گزارش سوالات تحلیلی :

سوال اول:

انتخاب های مساله(actions): در ابتدای هر ماه دو انتخاب داریم. 1- قرار دادن تمام حقوق در بورس 2- قرار دادن در بانک

پاداش و هزینه:

- قرار دادن پول در بورس:
  پاداش: دریافت سود ناشی از سرمایه گذاری
  هزینه: سقوط کردن بورس و از دست رفتن تمام سرمایه یا بخشی از آن
- قرار دادن در بانک:
  پاداش: دریافت سود ماهانه
  هزینه: ندارد(اگر بانک در پایان ماه سود پول را ندهد)
  ** با فرض وجود تورم ضرر هم خواهیم کرد.

سیاست:
میبایست سیاست را طوری تعیین کنیم که Exploration-Exploitation balance داشته باشیم.
در ماه های اول دریافت حقوق باید explore کنیم و در ادامه با کسب اطلاعات exploit انجام دهیم.
می‌توانیم میانگین پاداش های هر انتخاب تا این لحظه را محاسبه و بیشترین را انتخاب کنیم.(greedy)
با تغییر مناسب اپسیلون با توجه به exploration به پاداش مناسب برسیم.
البته به نوع مسئله نیز خیلی بستگی دارد. برای مثال اگر در ابتدا از بورس خیلی ضرر کنیم احتمالا زودتر اپسیلون(ریسک پذیری) را کاهش می‌دهیم تا اکشن مطمئن تر را انتخاب کنیم.

این مسئله به مقدار پولی که در ماه می‌گیریم نیز بستگی دارد:
اگر حقوق این ماه بیشتر باشد(از یک حدی) احتمالا بانک را انتخاب خواهیم کرد و اگر حقوق کمتری داشته باشیم شاید ریسک بورس را به جان بخریم که بدین صورت risk-averse بودن یا نبودن را هم وارد سیاست میکنیم.

سوال دوم:

انتخاب های مساله(actions): چهار انتخاب داریم: 1- انجام وظایف دانشگاه 2- انجام کارهای شرکت 3- روابط اجتماعی و گذراندن وقت با دوستان و آشنایان 4- وقت گذاشتن برای خود

پاداش ها و هزینه ها:

- انجام وظایف دانشگاه:
  پاداش: نمره خوب، موقعیت های بهتر در آینده، خوشحالی و رضایت
  هزینه: سختی، وقت نداشتن برای کارهای دیگر یا حتی خانواده
- انجام کارهای شرکت:
  پاداش: حقوق ماهانه، ترفیع رتبه در آینده
  هزینه: تنزل رتبه(مثلا وظایف را خوب انجام ندهیم)، وقت نداشتن برای دوستان، دور شدن از درس
- روابط اجتماعی:
  پاداش: خوشحالی و حس خوب، گسترش روابط
  هزینه: ناراحتی(مثلا اگر روابط تنش زا باشند)
- وقت گذاشتن برای خود:
  پاداش: درک نیازهای خود، آرامش و تمرکز
  هزینه: –

سیاست:
به نظر من این مسئله تک state نیست. چون در دنیای واقعی انتخاب هر کدام روی دیگری تاثیرگذار است. برای مثال اگر برای خود وقت بگذاریم احتمالا تمرکز بیشتری در انجام کارهای خود داریم و پاداش های بیشتری می‌گیریم. یا مثلا اگر بیش از حد برای وظایف دانشگاه وقت گذاشته شود احتمالا روابط با دوستان ضعیف تر می‌شود و پاداش کمتری در آینده از آن می‌گیریم.
ضمن اینکه پاداش هر کدام از این انتخاب ها در هر دفعه یکسان نیست یا حتی پاداش ها کلا تغییر می‌کنند. مثلا زمان‌هایی هستند که انسان دوست دارد تنها باشد بنابراین روابط اجتماعی باعث خوشحالی نخواهد شد یا مثلا شاید آشنایی با یک دوست جدید ما را از بیرون رفتن با یک دوست قدیمی بیشتر خوشحال کند.

با وجود این موارد فکر می‌کنم همه ما به طور متوسط در زندگی یکنواخت عمل میکنیم. به این معنی که الگوهای یکسانی در زندگی تکرار می‌کنیم و پاداش هر کار در هر شرایط را می‌توانیم حتی حدس بزنیم.(پاداش subjective)
مثلا بعضی از ما اگر اول کارهای شرکت را انجام دهیم، بهتر و با تمرکز تر کارهای دانشگاه را انجام خواهیم داد، یا بعضی‌ها در موازی پیش بردن کارها بهتر و راحت تر هستند.

پس بهتر است با explore کردن ببینیم در چه شرایطی بهتر عمل می‌کنیم تا این الگوها را تخمین بزنیم. برای مثال در یک روز که می‌خواهیم کارهای دانشگاه را انجام دهیم ببینیم تاثیر وقت گذاشتن برای خود(مثلا یک ساعت) چقدر موثر روی انجام آنها موثر است.
یا مثلا ببینیم با چقدر زمان گذاشتن روی هر کدام پاداش خوبی دریافت می‌کنیم. مثلا برای یک شخص بیرون رفتن با دوستان شاید پاداش کمتری از صحبت کردن با دوست صمیمی پاداش داشته باشد.

با فرض بدست آمدن این الگوها(چیزی شبیه utility function) می‌توانیم در ادامه از الگوهای بدست آمده استفاده کنیم(exploitation)
که این روش خیلی طول می‌کشد ولی اگر زمانی دیدیم که با تکرار یک الگو پاداش خوبی دریافت می‌کنیم باید همان را دوباره انتخاب کنیم که احتمالا دفعه بعد پاداش متفاوت می‌شود و دوباره باید کمی تغییرات ایجاد کنیم.

سوال سوم:

انتخاب های مساله(actions): روش های تبلیغاتی شامل: 1- شبکه اجتماعی 2- تبلیغات تلویزیونی 3- تبلیغ در سطح شهر

پاداش ها و هزینه ها:

- شبکه اجتماعی:
  پاداش: پخش(وایرال) شدن تبلیغ
  هزینه: هزینه تبلیغ(کم)
- تبلیغ تلویزیونی:
  پاداش: بیشتر دیده شدن(در سطح کشور)
  هزینه: هزینه تبلیغ(زیاد)
- سطح شهر:
  پاداش: جلب توجه، احتمال دیده شدن بالا
  هزینه: هزینه تبلیغ(متوسط)

در ابتدای یادگیری باید explore کنیم پس با greedy عمل کردن در ابتدا شروع می‌کنیم و در ادامه تبلیغ با پاداش بیشتر را دریافت می‌کنیم و مثلا اگر از eps greedy استفاده کنیم باید در ادامه آن را کم کنیم.

با توجه به اینکه تابع پاداش بازوها را می‌دانیم می‌توانیم با انتخاب های بهتر در exploration زودتر به پاداش بهینه همگرا شویم و هزینه را کم کنیم.

لازم به ذکر است که این بازوها روی یکدیگر می‌توانند تاثیر بگذارند. مثلا تبلیغ در تلویزیون و در ادامه در سطح شهر شاید پاداش بیشتری از تبلیغ تنها در تلویزیون داشته باشد.

گزارش سوال پیاده‌سازی :

بخش 1)
در این مساله 9 بازو داریم که مطابق با جدول داده شده در صورت سوال است.
بازوی1,2,3: junior: شبکه های اجتماعی، تبلغیات محیطی، آگهی کاریابی
بازوی 4و5و6: mid-level: شبکه های اجتماعی، تبلغیات محیطی، آگهی کاریابی
بازوی 7و8و9: senior: شبکه های اجتماعی، تبلغیات محیطی، آگهی کاریابی

پاداش ها:

- همیشه یک هزینه ثابت تبلیغ داریم : شبکه اجتماعی: 2 میلیون و تبلیغات محیطی: 14 میلیون و آگهی کاریابی: 10 میلیون
- اگر کلاس برگزار شود(با توجه به احتمال های گفته شده):
  Junior: هر نفر 8 میلیون—> مجموع کلاس: 80 میلیون
  Senior: هر نفر 3 میلیون—> مجموع کلاس: 30 میلیون
  Junior: هر نفر 6 میلیون—> مجموع کلاس: 60 میلیون
- پاداش مسئله استخدام شدن هر کدام از افراد در صنعت است(با توجه به احتمال):

  Junior: 27 Mil
  Senior: 25 Mil
  mid-level: 26 million

**من احتمال ها را اینطور در نظر گرفتم که احتمال استخدام هر شخص در صنعت برابر 0.7 یا 0.5 یا 0.3 است.

در واقع احتمال انتخاب نشدن اشخاص را نیز در نظر گرفتم و هزینه آن را کم کردم.
برای مثال:
اگر نوع تبلیغ شبکه اجتماعی باشد و گروه junior:
هزینه تبلیغ: 2 میلیون
احتمال برگزاری کلاس: 0.8 — > در صورت برگزاری : 80 میلیون هزینه کلاس
احتمال استخدام شدن (مثلا) 5 نفر:

(0.7)^5*(0.3)^5*C(10,5)
انتخاب 5 نفر از دو نفر:C(10,5)

پاداش استخدام : 135=27*5

احتمال کل:

(0.7)^5*(0.3)^5*C(10,5)*0.8

پاداش کل:

(135-82)=53

چون این موضوع را در نظر گرفتم، نمودار ریوارد نویزی شد و سعی کردم با رسم آن برای شرکت های بیشتر شکل کلی آن را بهتر نمایش دهم.
در الگوریتم UCB ابتدا همه بازوها را یک بار انتخاب کردم تا به مشکل صفر شدن آرگومان لگاریتم برنخورم و به همین دلیل در نمودار ریوارد یک جهش در ابتدای آن مشاهده می‌شود.

نمودار Reward: در نمودار ریوارد می‌بینیم که الگوریتم eps-greedy زودتر به اکشن بهینه که اکشن اول باشد میل کرده است و الگوریتم UCB با سرعت کمتری به اکشن بهینه میل کرده است.

نمودار Regret: در نمودار ریگرت، شیب مربوط به هر دو الگوریتم کم شده است ولی با توجه به انتهای نمودار می‌بینیم رشد نمودار UCB کمتر از نمودار eps-greedy است و این یعنی در horizon بیشتر همان رفتار لگاریتمی مورد انتظار را خواهیم دید.