Progetto di

Intelligenza artificiale

PAC Identification of Many Good Arms in Stochastic Multi-Armed Bandits
(Arghya Roy Chaudhuri, Shivaram Kalyanakrishnan)

Samuele Ferri [1045975]

a.a. 2019-2020 (Sessione di settembre)

Università degli studi di Bergamo
Scuola di Ingegneria
Corso di laurea magistrale in Ingegneria Informatica
v 0.0.1

Indice

1. Introduzione

In questo progetto verrà analizzato approfonditamente il paper «*PAC Identification of Many Good Arms in Stochastic Multi-Armed Bandits*» pubblicato da Arghya Roy Chaudhuri e Shivaram Kalyanakrishnan a inizio 2019 e mostrato in *Proceedings of the 36 th International Conference on Machine Learning*, Long Beach, California, PMLR 97.

Nel prossimo paragrafo verrà presentato brevemente il contenuto sotto forma di abstract. Qui di seguito una breve descrizione dei vari capitoli presenti nell'elaborato svolto.

- Nel **capitolo 2** verrà descritto il contesto in cui si colloca il lavoro, quale è il problema trattato e quali sono i possibili scenari applicativi.
- Nel **capitolo 3** verranno descritti e commentati i lavori precedenti in merito allo stesso problema. Oltre ai lavori citati nei *related works* e nei riferimenti bibliografici del lavoro analizzato verranno descritti anche altri lavori presenti nella letteratura che sono correlati al problema analizzato. Inoltre, verranno elencate anche quali delle tecniche viste in classe potrebbero essere utilizzate per risolvere il problema in analisi.
- Nel **capitolo 4** verrà descritto dettagliatamente il lavoro presentato nel paper, comprese le proprietà teoriche usate.
- Nel **capitolo 5** verranno descritti e replicati gli esperimenti svolti nel paper; non essendoci codice è stato richiesto di limitarsi a replicare gli esperimenti presenti nel paper.
- Nel **capitolo 6** verranno elencate le conclusioni sulle proprietà teoriche e sperimentali dei metodi analizzati nel paper ed eventuali scenari applicativi nella vita reale.
- Nel **capitolo 7** verranno fatte delle considerazioni personali sull'impatto che questo progetto ha avuto sia in ambito universitario/lavorativo che nella quotidianità.

Nelle **appendici A, B e C** verranno dimostrate le proprietà teoriche presenti nel paper.

Infine, sono presenti anche gli indici degli algoritmi, delle figure e i riferimenti bibliografici citati nell'elaborato.

1.1. Abstract

Nell'ambito *PAC*[1] verrà considerato il problema di identificare un numero k qualsiasi tra le migliori m arms in un *n-armed stochastic multi-armed bandit*. Questo particolare problema generalizza sia il problema della "migliore selezione del sottoinsieme" [KS10] sia quello della selezione di "una delle migliori m-arms" [CK17]. In applicazioni come il crowdsourcing e la progettazione di farmaci, identificare una singola buona soluzione spesso non è sufficiente. Inoltre, trovare il sottoinsieme migliore potrebbe essere difficile a causa della presenza di molte soluzioni indistinguibilmente vicine. La generalizzazione di identificare esattamente k arms dalle migliori m, dove $1 \leq k \leq m$, è un'alternativa più efficace. Verrà presentato un limite inferiore alla complessità del caso peggiore per il generico k e un algoritmo *PAC* completamente sequenziale molto più efficiente in casi semplici. Inoltre, estendendo l'analisi a *infinite-armed bandit*, verrà presentato un algoritmo *PAC* che è indipendente da n, che identifica un arm dalla migliore frazione ρ di arms usando al massimo un numero (polinomiale-logatrimico) addizionale di campioni rispetto al limite inferiore, migliorando così rispetto a [CK17]; [Azi+18]. Il problema di identificare $k > 1$ arms distinte dalla frazione ρ migliore non è sempre ben definito; per una classe speciale di questo problema verranno presentati i limiti inferiore e superiore. Infine, attraverso una riduzione, verrà stabilita una relazione tra i limiti superiori per il problema "una delle migliori ρ" per istanze infinite e quello "una delle migliori m" per le istanze finite. Verrà ipotizzato che sia più efficiente risolvere istanze "piccole" finite usando quest'ultima formulazione, piuttosto che passare attraverso la prima.

[1] ***Probably Approximately Correct (PAC) Learning*:** nella teoria dell'apprendimento computazionale, l'apprendimento approssimativamente corretto (*PAC*) è un framework per l'analisi matematica del machine learning proposto nel 1984 da Leslie Valiant. In questo framework, il learner riceve campioni e deve selezionare una funzione di generalizzazione (chiamata ipotesi) da una certa classe di possibili funzioni. L'obiettivo è che, con alta probabilità, la funzione selezionata abbia un errore di generalizzazione basso. Il learner deve essere in grado di apprendere il concetto dato qualsiasi rapporto di approssimazione arbitrario, probabilità di successo o distribuzione dei campioni.

2. Contesto

Prima di illustrare il problema centrale analizzato dal paper, definisco il problema dello *stochastic multi-armed bandit* e descrivo i lavori integrativi svolti nel corso degli anni riguardanti questo ambito.

Il problema dello *stochastic multi-armed bandit* [Rob52]; [BF85] è un problema ben studiato riguardante decisioni in condizioni di incertezza. Ogni leva (*arm*) di un bandit rappresenta una decisione. Un *pull* della leva rappresenta prendere la decisione associata che produce una ricompensa effettiva. La ricompensa è determinata da una distribuzione i.i.d. corrispondente all'arm selezionato, indipendente dai pull delle altre arms. Ogni possibile alternativa deve essere indipendente dalle altre, ossia che le decisioni prese precedentemente non condizionino il reward della scelta attuale. Ad ogni turno, il giocatore può consultare la precedente storia dei pull effettuati e delle ricompense ricevute per decidere quale arm tirare.

Il nome deriva dalle slot machines: il problema può essere visto come un giocatore d'azzardo avente di fronte una fila di slot machine: il giocatore deve decidere quali macchine giocare, quante volte giocare ogni macchina e in quale ordine giocarle, e se continuare con la macchina corrente o provare un'altra macchina. Oppure può essere visto come una sola slot machine con più leve (*multi-armed*) che possono essere tirate, ognuna con una propria probabilità di vincere denaro; il giocatore, inizialmente ignoto di qualsiasi caratteristica delle leve presenti, deve trovare e scegliere la leva che gli porti ad ottenere un maggiore quantitativo di denaro.

Il giocatore deve elaborare una strategia, deve capire quando conviene provare nuove scelte (*exploration*) oppure continuare a scegliere di tirare la leva più promettente in base a quanto ha appreso (*exploitation*). Vi è quindi un compromesso tra il continuare a sfruttare la leva che ha il profitto più alto atteso oppure continuare a provare nuove leve ad ogni turno cercando di esplorare e conoscere maggiori informazioni sui reward che possono dare le altre leve.

È quindi un problema di *reinforcement learning*, si vuole massimizzare il reward medio ottenibile. L'obiettivo del giocatore è massimizzare la ricompensa cumulativa attesa (*reward*) data una serie di pull, oppure equivalentemente minimizzare il rimpianto (*regret*) tirando sempre un solo arm.

Un problema a parte è quello di identificare un arm con la più alta ricompensa media [Bec58]; [Pau64]; [EMM02] sotto quello che viene chiamato *pure exploration regime*. Per applicazioni come product testing [AB10] e strategy selection [Gos+13], c'è una fase dedicata nell'esperimento in cui i premi ottenuti sono irrilevanti. Piuttosto,

l'obiettivo è quello di identificare l'arm migliore (1) in numero minimo di prove, data una determinata soglia di confidenza [EMM02]; [KS10], o in alternativa, (2) con errore minimo, dopo un determinato numero di prove [AB10]; [CV15]. Lo studio presente nel paper rientra nella prima categoria, che viene definita *fixed confidence setting*. Concepito da [Bec58], l'identificazione del arm migliore in *fixed confidence setting* ha ricevuto una notevole attenzione nel corso degli anni [EMM02]; [Gab+11]; [KKS13]; [JN14]. Il problema è stato anche generalizzato per identificare il miglior sottoinsieme di arms [Kal+12].

Più recentemente, Roy Chaudhuri e Kalyanakrishnan [CK17] hanno introdotto il problema di identificare un singolo arm tra i migliori m in un *n-armed-bandit*. Questa formulazione è particolarmente utile quando il numero di arms è grande, e in effetti è una valida alternativa anche quando il numero di arms è *infinito*. In molti scenari pratici, tuttavia, è necessario identificare più di un singolo arm buono. Per ad esempio, si immagini che un'azienda debba completare un lavoro che è troppo grande per essere realizzato da un singolo lavoratore, ma che può essere suddiviso in 5 sotto attività, ciascuna capace di essere completata da un solo lavoratore. Supponiamo che ci siano un totale di 1000 lavoratori e, grazie a un sondaggio, si è rilevato che almeno il 15% dei lavoratori ha le competenze per completare la sotto attività. Per rispondere alle esigenze dell'azienda, sicuramente sarebbe sufficiente identificare i 5 migliori lavoratori per la sotto attività. Tuttavia, se i lavoratori devono essere identificati sulla base di un test di abilità che ha risultati stocastici, questo test sarebbe inutilmente costoso se il fine è identificare il miglior sottoinsieme di lavoratori (*best subset* selection). Piuttosto sarebbe sufficiente identificare 5 lavoratori tra i migliori 150. Questo è precisamente il problema che trattato nel paper: l'identificazione di k qualsiasi tra le migliori m arms di un *n-armed bandit.*

Il problema assume uguale significato da un punto di vista teorico, dal momento che generalizza sia il problema di selezione del miglior sottoinsieme (*best subset selection*) [KS10] (dato $k = m$) e il problema di selezionare un arm singolo da miglior sottoinsieme (*single arm from the best subset*) [CK17] (dato $k = 1$). A differenza del *best subset selection*, il problema rimane fattibile da risolvere anche quando n è grande o infinito, fintanto che il rapporto m/n è una costante $\rho > 0$. Tradizionalmente, *infinite-armed bandits* sono stati affrontati ricorrendo a informazioni secondarie come le distanze tra gli arms [Agr95]; [Kle05] o la struttura della loro distribuzione dei rewards [WAM09]. Questo approccio introduce parametri aggiuntivi, che potrebbero non essere facili da settare in pratica. In alternativa, buoni arms possono essere raggiunti semplicemente selezionando gli arms a caso e testando facendo il pull. Quest'ultimo approccio è stato applicato con successo sia in *regret minimization setting* [HPR96] che in *fixed confidence setting* [Gos+13]; [CK17]. La formulazione presente nel testo apre la strada all'identificazione di "molti (k) buoni" (tra i migliori m degli n) arms in questo modo.

Nel paper verranno proposti diversi algoritmi *Probably Approximately Correct Learning* (*PAC*): l'apprendimento approssimativamente corretto (*PAC*) è un framework per l'analisi matematica del machine learning proposto nel 1984 da Leslie Valiant. In

questo framework, il learner riceve campioni e deve selezionare una funzione di generalizzazione (chiamata ipotesi) da una certa classe di possibili funzioni. L'obiettivo è che, con alta probabilità, la funzione selezionata abbia un errore di generalizzazione basso. Il learner deve essere in grado di apprendere il concetto dato qualsiasi rapporto di approssimazione arbitrario, probabilità di successo o distribuzione dei campioni.

Nella tabella presente in figura 2.1 vi è un riepilogo dei risultati teorici che saranno trattati nel paper.

Problem	Lower Bound	Previous Upper Bound	Current Upper Bound
$(1,1,n)$ Best-Arm	$\Omega\left(\frac{n}{\epsilon^2}\log\frac{1}{\delta}\right)$ (Mannor & Tsitsiklis, 2004)	$O\left(\frac{n}{\epsilon^2}\log\frac{1}{\delta}\right)$ (Even-Dar et al., 2002)	Same as previous
(m,m,n) SUBSET	$\Omega\left(\frac{n}{\epsilon^2}\log\frac{m}{\delta}\right)$ (Kalyanakrishnan et al., 2012)	$O\left(\frac{n}{\epsilon^2}\log\frac{m}{\delta}\right)$ (Kalyanakrishnan & Stone, 2010)	Same as previous
$(1,m,n)$ Q-F	$\Omega\left(\frac{n}{m\epsilon^2}\log\frac{1}{\delta}\right)$ (Roy Chaudhuri & Kalyanakrishnan, 2017)	$O\left(\frac{n}{m\epsilon^2}\log^2\frac{1}{\delta}\right)$	$O\left(\frac{1}{\epsilon^2}\left(\frac{n}{m}\log\frac{1}{\delta}+\log^2\frac{1}{\delta}\right)\right)$ **This paper**
(k,m,n) Q-F$_k$	$\Omega\left(\frac{n}{(m-k+1)\epsilon^2}\log\frac{\binom{m}{k-1}}{\delta}\right)$ **This paper**	-	$O\left(\frac{k}{\epsilon^2}\left(\frac{n\log k}{m}\log\frac{k}{\delta}+\log^2\frac{k}{\delta}\right)\right)^*$ **This paper** (*for $k\geq 2$)
$(1,\rho)$ $(\lvert\mathcal{A}\rvert=\infty)$ Q-P	$\Omega\left(\frac{1}{\rho\epsilon^2}\log\frac{1}{\delta}\right)$ (Roy Chaudhuri & Kalyanakrishnan, 2017)	$O\left(\frac{1}{\rho\epsilon^2}\log^2\frac{1}{\delta}\right)$	$O\left(\frac{1}{\epsilon^2}\left(\frac{1}{\rho}\log\frac{1}{\delta}+\log^2\frac{1}{\delta}\right)\right)$ **This paper**
(k,ρ) $(\lvert\mathcal{A}\rvert=\infty)$ Q-P$_k$	$\Omega\left(\frac{k}{\rho\epsilon^2}\log\frac{k}{\delta}\right)$ **This paper**	-	$O\left(\frac{k}{\epsilon^2}\left(\frac{\log k}{\rho}\log\frac{k}{\delta}+\log^2\frac{k}{\delta}\right)\right)^*$ **This paper** (*for a special class with $k\geq 2$)

Figura 2.1.: Limiti inferiore e superiore sulla complessità del campione attesa (ponendosi nel caso peggiore). I limiti per $(k;\rho)$, $k>1$ sono per la classe speciale di istanze "al massimo k-equiprobabili".

2.1. Scenari applicativi

In pratica, il problema dello *stochastic multi-armed bandit* è stato usato per modellare problemi come la gestione di progetti di ricerca di grandi organizzazioni sia in ambito scientifico che farmaceutico. Descrivo solo alcuni scenari citati anche dal paper; ulteriori applicazioni pratiche saranno presentate nella sezione 6.1 così come alcune applicazioni in ambito *machine learning* che saranno presentate nella sezione 6.2.

Clinical Trials

Il trial clinico si riferisce a uno studio medico farmacologico, biomedico o correlato alla salute sull'uomo ben definito da protocolli. L'obiettivo è quello di verificare che una nuova terapia sia più efficace, migliore e soprattutto sicura di quella normalmente impiegata. Assume notevole importanza l'insieme di campioni sui quali poter testare nuove cure come descritto da Robbins, Herbert in *«Some aspects of the sequential design of experiments»* [Rob52].

Figura 2.2.: Clinical Trials

Un'altra applicazione sempre in questo ambito è quella presentata da Durand, A., Achilleos, C., Iacovides, D., Strati, K., Mitsis, G. D., & Pineau, J. in *«Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis»* [Dur+18] dove si vuole progettare una strategia adattiva di allocazione per migliorare l'efficienza della raccolta dati allocando più campioni per esplorare trattamenti promettenti. Questa applicazione è stata vista come un *contextual bandit problem* e hanno introdotto un algoritmo pratico per il trade-off tra *exploration* e *exploitation*.

Drug Design

Vi sono applicazioni anche riguardanti la progettazione di farmaci come descritto da Will, Yvonne, McDuffie, J Eric, Olaharski, Andrew J, and Jeffy, Brandon D in *«Drug Discovery Toxicology: From Target Assessment to Translational Biomarkers»* [Wil+16]. È una guida per i professionisti farmaceutici ai problemi e alle pratiche della tossicologia della scoperta di farmaci in cui vengono usati anche alcuni algoritmi per risolvere il problema dei *stochastic multi-armed bandits*.

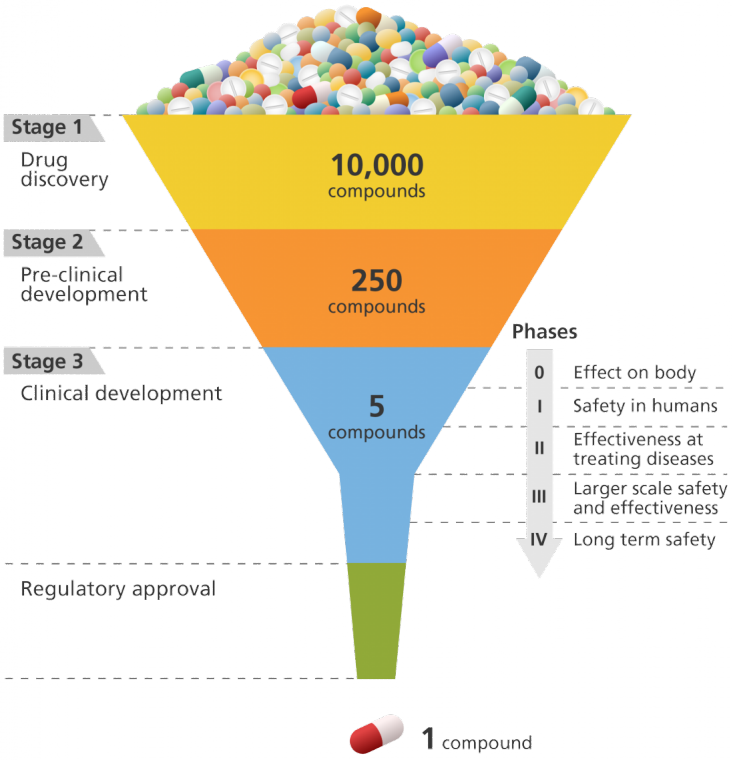

Figura 2.3.: Drug Design stages

Internet Advertising

Ogni volta che un utente visita un sito è necessario scegliere di mostrare una delle k pubblicità possibili; la ricompensa si ottiene se un utente seleziona col mouse la pubblicità. Inizialmente non si ha nessuna conoscenza dell'utente, del contenuto dell'annuncio e del contenuto della pagina web richiesta. Questo è un caso dei *recommender system,* descritti in *«A contextual-bandit approach to personalized news article recommendation»* da Li, Lihong, Chu, Wei, Langford, John, and Schapire, Robert E. [Li+10]; altre applicazioni possibili sono sul consigliare video correlati a quello che l'utente sta guardando su piattaforme come *YouTube*.

La *online recommendation* è una caratteristica importante in molte applicazioni. In pratica, l'interazione tra gli utenti e il sistema di raccomandazione potrebbe essere scarsa, vale a dire che gli utenti non interagiscono sempre con il sistema di segnalazione. Ad esempio, alcuni utenti preferiscono scorrere la raccomandazione anziché fare click sui dettagli. Pertanto, una risposta nulla «0» potrebbe non essere necessariamente una risposta negativa, ma una non risposta. È peggio distinguere queste due situazioni quando si consiglia all'utente un solo elemento alla volta e sono disponibili solo poche ulteriori informazioni. La maggior parte delle strategie di raccomandazione esistenti ignorano la differenza tra mancate risposte e risposte negative. In questo documento [Li+10], viene proposto un nuovo approccio, denominato *SAOR*, per formulare

raccomandazioni online tramite interazioni sparse. *SAOR* utilizza risposte positive e negative per creare il modello delle preferenze dell'utente, ignorando tutte le non risposte. Viene fornita un'analisi del regret di *SAOR*, gli esperimenti su set di dati reali mostrano anche che *SAOR* supera i metodi concorrenti.

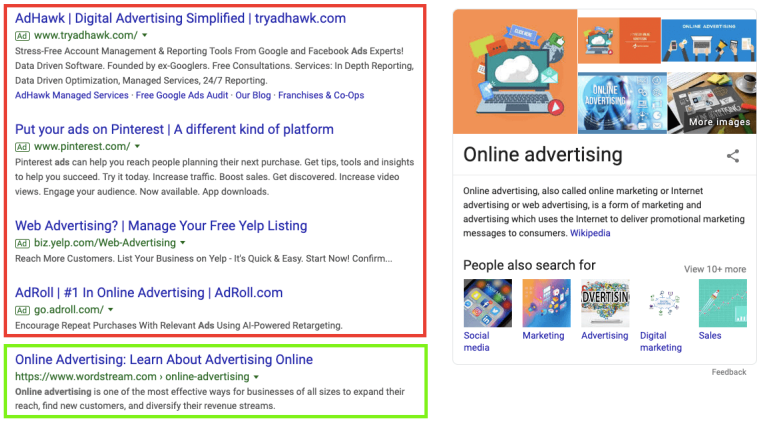

Figura 2.4.: Google Ads

Network Server Selection

Nel caso in cui un lavoro deve essere elaborato su uno dei numerosi server, ognuno dei quali ha differenti velocità di processo dovute a distanza geografica, carico ecc., ogni server può essere visto come un arm; nel tempo si vuole apprendere quale sia il miglior arm da usare. Questo problema è stato applicato nel routing (*adaptive routing*) [AK08], nel *DNS server selection* e nel *cloud computing*.

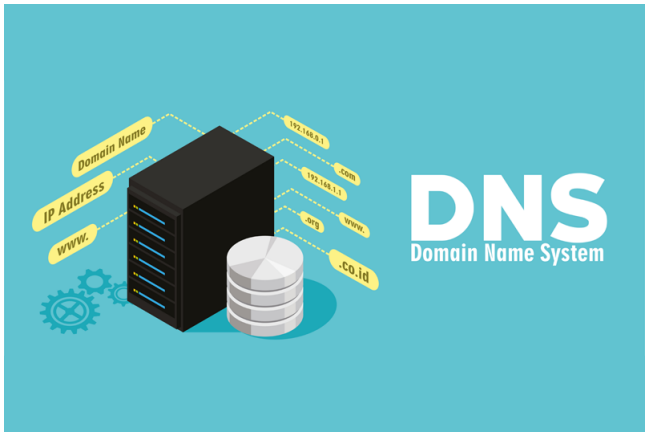

Figura 2.5.: Domain Name System (DNS)

Nell'articolo *«Online linear optimization and adaptive routing»* di Baruch Awerbuch e Robert Kleinberg [AK08] si studia un problema di ottimizzazione lineare online

generalizzando il problema del *stochastic multi-armed bandit*. Motivati principalmente dal compito di progettare algoritmi di routing adattivi per reti sovrapposte, gli autori presentano due *randomized online algorithms* per selezionare una sequenza di percorsi di routing in una rete con ritardi alle frontiere sconosciuti che variano nel tempo in modo imprevedibile. Contrariamente ai precedenti lavori su questo problema, viene supposto che l'unico feedback dopo aver scelto un determinato percorso sia il ritardo totale end-to-end del percorso selezionato. Vengono presentati due algoritmi il cui regret è sub lineare nel numero di prove e polinomiale nelle dimensioni della rete. Il primo di questi algoritmi generalizza per risolvere qualsiasi problema di ottimizzazione lineare online, dato un oracolo per l'ottimizzazione delle funzioni lineari sull'insieme delle strategie; il lavoro degli autori può quindi essere interpretato come una riduzione generalizzata dall'ottimizzazione lineare offline a quella online. Un elemento chiave di questo algoritmo è la nozione di *barycentric spanner*, un tipo speciale di base per lo spazio vettoriale che consente a qualsiasi strategia possibile di essere espressa come una combinazione lineare di vettori di base utilizzando coefficienti limitati. Inoltre è presentato anche un secondo algoritmo per il problema del percorso più breve (online), che risolve il problema utilizzando una catena di oracoli decisionali (online), uno su ciascun nodo del grafico. Ciò presenta numerosi vantaggi rispetto all'approccio di ottimizzazione lineare online. In primo luogo, è efficace contro un avversario adattivo, mentre l'algoritmo sviluppato di ottimizzazione lineare assume un avversario inconsapevole. In secondo luogo, anche nel caso di un avversario inconsapevole, il secondo algoritmo si comporta leggermente meglio del primo, come misurato dal loro regret additivo.

Gestione di grandi reti di sensori

Le applicazioni di questo problema includono la gestione di grandi reti di sensori come descritto da Mousavi, Seyed Hamed, Haghighat, Javad, Hamouda, Walaa, and Dastbasteh, Reza in *«Analysis of a Subset Selection Scheme for Wireless Sensor Networks in Time-Varying Fading Channels»* [Mou+16], in cui più sensori affidabili devono essere identificati facendo il minor numero di tests possibili.

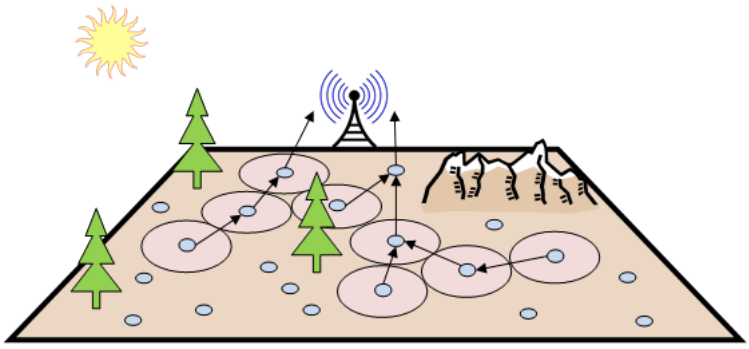

Figura 2.6.: Wireless Sensor Network (WSN)

Una delle principali sfide che affrontano le reti di sensori wireless (*WSN*) sono le limitate risorse di energia disponibili nei piccoli nodi dei sensori. Si desidera pertanto ridurre il consumo di energia dei sensori mantenendo la distorsione tra la sorgente e la sua stima nel centro di fusione (*FC*) al di sotto di una soglia specifica. In questo documento [Mou+16], viene analizzata una strategia di selezione dei sottoinsiemi per ridurre la potenza di trasmissione media della *WSN*. Si considera una rete a due hop e ipotizzando che i canali tra la sorgente e i *relay sensor* siano *time-varying fading channels*, modellati come canali *Gilbert-Elliott*. Viene mostrato che quando questi canali sono noti all'*FC*, l'*FC* può selezionare un sottoinsieme di sensori per ridurre al minimo la potenza di trasmissione soddisfacendo al contempo il criterio di distorsione. Attraverso l'analisi, viene ricavata la distribuzione di probabilità della dimensione di questo sottoinsieme. Vengono anche considerati aspetti pratici dell'attuazione del metodo proposto, compresa la stima dei canali ai *relay*. Attraverso simulazioni, vengono confrontate le prestazioni dello schema proposto con gli schemi che appaiono in letteratura. I risultati della simulazione confermano che per un certo intervallo di *end-to-end bit-error rates (BERs)*, lo schema proposto riesce a ottenere una riduzione di potenza superiore rispetto ad altri schemi.

Distribuited Crowdsouring

Vi sono applicazioni riguardanti il crowdsourcing distribuito come presentato da Tran-Thanh, Long, Stein, Sebastian, Rogers, Alex, and Jennings, Nicholas R. in *«Efficient crowdsourcing of unknown experts using bounded multi-armed bandits»* [Tra+14].

Figura 2.7.: Crowdsourcing

Sempre più organizzazioni esternalizzano in modo flessibile il lavoro su base temporanea a un pubblico globale di lavoratori. Il crowdsourcing è stato applicato con successo a una serie di lavori, dalla traduzione di testi e annotazioni di immagini, alla raccolta di informazioni durante le situazioni di crisi e all'assunzione di lavoratori qualificati per creare software complessi. Mentre tradizionalmente questi compiti sono stati piccoli e potrebbero essere completati da non professionisti, le organizzazioni stanno ora iniziando a fare crowdsourcing di compiti più grandi e più complessi agli esperti nei loro rispettivi campi. Queste attività includono, ad esempio, lo sviluppo e

test del software, web design e marketing del prodotto. Mentre questo crowdsourcing di esperti emergenti offre flessibilità e costi potenzialmente inferiori, solleva anche nuove sfide, poiché i lavoratori possono essere altamente eterogenei, sia nei costi che nella qualità del lavoro che producono. In particolare, l'utilità di ciascuna attività esternalizzata è incerta e può variare in modo significativo tra lavoratori distinti e persino tra compiti successivi assegnati allo stesso lavoratore. Inoltre, in contesti realistici, i lavoratori hanno limiti alla quantità di lavoro che possono svolgere e il datore di lavoro avrà un budget fisso per i lavoratori paganti. Data questa incertezza e i relativi vincoli, l'obiettivo del datore di lavoro è quello di assegnare compiti ai lavoratori al fine di massimizzare l'utilità complessiva raggiunta. Per formalizzare questo problema di crowdsourcing, viene introdotto un nuovo multi-armed bandit (MAB), il *bounded MAB*. Inoltre, viene sviluppato un algoritmo per risolvere il problema in modo efficiente, chiamato *bounded ε-first*, che procede in due fasi: *exploration* e *exploitation*. Durante l'*exploration*, l'algoritmo usa prima εB del suo budget totale B per apprendere stime delle caratteristiche di qualità dei lavoratori. Quindi, durante l'*exploitation*, utilizza il rimanente $(1 - \varepsilon)B$ per massimizzare l'utilità totale in base a tali stime. L'utilizzo di questa tecnica ci consente di ricavare un limite superiore $O(B^{\frac{2}{3}})$ dal suo regret di prestazione (ovvero, la differenza attesa nell'utilità tra l'algoritmo e l'ottimale), il che significa che quando il budget B aumenta, il regret tende a 0. Oltre a questo approccio teorico, l'algoritmo viene applicato ai dati del mondo reale usando *oDesk*, un importante sito di crowdsourcing. Utilizzando i dati di progetti reali, inclusi budget di progetti storici, costi di esperti e valutazioni di qualità, viene dimostrato che l'algoritmo supera i metodi di crowdsourcing esistenti fino al 300%, ottenendo al contempo un massimo ipotetico con informazioni complete.

3. Studi correlati

3.1. Lavori citati nel paper

In questa sezione verranno presentati in ordine alfabetico tutti i lavori citati nel paper corredati da una descrizione riguardante il contenuto del problema principale trattato da ciascuno di essi.

[Agr95] Agrawal, R. *The continuum-armed bandit problem.* SIAM J. Control Optim., 33(6):1926–1951, 1995.

In questo articolo viene considerato il problema del *stochastic multi-armed bandit* in cui le arms sono scelte da un sottoinsieme dei numeri reali e si presume che le ricompense medie siano una funzione continua delle arms. Il problema con un numero infinito di arms è molto più difficile del solito problema con un numero finito di arms perché il *built-in learning* è ora di dimensione infinita. Viene elaborato uno schema di apprendimento basato sullo stimatore a kernel per la ricompensa media in funzione degli arms. Usando questo schema di apprendimento, viene costruita una classe di *certainty equivalence control with forcing schemes* e successivamente vengono derivati i limiti superiori asintotici rispetto alla loro perdita di apprendimento. In base ai dati in loro possesso, questi limiti sono i rates più restrittivi finora disponibili.

[AB10] Audibert, J.-Y., Bubeck, S., and Munos, R. *Best arm identification in multi-armed bandits.* In Proc. COLT 2010, pp. 41–53. Omnipress, 2010.

In questo articolo viene trattato il problema di trovare l'arms migliore in un *stochastic multi-armed bandit.* Il rimpianto (*regret*) di un previsore è qui definito dal gap tra la ricompensa media del arm ottimale e la ricompensa media del arm scelto. Gli autori propongono una *UCB policy*[1] altamente esplorativa e un nuovo algoritmo basato su scarti successivi. Viene mostrato che questi algoritmi sono essenzialmente ottimali poiché il loro rimpianto diminuisce esponenzialmente a una velocità che è, fino a un fattore logaritmico, il migliore possibile. Tuttavia, mentre la *UCB policy* richiede l'ottimizzazione di un parametro in base alla complessità non osservabile dell'attività, la *successive rejects policy* beneficia di essere priva di parametri e indipendente dal ridimensionamento dei premi. Come sottoprodotto della loro analisi, viene mostrato che l'identificazione del arm migliore (quando è unico) richiede un numero di campioni

[1] ***Upper Confidence Bound (UCB) Policy*:** pone un limite superiore al valore della ricompensa ottenibile

di ordine (fino a un fattore *log(K)*) $\Sigma i \frac{1}{\Delta_i^2}$, dove la somma è sugli arms non ottimali e Δ_i rappresenta la differenza tra la ricompensa media del arm migliore e quella del arm *i*. Ciò generalizza il fatto ben noto che è necessario un ordine di $1/\Delta^2$ campioni per differenziare le medie di due distribuzioni con gap Δ.

[AK08] Awerbuch, B. and Kleinberg, R. *Online linear optimization and adaptive routing*. In J. Comput. Syst. Sci., volume 74, pp. 97–114. Academic Press, Inc., 2008.

In questo articolo si studia un problema di ottimizzazione lineare online generalizzando il problema del *stochastic multi-armed bandit*. Motivati principalmente dal compito di progettare algoritmi di routing adattivi per reti sovrapposte, gli autori presentano due *randomized online algorithms* per selezionare una sequenza di percorsi di routing in una rete con ritardi alle frontiere sconosciuti che variano nel tempo in modo imprevedibile. Contrariamente ai precedenti lavori su questo problema, viene supposto che l'unico feedback dopo aver scelto un determinato percorso sia il ritardo totale end-to-end del percorso selezionato. Vengono presentati due algoritmi il cui regret è sub lineare nel numero di prove e polinomiale nelle dimensioni della rete. Il primo di questi algoritmi generalizza per risolvere qualsiasi problema di ottimizzazione lineare online, dato un oracolo per l'ottimizzazione delle funzioni lineari sull'insieme delle strategie; il lavoro degli autori può quindi essere interpretato come una riduzione generalizzata dall'ottimizzazione lineare offline a quella online. Un elemento chiave di questo algoritmo è la nozione di *barycentric spanner*, un tipo speciale di base per lo spazio vettoriale che consente a qualsiasi strategia possibile di essere espressa come una combinazione lineare di vettori di base utilizzando coefficienti limitati. Inoltre è presentato anche un secondo algoritmo per il problema del percorso più breve (online), che risolve il problema utilizzando una catena di oracoli decisionali (online), uno su ciascun nodo del grafico. Ciò presenta numerosi vantaggi rispetto all'approccio di ottimizzazione lineare online. In primo luogo, è efficace contro un avversario adattivo, mentre l'algoritmo sviluppato di ottimizzazione lineare assume un avversario inconsapevole. In secondo luogo, anche nel caso di un avversario inconsapevole, il secondo algoritmo si comporta leggermente meglio del primo, come misurato dal loro regret additivo.

[Azi+18] Aziz, M., Anderton, J., Kaufmann, E., and Aslam, J. *Pure exploration in infinitely-armed bandit models with fixed confidence*. In Proc. ALT 2018, volume 83 of PMLR, pp. 3–24. PMLR, 2018.

Consideriamo il problema dell'identificazione del arm quasi ottimale in *fixed confidence setting* del problema *infinite-armed bandits* quando non si sa nulla sulla distribuzione degli arms. Viene introdotto un framework simile al PAC[2] all'interno del quale

[2] ***Probably Approximately Correct (PAC) Learning*:** nella teoria dell'apprendimento computazionale, l'apprendimento approssimativamente corretto (*PAC*) è un framework per l'analisi matematica del machine learning proposto nel 1984 da Leslie Valiant. In questo framework, il learner riceve campioni e deve selezionare una funzione di generalizzazione (chiamata ipotesi) da una certa classe di possibili funzioni. L'obiettivo è che, con alta probabilità, la funzione

derivare e trasmettere i risultati; hanno derivato un limite inferiore sulla complessità del campione per l'identificazione del arm quasi ottimale; hanno proposto un algoritmo che identifica un arm quasi ottimale con alta probabilità e deriva un limite superiore sulla complessità del campione che è compreso entro un fattore *log* del loro limite inferiore calcolato; hanno discusso se la dipendenza $log^2(\frac{1}{\Delta})$ è inevitabile per gli algoritmi "a due fasi" (prima selezionano gli arms, poi identificano il migliore) nell'impostazione infinita. Questo lavoro consente l'applicazione dei *bandit models* a una classe più ampia di problemi in cui valgono meno ipotesi.

[Bec58] Bechhofer, R. E. *A sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance, and its use with various experimental designs.* In Biometrics, volume 14, pp. 408–429. Wiley International Biometric Society, 1958.

In questo articolo sono presentati diversi risultati importanti per l'applicazione pratica della procedura sequenziale con decisioni multiple che consiste nel selezionare da un gruppo di k aventi distribuzione normale con una varianza sconosciuta quello con la media della popolazione più grande. Sono state fatte anche delle simulazioni di Monte Carlo.

[BF85] Berry, D. and Fristedt, B. *Bandit Problems: Sequential Allocation of Experiments.* Chapman & Hall, 1985.

In questo paper sono stati presentati ulteriori nuovi risultati riguardanti il problema *stochastic multi-armed bandit.* Tuttavia molti risultati non sono stati dimostrati perché semplici da capire oppure solamente attraverso una dimostrazione concettuale.

[Cap+13] Cappè, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. *Kullback-Leibler upper confidence bounds for optimal sequential allocation.* The Annals of Stat., 41(3): 1516–1541, 2013.

Viene considerata l'allocazione sequenziale ottimale nel contesto del modello *stochastic multi-armed bandit.* Viene descritta una *generic index policy*, esposta da *Gittins*[3], basata sui limiti superiori di confidenza dei payoffs degli arms calcolati usando la divergenza di *Kullback-Leibler.* Vengono considerate due classi di distribuzioni per le quali vengono analizzate le istanze: l'algoritmo *kl-UCB* è progettato per famiglie esponenziali a un parametro e l'algoritmo *KL-UCB* empirico per distribuzioni limitate e finite. Il contributo portato dal paper è un'analisi unificata a tempo finito del regret di questi algoritmi che corrisponde asintoticamente ai limiti inferiori di *Lai e Robbins*[4] e di *Burnetas e Katehakis*[5], rispettivamente. Viene studiato anche

selezionata abbia un errore di generalizzazione basso. Il learner deve essere in grado di apprendere il concetto dato qualsiasi rapporto di approssimazione arbitrario, probabilità di successo o distribuzione dei campioni.

[3] Gittins, John C. *«Bandit processes and dynamic allocation indices»* Journal of the Royal Statistical Society: Series B (Methodological) 41.2 (1979): 148-164

[4] Adv. in Appl. Matematica. 6 (1985) 4–22

[5] Adv. in Appl. Matematica. 17 (1996) 122-142

il comportamento di questi algoritmi quando usati con *general bounded rewards*, dimostrando in particolare che forniscono miglioramenti significativi rispetto allo stato dell'arte.

[CV15] Carpentier, A. and Valko, M. *Simple regret for infinitely many armed bandits*. In Proc. ICML 2015, pp. 1133–1141. JMLR, 2015.

Viene considerato il problema *stochastic multi-armed bandit*. In questo contesto, il learner non ha alcuna possibilità di provare tutti gli arms e deve dedicare il suo numero limitato di prove solo a un certo numero di arms. Tutti gli algoritmi precedenti per questa impostazione sono stati progettati minimizzando il *cumulative regret*. In questo documento, viene proposto un algoritmo che mira a minimizzare il *simple regret*. Come nell'impostazione del *cumulative regret* in *infinitely multi-armed bandits*, il tasso del *simple regret* dipenderà dal parametro β che caratterizza la distribuzione degli arms ottimali. Viene dimostrato che, a seconda del β , l'algoritmo proposto è *minimax optimal* a meno di una costante moltiplicativa al massimo di fattore $alog(n)$. Vengono fornite anche estensioni in diversi casi importanti: quando il parametro β è sconosciuto, in un ambiente normale in cui gli arms quasi ottimali hanno una piccola varianza, e nel caso di orizzonte temporale sconosciuto.

[EMM02] Even-Dar, E., Mannor, S., and Mansour, Y. *PAC bounds for multi-armed bandit and Markov Decision Processes*. In Proc. COLT 2002, pp. 255–270. Springer, 2002.

Il problema *stochastic multi-armed bandit* viene rivisitato e considerato nel modello *PAC*. Il principale contributo del paper è mostrare che, dati n arms, è sufficiente tirare le arms $O(\frac{n}{\varepsilon^2}log\frac{1}{\delta})$ volte per trovare un arm ε-ottimale con probabilità di almeno $1-\delta$. Ciò è in contrasto con il limite $O(\frac{n}{\varepsilon^2}log\frac{n}{\delta})$. Viene costruito un altro algoritmo la cui complessità dipende dall'impostazione specifica delle ricompense, piuttosto che dal settaggio del caso peggiore. Viene fornito anche un limite inferiore corrispondente e mostrato come, dato un algoritmo per il problema *multi-armed bandit* del modello *PAC*, si possa derivare un algoritmo di apprendimento batch per i processi decisionali di Markov. Questo viene fatto essenzialmente simulando la *value iteration* del valore e in ogni iterazione viene invocato l'algoritmo *multi-armed bandit*. Usando il nostro algoritmo *PAC* per il problema del *multi-armed bandit*, miglioriamo la dipendenza dal numero di azioni da svolgere a ogni iterazione.

[Gab+11] Gabillon, V., Ghavamzadeh, M., Lazaric, A., and Bubeck, S. *Multi-bandit best arm identification*. In Adv. NIPS 24, pp. 2222–2230. Curran Associates, Inc., 2011.

Viene studiato il problema di identificare l'arm migliore in ciascuno dei bandits in *multi-bandit multi-armed setting*. Per prima cosa viene proposto un algoritmo chiamato *Gap-based Exploration (GapE)* che si concentra sugli arms la cui media è vicina alla media del arms migliore nello stesso bandit (cioè con un piccolo gap). Viene introdotto quindi un algoritmo, chiamato *GapE-V*, che tiene conto della varianza degli arms oltre al loro gap. Viene dimostrato un limite superiore alla probabilità di errore per entrambi gli algoritmi. Poiché *GapE* e *GapE-V* devono ottimizzare un

parametro di esplorazione che dipende dalla complessità del problema, molto spesso sconosciuto in anticipo, vengono introdotte anche variazioni di questi algoritmi che stimano questa complessità online. Infine, vengono valutate le prestazioni di questi algoritmi e confrontate con altre strategie di allocazione su una serie di problemi sintetici.

[Gos+13] Goschin, S., Weinstein, A., Littman, M. L., and Chastain, E. *Planning in reward-rich domains via PAC bandits*. In Proc. EWRL 2012, volume 24, pp. 25–42. JMLR, 2012.

In alcuni ambienti decisionali, le soluzioni di successo sono comuni. Se la valutazione delle soluzioni candidate è molto variabile, la sfida è sapere quando è stata trovata una soluzione "abbastanza buona". Viene formulato questo problema come *infinite-armed bandit* e vengono forniti dei limiti inferiori sul numero di valutazioni o di pull necessari per identificare una soluzione il cui valore superi una determinata soglia r_0. Vengono presentati diversi algoritmi e vengono usati per identificare strategie affidabili per la risoluzione di livelli di videogiochi come *Infinite Mario* e *Pitfall!*. Vengono mostrati i miglioramenti in ordine di grandezza nella complessità del campione su un approccio naturale che tira ogni arm fino a quando non si conosce una buona stima della sua probabilità di successo.

[HPR96] Herschkorn, S. J., Pekoz, E., and Ross, S. M. *Policies without memory for the infinite-armed Bernoulli bandit under the average-reward criterion*. Prob. in the Engg. and Info. Sc., 10(1):21–28, 1996.

Viene considerato il problema *infinite-armed bandit* i cui arms seguono una distribuzione di Bernoulli i cui parametri sconosciuti sono i.i.d. Vengono presentati due policy che massimizzano la ricompensa media quasi certa su un orizzonte infinito. Nessuna delle due policy ritorna mai a un arm precedentemente osservato dopo il passaggio a un nuovo arm o conserva informazioni dalle arms scartate; inoltre le serie di fallimenti indicano la selezione di un nuovo arm. La prima policy è non stazionaria e non richiede informazioni sulla distribuzione del parametro Bernoulli. La seconda policy è stazionaria e richiede solo informazioni parziali; la sua ottimalità è stabilita dalla *renewal theory*.[6] Vengono sviluppate anche policy stazionarie ε-ottimali che non richiedono informazioni sulla distribuzione del parametro sconosciuto e discusse policy stazionarie universalmente ottimali.

[Jam+14] Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. *lil' UCB : An optimal exploration algorithm for multi-armed bandits*. In Proc. COLT 2014, volume 35 of PMLR, pp. 423–439. PMLR, 2014.

Il documento propone un nuovo metodo del limite di confidenza superiore (UCB) per identificare l'arm con la media più grande in un *stochastic multi-armed bandit* con *fixed confidence setting* usando un piccolo numero di campioni rispetto al totale.

[6] **Renewal Theory:** è una branca della teoria della probabilità che generalizza il processo di Poisson per tempi arbitrari. Invece di tempi esponenzialmente distribuiti, un renewal process può avere tempi indipendenti e identicamente distribuiti che hanno una media finita.

Il metodo descritto non può essere migliorato nel senso che il numero di campioni necessari per identificare l'arm migliore rientra in un fattore costante del limite inferiore definito dalla legge del logaritmo iterato *(LIL)*. Ispirati dal *LIL*, vengono costruiti i loro limiti di confidenza dell'algoritmo con orizzonte temporale infinito. Inoltre, utilizzando un nuovo tempo di arresto per l'algoritmo, viene evitato un *union bound* rispetto agli altri arms visti in altri algoritmi di tipo UCB. Viene dimostrato che l'algoritmo è ottimale a meno di costanti e viene dimostrato che anche attraverso simulazioni che fornisce prestazioni superiori rispetto allo stato dell'arte.

[JN14] Jamieson, K. G. and Nowak, R. D. *Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting*. In Proc. 48th Annual Conf. on Information Sciences and Systems (CISS), pp. 1–6. IEEE, 2014.

Questo documento si occupa di identificare l'arm con la media più alta in un problema *stochastic multi-armed bandit* usando il minor numero possibile di campioni indipendenti dagli arms. Mentre il cosiddetto *best arm problem* risale agli anni '50, solo di recente sono stati proposti due algoritmi qualitativamente diversi che raggiungono la complessità ottimale del campione per il problema. Questo documento esamina questi recenti progressi e mostra che la maggior parte degli algoritmi *best-arm* possono essere descritti come varianti dei due recenti algoritmi ottimali. Per ogni tipo di algoritmo vengono considerate un'istanza specifica per analizzare sia teoricamente che empiricamente esponendo in tal modo i componenti principali dell'analisi teorica di questi algoritmi e l'intuizione su come funzionano gli algoritmi nella pratica. I limiti di complessità del campione derivato sono nuovi e in alcuni casi migliorano rispetto ai limiti precedenti. Inoltre, viene confrontato empiricamente una varietà di algoritmi all'avanguardia attraverso simulazioni per il problema del *best arm problem*.

[JHR16] Jamieson, K. G., Haas, D., and Recht, B. *On the detection of mixture distributions with applications to the most biased coin problem*. CoRR, abs/1603.08037, 2016.

Questo documento studia il trade-off tra due diversi tipi di *pure exploration*: ampiezza e profondità. Il problema *most biased coin* richiede quante lanci di monete totali sono necessari per identificare una moneta *heavy* da una borsa infinita contenente sia le monete *heavy* con media $\theta_1 \in (0, 1)$ e monete *light* con media $\theta_0 \in (0, \theta_1)$, dove vengono estratte monete *heavy* dal sacchetto con probabilità $\alpha \in (0, \frac{1}{2})$. La difficoltà principale di questo problema sta nel distinguere se i due tipi di monete hanno medie molto simili o se le monete *heavy* sono estremamente rare. Questo problema ha applicazioni nel crowdsourcing, nel rilevamento di anomalie e nella ricerca dello spettro radio. Vengono quindi costruiti algoritmi adattativi alla conoscenza parziale o assente dei parametri del problema. Inoltre, le tecniche sviluppate generalizzano anche per casi più generali di *infinite-armed bandit*. Vengono anche dimostrati i limiti inferiori che mostrano che i limiti superiori del nostro algoritmo sono strettamente compresi a meno di fattori logaritmici e sulla strada caratterizzata dalla complessità del campione che varia tra una singola distribuzione parametrica e un mix di tali due

distribuzioni. Di conseguenza, questi limiti hanno implicazioni sorprendenti sia per le soluzioni al problema *most biased coin* che per il rilevamento di anomalie quando sono note solo informazioni parziali sui parametri.

[Kal11] Kalyanakrishnan, S. *Learning Methods for Sequential Decision Making with Imperfect Representations*. PhD thesis, The University of Texas at Austin, 2011.

Questa tesi di dottorato fornisce contributi filosofici, analitici e metodologici allo sviluppo di metodi di apprendimento robusti e automatizzati per il processo decisionale sequenziale con rappresentazioni imperfette.

[KS10] Kalyanakrishnan, S. and Stone, P. *Efficient selection of multiple bandit arms: Theory and practice*. In Proc. ICML 2010, pp. 511–518. Omnipress, 2010.

Viene considerato il problema generale, ampiamente applicabile, di selezionare da n variabili casuali a valore reale un sottoinsieme di quelle con media più alta, sulla base del minor numero possibile di campioni. Questo problema, che denotiamo *Explore-m*, è un aspetto fondamentale di numerosi algoritmi di ottimizzazione stocastica e applicazioni in simulazione e ingegneria industriale. La base teorica per il nostro lavoro è un'estensione di una formulazione precedente che utilizza i *multi-armed bandit* che si dedica all'identificazione delle migliori variabili casuali (*Explore-1*). Oltre a fornire limiti *PAC* per il caso generale, adattiamo il nostro approccio teorico per lavorare in modo efficiente nella pratica. Confronti empirici del relativo algoritmo di campionamento rispetto ad altre strategie di selezione di sottogruppi presenti allo stato dell'arte dimostrano significativi guadagni nell'efficienza del campionamento.

[Kal+12] Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. *PAC subset selection in stochastic multi-armed bandits*. In Proc. ICML 2012, pp. 655–662. Omnipress, 2012.

Viene considerato il problema di selezionare, tra gli arms di *n-armed bandit*, un sottoinsieme di dimensione m di arms con le più alte ricompense attese, basate sul campionamento efficiente delle arms. Questo problema di *sub-set selection* trova applicazione in diverse aree. Nel lavoro precedente degli autori [KS10], questo problema è inquadrato in un contesto *PAC* (indicata con "*Explore-m*") e vengono analizzati gli algoritmi di campionamento corrispondenti. Mentre l'analisi formale al suo interno è limitata alla complessità del caso peggiore degli algoritmi, in questo documento, viene progettato e analizzato un algoritmo *(LUCB)* con una migliore complessità del campionamento previsto. È interessante notare che *LUCB* assomiglia molto al noto algoritmo *UCB* per minimizzare il regret. Il limite di complessità del campione atteso che viene mostrato per *LUCB* è nuovo anche per la selezione a arm singolo (*Explore-1*). Viene dato anche un limite inferiore alla complessità del campione peggiore degli algoritmi *PAC* per *Explore-m*.

[KKS13] Karnin, Z., Koren, T., and Somekh, O. *Almost optimal exploration in multi-armed bandits*. In Proc. ICML 2013, volume 28, pp. 1238–1246. PMLR, 2013.

In questo paper viene studiato il problema dell'esplorazione nei *stochastic multi-armed bandit*s. Anche nella più semplice impostazione dell'identificazione dell'arm migliore, rimane un gap moltiplicativo logaritmico tra i limiti inferiore e superiore noti per il numero di pull del arm richiesti per il task. Questo ulteriore fattore logaritmico è abbastanza significativo nelle applicazioni su larga scala al giorno d'oggi. Vengono presentati due nuovi algoritmi privi di parametri per identificare l'arm migliore, in due diverse impostazioni: data una *target confidence* e dato un *budget of arm pulls*, per il quale vengono dimostrati i limiti superiori il cui gap dal limite inferiore è solo doppiamente logaritmico nei parametri del problema. Vengono confermati i loro risultati teorici con esperimenti che dimostrano che il nostro algoritmo supera lo stato dell'arte e si adatta meglio all'aumentare della dimensione del problema.

[KK13] Kaufmann, E. and Kalyanakrishnan, S. *Information complexity in bandit subset selection.* In Proc. COLT 2013, volume 30, pp. 228–251. JMLR, 2013.

Viene considerato il problema di esplorare efficacemente gli arms di un *stochastic multi-armed bandit*s per identificare il sottoinsieme migliore data una dimensione specificata. In base al *PAC* e alle formulazioni *fixed-budget*, otteniamo limiti migliorati utilizzando intervalli di confidenza basati sulla divergenza KL[7]. Mentre l'applicazione di un'idea simile nel contesto del regret ha prodotto dei limiti in termini di divergenza KL tra gli arms, i limiti trovati nel contesto dell'esplorazione pura implicano la *Chernoff information*[8] tra gli arms. Oltre a introdurre questa nuova quantità nella letteratura riguardante i bandits, gli autori hanno contribuito anche a un confronto tra strategie basate su campionamenti uniformi e adattivi per problemi di pura esplorazione, trovando prove a favore di questi ultimi.

[Kle05] Kleinberg, R. *Nearly tight bounds for the continuum-armed bandit problem.* In Adv. NIPS 17, pp. 697–704. MIT Press, 2005.

Nel problema dei *stochastic multi-armed bandit*s, un algoritmo online deve scegliere da una serie di strategie in una sequenza di prove in modo da ridurre al minimo il costo totale delle strategie scelte. Mentre i limiti superiore e inferiore sono noti nel caso in cui il set di strategie sia finito, non è noto quando esiste un set di strategie infinito. Qui viene considerato il caso in cui l'insieme di strategie è un sottoinsieme di $\mathbb{R}^d$ e le funzioni di costo sono continue. Nel caso $d = 1$, hanno migliorato sui limiti superiore e inferiore noti, riducendo il gap a un fattore sub logaritmico. Considerano anche il caso $d > 1$ e le funzioni di costo sono convesse, adattando un algoritmo di ottimizzazione convessa online di *Zinkevich* al modello *sparser feedback* del problema dei *multi-armed bandit*s.

[7] **Divergenza di Kullback–Leibler:** in teoria della probabilità, è una misura non simmetrica della differenza tra due distribuzioni di probabilità P e Q.

[8] **Chernoff Information:** da' limiti esponenzialmente decrescenti sulle distribuzioni di coda di somme di variabili casuali indipendenti

[Li+10] Li, L., Chu, W., Langford, J., and Schapire, R. E. *A contextual-bandit approach to personalized news article recommendation.* In Proc. WWW, pp. 661–670. ACM, 2010.

La *online recommendation* è una caratteristica importante in molte applicazioni. In pratica, l'interazione tra gli utenti e il sistema di raccomandazione potrebbe essere scarsa, vale a dire che gli utenti non interagiscono sempre con il sistema di segnalazione. Ad esempio, alcuni utenti preferiscono scorrere la raccomandazione anziché fare click sui dettagli. Pertanto, una risposta nulla «0» potrebbe non essere necessariamente una risposta negativa, ma una non risposta. È peggio distinguere queste due situazioni quando si consiglia all'utente un solo elemento alla volta e sono raggiungibili poche ulteriori informazioni. La maggior parte delle strategie di raccomandazione esistenti ignorano la differenza tra mancate risposte e risposte negative. In questo documento, viene proposto un nuovo approccio, denominato *SAOR*, per formulare raccomandazioni online tramite interazioni sparse. *SAOR* utilizza risposte positive e negative per creare il modello delle preferenze dell'utente, ignorando tutte le non risposte. Viene fornita un'analisi del regret di *SAOR*, gli esperimenti su set di dati reali mostrano anche che *SAOR* supera i metodi concorrenti.

[MT03] Mannor, S. and Tsitsiklis, J. N. *The sample complexity of exploration in the multi-armed bandit problem.* JMLR, 5: 623–648, 2004.

Viene considerato il problema dei *stochastic multi-armed bandit* nel modello *PAC*. È stato mostrato da [EMM02] che ha dati n arms, un totale di $O(\frac{n}{\varepsilon^2} log\frac{1}{\delta})$ è sufficiente per trovare un arm ε-ottimale con probabilità almeno $1-\delta$. Viene mostrato un limite inferiore corrispondente al numero previsto di prove nell'ambito di qualsiasi politica di campionamento. Viene generalizzato inoltre il limite inferiore e mostriamo una dipendenza esplicita dalle statistiche (sconosciute) degli arms. Viene fornito anche un limite simile all'interno di un ambiente bayesiano. Viene anche discusso il caso in cui le statistiche sugli arms sono note ma non le identità degli arms. Per questo caso, viene fornito un limite inferiore di $\Theta(\frac{1}{\varepsilon^2}(n + log\frac{1}{\delta}))$ sul numero previsto di prove, nonché una politica di campionamento con un limite superiore corrispondente. Se invece del numero previsto di prove, viene considerato il numero massimo (su tutti i percorsi di campionamento) di prove, viene stabilito un limite superiore e inferiore corrispondente della forma $\Theta(\frac{1}{\varepsilon^2} log\frac{1}{\delta})$. Infine, vengono derivati i limiti inferiori sul regret atteso, come avevano fatto *Lai* e *Robbins*.

[Mou+16] Mousavi, S. H., Haghighat, J., Hamouda, W., and Dastbasteh, R. *Analysis of a subset selection scheme for wireless sensor networks in time-varying fading channels.* IEEE Trans. Signal Process., 64(9):2193–2208, 2016.

Una delle principali sfide che affrontano le reti di sensori wireless (*WSN*) sono le limitate risorse di energia disponibili nei piccoli nodi dei sensori. Si desidera pertanto ridurre il consumo di energia dei sensori mantenendo la distorsione tra la sorgente e la sua stima nel centro di fusione (*FC*) al di sotto di una soglia specifica. In questo documento, viene analizzata una strategia di selezione dei sottoinsiemi per ridurre

la potenza di trasmissione media della *WSN*. Si considera una rete a due hop e ipotizzando che i canali tra la sorgente e i *relay sensor* siano *time-varying fading channels*, modellati come canali *Gilbert-Elliott*. Viene mostrato che quando questi canali sono noti all'*FC*, l'*FC* può selezionare un sottoinsieme di sensori per ridurre al minimo la potenza di trasmissione soddisfacendo al contempo il criterio di distorsione. Attraverso l'analisi, viene ricavata la distribuzione di probabilità della dimensione di questo sottoinsieme. Vengono anche considerati aspetti pratici dell'attuazione del metodo proposto, compresa la stima dei canali ai *relay*. Attraverso simulazioni, vengono confrontate le prestazioni dello schema proposto con gli schemi che appaiono in letteratura. I risultati della simulazione confermano che per un certo intervallo di *end-to-end bit-error rates (BERs)*, lo schema proposto riesce a ottenere una riduzione di potenza superiore rispetto ad altri schemi.

[Pau64] Paulson, E. *A sequential procedure for selecting the population with the largest mean from k normal populations*. The Annals of Mathematical Stat., 35(1):174–180, 1964.

In questo articolo vengono fornite procedure sequenziali per selezionare la popolazione normale con la media maggiore quando (1) le popolazioni k hanno una varianza nota comune o (2) le popolazioni k hanno una varianza comune ma sconosciuta, in modo che in ogni caso la probabilità di effettuare la selezione corretta supera un valore specificato quando la media massima supera tutte le altre medie di almeno un valore specificato. Le procedure presenti nel documento hanno tutte la proprietà che le popolazioni con media bassa possono essere eliminate da ulteriori considerazioni man mano che l'esperimento procede.

[RLS19] Ren, W., Liu, J., and Shroff, N. B. *Exploring k out of top fraction of arms in stochastic bandits*. CoRR, abs/1810.11857, 2018.

Questo documento studia il problema dell'identificazione di qualsiasi k arms distinti tra la frazione ρ superiore (ad esempio, il 5% superiore) di arms da un insieme finito o infinito con una tolleranza probabilmente approssimativamente corretta (PAC) ϵ. Vengono considerati due casi: (1) quando è nota la soglia dei premi previsti per le migliori arms e (2) quando è sconosciuta. Vengono dimostrati i limiti inferiori per le quattro varianti (*finite-arms* o *infinite-arms* e soglia nota o sconosciuta) e vengono proposti alcuni algoritmi per ciascuna variante. Due di questi algoritmi si dimostrano ottimali per la complessità del campione (a meno di fattori costanti) e gli altri due sono ottimali a meno di un fattore *log*. I risultati in questo documento forniscono riduzioni fino a $\frac{\rho n}{k}$ rispetto agli algoritmi di *k-exploration* che si concentrano sulla ricerca dei migliori k arms (PAC) da n arms. Vengono mostrati numericamente miglioramenti rispetto allo stato dell'arte.

[Rob52] Robbins, H. *Some aspects of the sequential design of experiments*. Bulletin of the AMS, 58(5):527–535, 1952.

È un articolo che non ha ricevuto molte attenzioni al tempo della stesura ma è stato rivalutato negli ultimi anni e viene spesso citato. Studia la teoria dell'analisi sequenziale, soffermandosi anche sulla numerosità campionaria che deve essere raggiunta.

Mostra quindi alcuni semplici problemi modellizzandoli con la tecnica del design sequenziale.

[CK17] Roy Chaudhuri, A. and Kalyanakrishnan, S. *PAC identification of a bandit arm relative to a reward quantile*. In Proc. AAAI 2017, pp. 1977–1985. AAAI Press, 2017.

Viene proposta una formulazione PAC per identificare un arm in un *n-armed bandit*s la cui media rientra in una *fixed tolerance of the m-th highest mean*. Questo setup generalizza una formulazione precedente con $m = 1$ e differisce da un'altra ancora che richiede l'identificazione di tali arms. L'implicazione chiave dell'approccio proposto è la capacità di derivare limiti superiori dalla complessità del campione che dipendono da $\frac{n}{m}$ al posto di n. Di conseguenza, anche quando il numero di arms è infinito, si ha solo bisogno di un numero finito di campioni per identificare un arm che si confronta favorevolmente con un quantile di ricompensa fisso. Questa funzione rende l'approccio presentato attraente per applicazioni come la scoperta di farmaci, in cui il numero di arm (configurazioni molecolari) può incorrere in alcune migliaia. Sono presentati algoritmi di campionamento sia per i casi finiti che per i casi infiniti e ne viene convalidata l'efficienza attraverso l'analisi teorica e sperimentale. Sono presentati anche un limite inferiore alla peggiore complessità del campione di algoritmi *PAC* per il loro problema, che corrisponde al loro limite superiore a meno di un fattore logaritmico.

[Tra+14] Tran-Thanh, L., Stein, S., Rogers, A., and Jennings, N. R. *Efficient crowdsourcing of unknown experts using bounded multi-armed bandits*. Artif. Intl., 214:89 – 111, 2014.

Sempre più organizzazioni esternalizzano in modo flessibile il lavoro su base temporanea a un pubblico globale di lavoratori. Il crowdsourcing è stato applicato con successo a una serie di lavori, dalla traduzione di testi e annotazioni di immagini, alla raccolta di informazioni durante le situazioni di crisi e all'assunzione di lavoratori qualificati per creare software complessi. Mentre tradizionalmente questi compiti sono stati piccoli e potrebbero essere completati da non professionisti, le organizzazioni stanno ora iniziando a fare crowdsourcing di compiti più grandi e più complessi agli esperti nei loro rispettivi campi. Queste attività includono, ad esempio, lo sviluppo e test del software, web design e marketing del prodotto. Mentre questo crowdsourcing di esperti emergenti offre flessibilità e costi potenzialmente inferiori, solleva anche nuove sfide, poiché i lavoratori possono essere altamente eterogenei, sia nei costi che nella qualità del lavoro che producono. In particolare, l'utilità di ciascuna attività esternalizzata è incerta e può variare in modo significativo tra lavoratori distinti e persino tra compiti successivi assegnati allo stesso lavoratore. Inoltre, in contesti realistici, i lavoratori hanno limiti alla quantità di lavoro che possono svolgere e il datore di lavoro avrà un budget fisso per i lavoratori paganti. Data questa incertezza e i relativi vincoli, l'obiettivo del datore di lavoro è quello di assegnare compiti ai lavoratori al fine di massimizzare l'utilità complessiva raggiunta. Per formalizzare questo problema di crowdsourcing, viene introdotto un nuovo multi-armed bandit

(*MAB*), il *bounded MAB*. Inoltre, viene sviluppato un algoritmo per risolvere il problema in modo efficiente, chiamato *bounded ε-first*, che procede in due fasi: *exploration* e *exploitation*. Durante l'*exploration*, l'algoritmo usa prima εB del suo budget totale B per apprendere stime delle caratteristiche di qualità dei lavoratori. Quindi, durante l'*exploitation*, utilizza il rimanente $(1-\varepsilon)B$ per massimizzare l'utilità totale in base a tali stime. L'utilizzo di questa tecnica ci consente di ricavare un limite superiore $O(B^{\frac{2}{3}})$ dal suo regret di prestazione (ovvero, la differenza attesa nell'utilità tra l'algoritmo e l'ottimale), il che significa che quando il budget B aumenta, il regret tende a 0. Oltre a questo approccio teorico, l'algoritmo viene applicato ai dati del mondo reale usando *oDesk*, un importante sito di crowdsourcing. Utilizzando i dati di progetti reali, inclusi budget di progetti storici, costi di esperti e valutazioni di qualità, viene dimostrato che l'algoritmo supera i metodi di crowdsourcing esistenti fino al 300%, ottenendo al contempo un massimo ipotetico con informazioni complete.

[WAM09] Wang, Y., Audibert, J.-Y., and Munos, R. *Algorithms for infinitely many-armed bandits*. In Adv. NIPS 21, pp. 1729– 1736. Curran Associates Inc., 2008.

Viene considerato il problema dei *stochastic multi-armed bandit* in cui il numero di arms è maggiore del possibile numero di esperimenti. Viene fatta un'ipotesi stocastica sulla ricompensa media di un nuovo arm selezionato che caratterizza la sua probabilità di essere un arm quasi ottimale. La loro ipotesi è più debole rispetto ad altri paper precedenti presenti in letteratura. Vengono descritti algoritmi basati su limiti di confidenza superiore applicati a un insieme limitato di arms selezionati casualmente e vengono forniti i limiti superiori sul regret atteso risultante. Viene derivato anche un limite inferiore che corrisponde (a meno di fattori logaritmici) al limite superiore in alcuni casi.

[Wil+16] Will, Y., McDuffie, J. E., Olaharski, A. J., and Jeffy, B. D. *Drug Discovery Toxicology: From Target Assessment to Translational Biomarkers*. Wiley, 2016.

È una guida per i professionisti farmaceutici ai problemi e alle pratiche della tossicologia della scoperta di farmaci in cui vengono usati anche alcuni algoritmi per risolvere il problema dei *stochastic multi-armed bandits*.

3.2. Lavori citati nella letteratura

In questa sezione sono analizzati ulteriori paper presenti nella letteratura che sono serviti sia per la comprensione che per l'approfondimento del paper assegnato. Essi sono cercati tramite questi due motori di ricerca:

- *Google Scholar*[9]: motore di ricerca accessibile liberamente che tramite parole chiave specifiche consente di individuare testi della letteratura accademica come articoli sottoposti a revisione paritaria, tesi di laurea e dottorato, libri, prestampe, sommari, recensioni e rapporti tecnici di tutti i settori della ricerca scientifica e tecnologica.
- *Proceedings of Machine Learning Research,* in particolare in riferimento al *Volume 97: International Conference on Machine Learning, 9-15 June 2019, Long Beach, California, USA*[10].

Ulteriori lavori presenti nella letteratura riguardanti il problema *stochastic multi-armed bandit* possono essere trovati anche nei related works del paper *Batched Multi-armed Bandits Problem* di Zijun Gao, Yanjun Han, Zhimei Ren, Zhengqing Zhou che avevo in parte letto e analizzato durante la sessione estiva.

[Gao+19] Gao, Z., Han, Y., Ren, Z., & Zhou, Z. (2019). *Batched multi-armed bandits problem.* In Advances in Neural Information Processing Systems (pp. 503-513).

In questo paper ci si concentra sul problema del *multi-armed bandit* a impostazione batched detto *batched multi-armed bandit* (*BMaB*), in cui i dati vengono suddivisi in un piccolo numero di batches. La motivazione alla base di tale studio è che mentre il *minimax regret* per il problema del *two-armed stochastic bandit* è stato caratterizzato a pieno in *Batched bandit problems*[11] da Perchet, Rigollet, Chassang e Snowberg, l'effetto del numero degli arms nel regret nel caso multi-armed è ancora un argomento aperto. Inoltre, rimane ancora inesplorata la domanda se le dimensioni dei batch scelti in modo adattivo aiutano o meno a ridurre il regret. Nel documento si propone la policy *Batched Successive Elimination* (*BaSE*) per ottenere *rate-optimal regrets* (a meno di fattori logaritmici) per il *batched multi-armed bandit*, con matching lower bounds anche se le dimensioni dei batches sono determinate in modo adattivo.

Altri lavori correlati presenti in letteratura che hanno aiutato alla comprensione del problema sono i seguenti:

[Aga+17] Agarwal, A., Agarwal, S., Assadi, S., & Khanna, S. (2017, June). *Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons.* In Conference on Learning Theory (pp. 39-75).

[9] **Google Scholar** (https://scholar.google.it/)
[10] **Proceedings of Machine Learning Research v97** (http://proceedings.mlr.press/v97/)
[11] Perchet, V., Rigollet, P., Chassang, S., & Snowberg, E. (2016). *Batched bandit problems.* The Annals of Statistics, 44(2), 660-681.

Viene analizzata la relazione fra la complessità e l'adattabilità nel richiedere attivamente nuovi dati per identificare le k monete che hanno una probabilità maggiore di ottenere il risultato desiderato (testa), in un insieme di n monete. Successivamente si passa poi a considerare il problema delle migliori k arms nel problema *multi-armed bandit* e al problema dell'ordinamento con confronti a coppie dei primi k oggetti in un insieme finito.

[ACF02] Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). *Finite-time analysis of the multiarmed bandit problem.* Machine learning, 47(2-3), 235-256.

Viene studiato il problema *multi-armed bandit* focalizzandosi sulla variabile tempo e quindi sul numero di turni. Viene dimostrato che il regret ottimo cresce almeno logaritmicamente non solo in relazione al numero di turni ma anche uniformemente in base al tempo con semplici ed efficienti politiche e per tutte le distribuzioni del reward con supporto limitato (cioè l'insieme dei punti in cui la distribuzione non è una funzione liscia, che vuol dire derivabile infinite volte in quel punto).

[AMS09] Audibert, J. Y., Munos, R., & Szepesvári, C. (2009). *Exploration–exploitation tradeoff using variance estimates in multi-armed bandits.* Theoretical Computer Science, 410(19), 1876-1902.

Viene studiato il trade-off tra *exploration* e *exploitation* in una variante dell'algoritmo base per il problema *multi-armed bandit* usando la varianza empirica degli arms. Viene discusso di come l'upper bound per il regret sia logaritmico e che possa quindi non essere adatto a problemi reali con decisori avversi al rischio.

[SC12] Sébastien, B., & Cesa-Bianchi, N. (2012). *Regret analysis of stochastic and non stochastic multi-armed bandit problems.* Foundations and Trends in Machine Learning, 5(1), 1-122.

Viene analizzato il regret nei problemi *stochastic multi-armed bandits* evidenziando le differenze fra *exploration* e *exploitation.* In particolare, confronta due casi estremi: ricompense indipendenti e identicamente distribuite e il caso in cui non siano indipendenti. Analizza alcune varianti ed estensioni, come il *contextual bandit model,* dove ad ogni turno il giocatore può scegliere solo un sottoinsieme delle possibili scelte.

[BPR13] Bubeck, S., Perchet, V., & Rigollet, P. (2013, June). *Bounded regret in stochastic multi-armed bandits.* In Conference on Learning Theory (pp. 122-134).

Viene studiato il problema *stochastic multi-armed bandit* nel caso in cui si conosca quale sia il valore della scelta ottima e il lower bound sulla più piccola differenza fra valore di una scelta e scelta ottima. Propone quindi una politica di randomizzazione che porta a un regret uniformemente limitato nel tempo, e mostra diversi lower bound, dimostrando che conoscere solo una delle due ipotesi sopra dichiarate renda impossibile ottenere bound più bassi.

[EMM06] Even-Dar, E., Mannor, S., & Mansour, Y. (2006). *Action elimination and stopping conditions for the multi-armed bandit and reinfor-*

***cement learning problems*. Journal of machine learning research, 7(Jun), 1079-1105.**

Mostra gli intervalli di confidenza nel problema dei banditi dimostrando quale sia il numero minimo di scelte per trovare il braccio ottimale con una probabilità definita. Propone un framework che cerca di eliminare le azioni che non sono ottime con alta probabilità. Inoltre mostra una variante basata su un modello e una senza modello per il metodo di eliminazione, derivando anche le condizioni di stop che garantiscono che la politica imparata sia approssimativamente ottima con alta probabilità.

[PR+13] Perchet, V., & Rigollet, P. (2013). *The multi-armed bandit problem with covariates*. The Annals of Statistics, 41(2), 693-721.

Viene considerato il problema *stochastic multi-armed bandit* dove ogni arm presenta una ricompensa con rumore che dipende da una variabile casuale. Al contrario del problema classico, questa variante permette cambiamenti dinamici dei reward che descrivono meglio scenari in cui l'informazione non è certa. Utilizza un modello non parametrico e introduce una politica chiamata *Adaptively Binned Successive Elimination (ABSE)* che decompone adattativamente il problema in più piccoli problemi dei *bandits* di tipo statico.

Altri lavori analizzati nella ricerca dei possibili scenari applicativi del multi *armed bandit* nella vita reale sono i seguenti:

[Dur+18] Durand, A., Achilleos, C., Iacovides, D., Strati, K., Mitsis, G. D., & Pineau, J. (2018, November). *Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis*. In Machine Learning for Healthcare Conference (pp. 67-82).

Si vuole progettare una strategia adattiva di allocazione per migliorare l'efficienza della raccolta dati allocando più campioni per esplorare trattamenti promettenti. Questa applicazione è stata vista come un *contextual bandit problem* e hanno introdotto un algoritmo pratico per il trade-off tra *exploration* e *exploitation*.

[She+15] Shen, W., Wang, J., Jiang, Y. G., & Zha, H. (2015, June). *Portfolio choices with orthogonal bandit learning*. In Twenty-fourth international joint conference on artificial intelligence.

Nell'ambito della gestione del portfolio finanziario (collezione di investimenti) di una compagnia di investimenti per massimizzare il reward cumulativo, gli autori hanno proposto un algoritmo bandit per fare scelte online sul portfolio sfruttando le correlazioni tra arms multiple. Costruendo portafogli ortogonali da più assets e integrando il loro approccio col *upper-confidence-bound bandit framework,* gli autori hanno ottenuto la strategia ottimale del portfolio rappresentata da una combinazione di investimenti passivi e attivi in accordo con la *risk-adjusted reward function*.

[BHF10] Brochu, E., Hoffman, M. W., & de Freitas, N. (2010). *Portfolio allocation for Bayesian optimization*. arXiv preprint arXiv:1009.5419.

Ulteriori esempi riguardanti la gestione del portfolio finanziario.

[Tro+15] Trovo, F., Paladino, S., Restelli, M., & Gatti, N. (2015). *Multi–armed bandit for pricing*. In Proceedings of the European Workshop on Reinforcement Learning (EWRL).

Viene proposta una modifica dell'algoritmo upper confidence bound (*UCB*) per i bandit sfruttando due peculiarità del *pricing*: (1) come il prezzo di vendita sale, è razionale assumere che la probabilità di vendere un oggetto diminuisce; (2) dato che solitamente le persone comparano prezzi da differenti venditori e tengono traccia dei cambiamenti di prezzo nel tempo prima di acquistare (specialmente per acquisti online, esistono anche vari tools che permettono di tenere traccia automatica dei prezzi di prodotti online), il numero di volte che un certo tipo di oggetto è acquistato è solo una piccola frazione del numero di volte che il prezzo è visualizzato da acquirenti possibili. Facendo leva su queste assunzioni, gli autori migliorano la *concentration inequality* usata nell'algoritmo *UCB1,* che porta ad avere risultati migliori rispetto all'originale, specialmente nei primi passi del learning dove solo pochi campioni sono disponibili. Gli autori hanno presentato un'evidenza empirica sull'efficacia delle variazioni proposte in termini di aumento della velocità del processo di learning dell'algoritmo *UCB1* nelle applicazioni di *pricing*.

[Zho+17] Zhou, Q., Zhang, X., Xu, J., & Liang, B. (2017, November). *Large-scale bandit approaches for recommender systems*. In International Conference on Neural Information Processing (pp. 811-821). Springer, Cham.

I sistemi di raccomandazione sono frequentemente usati in varie applicazioni per predire le preferenze degli utenti. In questi sistemi vi è il dilemma *exploration-exploitation* dal momento che vi è bisogno di avere conoscenza rispetto agli oggetti precedentemente selezionati (oggetti interessanti per gli utenti) ed esplorare nuovi oggetti che magari possono piacere agli utenti. Questo tipo di problema è stato affrontato dagli autori che soprattutto per dei sistemi di raccomandazione su larga scala che hanno tanti o addirittura infiniti elementi.

[Vas+17] Vaswani, S., Kveton, B., Wen, Z., Ghavamzadeh, M., Lakshmanan, L., & Schmidt, M. (2017). *Model-independent online learning for influence maximization*. arXiv preprint arXiv:1703.00557.

Gli autori considerano la massimizzazione dell'influenza (*IM*) nei social networks, in cui il problema è massimizzare il numero di utenti che diventano interessati a un prodotto selezionando un set di *seed users* per esporre il prodotto. Gli autori hanno proposto un nuova parametrizzazione che rende il framework indipendente dal sottostante modello di diffusione ma anche staticamente efficiente per imparare dai dati.

[LPB17] Losada, D. E., Parapar, J., & Barreiro, A. (2017). *Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems*. Information Processing & Management, 53(5), 1005-1025.

Gli autori hanno modellato il processo iterativo di selezione e recupero dell'informazione come un *contextual bandit problem*. In particolare hanno prodotto numerosi metodi per la *document adjudication* (cinque stazionari e due non stazionari) per effettuare un confronto fra loro per la valutazione *pooling-based* e *metasearch*.

[Liu+17] Liu, B., Yu, T., Lane, I., & Mengshoel, O. J. (2017). *Customized nonlinear bandits for online response selection in neural conversation models*. arXiv preprint arXiv:1711.08493.

Nell'ambito della *dialogue response selection*, gli autori hanno proposto un *contextual multi-armed bandit model* con un *non-linear reward function* che usa una rappresentazione distribuita di testo per una selezione online della risposta.

[Sil+18] Silander, T. (2018). *Contextual memory bandit for pro-active dialog engagement*.

Nell'ambito della *dialogue response selection*, l'autore ha proposto un *contextual multi-armed bandit model* per il problema della massimizzazione del reward con feedback parziale.

[Upa+19] Upadhyay, S., Agarwal, M., Bounneffouf, D., & Khazaeni, Y. (2019). *A Bandit Approach to Posterior Dialog Orchestration Under a Budget*. arXiv preprint arXiv:1906.09384.

Nell'ambito dei *multi-domain dialogue systems*, gli autori hanno studiato il compito dell'*online posterior dialogue orchestration* che hanno definito come il compito di seleziona un subset di skills che più appropriatamente rispondono agli input dell'utente e delle skills individuali.

[Bol+18] Boldrini, S., De Nardis, L., Caso, G., Le, M. T., Fiorina, J., & Di Benedetto, M. G. (2018). *muMAB: A multi-armed bandit model for wireless network selection*. Algorithms, 11(2), 13.

Nell'ambito delle telecomunicazioni gli autori hanno usato un modello *multi-armed* per descrivere il problema della selezione del miglior rete wireless eseguita da un dispositivo *multi-Radio Access Technology (multi-RAT)* con il fine di massimizzare la qualità percepita dall'utente finale.

[DLL19] Ding, K., Li, J., & Liu, H. (2019, January). *Interactive anomaly detection on attributed networks*. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (pp. 357-365).

Nell'ambito dell'identificazione delle anomalie gli autori hanno sviluppato un *collaborative contextual bandit algorithm* che modella gli attributi e le dipendenze dei nodi e gestisce il dilemma *exploration-exploitation* quando si investigano anomalie di differenti tipi.

Altri lavori analizzati nella ricerca dei possibili scenari applicativi del multi *armed bandit* nel *machine learning* sono i seguenti:

[GS10] Gagliolo, M., & Schmidhuber, J. (2010, January). *Algorithm selection as a bandit problem with unbounded losses*. In International Conference on Learning and Intelligent Optimization (pp. 82-96). Springer, Berlin, Heidelberg.

L'*agorithm selection* solitamente era basato su modelli di performance degli algoritmi costruiti con sequenze di training offline; attraverso un approccio online si può iterativamente aggiornare il modello e usarlo per la selezione di una sequenza di istanze del problema analizzato. Il trade-off *exploration-exploitation* è rappresentato da un bandit usando informazioni parziali e un limite sconosciuto sulle perdite.

[Li+17] Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). *Hyperband: A novel bandit-based approach to hyperparameter optimization*. The Journal of Machine Learning Research, 18(1), 6765-6816.

L'ottimizzazione degli iperparametri è fondamentale per le performance degli algoritmi di machine learning. Gli autori hanno formulato un'ottimizzazione degli iperparametri come un infinite-armed bandit di pura esplorazione non stocastico dove le risorse predefinite come le iterazioni, i campioni di dati e le features sono allocati come configurazione di campioni random.

[Bou+17] Bouneffouf, D., Rish, I., Cecchi, G. A., & Féraud, R. (2017). *Context attentive bandits: Contextual bandit with restricted context*. arXiv preprint arXiv:1705.03821.

Nell'ambito della *feature selection*, gli autori hanno affrontato il problema dell'*online feature selection* associando il problema dell'ottimizzazione combinatoria in un bandit stocastico con bandit feedback usando l'algoritmo *Active Thompson Sampling (ATS).*

[Bou+14] Bouneffouf, D., Laroche, R., Urvoy, T., Féraud, R., & Allesiardo, R. (2014, November). *Contextual bandit for active learning: Active thompson sampling*. In International Conference on Neural Information Processing (pp. 405-412). Springer, Cham.

Nell'ambito dei bandits per *active learning*, etichettare tutti gli esempi di training nella classificaizone supervisionata può essere complicato. Le strategie di *active learning* risolvono questo problema selezionando solo i campioni non etichettati più utili per ottenere la label e per allenare il modello predittivo. La scelta dei campioni da etichettare può essere vista come un dilemma tra l'*exploration* e l'*exploitation* su uno spazio di input. Gli autori hanno proposto di modellare questo problema come un *contextual bandit problem* e proposto un algoritmo sequenziale chiamato *Active Thompson Sampling (ATS).*

[SL18] Sublime, J., & Lefebvre, S. (2018, July). *Collaborative clustering through constrained networks using bandit optimization*. In 2018 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

Per quanto riguarda il clustering collaborativo, gli autori hanno proposto un algoritmo collaborativo *peer to peer* per il clustering basato sul principio del *non stochastic multi-arm bandits* per assicurarsi in tempo reale quali algoritmi o views potevano portare informazioni utili.

[Noo+18] Noothigattu, R., Bouneffouf, D., Mattei, N., Chandra, R., Madan, P., Varshney, K., ... & Rossi, F. (2018). *Interpretable multi-objective reinforcement learning through policy orchestration.* arXiv preprint arXiv:1809.08343.

Nell'ambito del *reinforcement learning*, gli agenti devono sviluppare tecniche per massimizzare il loro rewards e seguire delle limitazioni implicite presenti nell'ambiente. Gli autori hanno usato un *contextual bandit-based orchestrator* che prende una scelta appropriata tra due policies: *constraint-based* e *environment reward-based*.

Molti di queste applicazioni pratiche sono state trovate grazie al paper *«A survey on practical applications of multi-armed and contextual bandits»* di Bouneffouf, D., & Rish, I [BR19].

4. Descrizione

Inizialmente verrà definito quale è il problema trattato dal paper e i contributi forniti dagli autori e successivamente mostrati algoritmi sia per istanze finite che infinite.

4.1. Definizione del problema

Sia A un set di arms nell'istanza del bandit data. Ogni arm $a \in A$ ha associata una distribuzione dei reward compresa nell'intervallo $[0, 1]$, con media μ_a. Quando tirato, un arm $a \in A$ produce un reward i.i.d pescato dalla distribuzione corrispondente e indipendente dagli altri pull degli altri arms. A ogni turno, basandosi sulla precedente sequenza di pulls e rewards, un algoritmo decide quale arm tirare, fermarsi oppure ritornare un set di arms.

Per un istanza finita del bandit con n arms, prendiamo $A = \{a_1, a_2, \ldots, a_n\}$, e assumiamo, senza perdita di generalità, che per le arms $a_i, a_j \in A$, vale $\mu_{a_i} \geq \mu_{a_j}$ dove $i \leq j$. Data una tolleranza $\epsilon \in [0, 1]$ e $m \in \{1, 2, \ldots, n\}$, possiamo chiamare un arm $a \in A$ (ϵ, m)-optimal se $\mu_a \geq \mu_{a_m} - \epsilon$. Denotiamo il set di tutte le (ϵ, m)-optimal arms come $\mathcal{TOP}_m(\epsilon) = \{a : \mu_a \geq \mu_{a_m} - \epsilon\}$. Per semplicità denotiamo$\mathcal{TOP}_m(0)$ come $\mathcal{TOP}_m$.

Definizione 4.1. [Problema (k, m, n)] Un'istanza del problema (k, m, n) è nella forma $(A, n, m, k, \epsilon, \delta)$, dove A è il set di arms con $|A| = n \geq 2$; $m \in \{1, 2, \ldots, n-1\}$; $k \in \{1, \ldots, m\}$; tolleranza $\epsilon \in (0, 1]$; e probabilità d'errore $\delta \in (0, 1]$. Un algoritmo $\mathcal{L}$ si dice in grado di risolvere (k, m, n) se per ogni istanza di (k, m, n), l'algoritmo termina con probabilità 1 e restituisce k *distinct* (ϵ, m)-optimal arms con probabilità almeno $1 - \delta$.

Il problema (k, m, n) è interessante da un punto di vista teorico perché copre un intero range di problemi, dall'identificazione di una singola arm $(m = 1)$ fino alla selezione di un sottoinsieme $(k = m)$. Così, ogni bound basato sulla complessità del campione (k, m, n) può essere applicato a *Q-F [CK17]* e *SUBSET [KS10].* In questo paper, verrà mostrato un algoritmo che risolve il problema (k, m, n) in cui valga un limite inferiore di pulls $\Omega\left(\frac{n}{(m-k+1)\epsilon^2} \log\left(\frac{\binom{m}{k-1}}{\delta}\right)\right)$ per qualche istanza del problema. Non si è a conoscenza dei limiti in *fixed confidence setting* che coinvolgono un tale termine combinatorio all'interno del logaritmo.

Nella tabella presente in figura 4.1 sono confrontati i limiti che verranno dimostrati in questo paper rispetto ai limiti stabiliti dei lavori precedenti. I limiti presenti nella tabella sono da considerare come il peggior caso tra tutte le istanze del problema; in pratica uno può sperare di fare meglio su istanze del problema più semplici adottando una strategia di campionamento sequenziale completa. Infatti adattiamo l'algoritmo *LUCB1 [Kal+12]* per risolvere il problema (k, m, n) con un algoritmo adattato chiamato *LUCB-k-m*. La nostra analisi dimostrerà che per $k = 1$ e $k = m$, l'upper bound della complessità del campione di questo algoritmo corrisponde, rispettivamente, a F_2 [CK17] e *LUCB1 [Kal+12]*, a meno di una costante moltiplicativa. Empiricamente, *LUCB-k-m* con $k = 1$ sembra più efficiente diF_2 per risolvere *Q–F*.

Problem	Lower Bound	Previous Upper Bound	Current Upper Bound
$(1, 1, n)$ Best-Arm	$\Omega\left(\frac{n}{\epsilon^2}\log\frac{1}{\delta}\right)$ (Mannor & Tsitsiklis, 2004)	$O\left(\frac{n}{\epsilon^2}\log\frac{1}{\delta}\right)$ (Even-Dar et al., 2002)	Same as previous
(m, m, n) SUBSET	$\Omega\left(\frac{n}{\epsilon^2}\log\frac{m}{\delta}\right)$ (Kalyanakrishnan et al., 2012)	$O\left(\frac{n}{\epsilon^2}\log\frac{m}{\delta}\right)$ (Kalyanakrishnan & Stone, 2010)	Same as previous
$(1, m, n)$ Q-F	$\Omega\left(\frac{n}{m\epsilon^2}\log\frac{1}{\delta}\right)$ (Roy Chaudhuri & Kalyanakrishnan, 2017)	$O\left(\frac{n}{m\epsilon^2}\log^2\frac{1}{\delta}\right)$	$O\left(\frac{1}{\epsilon^2}\left(\frac{n}{m}\log\frac{1}{\delta}+\log^2\frac{1}{\delta}\right)\right)$ **This paper**
(k, m, n) Q-F$_k$	$\Omega\left(\frac{n}{(m-k+1)\epsilon^2}\log\frac{\binom{m}{k-1}}{\delta}\right)$ **This paper**	-	$O\left(\frac{k}{\epsilon^2}\left(\frac{n\log k}{m}\log\frac{k}{\delta}+\log^2\frac{k}{\delta}\right)\right)^*$ **This paper** (*for $k \geq 2$)
$(1, \rho)$ ($\lvert\mathcal{A}\rvert = \infty$) Q-P	$\Omega\left(\frac{1}{\rho\epsilon^2}\log\frac{1}{\delta}\right)$ (Roy Chaudhuri & Kalyanakrishnan, 2017)	$O\left(\frac{1}{\rho\epsilon^2}\log^2\frac{1}{\delta}\right)$	$O\left(\frac{1}{\epsilon^2}\left(\frac{1}{\rho}\log\frac{1}{\delta}+\log^2\frac{1}{\delta}\right)\right)$ **This paper**
(k, ρ) ($\lvert\mathcal{A}\rvert = \infty$) Q-P$_k$	$\Omega\left(\frac{k}{\rho\epsilon^2}\log\frac{k}{\delta}\right)$ **This paper**	-	$O\left(\frac{k}{\epsilon^2}\left(\frac{\log k}{\rho}\log\frac{k}{\delta}+\log^2\frac{k}{\delta}\right)\right)^*$ **This paper** (*for a special class with $k \geq 2$)

Figura 4.1.: Limiti inferiore e superiore sulla complessità del campione attesa (ponendosi nel caso peggiore). I limiti per $(k; \rho)$, $k > 1$ sono per la classe speciale di istanze "al massimo k-equiprobabili".

Allo stesso modo [CK17] definisce il problema *Q-P* per istanze infinite, come una generalizzazione di *Q-P* per seleziona tanti buoni arms che denotiamo (k, ρ). Dato un set di arms A, una distribuzione del campione P_A, $\epsilon \in (0, 1]$, e $\rho \in [0, 1]$, un arm $a \in A$ è chiamata $[\epsilon, \rho]$-optimal se $P_{a' \sim P_A}\{\mu_a \geq \mu_{a'} - \epsilon\} \geq 1 - \rho$. Per $\rho, \epsilon \in [0, 1]$, possiamo definire un set di tutte $[\epsilon, \rho]$-optimal arms come $\mathcal{TOP}_\rho(\epsilon)$. Come prima possiamo definire $\mathcal{TOP}_\rho(0)$ come $\mathcal{TOP}_\rho$. Una semplice generalizzazione di *Q-P* è la seguente.

Definizione 4.2. [Problema (k, ρ)] Un'istanza del problema (k, ρ) è nella forma $(A, P_A, k, \rho, \epsilon, \delta)$, dove A è il set di arms; P_A è la probabilità di distribuzione su A; frazione quantile $\rho \in (0, 1]$; tolleranza $\epsilon \in (0, 1]$; e probabilità d'errore $\delta \in (0, 1]$. Una tale istanza è *valida* $|\mathcal{TOP}_\rho| \geq k$, e *non valida* altrimenti. Dato un algoritmo $\mathcal{L}$ si dice in grado di risolvere il problema (k, ρ) se per ogni istanza *valida* di (k, ρ), termina con probabilità 1 e restituisce k *distinct* $[\epsilon, \rho]$-optimal arms con probabilità di almeno $1 - \delta$.

Al massimo k istanze equiprobabili Si osservi il fatto che il problema (k, ρ) è ben definito soltanto se l'istanza data ha almeno k arms distinte in $\mathcal{TOP}_\rho$; possiamo

definire una tale istanza *valida*. Vale la pena notare che anche istanze valide possono richiedere un tempo di calcolo arbitrario per essere risolte. Per esempio, consideriamo un istanza con $k > 1$ arms in $\mathcal{TOP}_\rho$, una avrà probabilità γ di essere pescata da P_A, e le restanti hanno ognuna probabilità $(\rho - \gamma)/(k-1)$. Dal momento che le arms devono essere identificate campionando da P_A, la probabilità di identificare le ultime $k-1$ arms tende a 0 se $\gamma \to \rho$, chiedendo un numero infinito di tentativi. Per evitare questo scenario, restringiamo la nostra analisi a una classe speciale di istanze valide in cui P_A riserva non più di ρ/k probabilità per ogni arm in $\mathcal{TOP}_\rho$. Ci riferiamo a tali istanze come "al massimo k istanze equiprobabili". Formalmente, un'istanza del problema (k, ρ) data da $(A, P_A, k, \rho, \epsilon, \delta)$ è chiamata "al massimo k-equiprobable" se $\forall a \in \mathcal{TOP}_\subset$, $\Pr_{\mathbf{a}' \sim P_A}\{\mathbf{a}' = a\} \leq \frac{\rho}{k}$.[1]

Esempio 4.1. Per capire la nozione del problema (k, ρ) prendiamo un esempio concreto.

Dato un set di arms $A = [0, 1]$, tale che $\mu_a = a$. Ora per qualche $\gamma \in [0, 0.5]$, su cui definiamo una distribuzione dei campioni $P_A^{(\gamma)}$ su A tale che $\Pr_{a \sim P_A^{(\gamma)}}\{\mu_a \in [0, 0.5]\} = 0.4$, $\Pr_{a \sim P_A^{(\gamma)}}\{\mu_a \in (0.5, 0.95] \cup (0.98, 1)\} = 0$, $\Pr_{a \sim P_A^{(\gamma)}}\{\mu_a (0.95, 0.98]\} = \gamma$, $\Pr_{a \sim P_A^{(\gamma)}}\{\mu_a = 1\} = 0.6 - \gamma$.

$$\Pr_{a \sim P_A^{(\gamma)}}\{0 \leq \mu_a \leq 0.5\} = 0.4, \quad \Pr_{a \sim P_A^{(\gamma)}}\{0.5 < \mu_a \leq 0.95\} = 0,$$
$$\Pr_{a \sim P_A^{(\gamma)}}\{0.95 < \mu_a < 1\} = \gamma, \quad \Pr_{a \sim P_A^{(\gamma)}}\{\mu_a = 1\} = 0.6 - \gamma.$$

Per $\rho = 0.5$, $k = 2$, $\epsilon = 0.1$, e $\delta \in (0, 1)$, possiamo definire un istanza del problema del problema (k, ρ) come $I^{(\gamma)} = (A, P_A^{(\gamma)}, k, \rho, \epsilon, \delta)$. Ora in accordo con la definizione di del problema (k, ρ), l'istanza $I^{(0)}$ è *non valida*; per tutte le $\nu \in (0, 0.1)$ un istanza $I^{(\nu)}$ è *non valida*, ma devono esistere k distinct $[\epsilon, \rho]$-optimal arms; altrimenti, la sua soluzione cesserà di esistere per $\epsilon < 0.02$. Inoltre, tale istanza non è interessante visto che può mancare l'esistenza di una soluzione arbitrariamente approssimata. Nuovamente per ogni $\gamma > 0.1$, $I^{(\gamma)}$ è un'istanza *valida*, inoltre come $\gamma \downarrow 0.1$, l'istanza diventa arbitrariamente più complicata da risolvere $I^{(\gamma)}$.

Notare che ogni istanza del problema $(1, \rho)$ or *Q-P problem* [CK17] è necessariamente *valida* e al massimo 1-equiprobable. Si è migliorato rispetto alla versione esistente dell'upper bound di questo problema e va a corrispondere al lower bound del termine *adattive* $O\left(\frac{1}{\epsilon^2} \log^2 \frac{1}{\delta}\right)$.

[1] In un paper recente, Ren et al. (2018) cerca di risolvere il problema (k,). Tuttavia, loro non hanno notato che il problema può essere mal posto. Inoltre, anche con un'istanza al massimo k-equiprobable come input, il loro algoritmo fallisce nel tentativo di eliminare la dipendenza da $(1/\rho)log^2(1/\delta)$.

4.2. Contributi

Qui di seguito i contributi che gli autori hanno dato grazie a questo paper e che verranno analizzati nel dettaglio nelle prossime sezioni:

1. Hanno generalizzato due problemi precedenti, *Q-F [CK17]* e *SUBSET [KS10]* via (k, m, n). Nella sezione 4.3.1 hanno derivato un lower bound sul caso peggiore della complessità del campione per risolvere (k, m, n) che generalizzava gli esistenti lower bound per i problemi *Q-F* e *SUBSET.*
2. Nella sezione 4.3.2 hanno esteso l'algoritmo $LUCB1$ [Kal+12] per presentare un algoritmo completamente sequenziale (*LUCB for k out of m* or *LUCB-k-m*) per risolvere (k, m, n). Hanno dimostrato che per $k = 1$, e $k = m$ l'upper bound sulla complessità del campione attesa corrisponde a quelli di F_2, e $LUCB1$, rispettivamente, a meno di un fattore moltiplicativo.
3. Nella sezione 4.4 hanno presentato un algoritmo $\mathcal{P}_3$ per risolvere il problema *Q-P* con una complessità del campione che è un termina additivo $O((1/\epsilon^2)\log^2(1/\delta))$ lontano dal lower bound. Hanno esteso ciò a un algoritmo *KQP-1* per risolvere le istanze al massimo k-equiprobable (k, ρ). Inoltre P_3 e *KQP-1* possono risolvere *Q-F* e (k, m, n) rispettivamente e le loro complessità del campione sono sono indipendenti dall'istanza più stretta limiti superiori.
4. Nella sezione 4.4.3 hanno presentato una relazione generale tra l'upper bound sulle complessità dei campioni per risolvere *Q-F* e *Q-P.* Questo aiuta a trasferire qualsiasi miglioramento nell'upper bound dal primo problema a quest'ultimo. Inoltre ipotizziamo l'esistenza di una classe di istanze *Q-F* che possono essere risolte di più in modo efficiente rispetto alle loro istanze *Q-P* "corrispondenti".
5. Nel capitolo 5 , riguardante gli esperimenti, hanno dimostrato empiricamente che *LUCB-k-m* è più efficiente di F_2 per risolvere problemi *Q-F.*

4.3. Algortimi per istanze finite

Prima di passare all'analisi tecnica del paper, viene fornito un lower bound sulla complessità del campione degli algoritmi (k, m, n). Le dimostrazioni complete sono presenti nelle appendici a fine documento.

In figura 4.2 sono rappresentati le possibili istanze del problema (k, m, n).

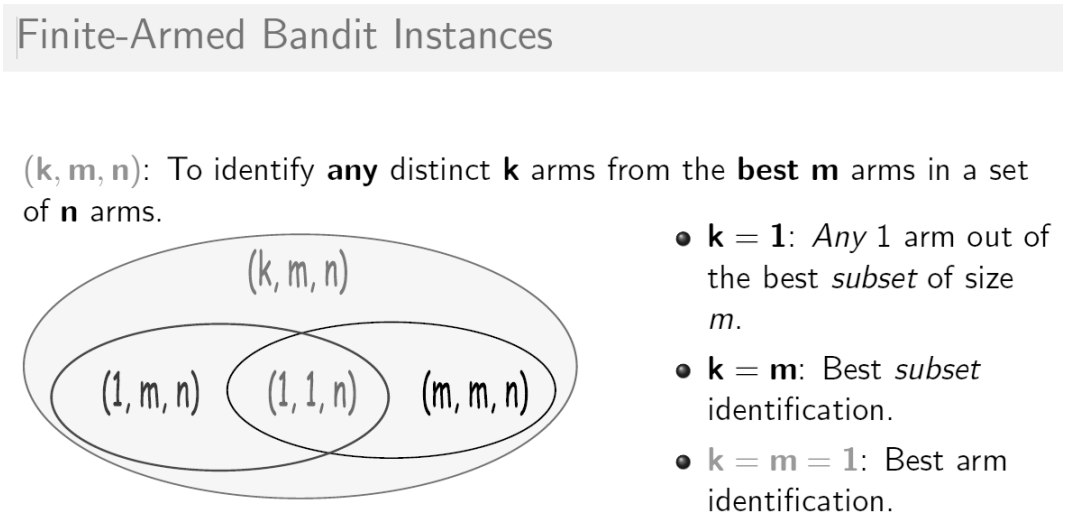

Figura 4.2.: Finite-Armed Bandit Instances

4.3.1. Lower Bound sulla complessità del campione

Teorema 4.1. *[Lower Bound per (k, m, n)] Sia $\mathcal{L}$ un algoritmo che risolve il problema (k, m, n). Allora, esiste un istanza $(A, n, m, k, \epsilon, \delta)$, con $0 < \epsilon \leq \frac{1}{\sqrt{32}}$, $0 < \delta \leq \frac{e^{-1}}{4}$, e $n \geq 2m$, $1 \leq k \leq m$, su cui il numero di pulls atteso fatto dall'algoritmo $\mathcal{L}$ è almeno $\frac{1}{18375} \cdot \frac{1}{\epsilon^2} \cdot \frac{n}{m-k+1} \ln \frac{\binom{m}{k-1}}{4\delta}$.*

La dimostrazione completa è presente nell'appendice A. La dimostrazione generalizza sia il lower bound per (m, m, n) [Kal+12] che $(1, m, n)$ [CK17]. L'idea centrale in queste dimostrazioni è considerare due set di istanze di bandit, $\mathcal{I}$ e $\mathcal{I}'$, tale che su traiettorie verbi, un istanza di $\mathcal{I}$ produrrà la stessa sequenza di reward corrispondente all'istanza presa da $\mathcal{I}'$, con alta probabilità. Inoltre, ogni algoritmo ritornerà lo stesso set di arms per entrambe le istanze, con alta probabilità. Tuttavia, per costruzione, nessun set di arms può essere corretto contemporaneamente per entrambe le istanze, il che implica che un algoritmo corretto deve incontrare traiettorie sufficientemente "lunghe". Il contributo principale degli autori è la progettazione di $\mathcal{I}$ e $\mathcal{I}'$ quando $k \in \{1, 2, \ldots, m\}$ (piuttosto che esattamente 1 or m) arms che devono essere ritornate.

Gli algoritmi proposti per raggiungere upper bounds migliorativi per Q-F e $(k; m, n)$ (across bandit instances) seguono i metodi che gli autori hanno progettato per

infinite-armed setting nella sezione 4.4 (vedi Corollario 4.1 e Corollario 4.2). Nel resto di questa sezione, viene presentato un algoritmo completamente sequenziale per $(k; m, n)$ la cui complessità del campione prevista varia con la complessità dell'istanza in input.

4.3.2. Algoritmo adattivo proposto

L'algoritmo 4.1 descrive *LUCB-k-m*, un algoritmo completamente sequenziale che generalizza *LUCB1* [Kal+12] che risolve (m, m, n), in cui per $k = 1$ ha la stessa complessità del campione di F_2 [CK17], ma empiricamente appare più economico.

Algoritmo 4.1 Algoritmo *LUCB-k-m* per selezionare k (ϵ, m)-optimal arms

Input: $\mathcal{A}$ (tale che $|\mathcal{A}|= n$), k, m, ϵ, δ.
Output: k distinct (ϵ, m)-optimal arms da $\mathcal{A}$.

Tira ogni arm $a \in \mathcal{A}$ una sola volta. Dato $t = n$.
while $ucb(l_*^t, t+1) - lcb(h_*^t, t+1) > \epsilon$. **do**
  $t = t + 1$.
  A_1^t= Set di k arms con la più alta media empirica.
  A_3^t = Set di $n - m$ arms con la media empirica più bassa.
  $A_2^t = \mathcal{A} \setminus (A_1^t \cup A_3^t)$.
  $h_*^t = \arg\max_{\{a \in A_1^t\}} lcb(a, t)$.
  $m_*^t = \arg\min_{\{a \in A_2^t\}} u_a^t$.
  $l_*^t = \arg\max_{\{a \in A_3^t\}} ucb(a, t)$.
  pull h_*^t, m_*^t, l_*^t.
end while
Return A_1^t.

Ad ogni turno t, partizioniamo A in tre subsets. Teniamo le k arms con la più grande media empirica in A_1^t, le $n-m$ arms con la più bassa media empirica in A_3^t, e il resto in A_2^t; i legami sono rotti arbitrariamente (uniformati a random negli esperimenti degli autori). Ad ogni turno si sceglie la *contentious* arm da ciascuno dei tre subset: da A_1^t si sceglie h_*^t, l'arm con il più basso lower confidence bound (LCB); da A_2^t l'arm che è stata tirata meno volte che chiamiamo m_*^t; da A_3^t scegliamo l_*^t, l' arm con la più alta upper confidence bound (UCB). L'algoritmo si ferma non appena la differenza tra il lower confidence bound di h_*^t, e l'upper confidence bound di l_*^t diventa più piccola della tolleranza ϵ.

Sia B_1, B_2, B_3 i corrispondenti sets basati sulla media reale: cioè i subsets di $\mathcal{A}$ tali che $B_1 = \{1, 2, \cdots, k\}$, $B_2 = \{k+1, k+2, \cdots, m\}$ e $B_3 = \{m+1, m+2, \cdots, n\}$. Per ogni due arms $a, b \in \mathcal{A}$ definiamo $\Delta_{ab} = \mu_a - \mu_b$. Per comodità appesantiamo leggermente questa notazione come

$$\Delta_a = \begin{cases} \mu_a - \mu_{m+1} \text{ if } a \in B_1 \\ \mu_k - \mu_{m+1} \text{ if } a \in B_2 \\ \mu_m - \mu_a \quad \text{if } a \in B_3 \end{cases} \tag{4.1}$$

Notiamo che $\Delta_k = \Delta_{k+1} = \cdots = \Delta_m = \Delta_{m+1}$. Sia $u^*(a,t) = \lceil \frac{32}{\max\{\Delta_a, \frac{\epsilon}{2}\}^2} \ln \frac{k_1 n t^4}{\delta} \rceil$ per ogni $a \in \mathcal{A}$, dove $k_1 = 5/4$. Ora definiamo *l'hardness term* come $H_\epsilon = \sum_{a \in \mathcal{A}} \frac{1}{\max\{\Delta_a, \epsilon/2\}^2}$.

Teorema 4.2. *[Complessità del campione attesa di LUCB-k-m] LUCB-k-m risolve il problema (k,m, n) usando un limite superiore alla complessità del campione attesa di* $O\left(H_\epsilon \log \frac{H_\epsilon}{\delta}\right)$.

Nella appendice A è descritta la dimostrazione nel dettaglio. Il punto centrale è simile all'algoritmo F_2 [CK17]. Tuttavia, differisce leggermente a causa della diversa strategia per la scelta delle arms poiché il set di output non è necessariamente singolo. In pratica, si possono usare bound di confidenza più stretti (gli autori usano limiti di confidenza basati sulla divergenza KL nei loro esperimenti) per ottenere una complessità del campione ancora migliore.

Successivamente si andrà a considerare le istanze *infinite-armed bandit* e descrivere algoritmi che le risolvono.

4.4. Algoritmi per istanze infinite

Prima di procedere all'identificazione delle k $[\epsilon, \rho]$-optimal arms in *infinite-armed bandits*, rivisitiamo il caso $k = 1$. Per trovare un $[\epsilon, \rho]$-optimal arm, la complessità del campione di tutti gli algoritmi esistenti [CK17]; [Azi+18] scala a $(1/\rho\epsilon^2)\log^2(1/\delta)$, per la probabilità d'errore data δ. In questa sezione viene presentato un algoritmo P_3 la cui complessità del campione è *additive* poly-log factor lontana dal lower bound di $\Omega((1/\rho\epsilon^2)\log 1/\delta)$ [CK17].

In figura 4.3 sono rappresentati le possibili istanze del problema (k, ρ).

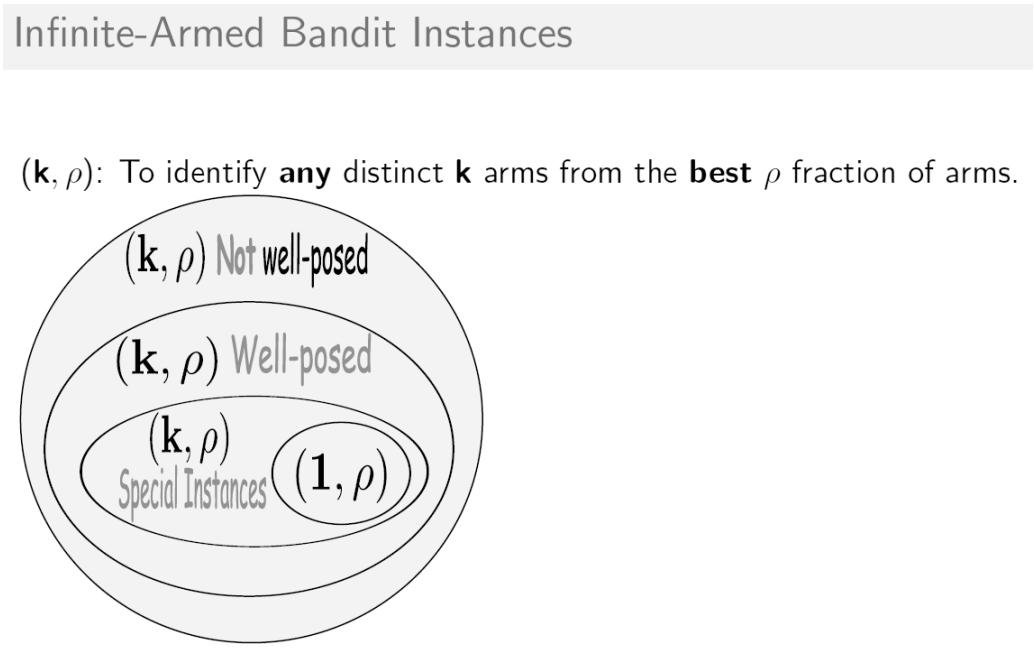

Figura 4.3.: Infinite-Armed Bandit Instances

4.4.1. Risolvere le istanze Q-P

$\mathcal{P}_3$ è un algoritmo a due fasi. Nella prima fase esegue un numero abbastanza grande di copie indipendenti di $\mathcal{P}_2$ e sceglie un subset grande di arms (grande u) in cui ogni arm è $[\epsilon, \rho]$-optimal con probabilità di almeno $1 - \delta'$, dove δ' è una costante piccola. Il valore u è scelto in modo che almeno una delle arms scelte è $[\epsilon/2, \rho]$-optimal con probabilità di almeno $\delta/2$. La seconda fase risolve il problema dell'identificazione dell'arm migliore $(1, 1, u)$ applicando la Median Elimination.

L'algoritmo 4.2 descrive $\mathcal{P}_3$. Usa $\mathcal{P}_2$ [CK17] con Median Elimination come subroutine, per selezionare una $[\epsilon, \rho]$-optimal arm con confidenza $1 - \delta'$. Abbiamo assunto che $\delta' = 1/4$, in pratica si può scegliere qualsiasi valore sufficientemente

piccolo per esso, che influenzerà semplicemente la costante moltiplicativa nel limite superiore.

Algoritmo 4.2 Algoritmo $\mathcal{P}_3$ per l'identificazione dell'arm migliore
Input: $\mathcal{A}, \epsilon, \delta$.
Output: Un $[\epsilon, \rho]$-optimal arm.

Set $\delta' = 1/4$, $u = \lceil \frac{1}{\delta'} \log \frac{2}{\delta} \rceil = \lceil 4 \log \frac{2}{\delta} \rceil$.
Esegue u copie di $\mathcal{P}_2(A, \rho, \epsilon/2, \delta')$ e costruisce il set S con le arms in output.
Identifica una $(\epsilon/2, 1)$-optimal arm in S usando la MEDIAN ELIMINATION con confidenza di almeno $1 - \delta/2$.

Teorema 4.3. *[Correttezza e complessità del campione di $\mathcal{P}_3$] $\mathcal{P}_3$ risolve il problema Q-P, con una complessità del campione $O(\epsilon^{-2}(\rho^{-1} \log(1/\delta) + \log^2(1/\delta)))$.*

Dimostrazione. Innanzitutto dimostriamo la correttezza e poi l'upper bound della complessità del campione.

Correttezza

Dimostrazione. Innanzitutto notiamo che ogni copia degli output di $\mathcal{P}_2$ è un $[\epsilon/2, \rho]$-optimal arm con probabilità di almeno $1 - \delta'$. Ora, $S \cap \mathcal{TOP}_\rho = \emptyset$ può succedere solo se tutte le u copie degli output di $\mathcal{P}_2$ sono sub-optimal arms. Allora, $\Pr\{S \cap \mathcal{TOP}_\rho = \emptyset\} = (1 - \delta')^u \leq \delta/2$. Dall'altra parte la probabilità d'errore della MEDIAN ELIMINATION ha un limite superiore di $\delta/2$. Inoltre, considerando l'union bound, otteniamo che la probabilità di errore sia limitata in alto da δ. Inoltre, la media dell'arm in output non è meno di $\frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$ dal $(1 - \rho)$-esimo quantile. □

Complessità del campione

Dimostrazione. Innanzitutto notiamo che, per qualche costante appropriata C, la *sample complexity* (*SC*) di ciascuna delle u copie di $\mathcal{P}_2$ è $\frac{C}{\rho(\epsilon/2)^2} \left(\ln \frac{2}{\delta'}\right)^2 \in O\left(\frac{1}{\rho\epsilon^2}\right)$. Quindi, SC di tutte le u copie $\mathcal{P}_2$ ha un limite superiore di $\frac{C_1 \cdot u}{\rho\epsilon^2}$, per qualche costante C_1. Inoltre, per qualche costante C_2, la complessità del campione della MEDIAN ELIMINATION ha un limite superiore di $\frac{C_2 \cdot u}{(\epsilon/2)^2} \ln \frac{2}{\delta} \leq \frac{C_3}{\epsilon^2} \ln^2 \frac{2}{\delta}$. Aggiungendo le complessità del campione e sostituendo u si ottiene il bound. □

□

Corollario 4.1. *$\mathcal{P}_3$ può risolvere qualsiasi istanza di Q-F $(A, n, m, \epsilon, \delta)$ con complessità del campione $O\left(\frac{1}{\epsilon^2}\left(\frac{n}{m} \log \frac{1}{\delta} + \log^2 \frac{1}{\delta}\right)\right)$.*

Dimostrazione. Sia $(A, n, m, \epsilon, \delta)$ la data istanza di *Q-F*. Partizioniamo il set $A^\infty = [0, 1]$ in n segmenti uguali e associamo a ciascuno un'arm unica in A, in modo tale che due differenti subsets non possono essere associati alla stessa arm. Ora, definiamo $P_{A^\infty} = Uniform[0, 1]$, e $\rho' = m/n$, realizziamo che risolvere l'istanza *Q-P* $(A^\infty, P_{A^\infty}, \rho', \epsilon, \delta)$ risolve anche l'istanza iniziale *Q-F*, dimostrando così il corollario. □

A questo punto è naturale trovare un algoritmo efficiente per risolvere (k, ρ). Successivamente, gli autori hanno discusso dell'estensione di *Q-P* a (k, ρ) e presentato un lower e un upper bound sulla complessità del campione necessaria per risolverlo.

4.4.2. Risolvere le istanze "al massimo k-equiprobabili"

Ora si vuole identificare le k $[\epsilon, \rho]$-optimal arms. Nel teorema 4.4 gli autori hanno dedotto il lower bound sulla complessità del campione per risolvere un istanza (k, ρ) riducendo a risolvere il SUBSET problem [CK17] che segue.

Teorema 4.4. *[Lower Bound sulla complessità del campione per risolvere (k, ρ)] Per ogni $\epsilon \in (0, \frac{1}{\sqrt{32}}]$, $\delta \in (0, \frac{1}{\sqrt{32}}]$, e $\rho \in (0, \frac{1}{2}]$, esiste un istanza di (k, ρ) data da $(A, P_A, \rho, \epsilon, \delta)$, tale che ogni algoritmo che risolver (k, ρ) necessita di almeno $C \cdot \frac{k}{\rho\epsilon^2} \ln \frac{k}{8\delta}$ campioni, dove $C = \frac{1}{18375}$.*

Dimostrazione. Dimostriamo il teorema per assurdo. Assumiamo che il teorema non sia corretto. Allora esiste un algoritmo ALG tale che ALG può risolvere qualsiasi istanza di (k, ρ) usando non più di $C \cdot \frac{k}{\rho\epsilon^2} \ln \frac{k}{8\delta}$ campioni, con $C = \frac{1}{18375}$. Ora, data $(n, A, m, \epsilon, \delta)$ un istanza del SUBSET [CK17] con $n \geq 2m$. Sia $P_A = Uniform\{1, 2, \ldots, n\}$, $k = m$, e $\rho = m/n$, creiamo un'istanza di (k, ρ) come $(A, P_A, \rho, k, \epsilon, \delta)$. Allora, risolvere questa (k, ρ) istanza significa risolvere l'istanza del SUBSET. In accordo a quanto abbiamo definito, ALG risolve il problema SUBSET usando al massimo $C \cdot \frac{k}{(k/n)\epsilon^2} \ln \frac{k}{8\delta} = C \cdot \frac{m}{(m/n)\epsilon^2} \ln \frac{m}{8\delta} = C \cdot \frac{n}{\epsilon^2} \ln \frac{m}{8\delta}$ campioni. Questa osservazione contraddice il lower bound della complessità del campione per risolvere il SUBSET. In questo modo abbiamo dimostrato il problema. □

Algoritmo per risolvere le istanze "al massimo k-equiprobabile" Sia, per ogni $\mathcal{S} \subseteq A$, $\nu(\mathcal{S}) = \Pr_{a \sim P_A}\{a \in \mathcal{S}\}$. Allora, $\nu(A) = 1$. Ora, presentiamo un algoritmo *KQP- 1* che può risolvere qualsiasi delle most k-equiprobable istanze di (k, ρ). L'algoritmo 4.3 descrive *KQP-1*. Ad ogni fase y, l'algoritmo risolve un istanza di *Q-P* avente in output un arm, chiamata $a^{(y)}$, da $\mathcal{TOP}_\rho(\epsilon)$. Nella prossima fase, si aggiorna l'istanza del bandit $A^{y+1} = A^y \setminus \{a^{(y)}\}$, la distribuzione dei campioni $P_{A^{y+1}} = \frac{1}{1-\nu(A \setminus A^{y+1})} P_{A^y}$, e il quantile target $\rho^{y+1} = \rho^y - \nu(a^{(y)})$. Tuttavia, visto che non diamo una forma esplicita di P_A, realizziamo $P_{A^{y+1}}$ con *rejection-sampling* (se $a' \in A \setminus A^{y+1}$ è scelta da P_A, noi semplicemente scartiamo a' e continuiamo a campionare P_A una

volta in più). Visto che $\nu(\{a^y\})$ non è conosciuto esplicitamente, ci basiamo sul fatto che $\nu(\{a^y\}) \leq \rho/k$: è per questa ragione che noi necessitiamo che l'istanza sia al massimo k-equiprobable. Allora, in ogni fase $y \geq 1$, $\rho^y - \rho/k \leq \rho^{y+1} \leq \rho^y - \nu\{a^y\}$, e quindi *KQP-1* risolve un istanza di *Q-P* data da $(A^y, P_{A^y}, \rho - (y-1)\rho/k, \epsilon, \delta)$.

Algoritmo 4.3 Algoritmo *KQP-1* per risolvere le most k-equiprobabile (k, ρ) istanze

Input: $A, P_A, k, \rho, \epsilon, \delta$.
Output: Set di k arms distinte da $\mathcal{TOP}_\rho(\epsilon)$.

$A^1 = A, \rho^1 = \rho$.
for $y = 1, 2, 3, \cdots, k$
  Run $\mathcal{P}_3$ per risolvere l'istanza Q-P data da $(A^y, P_{A^y}, \rho^y, \epsilon, \frac{\delta}{k})$, e sia $a^{(y)}$ l'output.
  $A^{y+1} = A^y \setminus \{a^{(y)}\}$.
  $\rho^{y+1} = \rho^y - ((y-1)\rho)/k$.
end for

Nel teorema 4.5 presentiamo un upper buond della complessità di campionamento attesa del problema *KQP-1*.

Teorema 4.5. *Data una qualsiasi al massimo k-equiprobable istanza di (k, ρ) con $k > 1$, KQP-1 risolve l'istanza con una complessità del campione attesa limitata superiormente da $O\left(\frac{k}{\epsilon^2}\left(\frac{\log k}{\rho}\log\frac{k}{\delta} + \log^2\frac{k}{\delta}\right)\right)$.*

Dimostrazione. Dividiamo la dimostrazione in due parti: upper-bounding della complessità del campione e la verifica di correttezza.

Complessità del campione

Dimostrazione. Nella fase y, la complessità del campione $\mathcal{P}_3$ ha come upper bound $\mathrm{SC}(y) \leq \frac{C}{\rho^y\epsilon^2}\log\frac{k}{\delta}$, per qualche costante C. Inoltre, la complessità del campione di *KQP-1* è limitata superiormente in questo modo:

$$\begin{aligned}
&\sum_{y=1}^{k}\mathrm{SC}(y) \leq \sum_{y=1}^{k}\frac{C}{\epsilon^2}\left(\frac{1}{\rho^y}\log\frac{k}{\delta} + \log^2\frac{k}{\delta}\right),\\
&\leq \frac{C}{\epsilon^2}\left(\log\frac{k}{\delta}\sum_{y=1}^{k}\frac{1}{\rho - (y-1)\frac{\rho}{k}} + k\log^2\frac{k}{\delta}\right),\\
&= \frac{Ck}{\epsilon^2}\left(\frac{1}{\rho}\log\frac{k}{\delta}\sum_{y=1}^{k}\frac{1}{k-y+1} + \log^2\frac{k}{\delta}\right),\\
&\leq \frac{C'k}{\epsilon^2}\left(\frac{\log k}{\rho}\log\frac{k}{\delta} + \log^2\frac{k}{\delta}\right),
\end{aligned}$$

per $k > 1$, e qualche costante C'. □

Correttezza

Dimostrazione. Sia E_y l'evento che $a^{(y)} \notin TOP_\rho(\epsilon)$, la probabilità d'errore di KQP-1 può essere limitata superiormente come $\Pr\{\text{Error}\} = \Pr\{\exists y \in \{1, \cdots, k\}\ E_y\} \leq \sum_{y=1}^{k} \Pr\{E_y\} \leq \sum_{y=1}^{k} \frac{\delta}{k} = \delta$. □

□

Corollario 4.2. *KQP-1 può risolvere qualsiasi istanza di (k, m, n) data da $(A, n, m, k, \epsilon, \delta)$ con $k \geq 2$, usando $O\left(\frac{k}{\epsilon^2}\left(\frac{n \log k}{m} \log \frac{k}{\delta} + \log^2 \frac{k}{\delta}\right)\right)$ campioni.*

Notiamo che sebbene la complessità del campione di *KQP-1* sia indipendente dalla dimensione dell'istanza del bandit A, e ogni istanza di (k, m, n) data da $(A, n, m, m, \epsilon, \delta)$, può essere risolta da KQP-1 ponendola come un istanza di (k, ρ) data da $(A, Uniform\{A\}, m/n, m, \epsilon, \delta)$. Tuttavia, per $k = m$, la complessità del campione di KQP-1 viene ridotta a $O\left(\frac{1}{\epsilon^2}\left(n \log m \cdot \log \frac{m}{\delta} + \log^2 \frac{m}{\delta}\right)\right)$, che è più grande della complessità del campione di Halving [KS10] che necessita solo di $O\left(\frac{n}{\epsilon^2} \log \frac{m}{\delta}\right)$ campioni. Quindi, per il problema del *best subset selection* in istanze finite Halving è meglio di *KQP-1*. Tuttavia, in istanze molto grandi, dove la probabilità di pescare una qualsiasi arm data da $\mathcal{TOP}_\rho$ è prossima allo zero, è il problema ideale da risolvere e *KQP-1* è la prima soluzione che proponiamo.

Corollario 4.3. *Ogni istanza di (k, ρ) data da $(A, P_A, k, \rho, \epsilon, \delta)$, tale che $|A| = \infty$, e per tutti i subset finiti $S \subset A$, $\Pr_{a \sim P_A}\{a \in S\} = 0$; può essere risolta con una complessità del campione compresa in $O\left(k\epsilon^{-2}\left(\rho^{-1} \log(k/\delta) + \log^2(k/\delta)\right)\right)$, risolvendo indipendentemente k istanze differenti di Q-P, ognuna data da $(A, P_A, k, \rho, \epsilon, \delta/k)$.*

La correttezza del corollario può essere provata notando che tutti i k outputs sono unici con probabilità 1 e poi prendendo l'union bound sulle probabilità d'errore. Prima di passare agli esperimenti, verranno presentati dei risultati importanti sulla complessità nella risoluzione di *Q-P*.

4.4.3. Complessità della risoluzione

Il teorema 4.6 presenta una relazione generale tra l'upper bound della complessità dei campioni per risolvere *Q-F* e *Q-P*.

Teorema 4.6. *Sia $\gamma : \mathbb{Z}^+ \times \mathbb{Z}^+ \times [0, 1] \times [0, 1] \mapsto \mathbb{R}^+$. Se ogni istanza di Q-F data da $(A, n, m, \epsilon, \delta)$, può essere risolta con la complessità dei campioni compresa in $O\left(\frac{n}{m\epsilon^2} \log \frac{1}{\delta} + \gamma(n, m, \epsilon, \delta)\right)$, allora, ogni istanza di Q-P data da $(A, P_A, \rho, \epsilon, \delta)$ può essere risolta con la complessità del campione compresa in $O\left((1/\rho\epsilon^2) \log(1/\delta) + \gamma\left(\lceil 8 \log(2/\delta) \rceil, \lfloor 4 \log(2/\delta) \rfloor, \epsilon/2, \delta/2\right)\right)$.*

Assumiamo che esista un algoritmo OptQF che risolve ogni istanza di Q-F data da $(A, n, m, \epsilon, \delta)$, usando $O\left(\frac{n}{m\epsilon^2}\log\frac{1}{\delta} + \gamma(n, m, \epsilon, \delta)\right)$ campioni. Stabiliamo un upper bound sulla complessità del campione per risolvere Q-P costruendo un algoritmo OptQP che segue un approccio similare a $\mathcal{P}_3$. In particolare, OptQP riduce l'istanza di Q-P in input in un istanza di Q-F usando $O\left(\frac{1}{\rho\epsilon^2}\log\frac{1}{\delta}\right)$ campioni. Poi, risolve Q-F usando OptQF come subroutine. La dimostrazione dettagliata è data nell'appendice C.

Corollario 4.4. *Il corollario 4.1 mostra che ogni istanza di Q-F può essere risolta con* $O\left(\frac{1}{\epsilon^2}\left(\frac{n}{m}\log\frac{1}{\delta} + \log^2\frac{1}{\delta}\right)\right)$ *campioni. Quindi,* $\gamma(n, m, \epsilon, \delta) \in O\left(\frac{1}{\epsilon^2}\log^2\frac{1}{\delta}\right)$*, e quindi, ogni Q-P è risolvibile in* $O\left(\frac{1}{\epsilon^2}\left(\frac{1}{\rho}\log\frac{1}{\delta} + \log^2\frac{1}{\delta}\right)\right)$ *campioni. Dall'altro lato, se il lower bound per risolvere Q-F dato da [CK17] corrisponde con l'upper bound a meno di un fattore costante, allora* $\gamma(n, m, \epsilon, \delta) \in \Theta\left(\frac{n}{m\epsilon^2}\log\frac{1}{\delta}\right)$*. In quel caso, Q-P è risolvibile usando* $\Theta\left(\frac{1}{\rho\epsilon^2}\log\frac{1}{\delta}\right)$ *campioni.*

È interessante trovare un $\gamma(\cdot)$ tale che l'upper bound presentato nel teorema 4.6 corrisponda con il lower bound a meno di un fattore costante. Notiamo che il teorema 4.6 garantisce che esiste una costante C, tale che per ogni ϵ, δ dati, e $m \leq n/2$, $\gamma(n, m, \epsilon, \delta) \leq C \cdot \gamma\left(\lceil 8\log(2/\delta)\rceil, \lfloor 4\log(2/\delta)\rfloor, \frac{\epsilon}{2}, \frac{\delta}{2}\right)$. Tuttavia, per $n < \lceil 8\log(2/\delta)\rceil$ crediamo che Q-F può essere risolto più efficientemente rispetto a porlo come Q-P.

Definizione 4.3. Per $g : \mathbb{Z}^+ \times \mathbb{Z}^+ \times [0,1] \times [0,1] \mapsto \mathbb{R}^+$ diciamo che Q-F è risolvibile in $\Theta(g(\cdot))$, se esiste un algoritmo che risolve ogni istanza di Q-F data da $(A, n, m, \epsilon, \delta)$ in $O(g(n, m, \epsilon, \delta))$ campioni, e esiste un istanza di Q-F data da $(\bar{A}, \bar{n}, \bar{m}, \bar{\epsilon}, \bar{\delta})$ tale che ogni algoritmo ha bisogno di almeno $\Omega(g(\bar{n}, \bar{m}, \bar{\epsilon}, \bar{\delta}))$ campioni per risolverla.

Congettura 4.1. *Esiste una costante* $C > 0$*, e le funzioni* $g : \mathbb{Z}^+ \times \mathbb{Z}^+ \times [0,1] \times [0,1] \mapsto \mathbb{R}^+$*,* $e\ h : \mathbb{Z}^+ \times \mathbb{Z}^+ \times [0,1] \times [0,1] \mapsto \mathbb{R}^+$*, tali che per ogni* $\delta \in (0,1]$*, esiste un intero* $n_0 < C\log\frac{2}{\delta}$*, tale che per ogni* $n \leq n_0$*, Q-F è risolvibile in* $\Theta(g(n, m, \epsilon, \delta))$ *campioni, ed è equivalente a Q-P (ottenuto ponendo l'istanza Q-F come istanza di Q-P, come è stato fatto dal corollario 4.1) e ha bisogno di almeno* $\Omega(h(n, m, \epsilon, \delta))$ *campioni, poi* $\lim_{\delta\downarrow 0}\frac{g(n,m,\epsilon,\delta)}{h(n,m,\epsilon,\delta)} \to 0$.

Successivamente, gli autori comparano empiricamente $LUCB$-k-m avente $k = 1$ con F_2 su istanze differenti e inoltre studiano empiricamente la performance di $LUCB$-k-m variando k.

5. Esperimenti

In questa sezione verranno replicati gli esperimenti fatti dagli autori. In particolare gli autori hanno confrontato le performance tra gli algoritmi presentati precedentemente.

5.1. Confronto F_2 e $LUCB - k - m$

Si inizia comparando l'algoritmo F_2 [CK17] con $LUCB - k - m$ basandosi sul numero di campioni estratti da istanze differenti di Q-F o $(1; m; n)$. F_2 è un algoritmo completamente sequenziale che assomiglia a $LUCB - k - m$ ma ha delle differenze sottili nel modo in cui l'algoritmo partiziona A e nel selezionare le arms da tirare e ciò porta a risultati differenti. Allo stesso turno t, F_2 partiziona A in $\bar{A}_1(t)$, $\bar{A}_2(t)$, e $\bar{A}_3(t)$. L'algoritmo mette le arms con il più alto LCB in $\bar{A}_1(t)$, inoltre mette le $m-1$ arms con i più alti UCBs in $\bar{A}_2(t)$; e il resto delle $n-m$ arms in $\bar{A}_3(t)$; i legami sono rotti a caso. A ogni turno t, l'algoritmo campiona tre arm: l'arm in $\bar{A}_1(t)$, l'ultima arm campionata in $\bar{A}_2(t)$, e l'arm avente l'UCB più alto in $\bar{A}_3(t)$.

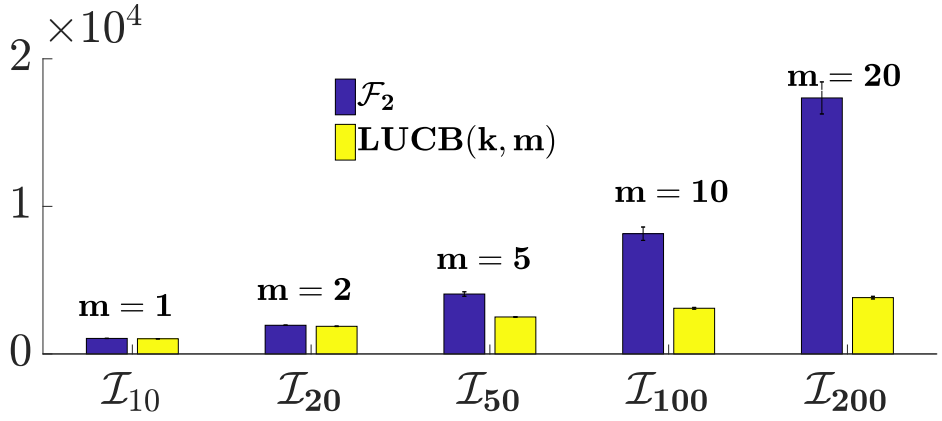

Figura 5.1.: Confronto delle complessità del campione di F_2 e $LUCB - k - m$ per risolvere Q-F con $m = n/10$, sulle cinque istanze descritte sopra. In questo grafico e anche nei successivi, l'asse y rappresenta la media della complessità del campione su 100 runs con barre di standard error.

Ora prendiamo cinque istanze di Bernoulli dalla dimensione $n = 10, 20, 50, 100$, e 200, con la media spaziata linearmente tra 0.999 e 0.001 (estremi inclusi), e queste

istanze vengono ordinate in ordine decrescente. Chiamiamo l'istanza del bandit di dimensione n come I_n. Ora, settato $\epsilon = 0.05, \delta = 0.001$, e $m = 0.1 \times n$, eseguiamo gli esperimenti e compariamo il numero di sample pescati da F_2 e $LUCB - k - m$ per risolvere queste cinque istanze per $k = 1$. Nella nostra implementazione abbiamo usato la *K-divergence* basata sui bounds dei coefficienti [Cap+13] sia per F_2 che $LUCB - k - m$. Come viene mostrato nella figura 5.1, come il numero di arms (n) aumenta, la complessità del campione di entrambi gli algoritmi aumenta a causa dell'aumento dell'hardness H_ϵ. Tuttavia, la complessità del campione di F_2 aumenta molto più velocemente di $LUCB - k - m$.

Come mostrato da [JN14], l'efficienza di *LUCB1* deriva dall'identificazione rapida dell'arm più ottimale dovuta alla grande separazione dalla $m + 1$-esima arm. Intuitivamente, la ragione per cui F_2 ha bisogno di più campioni è dovuta al ritardo nell'assegnazione delle priorità per tirare più frequentemente l'arm ottimale. Ciò dovrebbe risultare in una frazione più piccola di tutti i campioni totali presi dall'arm migliore. La figura 5.2 conferma quest'intuizione. Viene rappresentato il confronto tra F_2 e $LUCB-k-m$ sul numero di campioni ottenuti da tre *"ground-truth" groups* — B_1, B_2, e B_3 su I_{10}, tenendo $k = 1$ e variando m da 1 a 5. Notiamo che minore è la differenza tra k e m, maggiore è l'hardness (H_ϵ), e sia F_2 e $LUCB - k - m$ trovano difficoltà a identificare l'arm corretta. Quindi, per $k = m = 1$, entrambi spendono la stessa frazione dei pulls sulla miglior arm. Tuttavia, come m diventa più grande, tenendo $k = 1$, l'hardness del problema si riduce ma F_2 fa ancora fatica a identificare l'arm migliore e ciò si traduce in una frazione significativamente minore dei pulls totali per esso, rispetto a $LUCB - k - m$.

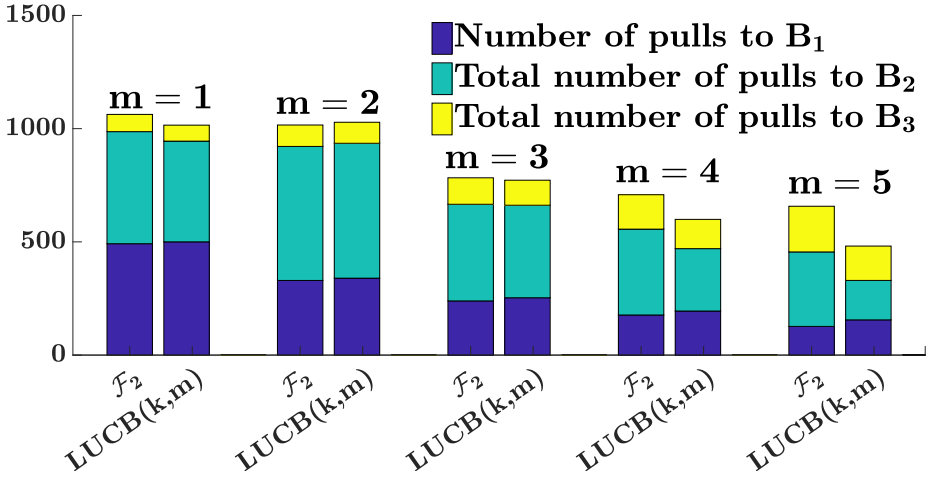

Figura 5.2.: Confronto tra F_2 e $LUCB - k - m$ sul numero di pull ricevuti dai campi B_1, B_2, e B_3, per risolvere diverse istanze di Q-F su I_{10}, variando m da 1 a 5. Ricorda che B_1 è l'insieme singleton, con l'arm migliore come unico membro.'asse y rappresenta la media della complessità del campione su 100 runs.

5.2. Confronto del numero di campioni usati per risolvere istanze differenti di $(k; m; n)$

In questo paper, gli autori hanno sviluppato algoritmi specifici per il problema $(k; m; n)$; precedentemente uno poteva risolvere il problema $(k; m; n)$ risolvendo (k, k, n) o (m, m, n): cioè scegliere il miglior k- o m-sized subset. Nella figura 5.3 viene presentato un confronto sulla complessità del campione per risolvere $(k; m; n)$ e i problemi *best subset-selection*. Fissando $A = I_{20}$, $n = 20$, $m = 10$, le istanze $(k; m; n)$ sono date variando $k \in \{1, 3, 5, 8, 10\}$, dove, per *best subset-selection problem* abbiamo settato $m = k$.

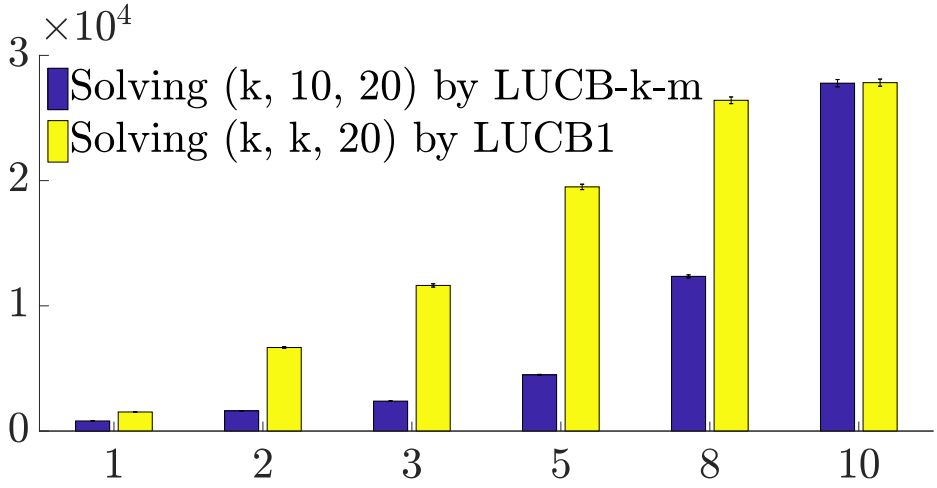

Figura 5.3.: Confronto del numero di campioni usati per risolvere istanze differenti di $(k; m; n)$ definite su I_{20}, settando $m = 10$, e variando $k \in \{1, 2, 3, 5, 8, 10\}$. L'asse x rappresenta k, l'asse y rappresenta il numero di campioni mediato su 100 runs, con barre di standard error.

Come ci si aspettava, il numero di campioni usati è significativamente inferiore per risolvere il problema delle istanze con $k < m$, convalidando l'uso di $LUCB - k - m$.

6. Conclusioni

Identificare un arm tra le migliori m, in un *n-armed stochastic bandit* è un problema molto interessate identificato da [CK17]. Loro hanno menzionato gli scenari dove identificare il miglior subset è praticamente impossibile. Tuttavia, vi sono numerosi esempi in pratica che richiedono l'identificazione efficiente di più soluzioni buone invece che una sola; per esempio, assegnando un compito di crowdsourcing distribuito, identificazione di buone combinazioni molecolari nel design dei farmaci, etc. In questo paper, gli autori hanno presentato il problema $(k; m; n)$: una generalizzazione del problema di identificare k tra le best m arms. Settato $k = 1$, $(k; m; n)$ viene ridotto alla selezione di uno tra le migliori m arms, mentre settato $k = m$, diventa uguale al subset selection [KS10]. Gli autori hanno presentato un lower bound alla complessità del campione per risolvere $(k; m; n)$. Inoltre hanno presentato un algoritmo completamente sequenziale adattivo *PAC, LUCB-k-m,* che risolve il problema $(k; m; n)$ con una complessità del campione attesa che corrisponde rispettivamente, a meno di un fattore costante, a F_2 [CK17] e *LUCB1* [Kal+12] per $k = 1$ e $k = m$. Inoltre hanno confrontato empiricamente i due algoritmi su differenti istanze del problema e mostrato che *LUCB1* è molto migliore di F_2 con un ampio margine del numero dei campioni come n cresce.

Per il problema dell'identificazione di una singola $[\epsilon, \rho]$-optimal arm [CK17] tra istanze infinite del bandit, gli esistenti upper bound sulla complessità del campione differiscono dal lower bound per un fattore moltiplicativo $\log \frac{1}{\delta}$. Non era chiaro se il limite inferiore fosse allentato o se quello superiore potesse essere migliorato e ciò ha lasciato un problema interessante da risolvere come descritto da [Azi+18]. In questo paper, gli autori hanno ridotto il gap fornendo un upper bound che è ottimale a meno di un termine polinomiale-logaritmico. Inoltre, hanno mostrato che il problema dell'identificazione di k distinct $[\epsilon, \rho]$-optimal arms non è ben posto in generale, ma quando lo è, gli autori hanno derivato un lower bound sulla stessa complessità del campione. Inoltre, hanno identificato una classe di istanze ben poste per cui gli autori hanno presentato un algoritmo efficiente. Alla fine hanno mostrato come il miglioramento del limite superiore della complessità del campione per la risoluzione di istanze *Q-F* può essere tradotto nel miglioramento del limite superiore della complessità del campione per la risoluzione di *Q-P*. Tuttavia, hanno ipotizzato che esista un insieme di istanze *Q-F* e un corrispondente insieme di istanze *Q-P*, in modo tale che ogni istanza di *Q-F* richieda un numero minore di campioni da risolvere rispetto alla corrispondente istanza *Q-P* nell'altro insieme. Hanno mostrato la correttezza dell'ipotesi e migliorato sia il lower che l'upper bound sulla stessa complessità del campione e altre interessanti strade da percorrere in futuri lavori.

Nella figura 6.1 vi è un riassunto di tutti i lower e upper bound migliorati da questo paper.

Problem	Lower Bound	Previous Upper Bound	Current Upper Bound
$(1,1,n)$ Best-Arm	$\Omega\left(\frac{n}{\epsilon^2}\log\frac{1}{\delta}\right)$ (Mannor & Tsitsiklis, 2004)	$O\left(\frac{n}{\epsilon^2}\log\frac{1}{\delta}\right)$ (Even-Dar et al., 2002)	Same as previous
(m,m,n) SUBSET	$\Omega\left(\frac{n}{\epsilon^2}\log\frac{m}{\delta}\right)$ (Kalyanakrishnan et al., 2012)	$O\left(\frac{n}{\epsilon^2}\log\frac{m}{\delta}\right)$ (Kalyanakrishnan & Stone, 2010)	Same as previous
$(1,m,n)$ Q-F	$\Omega\left(\frac{n}{m\epsilon^2}\log\frac{1}{\delta}\right)$ (Roy Chaudhuri & Kalyanakrishnan, 2017)	$O\left(\frac{n}{m\epsilon^2}\log^2\frac{1}{\delta}\right)$ (Roy Chaudhuri & Kalyanakrishnan, 2017)	$O\left(\frac{1}{\epsilon^2}\left(\frac{n}{m}\log\frac{1}{\delta}+\log^2\frac{1}{\delta}\right)\right)$ **This paper**
(k,m,n) Q-F$_k$	$\Omega\left(\frac{n}{(m-k+1)\epsilon^2}\log\frac{\binom{m}{k-1}}{\delta}\right)$ **This paper**	-	$O\left(\frac{k}{\epsilon^2}\left(\frac{n\log k}{m}\log\frac{k}{\delta}+\log^2\frac{k}{\delta}\right)\right)^*$ **This paper** (*for $k\geq 2$)
$(1,\rho)$ $(\lvert\mathcal{A}\rvert=\infty)$ Q-P	$\Omega\left(\frac{1}{\rho\epsilon^2}\log\frac{1}{\delta}\right)$ (Roy Chaudhuri & Kalyanakrishnan, 2017)	$O\left(\frac{1}{\rho\epsilon^2}\log^2\frac{1}{\delta}\right)$ (Roy Chaudhuri & Kalyanakrishnan, 2017)	$O\left(\frac{1}{\epsilon^2}\left(\frac{1}{\rho}\log\frac{1}{\delta}+\log^2\frac{1}{\delta}\right)\right)$ **This paper**
(k,ρ) $(\lvert\mathcal{A}\rvert=\infty)$ Q-P$_k$	$\Omega\left(\frac{k}{\rho\epsilon^2}\log\frac{k}{\delta}\right)$ **This paper**	-	$O\left(\frac{k}{\epsilon^2}\left(\frac{\log k}{\rho}\log\frac{k}{\delta}+\log^2\frac{k}{\delta}\right)\right)^*$ **This paper** (*for a special class with $k\geq 2$)

Figura 6.1.: Limiti inferiore e superiore sulla complessità del campione attesa (ponendosi nel caso peggiore). I limiti per $(k;\rho)$, $k>1$ sono per la classe speciale di istanze "al massimo k-equiprobabili".

6.1. Applicazioni nella vita reale

Nella sezione 2.1 erano già stati descritti i seguenti scenari applicativi citati anche dal paper analizzato:

- **Clinical Trials:** miglioramento delle cure su persone e animali volte a rendere più affidabili e migliori i trattamenti usati [Rob52], [Dur+18].
- **Drug Design**: progettazione di nuovi farmaci [Wil+16].
- **Internet Advertising**: ogni volta che un utente visita un sito è necessario scegliere di mostrare una delle k pubblicità possibili per massimizzare il reward complessivo [Li+10].
- **Network Server Selection:** selezione del miglior server in applicazioni quali *adaptive routing* [AK08], nel *DNS server selection* e nel *cloud computing.*
- **Gestione di grandi reti di sensori**: gestione di grandi reti di sensori in cui più sensori affidabili devono essere identificati facendo il minor numero di tests possibili [Mou+16].
- **Distribuited Crowdsouring:** crowdsourcing distribuito [Tra+14].

Ulteriori applicazioni nella vita reale sono le seguenti:

- **Financial Portfolio Design:** gestione del portfolio finanziario (collezione di investimenti) di una compagnia di investimenti per massimizzare il reward cumulativo. Ad esempio Shen, Weiwei, Wang, Jun, Jiang, Yu-Gang, e Zha, Hongyuan in *«Portfolio choices with orthogonal bandit learning»* [She+15] hanno proposto un algoritmo bandit per fare scelte online sul portfolio sfruttando le correlazioni tra arms multiple. Costruendo portafogli ortogonali da più assets e integrando il loro approccio col *upper-confidence-bound bandit framework,* gli autori hanno ottenuto la strategia ottimale del portfolio rappresentata da una combinazione di investimenti passivi e attivi in accordo con la *risk-adjusted reward function.* Ulteriori esempi si possono trovare nel paper *«Portfolio allocation for Bayesian optimization»* di Brochu, Eric, Hoffman, Matthew W, and de Freitas, Nando [BHF10].
- **Dynamic Pricing:** pricing dinamico, identificare il prezzo di vendita per un particolare tipo di prodotto che massimizza il profitto del venditore senza conoscere la domanda dei consumatori. Trovò, Francesco, Paladino, Stefano, Restelli, Marcello, e Gatti, Nicola in *«Multi--armed bandit for pricing»* [Tro+15] hanno proposto una modifica dell'algoritmo upper confidence bound (UCB) per i bandit sfruttando due peculiarità del *pricing*: (1) come il prezzo di vendita sale, è razionale assumere che la probabilità di vendere un oggetto diminuisce; (2) dato che solitamente le persone comparano prezzi da differenti venditori e tengono traccia dei cambiamenti di prezzo nel tempo prima di acquistare (specialmente per acquisti online, esistono anche vari tools che permettono di tenere traccia automatica dei prezzi di prodotti online), il numero di volte che

un certo tipo di oggetto è acquistato è solo una piccola frazione del numero di volte che il prezzo è visualizzato da acquirenti possibili. Facendo leva su queste assunzioni, gli autori migliorano la *concentration inequality* usata nell'algoritmo *UCB1,* che porta ad avere risultati migliori rispetto all'originale, specialmente nei primi passi del learning dove solo pochi campioni sono disponibili. Gli autori hanno presentato un'evidenza empirica sull'efficacia delle variazioni proposte in termini di aumento della velocità del processo di learning dell'algoritmo *UCB1* nelle applicazioni di *pricing*.

- **Recommender System:** i sistemi di raccomandazione sono frequentemente usati in varie applicazioni per predire le preferenze degli utenti. In questi sistemi vi è il dilemma *exploration-exploitation* dal momento che vi è bisogno di avere conoscenza rispetto agli oggetti precedentemente selezionati (oggetti interessanti per gli utenti) ed esplorare nuovi oggetti che magari possono piacere agli utenti. Questo tipo di problema è stato affrontato da Qian Zhou, XiaoFang Zhang, Jin Xu, e Bin Liang in *«Large-scale bandit approaches for recommender systems»* [Zho+17] attraverso un *multi-armed bandit*, costruito soprattutto per dei sistemi di raccomandazione su larga scala che hanno tanti o addirittura infiniti elementi. Un ulteriore esempio è l'ottimizzazione dei contenuti offerti ai lettori, basandosi sugli interessi dell'utente come l'ottimizzazione dei titoli di notizie e articoli svolta da Li, Lihong, Chu, Wei, Langford, John, and Schapire, Robert E. in *«A contextual-bandit approach to personalized news article recommendation»* [Li+10].
- **Influence Maximization:** Vaswani, S., Kveton, B., Wen, Z., Ghavamzadeh, M., Lakshmanan, L., & Schmidt, M. in *«Model-independent online learning for influence maximization»* [Vas+17] considerano la massimizzazione dell'influenza (*IM*) nei social networks, in cui il problema è massimizzare il numero di utenti che diventano interessati a un prodotto selezionando un set di *seed users* per esporre il prodotto. Gli autori hanno proposto un nuova parametrizzazione che rende il framework indipendente dal sottostante modello di diffusione ma anche staticamente efficiente per imparare dai dati.
- **Information Retrieval:** Losada, D. E., Parapar, J., & Barreiro, A. in *«Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems»* [LPB17] hanno modellato il processo iterativo di selezione e recupero dell'informazione come un *contextual bandit problem*. In particolare hanno prodotto numerosi metodi per la *document adjudication* (cinque stazionari e due non stazionari) per effettuare un confronto fra loro per la valutazione *pooling-based* e *metasearch*.
- **Dialogue Systems:** in particolare nei seguenti ambiti:
  - **Dialogue response selection**: Liu, B., Yu, T., Lane, I., & Mengshoel, O. J. in *«Customized nonlinear bandits for online response selection in neural conversation models»* [Liu+17] hanno proposto un *contextual multi-armed bandit model* con un *non-linear reward function* che usa

una rappresentazione distribuita di testo per una selezione online della risposta.

  - **Dialogue response selection:** Silander, T. in *«Contextual memory bandit for pro-active dialog engagement»* [Sil+18] ha proposto un *contextual multi-armed bandit model* per il problema della massimizzazione del reward con feedback parziale.
  - **Multi-domain dialogue systems:** Upadhyay, S., Agarwal, M., Bounneffouf, D., & Khazaeni, Y. in *«A Bandit Approach to Posterior Dialog Orchestration Under a Budget»* [Upa+19] hanno studiato il compito dell'*online posterior dialogue orchestration* che hanno definito come il compito di seleziona un subset di skills che più appropriatamente rispondono agli input dell'utente e delle skills individuali.

- **Telecommunication:** Boldrini, S., De Nardis, L., Caso, G., Le, M. T., Fiorina, J., & Di Benedetto, M. G. in *«muMAB: A multi-armed bandit model for wireless network selection»* [Bol+18] hanno usato un modello *multi-armed* per descrivere il problema della selezione del miglior rete wireless eseguita da un dispositivo *multi-Radio Access Technology (multi-RAT)* con il fine di massimizzare la qualità percepita dall'utente finale.
- **Anomaly Detection:** la ricerca delle anomalie su reti si occupa di identificare i nodi il cui comportamento è significativamente diverso da quello dei restanti nodi. Ding, K., Li, J., & Liu, H. in *«Interactive anomaly detection on attributed networks»* [DLL19] hanno sviluppato un *collaborative contextual bandit algorithm* che modella gli attributi e le dipendenze dei nodi e gestisce il dilemma *exploration-exploitation* quando si investigano anomalie di differenti tipi.

In figura 6.2 è presente un riassunto delle possibili applicazioni del *bandit problem* nella vita reale.

	MAB	Non-stationary MAB	CMAB	Non-stationary CMAB
Healthcare	√		√	
Finance	√			
Dynamic pricing		√		
Recommendr system	√	√	√	√
Maximization	√			
Dialogue system			√	
Telecomunication	√			
Anomaly detection	√			

Figura 6.2.: Bandit in applicazioni pratiche nella vita reale

6.2. Applicazioni nel machine learning

In ambito *machine learning* i possibili usi sono nei seguenti ambiti (alcuni visti a lezione):

- **Algorithm Selection:** solitamente era basato su modelli di performance degli algoritmi costruiti con sequenze di training offline; attraverso un approccio online si può iterativamente aggiornare il modello e usarlo per la selezione di una sequenza di istanze del problema analizzato. Il trade-off *exploration-exploitation* è rappresentato da un bandit usando informazioni parziali e un limite sconosciuto sulle perdite come proposto da Gagliolo, M., & Schmidhuber, J. in *«Algorithm selection as a bandit problem with unbounded losses»* [GS10].
- **Hyperparameter Optimization:** l'ottimizzazione degli iperparametri è fondamentale per le performance degli algoritmi di machine learning. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. in *«Hyperband: A novel bandit-based approach to hyperparameter optimization»* [Li+17] hanno formulato un'ottimizzazione degli iperparametri come un infinite-armed bandit di pura esplorazione non stocastico dove le risorse predefinite come le iterazioni, i campioni di dati e le features sono allocati come configurazione di campioni random.
- **Feature Selection:** Bouneffouf, D., Rish, I., Cecchi, G. A., & Féraud, R. in *«Context attentive bandits: Contextual bandit with restricted context»* [Bou+17] hanno affrontato il problema dell'*online feature selection* associando il problema dell'ottimizzazione combinatoria in un bandit stocastico con bandit feedback usando l'algoritmo *Active Thompson Sampling (ATS).*
- **Bandits per Active Learning:** etichettare tutti gli esempi di training nella classificaizone supervisionata può essere complicato. Le strategie di *active learning* risolvono questo problema selezionando solo i campioni non etichettati più utili per ottenere la label e per allenare il modello predittivo. La scelta dei campioni da etichettare può essere vista come un dilemma tra l'*exploration* e l'*exploitation* su uno spazio di input. Bouneffouf, D., Laroche, R., Urvoy, T., Féraud, R., & Allesiardo, R. in *«Contextual bandit for active learning: Active thompson sampling»* [Bou+14] hanno proposto di modellare questo problema come un *contextual bandit problem* e proposto un algoritmo sequenziale chiamato *Active Thompson Sampling (ATS).*
- **Clustering**: per quanto riguarda il clustering collaborativo, Sublime, J., & Lefebvre, S. in *«Collaborative clustering through constrained networks using bandit optimization»* [SL18] hanno proposto un algoritmo collaborativo *peer to peer* per il clustering basato sul principio del *non stochastic multi-arm bandits* per assicurarsi in tempo reale quali algoritmi o views potevano portare informazioni utili.
- **Reinforcement Learning:** gli agenti devono sviluppare tecniche per massimizzare il loro rewards e seguire delle limitazioni implicite presenti nell'ambiente.

Noothigattu, R., Bouneffouf, D., Mattei, N., Chandra, R., Madan, P., Varshney, K., & Rossi, F. in *«Interpretable multi-objective reinforcement learning through policy orchestration»* [Noo+18] hanno usato un *contextual bandit-based orchestrator* che prende una scelta appropriata tra due policies: *constraint-based* e *environment reward-based*.

In figura 6.3 è presente un riassunto delle possibili applicazioni del *bandit problem* nel *machine learning*.

	MAB	Non-stationary MAB	CMAB	Non-stationary CMAB
Algorithm Slection		√		
Parameter Optimization	√			
Features Selection	√	√		
Active Learning	√		√	
Clustering	√			
Reinforcement learning	√	√	√	

Figura 6.3.: Bandit in machine learning

Molti di queste applicazioni pratiche sono state trovate grazie al paper *«A survey on practical applications of multi-armed and contextual bandits»* di Bouneffouf, D., & Rish, I [BR19].

7. Considerazioni personali

Inizialmente mi ero trovato un po' disorientato di fronte al paper, mi sembrava troppo complesso analizzare una ricerca recente del 2019 in un ambito dell'intelligenza artificiale vasto quale è il *reinforcement learning*, ma dopo aver letto più volte il paper e soprattutto integrato con la lettura di alcuni related works si è chiarito molto. Fondamentale è stata l'analisi del paper *Batched Multi-armed Bandits Problem* di Zijun Gao, Yanjun Han, Zhimei Ren, Zhengqing Zhou [Gao+19], presentato tra le scelte disponibili nella sessione estiva. L'avevo già letto, mi incuriosiva ma per altri impegni ho dovuto rinunciare alla stesura dell'elaborato in estate. È stata una buona base per analizzare il problema dei *multi-armed bandits* come introduzione al paper a me assegnato.

È stato molto interessante seguire il processo di ricerca svolto dagli autori e come essi si approcciano con i lavori già esistenti in letteratura, basandosi sui precedenti risultati, producendo nuovi risultati migliorativi e aprendo nuove strade per possibili ricerche future. Ad esempio in questo paper, oltre a nuovi algoritmi proposti, vengono migliorati numerosi lower e upper bound su problemi già precedentemente analizzati (si veda la tabella 6.1). È maturata in me anche la consapevolezza che non esiste un algoritmo in grado di risolvere gran parte dei problemi rimanendo performante, ma in base al problema pratico che si deve risolvere bisogna scegliere la miglior strategia risolutiva o una tra le migliori sulla base della specifica istanza analizzata.

Concluso il lavoro, mi sento molto più sicuro nell'affrontare argomenti che inizialmente mi parevano troppo difficili e soddisfatto nel aver riuscito a produrre un elaborato, curato anche formalmente, analizzando nel modo più ordinato possibile il problema di identificare un numero k qualsiasi tra i migliori m arms in un *n-armed stochastic multi-armed bandit,* affrontato dal paper a me assegnato. Un peccato che non ci fossero porzioni di codice degli autori, avrei potuto replicare gli esperimenti e produrne di nuovi e apprendere anche in quel aspetto. A causa della mancanza dei codici, mi sono concentrato molto di più sull'analisi dei lavori correlati, nei possibili scenari applicativi e nel fare le **dimostrazioni teoriche** citate nella descrizione del paper che ho prodotto nelle appendici (vedi **appendici A, B e C**). Sicuramente, per l'anno prossimo, non è da escludere il fatto di basare il mio lavoro di tesi su un ambito dell'intelligenza artificiale, magari proprio il *reinforcement learning*, dopo quest'esperienza stimolante.

A. Lower bound sulla complessità del campione nel caso peggiore da risolvere

Teorema A.1. *[Lower bound per (k, m, n)] Sia $\mathcal{L}$ un algoritmo che risolve (k, m, n). Quindi, esiste un istanza $(A, n, m, k, \epsilon, \delta)$ con $0 < \epsilon \leq \frac{1}{\sqrt{32}}$, , $0 < \epsilon \leq \frac{e^{-1}}{14}$ e $n \geq 2m$, $1 \leq k \leq m$ su cui il numero di pulls attesi fatti da $\mathcal{L}$ sia almeno $\frac{1}{18375}\frac{1}{\epsilon^2}\frac{n}{m-k+1}ln\frac{\binom{m}{k-1}}{4\delta}$.*

La dimostrazione per il teorema A.1 è molto simile a quella del teorema 8 presente in [Kal+12] ma differisce per il fatto che qualsiasi delle m (ϵ, m)-optimal arms ha bisogno di essere ritornata al contrario di tutte le m.

A.1. Istanze del bandit

Ipotizziamo di avere un set di n arms $\mathcal{A} = \{0, 1, 2, \cdots, n-1\}$. Sia $I_0 = \{0, 1, 2, \cdots, m-k\}$ e $I_l = \{I : I \subseteq \{\mathcal{A} \setminus I_0\} \wedge |I| = l\}$. Inoltre per $I \subseteq \{m-k+1, m-k+2, \cdots, n-1\}$, definiamo:

$$\bar{I} = \{m - k + 1, m - k + 2, \cdots, n - 1\} \setminus I.$$

Con ogni $I \in I_{k-1} \cup I_m$ associamo un'istanza n-armed bandit $\mathcal{B}^I$, in cui ogni arm a produce un reward da una distribuzione di Bernoulli con media μ_a definita come:

$$\mu_a = \begin{cases} \frac{1}{2} & \text{se } a \in I_0 \\ \frac{1}{2} + 2\epsilon & \text{se } a \in I \\ \frac{1}{2} - 2\epsilon & \text{se } a \in \bar{I}. \end{cases} \tag{A.1}$$

Notare che tutte le istanze in $I_{k-1} \cup I_m$ hanno esattamente m (ϵ, m)-optimal arms. Per $I \in I_{k-1}$, tutte le arms in I_0 sono (ϵ, m)-optimal, ma per $I \in I_m$ non lo sono. Scriviamo $\mu(S)$ per indicare il multi-set composto dalle medie delle arms in $S \subseteq A$.

L'idea di base della dimostrazione è che senza un campionamento sufficiente di ogni arm, non è possibile identificare correttamente k delle (ϵ, m)-optimal arms con alta probabilità.

A.2. Limiti della probabilità d'errore

Dimostreremo il teorema facendo prima l'assunzione seguente, dimostreremo poi che porta a una contraddizione.

Assunzione A.1. *Assumiamo che esista un algoritmo $\mathcal{L}$ che risolve ogni istanza del problema (k, m, n) definite sull'istanza del bandit $\mathcal{B}^I$, $I \in I_{k-1}$, e incorre in una complessità del campione SC_I. Quindi per ogni $I \in I_{k-1}$, $\mathbb{E}SC_I < \frac{1}{18375} \cdot \frac{1}{\epsilon^2} \cdot \frac{n}{m-k+1} \ln\left(\frac{\binom{m}{m-k+1}}{4\delta}\right)$, per $0 < \epsilon \leq \frac{1}{\sqrt{32}}$, $0 < \delta \leq \frac{e^{-1}}{4}$, e $n \geq 2m$, dove $C = \frac{1}{18375}$.*

Per comodità, indichiamo con $\Pr_I$ la distribuzione di probabilità indotta dall'istanza bandit $\mathcal{B}^I$ e la possibile randomizzazione introdotta dall'algoritmo $\mathcal{L}$. Inoltre, sia $S_{\mathcal{L}}$ il set di arms ritornato come output da $\mathcal{L}$, e T_S il numero totale di volte che le arms in $S \subseteq \mathcal{A}$ vengono campionate finché $\mathcal{L}$ si ferma.

Quindi, come $\mathcal{L}$ risolve (k, m, n), per ogni $I \in I_{k-1}$

$$\Pr_I\{S_{\mathcal{L}} \subseteq I_0 \cup I\} \geq 1 - \delta. \tag{A.2}$$

Tuttavia, per ogni $I \in I_{k-1}$

$$\mathbb{E}_I[T_{\mathcal{A}}] \leq C \frac{n}{m-k+1} \ln\left(\frac{\binom{m}{m-k+1}}{4\delta}\right). \tag{A.3}$$

A.2.1. Cambiare $\Pr_I$ a $\Pr_{I \cup Q}$ dove $Q \in \bar{I}$ s.t. $|Q| = m - k + 1$

Consideriamo un $I \in I_{k-1}$ arbitrario ma fisso. Consideriamo una partizione fissa di A, in $\llcorner \frac{n}{m-k+1} \lrcorner$ subsets di dimensione $(m-k+1)$ ciascuno. Se l'assunzione A.1 è corretta, allora per l'istanza $\mathcal{B}^I$, ci sono almeno $\llcorner \frac{n}{4(m-k+1)} \rfloor - 1$ partizioni $B \subset \bar{I}$, tali che $\mathbb{E}_I[T_B] \geq \frac{4C}{\epsilon^2} \ln\left(\frac{1}{4\delta}\right)$. Ora, come $\llcorner \frac{n-m}{m-k+1} \lrcorner - \left(\llcorner \frac{n}{4(m-k+1)} \lrcorner - 1\right) \geq \llcorner \frac{n}{4(m-k+1)} \lrcorner + 1 > 0$; tuttavia, esiste almeno un subset $Q \in \bar{I}$ tale che $|Q| = m-k+1$, e $\mathbb{E}_I[T_Q] < \frac{4C}{\epsilon^2} \ln\left(\frac{\binom{m}{m-k+1}}{4\delta}\right)$. Definito $T^* = \frac{16C}{\epsilon^2} \ln\left(\frac{\binom{m}{m-k+1}}{4\delta}\right)$. Ora usando la *Markov's inequality* abbiamo:

$$\Pr_I\{T_Q \geq T^*\} < \frac{1}{4}. \tag{A.4}$$

Sia $\Delta = 2\epsilon T^* + \sqrt{T^*}$ e inoltre sia K_Q il numero totale di reward ottenuti da Q.

Lemma A.1. *Se $I \in I_{k-1}$ e $Q \in \bar{I}$ s.t. $|Q| = m - k + 1$, allora*

$$\Pr_I\left\{T_Q \leq T^* \wedge K_Q \leq \frac{T_Q}{2} - \Delta\right\} \leq \frac{1}{4}.$$

Dimostrazione. Sia $K_Q(t)$ la somma totale ottenuta da Q alla fine del turno t. Quanto a $\mathcal{B}^{I_0}$, $\forall j \in Q\ \mu_j = 1/2 - 2\epsilon$, quindi selezionando e tirando un braccio ad ogni prova da Q seguendo qualsiasi regola (deterministica o probabilistica) è equivalente a selezionare una singola arm da Q per volta e successivamente fare i pulls su di essa. Quindi qualsiasi sia la strategia di tirare un arma ad ogni turno da Q, il reward atteso da ogni pull è $1/2 - 2\epsilon$. Sia r_i la distribuzione i.i.d. del reward ottenuto dal i^{th} turno. Quindi $K_Q(t) = \sum_{i=1}^{t} r_i$ e $Var\left[r_i\right] = \left(\frac{1}{2} - 2\epsilon\right)\left(\frac{1}{2} + 2\epsilon\right) = \left(\frac{1}{4} - 4\epsilon^2\right) < \frac{1}{4}$. Visto che $\forall i : 1 \leq i \leq t$, r_i sono i.i.d., otteniamo $Var[K_Q(t)] = \sum_{i=1}^{t} Var(r_i) < \frac{t}{4}$. Ora possiamo scrivere::

$$\begin{aligned} & \Pr_I\left\{\min_{1\leq t\leq T^*}\left(K_Q(t) - t\left(\frac{1}{2} - 2\epsilon\right)\right) \leq -\sqrt{T^*}\right\} \\ \leq\ & \Pr_I\left\{\max_{1\leq t\leq T^*}\left|K_Q(t) - t\left(\frac{1}{2} - 2\epsilon\right)\right| \geq \sqrt{T^*}\right\} \\ \leq\ & \frac{Var[K_Q(T^*)]}{T^*} < \frac{1}{4}, \end{aligned} \tag{A.5}$$

in cui abbiamo usato la *disuguaglianza di Kolmogorov.* □

Lemma A.2. *Sia $I \in I_{k-1}$ e $Q \in I_{m-k+1}$ tale che $Q \subseteq \bar{I}$, e sia W una qualche sequenza di reward ottenuti da un singolo run dell'algoritmo $\mathcal{L}$ su $\mathcal{B}^I$ tale che $T_Q \leq T^*$ e $K_Q \geq \frac{T_Q}{2} - \Delta$, allora:*

$$\Pr_{I\cup Q}\{W\} > \Pr_I\{W\} \cdot \exp(-32\epsilon\Delta). \tag{A.6}$$

Dimostrazione. Riprendendo il fatto che tutte le arms in Q hanno la stessa media. Quindi, se scelta una ad ogni turno (seguendo una qualsiasi strategia), il reward atteso ad ogni turno rimane lo stesso. Quindi la probabilità di ottenere una sequenza dei reward generata da Q è indipendente dalla strategia di campionamento. Nuovamente come le arms in Q hanno una media alta in $\mathcal{B}^Q$, la probabilità di ottenere una serquenza di reward decresce in modo monotono man mano che le 1-rewards per $\mathcal{B}^{I_0}$ diventano più piccole. Perciò abbiamo:

$$\begin{aligned} \Pr_{I\cup Q}\{W\} &= \Pr_I\{W\}\frac{\left(\frac{1}{2} + 2\epsilon\right)^{K_Q}\left(\frac{1}{2} - 2\epsilon\right)^{T_Q - K_Q}}{\left(\frac{1}{2} - 2\epsilon\right)^{K_Q}\left(\frac{1}{2} + 2\epsilon\right)^{T_Q - K_Q}} \\ &\geq \Pr_I\{W\}\frac{\left(\frac{1}{2} + 2\epsilon\right)^{\left(\frac{T_Q}{2} - \Delta\right)}\left(\frac{1}{2} - 2\epsilon\right)^{\left(\frac{T_Q}{2} + \Delta\right)}}{\left(\frac{1}{2} - 2\epsilon\right)^{\left(\frac{T_Q}{2} - \Delta\right)}\left(\frac{1}{2} + 2\epsilon\right)^{\left(\frac{T_Q}{2} + \Delta\right)}} \\ &= \Pr_I\{W\} \cdot \left(\frac{\frac{1}{2} - 2\epsilon}{\frac{1}{2} + 2\epsilon}\right)^{2\Delta} \\ &> \Pr_I\{W\} \cdot \exp(-32\epsilon\Delta)\left[\text{ for } 0 < \epsilon \leq \frac{1}{\sqrt{32}}\right]. \end{aligned}$$

□

Lemma A.3. *Se* (A.4) *vale per un* $I \in I_{k-1}$ *e* $Q \in I_{m-k+1}$ *tale che* $Q \subseteq \bar{I}$*, e se* $\mathcal{W}$ *è il set di tutte le sequenze di reward possibili* W*, ottenute dall'algoritmo* $\mathcal{L}$ *su* $\mathcal{B}^I$*, allora* $\Pr_{I \cup Q}\{\mathcal{W}\} > \left(\Pr_I\{\mathcal{W}\} - \frac{1}{2}\right) \cdot 4\delta$*. In particolare:*

$$\Pr_{I \cup Q}\{S_{\mathcal{L}} \subseteq I_0 \cup I\} > \frac{\delta}{\binom{m}{m-k+1}}. \tag{A.7}$$

Dimostrazione. Sia per qualche sequenza fissa dei rewards W, T_Q^W e K_Q^W rispettivamente indichiamo il numero totale di campioni ricevuti dalle arms in Q e il numero totale di 1-rewards ottenuti prima che l'algoritmo $\mathcal{L}$ si fermi. Allora:

$$\begin{aligned}
&\Pr_{I \cup Q}\{W\} = \Pr_{I \cup Q}(W : W \in \mathcal{W}) \\
&\geq \Pr_{I \cup Q}\left\{W : W \in \mathcal{W} \bigwedge T_Q^W \leq T^* \bigwedge K_Q^W \geq \frac{T_Q^W}{2} - \Delta\right\} \\
&> \Pr_{I}\left\{W : W \in \mathcal{W} \bigwedge T_Q^W \leq T^* \bigwedge K_Q^W \geq \frac{T_Q^W}{2} - \Delta\right\} \cdot \exp(-32\epsilon\Delta) \\
&\geq \left(\Pr_I\left\{W : W \in \mathcal{W} \bigwedge T_Q^W \leq T^*\right\} - \frac{1}{4}\right) \cdot \exp(-32\epsilon\Delta) \\
&\geq \left(\Pr_I\{\mathcal{W}\} - \frac{1}{2}\right) \cdot \frac{4\delta}{\binom{m}{m-k+1}} \quad \text{for } C = \frac{1}{18375},\ \delta < \frac{e^{-1}}{4}.
\end{aligned}$$

Qui sopra, il terzo, il quarto e l'ultimo passaggio sono ottenuti usando rispettivamente il lemma A.2, il lemma A.1 e l'equazione (A.4). La disuguaglianza (A.7) è ottenuta usando la disuguaglianza (A.2), come $\Pr_I\{S_{\mathcal{L}} \in I_0\} > 1 - \delta \geq 1 - \frac{e^{-1}}{4} > \frac{3}{4}$. □

A.2.2. Sommando su I_{k-1} e I_m

Ora, sommiamo la probabilità degli errori rispetto a tutte le istanze in I_{k-1} e I_m. Se l'assunzione A.1 è vera, usando il principio *pigeon-hole principle,* mostriamo che esiste una qualche istanza per cui la probabilità d'errore è maggiore di δ.

$$
\begin{aligned}
&\sum_{J\in I_m} \Pr_J\{S_{\mathcal{L}} \nsubseteq J\} \\
&\geq \sum_{J\in I_m} \sum_{\substack{J'\subset J \\ :|J'|=m-k+1}} \Pr_J\{S_{\mathcal{L}} \subseteq \{J \setminus J'\} \cup I_0\} \\
&\geq \sum_{J\in I_m} \sum_{\substack{J'\subset J \\ :|J'|=m-k+1}} \Pr_J\{\exists a \in I_0 : S_{\mathcal{L}} = \{J \setminus J'\} \cup \{a\}\} \\
&= \sum_{J\in I_m} \sum_{\substack{J'\subset J \\ :|J'|=m-k+1}} \sum_{I\in I_{k-1}} \mathbb{I}[I \cup J' = J] \cdot \Pr_J\{S_{\mathcal{L}} \subseteq I \cup I_0\} \\
&= \sum_{J\in I_m} \sum_{\substack{J'\subset \mathcal{A}\setminus I_0 \\ :|J'|=m-k+1}} \sum_{I\in I_{k-1}} \mathbb{I}[I \cup J' = J] \cdot \Pr_J\{S_{\mathcal{L}} \subseteq I \cup I_0\} \\
&= \sum_{J\in I_m} \sum_{I\in I_{k-1}} \sum_{\substack{J'\subset \mathcal{A}\setminus I_0 \\ :|J'|=m-k+1}} \mathbb{I}[I \cup J' = J] \cdot \Pr_J\{S_{\mathcal{L}} \subseteq I \cup I_0\} \\
&= \sum_{I\in I_{k-1}} \sum_{J\in I_m} \sum_{\substack{J'\subset \bar{I} \\ :|J'|=m-k+1}} \mathbb{I}[I \cup J' = J] \cdot \Pr_J\{S_{\mathcal{L}} \subseteq I \cup I_0\} \\
&= \sum_{I\in I_{k-1}} \sum_{\substack{J'\subset \bar{I} \\ :|J'|=m-k+1}} \sum_{J\in I_m} \mathbb{I}[I \cup J' = J] \cdot \Pr_J\{S_{\mathcal{L}} \subseteq I \cup I_0\} \\
&= \sum_{I\in I_{k-1}} \sum_{\substack{J'\subset \bar{I} \\ :|J'|=m-k+1}} \Pr_{I\cup J'}\{S_{\mathcal{L}} \subseteq I \cup I_0\}
\end{aligned}
$$

Riprendendo il fatto che $\forall I \in I_{k-1}$ esista un set $Q \subset \mathcal{A} \setminus \{I \cup I_0\} : |Q| = (m-k+1)$, tale che $T_Q < T^*$. Allora,

$$
\begin{aligned}
&\sum_{J \in I_m} \Pr_J\{S_{\mathcal{L}} \nsubseteq J\} \\
&\geq \sum_{I \in I_{k-1}} \sum_{\substack{J' \subset \bar{I} \\ :|J'| = m-k+1}} \Pr_{I \cup J'}\{S_{\mathcal{L}} \subseteq I \cup I_0\} \\
&> \sum_{I \in I_{k-1}} \sum_{\substack{J' \subset \bar{I} \\ :|J'| = m-k+1}} \frac{\delta}{\binom{m}{m-k+1}} \\
&\geq \sum_{I \in I_{k-1}} \binom{n-m}{m-k+1} \cdot \frac{\delta}{\binom{m}{m-k+1}} \\
&\geq \binom{n-(m-k+1)}{k-1} \cdot \binom{n-m}{m-k+1} \cdot \frac{\delta}{\binom{m}{m-k+1}} \\
&= \binom{n-(m+k-1)}{m} \delta \\
&= |I_m| \delta.
\end{aligned}
$$

Quindi abbiamo ottenuto una contraddizione all'assunzione A.1, quindi abbiamo dimostrato il teorema.

B. Analisi di *LUCB-k-m*

Sia al tempo t, $\hat{p}_a^t$ la media empirica dell'arm $a \in \mathcal{A}$, e u_a^t il numero di volte che l'arm a è stata tirata fino al tempo t escluso. Per un dato $\delta \in (0, 1]$, definiamo $\beta(u_a^t, t, \delta) = \sqrt{\frac{1}{2u_a^t} \ln \frac{k_1 n t^4}{\delta}}$, dove $k_1 = 5/4$. Definiamo un upper e lower bound di confidenza sulla stima della media reale dell'arm $a \in \mathcal{A}$ come $ucb(a, t) = \hat{p}_a + \beta(u_a^t, t, \delta)$, e $lcb(a, t) = \hat{p}_a - \beta(u_a^t, t, \delta)$ rispettivamente.

Per analizzare la complessità del campione, prima di tutto definiamo qualche evento, almeno uno dei qualche deve accadere se l'algoritmo non si ferma al turno t.

Definizione B.1. (Probable Events) Siano $a, b \in \mathcal{A}$, tali che $\mu_a > \mu_b$. Durante l'esecuzione dell'algoritmo, qualsiasi dei seguenti cinque eventi possono accadere:

1. La media empirica di un arm a può cadere oltre l'upper o il lower bound di confidenza; definiamo l'evento in questo modo:

$$CROSS_a^t = \{ucb(a, t) < \mu_a \vee lcb(a, t) > \mu_a\}.$$

2. La media empirica di un arm a può essere meno di un arm b; definiamo l'evento in questo modo:

$$ErrA(a, b, t) = \{\hat{p}_a^t < \hat{p}_b^t\}.$$

3. Il lower e l'upper bound di confidenza di un arm a può cadere sotto a quella arms b; definiamo l'evento in questo modo:

$$ErrL(a, b, t) = \{lcb(a, t) < lcb(b, t)\},$$
$$ErrU(a, b, t) = \{ucb(a, t) < ucb(b, t)\}.$$

4. Se gli intervalli di confidenza di un arm sono sotto un certo raggio (detto d), definiamo l'evento in questo modo:

$$NEEDY_a^t(d) = \{\{lcb(a, t) < \mu_a - d\} \vee \{ucb(a, t) > \mu_a + d\}\}.$$

Mostriamo che ogni arm a, se campionata sufficientemente, tale che $u_a^t \geq u^*(a, t)$, allora l'avvenimento di un qualsiasi Probable Events implica l'accadimento di $CROSS_a^t$. Prima abbiamo mostrato che se $CROSS_a^t$ non accade per ogni $a \in A$, allora l'accadimento di uno dei qualsiasi Probable Events implica l'accadimento di $NEEDY_a^t(\cdot)$ o $NEEDY_b^t(\cdot)$.

Lemma B.1. *[Esprimere i PROBABLE EVENTS in termini di $NEEDY_a^t$ e $CROSS_a^t$] Per dimostrare ciò $\{\neg CROSS_a^t \wedge \neg CROSS_b^t\} \wedge \{ErrA(a,b,t) \vee ErrU(a,b,t) \vee ErrL(a,b,t)\} \implies \{NEEDY_a^t\left(\frac{\Delta_{ab}}{2}\right) \vee NEEDY_b^t\left(\frac{\Delta_{ab}}{2}\right)\}$.*

Dimostrazione. Procediamo con la dimostrazione:

ErrA$(\mathbf{a},\mathbf{b},\mathbf{t})$: Per dimostrare ciò $\neg\{CROSS_a^t \vee CROSS_b^t\} \wedge ErrA(a,b,t) \implies NEEDY_a^t\left(\frac{\Delta_{ab}}{2}\right) \vee NEEDY_b^t\left(\frac{\Delta_{ab}}{2}\right)$.

$$
\begin{aligned}
&ErrA(a,b,t) \implies \hat{p}_a^t < \hat{p}_b^t \\
&\implies \hat{p}_a^t - (p_a - \beta(u_a^t,t,\delta)) < \hat{p}_b^t - (p_b + \beta(u_b^t,t,\delta) + \\
&(\beta(u_a^t,t,\delta) + \beta(u_b^t,t,\delta)) - \Delta_{ab}/2) \\
&\implies NEEDY_a^t\left(\frac{\Delta_{ab}}{2}\right) \vee NEEDY_b^t\left(\frac{\Delta_{ab}}{2}\right).
\end{aligned}
$$

ErrU$(\mathbf{a},\mathbf{b},\mathbf{t})$: Per dimostrare ciò $\neg\{CROSS_a^t \vee CROSS_b^t\} \wedge ErrU(a,b,t) \implies NEEDY_b^t\left(\frac{\Delta_{ab}}{2}\right)$.
Assumendo $\neg CROSS_a^t \wedge \neg CROSS_b^t$ otteniamo:

$$
\begin{aligned}
&ErrU(a,b,t) \implies \{ucb(b,t) > ucb(a,t)\} \\
&\implies \{\hat{p}_b^t + \beta(u_b^t,t,\delta) > \hat{p}_a^t + \beta(u_a^t,t,\delta)\} \\
&\implies \{\hat{p}_b^t > \mu_b + \beta(u_b^t,t,\delta)\} \vee \{\hat{p}_a^t < \mu_a - \beta(u_a^t,t,\delta)\} \vee \\
&\quad \{2\beta(u_b^t,t,\delta) > \Delta_{ab}\} \\
&\implies NEEDY_b^t\left(\frac{\Delta_{ab}}{2}\right).
\end{aligned}
$$

ErrL$(\mathbf{a},\mathbf{b},\mathbf{t})$: Per dimostrare ciò $\neg\{CROSS_a^t \vee CROSS_b^t\} \wedge ErrL(a,b,t) \implies NEEDY_a^t\left(\frac{\Delta_{ab}}{2}\right)$.
Assumendo $\neg CROSS_a^t \wedge \neg CROSS_b^t$ otteniamo

$$
\begin{aligned}
&ErrL(a,b,t) \implies \{lcb(b,t) > lcb(a,t)\} \\
&\implies \{\hat{p}_b^t - \beta(u_b^t,t,\delta) > \hat{p}_a^t - \beta(u_a^t,t,\delta)\} \\
&\implies \{\hat{p}_b^t > \mu_b + \beta(u_b^t,t,\delta)\} \vee \{\hat{p}_a^t < \mu_a - \beta(u_a^t,t,\delta)\} \vee \\
&\quad \{2\beta(u_a^t,t,\delta) > \Delta_{ab}\} \\
&\implies NEEDY_a^t\left(\frac{\Delta_{ab}}{2}\right).
\end{aligned}
$$

□

Mostriamo che data una soglia d, se un arm a è campionata sufficientemente, tale che $\beta(u_a^t,t,\delta) \leq \frac{d}{2}$, allora $NEEDY_a^t$ deduce $CROSS_a^t$.

Lemma B.2. *Per ogni* $a \in A$, $\{NEEDY_a^t(d) | \beta(u_a^t, t, \delta) < d/2\} \implies CROSS_a^t$.

Dimostrazione. Innanzitutto mostriamo che $\{lcb(a,t) < \mu_a - d | \beta(u_a^t, t, \delta) < d/2\} \implies CROSS_a^t$,

$$\begin{aligned}
&\{lcb(a,t) < \mu_a - d | \beta(u_a^t, t, \delta) < d/2\} \\
&\implies \{\hat{p}_a^t - \beta(u_a^t, t, \delta) < \mu_a - d | \beta(u_a^t, t, \delta) < d/2\} \\
&\implies \{\hat{p}_a^t < \mu_a - d + \beta(u_a^t, t, \delta) | \beta(u_a^t, t, \delta) < d/2\} \\
&\implies \{\hat{p}_a^t < \mu_a - d/2 | \beta(u_a^t, t, \delta) < d/2\} \\
&\implies CROSS_a^t. \qquad \text{(B.1)}
\end{aligned}$$

L'altra direzione segue una dimostrazione simile. □

Dalla stessa definizione di intervallo di confidenza, ad ogni turno t, la probabilità che la media empirica di un arm cada oltre è molto bassa. In altre parole, la probabilità dell'evento $CROSS_a^t$ è molto bassa per ogni t e $a \in A$.

Lemma B.3. *[Upper bound alla probabilità di* $CROSS_a^t$*]* $\forall a \in \mathcal{A}$ *e* $\forall t \geq 0$, $\Pr\{CROSS_a^t\} \leq \frac{\delta}{knt^4}$. *Quindi,* $P\left[\exists t \geq 0 \wedge \exists a \in \mathcal{A} : CROSS_a^t | u_a^t \geq 0\right] \leq \frac{\delta}{k_1 t^3}$.

Dimostrazione. $\Pr\{CROSS_a^t\}$ è limitato superiormente dalla *disuguaglianza di Hoeffding*, e il prossimo paragrafo lo dimostra prendendo l'union bound su tutte le arms e t. □

Ora, riprendendo la definizione di h_*^t, e l_*^t dall'algoritmo B.1, presentiamo la logica chiave alla base dell'analisi di *LUCB-k m*. L'idea è mostrare che se l'algoritmo non si ferma, allora uno dei Probable Events deve accadere. Quindi usando il lemma B.1, il lemma B.2 e il lemma B.3 dimostriamo che oltre un certo numero di turni, la probabilità che *LUCB-k m* continui è molto bassa. Infine, usando l'argomento alla base del principio *pigeon-hole*, simile al lemma 5 di [Kal11], stabiliamo un upper bound sulla complessità del campione. Di seguito presentiamo la logica di base che mostra la necessità del verificarsi di uno dei Probable Events fintanto che l'algoritmo non si è fermato.

Algoritmo B.1 Caso 1: $h_*^t \in B_1 \wedge l_*^t \in B_1$

if $\exists b_3 \in A_1^t \cap B_3$ **then**
  Then $ErrL(h_*^t, b_3, t)$ si è verificato.
else
  $\exists b_3 \in A_2^t \cap B_3$
  Then $ErrA(h_*^t, b_3, t)$ si è verificato.
endif

Algoritmo B.2 Caso 2: $h_*^t \in B_1 \wedge l_*^t \in B_2$

if $\exists b_3 \in A_1^t \cap B_3$ **then**
  Then $ErrL(h_*^t, b_3^t, t)$ si è verificato.
else
  $\exists b_3 \in A_2^t \cap B_3$
  if $\Delta_{h_*^t l_*^t} \geq \frac{\Delta_{h_*^t}}{2}$ **then**
    Then $NEEDY_{h_*^t}^t(\Delta_{h_*^t}/4) \vee NEEDY_{l_*^t}^t(\Delta_{h_*^t}/4)$ si è verificato.
  else
    Then $ErrL(l_*^t, b_3^t, t)$ si è verificato.
  endif
endif

Algoritmo B.3 Caso 3: $h_*^t \in B_1 \wedge l_*^t \in B_3$

Then $NEEDY_{h_*^t}^t(\Delta_{h_*^t}/4) \vee NEEDY_{l_*^t}^t(\Delta_{l_*^t}/4)$ si è verificato.

Algoritmo B.4 Caso 4: $h_*^t \in B_2 \wedge l_*^t \in B_1$

if $\Delta_{h_*^t l_*^t} \geq \frac{\Delta_{h_*^t}}{2}$ **then**
  Then $ErrA(l_*^t, h_*^t, t)$ si è verificato.
else
  if $\exists b_3 \in A_1^t \cap B_3$ **then**
    Then $ErrL(h_*^t, b_3^t, t)$ si è verificato.
  else
    $\exists b_3 \in A_2^t \cap B_3$
    $\therefore ErrA(l_*^t, b_3, t)$ si è verificato.
  endif
endif

Algoritmo B.5 Caso 5a: $h_*^t \in B_2 \wedge l_*^t \in B_2$ *e* $\Delta_{h_*^t l_*^t} > 0$

Qui, $\exists b_1 \in (A_2^t \cup A_3^t) \cap B_1$ e $\exists b_3 \in (A_1^t \cup A_2^t) \cap B_3$

if $\Delta_{h_*^t l_*^t} < \Delta_{h_*^t}/2$ **then**
  if $\Delta_{b_1 h_*^t} > \Delta_{b_1}/4$ **then**
    if $b_1 \in A_2^t) \cap B_1$ **then**
      $ErrA(b_1, h_*^t, t)$ si è verificato.
    else
      $b_1 \in A_3^t \cap B_1$
      $ErrU(b_1, l_*^t, t)$ si è verificato.
    endif
  else
    $\Delta_{b_1 h_*^t} \leq \Delta_{b_1}/4$ e quindi $\Delta_{l_*^t b_3} \geq \Delta_{l_*^t}/4$
    if $b_3 \in A_2^t \cap B_3$ **then**
      $ErrA(l_*^t, b_3, t)$ si è verificato.
    else
      $b_3 \in A_1^t \cap B_3$
      $ErrL(h_*^t, b_3, t)$ si è verificato.
    endif
  endif
else
  $|\Delta_{h_*^t l_*^t}| > \Delta_{h_*^t}/2$
  $NEEDY_{h_*^t}^t(\Delta_{h_*^t}/4) \vee NEEDY_{l_*^t}^t(\Delta_{h_*^t}/4)$ si è verificato.
endif

Algoritmo B.6 Caso 5b: $h_*^t \in B_2 \wedge l_*^t \in B_2 and \Delta_{h_*^t l_*^t} \leq 0$

Qui, $\exists b_1 \in (A_2^t \cup A_3^t) \cap B_1$ e $\exists b_3 \in (A_1^t \cup A_2^t) \cap B_3$

if $|\Delta_{h_*^t l_*^t}| < \Delta_{h_*^t}/2$ **then**
  if $\Delta_{b_1 l_*^t} > \Delta_{b_1}/4$ **then**
    if $b_1 \in A_2^t \cap B_1$ **then**
      $ErrA(b_1, h_*^t, t)$ si è verificato.
    else
      $b_1 \in A_3^t \cap B_1$
      $ErrU(b_1, l_*^t, t)$ si è verificato.
    endif
  else
    $\Delta_{b_1 l_*^t} \leq \Delta_{b_1}/4$ e quindi $\Delta_{h_*^t b_3} \geq \Delta_{h_*^t}/4$
    if $b_3 \in A_2^t \cap B_3$ **then**
      $ErrA(l_*^t, b_3, t)$ si è verificato.
    else
      $b_3 \in A_1^t \cap B_3$
      $ErrL(h_*^t, b_3, t)$ si è verificato.
    endif
  endif
else
  $|\Delta_{h_*^t l_*^t}| > \Delta_{h_*^t}/2$
  $NEEDY_{h_*^t}^t(\Delta_{h_*^t}/4) \vee NEEDY_{l_*^t}^t(\Delta_{h_*^t}/4)$ si è verificato.
endif

Algoritmo B.7 Caso 6: $h_*^t \in B_2 \wedge l_*^t \in B_3$

if $\Delta_{h_*^t l_*^t} \geq \frac{\Delta_{l_*^t}}{2}$ **then**
  Then $NEEDY_{h_*^t}^t(\Delta/4) \vee NEEDY_{l_*^t}^t(\Delta_{l_*^t}/4)$ si è verificato.
else
  $\Delta_{h_*^t l_*^t} < \frac{\Delta_{l_*^t}}{2}$
  $\forall b_1 \in \{A_2^t \cup A_3^t\} \cap B_1$, $\Delta_{b_1 h_*^t} > \frac{\Delta_{b_1}}{2}$
  if $\exists b_1 \in A_2^t \cap B_1$ **then**
    $ErrA(b_1, h_*^t, t)$ si è verificato.
  else
    $\exists b_1 \in A_3^t \cap B_1$
    Then $ErrU(b_1^t, l_*^t, t)$ si è verificato.
  endif
endif

Algoritmo B.8 Caso 7: $h_*^t \in B_3 \wedge l_*^t \in B_1$

$ErrA(l_*^t, h_*^t, t)$ si è verificato.

Algoritmo B.9 Caso 8: $h_*^t \in B_3 \wedge l_*^t \in B_2$

if $\Delta_{h_*^t l_*^t} \geq \frac{\Delta_{h_*^t}}{2}$ **then**
  $ErrA(l_*^t, h_*^t, t)$ si è verificato.
else
  $\Delta_{h_*^t l_*^t} < \frac{\Delta_{h_*^t}}{2}$
  $\forall b_1 \in \{A_2^t \cup A_3^t\} \cap B_1$, $\Delta_{b_1 l_*^t} > \frac{\Delta_{b_1}}{2}$
  if $\exists b_1 \in A_2^t \cap B_1$ **then**
    $ErrA(b_1, h_*^t, t)$ si è verificato.
  else
    $\exists b_1 \in A_3^t \cap B_1$
    $ErrU(b_1, l_*^t, t)$ si è verificato.
  endif
endif

Algoritmo B.10 Caso 9: $h_*^t \in B_3 \wedge l_*^t \in B_3$

$\exists b_1 \in \{A_2^t \cup A_3^t\} \cap B_1$
if $\exists b_1 \in A_2^t \cap B_1$ **then**
  $ErrA(b_1, h_*^t, t)$ si è verificato.
else
  $\exists b_1 \in A_3^t \cap B_1$
  $ErrA(b_1, l_*^t, t)$ si è verificato.
endif

Lemma B.4. *Se* $T = CH_\epsilon \ln\left(\frac{H_\epsilon}{\delta}\right)$*, allora per* $C \geq 2732$*, vale:*

$$T > 2 + 2\sum_{a\in\mathcal{A}} u^*(a,T).$$

Dimostrazione. La dimostrazione è presa dall'appendice B.3 di [Kal11].

$$\begin{aligned}
&2 + 2\sum_{a\in\mathcal{A}} u^*(a,T) = 2 + 64\sum_{a\in\mathcal{A}} \ulcorner \frac{1}{\max(\Delta_a,(\epsilon/2))^2}\ln\frac{knt^4}{\delta}\urcorner \\
&\leq 2 + 64n + 64H_\epsilon \ln\frac{knT^4}{\delta} \\
&= 2 + 64n + 64H_\epsilon \ln k + 64H_\epsilon \ln\frac{n}{\delta} + 256H_\epsilon \ln T \\
&< (66 + 64\ln k)H_\epsilon + 64H_\epsilon \ln\frac{n}{\delta} + 256H_\epsilon\left[\ln C + \ln H_\epsilon + \ln\ln\frac{H_\epsilon}{\delta}\right] \\
&< (66 + 64\ln k)H_\epsilon + 64H_\epsilon \ln\frac{n}{\delta} + 256H_\epsilon\left[\ln C + \ln H_\epsilon + \ln\ln\frac{H_\epsilon}{\delta}\right] \\
&< 130H_\epsilon + 64H_\epsilon \ln\frac{n}{\delta} + 256H_\epsilon\left[\ln C + \ln H_\epsilon + \ln\frac{H_\epsilon}{\delta}\right] \\
&< 130H_\epsilon + 64H_\epsilon \ln\frac{H_\epsilon}{\delta} + 256H_\epsilon\left[\ln C + 2\ln\frac{H_\epsilon}{\delta}\right] \\
&< (706 + 256\ln C)H_\epsilon \ln\frac{H_\epsilon}{\delta} < CH_\epsilon \ln\frac{H_\epsilon}{\delta} \text{ [For } C \geq 2732\text{]}.
\end{aligned}$$

□

Lemma B.5. *Sia* $T^* = \ulcorner 2732 H_\epsilon \ln\left(\frac{H_\epsilon}{\delta}\right)\urcorner$. *Per ogni* $T > T_1^*$, *la probabilità che l'algoritmoB.1 non termini dopo* T *turni di campionamento e circa* $\frac{8\delta}{T^2}$.

Dimostrazione. Sia $\bar{T} = \frac{T}{2}$ definiamo due eventi per $\bar{T} \leq t \leq T-1$: $E^{(1)} = \exists a \in \mathcal{A} : CROSS_a^t$ e $E^{(2)} = \exists NEEDY_a^t\left(\frac{\Delta_a}{4}\right)$. Se l'algoritmo si ferma per $t < \bar{T}$, non c'è nulla da dimostrare. Al contrario, se l'algoritmo non ha terminato dopo $t > \bar{T}$, né $E^{(1)}$ né $E^{(2)}$ si sono verificati. Sia N_{rounds} il numero di round richiesti prima di $\bar{T}$, possiamo limitarlo superiormente come:

$$
\begin{aligned}
N_{rounds} &= \sum_{t=\bar{T}} \left\{ \mathbb{I}\left[NEEDY_{h_*^t}^t\left(\frac{\Delta_{h_*^t}}{4}\right) \vee NEEDY_{m_*^t}^t\left(\frac{\Delta_{m_*^t}}{4}\right) \vee NEEDY_{l_*^t}^t\left(\frac{\Delta_{l_*^t}}{4}\right)\right]\right\} \\
&\leq \sum_{\bar{T}}^{T-1} \sum_{a\in\mathcal{A}} \mathbb{I}\left[a \in \{h_*^t, m_*^t, l_*^t\} \wedge NEEDY_a^t\left(\frac{\Delta_a}{4}\right)\right] \\
&= \sum_{\bar{T}}^{T-1} \sum_{a\in\mathcal{A}} \mathbb{I}[a \in \{h_*^t, m_*^t, l_*^t\} \wedge (u_a^t < u^*(a,t))] \\
&\leq \sum_{\bar{T}}^{T-1} \sum_{a\in\mathcal{A}} \mathbb{I}[a \in \{h_*^t, m_*^t, l_*^t\} \wedge (u_a^t < u^*(a,t))] \\
&\leq \sum_{a\in\mathcal{A}} \sum_{\bar{T}}^{T-1} \mathbb{I}[(a \in \{h_*^t, m_*^t, l_*^t\}) \wedge (u_a^t < u^*(a,t))] \\
&\leq \sum_{a\in\mathcal{A}} u^*(a,t).
\end{aligned}
$$

Usando il lemma B.4, $T \geq T^* \Rightarrow T > 2 + 2\sum_{a\in\mathcal{A}} u^*(a,t)$. Quindi, se né $E^{(1)}$ né $E^{(2)}$ si verificano allora l'algoritmo continua ad essere eseguito per al massimo $\bar{T} + N_{rounds} \leq \ulcorner T/2 \urcorner + \sum_{a\in\mathcal{A}} 16 u^*(a,t) < T$ numero di turni..

La probabilità che l'algoritmo non si fermi entro T turni, è limitata superiormente da $P[E^{(1)} \vee E^{(2)}]$. Applicando il lemma B.2 e il lemma B.3,

$$
P[E^{(1)} \vee E^{(2)}] \leq \sum_{t=\bar{T}}^{T-1} \left(\frac{\delta}{k_1 t^3} + \frac{\delta}{k t^4}\right) \leq \sum_{t=\bar{T}}^{T-1} \frac{\delta}{k_1 t^3}\left(1 + \frac{2}{t}\right) \leq \left(\frac{T}{2}\right) \frac{8\delta}{k_1 T^3}\left(1 + \frac{4}{T}\right) < \frac{8\delta}{T^2}.
$$

□

Teorema B.1. *[Complessità del campione attesa di LUCB-k-m] LUCB-k-m risolve il problema (k,m, n) usando una complessità del campione attesa limitata superiormente da* $O(H_\epsilon log\frac{H_\epsilon}{\delta})$.

Usando il lemma B.4, e il lemma B.5 la complessità del campione attesa dell'algoritmo B.1 può essere limitata superiormente in questo modo:

$$
E[SC] \leq 2\left(T_1^* + \sum_{t=T_1^*}^{\infty} \frac{8\delta}{T^2}\right) \leq 5464 \cdot \left(H_\epsilon \ln\left(\frac{H_\epsilon}{\delta}\right)\right) + 32. \tag{B.2}
$$

C. Dimostrazione del teorema 4.6

L'algoritmo C.1 descrive OptQP. Usa $\mathcal{P}_2$ [CK17] con Median Elimination come subroutine all'interno di $\mathcal{P}_2$ per selezionare un $[\epsilon, \rho]$-optimal arm con confidenza $1-\delta'$. Abbiamo assunto $\delta' = 1/4$, in pratica si può scegliere un qualsiasi valore sufficientemente piccolo per esso, tanto influenzerà semplicemente la costante moltiplicativa nel limite superiore.

Algoritmo C.1 Algoritmo OptQP

Input: $\mathcal{A}, \epsilon, \delta$, e OptQF.
Output: Una singola $[\epsilon, \rho]$-optimal arm.

Sia $\delta' = 1/4$, $u = \ulcorner \frac{1}{2(0.5-\delta')^2} \cdot \log \frac{2}{\delta} \urcorner = \ulcorner 8 \log \frac{2}{\delta} \urcorner$

Teorema C.1. *[Correttezza e complessità del campione di OptQP] Se OptQF esiste, allora OptQP risolve Q-P, con una complessità del campione compresa in* $\Theta\left(\frac{1}{\rho\epsilon^2} \log \frac{1}{\delta} + \gamma(\cdot)\right)$.

Dimostrazione. Prima di tutto dimostriamo la correttezza e poi l'upper bound sulla complessità del campione.

Correttezza

Dimostrazione. Inizialmente notiamo che ogni copia degli outputs di $\mathcal{P}_2$ è un $[\epsilon/2, \rho]$-optimal arm con probabilità di almeno $1-\delta'$. Inoltre, anche gli outputs di OptQF sono un $[\epsilon/2, \rho]$-optimal arm con probabilità $1-\delta$. Sia, $\hat{X}$ la frazione di sub-optimal arms in S. Quindi $\Pr\{\hat{X} \geq \frac{1}{2}\} = \Pr\{\hat{X} - \delta' \geq \frac{1}{4}\} \leq \exp(-2 \cdot (\frac{1}{4})^2 \cdot u) = \exp(-2 \cdot \frac{1}{16} \cdot 8 \log \frac{2}{\delta}) < \frac{\delta}{2}$. Dall'altro lato, la probabilità d'errore di OptQF ha un bound superiore di $\delta/2$. Tuttavia, facendo l'union bound, otteniamo che la probabilità d'errore è limitata superiormente da δ. Inoltre, la media dell'arm di output non è meno di $\frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$ presa dal $(1-\rho)$-esimo quantile. □

Complessità del campione

Dimostrazione. Innanzitutto notiamo che, per qualche costante C, la complessità del campione (SC) di ognuna delle u copie di$\mathcal{P}_2$ è $\frac{C}{\rho(\epsilon/2)^2}\left(\log \frac{2}{\delta'}\right)^2 \in O\left(\frac{1}{\rho\epsilon^2}\right)$. Quindi, la

SC di tutte le u copie $\mathcal{P}_2$ insieme è limitata superiormente da $\frac{C_1 \cdot u}{\rho\epsilon^2}$, per qualche costante C_1. Inoltre, per qualche costante C_2, la SC di OptQF è limitata superiormente da $C_2\left(\frac{u}{(u/2)(\epsilon/2)^2}\log\frac{2}{\delta}+\gamma(\cdot)\right) = C_2\left(\frac{8}{\epsilon^2}\log\frac{2}{\delta}+\gamma(\cdot)\right)$. Ora, aggiungendo le complessità dei campioni e sostituendo u dimostriamo tale bound. □

□

Elenco degli algoritmi

Elenco delle figure

Bibliografia

[AB10] Jean-Yves Audibert e Sébastien Bubeck. “Best arm identification in multi-armed bandits”. In: 2010.

[ACF02] Peter Auer, Nicolo Cesa-Bianchi e Paul Fischer. “Finite-time analysis of the multiarmed bandit problem”. In: *Machine learning* 47.2-3 (2002), pp. 235–256.

[Aga+17] Arpit Agarwal et al. “Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons”. In: *Conference on Learning Theory*. 2017, pp. 39–75.

[Agr95] Rajeev Agrawal. “The Continuum-Armed Bandit Problem”. In: *SIAM Journal on Control and Optimization* 33.6 (nov. 1995), pp. 1926–1951. ISSN: 1095-7138. DOI: 10.1137/s0363012992237273. URL: http://dx.doi.org/10.1137/S0363012992237273.

[AK08] Baruch Awerbuch e Robert Kleinberg. “Online linear optimization and adaptive routing”. In: *Journal of Computer and System Sciences* 74.1 (feb. 2008), pp. 97–114. ISSN: 0022-0000. DOI: 10.1016/j.jcss.2007.04.016. URL: http://dx.doi.org/10.1016/j.jcss.2007.04.016.

[AMS09] Jean-Yves Audibert, Rémi Munos e Csaba Szepesvári. “Exploration–exploitation tradeoff using variance estimates in multi-armed bandits”. In: *Theoretical Computer Science* 410.19 (2009), pp. 1876–1902.

[Azi+18] Maryam Aziz et al. *Pure Exploration in Infinitely-Armed Bandit Models with Fixed-Confidence*. 2018. arXiv: 1803.04665 [stat.ML].

[Bec58] Robert E. Bechhofer. “A Sequential Multiple-Decision Procedure for Selecting the Best One of Several Normal Populations with a Common Unknown Variance, and Its Use with Various Experimental Designs”. In: *Biometrics* 14.3 (set. 1958), p. 408. ISSN: 0006-341X. DOI: 10.2307/2527883. URL: http://dx.doi.org/10.2307/2527883.

[BF85] Donald A Berry e Bert Fristedt. “Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability)”. In: *London: Chapman and Hall* 5.71-87 (1985), pp. 7–7.

[BHF10] Eric Brochu, Matthew W Hoffman e Nando de Freitas. “Portfolio allocation for Bayesian optimization”. In: *arXiv preprint arXiv:1009.5419* (2010).

[Bol+18] Stefano Boldrini et al. “muMAB: A multi-armed bandit model for wireless network selection”. In: *Algorithms* 11.2 (2018), p. 13.

[Bou+14] Djallel Bouneffouf et al. “Contextual bandit for active learning: Active thompson sampling”. In: *International Conference on Neural Information Processing*. Springer. 2014, pp. 405–412.

[Bou+17] Djallel Bouneffouf et al. “Context attentive bandits: Contextual bandit with restricted context”. In: *arXiv preprint arXiv:1705.03821* (2017).

[BPR13] Sébastien Bubeck, Vianney Perchet e Philippe Rigollet. “Bounded regret in stochastic multi-armed bandits”. In: *Conference on Learning Theory*. 2013, pp. 122–134.

[BR19] Djallel Bouneffouf e Irina Rish. “A survey on practical applications of multi-armed and contextual bandits”. In: *arXiv preprint arXiv:1904.10040* (2019).

[Cap+13] Olivier Cappé et al. “Kullback–Leibler upper confidence bounds for optimal sequential allocation”. In: *The Annals of Statistics* 41.3 (giu. 2013), pp. 1516–1541. ISSN: 0090-5364. DOI: 10.1214/13-aos1119. URL: http://dx.doi.org/10.1214/13-AOS1119.

[CK17] Arghya Roy Chaudhuri e Shivaram Kalyanakrishnan. “PAC identification of a bandit arm relative to a reward quantile”. In: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[CV15] Alexandra Carpentier e Michal Valko. *Simple regret for infinitely many armed bandits*. 2015. arXiv: 1505.04627 [cs.LG].

[DLL19] Kaize Ding, Jundong Li e Huan Liu. “Interactive anomaly detection on attributed networks”. In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 2019, pp. 357–365.

[Dur+18] Audrey Durand et al. “Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis”. In: *Machine Learning for Healthcare Conference*. 2018, pp. 67–82.

[EMM02] Eyal Even-Dar, Shie Mannor e Yishay Mansour. “PAC Bounds for Multi-armed Bandit and Markov Decision Processes”. In: *Computational Learning Theory* (2002), pp. 255–270. ISSN: 0302-9743. DOI: 10.1007/3-540-45435-7_18. URL: http://dx.doi.org/10.1007/3-540-45435-7_18.

[EMM06] Eyal Even-Dar, Shie Mannor e Yishay Mansour. “Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems”. In: *Journal of machine learning research* 7.Jun (2006), pp. 1079–1105.

[Gab+11] Victor Gabillon et al. “Multi-bandit best arm identification”. In: *Advances in Neural Information Processing Systems*. 2011, pp. 2222–2230.

[Gao+19] Zijun Gao et al. “Batched multi-armed bandits problem”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 503–513.

[Gos+13] Sergiu Goschin et al. “Planning in reward-rich domains via PAC bandits”. In: *European Workshop on Reinforcement Learning*. 2013, pp. 25–42.

[GS10] Matteo Gagliolo e Jürgen Schmidhuber. “Algorithm selection as a bandit problem with unbounded losses”. In: *International Conference on Learning and Intelligent Optimization*. Springer. 2010, pp. 82–96.

[HPR96] Stephen J. Herschkorn, Erol Peköz e Sheldon M. Ross. “Policies without Memory for the Infinite-Armed Bernoulli Bandit under the Average-Reward Criterion”. In: *Probability in the Engineering and Informational Sciences* 10.1 (gen. 1996), pp. 21–28. ISSN: 1469-8951. DOI: 10.1017/s0269964800004149. URL: http://dx.doi.org/10.1017/S0269964800004149.

[Jam+14] Kevin Jamieson et al. “lil’ucb: An optimal exploration algorithm for multi-armed bandits”. In: *Conference on Learning Theory*. 2014, pp. 423–439.

[JHR16] Kevin Jamieson, Daniel Haas e Ben Recht. *On the Detection of Mixture Distributions with applications to the Most Biased Coin Problem*. 2016. arXiv: 1603.08037 [cs.LG].

[JN14] Kevin Jamieson e Robert Nowak. “Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting”. In: *2014 48th Annual Conference on Information Sciences and Systems (CISS)* (mar. 2014). DOI: 10.1109/ciss.2014.6814096. URL: http://dx.doi.org/10.1109/CISS.2014.6814096.

[Kal+12] Shivaram Kalyanakrishnan et al. “PAC Subset Selection in Stochastic Multi-armed Bandits.” In: *ICML*. Vol. 12. 2012, pp. 655–662.

[Kal11] Shivaram Kalyanakrishnan. “Learning methods for sequential decision making with imperfect representations”. In: (2011).

[KK13] Emilie Kaufmann e Shivaram Kalyanakrishnan. “Information complexity in bandit subset selection”. In: *Conference on Learning Theory*. 2013, pp. 228–251.

[KKS13] Zohar Karnin, Tomer Koren e Oren Somekh. “Almost optimal exploration in multi-armed bandits”. In: *International Conference on Machine Learning*. 2013, pp. 1238–1246.

[Kle05] Robert D Kleinberg. “Nearly tight bounds for the continuum-armed bandit problem”. In: *Advances in Neural Information Processing Systems*. 2005, pp. 697–704.

[KS10] Shivaram Kalyanakrishnan e Peter Stone. “Efficient Selection of Multiple Bandit Arms: Theory and Practice.” In: *ICML*. Vol. 10. 2010, pp. 511–518.

[Li+10] Lihong Li et al. “A contextual-bandit approach to personalized news article recommendation”. In: *Proceedings of the 19th international conference on World wide web - WWW ’10* (2010). DOI: 10.1145/1772690.1772758. URL: http://dx.doi.org/10.1145/1772690.1772758.

[Li+17] Lisha Li et al. “Hyperband: A novel bandit-based approach to hyperparameter optimization”. In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 6765–6816.

[Liu+17] Bing Liu et al. “Customized nonlinear bandits for online response selection in neural conversation models”. In: *arXiv preprint arXiv:1711.08493* (2017).

[LPB17] David E Losada, Javier Parapar e Alvaro Barreiro. “Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems”. In: *Information Processing & Management* 53.5 (2017), pp. 1005–1025.

[Mou+16] Seyed Hamed Mousavi et al. “Analysis of a Subset Selection Scheme for Wireless Sensor Networks in Time-Varying Fading Channels”. In: *IEEE Transactions on Signal Processing* 64.9 (mag. 2016), pp. 2193–2208. ISSN: 1941-0476. DOI: 10.1109/tsp.2016.2515067. URL: http://dx.doi.org/10.1109/TSP.2016.2515067.

[MT03] Shie Mannor e John N. Tsitsiklis. “Lower Bounds on the Sample Complexity of Exploration in the Multi-armed Bandit Problem”. In: *Lecture Notes in Computer Science* (2003), pp. 418–432. ISSN: 1611-3349. DOI: 10.1007/978-3-540-45167-9_31. URL: http://dx.doi.org/10.1007/978-3-540-45167-9_31.

[Noo+18] Ritesh Noothigattu et al. “Interpretable multi-objective reinforcement learning through policy orchestration”. In: *arXiv preprint arXiv:1809.08343* (2018).

[Pau64] Edward Paulson. “A Sequential Procedure for Selecting the Population with the Largest Mean from k Normal Populations”. In: *The Annals of Mathematical Statistics* 35.1 (mar. 1964), pp. 174–180. ISSN: 0003-4851. DOI: 10.1214/aoms/1177703739. URL: http://dx.doi.org/10.1214/aoms/1177703739.

[PR+13] Vianney Perchet, Philippe Rigollet et al. “The multi-armed bandit problem with covariates”. In: *The Annals of Statistics* 41.2 (2013), pp. 693–721.

[RLS19] Wenbo Ren, Jia Liu e Ness B Shroff. “Exploring k out of Top rho Fraction of Arms in Stochastic Bandits”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. 2019, pp. 2820–2828.

[Rob52] Herbert Robbins. “Some aspects of the sequential design of experiments”. In: *Bulletin of the American Mathematical Society* 58.5 (1952), pp. 527–535.

[SC12] Bubeck Sébastien e Nicolò Cesa-Bianchi. “Regret analysis of stochastic and non stochastic multi-armed bandit problems”. In: *Foundations and Trends in Machine Learning* 5.1 (2012), pp. 1–122.

[She+15] Weiwei Shen et al. “Portfolio choices with orthogonal bandit learning”. In: *Twenty-fourth international joint conference on artificial intelligence*. 2015.

[Sil+18] Tomi Silander et al. “Contextual memory bandit for pro-active dialog engagement”. In: (2018).

[SL18] Jérémie Sublime e Sylvain Lefebvre. “Collaborative clustering through constrained networks using bandit optimization”. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2018, pp. 1–8.

[Tra+14] Long Tran-Thanh et al. “Efficient crowdsourcing of unknown experts using bounded multi-armed bandits”. In: *Artificial Intelligence* 214 (set. 2014), pp. 89–111. ISSN: 0004-3702. DOI: 10.1016/j.artint.2014.04.005. URL: http://dx.doi.org/10.1016/j.artint.2014.04.005.

[Tro+15] Francesco Trovò et al. “Multi–armed bandit for pricing”. In: *Proceedings of the European Workshop on Reinforcement Learning (EWRL)*. 2015.

[Upa+19] Sohini Upadhyay et al. “A Bandit Approach to Posterior Dialog Orchestration Under a Budget”. In: *arXiv preprint arXiv:1906.09384* (2019).

[Vas+17] Sharan Vaswani et al. “Model-independent online learning for influence maximization”. In: *arXiv preprint arXiv:1703.00557* (2017).

[WAM09] Yizao Wang, Jean-Yves Audibert e Rémi Munos. “Algorithms for infinitely many-armed bandits”. In: *Advances in Neural Information Processing Systems*. 2009, pp. 1729–1736.

[Wil+16] Yvonne Will et al. *Drug Discovery Toxicology: From Target Assessment to Translational Biomarkers*. John Wiley & Sons, 2016.

[Zho+17] Qian Zhou et al. “Large-scale bandit approaches for recommender systems”. In: *International Conference on Neural Information Processing*. Springer. 2017, pp. 811–821.