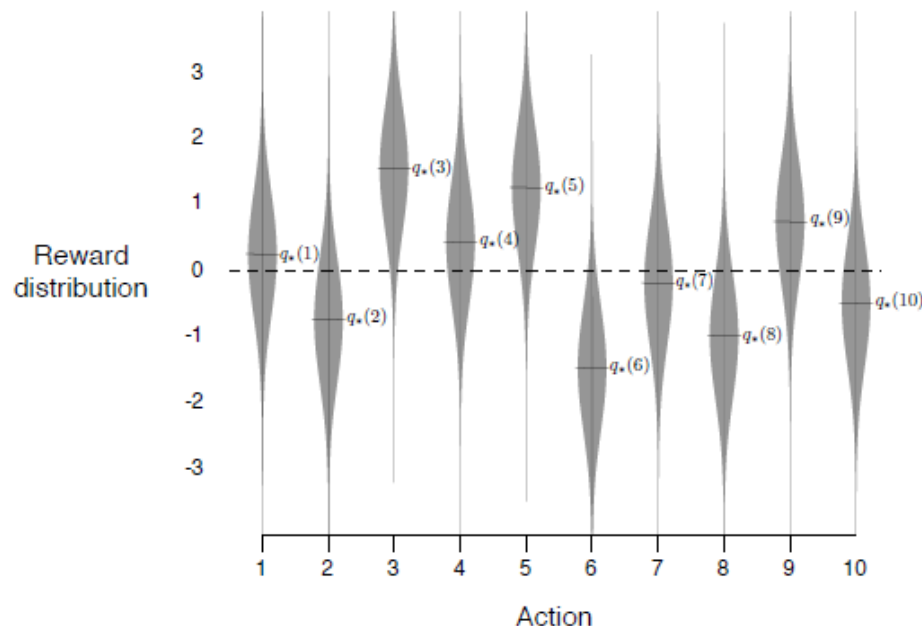


Figure 2.1: An example bandit problem from the 10-armed testbed. The true value $q_*(a)$ of each of the ten actions was selected according to a normal distribution with mean zero and unit variance, and then the actual rewards were selected according to a mean $q_*(a)$ unit variance normal distribution, as suggested by these gray distributions.

ده کنش روی کارهایی با ارزش‌های دارای توزیع نرمال استاندارد

- تاکنون مناسب برای مسئله راهزن مانا
- در عمل بیشتر محیط‌ها نامانا
- ضرورت بررسی بهره‌برداری و کاوش در چنین محیطی

محیط‌های پویا (دارای تغییر مداوم)

- غیرمانا
- عامل باید هم‌پای آن مدام در حال یادگیری
- در $\frac{1}{n}$ $Q_{n+1} = Q_n + \frac{1}{n}[R_n - Q_n]$ موجب عدم تغییر در زمان‌های طولانی
- نامناسب بودن چنین عدم تغییری در محیط پویا
- امر مطلوب در وزن‌دهی به پاداش‌ها؟
- وزن‌دهی بیشتر به پاداش‌های فعلی
- جانشینی با ضریب دیگر
- معروف به طول قدم یا سرعت یادگیری

$$Q_{n+1} = Q_n + \alpha[R_n - Q_n]$$

الگوریتم اندازه-گام «ثابت» یا ضریب ثابت

- ادامه فرایند یادگیری

معادله ضريب ثابت

$$Q_{n+1}$$

معادله ضريب ثابت

$$Q_{n+1} = Q_n + \alpha[R_n - Q_n]$$

معادله ضریب ثابت

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha[R_n - Q_n] \\ &= \alpha R_n + (1 - \alpha)Q_n \end{aligned}$$

معادله ضریب ثابت

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha[R_n - Q_n] \\ &= \alpha R_n + (1 - \alpha)Q_n \\ &= \alpha R_n + (1 - \alpha)[\alpha R_{n-1} + (1 - \alpha)Q_{n-1}] \end{aligned}$$

معادله ضریب ثابت

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha[R_n - Q_n] \\ &= \alpha R_n + (1 - \alpha)Q_n \\ &= \alpha R_n + (1 - \alpha)[\alpha R_{n-1} + (1 - \alpha)Q_{n-1}] \\ &= \alpha R_n + (1 - \alpha)\alpha R_{n-1} + (1 - \alpha)^2 Q_{n-1} \end{aligned}$$

معادله ضریب ثابت

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha[R_n - Q_n] \\ &= \alpha R_n + (1 - \alpha)Q_n \\ &= \alpha R_n + (1 - \alpha)[\alpha R_{n-1} + (1 - \alpha)Q_{n-1}] \\ &= \alpha R_n + (1 - \alpha)\alpha R_{n-1} + (1 - \alpha)^2 Q_{n-1} \\ &= \alpha R_n + (1 - \alpha)\alpha R_{n-1} + (1 - \alpha)^2 \alpha R_{n-2} + \cdots + (1 - \alpha)^{n-1} \alpha R_1 + (1 - \alpha)^n Q_1 \end{aligned}$$

معادله ضریب ثابت

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha[R_n - Q_n] \\ &= \alpha R_n + (1 - \alpha)Q_n \\ &= \alpha R_n + (1 - \alpha)[\alpha R_{n-1} + (1 - \alpha)Q_{n-1}] \\ &= \alpha R_n + (1 - \alpha)\alpha R_{n-1} + (1 - \alpha)^2 Q_{n-1} \\ &= \alpha R_n + (1 - \alpha)\alpha R_{n-1} + (1 - \alpha)^2 \alpha R_{n-2} + \cdots + (1 - \alpha)^{n-1} \alpha R_1 + (1 - \alpha)^n Q_1 \end{aligned}$$

معادله ضریب ثابت

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha[R_n - Q_n] \\ &= \alpha R_n + (1 - \alpha)Q_n \\ &= \alpha R_n + (1 - \alpha)[\alpha R_{n-1} + (1 - \alpha)Q_{n-1}] \\ &= \alpha R_n + (1 - \alpha)\alpha R_{n-1} + (1 - \alpha)^2 Q_{n-1} \\ &= \alpha R_n + (1 - \alpha)\alpha R_{n-1} + (1 - \alpha)^2 \alpha R_{n-2} + \cdots + (1 - \alpha)^{n-1} \alpha R_1 + (1 - \alpha)^n Q_1 \\ &= (1 - \alpha)^n Q_1 + \sum_{i=1}^n \alpha(1 - \alpha)^{n-i} R_i \end{aligned}$$

معادله ضریب ثابت

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha[R_n - Q_n] \\ &= \alpha R_n + (1 - \alpha)Q_n \\ &= \alpha R_n + (1 - \alpha)[\alpha R_{n-1} + (1 - \alpha)Q_{n-1}] \\ &= \alpha R_n + (1 - \alpha)\alpha R_{n-1} + (1 - \alpha)^2 Q_{n-1} \\ &= \alpha R_n + (1 - \alpha)\alpha R_{n-1} + (1 - \alpha)^2 \alpha R_{n-2} + \cdots + (1 - \alpha)^{n-1} \alpha R_1 + (1 - \alpha)^n Q_1 \\ &= (1 - \alpha)^n Q_1 + \sum_{i=1}^n \alpha(1 - \alpha)^{n-i} R_i \end{aligned}$$

$$(1 - \alpha)^n + \sum_{i=1}^n \alpha(1 - \alpha)^{n-i} = ? = 1$$

▪ «میانگین وزنی»

Q_1 یا صفر یا مقدار تصادفی

معادله ضریب ثابت

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha[R_n - Q_n] \\ &= \alpha R_n + (1 - \alpha)Q_n \\ &= \alpha R_n + (1 - \alpha)[\alpha R_{n-1} + (1 - \alpha)Q_{n-1}] \\ &= \alpha R_n + (1 - \alpha)\alpha R_{n-1} + (1 - \alpha)^2 Q_{n-1} \\ &= \alpha R_n + (1 - \alpha)\alpha R_{n-1} + (1 - \alpha)^2 \alpha R_{n-2} + \dots + (1 - \alpha)^{n-1} \alpha R_1 + (1 - \alpha)^n Q_1 \\ &= (1 - \alpha)^n Q_1 + \sum_{i=1}^n \alpha(1 - \alpha)^{n-i} R_i \end{aligned}$$

$$(1 - \alpha)^n + \sum_{i=1}^n \alpha(1 - \alpha)^{n-i} = ? = 1$$

▪ «میانگین وزنی»

Q_1 یا صفر یا مقدار تصادفی

▪ در صورت $Q_1 = 0$ میانگین کل پاداش‌ها

معادله ضریب ثابت

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha[R_n - Q_n] \\ &= \alpha R_n + (1 - \alpha)Q_n \\ &= \alpha R_n + (1 - \alpha)[\alpha R_{n-1} + (1 - \alpha)Q_{n-1}] \\ &= \alpha R_n + (1 - \alpha)\alpha R_{n-1} + (1 - \alpha)^2 Q_{n-1} \\ &= \alpha R_n + (1 - \alpha)\alpha R_{n-1} + (1 - \alpha)^2 \alpha R_{n-2} + \cdots + (1 - \alpha)^{n-1} \alpha R_1 + (1 - \alpha)^n Q_1 \\ &= (1 - \alpha)^n Q_1 + \sum_{i=1}^n \alpha(1 - \alpha)^{n-i} R_i \end{aligned}$$

$$(1 - \alpha)^n + \sum_{i=1}^n \alpha(1 - \alpha)^{n-i} = ? = 1$$

▪ «میانگین وزنی»

Q_1 یا صفر یا مقدار تصادفی
▪ در صورت $Q_1 = 0$ میانگین کل پاداش‌ها

ارزش حالت معادل با ترکیب خطی وزن‌دار از پاداش‌هایی رخ داده در حالت مذکور

اهمیت بیشتر پاداش‌های نزدیکتر از لحاظ زمان

$$Q_{n+1} = (1 - \alpha)^n Q_1 + \sum_{i=1}^n \alpha (1 - \alpha)^{n-i} R_i$$

اهمیت بیشتر پاداش‌های نزدیکتر از لحاظ زمان

$$Q_{n+1} = (1 - \alpha)^n Q_1 + \sum_{i=1}^n \alpha (1 - \alpha)^{n-i} R_i$$

▪ $1 - \alpha$ ؟

اهمیت بیشتر پاداش‌های نزدیکتر از لحاظ زمان

$$Q_{n+1} = (1 - \alpha)^n Q_1 + \sum_{i=1}^n \alpha (1 - \alpha)^{n-i} R_i$$

• $1 - \alpha = 0$ ؟

اهمیت بیشتر پاداش‌های نزدیکتر از لحاظ زمان

$$Q_{n+1} = (1 - \alpha)^n Q_1 + \sum_{i=1}^n \alpha (1 - \alpha)^{n-i} R_i$$

▪ $1 - \alpha = 0$ ؟

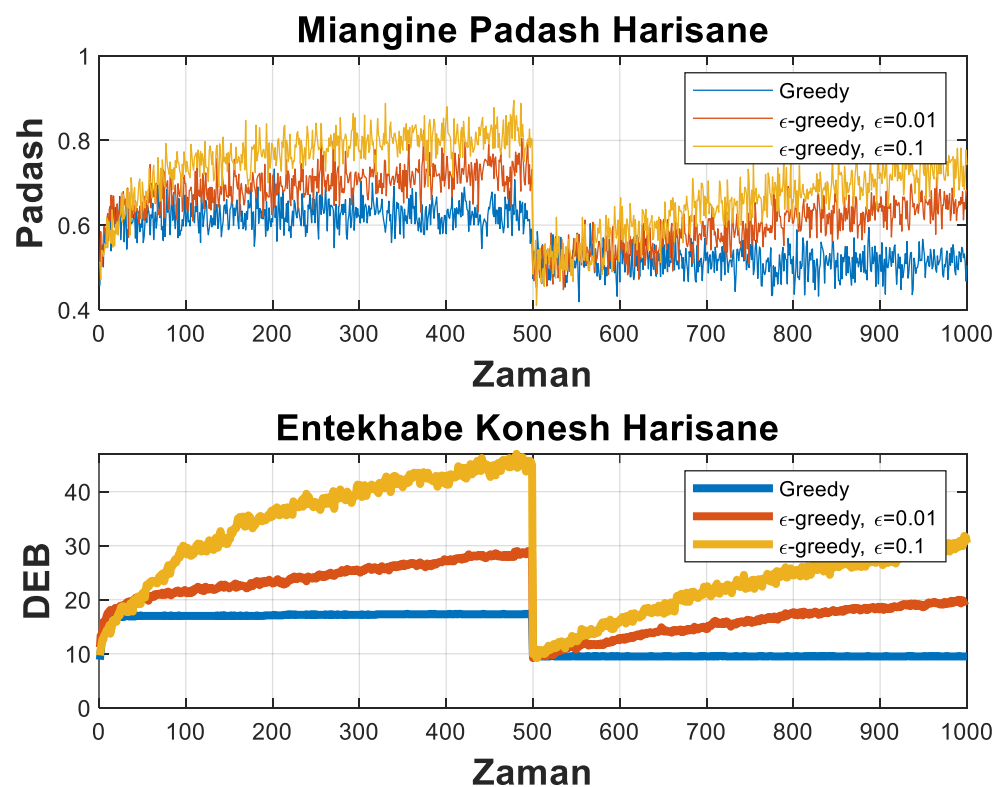
▪ میانگین وزن‌دار نمایی متاخر exponential recency-weighted average

وضعیت محیط پویا (روش میانگین نمونه) - مثال

شبیه سازی

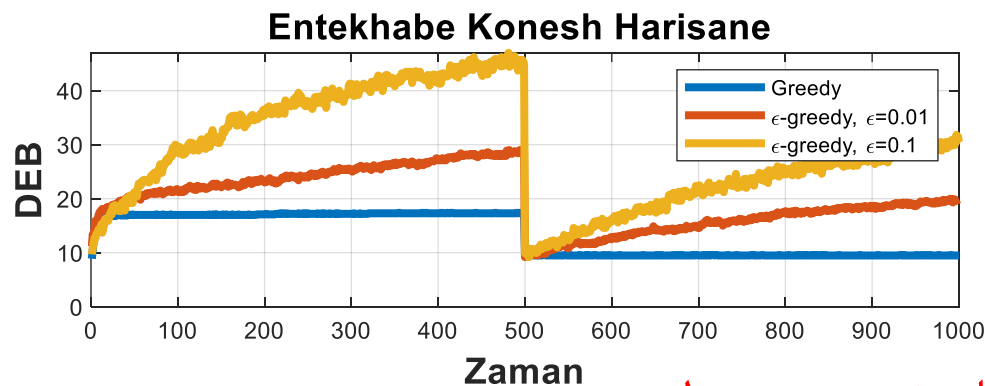
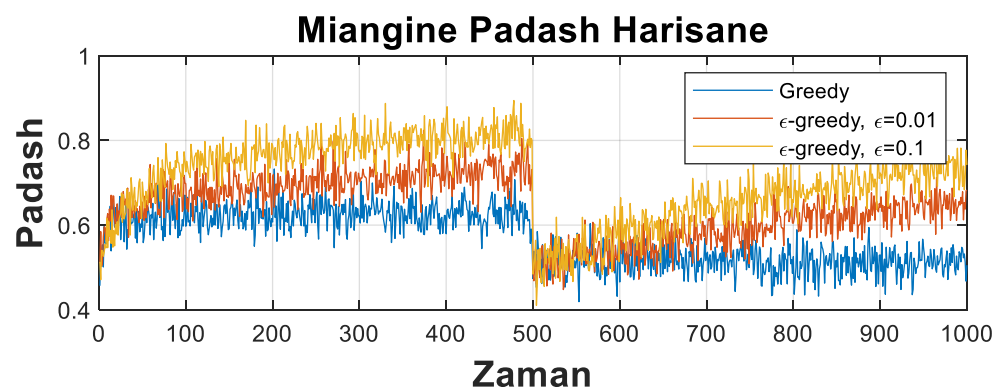
وضعیت محیط پویا (روش میانگین نمونه) - مثال

شبیه سازی



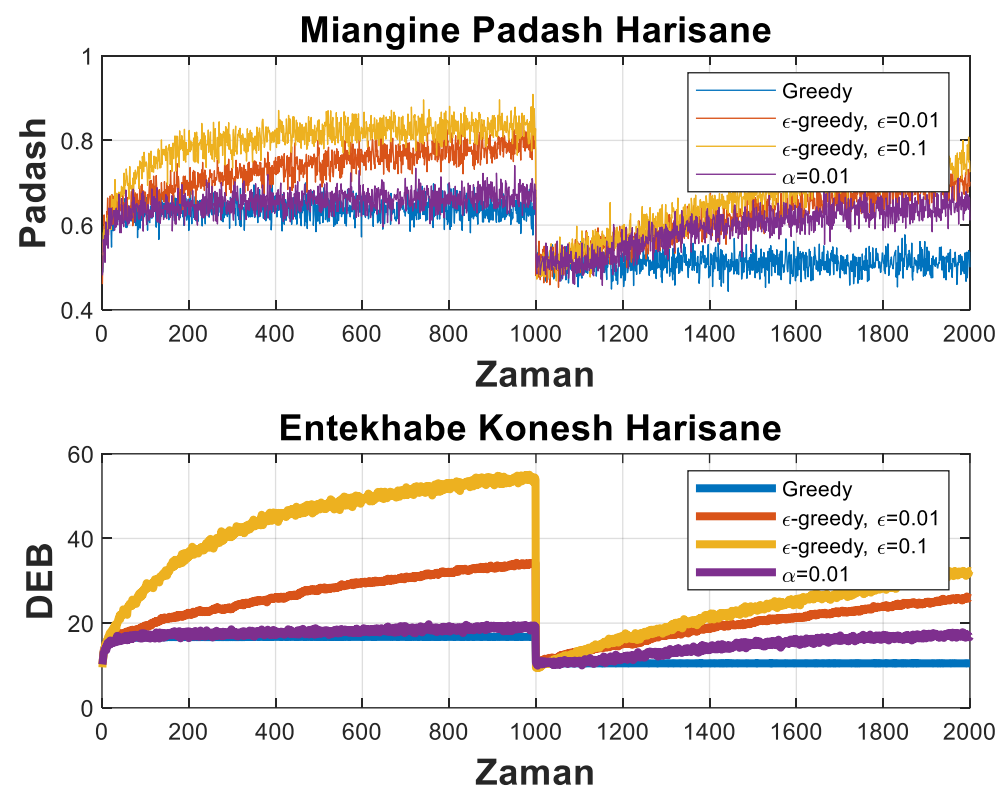
وضعیت محیط پویا (روش میانگین نمونه) - مثال

شبیه سازی



زمانبر بودن همگرایی پس از تغییر در محیط

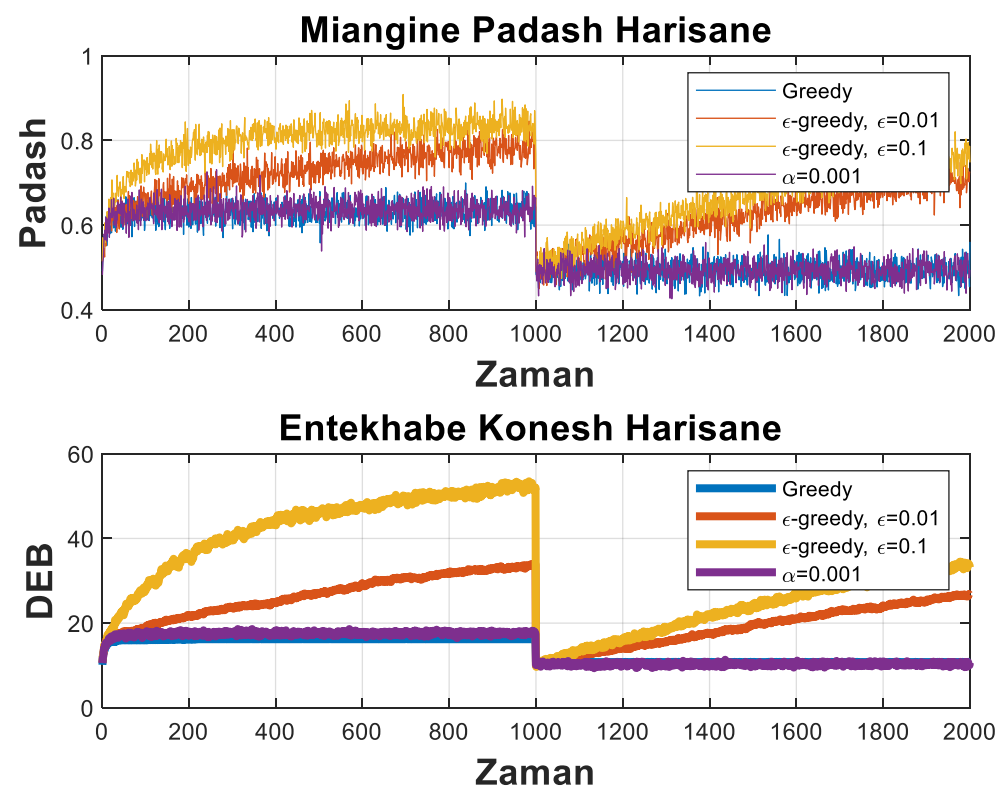
وضعیت محیط پویا (روش ضریب ثابت)-مثال



شبیه سازی

$$\alpha = 0.01$$

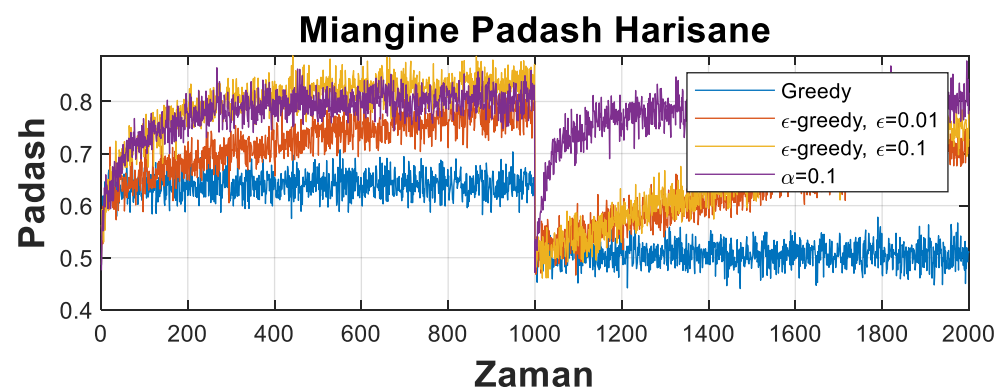
وضعیت محیط پویا (روش ضریب ثابت)-مثال



شبیه سازی

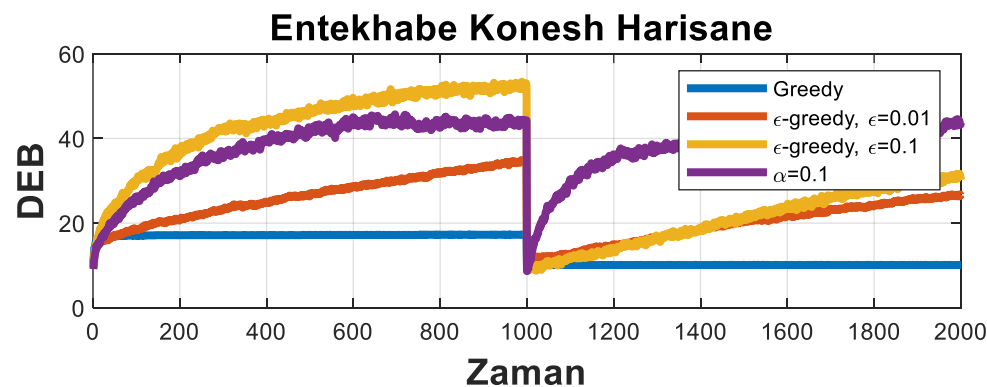
$$\alpha = 0.001$$

وضعیت محیط پویا (روش ضریب ثابت)-مثال



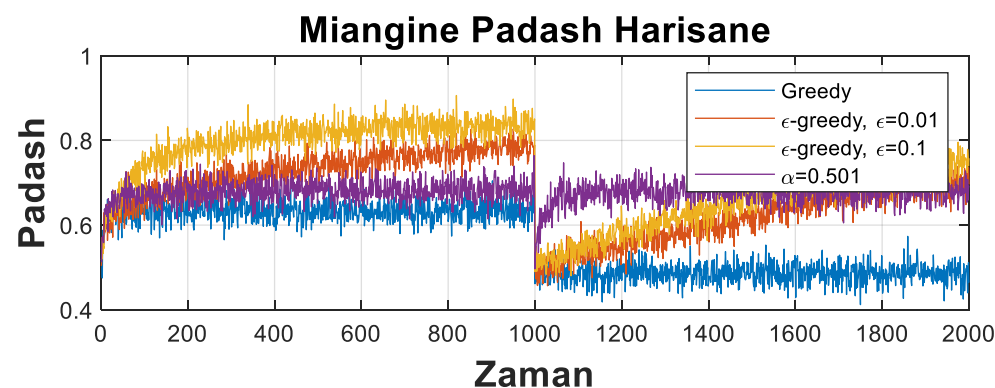
شبیه سازی

$$\alpha = 0.1$$



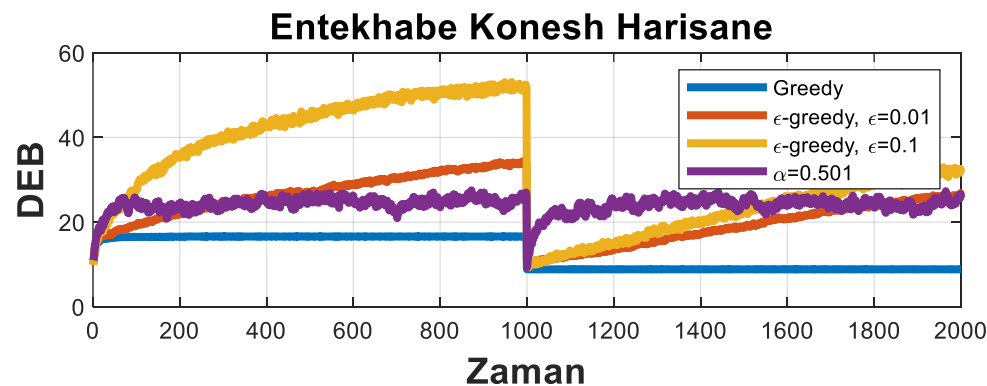
آلفای بزرگتر سرعت همگرایی بیشتر

وضعیت محیط پویا (روش ضریب ثابت)-مثال



شبیه‌سازی

$$\alpha = 0.5$$



آلفای بزرگتر سرعت همگرایی بیشتر

فراموشی پاداش‌های اولیه به مرور

▪ استفاده بر اساس مقادیر متاخر

▪ یاری‌رسان در محیط غیرمانا

الگوریتم با ضریب ثابت سرعت یادگیری حساس به شرایط اولیه

▪ هر چند حذف سوگیری در روش «میانگین نمونه»

$$Q_{n+1} = Q_n + \frac{1}{n} [R_n - Q_n]$$

بهتر بودن وابستگی آهنگ یادگیری به زمان در برخی مواقع

$$\alpha_n(a) = \frac{1}{n}$$

▪ میانگین نمونه

▪ $\Leftarrow$ در الگوریتم سرعت ثابت همواره سوگیری اما در الگوریتم میانگین نمونه در صورتی که همه نمونه‌ها یکبار انتخاب شوند سوگیری حذف خواهد شد

▪ Q_1 موجب کاوش زیاد در ابتدا

شروط کافی همگرایی الگوریتم

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty,$$
$$\sum_{n=1}^{\infty} \alpha_n^2(a) < \infty$$

- شرط اول به معنای واگرایی و حذف اثر شرایط اولیه در بلندمدت
- شرط دوم همگرایی به معنای ضمانت همگرایی در بلندمدت

$$\alpha_n = \frac{1}{n}$$

- رعایت هر دو شرط (شرط اول با انتگرال و شرط دوم با آنالیز فوریه)



شروط کافی همگرایی الگوریتم

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty,$$
$$\sum_{n=1}^{\infty} \alpha_n^2(a) < \infty$$

- شرط اول به معنای واگرایی و حذف اثر شرایط اولیه در بلندمدت
- شرط دوم همگرایی به معنای ضمانت همگرایی در بلندمدت

$$\alpha_n = \frac{1}{n}$$

- رعایت هر دو شرط (شرط اول با انتگرال و شرط دوم با آنالیز فوریه)

$$\alpha_n = \alpha$$

- برقرار نبودن یکی از شرطها؟
- به معنای
 - همگرای کامل نشدن تخمین
 - تغییر در لبیک به پاداش‌های دریافتی اخیر
- مناسب برای محیط نامانا و بالتبع اکثر مسائل یادگیری تقویتی
- ممکن است به بهینه میل نکند.

شروط کافی همگرایی الگوریتم

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty,$$
$$\sum_{n=1}^{\infty} \alpha_n^2(a) < \infty$$

- شرط اول به معنای واگرایی و حذف اثر شرایط اولیه در بلندمدت
- شرط دوم همگرایی به معنای ضمانت همگرایی در بلندمدت

دربارهٔ وزن‌های مرعی شروط دوگانه

- همگرایی کند یا نیاز به تنظیمات فراوان جهت دستیابی به سرعت همگرایی مناسب
- استفادهٔ غالب از چنین وزن‌هایی در نظر و نادر در عمل و کاربردها و پژوهش‌های تجربی

مقداردهی خوشبینانه

روش‌های معرفی شده تاکنون وابسته به تخمین ارزش-عمل آغازین یا $Q_1(a)$

- در آئین آمار دچار سوگیری به مقدار اولیه
- روش میانگین نمونه
- ناپدیدى سوگیری با انتخاب حداقل یکباره هر کنش
- روش میانگین وزن‌دار
- باقی‌ماندن سوگیری در طول زمان هرچند با کوچکتر شدن تاثیر

عیب می‌جمله چو گفتم هنرش نیز بگو

- در عمل خیلی مسئله نیست و حتی دارای فایده در پاره‌ای از موارد و شرایط
- هرچند نیاز به کاربر جهت مقداردهی
- راهی ساده برای افزودن دانش پیشین از سطح امید پاداش

مقداردهی خوشبینانه

روش‌های معرفی شده تاکنون وابسته به تخمین ارزش-عمل آغازین یا $Q_1(a)$

- در آئین آمار دچار سوگیری به مقدار اولیه
- روش میانگین نمونه
- ناپذیری سوگیری با انتخاب یکباره

افزایش سرعت جستجو با استفاده از مقداردهی خوشبینانه

- شرط اولیه ارزش حالت‌ها بیشتر از بیشترین مقدار ممکن برای آن‌ها
- موجب واریسی تمامی کنش‌ها و باقی نگذاشتن کنش آزمون‌نشده در ابتدا
- صرفاً مناسب در محیط‌های مانا
- به دلیل موقتی بودن کاوش

▪ زیرا که Q_1 در روش ضریب ثابت در مراحل بعد $Q_1(1 - \alpha)^n$

- موجب سوگیری
- اما حذف در حالت مانا در طی یکی دو مرحله

امکان استفاده از مکاشفه‌هایی برای حساب شرایط اولیه خوش‌بینانه در محیط مانا

مقداردهی خوشبینانه

فرض توزیع پاداش $R \sim N(\mu, \sigma^2)$ تبعیت می کند

▪ لاجرم در بیش از ۹۹ درصد مواقع $R \in [\mu - 3\sigma, \mu + 3\sigma]$ در روش میانگین نمونه

$$n = 1: Q_2 = R_1$$

نیاز به انتخاب Q_1 به ازای کمتر از مقدار پاداش واقعی

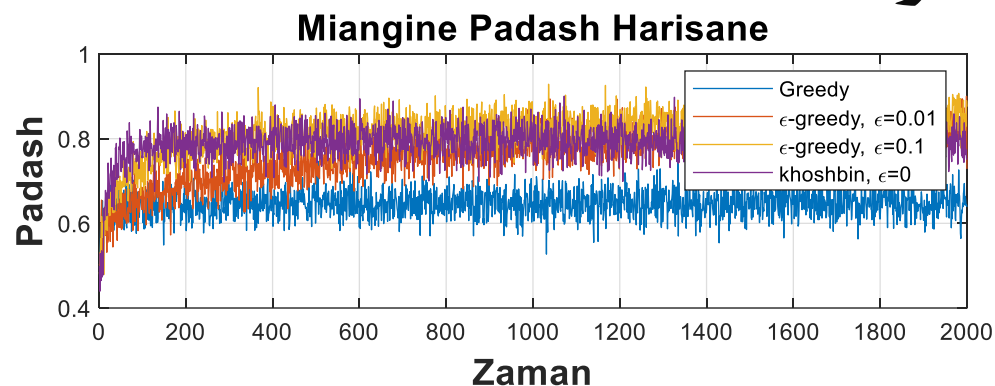
$$Q_1 \approx \mu + 3\sigma + c$$

در کد خود ۵ قرار دهیم.

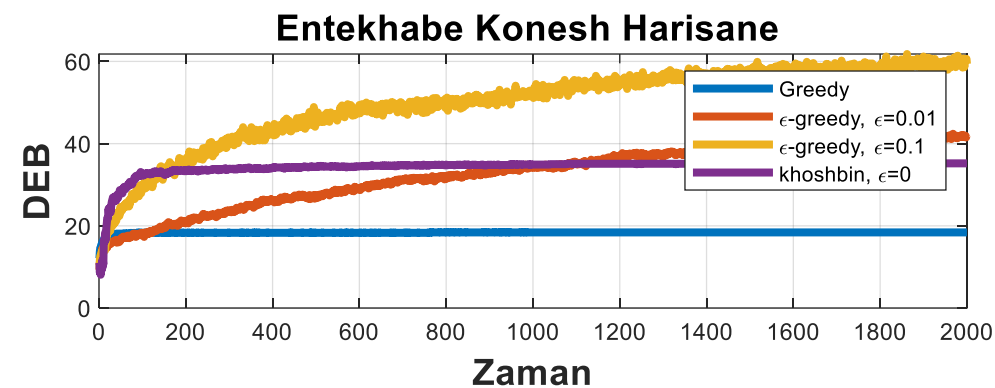
مقداردهی خوشبینانه

شبیه سازی

مقداردهی خوشبینانه

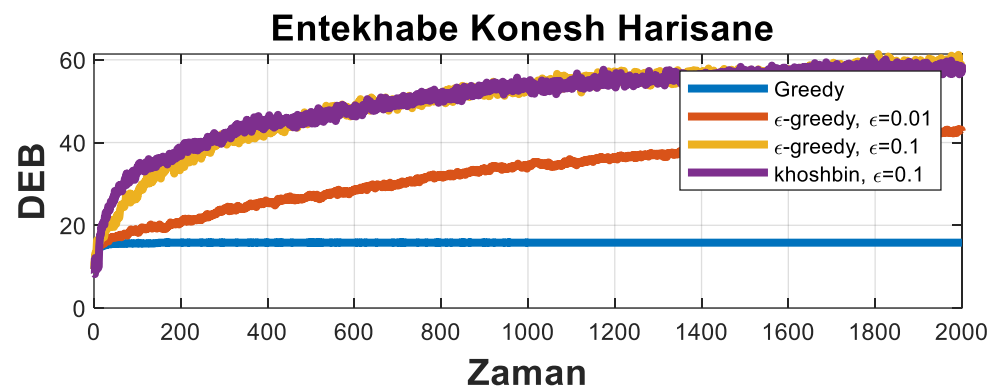
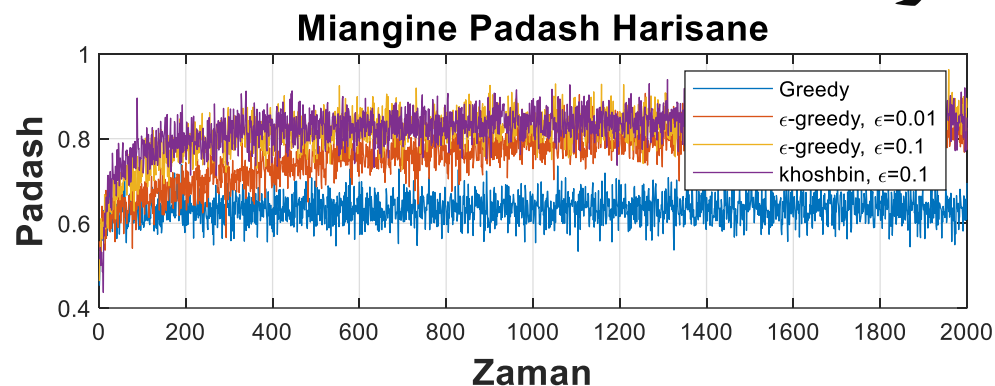


$$\epsilon = 0$$



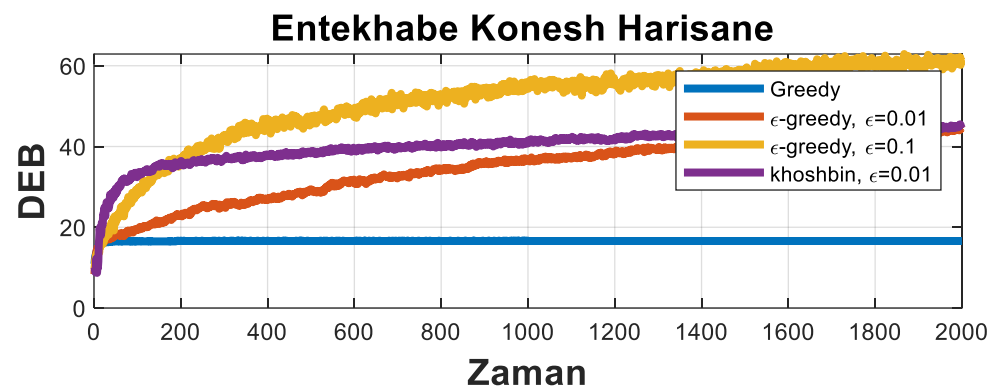
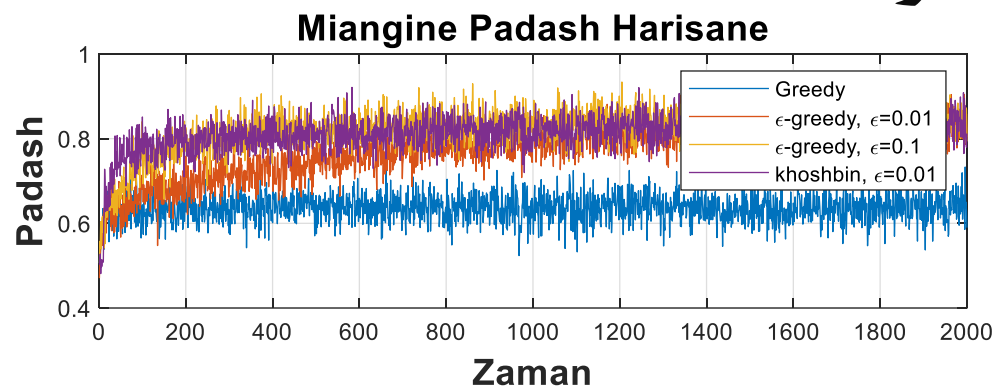
مقداردهی خوشبینانه

$$\epsilon = 0.1$$



مقداردهی خوشبینانه

$$\epsilon = 0.01$$



انتخاب خوشبینانه در آغاز

سرعت همگرایی کمتر در گام‌های آغازین؟

باقی نماندن کنش آزمایش نشده

کاوش بیشتر

سیاست محدوده بالای اعتماد

نیاز به کاوش همیشگی

- به دلیل وجود عدم قطعیت دائمی در تخمین ارزش حالت‌ها
- روش‌های حریصانه
- صرفاً دانش فعلی
- روش‌های حریصانه با اپسیلون
- کاوش ولی عدم تفاوت بین انتخاب‌ها
- روش مقداردهی خوش‌بینانه
- مقداردهی به زمان‌های نخست
- مشاهده افزایش در زمان‌های اولیه جستجو مقدار اولیه خوشبینانه

مناسب بودن انتخاب کنش‌ها بر اساس الگویی

- بر اساس ترکیبی از بهینگی بالقوه و میزان عدم قطعیت
- به دیگر سخن، انتخاب کنش بهینه‌تر و دارای عدم قطعیت کمتر
- $\Leftarrow$ سیاست محدوده بالای اعتماد
- معادله انتخاب کنش سیاست

$$A_t = \arg \max_a \left[Q_t(a) + c \sqrt{\frac{\ln t}{N_t(a)}} \right]$$

سیاست محدوده بالای اعتماد

$$A_t = \arg \max_a \left[Q_t(a) + c \sqrt{\frac{\ln t}{N_t(a)}} \right]$$

انتخاب کنش بیشینه‌ساز عبارت مذکور

در صورت حذف بخش دوم (یا قرار دادن $c = 0$) به روش حریصانه

بخش ریشه دوم

- اندازه‌گیری عدم قطعیت یا وردائی در تخمین ارزش کنش a
- c نمایش سطح اطمینان
- $N(a)$ شمارنده کنش a
- $N(a) = 0$ به مثابه در نظر گرفتن a چون کنش بیشینه‌ساز
- با انتخاب عمل مزبور افزایش شمارنده متناظر
- کاهش عدم قطعیت
- $\ln t$ سبب‌ساز انتخاب تمامی کنش‌ها در طول زمان
- با عدم انتخاب a افزایش t منجر به افزایش صورت
- دلیل لگاریتم‌گیری؟

سیاست محدوده بالای اعتماد

$$A_t = \arg \max_a \left[Q_t(a) + c \sqrt{\frac{\ln t}{N_t(a)}} \right]$$

انتخاب کنش بیشینه‌ساز عبارت مذکور

در صورت حذف بخش دوم (یا قرار دادن $c = 0$) به روش حریصانه

بخش ریشه دوم

- اندازه‌گیری عدم قطعیت یا وردائی در تخمین ارزش کنش a

- c نمایش سطح اطمینان

- $N(a)$ شمارنده کنش a

- $N(a) = 0$ به مثابه در نظر گرفتن a چون کنش بیشینه‌ساز

- با انتخاب عمل مزبور افزایش شمارنده متناظر

- کاهش عدم قطعیت

- $\ln t$ سبب‌ساز انتخاب تمامی کنش‌ها در طول زمان

- با عدم انتخاب a افزایش t منجر به افزایش صورت

- دلیل لگاریتم‌گیری؟

- کوچک شدن افزایش‌ها در طول زمان ولی بدون محدوده

- انتخاب کمتر کنش‌های با تخمین ارزش کمتر یا کنش‌های با بسامد بالای انتخاب در طول زمان

- امکان انتخاب دیگر کنش‌ها در زمان دستیابی بهینگی

سیاست محدوده بالای اعتماد

$$A_t = \arg \max_a \left[Q_t(a) + c \sqrt{\frac{\ln t}{N_t(a)}} \right]$$

انتخاب کنش در صورت $N_t = 0$

▪ در نتیجه در قدم اولیه انتخاب همه کنش‌ها شبیه روش مقداردهی خوش‌بینانه

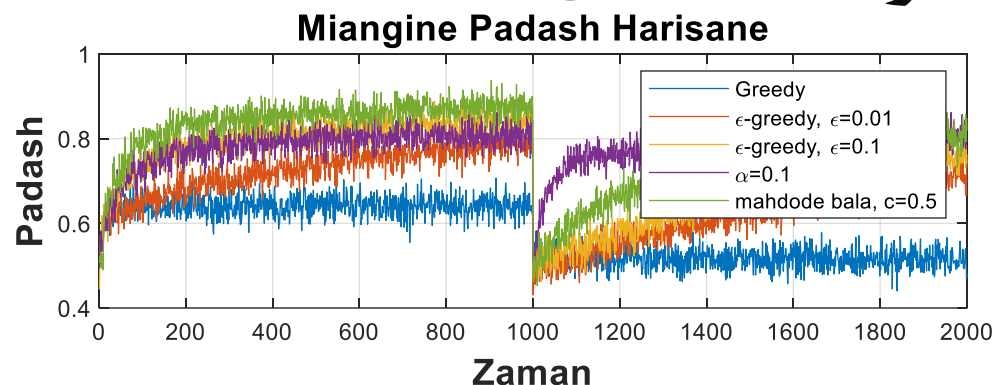
کاربرد روش صرفاً برای ماشین سکه‌ای

▪ مشکل بودن استفاده از آن در دیگر مسائل

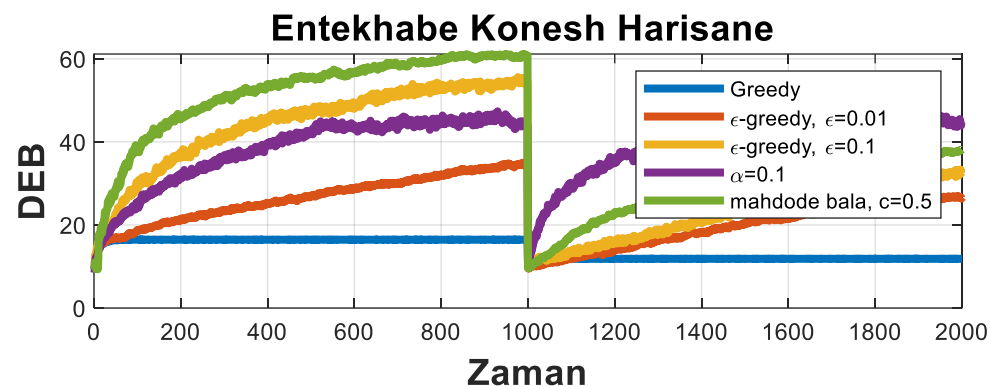
سیاست محدوده بالای اعتماد-شبیه سازی

شبیه سازی

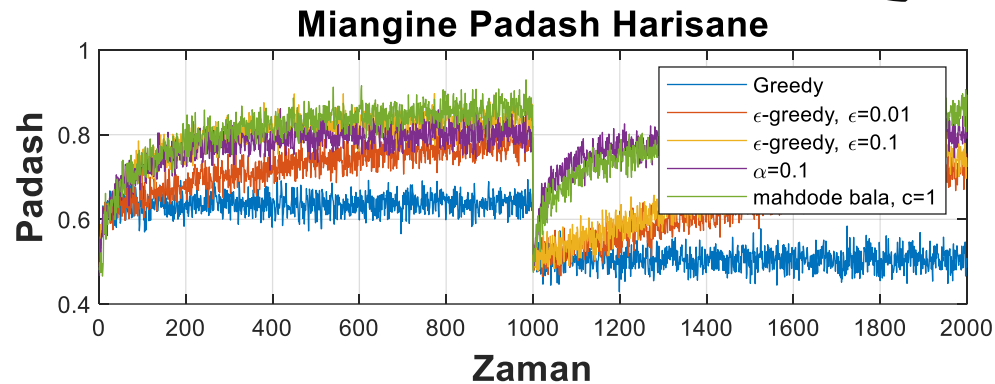
سیاست محدوده بالای اعتماد-شبیه سازی



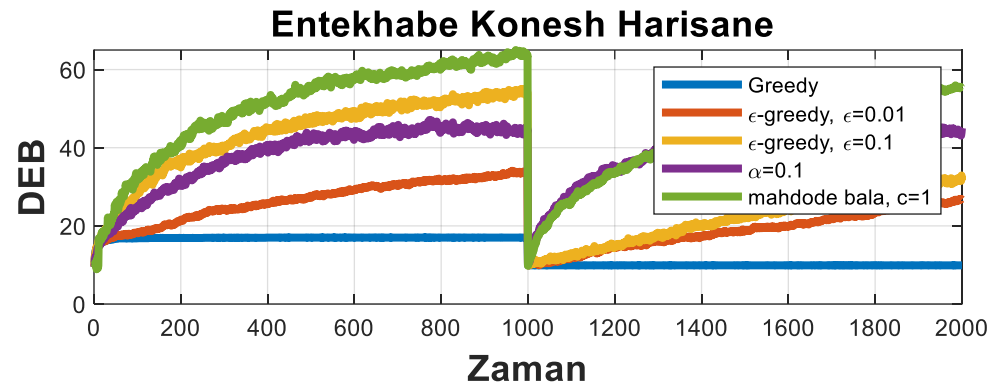
$$c = 0.5$$



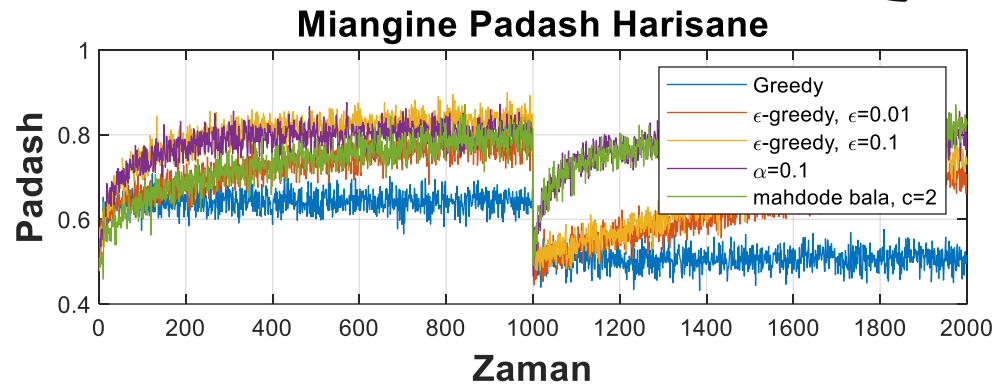
سیاست محدوده بالای اعتماد-شبیه سازی



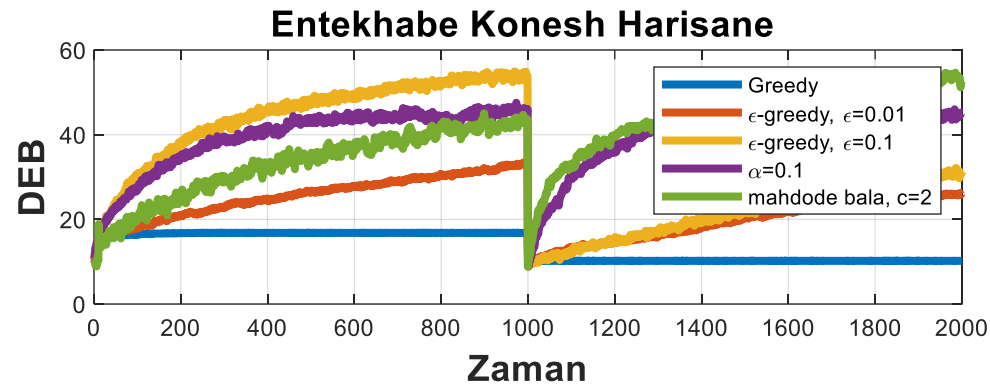
$$c = 1$$



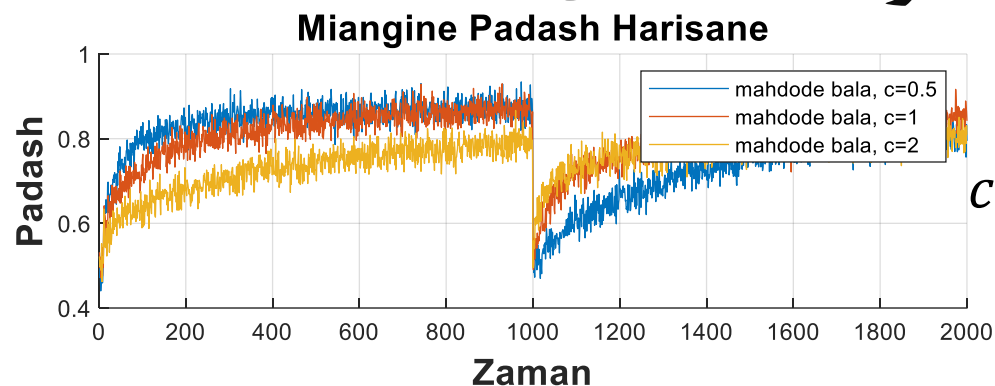
سیاست محدوده بالای اعتماد-شبیه سازی



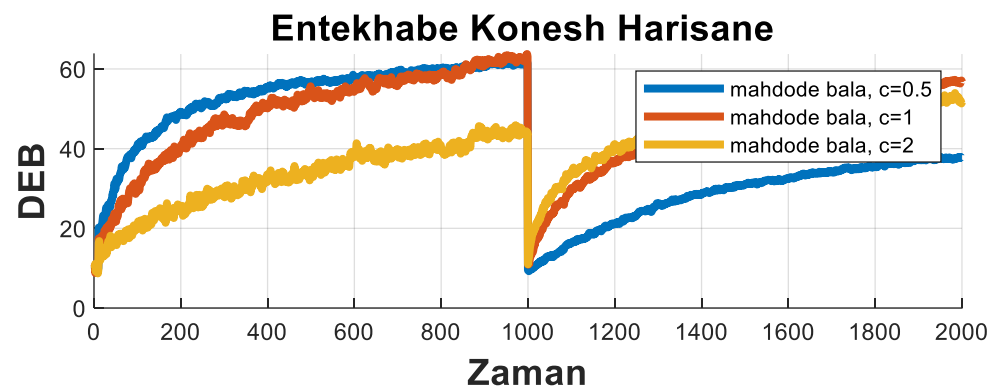
$$c = 2$$



سیاست محدوده بالای اعتماد-شبیه سازی



مقایسه همزمان سه مقدار c



سیاست محدوده بالای اعتماد

$$A_t = \arg \max_a \left[Q_t(a) + c \sqrt{\frac{\ln t}{N_t(a)}} \right]$$

انتخاب کنش در صورت $N_t = 0$

- در نتیجه در قدم اولیه انتخاب همه کنش‌ها شبیه روش مقداردهی خوش‌بینانه

کاربرد روش صرفاً برای ماشین سکه‌ای

- مشکل بودن استفاده از آن در دیگر مسائل

- ؟

- ناکارآمد در مسائل نامانا

- در مواجهه با فضای حالت‌های بزرگ

سیاست راهزن گرادانی

الگوریتم‌های حریصانه با اپسیلون و م‌با

- مبنی بر تخمین تجربی ارزش کنش‌ها
- احتمال انجام کنش انتخاب ۱۰۰ درصد و احتمال اجرای بقیه کنش‌ها صفر درصد در این دو روش

دسته روش‌های دیگر

- امکان انتخاب اولویت و ترجیح (غیرقطعی) صرفاً ترجیح یکی بر دیگری
 - اولویت بیشتر منجر به انتخاب بیشتر
 - عدم تفسیر ترجیح در قالب پاداش
 - صرفاً اهمیت ترجیح نسبی کنشی به دیگری

سیاست راهزن گرادانی

روش گردایان

- مبنی بر اولویت کنش‌ها نسبت به یکدیگر
- تعریف مجموعه احتمالات برای کنش‌های مختلف
- انتخاب یکی از کنش‌ها با توجه احتمال‌های متناظر و اعمال به محیط
- میزان اولویت کنش a برابر $H_t(a)$
- تغییر تدریجی احتمالات
- میل احتمال کنش بهینه به مقدار یک و بقیه به سمت صفر
- $\Leftarrow$ شانس انتخاب هر کنشی بسته به میزان احتمال اولویت آن

$$P\{A_t = a\} = \frac{e^{H_t(a)}}{\sum_{b=1}^k e^{H_t(b)}} = \pi_t(a)$$

$\Leftarrow$ توزیع سافت-مکس (بیشینه هموار) یا توزیع گیبز یا بولتزمان
توجه به تعریف $\pi_t(a)$

- احتمال انتخاب کنش a در زمان t
- اولویت یکسان تمامی کنش‌ها در آغاز

سیاست راهزن گرادانی

وجود الگوریتمی مخصوص و داتی
▪ استفاده از گردایان صعودی تصادفی

$$H_{t+1}(a) = H_t(a) + \alpha \frac{\partial E[R_t]}{\partial H_t(a)}$$
$$\begin{cases} H_{t+1}(A_t) = H_t(A_t) + \alpha(R_t - \bar{R}_t)(1 - \pi_t(A_t)) \\ H_{t+1}(a) = H_t(a) - \alpha(R_t - \bar{R}_t)\pi_t(a) \end{cases}$$

در ابتدا، برابری ارزش همه کنش‌ها

$$H_1(a) = 0$$

چرا صعودی و نه نزولی؟

سیاست راهزن گرادانی

وجود الگوریتمی مخصوص و داتی
▪ استفاده از گردایان صعودی تصادفی

$$H_{t+1}(a) = H_t(a) + \alpha \frac{\partial E[R_t]}{\partial H_t(a)}$$
$$\begin{cases} H_{t+1}(A_t) = H_t(A_t) + \alpha(R_t - \bar{R}_t)(1 - \pi_t(A_t)) \\ H_{t+1}(a) = H_t(a) - \alpha(R_t - \bar{R}_t)\pi_t(a) \end{cases}$$

در ابتدا، برابری ارزش همه کنش‌ها

$$H_1(a) = 0$$

چرا صعودی و نه نزولی؟

- صعودی برای بیشینه
- نزولی برای یافتن کمینه

سیاست راهزن گرادینانی

وجود الگوریتمی مخصوص و داتی
▪ استفاده از گردایان صعودی تصادفی

$$H_{t+1}(a) = H_t(a) + \alpha \frac{\partial E[R_t]}{\partial H_t(a)}$$
$$\begin{cases} H_{t+1}(A_t) = H_t(A_t) + \alpha(R_t - \bar{R}_t)(1 - \pi_t(A_t)) \\ H_{t+1}(a) = H_t(a) - \alpha(R_t - \bar{R}_t)\pi_t(a) \end{cases}$$

$\bar{R}_t$ میانگین پاداش‌ها تا زمان t
▪ به مثابه پایه مقایسه؟

سیاست راهزن گرادانی

وجود الگوریتمی مخصوص و داتی
▪ استفاده از گردایان صعودی تصادفی

$$H_{t+1}(a) = H_t(a) + \alpha \frac{\partial E[R_t]}{\partial H_t(a)}$$
$$\begin{cases} H_{t+1}(A_t) = H_t(A_t) + \alpha(R_t - \bar{R}_t)(1 - \pi_t(A_t)) \\ H_{t+1}(a) = H_t(a) - \alpha(R_t - \bar{R}_t)\pi_t(a) \end{cases}$$

$\bar{R}_t$ میانگین پاداش‌ها تا زمان t

▪ به مثابه پایه مقایسه؟

- در صورت بالاتر بودن پاداش از میانگین افزایش اولویت کنش منتخب در زمان‌های آینده
- در صورت پایین‌تر بودن پاداش از میانگین کاهش اولویت کنش منتخب در زمان‌های آینده

سیاست راهزن گرادانی

بیشینه ساز امید ریاضی پاداش ها

$$E[R_t] =$$

سیاست راهزن گرادانی

بیشینه ساز امید ریاضی پاداش ها

$$E[R_t] = \sum_x \pi_t(x) q_*(x)$$

سیاست راهزن گرادانی

بیشینه ساز امید ریاضی پاداش ها

$$E[R_t] = \sum_x \pi_t(x) q_*(x)$$

مشتق از

$$E[R_t] = \sum_x \sum_r r \pi_t(x) R_t$$

سیاست راهزن گرادانی

بیشینه‌ساز امید ریاضی پاداش‌ها

$$E[R_t] = \sum_x \pi_t(x) q_*(x)$$

مشتق از

$$E[R_t] = \sum_x \sum_r r \pi_t(x) R_t = \sum_x \pi_t(x) \sum_r r R_t$$

سیاست راهزن گرادانی

بیشینه‌ساز امید ریاضی پاداش‌ها

$$E[R_t] = \sum_x \pi_t(x) q_*(x)$$

مشتق از

$$E[R_t] = \sum_x \sum_r r \pi_t(x) R_t = \sum_x \pi_t(x) \sum_r r R_t = \sum_x \pi_t(x) q_*(x)$$

سیاست راهزن گرادینانی

بیشینه‌ساز امید ریاضی پاداش‌ها

$$E[R_t] = \sum_x \pi_t(x) q_*(x)$$

استفاده از گردایان صعودی تصادفی

$$H_{t+1}(a) = H_t(a) + \alpha \frac{\partial E[R_t]}{\partial H_t(a)}$$
$$\begin{cases} H_{t+1}(A_t) = H_t(A_t) + \alpha (R_t - \bar{R}_t) (1 - \pi_t(A_t)) \\ H_{t+1}(a) = H_t(a) - \alpha (R_t - \bar{R}_t) \pi_t(a) \end{cases}$$

در ابتدا، برابری ارزش همه کنش‌ها

$$H_1(a) = 0$$

سیاست راهزن گرادینانی - اثبات

$$E[R_t] = \sum \pi_t(x) q_*(x)$$
$$\frac{\partial E^x[R_t]}{\partial H_t(a)} =$$

سیاست راهزن گرادینانی - اثبات

$$E[R_t] = \sum_x \pi_t(x) q_*(x)$$
$$\frac{\partial E[R_t]}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \left[\sum_x \pi_t(x) q_*(x) \right]$$

سیاست راهزن گرادینانی - اثبات

$$E[R_t] = \sum_x \pi_t(x) q_*(x)$$
$$\frac{\partial E[R_t]}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \left[\sum_x \pi_t(x) q_*(x) \right] \overset{\text{جابجایی}}{\equiv} \sum_x q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)}$$

سیاست راهزن گرادینانی - اثبات

$$\begin{aligned} E[R_t] &= \sum_x \pi_t(x) q_*(x) \\ \frac{\partial E[R_t]}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)} \left[\sum_x \pi_t(x) q_*(x) \right] \stackrel{\text{جابجایی}}{\cong} \sum_x q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)} \\ &\stackrel{\text{چون } \frac{\partial \pi_t(x)}{\partial H_t(a)} = 0}{\cong} \sum_x (q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} \end{aligned}$$

سیاست راهزن گرادینانی - اثبات

$$\begin{aligned} E[R_t] &= \sum_x \pi_t(x) q_*(x) \\ \frac{\partial E[R_t]}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)} \left[\sum_x \pi_t(x) q_*(x) \right] \stackrel{\text{جابجایی}}{\cong} \sum_x q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)} \\ &\stackrel{\text{چون } \frac{\partial \pi_t(x)}{\partial H_t(a)} = 0}{\cong} \sum_x (q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} \end{aligned}$$

$$B_t = \overline{R_t} \text{ معمولاً}$$

سیاست راهزن گرادینانی - اثبات

$$\begin{aligned} E[R_t] &= \sum_x \pi_t(x) q_*(x) \\ \frac{\partial E[R_t]}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)} \left[\sum_x \pi_t(x) q_*(x) \right] \stackrel{\text{جابجایی}}{\equiv} \sum_x q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)} \\ &\stackrel{\text{چون } \frac{\partial \pi_t(x)}{\partial H_t(a)} = 0}{\equiv} \sum_x (q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} \end{aligned}$$

$$B_t = \overline{R_t} \text{ معمولاً}$$

$$\stackrel{* \frac{\pi_t(A_t)}{\pi_t(A_t)}}{\equiv} \sum_x (q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} * \frac{\pi_t(A_t)}{\pi_t(A_t)}$$

سیاست راهزن گرادینانی - اثبات

$$= \sum_x \pi_t(A_t)(q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} / \pi_t(A_t)$$

سیاست راهزن گرادینانی - اثبات

$$\begin{aligned} &= \sum_x \pi_t(A_t)(q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} / \pi_t(A_t) \\ &\stackrel{\cong}{=} E \left[\frac{(q_*(A_t) - B_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)}}{\pi_t(A_t)} \right] \end{aligned}$$

سیاست راهزن گرادینانی - اثبات

$$\begin{aligned} &= \sum_x \pi_t(A_t)(q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} / \pi_t(A_t) \\ &\stackrel{\cong}{=} E \left[\frac{(q_*(A_t) - B_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)}}{\pi_t(A_t)} \right] \end{aligned}$$

$$\begin{aligned} E(R_t | A_t = a) &= q_*(a) \text{ و } B_t = \bar{R}_t \\ &\stackrel{\cong}{=} \end{aligned}$$

سیاست راهزن گرادینانی - اثبات

$$\begin{aligned} &= \sum_x \pi_t(A_t)(q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} / \pi_t(A_t) \\ &\cong E \left[\frac{(q_*(A_t) - B_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)}}{\pi_t(A_t)} \right] \end{aligned}$$

$$\begin{aligned} E(R_t | A_t = a) &= q_*(a) \text{ و } B_t = \bar{R}_t \\ &\cong E \left[(R_t - \bar{R}_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t) \right] \quad (*) \end{aligned}$$

سیاست راهزن گرادینانی - اثبات

$$\begin{aligned} &= \sum_x \pi_t(A_t)(q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} / \pi_t(A_t) \\ &\cong E \left[\frac{(q_*(A_t) - B_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)}}{\pi_t(A_t)} \right] \end{aligned}$$

$$\begin{aligned} &E(R_t | A_t = a)_{=q_*(a)} \text{ و } B_t = \bar{R}_t \\ &\cong E \left[(R_t - \bar{R}_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t) \right] \quad (*) \\ &\cong E \left[(R_t - \bar{R}_t) (\mathbb{1}_{a=A_t} - \pi_t(a)) \right] \end{aligned}$$

سیاست راهزن گرادینی - اثبات

$$\frac{\partial}{\partial x} \left[\frac{f(x)}{g(x)} \right]$$

سیاست راهزن گرادینانی - اثبات

$$\frac{\partial}{\partial x} \left[\frac{f(x)}{g(x)} \right] = \frac{\frac{\partial f(x)}{\partial x} g(x) - f(x) \frac{\partial g(x)}{\partial x}}{g(x)^2} (**)$$

سیاست راهزن گرادینانی - اثبات

$$\frac{\partial}{\partial x} \left[\frac{f(x)}{g(x)} \right] = \frac{\frac{\partial f(x)}{\partial x} g(x) - f(x) \frac{\partial g(x)}{\partial x}}{g(x)^2} (**)$$
$$(**) \Rightarrow \frac{\partial \pi_t(x)}{\partial H_t(a)}$$

سیاست راهزن گرادینانی - اثبات

$$\frac{\partial}{\partial x} \left[\frac{f(x)}{g(x)} \right] = \frac{\frac{\partial f(x)}{\partial x} g(x) - f(x) \frac{\partial g(x)}{\partial x}}{g(x)^2} (**)$$
$$(**) \Rightarrow \frac{\partial \pi_t(x)}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \pi_t(x)$$
$$=$$

سیاست راهزن گرادینانی - اثبات

$$\begin{aligned}\frac{\partial}{\partial x} \left[\frac{f(x)}{g(x)} \right] &= \frac{\frac{\partial f(x)}{\partial x} g(x) - f(x) \frac{\partial g(x)}{\partial x}}{g(x)^2} (**) \\(**) \Rightarrow \frac{\partial \pi_t(x)}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)} \pi_t(x) \\&= \frac{\partial}{\partial H_t(a)} \left[\frac{e^{H_t(x)}}{\sum_{y=1}^k e^{H_t(y)}} \right] \\&= \end{aligned}$$

سیاست راهزن گرادینانی - اثبات

$$\begin{aligned}\frac{\partial}{\partial x} \left[\frac{f(x)}{g(x)} \right] &= \frac{\frac{\partial f(x)}{\partial x} g(x) - f(x) \frac{\partial g(x)}{\partial x}}{g(x)^2} (**) \\(**) \Rightarrow \frac{\partial \pi_t(x)}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)} \pi_t(x) \\&= \frac{\partial}{\partial H_t(a)} \left[\frac{e^{H_t(x)}}{\sum_{y=1}^k e^{H_t(y)}} \right] \\&= \frac{\frac{\partial e^{H_t(x)}}{\partial H_t(a)} \sum_{y=1}^k e^{H_t(y)} - e^{H_t(x)} \frac{\partial \sum_{y=1}^k e^{H_t(y)}}{\partial H_t(a)}}{\left(\sum_{y=1}^k e^{H_t(y)} \right)^2}\end{aligned}$$

سیاست راهزن گرادینانی - اثبات

$$\begin{aligned} \frac{\partial e^x}{\partial x} &\stackrel{=}{=} e^x \frac{\mathbb{I}_{a=x} e^{H_t(x)} \sum_{y=1}^k e^{H_t(y)} - e^{H_t(x)} e^{H_t(a)}}{\left(\sum_{y=1}^k e^{H_t(y)} \right)^2} \\ &= \frac{\mathbb{I}_{a=x} e^{H_t(x)}}{\sum_{y=1}^k e^{H_t(y)}} - \frac{e^{H_t(x)} e^{H_t(a)}}{\left(\sum_{y=1}^k e^{H_t(y)} \right)^2} \\ &= \mathbb{I}_{a=x} \pi_t(x) - \pi_t(x) \pi_t(a) \\ &= \pi_t(x) (\mathbb{I}_{a=x} - \pi_t(a)) \end{aligned}$$

سیاست راهزن گرادینانی - اثبات

$$\begin{aligned} H_{t+1}(a) &= H_t(a) + \alpha \frac{\partial E[R_t]}{\partial H_t(a)} \\ &= H_t(a) + \alpha E[(R_t - \bar{R}_t)(\mathbb{1}_{a=x} - \pi_t(a))] \\ &\approx H_t(a) + \alpha (R_t - \bar{R}_t)(\mathbb{1}_{a=x} - \pi_t(a)) \end{aligned}$$

توجه به خود پاداش نمونه به جای امید ریاضی

▪ بوتس تراپینگ! فرزند وصال خویشان!

▪ پاداش دریافتی اهمیت زیادی می دهیم ← خودراه اندازی

سیاست راهزن گرادانی

افزایش احتمال انتخاب دوباره کنشی با دریافت پاداشی بیشتر از مقدار پایه
▪ متقابلاً، کاهش یک اندازه احتمال سایر کنش‌ها

کاهش احتمال انتخاب دوباره کنشی با دریافت پاداشی کمتر از مقدار پایه
▪ متقابلاً، افزایش یک اندازه احتمال سایر کنش‌ها

جهت کاهش وردائی بهتر است میانگین پاداش‌ها غیر صفر

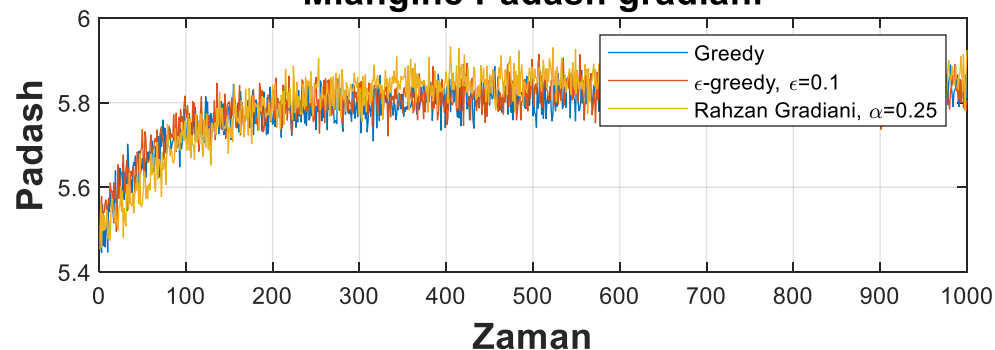
امکان اختیار هر مقدار دلخواه برای مقدار پایه
▪ معمولاً میانگین نمونه‌های ثبت شده

سیاست راهزن گرادانی

شبیه سازی

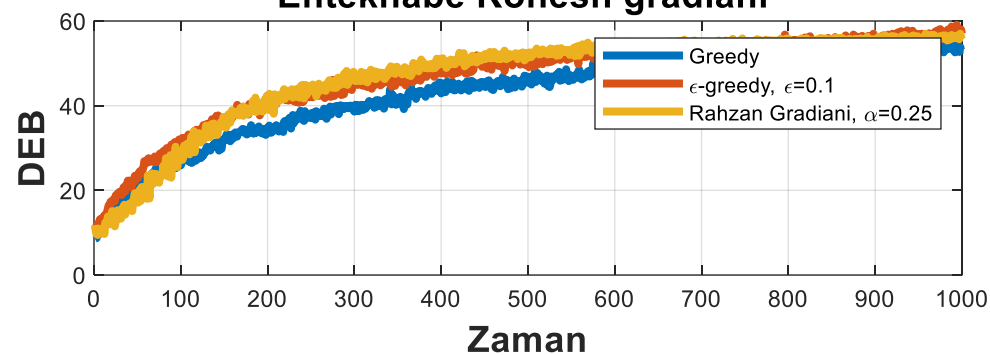
سیاست راهزن گرادینانی

Miangine Padash gradiani



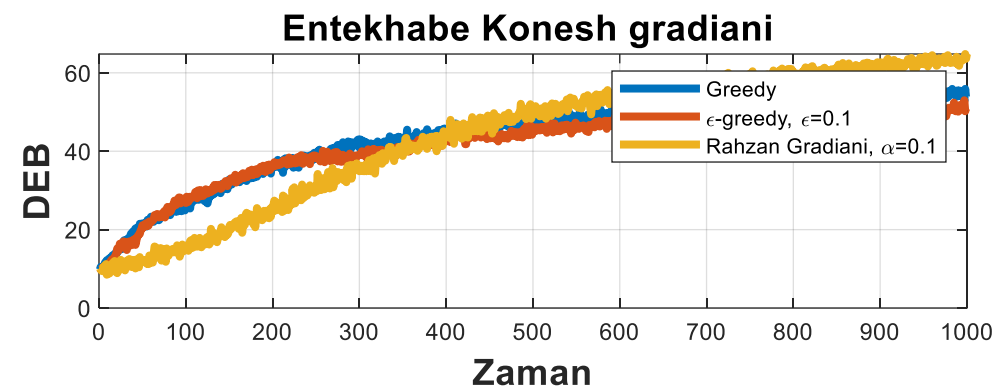
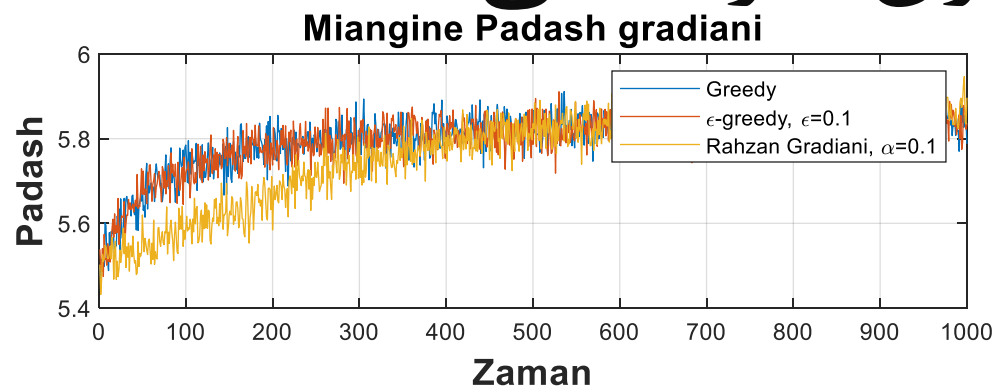
$$\alpha = 0.25$$

Entekhabe Konesh gradiani



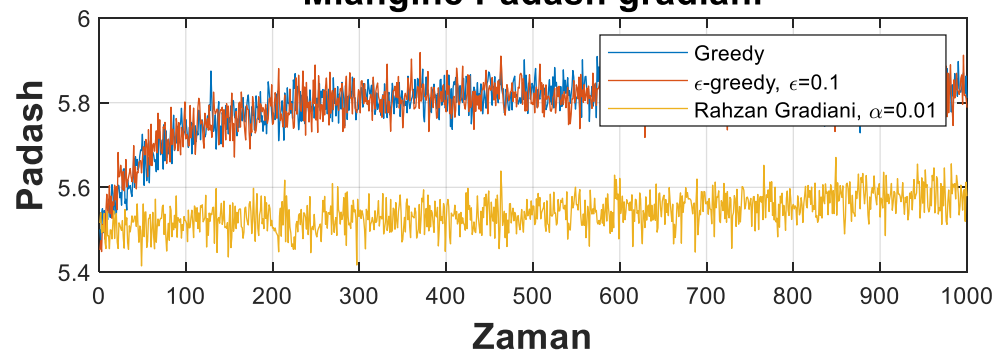
سیاست راهزن گرادیدانی

$$\alpha = 0.1$$



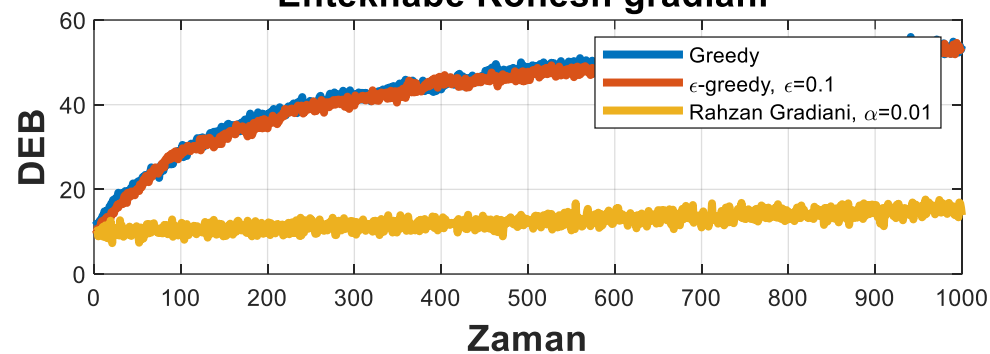
سیاست راهزن گرادینانی

Miangine Padash gradiani



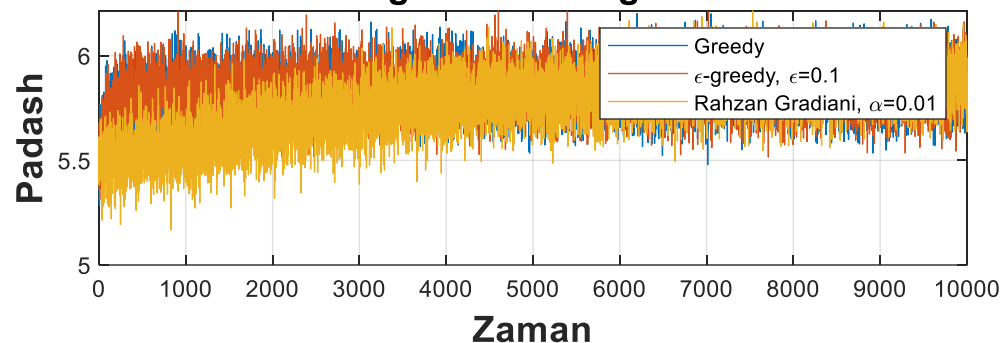
$$\alpha = 0.01$$

Entekhabe Konesh gradiani



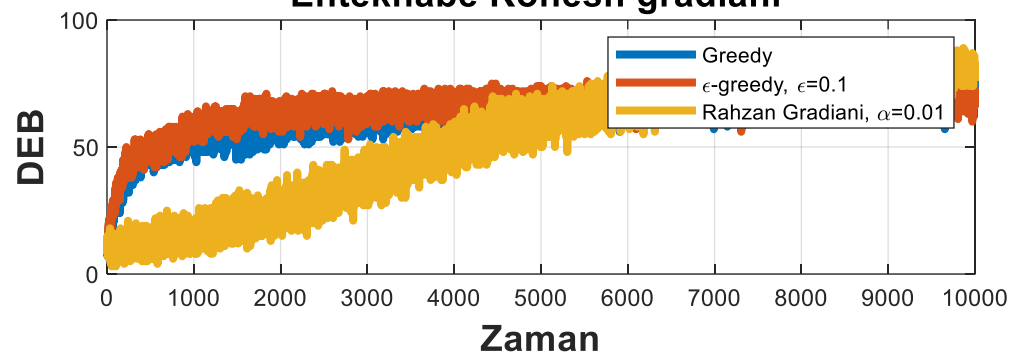
سیاست راهزن گرادینانی

Miangine Padash gradiani



$$\alpha = 0.01$$

Entekhabe Konesh gradiani



سیاست راهزن گرادانی

دارای زمان محاسبه بیشتر نسبت به حریصانه با اپسیلون
▪ دارای دردسر در تعمیم به دیگر مسائل

راهزن زمینه‌ای؟! جستجوی مرتبط؟!!

صرفاً پرداختن به جستجوی فارغ از محیط

- به دنبال یافتن بهترین کنش
- یا یافتن بهترین کنش و رهگیری آن در محیط نامانا

تغییر پاداش کنش‌ها (تغییر رفتار)
نامانا

- لزوماً دارای پاسخ مناسب نیست
- مگر در هنگام تغییرات محدود

فرض

▪ چشمکی با تغییر رفتار!

- سبز اهرم پنج
- سفید اهرم هفت
- سرخ اهرم یک

▪ جستجوی مرتبط associative search

▪ راهزن زمینه‌ای contextual bandit

▪ مرز میان راهزن معمول و یادگیری تقویتی

راهزن زمینه‌ای؟! جستجوی مرتبط؟!!

صرفاً پرداختن به جستجوی فارغ از محیط

- به دنبال یافتن بهترین کنش
- یا یافتن بهترین کنش و رهگیری آن در محیط نامانا

تغییر پاداش کنش‌ها (تغییر رفتار)

- نامانا
- لزوماً دارای پاسخ مناسب نیست
- مگر در هنگام تغییرات محدود

فرض

▪ چشمکی با تغییر رفتار!

- سبز اهرم پنج
- سفید اهرم هفت
- سرخ اهرم یک

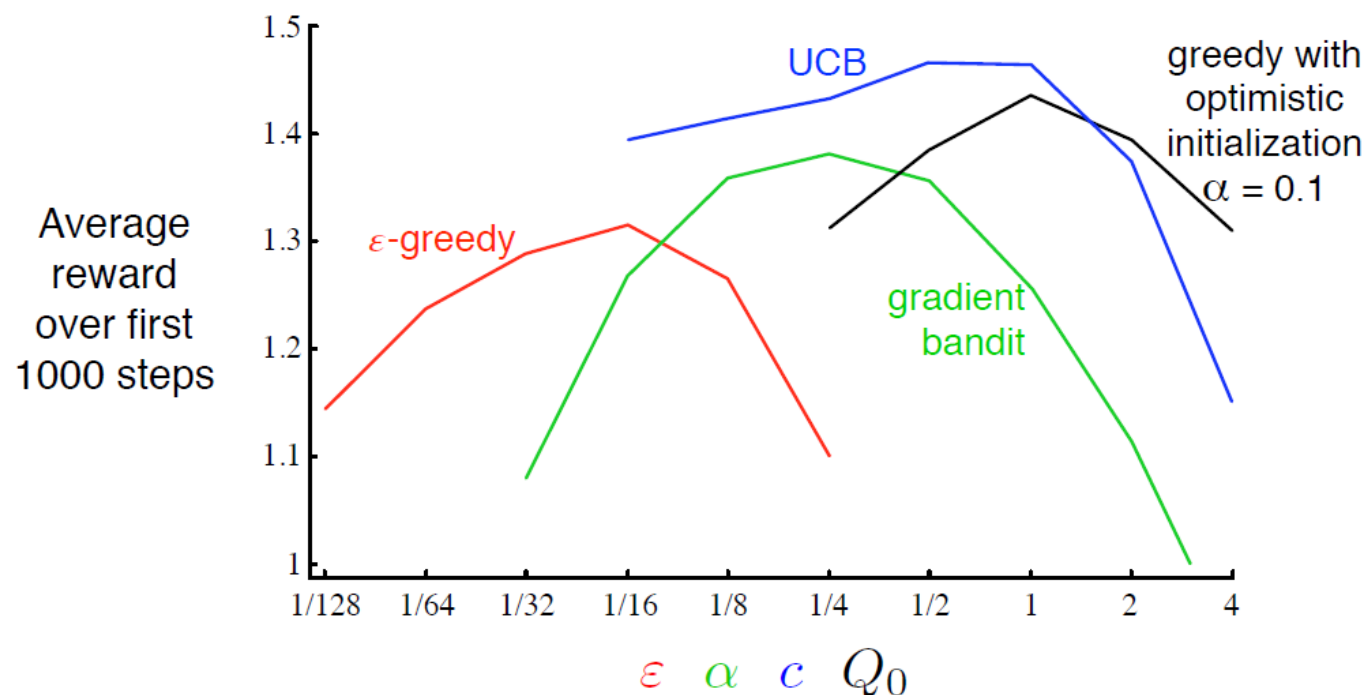
▪ جستجوی مرتبط associative search

▪ راهزن زمینه‌ای contextual bandit

▪ مرز میان راهزن معمول و یادگیری تقویتی

▪ یادگیری در صورتی که انتخاب فعلی بر آینده نیز تاثیر بگذارد

مقایسه بهره-کاوش یاب‌ها



معین در مقابل نامعین

▪ انتخاب نامعین روشمند در مقابل تصادفی

▪ توزیع گیبز

همه بر مبنای پارامتر

منحنی یادگیری

▪ مطالعات پارامتر

▪ منحنی درجه دو کاو

▪ حساسیت

سادگی روش‌ها

▪ مناسبتر از پیچیده‌ها

منابع

ساتن

زندى

ۋثرۋاكس