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.

Nella prossimo paragrafo verrà presentato brevemente il contenuto sottoforma di abstract.

- 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 descrittti e commentati i lavori precedenti in merito allo stesso problema. Oltre ai lavori citati nei "related works" e nei riferimenti bibliograci 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 quotidianetà.

Infine sono presenti anche i riferimenti bibliografici citati nell'elaborato.

1.1 Abstract

Nell'ambito *PAC*[1] verrà considerato il problema di identificare un numero k qualsiasi tra i 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 "uno dei 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 dai migliori m, dove $1 \leq k \leq m$, serve come 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 distinti 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 "uno dei migliori ρ" per istanze infinite e quello "uno dei 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 avrà 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 fatti 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 degli altri 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 piu 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 sottoattività, 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 sottoattività. Per rispondere alle esigenze dell'azienda, sicuramente sarebbe sufficiente identificare i 5 migliori lavoratori per la sottoattività. 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 qualsiasi k tra i migliori m arm 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 minimisation 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 avrà 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.

Network Server Selection

Nel caso in cui un lavoro deve essere elaborato su uno dei numerosi server, ognuno dei quali ha differenti velocita 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 2.2: Server

In questo articolo [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 sublineare 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.

Internet Advertising

Ogni volta che un utente visita il sito è necessario scegliere di visualizzare una delle k pubblicità possibili; la ricompensa si ottiene se un utente fa click sulla 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 [Li+10], altre applicazioni possibili sono sul consigliare video correlati a quello che l'utente sta guardando su piattaforme come YouTube.

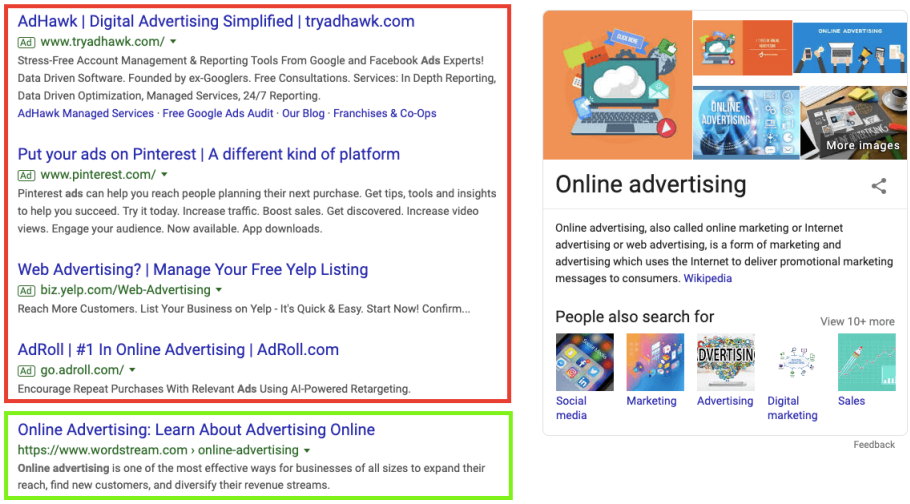

Figura 2.3: Google Ads

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 di «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 [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.

Drug Design

Figura 2.4: Drug design

Vi sono applicazioni anche riguardanti la progettazione di farmaci [Wil+16]. 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.

Clinical Trial

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 vericare che una nuova terapia sia piu efficace, migliore e soprattutto sicura di quella normalmente impiegata. Assume notevole importanza l'insieme di campioni sui quali poter testare nuove cure [Rob52].

Figura 2.5: Trial clinici

Gestione di grandi reti di sensori

Le applicazioni di questo problema includono la gestione di grandi reti di sensori [Mou+16], in cui più sensori affidabili devono essere identificati facendo il minor numero di test possibile.

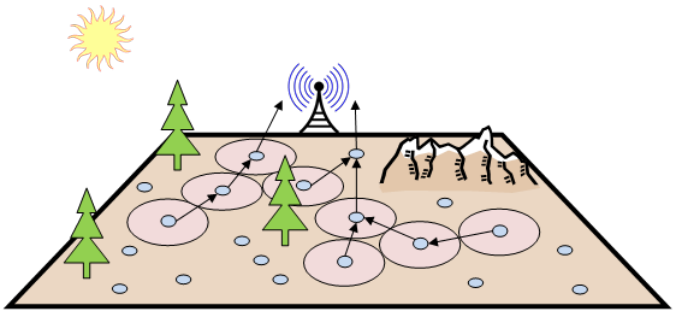

Figura 2.6: Wireless Sensor Network (*WSN*)

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 [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 controllo di equivalenza di certezza con schemi di forzatura e successivamente vengono derrivati 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 *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. Proponiamo 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 nostra analisi, mostriamo 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 rappresentana 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 sublineare 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 di bandit models ta 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 distribuiti con una 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 perchè semplici da capire oppure 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*, nel senso di *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 *Burnetas e Katehakis*[5], rispettivamente. Viene studiato anche il comportamento di

selezionata avrà 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

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 arnms. Tutti gli algoritmi precedenti per questa impostazione sono stati progettati forminimizzando 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 varaibile, 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. Il metodo descritto non può essere migliorato nel senso che il numero di campioni

[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.

necessari per identificare l'arm migliore rientra in un fattore costante del limite inferiore definito dalla legge del logaritmo iterato *(LIL)*. Ispirati dal *LIL*, vengon 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. Vengopno 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 *Ex-plore-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. Presentiamo 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 dimostriamo limiti superiori il cui gap dal limite inferiore è solo doppiamente logaritmico nei parametri del problema. Confermiamo i nostri 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 considerat 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 contribuit ancheo 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 sublogaritmico. 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:** dà 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 di «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 nel presente documento hanno tutte la proprietà che le popolazioni inferiori 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.

E' un articolo che non ha ricevuto molte attenzioni al tempo della stesura ma e stato rivalutato negli ultimi anni e viene spesso citato. Studia la teoria dell'analisi sequenziale, soermandosi anche sulla numerosita 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 armscci 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 viene usato anche alcuni algortimi 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[9] che sono serviti sia per la comprensione che per l'approfondimento del paper assegnato.

Alcuni lavori correlati possono essere trovati grazie a parole chiave in *Proceedings of Machine Learning Research,* in particolare riguardante il *Volume 97: International Conference on Machine Learning, 9-15 June 2019, Long Beach, California, USA*[10].

[9] Google Scholar (https://scholar.google.it/)
[10] *http://proceedings.mlr.press/v97/*

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. Qui di seguito un riassunto del contenuto del paper:

[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 del caso multi-armed è ancora un argomento aperto. Inoltre, rimane ancora inesplorata la domanda se dimensioni di batch scelti in modo adattivo aiutano 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.

[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).

Viene analizzata la relazione fra la complessità e l'adattivita nel richiedere attivamente nuovi dati per identicare 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 (cioe l'insieme dei punti in cui la distribuzione non è una funzione liscia, che vuol dire derivabile infinite volte in quel punto).

[11] Perchet, V., Rigollet, P., Chassang, S., & Snowberg, E. (2016). *Batched bandit problems.* The Annals of Statistics, 44(2), 660-681.

[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 dierenze 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 puo 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 piu piccola dierenza 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 piu bassi.

[EMM06] Even-Dar, E., Mannor, S., & Mansour, Y. (2006). *Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems.* Journal of machine learning research, 7(Jun), 1079-1105.

Mostra gli intervalli di condenza nel problema dei banditi dimostrando quale sia il numero minimo di scelte per trovare il braccio ottimale con una probabilita definita. Propone un framework che cerca di eliminare le azioni che non sono ottime con alta probabilità. Inoltre mostra una variante basata su 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 introudce una politica chiamata *Adaptively Binned Successive Elimination (ABSE)* che decompone adattativamente il problema in più piccoli problemi dei *bandits* di tipo statico.

ABSTRACT DI QUESTO SOPRA

KEY (PAC, BANDITS)

NOTE A PIE DI PAGINA PER OGNI TERMINE

GIA MESSO ([Rob52] Herbert Robbins)

4 Descrizione

VERSIONE ESTESA DEL PAPER

5 Esperimenti

ALGORITMI?

CODICE?

6 Conclusioni

COMMENTARE BENE I RISULTATI
SCENARI APPLICATIVI SPOSTARLI QUI?

7 Considerazioni personali

Il 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 mi incuriosiva già infatti avevo già letto e analizzato alcune parti che mi sono risultate poi utili nel produrre l'analisi del paper a me assegnato.

Un'opzione per la tesi.

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.

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

[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].

[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.

[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.

[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.

[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.

[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.

[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.