Thompson Sampling

Alvaro J. Riascos Villegas
Universidad de los Andes y Quantil

Septiembre de 2023

Contenido

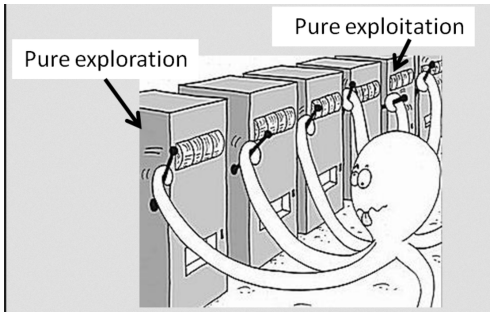
Aprendizaje e Interacciones Sociales

- Introducimos las ideas principales del aprendizaje basado en la intereaccion con el mundo (ambiente).
- Para esto utilizaremos la idea de **aprendizaje por refuerzo (RL)**.
- La característica principal es aprender de experimentar (explotacion y exploracion) y recompensas diferidas (credit assignment).
- Utilizaremos un método heurístico para gestionar el compromiso entre explotar y explorar: **Thompson Sampling**.
- Utilizaremos estas ideas modelar la forma en la que los individuos en una sociedad toman decisiones y aprenden del conocimiento de los demás.
- Aplicaciones: Efectos cascadas y competencia.

Contenido

Aprendizaje por Refuerzo

- El ambiente siempre es el mismo y no cambia con la interacción con el aprendiz.
- Las caracteristicas principales son: aprender de experimentar (explotacion y exploracion) y recompensas diferidas (credit assignment).
- Tenemos K-brazos (acciones):



Aprendizaje por Refuerzo

- La recompensa $R_t(a)$ de tomar una accion $A_t = a$ en el momento t es una variable aleatoria: $R_t(a)$. El valor esperado es:

$$q_t(a) = E[R_t(a) \mid A_t = a] \tag{1}$$

- El objetivo es elegir la mejor acción en cada periodo t.
- No conocemos la distribución de la recompensa $R_t(a)$ para ninguna acción, de lo contrario sería en principio fácil resolver:

$$\max_a q_t(a) \tag{2}$$

- Sin embargo, si $R_t(a)$ es un proceso estacionario podemos estimar $q_t(a)$ por:

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i(a) I_{[A_i=a]}}{\sum_{i=1}^{t-1} I_{[A_i=a]}} \tag{3}$$

  y cambiar el problema por:

$$\max_a Q_t(a) \tag{4}$$

- En el caso estacionario escribimos $q_t = q_*$

Estrategia ϵ-codiciosa

- Como $Q_t(a)$ es apenas una estimación del verdadero valor, la estrategia anterior (codiciosa) puede no ser óptima en el largo plazo.
- Si en cada periodo, con probabiliad ϵ exploramos otras acciones, esto puede mejorar la probabilidad de elegir de forma óptima en el largo plazo.
- ϵ es una medida de la incertidumbre que se tiene del estimador $Q_t(a)$.
- La estrategia que elige en cada periodo de forma codiciosa con probabilidad $1 - \epsilon$ y explorar con probabilidad ϵ la llamamos ϵ-codiciosa.

Experimento

- La siguiente figura muestra la distribución $R_t(a)$ para diez valores de $q_\star(a)$.

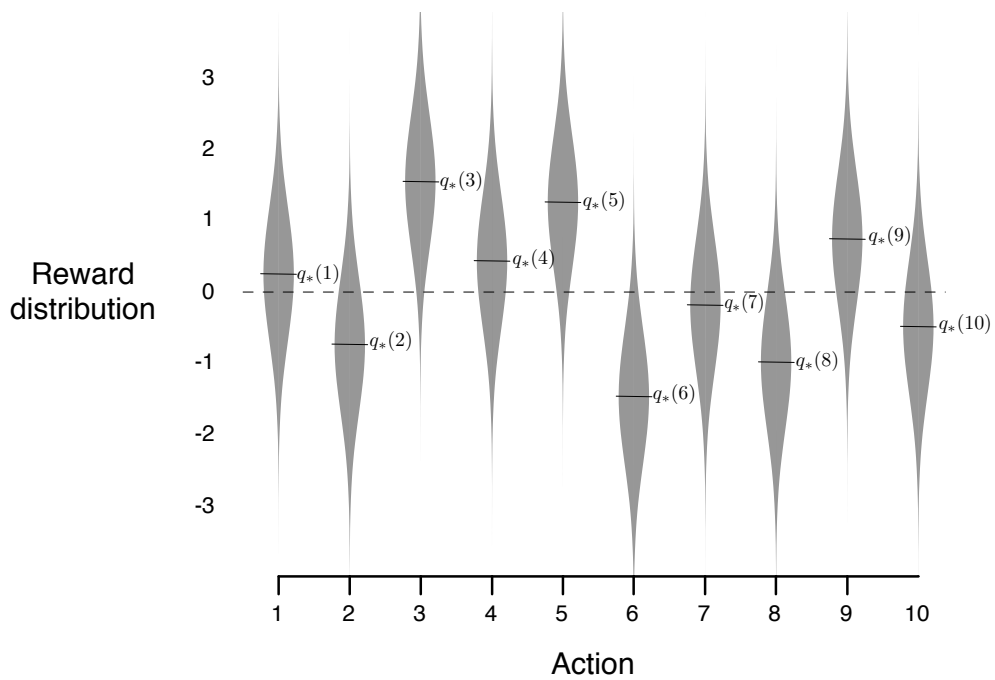


Figure 2.1: An example bandit problem from the 10-armed testbed. The true value $q_*(a)$ of each of the ten actions was selected according to a normal distribution with mean zero and unit variance, and then the actual rewards were selected according to a mean $q_*(a)$ unit variance normal distribution,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

Experimento: Resultados

- En la realidad no conocemos las anteriores distribuciones, solo que tenemos diez acciones disponibles (i.e., nos enfrentamos a un bandido de 10 brazos).
- La siguiente figura muestra los resultados de simular 2000 problemas (bandido de 10 brazos): cada simulacion se obtiene como indica el texto de la figura anterior.
- En cada problema de estos se usan estrategias ϵ-codiciosa por 1000 periodos.

Experimento: Resultados

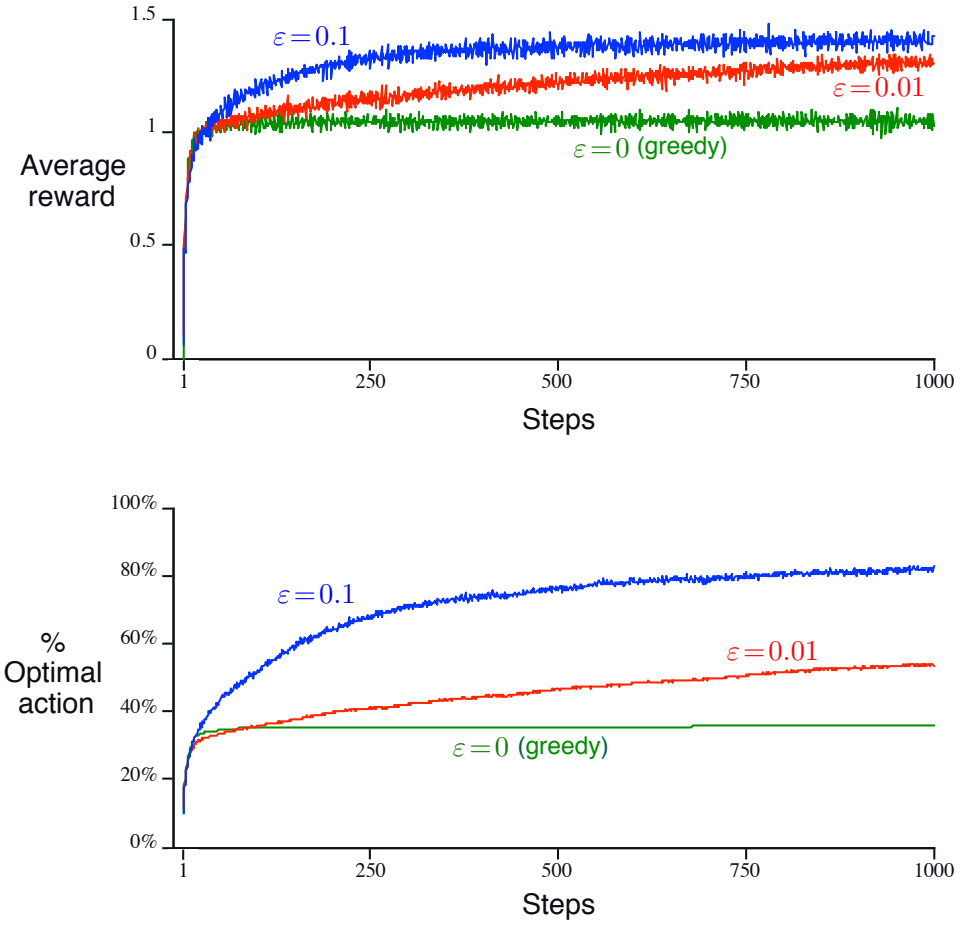


Figure 2.2: Average performance of ε-greedy action-value methods on the 10-armed testbed. These data are averages over 2000 runs with different bandit problems. All methods used sample averages as their action-value estimates.

- La primera gráfica muestra la recompensa promedio y la segunda el porcentaje de veces que cada estrategia seleccionó la estrategia óptima.

Implementación Incremental

- El objetivo es hacer la estrategia anterior computacionalmente eficiente:
  1. En el número de cálculos a realizar.
  2. En memoria.
- Es fácil ver que:

$$Q_{t+1}(a) = Q_t(a) + \frac{I_{[A_t=a]}}{\sum_{i=1}^{t} I_{[A_i=a]}}(R_t(a) - Q_t(a)) \qquad (5)$$

- La forma general de esta estrategia es:

$$\text{Nueva estimación} \leftarrow \text{Estimación anterior} + \text{Tamaño del salto} \times (\text{Recompensa} - \text{Estimación anterior})$$

Estrategia ϵ codiciosa: Pseudocódigo

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

$$A \leftarrow \begin{cases} \arg\max_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

$R \leftarrow bandit(A)$

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

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

Ejemplo: Bernoulli

- Supongamos que tenemos tres armas R_i que se distribuyen Bernoulli, $Bern(\theta_i)$, donde θ_i es desconocido y fijo en el tiempo.
- Cuando se dispara una arma se recibe una recompensa de 1 de lo contrario cero.
- Obsérvese que $E[R_i] = \theta_i$.
- Sigiendo un aproximación Bayesiana, supongamos que θ_i es una variable aleatoria. Esto es una estrategia que usaremos para resolver el problema, no queremos decir que en realidad θ_i sea una variable aleatoria.
- El objetivo es maximizar la suma de las recompensas hasta la ronda T.

Ejemplo: Bernoulli

- Supongamos que después de interactuar con el ambiente, elegir las acciones 1 y 2, 1000 veces y la acción 3, 3 veces, se tiene el siguiente conocimiento sobre la recompensa promedio de cada acción: $E[R_i] = \theta_i$ (i.e., la posterior de θ_i):

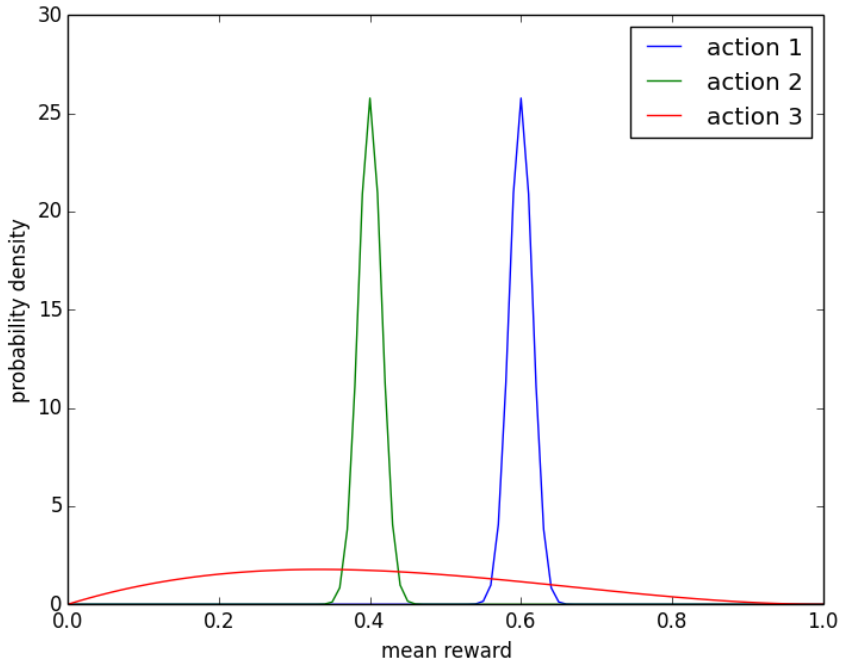


Figura: Densidad de probabilidad sobre recompensas promedio después de elegir las acciones 1 y 2, 1000 veces y la acción 3, 3 veces. Con 600, 400 y 1 ganancia acumulada respectivamente.

Ejemplo: Bernoulli

- En promedio las acción 2 es más alta pero la acción 3 tiene mucha incertidumbre.
- Esta incertidumbre se puede deber a que se ha disparado poco y un algoritmo que no tenga en consideración esto podría no elegir, eventualmente, la mejor acción.
- Un algoritmo ϵ-codicioso explora com la misma probabilidad cada una de la acciones. Esto puede ser ineficiente porque la acción 2 parece estar dominada or la acción 1 mientras que la acción 3 es promisoria.
- Thompson Sampling es una forma de atacar ese problema.

Ejemplo: Caminos más cortos en un grafo

- Una persona desea ir del punto 1 al 12 y los tiempos de desplazamiento son en **promedio** θ_e, donde e es un enlace entre nodos.

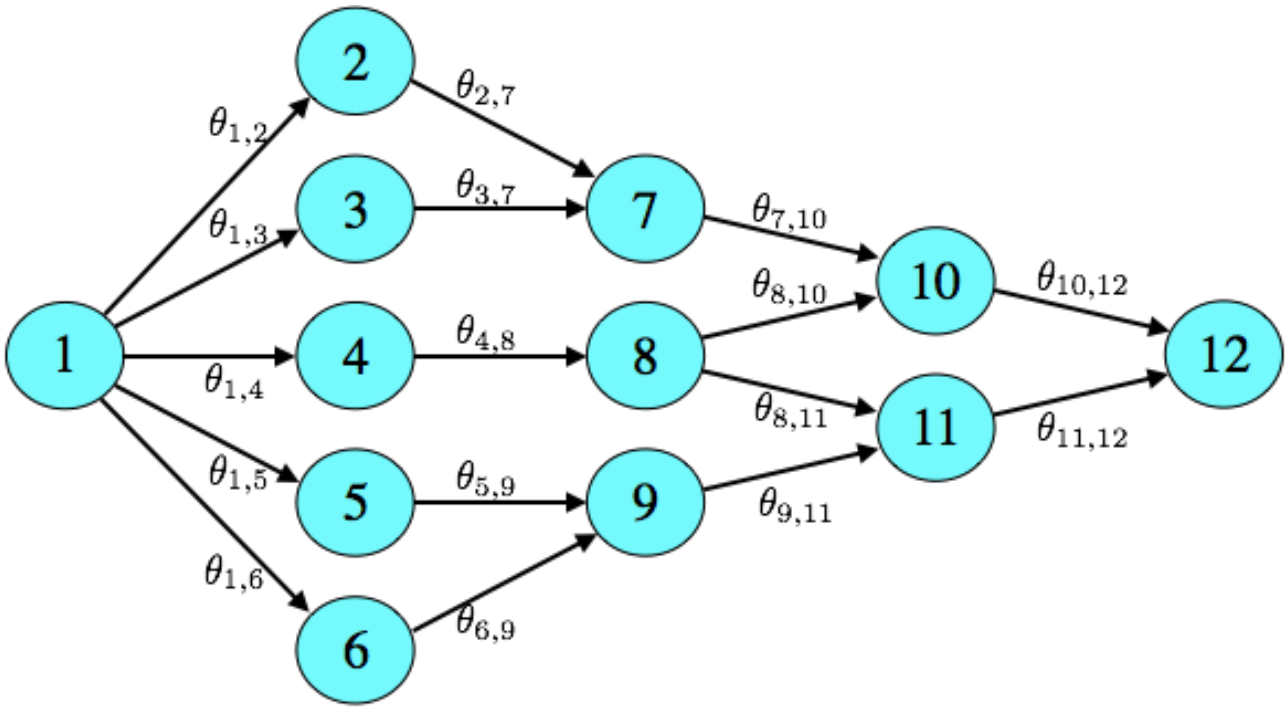


Figura: Camino más corto

- Las acciones son caminos en el grafo entre 1 y 12. Un camino a_t es una sucesion de enlaces $a = (e_1, ..., e_k)$.
- El objetivo es minimizar el valor esperado del tiempo de recorrido: $\sum_{e \in a} \theta_e$

Ejemplo: Caminos más cortos en un grafo

- Las acciones son caminos en el grafo entre 1 y 12. Un camino a_t es una sucesion de enlaces $a = (e_1, ..., e_k)$.
- El objetivo es minimizar el valor esperado del tiempo de recorrido: $\sum_{e \in a} \theta_e$
- Obsérvese que θ_e son los parámetros de interes. La estrategia que usaremos es suponer que son el valor esperado de una variable aleatoria.

Contenido

Ejemplo: Bernoulli

- Supongamos que modelamos θ_k como una distribución Beta con parámetros α_k, β_k ($Beta(\alpha_k, \beta_k)$):

$$p(\theta_k) = \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)} \theta_k^{\alpha_k} (1 - \theta_k)^{\beta_k - 1}$$

- Esta distribución juega un papel instrumental en nuestro objetivo. La actualización de esta distribución, usando el Teorema de Bayes, cuantifica la incertidumbre que se tiene de los parámetros de interes.

Ejemplo: Bernoulli

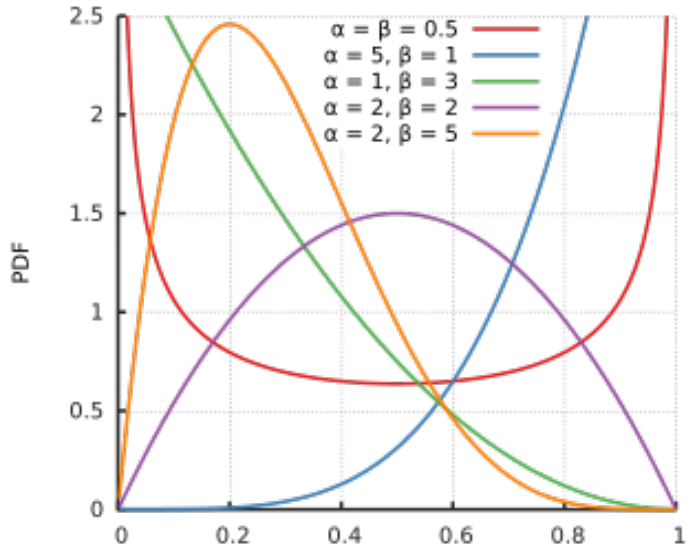


Figura: Beta distribution. By Horas based on the work of Krishnavedala - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=15404515

Ejemplo: Bernoulli

- Por el teorema de Bayes:

$$f(\theta_i \mid R_i) = \frac{f(R_i \mid \theta_i) f(\theta_i)}{f(R_i)}$$

- Si R_i se distribuye $Bern(\theta_i)$ y θ_i $Beta(\alpha_i, \beta_i)$ entonces:

$$f(\theta_i \mid R_i) \text{ se distribuye } Beta(\alpha_i', \beta_i')$$

  donde: $(\alpha_i', \beta_i') \leftarrow (\alpha_i, \beta_i) + (R_i, 1 - R_i)$ si $a = i$, caso contrario los parámetros permanencen invariantes (a es la acción que el agente toma antes de actualizar la distribución de θ_i).
- Observaciones:
  - El parámetro solo se actualiza si se dispara el arma.
  - La ventaja de cuantificar la incertidumbre con la distribucion Beta es que la posterior sigue siendo Beta. Decimos que la distribución de Bernoulli y Beta son distribuciones conjugadas.

Ejemplo: Bernoulli

- Una distribución $Beta(\alpha_i, \beta_i)$ tiene media $\frac{\alpha_i}{\alpha_i+\beta_i}$
- La figura corresponde a:
  $(\alpha_1, \beta_1) = (601, 401), (\alpha_2, \beta_2) = (401, 601), (\alpha_3, \beta_3) = (2, 3)$

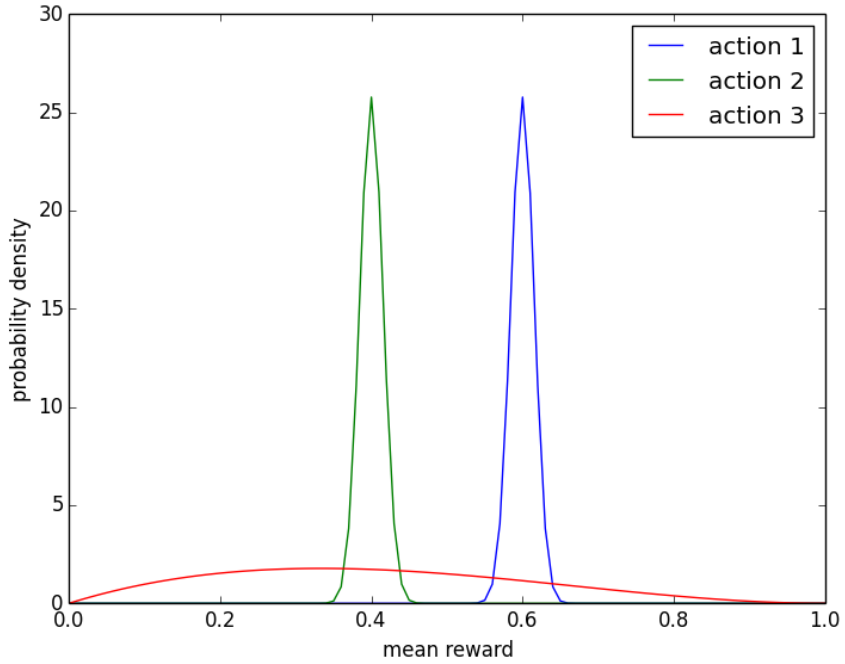


Figura: Densidad de probabilidad sobre recompensas promedio después de elegir las acciones 1 y 2, 1000 veces y la acción 3, 3 veces. Con 600, 400 y 1 ganancia acumulada respectivamente.

Ejemplo: Bernoulli

- Una estrategia greedy (codiciosa) para descubrir la política óptima es:

Algorithm 1 BernGreedy(K, α, β)

for $t = 1, 2, \ldots$ **do**
  #estimate model:
  for $k = 1, \ldots, K$ **do**
    $\hat{\theta}_k \leftarrow \alpha_k / (\alpha_k + \beta_k)$
  end for

  #select and apply action:
  $x_t \leftarrow \text{argmax}_k \, \hat{\theta}_k$
  Apply x_t and observe r_t

  #update distribution:
  $(\alpha_{x_t}, \beta_{x_t}) \leftarrow (\alpha_{x_t} + r_t, \beta_{x_t} + 1 - r_t)$
end for

Algorithm 2 BernTS(K, α, β)

for $t = 1, 2, \ldots$ **do**
  #sample model:
  for $k = 1, \ldots, K$ **do**
    Sample $\hat{\theta}_k \sim \text{beta}(\alpha_k, \beta_k)$
  end for

  #select and apply action:
  $x_t \leftarrow \text{argmax}_k \, \hat{\theta}_k$
  Apply x_t and observe r_t

  #update distribution:
  $(\alpha_{x_t}, \beta_{x_t}) \leftarrow (\alpha_{x_t} + r_t, \beta_{x_t} + 1 - r_t)$
end for

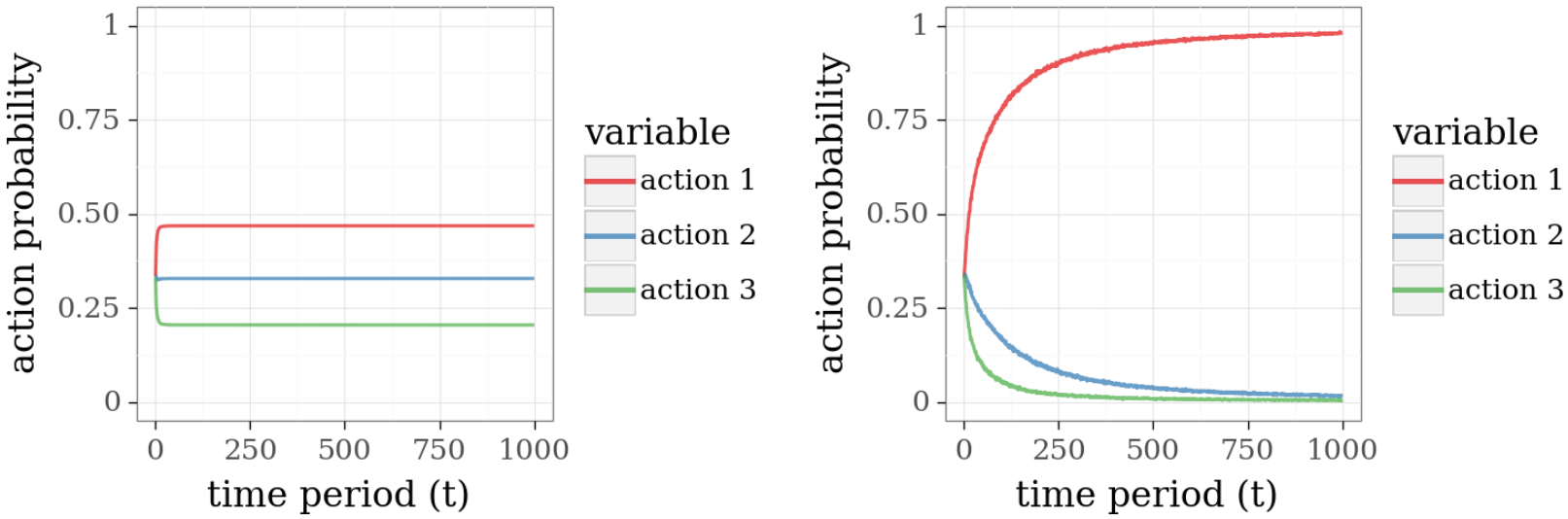
Figura: Bernoulli Codicioso y Bernoulli TS

Ejemplo: Bernoulli

- ¿Por que TS funcionaria?
- La estrategia codiciosa no eligiría la acción 3.
- La estrategia Bernoulli Greddy tampoco eligiría la acción 3.
- Las dos estrategias anteriores con exploración (ϵ codiciosa) le asignariá la misma probabilidad a las tres acciones.
- TS elige las estrategias $1, 2, 3$ con probabilidad $0{,}82, 0, 0{,}18$. Es decir, explora con una probabilidad alta aquellas acciones sobre las que se tiene más incertidumbre.

Ejemplo: Bernoulli

- La siguiente gráfica muestra el desempeño del modelo para los parámetros: $\theta_1 = 0{,}9, \theta_2 = 0{,}8, \theta_1 = 0{,}7$.



(a) greedy algorithm **(b)** Thompson sampling

Figure 3.1: Probability that the greedy algorithm and Thompson sampling selects an action.

Figura: 10,000 mil simulaciones de cada algortimo. Cada simulacion de 1,000 rondas. Cada punto representa la fracción de veces en una ronda específica que el algoritmo seleccionó una acción.

Modelo

- Supongamos que un agente toma una sucesión de acciones $x_1, x_2, ...,$ donde cada $x_i \in \Xi$.
- Depués de la acción i el agente observa un resultado y_i, $y_i \backsim q_{\theta_i}(. \mid x_t)$.
- θ es desconocido pero el agente cuantifica su incertidumbre usando una prior $p(\theta)$.
- El agente recibe una recompensa $r_t = r(y_t)$.
- El objetivo del agente es maximizar el valor esperado de la recompensa: $v_{x_t}(\theta) = E_{q_\theta(.|x_t)}[r_t]$
- El algoritmo general es:

Ejemplo: Bernoulli

Algorithm 3 Greedy$(\mathcal{X}, p, q, r)$

for $t = 1, 2, \ldots$ **do**
  #estimate model:
  $\hat{\theta} \leftarrow \mathbb{E}_p[\theta]$

  #select and apply action:
  $x_t \leftarrow \text{argmax}_{x \in \mathcal{X}} \mathbb{E}_{q_{\hat{\theta}}}[r(y_t)|x_t = x]$
  Apply x_t and observe y_t

  #update distribution:
  $p \leftarrow \mathbb{P}_{p,q}(\theta \in \cdot | x_t, y_t)$
end for

Algorithm 4 Thompson$(\mathcal{X}, p, q, r)$

for $t = 1, 2, \ldots$ **do**
  #sample model:
  Sample $\hat{\theta} \sim p$

  #select and apply action:
  $x_t \leftarrow \text{argmax}_{x \in \mathcal{X}} \mathbb{E}_{q_{\hat{\theta}}}[r(y_t)|x_t = x]$
  Apply x_t and observe y_t

  #update distribution:
  $p \leftarrow \mathbb{P}_{p,q}(\theta \in \cdot | x_t, y_t)$
end for

Figura: TS General

Ejemplo: Bernoulli

- $\Xi = \{1, 2, ...K\}$.
- $y_t = r_t$.
- $q_\theta(1 \mid k) = \theta_k$
- $p(\theta)$ es Beta.

Forma Equivalente de TS

- La siguiente es una forma equivalente del algortimo.
- Sea:

$$w_{xt} = \int I(x = argmax_{x'} v_{x'}(\theta)) p(\theta \mid y_t) d\theta$$

  intutivamente, la probabilidad de x ser óptimo dado y_t.
- Entonces TS se puede implementar de la siguinete forma:
  1. Para cada x estimar w_{xt} por ejemplo usando Montecarlo: muestrear θ de $p(\theta)$ y calcular el promedio.
  2. Para $t+1$ elegir x con probabilidad w_{xt}.
- Esto es equivalente al algoritmo TS introducido anteriormente.
- Con la nueva formulación se tiene una interpretación: si x es óptimo con una probabilidad de w_{xt} entonces con esa probabilidad se elige en la siguiente ronda.

Contenido

Recomendación de Artículos de Noticias

- Una pagina web interactua con una sucesion de usuarios: $t = 1, 2, 3..$
- En cada ronda, el administrador de la página web, observa un vector de características del usuario t, $z_t \in R^d$ (e.g., visitas anteriores, características socio demográficas, geográficas, etc). Este elige un artículo para mostrarle $\Xi = \{1, ..., k\}$ y se observa una recompensa $r_i \in \{0, 1\}$ (i.e., le gusto o no el artículo).
- Suponemos que:

$$P(r_i = 1 \mid x_t = i, \theta_i, z_i) = g(z_t^T \theta_i) = \frac{1}{1 + e^{-z_t^T \theta_i}}$$

- Definimos el arrepentimiento (i.e., error) como:

$$regret_t(\theta_1, ..., \theta_k) = \max_i g(z_t^T \theta_i) - g(z_t^T \theta_i)$$

Recomendación de Artículos de Noticias

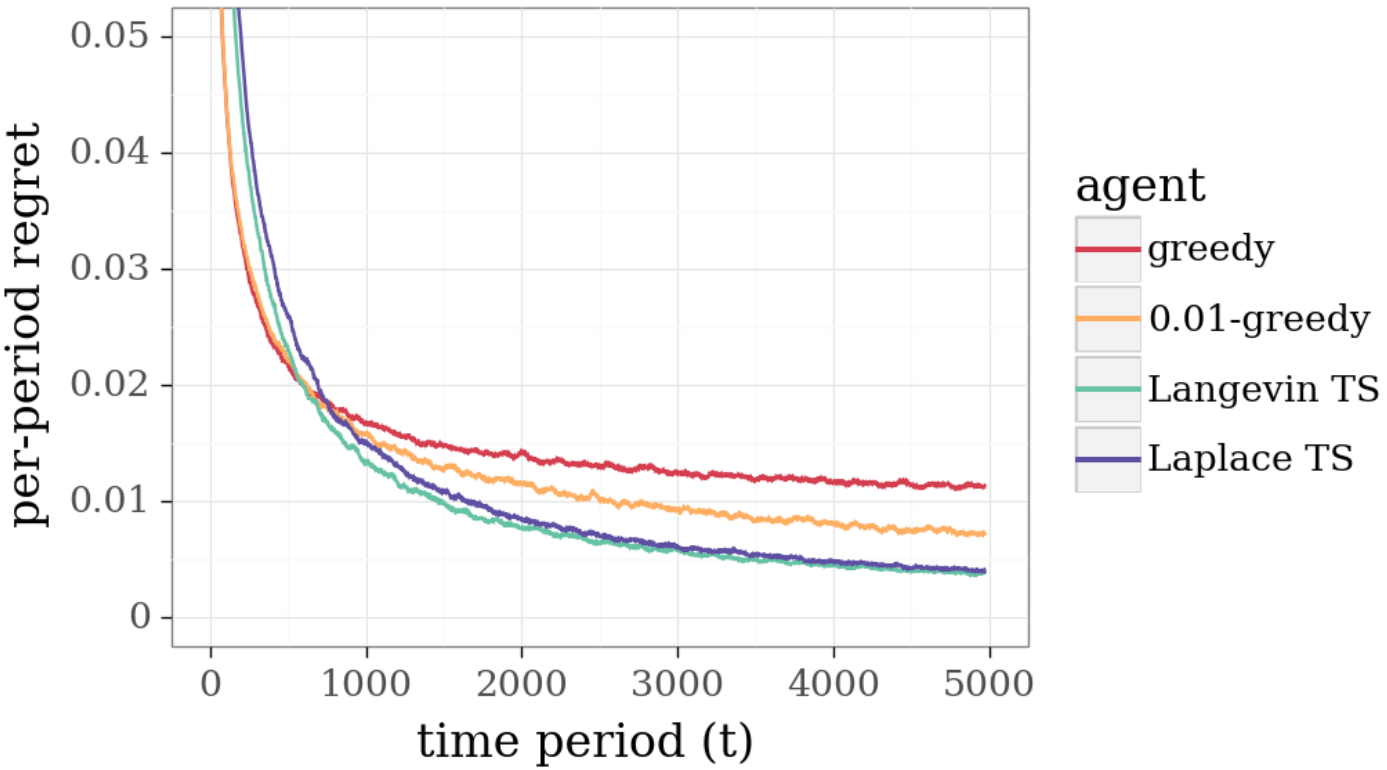


Figure 7.1: Performance of different algorithms applied to the news article recommendation problem.

Figura: $k = 3, d = 7, z_1 = 1$ y las otras seis características son binarias y se generan aleatoriamente con una distribución de Bernoulli con probabilidad de acierto $\frac{1}{6}$. Promedios de 2,000 simulaciones. $\theta_i \sim N(0, I)$.

Modelo

- Un agente tiene para la venta $i = 1, ..., n$ productos. El beneficio neto de vender i es p_i. Dado que los productos pueden ser sustitutos o complementos, el objetivo es armar canastas de productos que maximizen el beneficio neto.
- $x_i \in \{0, 1\}^n$. Variable indicador de que productos ofrece.
- Una vez se ofrecen los productos se observa una demanda d_i.
- Sea θ una matriz $k \times k$.
- Suponemos que: $log(d_i) \mid \theta, x \sim N((\theta x)_i, \sigma^2)$ donde la varianza es conocida.
- Obsérvese que $(\theta x)_i = \theta_{ii} + \sum_{j \neq i} \theta_{ij}$.
- Supongamos que $p(\theta)$ es Normal multivariada (prior conjugada).

Canastas de Productos

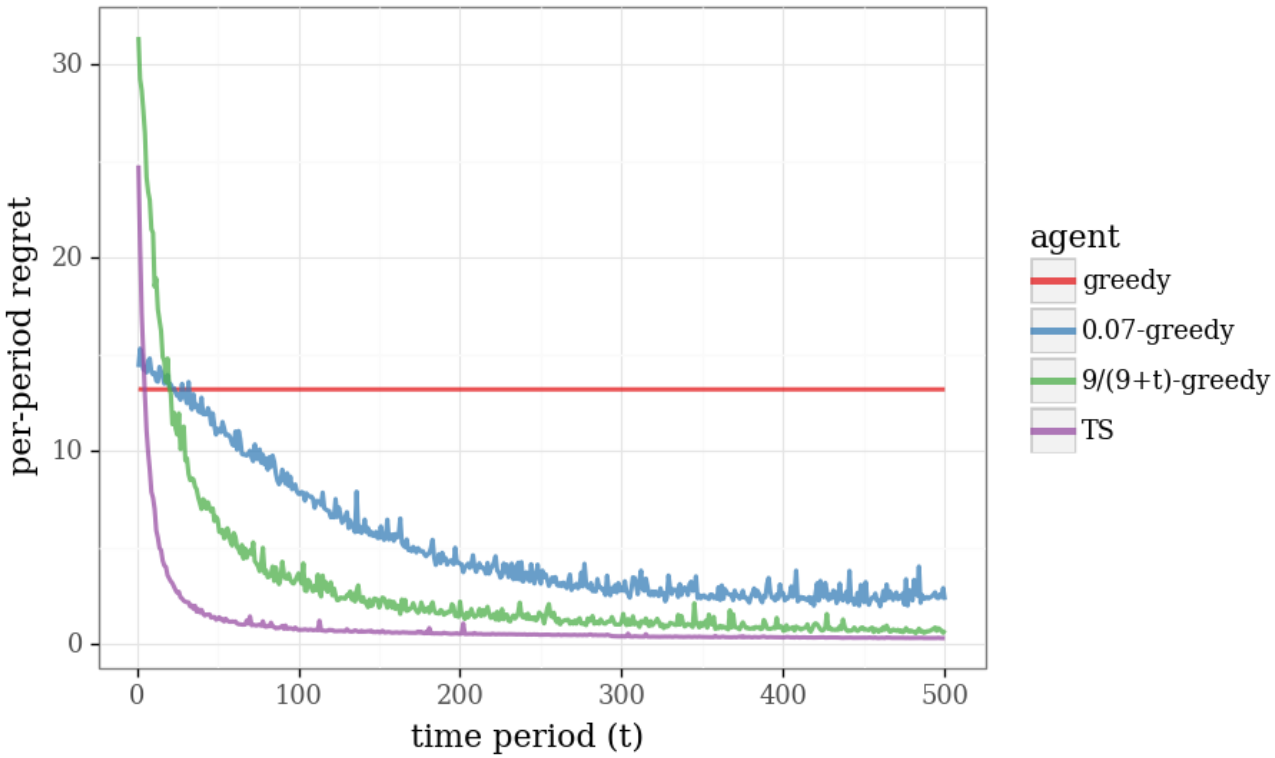


Figure 7.2: Regret experienced by different learning algorithms applied to product assortment problem.

Figura: Canastas de Productos

Introducción

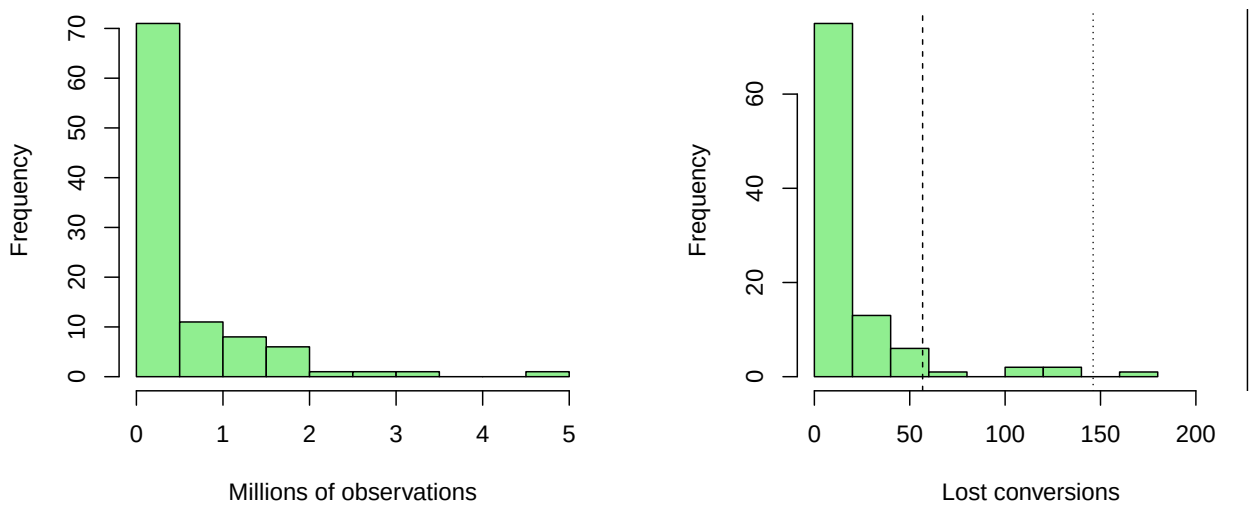
- El sector de servicios es el 80 % del PIB de los Estados Unidos. Una gran parte de este sector involucra transacciones en línea.
- El costo de experimentar lo domina el costo de no prestar el servicio óptimo por estar experimentado en contraposición al costo de obtener observaciones experimentales (usalment con costo marginal cero).
- Otra carcaterística importante es la posibilidad de hacer un monitoreo continuo de cualquier experimento.
- Ilustramos el uso de la teoría de bandidos multiarmados para descubrir el conocimiento de la forma más rápida y económica posible.

A/B Testing

- Dos versiones de una pagina web. La tasa de conversion en cada una es $p = 0{,}001$ y $p = 0{,}0011$.
- Las paginas se asignan de forma aleatoria a cada visitante.
- Para detectar esas diferencias en tasas de conversion aleatorizando 50-50, se necesitan muchas observaciones: Aproximadamente 2.5 millones de observaciones de cada arma para una confianza del 95 % o 0.5 millones cada una con 50 % de confianza.

A/B Testing

- La figura muestra el histograma de 100 simulaciones, 100 visitas a cada pagina por episodio, cada visita se actualiza.
- Cada episodio se interrumpe cuando el arrepentimiento relativo es menor al 5 %.
- En las simulaciones, en el 84 % se eligió la página óptima (B).



(a) (b)

Figure 1: *(a) Histogram of the number of observations required in 100 runs of the binomial bandit described in Section 4.1. (b) The number of conversions lost during the experiment period. The vertical lines show the number of lost conversions under the traditional experiment with 95% (solid), 50% (dashed), and 84% (dotted) power.*

Figura: AB Testing

Contenido

Introducción

- Thompson sampling es optimo en terminos de minimizacion de arrepentimiento.
- La prior se basa en el aprendizaje social.
- Esta puede ser influenciada o manipulada por efectos cascada.
- La dinamica depende de la distribucion completa.
- La clave es fitness para prevenir dominancia de ciertos agentes.
- Video Pentland.

Aprendizaje Social

- N agentes, M opciones cada uno. Recompensas $r_{it} \sim Bernoulli(\eta_i)$. Luego $p(r_{it} = 1 \mid \eta_i) = \eta_i$ (aqui denominamos η_i la calidad de la elección). Sea η^* el mayor. Obsérvese que no sabemos cual es la elección con mayor η_i.
- El agente puede observar la historia de recompensas de una acción en particular o todas las acciones.
- La popularidad de una acción j es: $p_{jt} = \sum_{i=1}^{N} I(x_{i,t-1} = j)$.

Aprendizaje Social: Dos etapas

- **Prior**: Cada agente i elige una opcion candidata j, $o_{it} = j$ de acuerdo a una probabilidad proporcional a la popularida de la opción (i.e., prior social):

$$p(o_{it} = j) = \frac{p_{jt}}{\sum_{k=1}^{M} p_{kt}}$$

- **Aceptar/Rechazar**: El agente i acepta o rechaza la opción j con probabilidad:

$$p(a_{it} = j \mid o_{it} = j) = p(r_{jt} \mid \eta_j = \eta^*) = (\eta^*)^{r_{jt}}(1 - \eta^*)^{(1-r_{jt})}$$

Aprendizaje Social: eToro

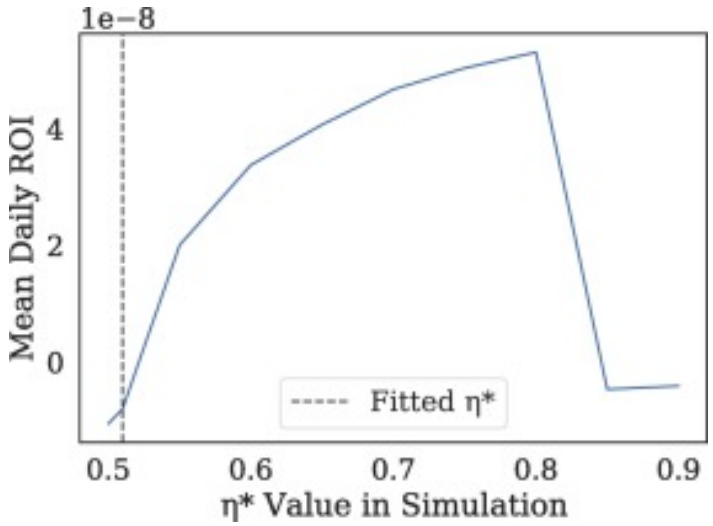


Figura: Fig. 8. Simulated mean daily ROI within a population of ideal social samplers following the traders on eToro over the time period we study, for different values of . These simulations check how well the social sampling model balances exploration versus exploitation. The fitted value of that achieves the best predictive accuracy of eToro follow decisions is suboptimal in terms of mean daily ROI in these simulations.