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 nella conferenza *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 senza farne di nuovi.
- Nel **capitolo 6** verranno elencate le conclusioni sulle proprietà teoriche e sperimentali del metodo analizzato 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 bibliograci 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 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 effettuato e 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 scelte possibili, 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]; [EDMM02] 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 [EDMM02]; [KS10], o in alternativa, (2) con errore minimo, dopo un determinato numero di prove [AB10]; [CV15]. La nostra indagine 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 [EDMM02]; [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, sarebbe inutilmente costoso per identificare il miglior sottoinsieme (*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 nella minimizzazione del rimpianto [HPR96] che in *fixed confidence setting* [Gos+13]; [CK17]. La nostra formulazione apre la strada all'identificazione di "molti (k) good" (tra i migliori m degli n) arms in questo modo.

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;), 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

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, nel DNS server selection e nel cloud computing. Approfondire l'adaptive routing [AK08].

Figura 2.2: Server

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. Nessuna conoscenza dell'utente, del contenuto dell'annuncio e del contenuto della pagina web richiesta. Approfondire i recommender system [Li+10].

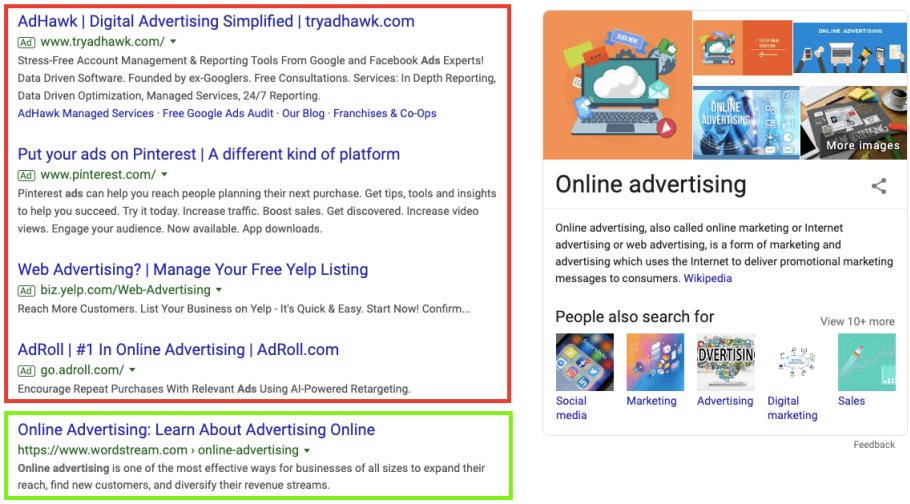

Figura 2.3: Google Ads

Design di trial clinici

Il trial clinico si riferisce a uno studio medico farmacologico, biomedico o correlato alla salute sull'uomo. Vengono adottati protocolli ben definiti. L'obiettivo e quello di vericare che una nuova cura sia piu efficace, migliore e soprattutto sicura di quella normalmente impiegata. Assume notevole importanza l'insieme di campioni sui quali poter testare nuove cure. Approfondire i trial clinici [Rob52].

Figura 2.4: Trial clinici

Progettazione di farmaci

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

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.

Crowdsouring distribuito

Vi sono applicazioni riguardanti il crowdsourcing distribuito [TT+14].

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.

[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 di ordine (fino a un fattore *log(K)*) $\Sigma i\, 1/\Delta_i^2$, dove la somma è sugli arms non

[1] Upper Confidence Bound (UCB) policy

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 dei *stochastic multi-armed bandits*. Motivati principalmente dal compito di progettare algoritmi di routing adattivi per reti sovrapposte, presentano due *randomized online algorithms* per selezionare una sequenza di percorsi di routing in una rete con ritardi ai bordi 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 loro lavoro 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 il nostro algoritmo 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 derivare e trasmettere i risultati; hanno derivato un limite inferiore sulla

[2]Probably Approximately Correct 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 concet-

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 $log^2(1/\Delta)$ 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 bandits.* Tuttavia molti risultati non sono stati dimostrati perchè semplici da capire oppure attraverso una dimostrazione concettuale invece di usare i calcoli.

3.2 Lavori citati nella letteratura

In questa sezione sono analizzati ulteriori paper presenti nella letteratura[3] che sono serviti sia per la comprensione che per l'approfondimento del paper assegnato.

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 nella sessione estiva.

CORSIVO I TITOLI

to dato qualsiasi rapporto di approssimazione arbitrario, probabilità di successo o distribuzione dei campioni.

[3]Google Scholar (https://scholar.google.it/)

4 Descrizione

Versione estesa del file.

5 Esperimenti

6 Conclusioni

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 dei listati

Elenco delle figure

Bibliografia

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

[Agr95] Rajeev Agrawal. “The Continuum-Armed Bandit Problem”. In: *SIAM Journal on Control and Optimization* 33.6 (1995), 1926â1951. 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 (2008), 97â114. ISSN: 0022-0000. DOI: 10.1016/j.jcss.2007.04.016. URL: http://dx.doi.org/10.1016/j.jcss.2007.04.016.

[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 (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.

[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].

[EDMM02] Eyal Even-Dar, Shie Mannor e Yishay Mansour. “PAC Bounds for Multi-armed Bandit and Markov Decision Processes”. In: *Computational Learning Theory* (2002), 255â270. ISSN: 0302-9743. DOI: 10.1007/3-540-45435-7_18. URL: http://dx.doi.org/10.1007/3-540-45435-7_18.

[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 PekÃ¶z 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 (1996), 21â28. ISSN: 1469-8951. DOI: 10.1017/s0269964800004149. URL: http://dx.doi.org/10.1017/S0269964800004149.

[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)* (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.

[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 â10* (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 (2016), 2193â2208. ISSN: 1941-0476. DOI: 10.1109/tsp.2016.2515067. URL: http://dx.doi.org/10.1109/TSP.2016.2515067.

[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 (1964), 174â180. ISSN: 0003-4851. DOI: 10.1214/aoms/1177703739. URL: http://dx.doi.org/10.1214/aoms/1177703739.

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

[TT+14] Long Tran-Thanh et al. “Efficient crowdsourcing of unknown experts using bounded multi-armed bandits”. In: *Artificial Intelligence* 214 (2014), 89â111. 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.