Journal of Machine Learning **ISSN**: 2790-2048(e), 2790-203X(p)

A Note on Continuous-Time Online Learning

Lexing Ying *

Department of Mathematics, Stanford University, Stanford, CA 94305, USA.

Abstract. In online learning, the data is provided in sequential order, and the goal of the learner is to make online decisions to minimize overall regrets. This note is concerned with continuous-time models and algorithms for several online learning problems: online linear optimization, adversarial bandit, and adversarial linear bandit. For each problem, we extend the discrete-time algorithm to the continuous-time setting and provide a concise proof of the optimal regret bound.

Keywords:
Online learning,
Online optimization,
Adversarial bandits,
Adversarial linear bandits.

Article Info.:
Volume: X
Number: X
Pages: 1 - 10
Date: /2024
doi.org/10.4208/jml.240605

Article History:
Received: 05/06/2024
Accepted: 30/09/2024

Communicated by:
Qianxiao Li

1 Introduction

In online learning, the data is provided in sequential order, and the goal of the learner is to make online decisions to minimize overall regrets. This is particularly relevant for problems with a dynamic aspect. This topic has produced many surprisingly efficient algorithms that are nothing short of magic.

This note is concerned with several important online learning problems:

- online linear optimization,
- adversarial bandit,
- adversarial linear bandit.

For each problem, we define a regret that quantifies how much worse a learning algorithm's performance is compared to the best fixed strategy in hindsight. In most of the existing literature, online learning problems are often placed in the discrete-time setting, and many discrete-time algorithms have been developed to achieve optimal regret bounds. However, there has been relatively little work for online learning in the continuous-time setting. In this note, for each of these problems, we propose a continuous-time model, describe an algorithm motivated by the discrete-time version, and provide a simple proof for the optimal regret bound. The main technical tools are Legendre transform and Ito's lemma.

Several books, reviews, and lecture notes are devoted to online learning [1, 2, 4, 6, 8, 11, 12] in the discrete-time setting. In recent years, there has been a growing interest in

*Corresponding author. lexing@stanford.edu

the continuous-time setting [3, 5, 10, 13, 14]. Among them, [3, 13] proposed diffusion approximations for Thompson sampling algorithms for multi-arm bandits; [5, 14] developed continuous models based on Hamilton-Jaboi-Bellman equation for two-armed bandits; and [10] proposed the first continuous prediction models for the experts' advice setting. Our result for the adversarial bandit problem is closely related to the work in [10].

The rest of this note is organized as follows. Section 2 summarizes the main results of the Legendre transform. Section 3 discusses the online linear optimization problem. Section 4 presents the continuous-time model for the adversarial bandit. Section 5 extends the result to the adversarial linear bandit.

2 Legendre transform

Let X be a convex set in $\mathbb{R}^d$ and $F(x)$ be a convex function defined on X. To simplify the discussion, we assume that $F(x)$ is strictly convex.

The Legendre transform [7], denoted by $G(y)$, of $F(x)$, is defined as

$$G(y) \equiv \max_{x \in X} y^\top x - F(x),$$

where the domain Y of the set where $G(y)$ is bounded.

Let $x(y)$ be the point where the maximum is achieved for a given y. Then

$$y = \nabla F\big(x(y)\big).$$

A key result of Legendre transform is that $F(x)$ is also the Legendre transform of $G(y)$

$$F(x) = \max_{y \in Y} x^\top y - G(y)$$

and, similarly for a given x, the maximizer $y(x)$ satisfies

$$x = \nabla G\big(y(x)\big).$$

A trivial but useful inequality is

$$F(x) + G(y) \geq x^\top y.$$

In this note, we are concerned with the following case:

$$X = \Delta^d \equiv \left\{ (x_1, \dots, x_d) : x_a \geq 0, \ \sum_{i=1}^{d} x_i = 1 \right\}, \quad Y = \mathbb{R}^d$$

with $F(x)$ and $G(y)$ given by

$$F(x) = \beta^{-1} \sum_{i=1}^{d} x_i \ln x_i, \quad G(y) = \beta^{-1} \ln \left(\sum_{i=1}^{d} \exp(\beta y_i) \right) \tag{2.1}$$

with $\beta > 0$.

Notations. The calculations in this note involve explicit manipulations of the entries of vectors and matrices. For a vector v, we will use v_i to denote its i-th coordinate. For a matrix M, M_{ij} denotes its (i,j)-th component. On the other hand, when given an array M_{ij} of scalars, we use $[M_{ij}]$ to denote the associated matrix. For example, we shall encounter at several places the gradient and the Hessian of $G(y)$ in (2.1). For example, a direct calculation shows that, in the component form,

$$\begin{aligned}\nabla G(y)_i &= \frac{\exp(\beta y_i)}{\sum_j \exp(\beta y_j)}, && 1 \le i \le d,\\ \nabla^2 G(y)_{ij} &= \beta\big(\delta_{ij}\nabla G(y)_i - \nabla G(y)_i \nabla G(y)_j\big), && 1 \le i,j \le d.\end{aligned}$$

3 Online linear optimization

Discrete-time problem. The discrete-time online linear optimization [15] is stated as follows. At each round $t = 1, \dots, T$,

- the learner picks $x_t \in X = \Delta^d$,
- the adversary picks a reward vector $r_t \in [0,1]^d$,
- the learner observes r_t and gets reward $x_t^\top r_t$.

The regret is defined as

$$R = \max_{x \in X} \sum_t (x - x_t)^\top r_t.$$

Remark 3.1. In most of the literature, the problem is formulated as minimizing the loss rather than maximizing the reward. These two formulations are equivalent. We choose the latter to put the problem into a Legendre transform setting.

Discrete-time algorithm. An optimal algorithm for this problem is called follow-the-regularized-leader [15]. At each round $t = 1, \dots, T$, the learner chooses

$$x_t = \operatorname*{argmax}_{x \in X} \left(x^\top \left(\sum_{z=1}^{t-1} r_z \right) - F(x) \right)$$

for $F(x)$ in (2.1). A direct computation shows that

$$x_t \propto \exp\left(\beta \sum_{z=1}^{t-1} r_z \right).$$

This algorithm is also called the multiplicative weights method, since the exponent of x_t can also be updated as $x_t \propto x_{t-1} \exp(\beta r_{t-1})$ at each round t. By setting $\beta = \sqrt{2 \ln d / T}$ in (2.1), the discrete-time regret can be bounded by $\mathcal{O}(\sqrt{T \ln d})$ [6].

Continuous-time problem. The continuous-time model is stated as follows. At each time $t \in [0, T]$,

- the learner picks $x(t) \in X$,
- the adversary picks reward $r(t) \in [0,1]^d$,
- the learner observes $r(t)$ and gets reward $x(t)^\top r(t)$.

By introducing the cumulative reward

$$s(t) \equiv \int_0^t r(z)dz,$$

the continuous-time regret is

$$R = \max_{x\in X} \left(x^\top s(T) - \int_0^T x(t)^\top ds(t) \right).$$

Continuous-time algorithm. Motivated by the discrete-time setting, we set the action at time t as

$$x(t) = \text{argmax}_{x\in X} \left(x^\top s(t) - F(x) \right).$$

By Legendre transform,

$$x(t) = \nabla G\big(s(t)\big).$$

Continuous-time regret bound. The regret analysis is particularly simple in the continuous-time setting.

Theorem 3.1. *For any $\beta > 0$, the continuous-time regret is bounded by $\beta^{-1} \ln d$.*

Proof. For any $x \in X$,

$$\begin{aligned}
&x^\top s(T) - \int_0^T x^\top ds(t) \\
&= x^\top s(T) - \int_0^T \nabla G\big(s(t)\big)^\top ds(t) \\
&= x^\top s(T) - G\big(s(T)\big) + G(0) \\
&\le F(x) + G(0) \le \beta^{-1} \ln d.
\end{aligned}$$

Here we used the facts that $x^\top s \le G(s) + F(x)$, $F(x) \le 0$, and $G(0) = \beta^{-1} \ln d$. □

Remark 3.2. By letting β approach infinity, the regret goes to zero. This shows that in the continuous-time case, following the leader (i.e. $\beta = \infty$) is, in fact, the optimal strategy. This is different from the discrete-time setting.

4 Adversarial bandit

Discrete-time problem. The discrete-time setting is stated as follows. The arms are indexed by $\{1, \ldots, d\}$. At the beginning, the adversarial chooses the rewards $r_1, \ldots, r_t \in [0,1]^d$. In each round $t = 1, \ldots, T$,

- the learner picks arm a_t,
- the learner gets reward r_{t,a_t} (but without knowing other components of r_t).

Since the arm can be chosen randomly, the regret is defined as

$$R = \max_i \mathbb{E} \sum_t (r_{t,i} - r_{t,a_t}).$$

Discrete-time algorithm. The algorithm performs two tasks at each round t:

(1) Computing a probability distribution p_t for selecting the arm.

(2) Forming an estimate $\hat{r}_t$ of r_t based on the only reward r_{t,a_t} received.

Assuming that $\hat{r}_z$ are available for time $z < t$, the algorithm defines the probability

$$p_{t,i} \propto \exp\left(\beta \sum_{z=1}^{t-1} \hat{r}_{z,i}\right)$$

for an appropriate β value and the reward estimate

$$\hat{r}_{t,i} = \begin{cases} \dfrac{r_{t,a_t}}{p_{t,a_t}}, & i = a_t, \\ 0, & \text{otherwise.} \end{cases}$$

From the properties of the Legendre transform,

$$p_t = \nabla G\left(\sum_{z=1}^{t-1} \hat{r}_z\right).$$

The random estimate $\hat{r}_t$ is unbiased

$$\mathbb{E}\hat{r}_t = r_t,$$

and its covariance matrix Σ_t has entries given by

$$\Sigma_{t,ij} = \frac{r_{t,i}^2}{p_i}\delta_{ij} - r_{t,i}r_{t,j}.$$

By setting $\beta = \sqrt{2\ln d/dT}$, the discrete-time regret can be bounded by $\sqrt{2Td\ln d}$ [6].

Continuous-time problem. The continuous-time model is stated as follows. At the beginning, the adversarial chooses the rewards $r(t) \in [0,1]^d$ for $0 \le t \le T$. At each time $t \in [0,T]$,

- the learner picks arm $a(t)$,
- the learner gets reward $r(t)_{a(t)}$ (but without knowing other components of $r(t)$).

The continuous-time regret is defined as

$$R = \max_i \mathbb{E} \int_0^T \left(r(t)_i - r(t)_{a(t)}\right) dt.$$

Continuous-time algorithm. Motivated by the discrete-time algorithm, we adopt the reward estimate

$$\hat{r}(t)_i = \begin{cases} \dfrac{r(t)_{a(t)}}{p(t)_{a(t)}}, & i = a(t), \\ 0, & \text{otherwise.} \end{cases}$$

Then the cumulative reward estimate $s(t) \in \mathbb{R}^d$ follows the following stochastic differential equation (SDE) [9]:

$$ds(t) = r(t)dt + \sigma(t)dB(t),$$

where $\sigma(t)\sigma(t)^\top = \Sigma(t)$ with

$$\Sigma(t)_{ij} = \frac{r(t)_i^2}{p(t)_i}\delta_{ij} - r(t)_i r(t)_j.$$

The probability of choosing arm i at time t is

$$p(t)_i \propto \exp\left(\beta s(t)_i\right).$$

Notice that $p(t) = \nabla G(s(t))$.

Continuous-time regret bound. The following theorem states the regret bound of the continuous-time algorithm after optimizing β.

Theorem 4.1. *For $\beta = \sqrt{2\ln d / dT}$, the continuous-time regret is bounded by $\sqrt{2Td\ln d}$.*

Proof. For an arbitrary arm a, let $x = (0, \ldots, 1, \ldots, 0)^\top$ with 1 at the a-th slot. The regret with respect to a can be recast as

$$\mathbb{E}\left(x^\top s(T) - \int_0^T \nabla G\big(s(t)\big)^\top ds(t)\right).$$

In order to bound $\int_0^T \nabla G(s(t))^\top ds(t)$, we use Ito's lemma

$$dG(s) = \nabla G(s)^\top ds + \frac{1}{2} ds^\top \nabla^2 G(s) ds.$$

The second (quadratic variation) term can be written as

$$\begin{aligned} \frac{1}{2}\operatorname{tr}\left(ds ds^\top \nabla^2 G\right) &= \frac{1}{2}\operatorname{tr}\left(\Sigma(t)\nabla^2 G\right)dt \\ &= \frac{\beta}{2}\operatorname{tr}\left(\left[\frac{r(t)_i^2}{p(t)_i}\delta_{ij} - r(t)_i r(t)_j\right]\left[\delta_{ij}p(t)_i - p(t)_i p(t)_j\right]\right)dt. \end{aligned}$$

Using the facts that both matrices are symmetric nonnegative definite and that the product only increases if one makes the matrices more positive definite, we can bound this by

$$\frac{\beta}{2}\operatorname{tr}\left(\left[\frac{r(t)_i^2}{p(t)_i}\delta_{ij}\right]\left[\delta_{ij}p(t)_i\right]\right)dt \le \frac{\beta d}{2}dt,$$

where we used $r(t) \in [0,1]^d$. From this, we can bound the regret by

$$\begin{aligned}
&\mathbb{E}\left(x^\top s(T) - \int_0^T dG\big(s(t)\big)\right) + \frac{\beta d T}{2} \\
&= \mathbb{E}\left(x^\top s(T) - G\big(s(T)\big)\right) + G(0) + \frac{\beta d T}{2} \\
&\le F(x) + G(0) + \frac{\beta d T}{2} \le \beta^{-1}\ln d + \frac{\beta d T}{2}.
\end{aligned}$$

By choosing $\beta = \sqrt{2\ln d/dT}$, the regret can be bounded by $\sqrt{2Td\ln d}$. □

5 Adversarial linear bandit

In practice, there might be many but correlated arms. A common setting is an arm set $\mathcal{A} = \{a\} \subset \mathbb{R}^d$ with $|\mathcal{A}| = k \gg d$. Assume that each arm $a \in \mathbb{R}^d$ satisfies $\|a\|_1 \le 1$.

Discrete-time problem. At the beginning, the adversarial chooses the rewards $r_1, \ldots, r_T \in [0,1]^d$. In each round $t = 1, \ldots, T$,

- the learner picks arm $a_t \in \mathcal{A}$,
- the learner gets reward $a_t^\top r_t$ (but without knowing other information about r_t).

The discrete-time regret is defined as

$$R = \max_{a\in\mathcal{A}} \mathbb{E}\sum_t (a - a_t)^\top r_t.$$

Discrete-time algorithm. The discrete-time algorithm performs two tasks at round t:

(1) Computing a probability distribution p_t for selecting an arm.

(2) Forming an estimate $\hat{r}_t$ for r_t based only on $a_t^\top r_t$.

Assuming that $\hat{r}_z$ are available for time $z < t$, the algorithm defines the probability

$$p_{t,a} \propto \exp\left(\beta \sum_{z=1}^{t-1} \hat{r}_z^\top a\right)$$

for some $\beta > 0$ and the reward estimate

$$\hat{r}_t = Q_t^{-1} a_t \big(a_t^\top r_t\big), \quad Q_t = \sum_{a\in\mathcal{A}} p_{t,a} a a^\top.$$

The random estimate $\hat{r}_t$ is unbiased

$$\mathbb{E}\hat{r}_t = r_t,$$

and it has a covariance matrix $\Sigma_t \in \mathbb{R}^{d\times d}$ given by

$$\Sigma_t = \sum_{a\in\mathcal{A}} p_{t,a} Q_t^{-1} a \big(a^\top r_t\big)\big(r_t^\top a\big) a^\top Q_t^{-1} - r_t r_t^\top.$$

By setting $\beta = \mathcal{O}(\sqrt{\ln k/dT})$ and including appropriate exploration, the regret can be bounded by $\mathcal{O}(\sqrt{Td\ln k})$ [6]. Notice that it depends only logarithmically on the number of arms k.

Continuous-time problem. The continuous-time model is given as follows. At the beginning, the adversarial chooses the rewards $r(t) \in [0,1]^d$ for $0 \le t \le T$. At each time $t \in [0,T]$,

- the learner picks arm $a(t)$,
- the learner gets reward $a(t)^\top r(t)$ (but without knowing other information of $r(t)$).

The continuous-time regret is defined as

$$R = \max_{a\in A} \mathbb{E} \int_0^T \big(a - a(t)\big)^\top r(t)dt.$$

Continuous-time algorithm. Motivated by the discrete-time algorithm, we use the reward estimate

$$r(t) = Q(t)^{-1}a(t)\big(a(t)^\top r(t)\big), \quad Q(t) = \sum_{a\in\mathcal{A}} p(t)_a a a^\top.$$

Then the cumulative reward estimate $s(t) \in \mathbb{R}^d$ follows the following stochastic differential equation:

$$ds(t) = r(t)dt + \sigma(t)dB(t),$$

where $\sigma(t)\sigma(t)^\top = \Sigma(t)$ with

$$\Sigma(t) = \sum_a p(t)_a Q(t)^{-1} a\big(a^\top r(t)\big)\big(r(t)^\top a\big) a^\top Q(t)^{-1} - r(t)r(t)^\top.$$

The probability of choosing arm a at time t is

$$p(t)_a \propto \exp\big(\beta a^\top s(t)\big).$$

Continuous-time regret bound. Let A be the $k \times d$ matrix with rows given by $a^\top$. Instead of defined over $\mathbb{R}^d$, we redefine $F(x)$ and $G(x)$ over $\mathbb{R}^k$

$$F(x) = \beta^{-1}\sum_{i=1}^k x_i \ln x_i, \quad G(y) = \beta^{-1}\ln\left(\sum_{i=1}^k \exp(\beta y_i)\right).$$

Notice that $G(0) = \beta^{-1}\ln k$ and now $p(t) = \nabla G(As(t)) \in \mathbb{R}^k$.

Theorem 5.1. *For $\beta = \sqrt{2\ln k/dT}$, the continuous-time regret is bounded by $\sqrt{2Td\ln k}$.*

Proof. For an arbitrary arm a, let $x = (0,\ldots,1,\ldots,0)^\top$ with 1 at the a-th slot. The regret with respect to a can be written as

$$\mathbb{E}\left(x^\top As(T) - \int_0^T \nabla G\big(As(t)\big)^\top d\big(As(t)\big)\right).$$

To bound $\int_0^T \nabla G(As(t))^\top d(As(t))$, we again invoke Ito's lemma

$$dG(As) = \nabla G(s)^\top A ds + \frac{1}{2} ds^\top A^\top \nabla^2 G(As) A ds.$$

The second (quadratic variation) term can be written as (hiding the t dependence)

$$\begin{aligned}
&\frac{1}{2} \operatorname{tr}\left(A ds ds^\top A^\top \nabla^2 G(As)\right) \\
=& \frac{1}{2} \operatorname{tr}\left(A \Sigma(t) A^\top \nabla^2 G(As)\right) dt \\
=& \frac{\beta}{2} \operatorname{tr}\left(A^\top [\delta_{ab} p_a - p_a p_b] A \left(\sum_a p_a Q^{-1} a (a^\top r)(r^\top a) a^\top Q^{-1}\right)\right) dt \\
\leq& \frac{\beta}{2} \operatorname{tr}\left(A^\top [\delta_{ab} p_a] A \left(Q^{-1} \sum_a a p_a a^\top Q^{-1}\right)\right) dt \leq \frac{\beta d}{2} dt.
\end{aligned}$$

Here we used $\|a^\top r\| \leq 1$, $A^\top [\delta_{ab} p_a] A = Q$, and $\sum_a a p_a a^\top = Q$.

From this, we can bound the regret by

$$\begin{aligned}
&\mathbb{E}\left(x^\top As(T) - \int_0^T dG\big(As(t)\big)\right) + \frac{\beta d T}{2} \\
=& \mathbb{E}\left(x^\top As(T) - G\big(As(T)\big)\right) + G(0) + \frac{\beta d T}{2} \\
\leq& F(x) + G(0) + \frac{\beta d T}{2} \leq \beta^{-1} \ln k + \frac{\beta d T}{2}.
\end{aligned}$$

By choosing $\beta = \sqrt{2 \ln k / dT}$, the regret can be bounded by $\sqrt{2Td \ln k}$. □

Notice again that the bound depends only logarithmically on the number of arms k.

6 Discussions

This note discusses continuous-time formulations and algorithms for several online learning problems. The main advantage of the continuous-time approach is that the proof of the regret bounds can be made concise. Many other online learning problems can be addressed similarly, including online convex optimization, semi-bandits, combinatorial bandits, and stochastic bandits.

Acknowledgments

This work is partially supported by the NSF (Grant Nos. DMS-2011699, DMS-2208163).

References

[1] S. Bubeck, Introduction to online optimization, *Lecture Notes*, 2:1–86, 2011.

[2] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*, Cambridge University Press, 2006.
[3] L. Fan and P. W. Glynn, Diffusion approximations for Thompson sampling, *arXiv:2105.09232*, 2021.
[4] E. Hazan, *Introduction to Online Convex Optimization*, MIT Press, 2022.
[5] V. A. Kobzar and R. V. Kohn, A PDE-based analysis of the symmetric two-armed Bernoulli bandit, *arXiv:2202.05767*, 2022.
[6] T. Lattimore and C. Szepesvári, *Bandit Algorithms*, Cambridge University Press, 2020.
[7] M. Levi, *Classical Mechanics with Calculus of Variations and Optimal Control: An Intuitive Introduction*, in: *Student Mathematical Library*, AMS, Vol. 69, 2014.
[8] P. Liang, Statistical Learning Theory (2016), `https://web.stanford.edu/class/cs229t/notes.pdf`, 2016.
[9] B. Oksendal, *Stochastic Differential Equations: An Introduction with Applications*, Springer Science & Business Media, 2013.
[10] V. S. Portella, C. Liaw, and N. J. Harvey, Continuous prediction with experts' advice, *arXiv:2206.00236*, 2022.
[11] A. Rakhlin, J. Abernethy, A. Agarwal, P. Bartlett, E. Hazan, and A. Tewari, *Lecture Notes on Online Learning DRAFT*, `https://www.mit.edu/~rakhlin/papers/online_learning.pdf`, 2009.
[12] S. Shalev-Shwartz et al., Online learning and online convex optimization, *Found. Trends Mach. Learn.*, 4(2):107–194, 2012.
[13] S. Wager and K. Xu, Diffusion asymptotics for sequential experiments, *arXiv:2101.09855*, 2021.
[14] Y. Zhu, Z. Izzo, and L. Ying, Continuous-in-time limit for Bayesian bandits, *J. Mach. Learn. Res.*, 24(305):1–35, 2023.
[15] M. Zinkevich, Online convex programming and generalized infinitesimal gradient ascent, in: *Proceedings of the 20th International Conference on Machine Learning*, AAAI Press, 928–936, 2003.