UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

PROBLEMA DE INTERDICCIÓN SECUENCIAL DE CAMINO MÁS CORTO ESTOCÁSTICO

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

NATALIA NICOLE TRIGO TOMASEVICH

PROFESOR GUÍA:

Denis Sauré Valenzuela

MIEMBROS DE LA COMISIÓN:

Juan Borrero Angarita

Fernando Ordoñez Pizarro

Charles Thraves Cortés-Monroy

Este trabajo ha sido parcialmente financiado por:

Asociación Nacional de Investigación y Desarrollo (ANID)

Instituto de Sistemas Complejos de Ingeniería (ISCI)

SANTIAGO DE CHILE

2023

PROBLEMA DE INTERDICCIÓN SECUENCIAL DE CAMINO MÁS CORTO ESTOCÁSTICO

Esta tesis se enmarca en programación binivel, que corresponde a un problema de optimización que depende de las decisiones de dos partes, *líder* y *seguidor* y, por otro lado, se relaciona con el problema *Multi-armed Bandit*, donde el principal objetivo es escoger qué brazo jalar en una máquina con n opciones y recibir una recompensa aleatoria según la decisión. En el Bandit, el término *Regret* caracteriza la diferencia entre el brazo que se escoge y haber elegido el brazo óptimo, que en principio es desconocido. La linealidad del Regret (Lai y Robbins, 1985) permite extrapolar dicho problema a una configuración de caminos, donde el costo de cada camino viene de una distribución de probabilidad. El objetivo principal en este caso es bloquear caminos (escoger un brazo) para encarecer la ruta de un evasor (recompensa). A esto se le denomina el problema de "Interdicción de camino más corto estocástico". Este tipo de configuración permite, por ejemplo, simular el bloqueo de caminos de contrabando, adoptando una estrategia para hacer menos atractivo el negocio desde los costos.

En la configuración de Bandit clásico, surge la disyuntiva de *exploración* o *explotación* que trata de decidir si se exploran distintos brazos hasta encontrar el óptimo o si se explota algún brazo cuya ganancia ya se conoció en algún periodo. Este *trade off* entre exploración y explotación se extiende al problema de interdicción, motivando a descubrir una cota inferior de desempeño de políticas de decisión y estimar el costo de la exploración. En otras palabras, cuánto se debe explorar para asegurar la optimalidad de una política de interdicción que se traduce en términos de Regret. Se establece un límite fundamental para el desempeño asintótico de políticas de decisión admisibles y se comparan distintas políticas mediante simulaciones.

A la Nati de 5 años...

Agradecimientos

Quiero agradecer primero a mi familia, en especial a mis padres por apoyarme siempre y por darme la oportunidad de estudiar. A mi mamá por enviarme comida todas las semanas y a mi papá por venir a dejarme todos los domingos en la noche aunque estuviese cansado. Son, sin duda alguna, unos padres únicos y los amo mucho.

También agradezco a mi pareja, Francisco, por creer en mí y por destacar siempre que puedo hacer todo lo que me proponga. Gracias por ser mi compañero mi y mi mejor amigo. Te amo. Gracias también a su familia hermosa por su apoyo.

Agradezo a mis perritos hermosos, mis bebés, por darme amor y por renovar mi energía cada vez que los veía los fines de semana.

Gracias a mi Maestro Jedi, Denis Sauré, por apostar por mí, por potenciarme a ser mejor alumna, por brindarme todas las oportunidades de crecimiento en la universidad y por inspirarme a seguir un futuro académico.

Gracias a ANID por otorgarme la Beca de Magíster Nacional y gracias a ISCI por financiar mi viaje a CLAIO 2022.

Gracias Martín Valdevenito, por ser como un hermano en todo mi proceso universitario. A Dani Shcumacher, por ser inspiradora y sabia. Gracias a Benja por dar la hora; Coni por tu dulzura; Dani Iriarte por tu energía; Exe por ser un genio; Jo por ser honesta y siempre dar buenos consejos; Maite por siempre estar ahí para mi; Pablo por tu buena onda y experiencia.

Gracias a todos los amigos y amigas que fueron parte de mi vida universitaria.

Gracias a Hans Zimmer por acompañar mis horas de estudio con su música inspiradora. A la Queen Taylor Swift por inspirarme a ser una mujer fuerte y por sus letras infinitamente profundas.

And last but not least, I wanna thank me (Snoop Dog).

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

En esta tesis se estudia el problema de interdicción de camino más corto estocástico en el contexto de programación binivel. En este tipo de problemas, periodo a periodo, un líder decide sobre un conjunto de acciones y, luego de observar dicha decisión, un seguidor responde. El líder quiere maximizar su ganancia, que dependerá de la respuesta del seguidor, mientras que el seguidor quiere minimizar costos, que dependerán de la acción del líder.
La interdicción de camino más corto se considera binivel ya que existe un interdictor que pretende bloquear caminos a un evasor. La ganancia del interdictor está en maximizar el costo del camino total recorrido por el evasor, mientras que el evasor busca minimizar dicho costo. Por otro lado, esta configuración puede ser tratada como un problema de *Multi-Armed Bandit (MAB)*, que describe una secuencia de decisiones sobre una máquina de n brazos, donde cada brazo entrega una recompensa cuyo valor proviene de una distribución de variable aleatoria, y en cada periodo se decide qué brazo jalar. A través de ello, se define el término Regret que caracteriza la diferencia en ganancias entre el brazo que se escoge y haber elegido el brazo óptimo, que en principio es desconocido. La linealidad del Regret (Lai y Robbins, 1985) permite extrapolar dicho problema a una configuración de caminos, donde el costo de cada camino viene de una distribución de probabilidad, análogo a los brazos en el problema clásico de MAB.

Mediante la adaptación de los problemas binivel y Multi-Armed Bandit en la configuración de interdicción de caminos, se busca encontrar una cota inferior para el desempeño de diversas políticas de decisión de manera teórica y aplicarla a resultados numéricos mediante simulaciones. Esta cota se obtiene a través de la adaptación de políticas de los problemas binivel

y Multi-Armed Bandit. Para esto, se pide que las políticas de interdicción sean consistentes, es decir, que mantengan un desempeño similar para todas las configuraciones. Además, se estudian políticas que se limitan a un conjunto de vectores de costos aleatorios que logran que, ante cierta interdicción, el evasor mantenga el mismo recorrido bajo distintos elementos de este conjunto, con el fin de comparar el desempeño frente a otras políticas que consideran todo tipo de vectores aleatorios.

A través de este trabajo se generan las siguientes contribuciones:

- Se encuentra una cota inferior para el desempeño de diversas políticas de interdicción. Esta cota permite obtener un parámetro acerca de la cantidad mínima de exploración de arcos que hay que hacer para garantizar optimalidad de una política de interdicción y, por ende, estimar el costo de dicha exploración.

- Adaptar políticas del MAB clásico al contexto de interdicción de camino más corto con estocasticidad y darle garantías de optimalidad.

- Se modela del conjunto de vectores que logran consistencia en las respuetas del evasor, es decir, el conjunto de vectores que, mientras no cambie la decisión de interdicción, hace que el evader prefiera el mismo camino. Esto permite que a medida que se simulan costos, las acciones se mantengan consistentes. El modelamiento se hace mediante formulación de problemas de optimización mixta entera.

- Se simulan instancias de grafos (caminos) tomando mil periodos de ejecución, para los cuales se aplican políticas de decisión en contexto de interdicción de camino más corto y se resuelven los problemas de optimización mixta entera mencionados más adelante. Se presentan resultados en términos de Regret acumulado con su respectivo error, haciendo un benchmark entre las políticas.

Capítulo 2

Literatura relevante

Multi-Armed Bandit: El trabajo de Lai y Robbins (1985) describe el problema tradicional de Multi-Armed Bandit. Un problema clásico de decisión bajo incertidumbre. En esta formulación un apostador intenta maximizar su ganancia acumulada jalando brazos de una máquina tragamonedas secuencialmente, donde la información previa que tiene acerca de la distribución de ganancias de cada brazo es limitada. El apostador enfrenta el *trade off* de jalar el brazo que le ha dado la mejor recompensa hasta un determinado periodo (explotación) o de jalar otro brazo que no ha escogido antes (exploración), lo cual permitiría identificar otra opción con una recompensa mayor pero a costas de arriesgar la maximización de ganancia total.

Gittins (1979) muestra que en el caso de recompensas de brazo independientes y horizonte infinito descontado, la política óptima es del tipo índice. Desafortunadamente, este tipo de políticas no siempre son óptimas (ver Berry y Fristedt (1985)y Whittle (1982)) o no se pueden calcular de forma cerrada. En su trabajo seminal,Lai y Robbins (1985) estudian políticas asintóticamente eficientes para el caso no descontado. Establecen un límite fundamental sobre el rendimiento alcanzable, lo que implica la optimalidad (asintótica) de la dependencia del orden $ln(N)$ (donde N es el número total de períodos) en el Regret (ver Kulkarni y Lugosi (1997)).En el mismo contexto, Auer, Cesa-Bianchi, y Fischer (2002) introducen la celebrada política UCB1 basada en índices, que es eficiente y se puede implementar.

El trabajo de Modaresi, Sauré, y Vielma (2020) corresponde a una aplicación combinatorial del problema de Bandit clásico. En este contexto, se busca aplicar dicho problema a una configuración de elección de caminos con el fin de saber qué se debe explorar y cómo hacerlo

de manera eficiente, resolviendo un problema de cota inferior (*Lower Bound Problem*) al igual que esta tesis.

Otros trabajos influyentes en el tema son Chen, Wang, y Yuan (2013) que proporcionan un límite de rendimiento más ajustado para la política tipo UCB1 de Gai, Krishnamachari, y Jain (2012), la cual extienden al ajuste combinatorio de Modaresi et al. (2020). También está Liu, Vakili, y Zhao (2012) que propone políticas de optimización de redes estocásticas, Cesa-Bianchi y Lugosi (2012) que estudian Bandit combinatorial adversarial y Anantharam, Varaiya, y Walrand (1987) que estudian Bandit combinatorial donde el elemento a estudiar se puede tomar como jalar muchos brazos simultáneamente.

Programación Binivel: La programación binivel ha tenido diversas aplicaciones incluído en áreas como defensa (Brown, Carlyle, Salmerón, y Wood, 2006), economía (Sherali, Soyster, y Murphy, 1983), transporte (Lucotte y Nguyen, 2013), revenue management (Côté, Marcotte, y Savard, 2003) (ver trabajo de Colson, Marcotte, y Savard (2005) y sus referencias) y bloqueo de rutas de contrabando o tráfico como señalan Borrero, Prokopyev, y Sauré (2021) o Steinrauf (1991).

El trabajo de Borrero et al. (2021) se enmarca en la programación binivel donde un líder y un seguidor interactúan a lo largo de múltiples períodos de tiempo. En cada período, el seguidor observa las acciones tomadas por el líder y reacciona de manera óptima, de acuerdo con su propia función objetivo, que inicialmente es desconocida para el líder. Al observar varias formas de retroalimentación de información a partir de las acciones del seguidor, el líder es capaz de refinar su conocimiento sobre la función objetivo del seguidor y, por lo tanto, ajustar sus acciones en períodos de tiempo subsiguientes, lo que debería ayudar a maximizar el beneficio acumulado del líder. En este contexto, proponen diversas políticas que se pueden resolver mediante programación lineal entera-mixta. Similar al trabajo que se verá más adelante.

Israeli y Wood (2002) presenta un trabajo de programación Binivel, donde estudian el problema de interdictar arcos con el fin de maximizar el camino más corto de una red. Esto lo resuelven mediante un problema de programación entera-mixta. Otros trabajos en este tema lo presentan Chern y Lin (1995); Wood (1993); Lim y Smith (2007); Bayrak y Bailey (2008) y Smith y Song (2020).

Políticas de interdicción: En la investigación de Auer et al. (2002) y de Russo y Van Roy (2014) se muestran distintas políticas aplicadas a problemas de Multi-Armed Bandit. Auer

et al. (2002) propone políticas UCB mientras que Russo y Van Roy (2014) propone tanto políticas UCB como de Thompson Sampling.

Capítulo 3

Formulación del problema

El problema de interdicción de camino más corto secuencial se modela sobre un grafo G en un horizonte finito de T periodos. En cada periodo $t \in T$ el interdictor (líder) bloquea una cantidad finita de arcos B^t, a lo que el evasor (seguidor) responde recorriendo un camino $S^t \in \mathcal{P}(B^t)$. Se denota $G^t \equiv G(B^t)$ como el grafo disponible al evasor en el periodo $t \in T$, es decir, luego del bloqueo del interdictor.

3.1. Respuesta del evasor

Se asume que en cada periodo $t \in T$ el evasor enfrenta una función de costos lineal, regulada por un vector de costos aleatorio $\boldsymbol{c}^t(\omega) := (c_a^t(\omega) : a \in A)$, por ende, el costo asociado al camino S en periodo t está dado por

$$r(S, \boldsymbol{c}^t) := \sum_{a \in S} c_a^t \qquad \text{con } S \in \mathcal{P}(B), t \in T.$$

En adelante se obviará la dependencia en $\omega \in \Omega$ de elementos aleatorios cuando el contexto lo aclare.

Se asume que los vectores $\{\mathbf{c}^t : t \in T\}$ forman una secuencia i.i.d y se denota F como la distribución (común) de $\{\mathbf{c}^t, t \in T\}$. Sea $\mathbb{C} := \prod_{a \in A}[l_a, u_a]$ el soporte de F, donde l_a y u_a son las cotas superior e inferior en los costos del arco $a \in A$, respectivamente. Se considera el siguiente supuesto de la información que conoce el evasor inicialmente.

Supuesto 3.1 *El evasor conoce F pero no observa $\mathbf{c}^t$, por ende, escoge el camino minimi-*

zando costo esperado en $t \in T$.

Considerando el Supuesto 3.1, se asume que luego de observar G^t, el evasor escoge ($S^t \in \mathcal{P}(B^t)$) $\equiv \mathcal{S}(B^t)$, minimizando su costo esperado. Sea $\mu := \mathbb{E}_F[\mathbf{c}^t]$, se tiene que el evasor escoge

$$S^t \in argmin\{r(S, \mu) : S \in \mathcal{P}(B^t)\}.$$

Es importante notar que, condicional en la acción del interdictor, el evasor resuelve el problema determinista de camino más corto en cada periodo. A pesar de este hecho, la información del setup es tal que el interdictor no es capaz de anticipar la respuesta del evasor, como se detalla en la siguiente sección.

3.2. Decisión del interdictor

Se considera una cantidad máxima de bloqueos $K < \infty$. En periodo $t \in T$ el interdictor escoge $B^t \in \mathcal{B} := \{B \subset A : |B| \leq K\}$, es decir, $\mathcal{B}$ es el conjunto factible de acciones de interdicción. Se asume que el interdictor está interesado en maximizar el costo acumulado en el que incurre el evasor y también lo detallado en el Supuesto 3.2 acerca de la información disponible para el interdictor al momento de escoger B^t.

Supuesto 3.2 *El interdictor no conoce* F*, pero conoce su soporte* $\mathbb{C}$*. Observa* S^t *y* $\{c_a^t : a \in S^t\}$ *inmediatamente después de que el evasor recorre* S^t *en* $t \in T$.

Se debe notar que si el interdictor tiene acceso a F implementaría $B^t = B^*$ para todo $t \in T$, donde

$$B^* \in argmax\{min\{r(S, \mu) : S \in \mathcal{P}(B)\} : B \in \mathcal{B}\}.$$

La formulación anterior se denomina como *full information problem* o problema de información completa, donde B^* corresponde a la solución al problema en el cual se conocen los costos medios del grafo y el líder es capaz de predecir cuál será la respuesta del evasor. De manera similar, se define la respuesta del evasor al problema de información completa como

$$S^* := argmin\{r(S, \mu) : S \in \mathcal{P}(B^*)\},$$

donde se asume que es una solución única, para evitar problemas de notación.

El interdictor no puede implementar la solución del problema de información completa desde el principio dado que inicialmente F es desconocida, y por ende, también lo es μ. En este sentido, las acciones del interdictor están ajustadas a la información que esté disponible al inicio del periodo t.

Sea

$$\mathcal{F}^t := \sigma((S^s, \{c_a^s : a \in S^s\}) : s < t), \quad t \in T,$$

y sea $\mathcal{F} := \{\mathcal{F}^t : t \in T\}$ la filtración generada por la información revelada al interdictor en su interacción con el evasor, se dice que $\pi := \{\pi^t : t \in T\}$ es una política de interdicción **admisible** si es un proceso estocástico adaptado a $\mathcal{F}$, tal que para todo $t \in T$, se tiene que $B^{t,\pi} \equiv \pi^t \in \mathcal{B}$.

Observación Una política de interdicción que implementa B^* en cada periodo es admisible podría tener un mal desempeño si se implementa una distribución alternativa de costos para la cual B^* no es la solución al problema de información completa. En el siguiente capítulo, el problema se restringe a las políticas que tienen **consistentemente** un buen desempeño para todas las instancias.

Siguiendo la literatura tradicional, se asume que el líder está interesado en minimizar el **Regret esperado**, asociado a sus acciones de interdicción. Esto es, para una distribución F y un horizonte T, se define el Regret esperado asociado a una política admisible π como

$$\mathcal{R}^\pi(T, F) := T \cdot r(S^*, \mu) - \sum_{t=1}^{T} r(S^t, \mu)$$

que depende de F mediante $\mu = \mathbb{E}_F\{\mathbf{c^t}\}$. Cuando el contexto sea claro, se omitirá dicha dependencia.

Capítulo 4

Límite fundamental de desempeño

En este capítulo se establece un límite fundamental de desempeño para cada política admisible y *consistente* (definido más adelante). Para esto, se debe considerar la información que se puede recuperar sobre μ mediante la observación de las respuestas del evasor ante las acciones de interdicción empleadas en tiempo finito y de los costos de dichas respuestas. Para esto, se considera el Supuesto 4.1 acerca del comportamiento del evasor, declarando que sus decisiones son consistentes en el tiempo.

Supuesto 4.1 *Si $B^t = B^s$ para $t \neq s$ entonces $S^t = S^s, s, t \in T$.*

Cuando el interdictor bloquea los arcos en B^t en periodo $t \in T$, independiente del vector $\mathbf{c^t}(\omega)$, el evasor escoge una respuesta S^t que, bajo el Supuesto 4.1, es el mismo para todos los periodos $s \in T$ para los cuales $B^s = B^t$. Con esto en mente, se considera que se puede recolectar información acerca de μ mediante la observación de las respuestas del evasor para cada acción de interdicción en tiempo finito.

Sea $S(B)$ la respuesta del evasor cuando el interdictor bloquea los arcos en $B \subseteq \mathcal{B}$, luego de implementar todas las soluciones de interdicción posibles, el interdictor conoce que

$$\mu \in \mathcal{U} := \{\nu \in \mathbb{C} : r(S(B), \nu) \leq r(S, \nu),\ S \in \mathcal{P}(B),\ B \in \mathcal{B}.\} \tag{4.1}$$

Se permite asumir que el interdictor conoce que $\mu \in \mathcal{U}$, ya que esta información es determinista (por el Supuesto 4.1) y puede ser obtenida en tiempo finito.

Se puede capturar información adicional de μ mediante la observación iterada de realizaciones de costos de los caminos escogidos. Por antecedentes de otros trabajos en problemas

de Bandit, se sabe de antemano que recolectar esa información requiere incurrir en un costo en términos de Regret, y será útil para identificar configuraciones en las cuales B^* es óptimo y en las que no lo es.

4.1. Cota inferior de desempeño

Para establecer cota, se consideran aquellas políticas que tienen **consistentemente** un buen desempeño, siguiendo el trabajo hecho por investigadores en problemas de Multi-Armed Bandit, Lai y Robbins (1985). En particular, se consideran aquellas políticas π que para todas las distribuciones F y para todo $\alpha > 0$,

$$\mathcal{R}^{\pi}(t, F) = o(t^{\alpha}).$$

Este set de políticas consistentes excluye aquellas que tienen un buen desempeño para una configuración en particular pero un mal desempeño en otras. Considerando lo anterior, se define la siguiente clase de arcos minimal tal que, si los costos de dichos arcos no son conocidos, el líder no puede asegurar la optimalidad de la solución al problema de información completa. Se define entonces $\mathcal{E} \subseteq 2^A$ que contiene a todos los conjuntos $E \subseteq A \setminus S^*$ tal que si los componentes de μ asociados a los arcos en E cambian, es posible que B^* no sea óptimo, i.e

$$E \in \mathcal{E} \Longleftrightarrow \exists \nu \in \mathcal{U} : \nu_a = \mu_a \text{ para } a \in A \setminus E \text{ y } r(S^*, \nu) < \max\{r(S(B), \nu) : B \in \mathcal{B}\}$$

El siguiente ejemplo ilustra la estructura de la clase $\mathcal{E}$.

Ejemplo 4.1 Considerar la configuración de interdicción mostrada en la Figura 4.1. Bajo el supuesto de que se pueden bloquear $K < k-1$ arcos, con $l_a = 0$ y $u_a = \infty$ para todo $a \in A$. Se definen los subrecorridos $l(j) \equiv (1, j+1) \to (j+1, k+1)$ y $r(j) \equiv (k+1, j+k+1) \to (j+k+1, 2k+1)$, para todo $j < k$. Sin pérdida de generalidad, se asume que

$$\sum_{a \in l(j)} \mu_a < \sum_{a \in l(j+1)} \mu_a \quad \text{y} \quad \sum_{a \in r(j)} \mu_a < \sum_{a \in r(j+1)} \mu_a, \quad \text{j < k-1.}$$

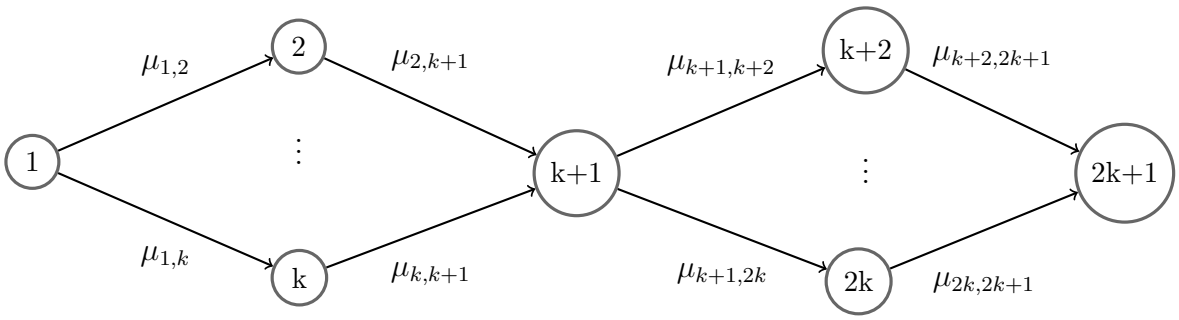


Figura 4.1: Configuración de interdicción del Ejemplo 4.1. Cada arco $a \in A$ está etiquetado como μ_a, su costo esperado bajo F.

Para cualquier $j < K$ se define

$$\mathcal{B}(j) := \{B \in \mathcal{B} : B \cap l(i) \neq \emptyset, i \leq j \wedge B \cap r(i) \neq \emptyset, i \leq K - j\}.$$

Fijando $j \leq K$, se escoge μ tal que $B^* \in \mathcal{B}(j)$, i.e. es óptimo bloquear el subrecorrido j más corto entre los $l(\cdot)'s$ y el subrecorrido más corto $K - j$ entre los $r(\cdot)'s$. Esto implica que $S^* = l(j+1) \to r(K-j+1)$.

Se define $E_{i,j} := \{a \in A : a = (1, j'), j' > i\}$ para $i > j$ y $E_{i,j} := \{a \in A : a = (k+1, k+1+j'), j' > K-i\}$ para $i < j$. Finalmente, para $i \neq j$ se considera como vector de costos medios ν^i donde

$$\nu^i_a = \begin{cases} \mu_a + M & a \in E_{i,j} \\ \mu_a & \sim, \end{cases}$$

Para algún $M >> 0$.

Notar que ν^i pertenece $\mathcal{U}$ y es tal que la decisión de interdicción óptima bajo el vector de costos ν^i pertenece a $\mathcal{B}(i)$. Luego, se concluye que $\mathcal{E}_{i,j} \in \mathcal{E}$. Sumado a esto, dada la estructura de G, es posible definir $2^{|E_{i,j}|}$ conjuntos equivalentes a $E_{i,j}$, pues como ν^i permanece en $\mathcal{U}$, al incluir los arcos $(1, j')$ o $(j', k+1)$ cuando $i > j$ y los arcos $(k+1, k+1+j')$ o $(k+1+j', 2k+1)$ cuando $i < j$, la decisión de interdicción óptima bajo ν^i sigue permaneciendo en $\mathcal{B}(i)$. Llamando a estas clases equivalentes como $\mathcal{E}_{i,j}$, se concluye que $\bigcup_{i \neq j} \mathcal{E}_{i,j} \in \mathcal{E}$.

Se impone que $E \in \mathcal{E}$ es minimal, es decir, que no existe un conjunto $E' \in \mathcal{E}$ tal que $E' \subset E$. En el contexto del Ejempo 4.1, esto implica que solo se puede asegurar que no existe

$E' \subset E_{0,j}$ y $E'' \subset E_{K,j}$ tal que $E' \in \mathcal{E}$ y $E'' \in \mathcal{E}$. En cualquier otro caso, la minimalidad del conjunto $E_{i,j}$ depende de μ de forma no trivial.

Dado que la información de los costos de los arcos en $E \in \mathcal{E}$ ayuda a descartar decisiones que son subóptimas, es imperativo que las políticas consistentes recolecten información de manera persistente sobre al menos un arco en E, para todo $E \in \mathcal{E}$. De esta forma, se puede chequear la optimalidad de B^* y, por consiguiente, alcanzar optimalidad asintótica.

Se usa esta construcción para establecer una cota inferior de desempeño para la frecuencia a la cual se deben observar los costos de ciertos arcos en E, para todo $E \in \mathcal{E}$.

Por construcción, para todo $E \in \mathcal{E}$ existe al menos un costo alternativo bajo el cual B^* deja de ser óptima. El siguiente lema establece que la decisión de interdicción óptima bajo este costo alternativo debe, necesariamente, bloquear el camino S^*, y viceversa, la respuesta óptima del evasor bajo esta nueva configuración debe ser interdictada por B^*.

Lema 4.1 *Sea $\mathcal{E} \neq \emptyset$. Para cada $E \in \mathcal{E}$, existe $\nu \in \mathcal{U}$ tal que $B^*(\nu) \neq B^*(\mu), S(B^*(\nu)) \cap E \neq \emptyset, S(B^*(\nu)) \notin \mathcal{P}(B^*(\mu))$ y $S(B^*(\mu)) \notin \mathcal{P}(B^*(\nu))$.*

En el lema anterior, se hace un abuso de notación denotando como $B^*(\nu)$ a la solución de interdicción óptima cuando el vector de costos medios está dado por ν.

Para cada $E \in \mathcal{E}$ y $t \in T$, se define $\tau(E,t)$ como el número de periodos hasta t (incluído), tal que la respuesta del evasor pasa por un arco en E. Esto es,

$$\tau(E,t) := \sum_{s=1}^{t} \mathbf{1}\{S^t \cap E \neq \emptyset\}.$$

El siguiente resultado formaliza la noción de que cualquier política consistente debe explorar arcos en E persistentemente y especifica la frecuencia asintótica óptima para hacerlo.

Teorema 4.1 *Sea $\mathcal{E} \neq \emptyset$. para cada $E \in \mathcal{E}$ existe una constante $\kappa_E < \infty$ tal que para cada política consistente π se tiene que,*

$$\limsup_{t\to\infty} \mathbb{P}\left\{\frac{\tau^\pi(E,t)}{\ln t} \leq \kappa_E\right\} = 0.$$

El Teorema 4.1 implica que la consistencia de una política requiere de la observación persistente de realización de costos de los arcos que no son observables cuando se escoge la

solución al problema de información completa, trayendo como consecuencia un mínimo crecimiento del Regret conforme crece t. Dicho teorema puede ser escalado para establecer una cota fundamental de desempeño alcanzable. Para esto, se define

$$\Delta_B := r(S^*, \mu) - r(S(B), \mu), \quad B \in \mathcal{B}.$$

El siguiente corolario es resultado del Teorema 4.1 por lo que no se adjunta demostración.

Corolario 4.1 *Para toda política consistente π se tiene que*

$$\liminf_{t\to\infty} \frac{\mathcal{R}^\pi(t, F)}{\ln t} \geq \kappa,$$

donde κ es el valor objetivo del siguiente Lower Bound Problem (LBP),

$$\begin{aligned} \kappa := \quad \min \quad & \sum_{B\in\mathcal{B}} x_B \, \Delta_B \\ s.t. \quad & y_a \leq \sum_{B\in\mathcal{B}:\, a\in S(B)} x_B, \quad a \in A \\ & \sum_{a\in E} y_a \geq k_E, \quad E \in \mathcal{E} \\ & y_a \in \mathbb{R}_+,\, a \in A,\, x_B \in \mathbb{R}_+,\, B \in \mathcal{B} \end{aligned}$$

En la literatura de problemas de Bandit combinatorial, los límites de desempeño alcanzable son expresados en términos de la cardinalidad de sus conjuntos, con la excepción en el trabajo de Modaresi et al. (2020), donde las cotas presentadas dependen de formulaciones similares al *LBP* presentado anteriormente. En este sentido, dado que el crecimiento logarítmico en T de la cota inferior del Regret es común en la literatura de Bandit, se puede tomar el Teorema 4.1 como prueba de que esta dependencia en el tiempo, en el caso de interdicción de camino más corto, será también de caracter logarítmico, bajo el contexto de que cada potencial decisión del interdictor ($B \in \mathcal{B}$) se trata como un brazo de la configuración clásica. Sin embargo, la constante que acompaña al término del logarítmo (en este caso κ) en políticas de bandit tradicional resulta ser proporcional a la cardinalidad de $\mathcal{B}$, mientras que las políticas diseñadas para configuraciones con muchos arcos en red combinatorial (que calculan costos estimados a nivel de arcos), muestran un mejor desempeño, alcanzando una constante κ proporcional (polinomial) a la cardinalidad de A.

La dependencia en los tamaños de $\mathcal{B}$ y A en cotas superiores de desempeño en tiempo finito para el caso de, por ejemplo, políticas UCB (descrito más adelante), está dada por la necesidad de estimar recompensas medias asociadas a los elementos de $\mathcal{B}$ o costos medios para los arcos en A, respectivamente. Analizando esto en detalle, el resultado del Teorema 4.1 y el Corolario 4.1 sugieren que, en el setting de interdicción de camino más corto, el Regret está dado por la necesidad de estimar costos medios de elementos en cada conjunto $E \in \mathcal{E}$. No obstante, esa estimación puede ser realizada implementando a lo más $\left(|\mathcal{E}| \wedge |\bigcup_{E\in\mathcal{E}} E|\right)$ soluciones en $\mathcal{B}$.

La cota en el Teorema 4.1 no considera explícitamente la dependencia en el conjunto de soluciones implementadas, ya que intenta alcanzar el menor Regret posible.

El siguiente ejemplo ilustra la diferencia entre la cota inferior señalada anteriormente y la manera en la cual se expresan cotas superiores de desempeño en tiempo finito.

Ejemplo 4.2 Considere la configuración de interdicción de camino más corto mostrado en la Figura 4.2. Sea un budget de interdicción $K < k-1$, $l_a = 0$ y $u_a = \infty$ para todo $a \in A$. Esta configuración es similar al Multi-Armed Bandit tradicional, en el sentido de que existen $k-1$ caminos independientes. Notar que, implementando un subconjunto de todas las posibles acciones de interdicción, el líder es capaz de identificar los $K+1$ caminos con el menor costo y hacer un ranking de dichas acciones. Tal información es suficiente para resolver el problema de información completa. En este setup, para cualquier vector de costo esperado, es siempre óptimo bloquear los K caminos con el menor costo promedio. Sin pérdida de generalidad, se puede suponer que

$$\mu_{1,j} + \mu_{j,k+1} \leq \mu_{1,j+1} + \mu_{j+1,k+1}, \quad j < k,$$

y considerar una política que inicialmente implementa $B^1 = \emptyset$, y $B^t = B^{t-1} \cup \{(1,t)\}$ para $t = 2, \ldots, K+1$. Para el periodo $T = K+2$ el interdictor ya es capaz de concluir que $B^*(\mu) = B^{K+1}$ y, por ende, la política alcanza un Regret finito (independiente de T). En términos de la cota inferior, se puede notar que $\mathcal{U}$ es la clase de todos los vectores de costos esperados para los cuales los $K+1$ caminos más cortos coinciden con los caminos más cortos bajo el vector de costos medios μ. Esto implica que $\mathcal{E} = \emptyset$ y por ende

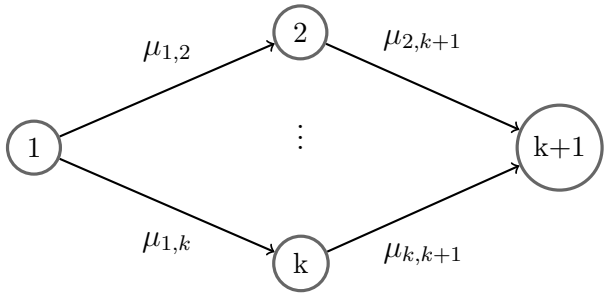


Figura 4.2: Configuración de interdicción del Ejemplo 4.2. Cada arco $a \in A$ está etiquetado como μ_a, su costo esperado bajo F

$$\liminf_{t\to\infty} \frac{\mathcal{R}(t,F)}{\ln t} = 0.$$

Lo anterior se puede comparar a las cotas superiores de desempeño en tiempo finito para políticas de referencia tradicionales (vistas más adelante) que aseguran que, a lo más, $\liminf_{t\to\infty} \frac{\mathcal{R}(t,F)}{\ln t} \leq o(|A|) \sim o(k)$ en la configuración actual.

Capítulo 5

Adaptación de políticas tradicionales

En este capítulo, se adaptan políticas del Multi-Armed Bandit clásico a la configuración de interdicción secuencial de camino más corto que se ha expuesto en esta tesis. Se presentan adaptaciones a las políticas Upper Confidence Bound (UCB) y Thompson Sampling (TS). Políticas como TS y sus derivadas están inmersas en una configuración Bayesiana y por ende, requiere de samplear a partir de una distribución a posteriori. Con esto en mente, se considera una distribución a priori $\lambda(\cdot)$ en μ tal que $\lambda(\mathcal{U}) = 1$, y se denota la distribución a posteriori condicional en el feedback observado hasta $t \in [T]$ (incluído) como $\lambda^t \equiv \lambda_{|\mathcal{F}^t}$.

Ambos tipos de políticas adaptadas comienzan con una fase de inicialización, que se describe a continuación. Para esto, se presenta la siguiente clase $\mathcal{X}$ que tiene una relación uno a uno entre sus elementos y las acciones de interdicción,

$$\mathcal{X} := \left\{ \{x_a : a \in A\} : x_a \in \{0,1\}, \sum_{a \in A} x_a \leq K, \, a \in A \right\}.$$

Con esto, en este capítulo, se hará referencia a $x \in \mathcal{X}$ y $B \in \mathcal{B}$ como acciones de interdicción de igual forma.

5.1. Fase de inicialización

El objetivo de esta fase es formar el set $\mathcal{U}$ y descartar aquellos arcos que están fuera de $\bigcup_{B \in \mathcal{B}} S(B)$. Se define $\mathcal{B}^t$ como el conjunto de acciones de interdicción $B \in \mathcal{B}$ para los cuales, en periodo t, $S(B)$ no es conocido ($\mathcal{B}^1 \equiv \mathcal{B}$). Se tiene que

$$\mathcal{B}^t := \left\{ B \subseteq \mathcal{B} : \nexists\, s < t \text{ t.q. } \mathcal{P}(B) \subseteq \bigcup_{h<t : S^h \in \mathcal{P}(B^s)} \mathcal{P}(B^h) \right\}, \quad t > 1. \tag{5.1}$$

La ecuación 5.1 defie $\mathcal{B}$ como todas las soluciones de interdicción $\mathcal{B}$ que hacen que el recorrido que escoja el evasor no contenga arcos que han sido explorados con otra solución de interdicción en algún periodo anterior, es decir, son todas las soluciones de interdicción que permiten explorar nuevos arcos.

En la fase de inicialización, una política itera escogiendo $B \in \mathcal{B}$ (arbitrariamente) hasta que $\mathcal{B}^t = \emptyset$. Cuando $\mathcal{B}^t = \emptyset$ significa que se exploran todas las soluciones de interdicción y aquellos arcos que no fueron vistos (el evasor nunca recorrió) se eliminan del grafo.

Este proceso se resume en el Algoritmo 1, donde se hace un abuso de notación y se denota como $S(B, \nu)$ a la respuesta del evasor ante la interdicción $B \in \mathcal{B}$ cuando el vector de costos medios es $\nu \in \mathcal{U}$.

Algorithm 1 Fase de Inicialización

Fijar $t = 1$ y $\mathcal{B}^t = \mathcal{B}$.
while $\mathcal{B}^t \neq \emptyset$ **do**
  Escoger $B^t \in \mathcal{B}^t$ arbitrariamente, y observar S^t.
  Fijar $t = t + 1$ y actualizar $\mathcal{B}^t$ acorde a (5.1)
end while
Fijar $A \equiv \bigcup_{B \in \mathcal{B}} S(B)$.
Fijar $\mathcal{U} \equiv \{\nu := \langle \nu_a,\, a \in A | : S(B, \mu) = S(B, \nu),\, B \in \mathcal{B}\}$.

Sea t_0 el primer periodo después de la fase de inicialización. El Algoritmo 1 requiere encontrar $B^t \in \mathcal{B}^t$ para $t < t_0$. Esta tarea puede ser realizada mediante la resolución del siguiente problema,

$$
\begin{aligned}
& & \max \quad & (\nu^2)^\top \bar{\boldsymbol{z}}^1 & - \bar{y}_n^2 + \bar{y}_1^2 & & (5.2a)\\
\text{s.t.} \quad & y_j^{k,s} - y_i^{k,s} & \leq \quad & \nu_{i,j}^k, \quad (i,j) \in A \setminus B^s, & s < t,\, k = 1,2 & (5.2b)\\
& y_n^{k,s} - y_1^{k,s} & = \quad & (\nu^k)^\top \boldsymbol{z}^s, & s < t,\, k = 1,2 & (5.2c)\\
& \mathcal{A}\, \bar{\boldsymbol{z}}^k & = \quad & b, & k = 1,2 & (5.2d)\\
& \bar{z}_{i,j}^k + x_{i,j} & \leq \quad & 1, & (i,j) \in A,\, k = 1,2 & (5.2e)\\
& \bar{y}_j^k - \bar{y}_i^k & \leq \quad & \nu_{i,j}^k + M\, x_{i,j}\,, & k = 1,2 & (5.2f)\\
& \bar{y}_n^k - \bar{y}_1^k & = \quad & (\nu^k)^\top \bar{\boldsymbol{z}}^k, & k = 1,2 & (5.2g)\\
|A| \cdot \Big(\sum_{(i,j)\in B^s} \bar{z}_{i,j}^k + \sum_{(i,j)\in z^s} x_{i,j} \Big) & & \geq \quad & \sum_{(i,j)\notin z^s} \bar{z}_{i,j}^k & s < t,\, k = 1,2 & (5.2h)\\
& \boldsymbol{y}^{k,s}, & \nu^k, \quad & \bar{\boldsymbol{y}}^k,\, \bar{\boldsymbol{z}}^k \in \mathbb{R}_+,\ x \in \mathcal{X}. & & (5.2i)
\end{aligned}
$$

En la formulación anterior, (5.2b) y (5.2c) fuerzan, mediante dualidad fuerte, que ambos vectores ν^1 y ν^2 sean consistentes con el feedback observado antes del periodo t, donde $\boldsymbol{z}^t$ es una versión vectorizada de $\mathcal{S}^t$ y, por lo tanto, conocido. La función objetivo y el resto de restricciones tienen como propósito encontrar una solución de interdicción $x \in \mathcal{X}$ para la cual el evasor responda diferente bajo los costos ν^1 y ν^2. Las respuestas del evasor, $\bar{\boldsymbol{z}}^1$ y $\bar{\boldsymbol{z}}^2$ respectivamente, se encuentran aplicando dualidad fuerte en las restricciones (5.2d) y (5.2g). En la restricción (5.2d), se impone factibilidad primal de $\bar{\boldsymbol{z}}^k$, donde $\mathcal{A}$ es la matriz de adyacencia y $\boldsymbol{b}$ tiene la forma $\boldsymbol{b} = (-1, 0, \ldots, 0, 1)$. La restricción (5.2h) impone que no existan empates entre respuestas, es decir, si no se bloquea un arco de la respuesta en s ($\sum_{(i,j)\in z^s} x_{i,j} = 0$)y no pasa por un arco bloqueado en s ($\sum_{(i,j)\in B^s} \bar{z}_{i,j}^k = 0$), la respuesta que se busca debe ser igual a la respuesta del evasor en s ($\sum_{(i,j)\notin z^s} \bar{z}_{i,j}^k = 0$), es decir, no puede escoger otro camino distinto porque sería un empate.

Notar que los términos $(\nu^k)^\top \bar{z}^l$ son la suma de productos de variables continuas y binarias, por lo que deben ser linealizadas.

5.2. Política de sampleo a posteriori

Se propone una adaptación de la política Thompson Sampling (TS). Considerar una aproximación Bayesiana en la cual $\lambda^t := \lambda_{|\mathcal{F}^t}$ denota la distribución a posterioi de μ en tiempo

$t \in [T]$. Como se explicó anteriormente, la política comienza con la fase de inicialización descrita en el Algoritmo 1, que define $\mathcal{U}$ y redefine A dejando solo los arcos relevantes del problema. Luego de la inicialización, para $t \geq t_0$, TS samplea un vector de costos $\nu^t \sim \lambda^t$ e implementa $B^t := B^*(\nu^t)$. El procedimiento se describe en el Algoritmo 2.

Algorithm 2 Fase TS

Fijar $t = t_0$ y $\lambda^t = \lambda$.
while $t \leq T$ **do**
  Samplear ν^t desde λ^t.
  Escoger $B^t \in argmax_{B \in \mathcal{B}} \{r(S(B), \nu^t)\}$ y observar S^t.
  Fijar $t = t + 1$ y actualizar $\lambda^t := \lambda_{|\mathcal{F}^t}$.
end while

Cabe destacar que TS requiere de una distribución a posteriori λ^t, por lo que a continuación, se propone un algoritmo para aproximar dicha distribución.

5.2.1. Aproximación de la distribución a posteriori

En la política descrita anteriormente, se considera un enfoque Bayesiano en el cual se asume que F pertenece a alguna familia paramétrica tal que $F(\cdot) \equiv F(\cdot|\mu)$ y que dicha familia está parametrizada por su promedio $\mu \in \mathbb{R}^{|A|}$. Además, se considera que μ es desconocido en principio y que $\mu \sim \bar{\lambda}(\cdot)$ para alguna distribución $\bar{\lambda} \in \mathbf{B}(\mathbb{R}^{|A|})/\mathbb{R}$. Luego de la etapa de inicialización, si algún arco no es obsevado en alguna respuesta de $S(B)$ para algún $B \in \mathcal{B}$ se elimina de A, por lo que se asume que $\bar{\lambda}(\cdot)$ es definida sobre el grafo resultante. Por otro lado, en tiempo $t = t_0$ es conocido que $\mu \in \mathcal{U}$ c.s, por lo que la distribución a priori considerada en TS corresponde a la restricción de $\bar{\lambda}$ sobre $\mathcal{U}$, i.e.

$$\lambda(U) := \frac{\bar{\lambda}(U \cap \mathcal{U})}{\bar{\lambda}(\mathcal{U})}, \quad U \in \mathbf{B}(\mathbb{R}^{|A|}).$$

5.2.2. Samplear desde λ^t

Sea $\bar{\lambda}^t(\cdot)$ la distribución a posteriori de μ dada $\mathcal{F}^t$ relativa a la distribución a priori $\bar{\lambda}^t(\cdot)$, i.e. $\bar{\lambda}^t := \bar{\lambda}(\cdot)|_{\mathcal{F}^t}$. Según lo anterior, se puede corroborar que $\lambda^t(U) = \frac{\bar{\lambda}^t(U \cap \mathcal{U})}{\bar{\lambda}^t(\mathcal{U})}$ para $U \in \mathbf{B}(\mathbb{R}^{|A|})$. Por ende, se tiene que

$$\frac{d\lambda^t(\omega)}{d\bar{\lambda}^t} = \frac{\mathbf{1}\{\omega \in \mathcal{U}\}}{\bar{\lambda}^t(\mathcal{U})}, \quad \omega \in \Omega,$$

Dado lo anterior, es posible samplear de λ^t usando *acceptance-rejection sampling* (sampleo aceptación-rechazo). En un esquema de este estilo, la probabilidad de rechazo es 0 si el sampleo de $\bar{\lambda}^t$ está en $\mathcal{U}$ y 1 si no. Este proceso está detallado en el Algoritmo 3.

Algorithm 3 AR sampling

Para $t \in \mathbb{N}$, calcular $\bar{\lambda}^t$ y samplear ν desde $\bar{\lambda}^t$.
while $\nu \notin \mathcal{U}$ **do**
  Samplear ν desde $\bar{\lambda}^t$.
end while
Fijar $\nu^t = \nu$.

Es importante notar que el Algoritmo 3 no necesita calcular $\bar{\lambda}^t(\mathcal{U})$, que es la principal dificultad de samplear directamente de λ^t. En lugar de eso, se samplea de $\bar{\lambda}^t$, que es manejable computacionalmente si, por ejemplo, se escogen distribuciones F y λ conjugadas, donde $\bar{\lambda}^t$ tiene una forma cerrada, como se muestra en los ejemplos numéricos más adelante.

Con este tipo de modelamiento, la mayor dificultad computacional es corroborar si el sample ν está o no en $\mathcal{U}$. Para esto, se plantean los siguientes problemas de programación entera-mixta.

Sea $\bar{\nu}$ algún vector conocido que cumple que $\bar{\nu} \in \mathcal{U}$ y sea ν un vector que ha sido sampleado de $\bar{\lambda}^t$ en el contexto del Algoritmo 3. Para una acción de interdicción fija $x \in \mathcal{X}$ se tiene que, para $B \equiv x, S(B, \nu) = S(B, \bar{\nu})$ si existe $(y^\nu, y^{\bar{\nu}}, z)$ tal que

$$\begin{aligned}
\mathcal{A}\,\boldsymbol{z} &= \boldsymbol{b} && \text{(5.3a)}\\
x_a + z_a &\leq 1, \quad a \in A && \text{(5.3b)}\\
y^\nu_j - y^\nu_i &\leq \nu_{i,j} + M\,x_{i,j}, \quad (i,j) \in A && \text{(5.3c)}\\
y^{\bar{\nu}}_j - y^{\bar{\nu}}_i &\leq \bar{\nu}_{i,j} + M\,x_{i,j}, \quad (i,j) \in A && \text{(5.3d)}\\
y^\nu_n - y^\nu_1 &= \nu^\top \boldsymbol{z} && \text{(5.3e)}\\
y^{\bar{\nu}}_n - y^{\bar{\nu}}_1 &= \bar{\nu}^\top \boldsymbol{z} && \text{(5.3f)}\\
\boldsymbol{z}, \quad \boldsymbol{y}^\nu, \quad \boldsymbol{y}^{\bar{\nu}} &\in \mathbb{R}_+ && \text{(5.3g)}
\end{aligned}$$

Al igual que antes, las restricciones (5.3a) y (5.3b) corresponde a la factibilidad primal de $\boldsymbol{z}$ para el problema de camino más corto, mientras que (5.3c) y (5.2c) corresponden a la factibilidad dual bajo el costo ν y $\bar{\nu}$, respectivamente. Por otro lado, (5.3e)y (5.3f) imponen

dualidad fuerte asegurando que $\boldsymbol{z}$ sea óptimo bajo ν y $\bar{\nu}$. Cabe destacar que ((5.3)) es lineal en $\boldsymbol{y}$ y $\boldsymbol{z}$ (para $\boldsymbol{x}$ dado), pero infactible si $S(B,\nu) \neq S(B,\bar{\nu})$, por ende, se formula el siguiente problema para chequear si $\nu \in \mathcal{U}$, dado un $\bar{\nu} \in \mathcal{U}$ conocido.

$$
\begin{aligned}
w(\nu) := \max \quad & \nu^\top \quad (\boldsymbol{z}^{\bar{\nu}} - \boldsymbol{z}^{\nu}) && \text{(5.4a)}\\
\mathcal{A}\,\boldsymbol{z}^{\nu} \quad &= \quad \boldsymbol{b} && \text{(5.4b)}\\
x_a + z_a^{\nu} \quad &\leq \quad 1, \quad a \in A && \text{(5.4c)}\\
\mathcal{A}\,\boldsymbol{z}^{\bar{\nu}} \quad &= \quad \boldsymbol{b} && \text{(5.4d)}\\
x_a + z_a^{\bar{\nu}} \quad &\leq \quad 1, \quad a \in A && \text{(5.4e)}\\
y_j^{\nu} - y_i^{\nu} \quad &\leq \quad \nu_{i,j} + M\,x_{i,j}, \quad (i,j) \in A && \text{(5.4f)}\\
y_j^{\bar{\nu}} - y_i^{\bar{\nu}} \quad &\leq \quad \bar{\nu}_{i,j} + M\,x_{i,j}, \quad (i,j) \in A && \text{(5.4g)}\\
y_n^{\nu} - y_1^{\nu} \quad &= \quad \nu^\top \boldsymbol{z}^{\nu} && \text{(5.4h)}\\
y_n^{\bar{\nu}} - y_1^{\bar{\nu}} \quad &= \quad \bar{\nu}^\top \boldsymbol{z}^{\bar{\nu}} && \text{(5.4i)}\\
\boldsymbol{z}^{\nu}, \boldsymbol{z}^{\bar{\nu}}, \quad \boldsymbol{y}^{\nu}, \quad & \boldsymbol{y}^{\bar{\nu}} \in \mathbb{R}_+,\, x \in \mathcal{X}. && \text{(5.4j)}
\end{aligned}
$$

El siguiente lema formaliza 5.4

Lema 5.1 *Sea* $\nu \in \mathbb{R}^{|A|}$, $\nu \in \mathcal{U} \iff w(\nu) = 0$.

Se debe notar que 5.4 es lineal en $\boldsymbol{y}, \boldsymbol{z}$ y $\boldsymbol{x}$, por lo que más concretamente es un **MIP**, que puede ser resuelto mediante un solver.

Para encontrar un vector $\bar{\nu} \in \mathcal{U}$ inicial en tiempo $t = t_0$ se puede resolver el siguiente **MIP** por una única vez

$$
\begin{aligned}
\bar{\nu} \in argmax \quad & 0 && \text{(5.5a)}\\
y_j^t - y_i^t \quad &\leq \quad \nu_{i,j}, \quad (i,j) \in A \setminus B^t,\, t < t_0 && \text{(5.5b)}\\
y_n^t - y_1^t \quad &= \quad \nu^\top \boldsymbol{z}^t, \quad t < t_0 && \text{(5.5c)}\\
\boldsymbol{y}^t, \quad \nu \quad &\in \mathbb{R}_+ && \text{(5.5d)}
\end{aligned}
$$

En lo anterior, $\{(\boldsymbol{z}^t \equiv S^t, B^t),\, t < t_0)\}$ representa el set de interdicciones y recorridos to-

mados por el evasor durante la inicialización, donde el feedback generado es suficiente para caracterizar la respuesta del evasor ante todas las acciones de interdicción posibles. Por último, es importante notar que μ es una solución factible a la formulación anterior, probando que está bien definido.

5.3. Desempeño en tiempo finito de políticas UCB

En esta sección se presenta una adaptación de la política UCB de Auer et al. (2002) para este setting, en la cual se establece una garantía de desempeño en tiempo finito que se compara con aquellas que vienen de literatura previa en problemas de Bandit lineal.

5.3.1. Política UCB simple

Al igual que la política TS, UCB simple comienza con la fase de inicialización descrita en el Algoritmo 1, donde al finalizar este proceso se eliminan los arcos que no se utilizan actualizando A y el conjunto $\mathcal{U}$ es calculado. Luego, haciendo abuso de notación, sea $\tau(a,t) \equiv \tau(\{a\},t)$ el número de veces que un arco $a \in A$ ha sido incluído en la respuesta del evasor previo al periodo t (incluído), esto es,

$$\tau(a,t) = \sum_{s \leq t} \mathbf{1}\{a \in S^s\}, \quad a \in A,\, t \geq t_0.$$

Dada la fase de inicialización, se tiene que $\tau(a,t_0) > 0$ para todo $a \in A$, ya que los arcos que permanecen en A son aquellos que fueron recorridos al menos una vez.

Para $t \geq t_0$, UCB implementa $B^t \in argmax\{r(S(B),\nu^t)\colon B \in \mathcal{B}\}$, donde $\nu^t := \{\nu_a^t \colon a \in A\}$, con

$$\nu_a^t := \frac{\sum_{s<t} \mathbf{1}\{a \in S^s\}\, c_a^s}{\tau(a,t-1)} + \sqrt{\frac{(\mathcal{L}+1)\ln t}{\tau(a,t-1)}} \quad a \in A,\, t \geq t_0, \tag{5.6}$$

y $\mathcal{L} := \max\{|P|\colon P \in \mathcal{P}\}$. El Algoritmo 4 describe paso a paso de la política.

Adaptando argumentos de Auer et al. (2002) para el Multi-armed Bandit tradicional se puede mostrar la siguiente cota de desempeño en tiempo finito.

Algorithm 4 Fase UCB Simple

Fijar $t = t_0$ y $\lambda^t = \lambda$.
while $t \leq T$ **do**
  Escoger $B^t \in argmax\,\{r(S(B), \nu^t) : B \in \mathcal{B}\}$ y observar S^t.
  Fijar $t = t + 1$ y actualizar ν^t acorde a (5.6).
end while

Teorema 5.1 *El regret asociado a la política UCB simple es tal que*

$$\mathbb{E}_F\left[\mathcal{R}^{nUCB}(t, F)\right] \leq |A \setminus S^*|\Delta_{\max}\Big(\Big\lceil \frac{4\mathcal{L}^2(\mathcal{L}+1)\ln t}{\Delta_{\min}^2}\Big\rceil + \frac{\pi^2|\mathcal{B}|}{6}\Big),$$

donde $\Delta_{\max} := \max\{\Delta_B \colon B \in \mathcal{B}\}$ *y* $\Delta_{\min} := \min\{\Delta_B \colon B \in \mathcal{B} \setminus \{B^*\}\}$.

Observación Es importante destacar que, a diferencia de TS, UCB simple no verifica si $\nu^t \in \mathcal{U}$, si no que asume directamente que el evasor responde a B con $S(B)$ con $B \in \mathcal{B}$, por ende, implementando el Algoritmo 4 como está descrito, requiere de un seguimiento del conjunto $\{(B, S(B)), B \in \mathcal{B}\}$, que es de cardinalidad exponencial (en A) en un caso pesimista. Es por esto que en los experimentos numéricos de gran tamaño se selecciona $B^t \in argmax\,\{r(S(B, \nu^t), \nu^t) : B \in \mathcal{B}\}$ donde, abusando de la notación, se denota la respuesta del evasor bajo la interdicción $B \in \mathcal{B}$ y el vector ν como $S(B, \nu)$. Usando técnicas de modelamiento tradicionales, se calcula B^t resolviendo el siguiente MIP

$$B^t \in argmax\ y_n - y_1 \tag{5.7a}$$
$$\text{s.t. } y_j - y_i \ \leq \ \nu_{i,j}^t + M\,x_{i,j}, \quad (i,j) \in A \tag{5.7b}$$
$$\boldsymbol{y}, \qquad x \in \mathcal{X} \tag{5.7c}$$

5.3.2. UCB recorrido completo

UCB simple calcula por separado una cota superior de confianza para cada arco $a \in A$ en lugar de seleccionarlos en conjunto. Este mecanismo simplifica el análisis del desempeño en tiempo finito de la política, pero dificulta su desempeño práctico. Dado esto, se puede modificar el Algoritmo 4 con el fin de calcular una cota superior de confianza para el costo total del recorrido del evasor.

Para $t \geq t_0$ y $B \in \mathcal{B}$ se define

$$\hat{r}^t(B) := \sum_{a \in S(B)} \frac{1}{\tau(a, t-1)} \sum_{s<t} c_a^s \mathbf{1}\{a \in S^s\} + (\mathcal{L}+1)\sqrt{\ln t \sum_{a \in S(B)} \frac{1}{\tau(a, t-1)}},$$

Luego, se puede modificar UCB simple seleccionando $B^t \in argmax\{r^t(B) : B \in \mathcal{B}\}$ en lugar de $B^t \in argmax\{r(S(B), \nu^t) : B \in \mathcal{B}\}$, respaldando esto en los argumentos de la demostración del Teorema 5.1

Teorema 5.2 *Sea π la política UCB de recorrido completo. El regret asociado a dicha política π es tal que*

$$\mathbb{E}_F[\mathcal{R}^\pi(t, F)] \leq |\tilde{A}|\Delta_{\max}\left(\left\lceil \frac{4\mathcal{L}(\mathcal{L}+1)\ln t}{\Delta_{\min}^2} \right\rceil + \frac{\pi^2|\mathcal{B}|}{6}\right).$$

Si bien esta política significa una mejora más bien modesta en la garantía teórica del resultado del Teorema 5.1, resolver B^t no es posible a través de un MIP, ya que el segundo término de la definición de $r^t(\cdot)$ deja de ser lineal en los elementos de $S(B)$. Dado lo anterior, se utiliza un solver más complejo que resuelve problemas MISOCP (Mixed integer second order cone programming).

5.3.3. UCB modificado

Como se mencionó anteriormente, UCB simple no chequea si $\nu^t \in \mathcal{U}$, si no que evalúa $r(S(B), \nu^t)$ para todo $B \in \mathcal{B}$, lo cual puede ser computacionalmente complicado dado el tamaño del set $\mathcal{B}$. En esta sección se modifica UCB simple de forma que, mientras sostiene la premisa de optimismo frente a la incertidumbre, chequea si ν^t está en $\mathcal{U}$, evitado enumerar explícitamente.

Para algún $t > t_0$ dado y $a \in A$ se define

$$\hat{\mu}_a^t := \frac{\sum_{s<t} \mathbf{1}\{a \in S^s\}\ c_a^s}{\tau(a, t-1)}, \qquad L_a^t := \sqrt{\frac{(\mathcal{L}+1)\ln t}{\tau(a, t-1)}}.$$

Luego, se puede interpretar que UCB simple selecciona de forma **optimista** un vector de costos medios $\nu^t \in [\bar{\mu}^t - L^t,\ \bar{\mu}^t + L^t]$, considerando que $|\mu_a^t - \bar{\mu}_a^t| \leq L_a^t$ tiene una probabilidad alta. Sin embargo, como UCB simple no verifica que ν^t está en $\mathcal{U}$, escoge $\nu^t = \hat{\mu}^t + L^t$.

Considerando lo anterior, se propone la siguiente versión que restringe su elección de vector

de costos a uno que esté en $\mathcal{U}$. Es decir

$$B^t \in argmax \left\{ \max \left\{ r(S(B), \nu) \colon \left| \nu - \bar{\nu}^t \right| \leq L^t,\, \nu \in \mathcal{U} \right\},\, B \in \mathcal{B} \right\}. \tag{5.8}$$

Una propiedad importante de UCB modificado es que en tiempo $t > t_0$, B^t puede ser calculado mediante el siguiente MIP, que es una modificación a la formulación 5.5.

$$\begin{aligned}
\max\ & \bar{y}_n - \bar{y}_1 && && && \text{(5.9a)}\\
\text{s.t.}\ & y^t_j - y^t_i && \leq && \nu_{i,j}, && (i,j) \in A \setminus B^t,\, t < t_0 && \text{(5.9b)}\\
& y^t_n - y^t_1 && = && \nu^\top \boldsymbol{z}^t, && t < t_0 && \text{(5.9c)}\\
& \bar{y}_j - \bar{y}_i && \leq && \nu_{i,j} + M\, x_{i,j}, && (i,j) \in A && \text{(5.9d)}\\
& \nu - \bar{\nu}^t && \leq && L^t && t < t_0 && \text{(5.9e)}\\
& \boldsymbol{y}^t, \bar{\boldsymbol{y}}, && \nu && \in \mathbb{R}_+,\, x \in \mathcal{X} && && \text{(5.9f)}
\end{aligned}$$

Al igual que en la formulación (5.5), (5.9b) y (5.9c) imponen que la elección de ν esté en $\mathcal{U}$. Las restricción (5.9d) y la función objetivo corresponden al problema dual del problema de camino más corto, que transforma la formulación del problema binivel de interdicción en una formulación de un solo nivel, incluso cuando ν se incluye como una variable de decisión. Por último, la restricción (5.9e) impone que el vector ν a decidir y el vector $\bar{\nu}$ que corresponde a los valores observados en el periodo de inicialización difieran a lo más en L como indica (5.8). Cabe notar que el número de restricciones y de variables depende de t_0, esto es, el número de periodos en la fase de inicialización.

Si bien este problema tiene un número exponencial de variables y restricciones en un caso pesimista, este tipo de formulación puede ser resuelto utilizando algoritmos de descomposición (por ejemplo: Branch and cut, branch and price), que funcionan rápido en la práctica y en escasas ocasiones necesita añadir variables y restricciones extra al problema.

Capítulo 6

Resultados numéricos

En este capítulo se presentan resultados de la implementación de las políticas señaladas anteriormente, es decir TS, UCB simple, UCB recorrido y UCB modificado. Dicha implementación se hace sobre cien grafos en capas (*layered graphs*) aleatorios, con costos sampleados de una distribución $Beta(1, 1)$ y considerando mil periodos de ejecución de exploración.

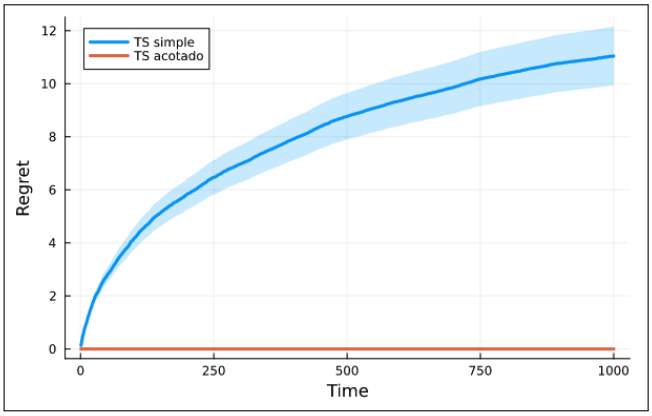
6.1. Thomson Sampling

Como se señaló anteriormente, Thompson Sampling samplea costos aleatorios de una distribución a posteriori y resuelve el problema de interdicción utilizando dichos costos. Posteriormente, ajusta los parámetros de la distribución. Esta política se aplica en las simulaciones utilizando una distribución Beta-Binomial, esto es, los costos provienen de una distribución Binomial con parámetro p que distribuye Beta de parámetros α y β (inicialmente $\alpha = 1$ y $\beta = 1$).

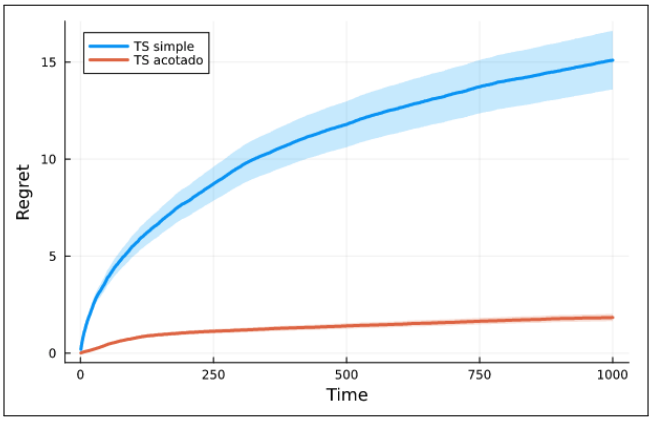
Se generan dos experimentos con TS. El primero es TS Simple, que no se limita a vectores de costos en el conjunto $\mathcal{U}$, es decir, considera cualquier vector de costos sampleado de la distribución (Algoritmo 2). El segundo experimento corresponde a TS Acotado que sí verifica que el vector de costos sampleado esté en $\mathcal{U}$ (Algoritmo 3).

En las Figuras 6.1.a y 6.1.b se muestra el Regret acumulado para ambas políticas, considerando grafos de una capa y dos capas, respectivamente, con un bloqueo disponible. Se puede notar que, en ambas figuras, la política TS Acotada alcanza un Regret logarítmico más rápido que TS Simple y mucho menor (cercano a cero) y con un error menor (área sombreada es cero). Esto indicaría que el desempeño de TS Acotado es en principio mejor que TS Simple

para estas configuraciones.



(a) Una capa y un bloqueo



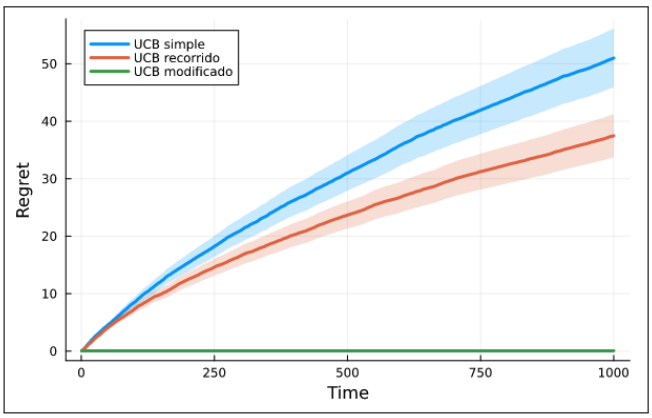
(b) Dos capas y un bloqueo

Figura 6.1: Thompson Sampling

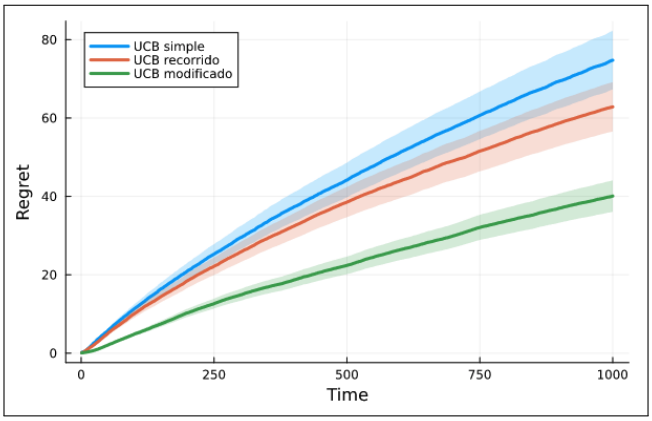
6.2. UCB

Para las políticas UCB Simple, UCB Recorrido y UCB Modificado, se implementa el algoritmo 4 con las variaciones descritas en la sección 5.3.

A continuación se presentan los resultados en la Figura 6.2, donde se utilizan la mismas configuraciones anteriores, es decir, gráfos de una y dos capas con un bloqueo disponible. Es posible apreciar que el Regret en UCB modificado en la Figura 6.2.a alcanza el comportamiento logarítmico característico rápidamente al igual que TS Acotado y con una magnitud mínima de Regret, superando a UCB Simple y UCB Recorrido reflejando, en promedio, un mejor desempeño en este tipo de grafos. Por otro lado, en la Figura 6.2.b UCB modificado, aunque sigue superando al resto de UCB's, aumenta su magnitud en Regret.



(a) Una capa y un bloqueo

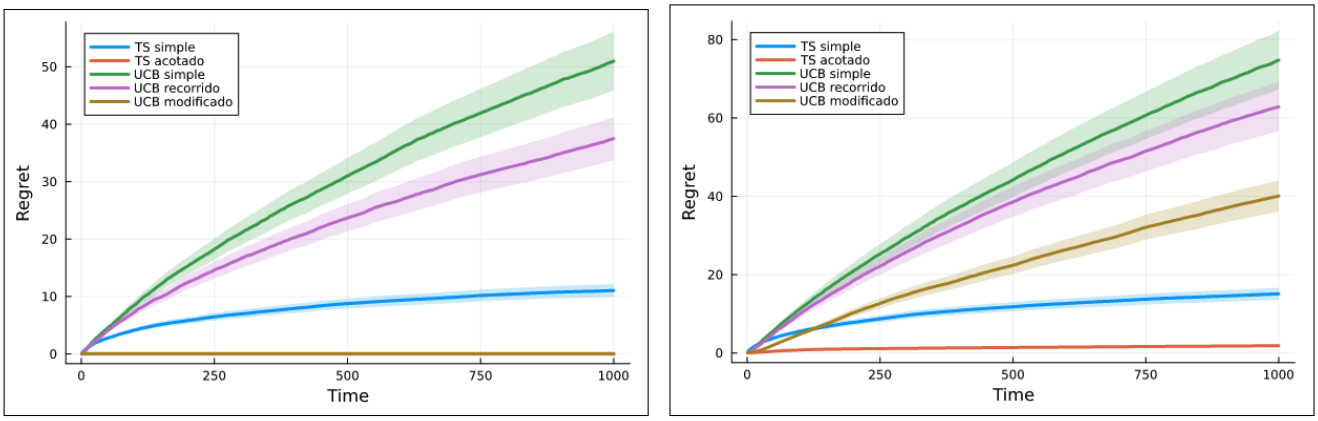


(b) Dos capas y un bloqueo

Figura 6.2: UCB

6.3. Benchmark general

La Figura 6.3 muestra el desempeño de todas las políticas para cien grafos de una y dos capas, con un bloqueo disponible. En dicha figura se puede apreciar que TS Acotado y UCB modificado presenta una diferencia significativa en desempeño con respecto al resto de las políticas.



(a) Una capa y un bloqueo

(b) Dos capas y un bloqueo

Figura 6.3: Políticas UCB y TS en grafos de una capas y un bloqueo

Por otro lado, en la Figura 6.3.b, que compara políticas en grafos de dos capas, se puede notar que UCB modificado se aleja del desempeño de TS Acotado, por lo que se podría inferir que en grafos más grandes UCB modificado empeora su desempeño y que TS Acotado se mantiene como la política con mejor desempeño. Además, es importante notar que para estos grafos más grandes el Regret, para todas las políticas, crece en magnitud.

Sumado a lo anterior, cabe notar que si bien TS Simple es una política que relaja la restricción de que los costos pertenezcan a $\mathcal{U}$, su desempeño supera las políticas UCB desde el corto plazo.

Desde otro ángulo, aunque TS Acotado supera a la mayoría de las políticas en desempeño, la variación en tiempo de implementación es mucho mayor que el resto. En particular, en el caso de grafos de una capa, es mayor que el tiempo de implementación de UCB modificado, que entrega resultados similares. El detalle se describe en la Tabla 6.1 y se debe tomar en cuenta que las implementaciones se ejecutaron paralelizando.

Tabla 6.1: Tiempos de implementación para un bloqueo disponible

(seg)	TS Simple	TS Acotado	UCB Simple	UCB Rec	UCB Mod
1 capa 1 bloqueo	363.99	4060.67	1576.55	1839.55	497.51
2 capas 1 bloqueo	267.93	12184.72	1056.88	1509.24	609.12

Capítulo 7

Conclusión

En vista de lo expuesto, mediante los resultados numéricos se observa un comportamiento logarítmico del Regret acumulado, lo que indica, al igual que en la demostración teórica de la cota inferior de desempeño, que el Regret crece conforme al logaritmo de t. En conclusión, para asegurar optimalidad de una solución de interdicción es preciso explorar al menos $\ln(t)$ veces.

En el estudio de las distintas políticas de interdicción se puede concluir que la que tiene un mejor desempeño es TS Acotado al conjunto $\mathcal{U}$, sin embargo, el costo de tiempo de implementación frente a las otras políticas es mucho mayor. Esto se puede mejorar en simulaciones futuras optimizando los recursos pero sin duda es importante mencionarlo.

En el futuro se desea explorar una política de interdicción adicional que corresponde a UCB pero con un enfoque bayesiano y comparar con el comportamiento de las políticas presentadas en esta tesis. En particular, se espera que su desempeño sea mejor al resto de políticas UCB y compita con TS Acotado.

Por último, si bien esta tesis se basa en el estudio teórico del problema de interdicción de camino más corto, es importante destacar que se este tipo de problemas se puede aplicar a situaciones reales como, por ejemplo, bloqueo de caminos de contrabando con recursos policiales limitados, lo cual puede llevar a hacer un mejor uso de dichos recursos.

Bibliografía

Anantharam, V., Varaiya, P., y Walrand, J. (1987). Asymptotically efficient allocation rules for the multi-armed bandit problem with multiple plays-part i: Iid rewards. *Automatic Control, IEEE Transactions on*, *32*(11), 968–976.

Auer, P., Cesa-Bianchi, N., y Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*(2-3), 235–256.

Bayrak, H., y Bailey, M. (2008). Shortest path network interdiction with asymmetric information. *Networks*, *52*(3), 133–140.

Berry, D. A., y Fristedt, B. (1985). *Bandit problems*. Chapman and Hall.

Borrero, J. S., Prokopyev, O. A., y Sauré, D. (2021). Learning in sequential bilevel linear programming. *INFORMS Journal on Optimization*, *To appear*.

Brown, G., Carlyle, M., Salmerón, J., y Wood, K. (2006). Defending critical infrastructure. *Interfaces*, *36*(6), 530–544.

Cesa-Bianchi, N., y Lugosi, G. (2012). Combinatorial bandits. *Journal of Computer and System Sciences*, *78*(5), 1404–1422.

Chen, W., Wang, Y., y Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework, results and applications. , 151–159.

Chern, M.-S., y Lin, K.-H. (1995). Interdicting the activities of a linear program: A parametric analysis. *European Journal of Operational Research*, *86*(3), 580–591.

Colson, B., Marcotte, P., y Savard, G. (2005). Bilevel programming: A survey. *4OR*, *3*(2), 87–107.

Côté, J.-P., Marcotte, P., y Savard, G. (2003). A bilevel modelling approach to pricing and fare optimisation in the airline industry. *Journal of Revenue and Pricing Management*, *2*(1), 23–36.

Gai, Y., Krishnamachari, B., y Jain, R. (2012). Combinatorial network optimization with

unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking (TON)*, *20*(5), 1466–1478.

Gittins, J. (1979). Bandit processes and dynamic allocation rules. *Journal of the Royal Statistical Society*, *41*, 148–177.

Israeli, E., y Wood, K. R. (2002). Shortest-path network interdiction. *NERWORKS*, *40(2),97–111 2002*.

Kulkarni, S. R., y Lugosi, G. (1997). Minimax lower bounds for the two-armed bandit problem. En *Decision and control, 1997., proceedings of the 36th ieee conference on* (Vol. 3, pp. 2293–2297).

Lai, T. L., y Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, *6*(1), 4–22.

Lim, C., y Smith, J. C. (2007). Algorithms for discrete and continuous multicommodity flow network interdiction problems. *IIE Transactions*, *39*(1), 15–26.

Liu, K., Vakili, S., y Zhao, Q. (2012). *Stochastic online learning for network optimization under random unknown weights.* (Working paper)

Lucotte, M., y Nguyen, S. (2013). *Equilibrium and advanced transportation modelling*. Springer Science Business Media.

Modaresi, S., Sauré, D., y Vielma, J. P. (2020). Learning in combinatorial optimization: What and how to explore. *Operations Research*, *68*(5), 1585–1604.

Russo, D., y Van Roy, B. (2014). Learning to optimize via posterior sampling. *INFORMS Journal on MATHEMATICS OF OPERATIONS RESEARCH*, *To appear*.

Sherali, H. D., Soyster, A. L., y Murphy, F. H. (1983). Stackelberg-nash-cournot equilibria: Characterizations and computations. *Operations Research*, *31*(2), 253–276.

Smith, J. C., y Song, Y. (2020). A survey of network interdiction models and algorithms. *European Journal of Operational Research*, *283*(3), 797–811.

Steinrauf, R. (1991). *Network interdiction models* (Tesis Doctoral no publicada). Naval Postgraduate School.

Whittle, P. (1982). *Optimization over time: Vol i*. John Wiley and Sons Ltd.

Wood, R. K. (1993). Deterministic network interdiction. *Mathematical and Computer Modelling*, *17*(2), 1–18.

Anexo

Anexo A. Demostraciones

Demostración. **Lema 4.1.** Por contradicción. Supongamos que existe E tal que $\forall \nu \in \mathcal{U}$, $\mathcal{B}^*(\nu) = B^*(\mu)$, $S(\mathcal{B}^*(\nu)) \cap E = 0$,$S(\mathcal{B}^*(\nu)) \in \mathcal{P}(B^*(\mu))$ y $S(\mathcal{B}^*(\mu)) \in \mathcal{P}(B^*(\nu))$. Como $\mathcal{B}^*(\nu) = B^*(\mu)$, $\mathcal{P}(B^*(\nu)) = \mathcal{P}(B^*(\mu))$. Luego, $S(B^*(\nu)) \in \mathcal{P}(B^*(\mu)) = \mathcal{P}(B^*(\nu))$, lo que es una contradicción pues el interdictor no puede pasar por donde se bloquea, es decir, $S(B^*(\nu)) \neq \mathcal{P}(B^*(\nu))$. ■

Demostración. **Teorema 4.1.** Se asume que la distribución de costos está parametrizada en el vector de costo medio, es absolutamente continua y que sus componentes son independientes. Por lo tanto, se adopta la notación $f_a(\cdot|u_a)$ para representar la densidad del costo asociado con el arco a, donde u_a representa su valor medio, para $a \in A$. Sea $L_a(u_a|\nu_a)$ la divergencia Kullback-Leibler entre las distribuciones de vector de costos para el arco a cuando los costos medios son dados por u_a y ν_a, respectivamente, es decir,

$$L_a(u_a|\nu_a) := \int_{\mathbb{R}} \ln(f_a(x|u_a)/f_a(x|\nu_a)) f_a(x|u_a)\, dx.$$

Considere $E \in \mathcal{E}$: por construcción, existe $\nu \in \mathcal{U}$ con $\nu_a = \mu_a$ para todo $a \in A \setminus E$, tal que $max\{r(S(B'), \nu) : B' \in \mathcal{B}\} > r(S^*, \nu)$. Defina $B := argmax\{r(S(B'), \nu) : B' \in \mathcal{B}\}$, lo cual se asume que es único (para evitar una notación excesiva). Observe que $E \cap S^* = \emptyset$, de manera que

$$r(S(B), \mu) \leq r(S^*, \mu) = r(S^*, \nu) < r(S(B), \nu).$$

Se concluye que $S(B) \cap E \neq \emptyset$.

Para cualquier política consistente π tenemos que para cualquier $\alpha > 0$

$$\begin{aligned}
\mathcal{R}^{\pi}(t, F(\cdot|\nu)) &\geq \Delta\, \mathbb{E}_{\nu}\left[t - \sum_{s\leq t} \mathbf{1}\left\{B^s = B\right\}\right] \\
&\geq \Delta\, \mathbb{E}_{\nu}\left[t - \sum_{s\leq t} \mathbf{1}\left\{S^s = S(B)\right\}\right] \\
&\geq \Delta(t-K)\ln t \mathbb{P}_{\nu}\left\{\sum_{s\leq t} \mathbf{1}\left\{S^s = S(B)\right\} < K\, \ln t\right\} \\
&\geq \Delta(t-K)\ln t \mathbb{P}_{\nu}\left\{\tau(E,t) < K\, \ln t\right\} \quad = \quad o(t^{\alpha}).
\end{aligned} \tag{A.2}$$

para cualquier constante positiva $K \leq t/\ln t$, donde Δ denota la máxima brecha de optimalidad bajo ν, y $\mathbb{E}_u$ y $\mathbb{P}_u$ denotan los operadores de expectativa y probabilidad cuando el vector de costos medios subyacente es dado por u. (Aunque varios elementos aleatorios, por ejemplo, S^t, dependen de la política π, ignoramos tal dependencia, para simplificar la exposición). A partir de la consistencia de π, se concluye que

$$\mathbb{P}_{\nu}\left\{\tau(E,t) < K\, \ln t\right\} = o(t^{\alpha-1}). \tag{A.3}$$

Para $a \in E$ y $k \leq \tau(\{a\}, t)$, se define la variable aleatoria $\eta(a,k)$ como el periodo en el cual una realización de costos de un arco a es observada en el k-ésimo periodo, i.e.

$$\eta(a,k)(\omega) := \inf\left\{t \geq 1 : \sum_{s\leq t} \mathbf{1}\left\{a \in S^s(\omega)\right\} = k\right\}, \quad k \leq \tau(\{a\}, t)(\omega),\, a \in E$$

y la variable aleatoria log-likelihood parcial

$$\mathcal{L}_{a,k}(\omega) := \sum_{i=1}^{k} \ln\left(\frac{f_a(c_a^{\eta(a,i)(\omega)}(\omega)|\mu_a)}{f_a(c_a^{\eta(a,i)(\omega)}(\omega)|\nu_a)}\right).$$

Se define el evento

$$W := \left\{\omega \in \Omega : \max_{a\in E}\left\{\mathcal{L}_{a,\tau(\{a\},t)}(\omega)\right\} \leq \frac{(1-\alpha)\ln t}{|E|},\, \tau(E,t)(\omega) < K \ln t\right\}.$$

Bajo los supuestos sobre F se tiene que

$$\begin{aligned}\mathbb{P}_\nu\{W\} = \int_W d\,\mathbb{P}_\nu &= \int_W \prod_{a\in E}\prod_{k=1}^{\tau(\{a\},t)} \frac{f_a(c_a^{\eta(a,k)}|\nu_a)}{f_a(c_a^{\eta(a,k)}|\mu_a)} d\,\mathbb{P}_\mu \\ &= \int_W \prod_{a\in E} exp(-\mathcal{L}_{a,\tau(\{a\},t)}) d\,\mathbb{P}_\mu \geq exp(-(1-\alpha)\ln t)\mathbb{P}_\mu\{W\} = \frac{\mathbb{P}_\mu\{W\}}{t^{1-\alpha}}.\end{aligned}$$

de (A.3), se concluye que

$$\lim_{t\to\infty} \mathbb{P}_\mu\{W\} = 0. \tag{A.4}$$

De SLLN se tiene que $\lim_{k\to\infty} \frac{1}{k}\mathcal{L}_{a,k} = L_a(\mu_a|\nu_a)$ a.s. $(\mathbb{P}_\mu)$ para $a \in E$, por ende

$$\lim_{k\to\infty} \frac{1}{k} \max\{\mathcal{L}_{a,l} : l \leq k\} = L_a(\mu_a|\nu_a) \quad a.s.\,(\mathbb{P}_\mu),\ a \in E.$$

Esto implica que, para todo $\delta > 1$,

$$\lim_{k\to\infty} \mathbb{P}_\mu\left\{\frac{\mathcal{L}_{a,l}}{k} > \delta\, L_a(\mu_a|\nu_a) \text{ for some } l \leq k\right\} = 0,$$

que implica que, tomando $k = \frac{(1-\alpha)\ln t}{\delta\,|E|\,L_a(\mu_a|\nu_a)}$,

$$\lim_{t\to\infty} \mathbb{P}_\mu\left\{\mathcal{L}_{a,l} > \frac{(1-\alpha)\ln t}{|E|} \text{ for some } l \leq \frac{(1-\alpha)\ln t}{\delta\,|E|\,L_a(\mu_a|\nu_a)}\right\} = 0.$$

La ecuación anterior implica que

$$\lim_{t\to\infty} \mathbb{P}_\mu\left\{\mathcal{L}_{a,\tau(\{a\},t)} > \frac{(1-\alpha)\ln t}{|E|},\ \tau(\{a\},t) \leq \frac{(1-\alpha)\ln t}{\delta\,|E|L_a(\mu_a|\nu_a)}\right\} = 0.$$

Se define $\kappa_E := \frac{(1-\alpha)|E|}{\delta} \min\{L_a(\mu_a|\nu_a)^{-1} : a \in E\}$: notando que $\tau(\{a\},t) \leq \tau(E,t)$ for $a \in E$, y mediante la unión se concluye que

$$\lim_{t\to\infty} \mathbb{P}_\mu\left\{\mathcal{L}_{a,n(\{a\},t)} > \frac{(1-\alpha)\ln t}{|E|},\ \tau(E,t) \leq \kappa_E\ln t, \text{for some } a \in E\right\} = 0,$$

Usando $K = \kappa_E$, (A.4) y lo anterior implica que

$$\lim_{t\to\infty} \mathbb{P}_\mu\{\tau(E,t) \leq \kappa_E\ln t\} = 0.$$

Por la arbitrariedad de $\alpha > 0$ y $\delta < 1$, el resultado se mantiene cuando se redefine

$$\kappa_E := |E| \min \left\{ L_a(\mu_a|\nu_a)^{-1} : a \in E \right\}.$$

■

Demostración. **Lema 5.1.** Notar que $\nu^\top z^\nu \leq \nu^\top z^{\bar{\nu}}$ por la optimalidad de z^ν bajo ν, por lo tanto, $w(\nu) = 0$ implica que no hay $\boldsymbol{x} \in \mathcal{X}$ para el cual se pueda hacer que $z^{\bar{\nu}}$ sea subóptimo, es decir, $z^\nu = z^{\bar{\nu}}$ para todo $x \in \mathcal{X}$, lo que en términos de los elementos básicos de la configuración es equivalente a $S(B, \nu) = S(B, \bar{\nu})$ para todo $B \in \mathcal{B}$. ■

Demostración. **Teorema 5.1.** Definimos $\tilde{A} := A \setminus S^*$. Para $t \geq t_0$, considerar $a_t \in \{\tau(a, t-1) \colon a \in S^t \cap \tilde{A}\}$. Si el *argmin* tiene más de un elemento, la elección de a_t se hace de manera aleatoria (independiente del resto). Para $a \in \tilde{A}$ se define $\widetilde{\tau}_a(t) = 0$ para $t < t_0$, y para $t \geq t_0$

$$\widetilde{\tau}_a(t) = \begin{cases} \widetilde{\tau}_a(t-1) + 1, & \text{si } a = a_t \\ \widetilde{\tau}_a(t-1), & \text{si no} \end{cases}$$

Para $a \in A$ y $s \geq 0$, y $t \geq t_0$, definimos

$$\bar{\mu}_a^t := \frac{\sum_{s<t} \mathbf{1}\left\{a \in S^s\right\} c_a^s}{\tau(a, t-1)}, \quad L_{t,s} := \sqrt{\frac{(\mathcal{L}+1)\ln t}{s}}.$$

Con esto, desde (5.6), se tiene que $\nu_a^t := \bar{\mu}_a^t + L_{t,\tau(a,t-1)}$. Usando estas definiciones, se tiene que, para $a \in \tilde{A}$ y $t \geq t_0$, y $\ell \geq 1$,

$$\begin{aligned} \widetilde{\tau}_a(t) &= \sum_{s=t_0}^{t} \mathbf{1}\left\{a = a_s\right\} \\ &\leq \ell + \sum_{s=t_0}^{t} \mathbf{1}\left\{a = a_s,\, \widetilde{\tau}_a(s-1) \geq \ell\right\} \\ &\leq \ell + \sum_{s=t_0}^{t} \mathbf{1}\left\{ \sum_{a' \in S^*} \left(\bar{\mu}_{a'}^{s-1} + L_{s,\tau(a',s-1)}\right) \leq \sum_{a' \in S^s} \left(\bar{\mu}_{a'}^{s-1} + L_{s,\tau(a',s-1)}\right), \widetilde{\tau}_a(s-1) \geq \ell \right\}. \end{aligned}$$

Notar que para cualquier $s \geq t_0$, en el evento de la función indicatriz en el término de la derecha en la ecuación de arriba, tenemos que $\ell \leq \widetilde{\tau}_a(s-1) \leq \tau(a', s-1)$ para todo $a' \in S^s$.

Para $s \leq t$, definimos $\mathcal{N}^s := \{(n_{a'} \colon a' \in S^*) \colon 1 \leq n_{a'} \leq s-1\}$. De forma similar, para $s \leq t$

y $B \in \mathcal{B}$, definimos $\mathcal{M}_\ell^s(B) = \{(m_{a'} \colon a' \in S(B)) \colon \ell \leq m_{a'} \leq s-1\}$. Considerando todos los valores posibles para los $\tau(a', s)$'s en el término de la derecha anterior, tenemos que

$$\mathbf{1}\left\{\sum_{a' \in S^*}\left(\bar{\mu}_{a'}^{s-1} + L_{s,\tau(a',s-1)}\right) \leq \sum_{a' \in S^s}\left(\bar{\mu}_{a'}^{s-1} + L_{s,\tau(a',s-1)}\right), \widetilde{\tau}_a(s-1) \geq \ell\right\} \leq$$

$$\sum_{B \in \mathcal{B}\setminus\{B^*\}} \sum_{n \in \mathcal{N}^s} \sum_{m \in \mathcal{M}_\ell^s(B)} \mathbf{1}\Big\{\sum_{a' \in S^*}\left(\bar{\mu}_{a'}^{s-1} + L_{s,n_{a'}}\right) \leq \sum_{a' \in S(B)}\left(\bar{\mu}_{a'}^{s-1} + L_{s,m_{a'}}\right),$$

$$\tau(a', s-1) = n_{a'},\ a' \in S^*,\ n \in \mathcal{N}^s,\ \tau(a', s-1) = m_{a'},\ a' \in S(B),\ m \in \mathcal{M}_\ell^s(B),\ B \in \mathcal{B} \setminus \{B^*\},\ s \leq t\Big\}$$

Notar que bajo las condiciones del término de la derecha en la ecuación anterior, $\bar{\mu}_{a'}^{s-1} \stackrel{d}{=} \bar{c}_{a',n_{a'}}$, donde para $a' \in A$ y k, definimos $\bar{c}_{a',k} := \frac{1}{k}\sum_{j \leq k} c_{a',j}$. Tomando esperanza en lo anterior, se conlcuye que

$$\mathbb{E}_F[\widetilde{\tau}_a(t)] \leq \ell + \sum_{s=t_0}^{\infty} \sum_{B \in \mathcal{B}\setminus\{B^*\}} \sum_{n \in \mathcal{N}^s} \sum_{m \in \mathcal{M}_\ell^s(B)} \mathbb{P}\Big\{\sum_{a' \in S^*}\left(\bar{c}_{a',n_{a'}} + L_{s,n_{a'}}\right) \leq \sum_{a' \in S(B)}\left(\bar{c}_{a,m_{a'}} + L_{s,m_{a'}}\right)\Big\}.$$

Considere el evento en la probabilidad del término de la derecha anterior. Para que un evento así ocurre, al menos una de las siguientes debe suceder.

$$\sum_{a' \in S^*} \bar{c}_{a',n_{a'}} \leq r(S^*, \mu) - \sum_{a' \in S^*} L_{s,n_{a'}} \tag{A.5}$$

$$\sum_{a' \in S(B)} \bar{c}_{a',m_{a'}} \geq r(S(B), \mu) + \sum_{a' \in S(B)} L_{s,m_{a'}} \tag{A.6}$$

$$r(S^*, \mu) < r(S(B), \mu) + 2\sum_{a' \in S(B)} L_{s,m_{a'}}. \tag{A.7}$$

Luego, acotamos la probabilidad de estos eventos y concluímos aplicando cota de la unión. Para esto, usamos el siguiente lema que se prueba directamente de la desigualdad de Hoeffding por lo que se omite. (Desigualdad Hoeffding para la suma de sampleos medios). Sea X_{ij}, $j = 1, \ldots, n_i$ y $i = 1, \ldots, k$ variables aleatorias independientes en $[0, 1]$ tal que$\mathbb{E}[X_{ij}] = \mu_i$ para todo $j = 1, \ldots, n_i$, $i = 1, \ldots, k$. Luego

$$\mathbb{P}\Big(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\Big(\frac{X_{ij} - \mu_i}{n_i}\Big) \geq x\Big) \leq \exp\Big(-\frac{2x^2}{\sum_{i=1}^{k}\frac{1}{n_i}}\Big).$$

Considere la probabilidad que (A.5) sea cierta. Arreglando el orden de los términos y

utilizando el lema anterior se tiene que

$$\begin{aligned}
\mathbb{P}(\sum_{a'\in S^*}\bar{c}_{a',n_{a'}} \leq r(S^*,\mu) - \sum_{a'\in S^*} L_{s,n_{a'}}) &= \mathbb{P}\Big(\sum_{a'\in S^*}\sum_{j=1}^{n'_a}\Big(\frac{c_{a',j}-\mu_{a'}}{n_{a'}}\Big) \leq -\sum_{a'\in S^*} L_{s,n_{a'}}\Big) \\
&\leq \exp\Big(-\frac{2(\sum_{a'\in S^*} L_{s,n_{a'}})^2}{\sum_{a'\in S^*}\frac{1}{n_{a'}}}\Big) \\
&\quad \Big(-2(\mathcal{L}+1)\frac{\Big(\sum_{a'\in S^*}\frac{1}{\sqrt{n_{a'}}}\Big)^2}{\Big(\sum_{a'\in S^*}\frac{1}{n_{a'}}\Big)}\Big) \\
&= s \\
&\leq s^{-2(\mathcal{L}+1)},
\end{aligned}$$

donde la última desigualdad viene del hecho que $\Big(\sum_{a'\in S^*}\frac{1}{\sqrt{n_{a'}}}\Big)^2\Big(\sum_{a'\in S^*}\frac{1}{n_{a'}}\Big)^{-1} \geq 1$. Following the same arguments, it can be shown that the probability that (A.6) holds true admits the same bound, i.e.

$$\mathbb{P}\Big(\sum_{a'\in S(B)}\bar{c}_{a',m_{a'}} \geq r(S(B),\mu) + \sum_{a'\in S(B)} L_{s,m_{a'}}\Big) \leq s^{-2(\mathcal{L}+1)}.$$

Considere ahora (A.7). Dado que $m_a \geq \ell$, se tiene

$$\ell \geq \frac{4(\mathcal{L}+1)\ln s\,\mathcal{L}^2}{\Delta_{\min}^2} \;\Rightarrow\; \Delta_B^2 \geq \frac{4(\mathcal{L}+1)\ln s\,\mathcal{L}^2}{\ell} \Rightarrow \Delta_B \geq 2\sum_{a\in S(B)} L_{s,m_a} \quad \forall B \in \mathcal{B}\setminus\{B^*\}.$$

Notar que la desigualdad anterior implica que (A.7) no se tiene . Por ende, considerando

$$\ell = \left\lceil \frac{4(\mathcal{L}+1)\ln t\,\mathcal{L}^2}{\Delta_{\min}^2} \right\rceil,$$

la probabilidad de que suceda (A.7) es cero. Luego, usando cota de la unión, se concluye que

$$\mathbb{E}_F[\widetilde{\tau}_a(t)] \leq \left\lceil \frac{4(\mathcal{L}+1)\ln t\,\mathcal{L}^2}{\Delta_{\min}^2} \right\rceil + \sum_{s=t_0}^{\infty}\sum_{B\in\mathcal{B}\setminus\{B^*\}}\sum_{n\in\mathcal{N}^s}\sum_{m\in\mathcal{M}_\ell^s(B)} s^{-2(\mathcal{L}+1)}.$$

Observar que el segundo termino en el lado derecho de la desigualdad es finito y no depende de t. En efecto, notar que $|\mathcal{N}^s| \leq t^{\mathcal{L}}$ y $|\mathcal{M}^s(B)| \leq t^{\mathcal{L}}$ para cualquier $B \in \mathcal{B}$, por lo que el término de la derecha admite una cota superior $\sum_{s=t_0}^{\infty}\sum_{B\in\mathcal{B}\setminus\{B^*\}} s^{-2} \leq \frac{\pi^2|\mathcal{B}|}{6} < \infty$.

Finalmente, usamos la cota superior en tiempo finito sobre $\mathbb{E}_F[\widetilde{\tau}_a(t)]$ para proveer una cota

en el regret esperado. Para esto, para $B \in \mathcal{B} \setminus \{B^*\}$, se define $N_B^t := \sum_{s \leq t} \mathbf{1}\{B^t = B\}$, y notar que

$$\mathbb{E}_F\left[\mathcal{R}^{\text{nUCB}}(t, F)\right] \leq \Delta_{\max} \sum_{B \in \mathcal{B} \setminus \{B^*\}} \mathbb{E}[N_B^t] = \Delta_{\max} \sum_{a \in \tilde{A}} \mathbb{E}\left[\widetilde{\tau}(a, t)\right] \leq \Delta_{\max} |\tilde{A}| \left(\left\lceil \frac{4(\mathcal{L}+1) \ln t \, \mathcal{L}^2}{\Delta_{\min}^2} \right\rceil + \frac{\pi^2 |\mathcal{B}|}{6} \right).$$

■