Small-loss bounds for online learning with partial information (Extended abstract)

Thodoris Lykouris TEDDLYK@CS.CORNELL.EDU
107 Hoy Rd, Ithaca, NY, 14850

Karthik Sridharan SRIDHARAN@CS.CORNELL.EDU
107 Hoy Rd, Ithaca, NY, 14850

Éva Tardos EVA.TARDOS@CORNELL.EDU
107 Hoy Rd, Ithaca, NY, 14850

Editors: Sébastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We consider the problem of adversarial (non-stochastic) online learning with partial information feedback, where at each round, a decision maker selects an action from a finite set of alternatives. We develop a black-box approach for such problems where the learner observes as feedback only losses of a subset of the actions that includes the selected action. When losses of actions are non-negative, under the graph-based feedback model introduced by Mannor and Shamir, we offer algorithms that attain the so called "small-loss" $o(\alpha L^\star)$ regret bounds with high probability, where α is the independence number of the graph, and $L^\star$ is the loss of the best action. Prior to our work, there was no data-dependent guarantee for general feedback graphs even for pseudo-regret (without dependence on the number of actions, i.e. utilizing the increased information feedback). Taking advantage of the black-box nature of our technique, we extend our results to many other applications such as semi-bandits (including routing in networks), contextual bandits (even with an infinite comparator class), as well as learning with slowly changing (shifting) comparators.

In the special case of classical bandit and semi-bandit problems, we provide optimal small-loss, high-probability guarantees of $\widetilde{\mathcal{O}}(\sqrt{dL^\star})$ for actual regret, where d is the number of actions, answering open questions of Neu. Previous bounds for bandits and semi-bandits were known only for pseudo-regret and only in expectation. We also offer an optimal $\widetilde{\mathcal{O}}(\sqrt{\kappa L^\star})$ regret guarantee for fixed feedback graphs with clique-partition number at most κ.

Keywords: Online Learning, Bandits, Feedback Graphs, Regret, High probability

1. Introduction

The online learning paradigm (Littlestone and Warmuth, 1994; Cesa-Bianchi and Lugosi, 2006) has become a key tool for solving a wide spectrum of problems such as developing strategies for players in large multiplayer games (Blum et al., 2006, 2008; Roughgarden, 2015; Lykouris et al., 2016; Foster et al., 2016), designing online marketplaces and auctions (Blum and Hartline, 2005; Cesa-Bianchi et al., 2013; Roughgarden and Wang, 2016), portfolio investment (Cover, 1991; Freund and Schapire, 1997; Hazan et al., 2007), online routing (Awerbuch and Kleinberg, 2004; Kalai and Vempala, 2005). In each of these applications, the learner has to repeatedly select an action on every round. Different actions have different costs or losses associated with them on every round. The goal of the learner is to minimize her cumulative loss and the performance of the learner is evaluated by the notion of "*regret*", defined as the difference between the cumulative loss of the learner, and the cumulative loss $L^\star$ of the benchmark.

The term "*small-loss* regret bound" is often used to refer to bounds on regret that depend (or mostly depend) on $L^\star$, rather than the total number of rounds played T often referred to as the time horizon. For instance, for many classical online learning problems, one can in fact show that regret can be bounded by $\widetilde{O}(\sqrt{L^\star})$ rather than $\widetilde{O}(\sqrt{T})$. However, these algorithms use the *full information* model: assume that on every round, the learner receives as feedback the losses of all possible actions (not only the selected actions). In such full information settings, it is well understood when small-loss bounds are achievable and how to design learning algorithms that attain them. However, in most applications, full information about losses of all actions is not available. Unlike the full information case, the problem of obtaining small-loss regret bounds for partial information settings is poorly understood. Even in the classical multi-armed bandit problem, small-loss bounds are only known in expectation against the so called oblivious adversaries or comparing against the lowest expected cost of an arm (and not the actual lowest cost), referred to as pseudo-regret.

The goal of this paper is to develop robust techniques for extending the small-loss guarantees to a broad range of partial feedback settings where learner only observes losses of selected actions and some neighboring actions. In the basic online learning model, at each round t, the decision maker or *learner* chooses one action from a set of d actions, typically referred to as *arms*. Simultaneously an adversary picks a loss vector $\ell^t \in [0,1]^d$ indicating the losses for the d arms. The learner suffers the loss of her chosen arm and observes some feedback. The variants of online learning differ by the nature of feedback received. The two most prominent such variants are the *full information setting*, where the feedback is the whole loss vector, and the *bandit setting* where only the loss of the selected arm is observed. Bandits and full information represent two extremes. In most realistic applications, a learner choosing an action i, learns not only the loss ℓ_i^t associated with her chosen action i, but also some partial information about losses of some other actions. A simple and elegant model of this partial information is the *graph-based* feedback model (Mannor and Shamir, 2011; Alon et al., 2017), where at every round, there is a (possibly time-varying) undirected graph G^t representing the information structure, where the possible actions are the nodes. If the learner selects an action i and incurs the loss ℓ_i^t, she observes the losses of all the nodes connected to node i by an edge in G^t. Our main result is a general technique that allows us to use any full information learning algorithm as a black-box, and design a learning algorithm whose regret can be bounded with high probability as $o(\alpha L^\star)$, where α is the maximum independence number of the feedback graphs. This graph-based information feedback model is a very general setting that can encode all of full information, bandit, as well as a number of other applications.

1.1. Our contribution

Our results We develop a unified, black-box technique to achieve small-loss regret guarantees with high probability in various partial information feedback models. We obtain the following results.

- We first provide a generic black box reduction from any small-loss full information algorithm. When used with known algorithms it achieves actual regret guarantees of $\widetilde{\mathcal{O}}\big((L^\star)^{2/3}\big)$ that hold with high probability for any of pure bandits, semi-bandits, contextual bandits, or feedback graphs (with dependence on the information structure in the $\widetilde{\mathcal{O}}$ as $d^{1/3}$ for the first three, and $\alpha^{1/3}$ for feedback graphs). There are three novel features of this result. First, unlike most previous work in partial information that is heavily algorithm-specific, our technique is black-box in the sense that it takes as input a small-loss full information algorithm and, via a small modification, makes it work under partial information. Second, prior to our work, there was no data-dependent guarantee for general feedback graphs even for pseudo-regret (without dependence on the number of actions, i.e., taking advantage of the increased information feedback), while we provide a high probability small-loss guarantee. Last, our guarantees are not for pseudo-regret but actual regret guarantees that hold with high probability.

- We then show various applications. The black-box nature of our reduction allows us to use the full information learning algorithms best suited for each application. We obtain small-loss guarantees for semi-bandits (Kalai and Vempala, 2005) (including routing in networks), for contextual bandits (Langford and Zhang, 2007) (even with an infinite comparator class), as well as learning with slowly changing (shifting) comparators (Herbster and Warmuth, 1998) as needed in games with dynamic population (Lykouris et al., 2016; Foster et al., 2016).

- Finally, we focus on the special case of bandits, semi-bandits, graph feedback from fixed graphs, and shifting comparators. In each setting we take advantage of properties of a learning algorithm best suited in the application to alleviate the inefficiencies resulting from the black-box nature of our general reduction. For bandits and semi-bandits, we provide optimal small-loss actual regret high-probability guarantees of $\widetilde{\mathcal{O}}(\sqrt{dL^\star})$. Previous work for bandits and semi-bandits offered analogous bounds only for pseudo-regret and only in expectation. This answers an open question of Neu (2015a,b). In the case of fixed feedback graphs, we achieve optimal $\sqrt{L^*}$ dependence on loss, at the expense of the bound depending on clique-partition number of the graph, rather than the independence number.

Our techniques Our main technique is a dual-thresholding scheme that temporarily freezes low-performing actions, i.e. does not play them at the current round. Traditional partial information guarantees are based on creating an unbiased estimator for the loss of each arm and then running a full information algorithm on the estimated loses. The most prominent such unbiased estimator, called *importance sampling*, is equal to the actual loss divided by the probability with which the action is played. This division can make the estimated losses unbounded in the absence of a lower bound on the probability of being played. Algorithms like EXP3 (Auer et al., 2003) for the bandit setting or Exp3-DOM (Alon et al., 2017) for the graph-based feedback setting mix in a $1/\sqrt{T}$ amount of noise which ensures that the range of losses is bounded. Adding such uniform noise works well for learners maximizing utility, but can be very damaging when minimizing losses. In the case of utilities, playing low performing arms with a small ϵ probability, can only lose at most an ϵ fraction of the utility. In contrast, when the best arm has small loss, the losses incurred due to the noise can dominate. This approach can only return uniform bounds with $\mathcal{O}(\sqrt{T})$ regret since, even in the case that there is a perfect arm that has 0 loss, the algorithm keeps playing low-performing arms. Some specialized algorithms do achieve small-loss bounds for bandits, but these techniques extend neither to graph feedback nor to high probability guarantees (see also discussion below about related work).

Instead of mixing in noise, we take advantage of the freezing idea, originally introduced by Allenberg et al. (2006) with a single threshold γ offering a new way to adapt the multiplicative weights algorithm to the bandit setting. The resulting estimator is negatively biased for the arms that are frozen but is always unbiased for the selected arm. Using these expectations, the regret bound of the full information algorithm can be used to bound the expected regret compared to the expected loss of any fixed arm, achieving low pseudo-regret in expectation. To achieve good bounds, we need to guarantee that the total probability frozen is limited. By freezing arms with probability less than γ, the total probability that is frozen at each round is at most $d\gamma$ and therefore contributes to a regret term of $d\gamma$ times the loss of the algorithm which gives a dependence on d on the regret bound. This was analyzed in the context of multiplicative weights by Allenberg et al. (2006).

Our main technical contribution is to greatly expand the power of this freezing technique. We show how to apply it in a black-box manner with any full information learning algorithm and extend it to graph-based feedback. To deal with the graph-based feedback setting, we suggest a novel and technically more challenging dual-threshold freezing scheme. The natural way to apply importance sampling in the graph-based feedback is by dividing the actual loss with the probability of being observed, i.e. the sum of the probabilities that the action and its neighbors are played. An initial approach is to freeze an action if its probability of being observed is below some threshold γ. We show that the total probability frozen by this step is bounded by $\alpha\gamma$, where α is the size of the maximum independent number of the feedback graph. To see why, consider a maximal independent set S of the frozen actions and note that all frozen actions are observed by some node in S. This observation seems to imply that we can replace the dependence on d by a dependence on α. However there are externalities among actions as freezing one action may affect the probability of another being observed. As a result, the latter may need to be frozen as well to ensure that all active arms are observed with probability at least γ (and therefore obtain our desired upper bound on the range of the estimated losses). This causes a cascade of freezing, with possibly freezing a large amount of additional probability.

To limit this cascade effect, we develop a dual-threshold freezing technique: we initially freeze arms that are observed with probability less than γ, and subsequently use a lower threshold $\gamma' = \gamma/3$ and only freeze arms that are observed with probability less than γ'. This technique allows us to bound the total probability of arms that are frozen subsequently by the total probability of arms that are frozen initially. We prove this via an elegant combinatorial charging argument.

Last, to go beyond pseudo-regret and guarantee actual regret bounds with high probability, it does not suffice to have the estimator be negatively biased but we need to also obtain a handle on the variance. We prove that freezing also provides such a lever leading to a high-probability $\widetilde{\mathcal{O}}(\alpha^{1/3}(L^\star)^{2/3})$ regret guarantee that holds in a black-box manner. Interestingly, this freezing technique via a small modification enables the same guarantee for semi-bandits where the independent set is replaced by the number of elements (edges).

In order to obtain the optimal high-probability guarantee for bandits and semi-bandits, we need to combine our black box analysis with features of concrete full information learning algorithms. The black-box nature of the previous analysis is extremely useful in demonstrating where additional features are needed. Combining our analysis with the implicit exploration technique of Kocák et al. (2014) similarly as in the analysis of Neu (2015b), we develop an algorithm based on multiplicative weights, which we term *GREEN-IX*, which achieves the optimal high-probability small-loss bound $\widetilde{\mathcal{O}}(\sqrt{dL^\star})$ for the pure bandit setting. Using an alternative technique of Neu (2015a): truncation in the follow the perturbed leader algorithm, we also obtain the corresponding result for semi-bandits.

1.2. Related work

Online learning with partial information dates back to the seminal work of Lai and Robbins (1985). They consider a stochastic version, where losses come from fixed distributions. We focus on the case where the losses are selected adversarially, i.e. they do not come from a distribution and may be adaptive to the algorithm's choices. This was first studied by Auer et al. (2003) who provided the EXP3 algorithm for pure bandits and the EXP4 algorithm for learning with expert advice (a generalization of the contextual bandits of Langford and Zhang (2007)). They focus on uniform regret bounds, i.e. that grow as a function of time $o(T)$, and bound mostly the expected performance, but such guarantees can also be derived with high probability (Auer et al., 2003; Audibert and Bubeck, 2010; Beygelzimer et al., 2011). Data-dependent guarantees are easily derived from the above algorithms for the case of maximizing some reward as even getting reward 0 with probability of ϵ only causes an ϵ fraction of loss in utility. In contrast, incurring high cost with a small probability ϵ can dominate the loss of the algorithm, if the best arm has small loss. In this paper we develop data-dependent guarantees for partial information algorithm for the cases of losses. There are a few specialized algorithms that achieve such small-loss guarantees for the case of bandits for pseudo-regret, e.g. by ensuring that the estimated losses of all arms remain close (Allenberg et al., 2006; Neu, 2015a) or using a stronger regularizer (Rakhlin and Sridharan, 2013; Foster et al., 2016), but all of these methods neither offer high probability small-loss guarantees even for the bandit setting, nor extend to graph-based feedback. Our technique allows us to develop small-loss bounds on actual regret with high probability.

The graph-based partial information that we examine in this paper was introduced by Mannor and Shamir (2011) who provided ELP, a linear programming based algorithm achieving $\widetilde{\mathcal{O}}(\sqrt{\alpha T})$ regret for undirected graphs. Alon et al. (2013, 2017) provided variants of Exp3 (Exp3-SET) that recovered the previous bound via what they call *explicit exploration*. Following this work, there have been multiple results on this setting, e.g.(Alon et al., 2015; Cohen et al., 2016; Kocák et al., 2016; Tossou et al., 2017), but prior to our work, there was no small-loss guarantee for the feedback graph setting that could exploit the graph structure. To obtain a regret bound depending on the graph structure, the above techniques upper bound the losses of the arms by the maximum loss which results in a dependence on the time horizon T instead of $L^\star$. Addressing this, we achieve regret that scales with an appropriate problem dimension, the size of the maximum independent set α, instead of ignoring the extra information and only depending on the number of arms as all small-loss results of prior work.

Biased estimators have been used prior to our work for achieving better regret guarantees. The freezing technique of Allenberg et al. (2006) can be thought of as the first use of biased estimators. Their *GREEN* algorithm uses freezing in the context of the multiplicative weights algorithm for the case of pure bandits. Freezing keeps the range of estimated losses bounded and when used with the multiplicative weights algorithm, also keeps the cumulative estimated losses very close, which ensures that one does not lose much in the application of the full information algorithm. Using these facts, Allenberg et al. (2006) achieved small-loss guarantees for pseudo-regret in the classical multi-armed bandit setting. An approach very close to freezing is the *implicit exploration* of Kocák et al. (2014) that adds a term in the denominator of the estimator making the estimator biased, even for the selected arms. The *FPL-TrIX* algorithm of Neu (2015a) is based on the Follow the Perturbed Leader algorithm using implicit exploration together with truncating the perturbations to guarantee that the estimated losses of all actions are close to each other and the *geometric resampling* technique of Neu and Bartók (2013) to obtain these estimated losses. His analysis provides small-loss regret bounds for pseudo-regret, but does not extend to high-probability guarantees. The *EXP3-IX* algorithm of Kocák et al. (2014)combines implicit exploration with multiplicative weights to obtain, via the analysis of Neu (2015b), high-probability uniform bounds. Focusing on uniform regret bounds, exploration and truncation were presented as strictly superior to freezing. In this paper, we show an important benefit of the freezing technique: it can be extended to handle feedback graphs (via our dual-thresholding). We also combine freezing with multiplicative weights to develop an algorithm we term *GREEN-IX* which achieves optimal high-probability small-loss $\widetilde{\mathcal{O}}(\sqrt{dL^\star})$ for the pure bandit setting. Finally, combining freezing with the truncation idea, we obtain the corresponding result for semi-bandits; in contrast, the geometric resampling analysis does not seem to extend to high probability since it does not provide a handle on the variance of the estimated loss.

Full version

We refer the reader to the full version of the paper that can be found here https://arxiv.org/abs/1711.03639.

Acknowledgments

We thank Haipeng Luo for pointing out the connection to the contextual bandit setting, Jake Abernethy, Adam Kalai, and Santosh Vempala for a useful discussion about Follow the Perturbed Leader, and anonymous reviewers for their useful feedback. This work was supported by NSF grants CCF-1563714 and CDS&E-MSS 1521544, and an NSF-CAREER Award 1750575

References

Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *17th International Conference on Algorithmic Learning Theory (ALT)*, 2006.

Noga Alon, Nicol Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. From bandits to experts: A tale of domination and independence. In *27th Annual Conference on Neural Information Processing Systems (NIPS)*, 2013.

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, 2015.

Noga Alon, Nicol Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing (SICOMP)*, 46(6):1785–1826, 2017.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2785–2836, 2010. URL http://portal.acm.org/citation.cfm?id=1953023.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, January 2003. ISSN 0097-5397. doi: 10.1137/S0097539701398375. URL http://dx.doi.org/10.1137/S0097539701398375.

Baruch Awerbuch and Robert D. Kleinberg. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)*, 2004.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Avrim Blum and Jason D. Hartline. Near-optimal online auctions. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2005.

Avrim Blum, Eyal Even-Dar, and Katrina Ligett. Routing without regret: On convergence to nash equilibria of regret-minimizing algorithms in routing games. In *Proceedings of the 25th Annual ACM Symposium on Principles of Distributed Computing (PODC)*, 2006.

Avrim Blum, MohammadTaghi Hajiaghayi, Katrina Ligett, and Aaron Roth. Regret minimization and the price of total anarchy. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*, 2008.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Regret minimization for reserve prices in second-price auctions. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2013.

Alon Cohen, Tamir Hazan, and Tomer Koren. Online learning with feedback graphs without the graphs. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML)*, 2016.

Thomas M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, 1991. ISSN 1467-9965. doi: 10.1111/j.1467-9965.1991.tb00002.x. URL http://dx.doi.org/10.1111/j.1467-9965.1991.tb00002.x.

Dylan J. Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Éva Tardos. Learning in games: Robustness of fast convergence. In *30th Annual Conference on Neural Information Processing Systems (NIPS)*, 2016.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, August 1997. ISSN 0022-0000. doi: 10.1006/jcss.1997.1504. URL http://dx.doi.org/10.1006/jcss.1997.1504.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Mach. Learn.*, 69(2-3):169–192, December 2007. ISSN 0885-6125. doi: 10.1007/s10994-007-5016-8. URL http://dx.doi.org/10.1007/s10994-007-5016-8.

Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Mach. Learn.*, 32(2):151–178, August 1998. ISSN 0885-6125. doi: 10.1023/A:1007424614876. URL http://dx.doi.org/10.1023/A:1007424614876.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *J. Comput. Syst. Sci.*, 71(3):291–307, October 2005. ISSN 0022-0000. doi: 10.1016/j.jcss.2004.10.016. URL http://dx.doi.org/10.1016/j.jcss.2004.10.016.

Tomás Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.

Tomáš Kocák, Gergely Neu, and Michal Valko. Online learning with noisy side observations. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1):4–22, March 1985. ISSN 0196-8858. doi: 10.1016/0196-8858(85)90002-8. URL http://dx.doi.org/10.1016/0196-8858(85)90002-8.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS)*, 2007.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Inf. Comput.*, 108 (2):212–261, February 1994. ISSN 0890-5401. doi: 10.1006/inco.1994.1009. URL http://dx.doi.org/10.1006/inco.1994.1009.

Thodoris Lykouris, Vasilis Syrgkanis, and Éva Tardos. Learning and efficiency in games with dynamic population. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2016.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *25th Annual Conference on Neural Information Processing Systems (NIPS)*, 2011.

Gergely Neu. First-order regret bounds for combinatorial semi-bandits. In *Proceedings of the 27th Annual Conference on Learning Theory (COLT)*, 2015a.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2015b.

Gergely Neu and Gábor Bartók. An efficient algorithm for learning with semi-bandit feedback. In *24th International Conference on Algorithmic Learning Theory (ALT)*, 2013.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Proceedings of the 26th Conference on Learning Theory (COLT)*, 2013.

Tim Roughgarden. Intrinsic robustness of the price of anarchy. *Journal of the ACM*, 2015.

Tim Roughgarden and Joshua R. Wang. Minimizing regret with multiple reserves. In *Proceedings of the 17th ACM Conference on Economics and Computation (EC)*, 2016.

Aristide C. Y. Tossou, Christos Dimitrakakis, and Devdatt Dubhashi. Thompson sampling for stochastic bandits with graph feedback. In *14th International Conference on Artificial Intelligence (AAAI)*, 2017.