An Optimal Algorithm for Linear Bandits

Nicolò Cesa-Bianchi
DSI, Università degli Studi di Milano
Italy

Sham Kakade
Microsoft Research New England
Wharton School, University of Pennsylvania
USA

Abstract

We provide the first algorithm for online bandit linear optimization whose regret after T rounds is of order $\sqrt{Td \ln N}$ on any finite class $\mathcal{X} \subseteq \mathbb{R}^d$ of N actions, and of order $d\sqrt{T}$ (up to log factors) when $\mathcal{X}$ is infinite. These bounds are not improvable in general. The basic idea utilizes tools from convex geometry to construct what is essentially an optimal exploration basis. We also present an application to a model of linear bandits with expert advice. Interestingly, these results show that bandit linear optimization with expert advice in d dimensions is no more difficult (in terms of the achievable regret) than the online d-armed bandit problem with expert advice (where EXP4 is optimal).

1 Introduction and Related Work

The problem of bandit linear optimization [1, 5, 6, 8] can be described as a repeated game between a forecaster and an opponent. The game is parameterized by a finite set $\mathcal{X} = \{\boldsymbol{x}(1), \ldots, \boldsymbol{x}(N)\} \subseteq \mathbb{R}^d$ of actions (we discuss the case where $\mathcal{X}$ is infinite later in the introduction). At each round $t = 1, 2, \ldots$ of the game, the forecaster chooses an action index $K_t \in \{1, \ldots, N\}$ and, simultaneously, the opponent chooses a payoff vector $\boldsymbol{y}_t \in \mathbb{R}^d$. The only information the forecaster receives after each round is the obtained payoff $\boldsymbol{y}_t^\top \boldsymbol{x}(K_T)$. More formally, the game is described as follows:

For each step $t = 1, 2, \ldots$
1. The opponent secretly chooses a payoff vector $\boldsymbol{y}_t \in \mathbb{R}^d$
2. The forecaster chooses $K_t \in \{1, \ldots, N\}$
3. The payoff $\boldsymbol{y}_t^\top \boldsymbol{x}(K_t)$ is announced to the forecaster.

Throughout this paper, we assume that $|\boldsymbol{y}_t^\top \boldsymbol{x}(k)| \leq 1$ for all $k = 1, \ldots, N$ and $t \geq 1$.

We consider randomized forecasters that, at each round t, choose a distribution p_{t-1} over $\{1, \ldots, N\}$ and then draw an action index K_t from p_{t-1}. The forecaster's goal is to control the *regret*

$$R_T = \max_{k=1,\ldots,N} \sum_{t=1}^{T} \boldsymbol{y}_t^\top \boldsymbol{x}(k) - \sum_{t=1}^{T} \boldsymbol{y}_t^\top \boldsymbol{x}(K_t)$$

in probability with respect to the forecaster's internal randomization. In this paper we focus on the weaker notion of *pseudoregret*

$$\overline{R}_T = \max_{k=1,\ldots,N} \sum_{t=1}^{T} \mathbb{E}\left[\boldsymbol{y}_t^\top \boldsymbol{x}(k) - \boldsymbol{y}_t^\top \boldsymbol{x}(K_t)\right] .$$

Though $\overline{R}_T \leq R_T$, the techniques of [5] imply that the pseudoregret guarantees obtained in this work can be extended to give regret bounds that hold with high probability.

This work provides an algorithm with essentially optimal regret, both in terms of T and d. In particular, [8] shows that for all $d \geq 1$ there exists a set $\mathcal{X} \subseteq \mathbb{R}^d$ of size $N = 2^d$ such that $\overline{R}_T = \Omega\big(d\sqrt{T}\big)$. Hence, our upper bound $\mathcal{O}\big(\sqrt{Td\ln N}\big)$ is not improvable, in general.

1.1 A template algorithm based on hedge

Two previously proposed forecasting strategies for this problem are GeometricHedge [8] and its variant ComBand [6]. Both strategies extend to the linear bandit case the Exp3 strategy of [2], and choose the sampling distribution p_t as a mixture $(1-\gamma)q_t + \gamma\,\mu$. The fixed "exploration distribution" μ is used to control the range of the estimates of the payoff vectors, and the way μ is defined is the only difference between the two forecasting strategies. In particular, GeometricHedge chooses μ to be uniform over a specific subset of actions (the barycentric spanners of $\mathcal{X}$, defined later). ComBand, instead, sets μ to the uniform distribution over the entire action set.

We now introduce GGH, a template for both strategies in which the exploration distribution is a parameter of the algorithm.

```
Algorithm:       GGH (Generalized GeometricHedge)
Parameters:      Finite action set $\mathcal{X} = \{\boldsymbol{x}(1), \dots, \boldsymbol{x}(N)\} \subseteq \mathbb{R}^d$
                 Exploration distribution $\mu$ over $\{1, \dots, N\}$
                 Mixing coefficient $\gamma > 0$
                 Learning rate $\eta > 0$
Initialization:  $q_0 =$ uniform distribution over $\{1, \dots, N\}$
For $t = 1, 2, \dots$
1. Let $p_{t-1} = (1-\gamma)q_{t-1} + \gamma\,\mu$
2. Draw action $K_t$ from $p_{t-1}$
3. Gain and observe payoff $\boldsymbol{y}_t^\top \boldsymbol{x}(K_t)$
4. Let $P_{t-1} = \mathbb{E}\big[\boldsymbol{X}\,\boldsymbol{X}^\top\big]$ where $\boldsymbol{X}$ has law $p_{t-1}$
5. Let $\widehat{\boldsymbol{y}}_t = P_{t-1}^{+}\boldsymbol{x}(K_t)\,\boldsymbol{x}(K_t)^\top \boldsymbol{y}_t$
6. Update  $q_t(k) \propto q_{t-1}(k)\exp\Big(-\eta\,\widehat{\boldsymbol{y}}_t^\top \boldsymbol{x}(k)\Big)$ for all $k = 1, \dots, N$.
```

Note that the unknown payoff vector $\boldsymbol{y}_t$ at time t is approximated using the "least square" estimate $\widehat{\boldsymbol{y}}_t = P_{t-1}^{+}\boldsymbol{x}(K_t)\,\boldsymbol{x}(K_t)^\top \boldsymbol{y}_t$, where $P_{t-1} = \mathbb{E}\big[\boldsymbol{X}\,\boldsymbol{X}^\top\big]$ and $\boldsymbol{X}$ is distributed according to the sampling distribution p_{t-1} over $\mathcal{X}$. A key idea in this algorithm is that $\widehat{\boldsymbol{y}}_t^\top \boldsymbol{x}$ is an unbiased estimate of $\boldsymbol{y}_t^\top \boldsymbol{x}$. Namely, $\mathbb{E}\big[\widehat{\boldsymbol{y}}_t^\top \boldsymbol{x}\big] = \boldsymbol{y}_t^\top \boldsymbol{x}$ for $\boldsymbol{x} \in \mathcal{X}$, where the expectation is w.r.t. p_{t-1} (which is straightforward to show —see [6, 8]). The pseudoregret of GGH is controlled by the variance of this estimate, and the analysis in [6, 8] shows, for a specific choice of η, the bound

$$\max_{\boldsymbol{x}\in\mathcal{X}} \big|\boldsymbol{x}^\top \widehat{\boldsymbol{y}}_t\big| \leq \frac{B^2}{\gamma\lambda_{\min}} . \tag{1}$$

Here B is an upper bound on the norm of any vector in $\mathcal{X}$ and $\lambda_{\min}$ is the smallest nonzero eigenvalue of $A_\mu = \mathbb{E}\big[\boldsymbol{Z}\,\boldsymbol{Z}^\top\big]$ with $\boldsymbol{Z}$ distributed according to the fixed exploration distribution μ. As shown in [6, 8], a suitable choice of the mixing coefficient γ guarantees a pseudoregret bound of the form

$$\overline{R}_T \leq 2\sqrt{\left(\frac{2B^2}{d\,\lambda_{\min}} + 1\right) Td\ln N} . \tag{2}$$

Since the trace of A_μ is at most B^2, if the spectrum of A_μ is approximately uniform, then $\lambda_{\min}$ is close to B^2/d and GGH achieves the optimal pseudoregret bound of order $\sqrt{Td\ln N}$. ComBand chooses μ to be uniform over $\mathcal{X}$, which amounts to betting that $\mathcal{X}$ is already nearly isotropic. In [6] the authors show different concrete choices of $\mathcal{X}$ where $\lambda_{\min} = \Omega\big(B^2/d\big)$.

Unlike ComBand, GeometricHedge uses a preprocessing step to make $\mathcal{X}$ look isotropic. Since the regret does not depend on the choice of the basis for $\mathbb{R}^d$, one can pick a convenient basis in which to express both actions and payoff vectors. The hope is that, when the matrix A_μ is expressed in this basis and μ is uniform over the same basis, $\lambda_{\min}$ becomes as close as possible to B^2/d.

The basis used by GeometricHedge are the barycentric spanners [3] of $\mathcal{X}$. This is a subset $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_d \in \mathcal{X}$ (without loss of generality, assume $\mathcal{X}$ spans $\mathbb{R}^d$) with the following properties:

1. $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_d\}$ spans $\mathcal{X}$.
2. For all $i = 1, \ldots, d$ and $\boldsymbol{x} \in \mathcal{X}$, the i-th coordinate of $\boldsymbol{x}$ in the spanner basis, denoted by $\boldsymbol{x}(i)$, satisfies $\boldsymbol{x}(i) \in [-1, 1]$.

An elegant algebraic argument from [3] shows that barycentric spanners always exist.

GeometricHedge chooses μ to be the uniform distribution over the spanners and performs an eigendecomposition of the symmetric matrix A_μ with respect to the spanner basis and the induced scalar product $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_{i=1}^d \boldsymbol{u}(i)\, \boldsymbol{v}(i)$. Then, if $\boldsymbol{e}$ is the eigenvector associated with any eigenvalue λ of that matrix,

$$\lambda = \sum_{i=1}^{d} \frac{1}{d} \langle \boldsymbol{s}_i, \boldsymbol{e} \rangle^2 = \frac{1}{d}$$

due to the orthonormality of eigenvectors (irrespective of the orthonormality of the spanner basis). Thus $\mathcal{X}$ is isotropic under the uniform measure over the spanners. However, by definition of barycentric spanner, $\boldsymbol{x}(i) \in [-1, 1]$. This implies $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \le d$ for all $\boldsymbol{x} \in \mathcal{X}$. Hence $B = \sqrt{d}$, and, from (2), we have $\overline{R}_T \le 2\sqrt{(2d+1)Td \ln N}$, a suboptimal bound. Using ideas from convex geometry, this work provides an optimal exploration strategy which simultaneously makes $\mathcal{X}$ isotropic while controlling the maximal norm over the set of actions.

1.2 Infinite decisions sets and computational efficiency

If $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$, then, as shown in [8], any bandit forecaster for finite action classes can be applied to infinite action sets via discretization; in particular, [8] show this construction results in pseudoregret of order $d^{3/2}\sqrt{T}$ using $(4dT)^{d/2}$ discretized actions.

With regards to computationally efficiency, when the size of $\mathcal{X}$ is infinite or exponential in d, GGH may not have an efficient implementation in general, though for important special cases (such as shortest path problems and certain dynamic programming problems) GGH can be implemented efficiently —see [6, 8]. For computationally efficiency, a third, radically different, forecasting strategy for bandit linear optimization was proposed based on the use of interior-point methods with self-concordant functions [1]. The resulting bound on the pseudoregret is $16d\sqrt{\theta T \ln T}$, where $\theta = \mathcal{O}(d)$ is the parameter of the self-concordant function; the pseudoregret implied by [1] is of the order $d^{3/2}\sqrt{T}$ (ignoring log factors).

The results we provide here lead to a regret of $d\sqrt{T}$ (up to constants and log factors) for arbitrary decision spaces (with the same discretization argument from [8]), matching the lower bound in [8]. An open question is if this rate is achievable with an efficient algorithm, which we discuss later.

2 An optimal exploration distribution

We use the following result from convex geometry —see [4] for a proof.

Lemma 1 (John's theorem)**.** *Let $K \subseteq \mathbb{R}^d$ be a convex set. The volume-minimizing ellipsoid enclosing K is the (translated) unit ball in some norm $\|\cdot\|$ derived from a scalar product $\langle \cdot, \cdot \rangle$ if and only if there exists a probability distribution $p_1^*, \ldots, p_n^*$ over the contact points $\boldsymbol{x}_i^*, \ldots, \boldsymbol{x}_n^*$ between the ellipsoid and K such that*

$$\sum_{i=1}^{n} p_i^* \langle \boldsymbol{x}_i^*, \boldsymbol{v} \rangle^2 = \frac{1}{d} \qquad \textit{for all } \boldsymbol{v} \in \mathbb{R}^d \textit{ such that } \|\boldsymbol{v}\| = 1.$$

Let $\mathcal{E} = \left\{ \boldsymbol{v} \in \mathbb{R}^d : (\boldsymbol{v} - \boldsymbol{x}_0)^\top H^{-1}(\boldsymbol{v} - \boldsymbol{x}_0) \le 1 \right\}$ be the volume-minimizing ellipsoid enclosing $\mathcal{X}$. This is often called the Löwner-John ellipsoid in the literature. Define a new system of coordinates where $\boldsymbol{v} \in \mathbb{R}^d$ is represented as $\boldsymbol{v}' = H^{-1}(\boldsymbol{v} - \boldsymbol{x}_0)$. Define the scalar product $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_H = \boldsymbol{u}^\top H \boldsymbol{v}$. It is easy to see that in the new system of coordinates $\mathcal{E}$ is the unit ball with respect to the above scalar product. Namely, if $\boldsymbol{v} \in \mathcal{E}$ then $\langle \boldsymbol{v}', \boldsymbol{v}' \rangle_H = (\boldsymbol{v} - \boldsymbol{x}_0)^\top H^{-1} (\boldsymbol{v} - \boldsymbol{x}_0) \le 1$. Now Lemma 1 implies that we can perform exploration in an optimal way.

Theorem 2. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ with $|\mathcal{X}| = N$ and assume $|\boldsymbol{y}_t^\top \boldsymbol{x}| \le 1$ for all $\boldsymbol{x} \in \mathcal{X}$ and $t \ge 1$. Suppose GGH is run on $\mathcal{X}' \equiv \big\{ H^{-1}\big(\boldsymbol{x} - \boldsymbol{x}_0\big) \,:\, \boldsymbol{x} \in \mathcal{X} \big\}$ with scalar product $\langle \cdot, \cdot \rangle_H$ and exploration distribution $\mu = (p_1^*, \dots, p_n^*)$ over the contact points $\{\boldsymbol{x}_i^*, \dots, \boldsymbol{x}_n^*\} \subseteq \mathcal{X}'$. If GGH parameters are chosen as $\eta = \sqrt{(\ln N)\big/(3dT)}$, $\gamma = d\eta$, then its pseudoregret after T steps satisfies $\overline{R}_T \le \sqrt{12Td \ln N}$.*

Proof. Note that $\langle \boldsymbol{y}_t, \boldsymbol{x}' \rangle_H = \boldsymbol{y}_t^\top H \, H^{-1} (\boldsymbol{x} - \boldsymbol{x}_0) = \boldsymbol{y}_t^\top \boldsymbol{x} - \boldsymbol{y}_t^\top \boldsymbol{x}_0$ for any $\boldsymbol{x}' \in \mathcal{X}'$. Since the regret is invariant with respect to the choice of the origin, we do not lose any generality by running GGH on $\mathcal{X}'$ and computing payoffs as $\langle \boldsymbol{y}_t, \boldsymbol{x}' \rangle_H$. Since $\mathcal{X}'$ is enclosed by the unit ball, we immediately have $\langle \boldsymbol{x}', \boldsymbol{x}' \rangle_H \le 1$ for all $\boldsymbol{x}' \in \mathcal{X}'$, so that $B = 1$. Moreover, the choice of μ and Lemma 1 ensures that $\langle \boldsymbol{v}, A_\mu \boldsymbol{v} \rangle_H = \sum_{i=1}^n p_i \, \langle \boldsymbol{x}_i^*, \boldsymbol{v} \rangle_H^2 = \frac{1}{d}$ for all $\boldsymbol{v} \in \mathbb{R}^d$ such that $\langle \boldsymbol{v}, \boldsymbol{v} \rangle_H = 1$. This implies $\lambda_{\min} = \frac{1}{d}$. From (2) we then obtain the desired bound. □

3 Computational issues

If $\mathcal{X}$ is given by a finite set of points, then there is a polynomial time algorithm for computing a constant factor approximation to the Löwner-John ellipsoid [9] (and this approximate basis will provide the same order of regret). However, if $\mathcal{X}$ is specified by the intersection of half spaces, then obtaining such a constant factor approximation to this ellipsoid is NP-hard in general [10]. Here, it is possible to efficiently compute an ellipsoid where the factor of $\frac{1}{d}$ is replaced by $\frac{1}{d^{3/2}}$ in Lemma 1 [9], which leads to a slightly worse dependence on d in the regret bound.

In special cases, we conjecture that Löwner-John ellipsoid may be computed efficiently, as for certain problems, there are efficient implementations of GeometricHedge that lead to optimal rates (such as shortest path problems and other settings where dynamic programming solutions exists).

4 Application to bandits with experts

Consider the following model of linear bandits with N experts. At each time step $t = 1, 2, \dots$, each expert $i = 1, \dots, N$ suggests an action $\boldsymbol{x}_t(i) \in \mathbb{R}^d$. The pseudoregret of a randomized linear bandit algorithm choosing expert $I_t \in \{1, \dots, N\}$ a time t is defined by

$$\overline{R}_T^{\mathrm{exp}} = \max_{i=1,\dots,N} \sum_{t=1}^T \mathbb{E}\big[\boldsymbol{y}_t^\top \boldsymbol{x}_t(i) - \boldsymbol{y}_t^\top \boldsymbol{x}_t(I_t)\big]$$

where the expectation is w.r.t. the algorithm's internal randomization and $\boldsymbol{y}_1, \boldsymbol{y}_2, \dots \in \mathbb{R}^d$ is any sequence of payoff vectors such that $\big|\boldsymbol{y}_t^\top \boldsymbol{x}_t(i)\big| \le 1$ for all $t = 1, 2, \dots$ and $i = 1, \dots, N$.

Corollary 3. *There exists a randomized algorithm for linear bandits with experts such that $\overline{R}_T^{\mathrm{exp}} \le \sqrt{12Td \ln N}$.*

Proof. We apply a variant of GGH in which the exploration distribution μ changes over time. Let $\mathcal{X}_t = \{\boldsymbol{x}_t(1), \dots, \boldsymbol{x}_t(N)\} \subseteq \mathbb{R}^d$ and let $\mathcal{X}_t' \equiv \big\{ H_t^{-1}\big(\boldsymbol{x}_t(k) - \boldsymbol{x}_{t,0}\big) \,:\, k = 1, \dots, N \big\}$, where H_t and $\boldsymbol{x}_{t,0}$ are defined with respect to the the volume-minimizing ellipsoid $\mathcal{E}_t$ enclosing $\mathcal{X}_t$. The algorithm chooses μ_t defined by the contact points for $\mathcal{X}_t'$ and then draws an action by calling GGH with μ_t as exploration distribution. Then, it feeds to GGH the payoff vector estimate at time t using $\langle \cdot, \cdot \rangle_H$. Since condition (1) holds at each time step, the pseudoregret after T rounds has the same bound as GGH (Theorem 2). □

For example, let each expert $i = 1, \dots, N$ be associated with a fixed payoff vector $\boldsymbol{y}(i) \in \mathbb{R}^d$ and the suggested action be defined by $\boldsymbol{x}_t(i) = \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{S}_t} \boldsymbol{y}(i)^\top \boldsymbol{x}$ where $\mathcal{S}_t \subseteq \mathbb{R}^d$ is an arbitrary "context set". This can be viewed as a natural nonstochastic variant of the contextual linear bandit model of [7]. Another notable special case is the d-armed bandit problem with expert advice, where we can view the suggested actions as the corners of the d-dimensional simplex. Here, the EXP4 algorithm of [2] achieves a regret of order $\sqrt{Td \ln N}$. Interestingly, the regret achievable in the more general d-dimensional linear optimization setting is no worse than in the d-armed bandit setting.

Acknowledgements

We thank Sébastien Bubeck for suggesting improvements in the presentation. The first author gratefully acknowledges partial support by the PASCAL2 Network of Excellence under EC grant no. 216886.

References

[1] J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 263–274. Omnipress, 2008.

[2] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[3] B. Awerbuch and R.D. Kleinberg. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. In *Proceedings of the 36th ACM Symposium on the Theory of Computing*. ACM Press, 2004.

[4] K. Ball. *An Elementary Introduction to Modern Convex Geometry*, volume 31 of *Flavors of Geometry*. MSRI Publications, 1997.

[5] P. Bartlett, V. Dani, T. Hayes, S.M. Kakade, A. Rakhlin, and A. Tewari. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 335–342. Omnipress, 2008.

[6] N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 2011. To appear.

[7] W. Chu, L. Li, L. Reyzin, and R.E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Conference and Workshop Proceedings, 2011.

[8] V. Dani, T.P. Hayes, and S.M. Kakade. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems 22*, pages 345–352. MIT Press, 2008.

[9] Martin Grötschel, Lászlo Lovász, and Alexander Schrijver. *Geometric Algorithms and Combinatorial Optimization*, volume 2 of *Algorithms and Combinatorics*. Springer, second corrected edition edition, 1993.

[10] A. Nemirovski. Advances in convex optimiza- tion: Conic programming. In *Proceedings of the International Congress of Mathematicians, 2006*. EMS-European Mathematical Society Publishing House, 2007.