BISTRO: An Efficient Relaxation-Based Method for Contextual Bandits

Alexander Rakhlin RAKHLIN@WHARTON.UPENN.EDU
University of Pennsylvania

Karthik Sridharan SRIDHARAN@CS.CORNELL.EDU
Cornell University

Abstract

We present efficient algorithms for the problem of contextual bandits with i.i.d. covariates, an arbitrary sequence of rewards, and an arbitrary class of policies. Our algorithm BISTRO requires d calls to the empirical risk minimization (ERM) oracle per round, where d is the number of actions. The method uses unlabeled data to make the problem computationally simple. When the ERM problem itself is computationally hard, we extend the approach by employing multiplicative approximation algorithms for the ERM. The integrality gap of the relaxation only enters in the regret bound rather than the benchmark. Finally, we show that the adversarial version of the contextual bandit problem is learnable (and efficient) whenever the full-information supervised online learning problem has a non-trivial regret guarantee (and efficient).

1. Introduction

A multi-armed bandit with covariates (also known as a *contextual bandit*) is a generalization of the classical multi-armed bandit problem (Lai & Robbins, 1985). As the name suggests, in this natural formulation the quality of the arms may depend on the observed set of covariates. Contextual bandits arise in many application areas, from ad placement and news recommendation to personalized medical care and clinical trials. In recent years, there has been a strong push to develop computationally efficient regret minimization methods with respect to a given set of policies (Langford & Zhang, 2008; Dudik et al., 2011; Beygelzimer et al., 2011; Agarwal et al., 2014). The grand goal here would be to develop efficient and statistically optimal methods for large (and possibly uncountable) sets of policies, just as machine learning and statistics succeeded in developing methods that perform well relative to rich classes of predictors (linear separators, SVMs, and so forth). Compared to batch learning, however, the state of affairs at the moment is quite poor. It appears to be difficult to develop scalable methods even for a finite set of policies, as witnessed by the papers mentioned earlier. To some extent, the reason is not surprising: while in statistical learning the batch nature of the problem suggests the empirical objective to optimize, the scope of algorithms for contextual bandits is not at all clear.

Assuming access to an ERM (empirical risk minimization) optimization oracle, (Agarwal et al., 2014) exhibit a computationally attractive method for any finite class. The oracle model allows one to address the question of how much more difficult (computationally) the bandit problem is in comparison to the batch learning problem.

In the present paper, we introduce a family of efficient methods (and, more generally, a new algorithmic approach based on relaxations) for minimizing regret against a potentially uncountable class $\mathcal{F}$, given that the *value* of the ERM objective can be computed. In addition, we require access to i.i.d. draws of contexts (e.g. unlabeled data) — a realistic assumption in many application areas mentioned earlier. Our method requires only d oracle calls per round, irrespective of the size of the policy class. Furthermore, the results hold in the hybrid scenario where the contexts are i.i.d. but rewards evolve according to an arbitrary process.

Let us now describe the scenario in more detail. On each round $t = 1, \dots, n$, we observe covariates $x_t \in \mathcal{X}$, select an action $\widehat{y}_t \in \{1, \dots, d\} \triangleq [d]$, and observe the cost $c_t(\widehat{y}_t)$ of the chosen action. Here $c_t \in [0,1]^d$ is a cost assignment to all actions, chosen by Nature independently of $\widehat{y}_1, \dots, \widehat{y}_t$ (that is, c_t is oblivious to our randomization, but possibly a function of $x_1, \dots, x_t$). This cost vector remains unknown to us, except for the coordinate $c_t(\widehat{y}_t)$. Since we include randomized prediction methods, we denote the distribution over the d choices on round t by $q_t \in \Delta_d$, and draw $\widehat{y}_t \sim q_t$. The goal is to design a prediction method with small

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

expected cumulative cost $\sum_{t=1}^n q_t^\intercal c_t$.

We assume that $x_1, \dots, x_n$ are drawn i.i.d. from some unknown distribution P_x on $\mathcal{X}$. At the same time, we do not place any assumption on the sequence of costs $c_1, \dots, c_n$, which may evolve according to some arbitrary stochastic process, or be an "individual sequence," or even be chosen adaptively and adversarially. As such, our setting may be termed "hybrid i.i.d.-adversarial." Our results also hold in the so-called transductive setting, where the side information is presented ahead of time.[1]

We have in mind machine learning applications such as online ad or product placement, whereby the contextual information $x_1, \dots, x_n$ of website visitors may be viewed as an i.i.d. sequence, *yet the decisions made by these customers might be too complex to be described in a probabilistic form.*

A common way to encode the prior knowledge about the problem is to take a class $\mathcal{F}$ of functions (or, *deterministic policies*) $\mathcal{X} \to [d]$, with the hope that one of the functions will incur small cost on the presented contexts. With this "inductive bias," we then aim to make predictions as to minimize regret

$$\mathbf{Reg} = \sum_{t=1}^n q_t^\intercal c_t - \inf_{f \in \mathcal{F}} \sum_{t=1}^n f(x_t)^\intercal c_t, \tag{1}$$

where henceforth we abuse the notation by identifying the value $f(x) \in [d]$ with the standard basis vector $\boldsymbol{e}_{f(x)}$. This regret formulation encodes the prior knowledge of the practitioner. If the modeling choice $\mathcal{F}$ is good and (1) is small, the algorithm is guaranteed to incur small loss $\sum_{t=1}^n q_t^\intercal c_t$. Modeling the set of solutions $\mathcal{F}$ to the problem is a more direct approach (in the spirit of statistical learning) as compared to the harder problem of positing distributional assumptions on the relationship between contexts and the rewards. (The latter approach typically suffers from the curse of dimensionality.)

The difficulty of the contextual bandit problem arises from the form of the feedback. The customer seeking to buy a product different from what is presented by the recommendation engine may leave the site without revealing her valuation for all the items. Similarly, in personalized care, we may only observe the effect of the drug choice selected for the given patient. It is well recognized that exploration—or randomization—is required in these problems. Yet, in the contextual bandit setting the exploration-exploitation tradeoff is not simple, as the quality of the arms changes with the context in a way that is only indirectly captured by the benchmark term.

[1] In Section 6 we also discuss the fully-adversarial case (see (Auer et al., 2002; McMahan & Streeter, 2009) for the famous EXP4 algorithm for finite $\mathcal{F}$).

Online multiclass classification with one bit (correct-or-not) feedback can be seen as an example of our setting. In that case c_t is a standard basis vector e_{y_t} for some class $y_t \in [d]$, and the feedback is $c_t(\widehat{y}_t) = \mathbf{I}\{\widehat{y}_t \neq y_t\}$. Unlike (Kakade et al., 2008), we posit that side information is i.i.d.—an assumption that will play a key role in developing computationally efficient methods, even for the indicator (rather than the easier hinge) loss.

The hybrid i.i.d.-adversarial scenario has been studied in both the full information and contextual bandit settings in (Lazaric & Munos, 2009). Their algorithm, as well as the algorithm of (Beygelzimer et al., 2011), maintain distributions over the set of functions and, hence, computation can be linear in the size of $\mathcal{F}$.

For the case when $\mathcal{F}$ is finite, the upper bound for BISTRO provided in Theorem 2 is $O(n^{3/4}(\log|\mathcal{F}|)^{1/4})$. The work of (Agarwal et al., 2014) gives a better $O(n^{1/2}(\log|\mathcal{F}|)^{1/2})$ rate for the case when rewards are i.i.d. On the other hand, our results hold for

- arbitrary $\mathcal{F}$ and arbitrary reward sequences,
- approximate ERM values and a way to address the computational problem associated to ERM.

For finite $\mathcal{F}$, the rate of $O(n^{3/4}\sqrt{d\log|\mathcal{F}|})$ is obtained in the concurrent work (Syrgkanis et al., 2016) in the online transductive setting. The authors also extend their result to the semi-bandit setting.

We remark that if contexts are arbitrary as well, our setting subsumes the problem of multiclass prediction with bandit feedback and indicator loss, as described above. Even for the multiclass hinge loss, it is still unclear (at least to the authors) whether the rate $O(n^{2/3})$ for the linear classifier considered in (Kakade et al., 2008) can be improved.[2] It is, therefore, an open question whether the $O(n^{3/4})$ rates achieved by our method for the hybrid scenario for arbitrary classes $\mathcal{F}$ can be improved.

There are several new techniques that make it possible to develop computationally feasible prediction methods with nontrivial regret guarantees:

- First is the idea of *relaxations*, presented in (Rakhlin et al., 2012) for the *full-information* setting. An extension to partial information case has been a big roadblock for developing new bandit methods. We present this extension here.
- Second is the idea of a random playout, also employed in (Cesa-Bianchi & Shamir, 2013; Rakhlin & Sridha-

[2] The $O(n^{1/2})$ rate in (Hazan & Kale, 2011) is only proved for the case of log-loss.

ran, 2015). We show that by having access to unlabeled contexts, the computational (and statistical) difficulty of integrating with respect to the unknown distribution simply disappears.

- We extend the notion of classical Rademacher averages to the case of vector-valued functions. The symmetrization technique in this case is of independent interest.

- In many cases, the offline ERM optimization problem (which we assume away as an "oracle call") may be NP hard. Building on the technique of (Rakhlin & Sridharan, 2015), we employ optimization-based relaxations for integer programs. We prove that the regret bound of the resulting algorithm only worsens by a multiplicative factor that is related to the ratio of average widths of the relaxed and the original sets.

It is worth emphasizing again that the family of prediction methods presented in this work is drived from the partial-information extension of the relaxation framework, and the resulting algorithms are distinct from the ones appearing in the literature. We believe that this approach is systematic and can partially fill the gap in our understanding of the algorithmic possibilities for contextual bandits.

2. Notation

Denote $[d] \triangleq \{1,\dots,d\}$ and $a_{1:t} \triangleq \{a_1,\dots,a_t\}$. Let Δ_d be the probability simplex over d coordinates. The vector of ones is denoted by $\mathbf{1}$ and an indicator of event A by $\mathbf{I}\{A\}$. For a matrix M, we use M_t to refer to its t-th column.

3. Setup

Let us recall the online protocol. On each round $t \in [n]$, we observe side information $x_t \in \mathcal{X}$, predict $\widehat{y}_t \sim q_t \in \Delta_d$, and observe feedback $c_t(\widehat{y}_t)$ for some $c_t \in [0,1]^d$.

Given $x_{1:n}$, it is convenient to work with a matrix representation of the class $\mathcal{F}$ projected on these data. Each $f \in \mathcal{F}$ yields sequence $(f(x_1),\dots,f(x_n))$, which we collect as a $d \times n$ matrix M^f, defined as

$$M^f(j,t) = \mathbf{I}\{f(x_t) = j\}\,. \tag{2}$$

Let $\widehat{\mathcal{M}} = \widehat{\mathcal{M}}[x_{1:n}] = \{M^f : f \in \mathcal{F}\}$ denote the collection of matrices. (The hat on $\widehat{\mathcal{M}}$ will remind us of the dependence of this set on $x_{1:n}$, even if not explicitly mentioned).

We may now define the oracle employed by the prediction method:

Definition 1. Given a class $\mathcal{F}$ of policies $\mathcal{X} \to [d]$, a set of covariates $x_{1:n}$, and a real-valued $d \times n$ matrix Y, a *value-of-ERM* oracle returns the value

$$\inf_{M \in \widehat{\mathcal{M}}[x_{1:n}]} \sum_{t=1}^{n} M_t^{\intercal} Y_t\,. \tag{3}$$

The oracle is called δ-*approximate* if the reported value is within an additive δ from the minimum.

We may express the comparator term in (1) as an ERM objective (3) with $Y = [c_1,\dots,c_n]$. Closely related to this expression is a new (to the best of our knowledge) definition of Rademacher averages for vector-valued functions: given $x_{1:n}$, define

$$\mathfrak{R}(\mathcal{F}; x_{1:n}) \triangleq \mathfrak{R}(\widehat{\mathcal{M}}) \triangleq \mathbb{E}_{\boldsymbol{\epsilon}_{1:n}} \sup_{M \in \widehat{\mathcal{M}}} \sum_{t=1}^{n} M_t^{\intercal} \boldsymbol{\epsilon}_t \tag{4}$$

where $\boldsymbol{\epsilon}_1,\dots,\boldsymbol{\epsilon}_n$ are d-dimensional vectors with independent Rademacher random variables. We observe that Rademacher complexity is nothing but a (negative of) the expected ERM objective with the random matrix $[-\boldsymbol{\epsilon}_1,\dots,-\boldsymbol{\epsilon}_n]$. Indeed, as in the classical case, correlation of the vector-valued function class $\mathcal{F}$ with noise measures its complexity.

4. Relaxations for Partial Information

Let us write the information obtained on round t as a tuple

$$I_t(x_t, q_t, \widehat{y}_t, c_t) = (x_t, q_t, \widehat{y}_t, c_t(\widehat{y}_t)),$$

keeping in mind that x_t is revealed before q_t is chosen. In full information problems, I_t contains the vector c_t, but not so in our bandit case. For partial information problems, it turns out to be crucial to include q_t in the definition of I_t, in addition to the value $c_t(\widehat{y}_t)$.

A *partial-information relaxation* $\mathbf{Rel}()$ is a function that maps $(I_1,\dots,I_t)$ to a real value, for any $t \in [n]$. We say that the partial-information relaxation $\mathbf{Rel}(I_1,\dots,I_t)$ is *admissible* if for any $t \in [n]$, for all $I_1,\dots,I_{t-1}$,

$$\begin{aligned}\mathop{\mathbb{E}}_{x_t} \inf_{q_t} \max_{c_t} \mathop{\mathbb{E}}_{\widehat{y}_t \sim q_t} \{c_t(\widehat{y}_t) + \mathbf{Rel}(I_{1:t-1}, I_t(x_t, q_t, \widehat{y}_t, c_t))\} \\ \le \mathbf{Rel}(I_{1:t-1})\end{aligned} \tag{5}$$

and for all $x_{1:n}$,$c_{1:n}$, and $q_{1:n}$,

$$\mathop{\mathbb{E}}_{\widehat{y}_{1:n} \sim q_{1:n}} \mathbf{Rel}(I_{1:n}) \ge -\inf_{f \in \mathcal{F}} \sum_{t=1}^{n} f(x_t)^{\intercal} c_t\,. \tag{6}$$

In the above expressions, x_t follows the (unknown) distribution P_x, q_t ranges over distributions on $[d]$, and c_t over $[0,1]^d$.

Any randomized strategy $(q_t)_{t=1}^n$ that certifies the inequalities (5) and (6) is called *an admissible strategy*.

Lemma 1. *Let* $\mathbf{Rel}()$ *be an admissible relaxation and* $(q_t)_{t=1}^n$ *an admissible strategy. Then for any* $c_{1:n}$,

$$\mathbb{E}[\mathbf{Reg}] \leq \mathbf{Rel}(\varnothing).$$

The above partial-information relaxation setup appears to be "the right" analogue of the full-information relaxation framework. While we do not present it here, one may recover the EXP4 algorithm through the above approach, with the correct regret bound.

We will now present an admissible strategy for the contextual bandit problem, assuming we can sample from the distribution P_x, or have access to unlabeled data.

5. The BISTRO Algorithm

For any $t \in [n]$, define a $d \times n$ matrix $Y^{(t)}$ as

$$Y^{(t)} = [c_1, \ldots, c_{t-1}, c_t, 2\boldsymbol{\epsilon}_{t+1}, \ldots, 2\boldsymbol{\epsilon}_n]$$

with $\boldsymbol{\epsilon}_s \in \{\pm 1\}^d$ a vector of independent Rademacher random variables. At each step $t \in [n]$, the randomized method presented below calculates a distribution $q_t \in \Delta_d$ with each coordinate at least $\gamma > 0$ (a parameter of the algorithm) and defines an unbiased estimate $\tilde{c}_t$ of c_t in a usual manner as

$$\tilde{c}_t(j) = \mathbf{I}\{\widehat{y}_t = j\} \times c_t(\widehat{y}_t)/q_t(j).$$

It is standard to verify that $\mathbb{E}_{\widehat{y}_t \sim q_t} \tilde{c}_t = c_t$. We then define

$$\tilde{Y}^{(t)} = [\tilde{c}_1, \ldots, \tilde{c}_{t-1}, \tilde{c}_t, 2\gamma^{-1}\boldsymbol{\epsilon}_{t+1}, \ldots, 2\gamma^{-1}\boldsymbol{\epsilon}_n], \quad (7)$$

and recall that $\tilde{Y}_s^{(t)}$ denotes the s-th column of this matrix.

The next theorem is the main result of the paper.

Theorem 2. *The partial-information relaxation*

$$\mathbf{Rel}(I_{1:t}) = \mathop{\mathbb{E}}_{(x,\boldsymbol{\epsilon})_{t+1:n}} \sup_{M \in \widehat{\mathcal{M}}} \left\{ -\sum_{s=1}^{n} M_s^\top \tilde{Y}_s^{(t)} \right\} + (n-t)d\gamma \quad (8)$$

is admissible. An admissible randomized strategy for this relaxation is given by BISTRO (Algorithm 1). The expected regret of the algorithm with $\gamma = \sqrt{2\mathbb{E}\mathfrak{R}(\mathcal{F}; x_{1:n})/(nd)}$ *is upper bounded by*

$$2\sqrt{2d \cdot n \cdot \mathbb{E}\mathfrak{R}(\mathcal{F}; x_{1:n})}.$$

In particular, a growth rate of $\mathbb{E}\mathfrak{R}(\mathcal{F}; x_{1:n}) = O(\sqrt{n})$ yields an overall $O(n^{3/4})$ regret bound. Techniques for upper bounding classical Rademacher averages of particular function classes are well-established (Bartlett & Mendelson, 2003), and these can be extended to control the Rademacher complexity of vector-valued functions (4). A straightforward application of Hoeffding's inequality implies an $O(\sqrt{nd \log |\mathcal{F}|})$ bound on Rademacher complexity of a finite class.

The proof of the Theorem appears in the Supplementary Material. Let us give a high level intuition for the result. An admissible relaxation can be thought of as a potential function (of the observed data) that interpolates between the comparator and the regret bound in a manner specified by Eq. (5). In the full information case, it has been shown that a Rademacher complexity-based potential function is a near-optimal one (Rakhlin et al., 2012), and any upper bound on this Rademacher-based potential is a good candidate for an admissible relaxation. In the partial information case, we need to obtain a relaxation that only depends on observations. Since we can produce unbiased estimates, Jensen's inequality suggests that the relaxation (8) is a good candidate for being an admissible relaxation. Theorem 2 proves that this is indeed the case.

We now state the algorithm.

Algorithm 1 BISTRO: **B**and**I**t**S** wi**T**h **R**elaxati**O**ns

input Parameter $\gamma \in (0, 1/d)$

for $t = 1, \ldots, n$ **do**

Observe x_t. Draw $x_{t+1:n} \sim P_x$ and $\boldsymbol{\epsilon}_{t+1:n}$.

Construct $\tilde{Y}^{(t)}$ and define q_t^* to be a minimizer of

$$\max_{j \in [d]} \left\{ q^\top \boldsymbol{e}_j - \min_{M \in \widehat{\mathcal{M}}[x_{1:n}]} \left\{ \sum_{s \neq t} \gamma M_s^\top \tilde{Y}_s^{(t)} + M_t^\top \boldsymbol{e}_j \right\} \right\}$$

over $q \in \Delta_d$ and set

$$q_t = (1 - \gamma d) q_t^* + \gamma \mathbf{1}. \quad (9)$$

Predict $\widehat{y}_t \sim q_t$ and observe $c_t(\widehat{y}_t)$.

Create an estimate $\tilde{c}_t$:

$$\tilde{c}_t(j) = \mathbf{I}\{\widehat{y}_t = j\} \times c_t(\widehat{y}_t)/q_t(j).$$

end for

The draw $x_{t+1:n} \sim P_x$ can be realized by drawing from a pool of unlabeled data.

The random signs comprising the matrix $\tilde{Y}$ provide a form of "regularization". We remark that in experiments, one may obtain better performance by replacing the factor 2 in (7) with a smaller value, or even with zero. A theoretical justification for this (which is related to using a surrogate loss) is beyond the scope of this paper.

Lemma 3. *The calculation of* q_t^* *in BISTRO*[3] *can be done by a water-filling argument and requires* d *calls to the ERM*

[3] 'Bistro' means 'fast' in Russian.

oracle.

Proof of Lemma 3. The optimization problem in Algorithm 1 is of the form $\min_{q \in \Delta_d} \max_{j \in [d]} \{q_j - \psi_j\}$ where ψ_j is the value of the infimum over $\widehat{\mathcal{M}}$ corresponding to $\boldsymbol{e}_j$, and it is solved by a water-filling argument which we describe next. Each value ψ_j is a value-of-ERM oracle call. Let $\psi_{(1)} \geq \ldots \geq \psi_{(d)}$ be a sorted order of these values, and let $q_{(1)} = \ldots = q_{(d)} = 0$ be the initial values of the corresponding coordinates of the solution q. Start with a unit amount and assign $q_{(1)} = \psi(1) - \psi(2)$. Then add $\psi_{(2)} - \psi_{(3)}$ to both $q_{(1)}$ and $q_{(2)}$, and proceed until either the unit mass is exhausted, or the smallest coordinate (d) in the ordering is reached and filled. In the former case, q is the solution, and the latter case requires us to uniformly fill all the coordinates of q until they sum to one. It is easy to see that this procedure minimizes the maximum difference. □

The algorithm only requires the *value* of the ERM objective, not the solution. Furthermore, this value can be δ-approximate, and the additional error is $O(n\delta)$ over the n rounds. This provides extra flexibility, since approximate ERM values may be obtained via optimization methods. To see that the errors do not propagate through the relaxation, we point to Eq. (22) in the proof and observe that the inequality holds whenever we use any unbiased estimate for the cost, even the one based on an approximate solution.

Perhaps the most unusual aspect of the algorithm is the use of unlabeled data. It is an example of a general random playout idea. In the setting of online linear optimization, the Follow-the-Perturbed-Leader method is an example of such a random playout, yet the idea extends well beyond this scenario. As shown in (Rakhlin et al., 2012), the random playout technique can be applied whenever a certain worst-case-choice can be replaced with a known bad-enough distribution. However, when side information x_t is i.i.d., the step is not even required. Furthermore, an inspection of the proof shows that we may deal with x's coming from a non-i.i.d. stochastic process, as long as we are able to draw future samples from it.

We also remark that (9) may be applied only to the coordinates that are close to zero, if any. The potential suboptimality of the $O(n^{3/4})$ bound stems from the uniform exploration. It is an open question whether this can be improved systematically for all classes $\mathcal{F}$, or whether there is a different structural property that allows one to avoid this form of exploration.

6. Extensions

In this section, we outline several extensions of BISTRO. Specifically, we show how to incorporate additional data-based constraints, and how to use further optimization-based relaxations (such as LP or SDP), to obtain polynomial time methods for the ERM (or regularized ERM) solution. We show that one obtains a regret bound that only worsens by a factor related to the integrality gap of the integer program relaxation. With an eye on both computation and prediction performance, these techniques expand the applicability of BISTRO.

6.1. Data-dependent policy classes

An inspection of the proof reveals that all the steps go through if we define regret in (1) with respect to a data-dependent class $\mathcal{F}[x_{1:n}]$:

$$\sum_{t=1}^{n} q_t^\top c_t - \inf_{f \in \mathcal{F}[x_{1:n}]} \sum_{t=1}^{n} f(x_t)^\top c_t. \tag{10}$$

In this case, given $x_{1:n}$, to each $f \in \mathcal{F}[x_{1:n}]$ we associate M^f as defined in (2), and take

$$\widehat{\mathcal{M}} = \{M^f : f \in \mathcal{F}[x_{1:n}]\}.$$

The BISTRO algorithm is then identical, while the regret upper bound of Theorem 2 now replaces $\mathbb{E}\mathfrak{R}(\mathcal{F}; x_{1:n})$ with $\mathbb{E}\mathfrak{R}(\mathcal{F}[x_{1:n}]; x_{1:n})$.

The ability to change the set of policies according to the actual data allows an extra degree of flexibility. This flexibility can be realized via additional global constraints in terms of $x_{1:n}$, as we show in the next few sections. We also discuss a concrete example.

6.2. Data-based constraints

A particular way to define a data-dependent subset of $\mathcal{F}$ is via constraints. Suppose we let $C(f; x_{1:n})$ be the degree to which $f \in \mathcal{F}$ violates constraints with respect to the given data $x_{1:n}$. We then define

$$\mathcal{F}_K[x_{1:n}] = \{f \in \mathcal{F} : C(f; x_{1:n}) \leq K\}, \tag{11}$$

a pruning of the original class that keeps only those policies that do not violate the constraints by more than K. Let us give an example.

Example: Product Recommendation Suppose at each time step we are asked to recommend one of d products to a person, based on her covariate information x_t. Let $\mathcal{F}$ be a set of policies that map x_t to the particular choice of the product (e.g. the label achieving maximum projection of x_t onto d vectors w_j; here $\mathcal{F}$ may consist of all such unit vector d-tuples). The payoff is whether the person decided to buy the recommended product. However, suppose x_t also encodes the location (physical, or within a network), and we believe it is a good idea to focus recommendations such that near-by people are targeted with the same product. The marketing motivation here is two-fold: first, the recommendations would reinforce each other when individuals communicate, or if one of them buys the product; second, in a social network near-by individuals (friends) tend to have similar tastes, and thus a good policy would suggest similar items.

The objective of enforcing similarity of recommendations is a global constraint that can only be checked once we know all the $x_1, \ldots, x_n$. We can easily incorporate the constraint into the definition of $\mathcal{F}_K[x_{1:n}]$ as follows. Let $w(x_s, x_r)$ be the cost of providing different recommendations to x_s and x_r (which is smaller if the two individuals are "far"). In the case of a network, we may set, for instance, $w(x_s, x_r) = 0$ if the sth person is more than a hop away from the rth person. Define

$$C(f; x_{1:n}) = \sum_{s,r \in [n]} w(x_s, x_r) \mathbf{I}\{f(x_s) \neq f(x_r)\}, \quad (12)$$

the constraint violation by f in assigning products to the given set of individuals. Let $\mathcal{F}_K[x_{1:n}]$ be defined as in (11). Note that the constraint is not on the behavior of the recommendation engine, but on the set of policies that we hope will do well for the problem. If there is indeed the effect of reinforcement of recommendations or similarity of tastes within the local neighborhood, the restriction to a smaller set $\mathcal{F}_K[x_{1:n}]$ is justified.

Within the same setting of product recommendation, we might instead take a set of policies ensuring that within each neighborhood at least k individuals receive each particular product recommendation. This constraint, which roughly corresponds to "coverage" of the relevant population, can be written as

$$C(f; x_{1:n}) = \sum_{\ell} \sum_{j \in [d]} \left[k - \sum_{s \in T_\ell} f(x_s)[j] \right]_+$$

where $\{T_\ell\}_\ell$ is a partition of $[n]$ into neighborhoods according to information contained in $x_{1:n}$. The above two examples give a flavor of the constraints that can be encoded — the framework is flexible enough to fit a wealth of scenarios.

From the computational point of view, it might be difficult to obtain the ERM value over a constrained set $\mathcal{F}_K[x_{1:n}]$. Instead, we consider an additional form of relaxation, where the constraint is subtracted off as a Lagrangian term. We will then employ certain linear programming relaxations to solve the product recommendation problem. Notably, by going to a regularized version of relaxations we are not changing the regret definition, which is still with respect to the constrained set.

6.3. Regularized relaxation

Let $\mathcal{F}_K[x_{1:n}] = \{f \in \mathcal{F} : C(f; x_{1:n}) \leq K\}$ be the constrained set for some value K and a constraint function C, as in the previous section. Let us write $C(M; x_{1:n})$ for the matrix representation of the corresponding $f \in \mathcal{F}$. The following form of a relaxation may be better suited for approximation algorithms than the one where the constraint is strictly enforced.

Lemma 4. *For any $\lambda, K > 0$, the partial-information relaxation*

$$\mathop{\mathbb{E}}_{(x,\epsilon)_{t+1:n}} \sup_{M \in \widehat{\mathcal{M}}} \left\{ -\sum_{s=1}^{n} M_s^\mathsf{T} \tilde{Y}_s^{(t)} - \lambda C(M; x_{1:n}) \right\} + \lambda K + (n-t) d\gamma \quad (13)$$

is admissible, where $\widehat{\mathcal{M}}$ denotes the matrix representation of the original (unconstrained) *set $\mathcal{F}$ of policies.*

Proof of Lemma 4. We check that the initial condition is satisfied. For this purpose, let $\widehat{\mathcal{M}}_K$ be the set of matrices corresponding to the constrained set $\mathcal{F}_K[x_{1:n}]$. Similarly to (18) in the proof of Theorem 2,

$$\begin{aligned} -\inf_{f \in \mathcal{F}_K[x_{1:n}]} \sum_{t=1}^{n} f(x_t)^\mathsf{T} c_t &\leq \mathbb{E} \sup_{M \in \widehat{\mathcal{M}}_K} \sum_{t=1}^{n} -M_t^\mathsf{T} \tilde{Y}_t^{(n)} \\ &\leq \mathbb{E} \sup_{M \in \widehat{\mathcal{M}}} \left\{ \sum_{t=1}^{n} -M_t^\mathsf{T} \tilde{Y}_t^{(n)} - \lambda C(M; x_{1:n}) \right\} + \lambda K. \end{aligned}$$

The second inequality holds since all the matrices in the former supremum have the constraint value bounded by K. The recursive condition argument follows exactly as in the proof of Theorem 2. □

The only change required for BISTRO is to define the optimization objective in terms of *regularized* ERM values

$$\min_{M \in \widehat{\mathcal{M}}} \left\{ \sum_{s \neq t} \gamma M_s^\mathsf{T} \tilde{Y}_s^{(t)} + M_t^\mathsf{T} \boldsymbol{e}_j + \gamma^{-1} \lambda C(M; x_{1:n}) \right\} \quad (14)$$

over the *unconstrained* set of matrices corresponding to $\mathcal{F}$. While the required minimization problem is over an unconstrained set of policies, we can control the expected regret

$$\sum_{t=1}^{n} q_t^\mathsf{T} c_t - \inf_{f \in \mathcal{F}_K[x_{1:n}]} \sum_{t=1}^{n} f(x_t)^\mathsf{T} c_t \quad (15)$$

of the modified BISTRO with respect to the *constrained* set $\mathcal{F}_K[x_{1:n}]$, which is the original goal. The regret is given by $\mathbf{Rel}(\varnothing)$, which is at most

$$\mathbb{E} \sup_{M \in \widehat{\mathcal{M}}} \left\{ -\gamma^{-1} \sum_{t=1}^{n} M_t^\mathsf{T} \boldsymbol{\epsilon}_t - \lambda C(M; x_{1:n}) \right\} + nd\gamma + \lambda K.$$

It is possible to optimally balance λ with respect to K and the Rademacher averages in a data-driven manner, but we omit this step for brevity.

As we illustrate in the next section, optimization problems of the form (14) may admit a linear programming (or other) relaxation, offering an alternative to the optimization problem over the constrained set.

6.4. Optimization-based relaxations

To make the algorithm of this paper more applicable, we discuss here the situation where the ERM oracle or the regularized ERM oracle for the class $\mathcal{F}_K[x_{1:n}]$ (or the unconstrained set $\mathcal{F}$) is a difficult or even an NP-hard integer program. The idea is to choose a superset $\widetilde{\mathcal{M}} \supseteq \widehat{\mathcal{M}}$ for which the linear optimization problem is easier.

Lemma 5. *Let $\widetilde{\mathcal{M}} \supseteq \widehat{\mathcal{M}}$ be a set of matrices such that the column sum $\sum_{j=1}^{d} M_t(j) \le 1$ for any $M \in \widetilde{\mathcal{M}}$ and $t \in [n]$. Then the partial information relaxation*

$$\mathbf{Rel}\left(I_{1:t}\right) = \underset{(x,\boldsymbol{\epsilon})_{t+1:n}}{\mathbb{E}} \sup_{M \in \widetilde{\mathcal{M}}} \left\{ -\sum_{s=1}^{n} M_s^{\intercal} \tilde{Y}_s^{(t)} \right\} + (n-t)d\gamma$$

is admissible. BISTRO (with ERM over $\widetilde{\mathcal{M}}$ rather than $\mathcal{M}$) is an admissible strategy for this relaxation and the expected regret is upper bounded by

$$2\sqrt{2d \cdot n \cdot \mathbb{E}\mathfrak{R}(\widetilde{\mathcal{M}})}.$$

Similarly, using $\widetilde{\mathcal{M}}$ in (13) yields an admissible relaxation, and BISTRO with the corresponding regularized ERM is an admissible strategy.

The set $\widetilde{\mathcal{M}}[x_{1:n}]$ may be defined via linear programming or SDP relaxations for integer programs, or via Lasserre/Parrilo hierarchies (Lasserre, 2001; Parrilo, 2003). There is a large body of literature that aims at understanding the integrality gap in relaxing the integer program. These results are directly applicable to the present problem.

As a concrete example, consider the product recommendation example in the previous section, and consider the cost (12) for each policy and the restriction $\mathcal{F}_K[x_{1:n}]$ in (11). We assume here that $\mathcal{F}$ is the set of all possible labelings, since in general the optimization problem will depend on the structure of $\mathcal{F}$ and its description. Let us phrase the regularized ERM integer program (14) as a *Metric Labeling Constraint* (Kleinberg & Tardos, 2002) problem. The general form of this integer program is given for $z \in [d]^n$ by

$$g(z) = \sum_{v \in V} d_1(v, z_v) + \sum_{(u,v) \in E} W_{(u,v)} d_2(z_u, z_v) \qquad (16)$$

where $G = (V, E, W)$ is a graph with nonnegative weights, $|V| = n$, the value $d_1 : V \times [d] \to \mathbb{R}$ is a cost of assigning a label to a node, and the separation cost $d_2 : [d] \times [d] \to \mathbb{R}_{\ge 0}$ on the edges is a metric on the space of labels. The Metric Labeling Constraint problem asks for a solution that minimizes $g(z)$ over $[d]^n$.

For our application to product recommendation we convert the regularized minimization objective of (14) with the constraint (12) into the above form (16) by matching the assignment costs to the linear part and the separation costs to the constraint part (12). More precisely, let G be a fully connected graph with weights $W_{(s,r)} = \gamma^{-1}\lambda \cdot w(x_s, x_r)$ between nodes corresponding to x_s and x_r. The indices of vertices correspond to time steps in $[n]$, and z_v corresponds to the coordinate chosen by the particular M at time v. We take $d_1(v, z_v)$ to be the value $\gamma \boldsymbol{e}_{z_v}^{\intercal} \tilde{Y}_v^{(t)}$ if $v \ne t$ and $\boldsymbol{e}_{z_v}^{\intercal} \boldsymbol{e}_j$ if $v = t$. Define $d_2(a, b) = \mathbf{I}\{a \ne b\}$ to be the uniform metric. We may also define a metric on the space of products, assigning smaller distance to similar items.

(Kleinberg & Tardos, 2002) give an LP relaxation for the Metric Labeling Constraint problem. The set that defines the relaxation is precisely the set $\widetilde{\mathcal{M}}$ we seek. Furthermore, the authors prove a 2-approximation ratio for the uniform metric, which is the case here. (Chekuri et al. (2004) prove an integrality gap of $O(\log k)$ for the general case).

Given the 2-approximation ratio result, we conclude that the regret bound for BISTRO with the LP program as the relaxation of the regularized ERM is only a constant worse than the bound with the constrained set $\mathcal{F}_K[x_{1:n}]$. The exact optimization over the latter set may be computationally intractable, while we provide an efficient method to achieve a bound, optimal to within a constant. As already noted in (Rakhlin & Sridharan, 2015), such an approach that fuses approximation algorithms and online relaxations is able to produce polynomial-time methods with regret defined as $1\times$ the benchmark, while the benchmark itself may be NP-hard. This phenomenon can be attributed to the improper nature of the predictions, which need not be consistent with any particular policy in $\mathcal{F}$.

More generally, by obtaining a multiplicative approximation of gap for the integer program, one may derive

$$\mathbb{E}\mathfrak{R}(\widetilde{\mathcal{M}}[x_{1:n}]) \le O(\mathrm{gap}) \times \mathbb{E}\mathfrak{R}(\mathcal{M}[x_{1:n}]). \qquad (17)$$

Then one obtains a method with better computational properties and a regret bound which is only $O(\sqrt{\mathrm{gap}})$ worse. Once again, the factor in front of the comparator in the definition (1) of regret is still one when using $\widetilde{\mathcal{M}}$ as a relaxation.

Finally, we remark that (17) is comparing an *average* width of $\widetilde{\mathcal{M}}$ (largest projection onto noise) with an average width of $\mathcal{M}$. Such a comparison of average widths (and, therefore, "average gap") for useful sets of contextual bandit policies $\mathcal{F}$ appears to be an interesting area of further investigation. We refer to (Rakhlin & Sridharan, 2015), where some of these ideas have been developed in the context of cut-based constraints for node prediction on graphs.

6.5. Adversarial contexts

Suppose we place no assumption on the evolution of x_t's, which may now be treated as worst-case. This problem subsumes the full information online classification setting, and, hence, one cannot hope to have nontrivial regret against policy classes $\mathcal{F}$ with infinite Littlestone dimension. More generally, the best one can hope for is to say that the adversarial contextual bandit problem can be solved whenever the corresponding full information problem may be solved. We now present essentially this result: if there is a full-information relaxation, then one may use it to solve the adversarial contextual bandit problem. Moreover, based on the work of (Rakhlin et al., 2012; Foster et al., 2015), all the known online learning methods appear to be relaxation based. Hence, we essentially prove below that

> If a problem is online learnable in the full-information adversarial setting, then it is learnable in the adversarial contextual bandit setting. Furthermore, if the former is computationally tractable, then so is the latter.

To be precise, the full information version of contextual problem is as follows. On round t, we observe $x_t \in \mathcal{X}$, predict $\widehat{y}_t \in [d]$, and observe $c_t \in [0,1]^d$. The regret is defined as before, with our cumulative cost being $\sum c_t(\widehat{y}_t)$.

A full information relaxation $\mathbf{Rel}^\dagger(c_1, \dots, c_t)$ is admissible if

$$\sup_{x_t} \inf_{q_t} \max_{c_t} \mathbb{E}_{\widehat{y}_t \sim q_t} \left\{ c_t(\widehat{y}_t) + \mathbf{Rel}^\dagger(c_{1:t}) \right\} \leq \mathbf{Rel}^\dagger(c_{1:t-1})$$

and

$$\mathbf{Rel}^\dagger(c_{1:n}) \geq -\inf_{f \in \mathcal{F}} \sum_{t=1}^{n} f(x_t)^\intercal c_t \,.$$

Similarly, a partial information relaxation is admissible in this adversarial case when $c_{1:t}$ are replaced with $I_{1:t}$ in the above admissibility definition, as in Section 4.

Lemma 6. *If* $\mathbf{Rel}^\dagger()$ *is an admissible full-information relaxation for the adversarial scenario, then*

$$\mathbf{Rel}(I_{1:t}) \triangleq \gamma^{-1}\mathbf{Rel}^\dagger(\gamma\tilde{c}_1, \dots, \gamma\tilde{c}_t) + (n-t)d\gamma$$

is admissible for the partial information scenario. Prediction q_t *is obtained as* $q_t = (1-d\gamma)q_t^* + \gamma\mathbf{1}$ *where* q_t^* *is computed by solving for a full-information strategy with the scaled unbiased estimates of costs. The resulting regret upper bound is*

$$2\sqrt{d \cdot n \cdot \mathbf{Rel}^\dagger(\varnothing)}.$$

Proof of Lemma 6. Let us first check the initial condition

$$\mathbb{E}_{\widehat{y}_{1:n} \sim q_{1:n}} \mathbf{Rel}(I_{1:n}) = \mathbb{E}_{\widehat{y}_{1:n} \sim q_{1:n}} \gamma^{-1}\mathbf{Rel}^\dagger(\gamma\tilde{c}_1, \dots, \gamma\tilde{c}_n)$$
$$\geq \mathbb{E}_{\widehat{y}_{1:n} \sim q_{1:n}} -\inf_{f \in \mathcal{F}} \sum_{t=1}^{n} f(x_t)^\intercal \tilde{c}_t \geq -\inf_{f \in \mathcal{F}} \sum_{t=1}^{n} f(x_t)^\intercal c_t$$

where the first inequality is due to admissibility of the full-information relaxation, and the second is due to Jensen's inequality and unbiasedness of $\tilde{c}_t$. For the recursive part, we follow the proof of Theorem 2 and note that all the statements, until the end, are done conditionally on x_t. Define the strategy q_t^* as

$$q_t^* = \operatorname*{argmin}_{q \in \Delta_d} \sup_{\tilde{c} \in \gamma^{-1}[0,1]^d} \left\{ q^\intercal(\gamma\tilde{c}_t) + \mathbf{Rel}^\dagger(\gamma\tilde{c}_1, \dots, \gamma\tilde{c}_t) \right\}$$

and let $q_t = (1-d\gamma)q_t^* + \gamma\mathbf{1}$. Given x_t, (23) tells us

$$\max_{c_t \in [0,1]^d} \mathbb{E}_{\widehat{y}_t \sim q_t} \left\{ c_t(\widehat{y}_t) + \mathbf{Rel}(I_1, \dots, I_t) \right\}$$
$$\leq \sup_{\tilde{c}_t \in \gamma^{-1}[0,1]^d} \left\{ (q_t^*)^\intercal \tilde{c}_t + \mathbf{Rel}(I_1, \dots, I_t) \right\} + d\gamma$$

which is equal to

$$\gamma^{-1} \sup_{\tilde{c}_t} \left\{ (q_t^*)^\intercal(\gamma\tilde{c}_t) + \mathbf{Rel}^\dagger(\gamma\tilde{c}_1, \dots, \gamma\tilde{c}_t) \right\} + (n-t+1)d\gamma$$
$$\leq \gamma^{-1}\mathbf{Rel}^\dagger(\gamma\tilde{c}_1, \dots, \gamma\tilde{c}_{t-1}) + (n-t+1)d\gamma$$

by admissibility of the full-information relaxation. Observe that the use of the full-information relaxation on $\gamma\tilde{c}_t$'s is warranted since these vectors are in $[0,1]^d$. □

We remark that time complexity of adversarial contextual bandit solution in Lemma 6 is the same as the time complexity of the corresponding full information procedure.

7. Open Problems and Future Directions

The main open problem is whether the regret upper bound for BISTRO or a related method can be improved. In the inequality (23) we decouple the distribution q_t' from q_t, and this appears to be the source of looseness, at least in the analysis. A more precise analysis at this step might resolve the issue. It is unclear what kind of structure of $\mathcal{F}$ may be used to improve computation and/or regret of BISTRO.

Under structural assumptions on $\mathcal{F}$ one may come up with sufficient statistics for the information $I_{1:t}$ and, therefore, avoid keeping around all the estimates $\tilde{c}_t$. Of course, this is the case in non-contextual bandits, where the sum $\sum \tilde{c}_t$ is sufficient (at least as evidenced by existing near-optimal bandit methods).

An interesting avenue of investigation is to study the more general case when x's are drawn from a stochastic process with a parametrized form. One may then attempt to estimate the parameters of the process on-the-go and use the estimate to hallucinate future data for random playout.

Acknowledgements We gratefully acknowledge the support of NSF under grants AF 1116928, CDS&E-MSS 1521529 and 1521544.

References

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. E. Taming the monster: A fast and simple algorithm for contextual bandits. *arXiv preprint arXiv:1402.0555*, 2014.

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

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. E. Contextual bandit algorithms with supervised learning guarantees. In *AISTATS*, pp. 19–26, 2011.

Cesa-Bianchi, N. and Shamir, O. Efficient transductive online learning via randomized rounding. In *Empirical Inference*, pp. 177–194. Springer, 2013.

Chekuri, C., Khanna, S., Naor, J., and Zosin, L. A linear programming formulation and approximation algorithms for the metric labeling problem. *SIAM Journal on Discrete Mathematics*, 18(3):608–625, 2004.

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*, 2011.

Foster, D., Rakhlin, A., and Sridharan, K. Adaptive online learning. In *NIPS*, 2015.

Hazan, E. and Kale, S. Newtron: an efficient bandit algorithm for online multiclass prediction. In *Advances in Neural Information Processing Systems*, pp. 891–899, 2011.

Kakade, S. M., Shalev-Shwartz, S., and Tewari, A. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th international conference on Machine learning*, pp. 440–447. ACM, 2008.

Kleinberg, J. and Tardos, E. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and markov random fields. *Journal of the ACM (JACM)*, 49(5):616–639, 2002.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Langford, J. and Zhang, T. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pp. 817–824, 2008.

Lasserre, J. B. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

Lazaric, A. and Munos, R. Hybrid stochastic-adversarial on-line learning. In *Conference on Learning Theory*, 2009.

McMahan, H. B and Streeter, M. J. Tighter bounds for multi-armed bandits with expert advice. In *COLT*, 2009.

Parrilo, P. A. Semidefinite programming relaxations for semialgebraic problems. *Mathematical programming*, 96(2):293–320, 2003.

Rakhlin, A. and Sridharan, K. Hierarchies of relaxations for online prediction problems with evolving constraints. In *COLT*, 2015.

Rakhlin, A., Shamir, O., and Sridharan, K. Relax and randomize: From value to algorithms. In *Advances in Neural Information Processing Systems 25*, pp. 2150–2158, 2012.

Syrgkanis, V., Krishnamurthy, A., and Schapire, R. E. Efficient algorithms for adversarial contextual learning. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

A. Proof of Lemma 1

In the proof, we use the shorthand $\langle\!\langle \dots \rangle\!\rangle_{t=1}^{n}$ to denote repeated application of the operators within the brackets from $t = 1$ to n. As an example, the sequence of operators $\mathbb{E}_{x_1} \max_{c_1} \mathbb{E}_{x_2} \max_{c_2}[G(x_1, c_1, x_2, c_2)]$ acting on the function G is abbreviated as $\langle\!\langle \mathbb{E}_{x_t} \max_{c_t} \rangle\!\rangle_{t=1}^{2} [G(x_1, c_1, x_2, c_2)]$.

Let $q_1, \dots, q_n$ be an admissible strategy. The expected regret of this strategy can be upper bounded by

$$\mathbb{E}[\mathbf{Reg}] \le \sup_{c_{1:n}} \mathbb{E}[\mathbf{Reg}] \le \left\langle\!\!\left\langle \mathbb{E}_{x_t} \sup_{c_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} q_t^\intercal c_t - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} f(x_t)^\intercal c_t \right]$$

by Jensen's inequality (pulling $\mathbb{E}_{x_t}$ out of multiple suprema until its t-th position). The last expression is further upper bounded by

$$\left\langle\!\!\left\langle \mathbb{E}_{x_t} \sup_{c_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} q_t^\intercal c_t + \mathbb{E}_{\widehat{y}_{1:n} \sim q_{1:n}} \mathbf{Rel}\left(I_{1:n}\right) \right]$$

by admissibility of the partial information relaxation. By linearity of expectation for $\mathbb{E}_{\widehat{y}_t}$ and Jensen's inequality (to pull it out through multiple suprema as before), we obtain an upper bound of

$$\left\langle\!\!\left\langle \mathbb{E}_{x_t} \sup_{c_t} \mathbb{E}_{\widehat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} c_t(\widehat{y}_t) + \mathbf{Rel}\left(I_{1:n}\right) \right].$$

We now start from step n and observe that $\sum_{t=1}^{n-1} c_t(\widehat{y}_t)$ does not depend on $x_n, c_n, \widehat{y}_n$, and thus we rewrite the preceding expression as

$$\left\langle\!\!\left\langle \mathbb{E}_{x_t} \sup_{c_t} \mathbb{E}_{\widehat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n-1} \left[\sum_{t=1}^{n-1} c_t(\widehat{y}_t) + \mathbb{E}_{x_n} \sup_{c_n} \mathbb{E}_{\widehat{y}_n \sim q_n} \left\{ c_n(\widehat{y}_n) + \mathbf{Rel}\left(I_{1:n}\right) \right\} \right].$$

By admissibility of q_t and (5), we pass to the upper bound of

$$\left\langle\!\!\left\langle \mathbb{E}_{x_t} \sup_{c_t} \mathbb{E}_{\widehat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n-1} \left[\sum_{t=1}^{n-1} c_t(\widehat{y}_t) + \mathbf{Rel}\left(I_{1:n-1}\right) \right].$$

Continuing in this fashion leads to a bound of $\mathbf{Rel}\left(\varnothing\right)$.

B. Proof of Theorem 2

Admissibility: initial condition For any $c_{1:n}, q_{1:n}, x_{1:n}$, it holds that

$$- \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} f(x_t)^\intercal c_t = \sup_{M \in \mathcal{M}[x_{1:n}]} - \sum_{t=1}^{n} M_t^\intercal Y_t^{(n)} \le \mathbb{E}_{\widehat{y}_{1:n} \sim q_{1:n}} \sup_{M \in \mathcal{M}[x_{1:n}]} - \sum_{s=1}^{n} M_s^\intercal \tilde{Y}_s^{(n)} = \mathbb{E}_{\widehat{y}_{1:n} \sim q_{1:n}} \mathbf{Rel}\left(I_{1:n}\right). \tag{18}$$

In the remainder of the proof we will often write $\mathcal{M}$ instead of $\mathcal{M}[x_{1:n}]$ for brevity.

Admissibility: recursion Let $\mathcal{D} \triangleq \{\gamma^{-1} \boldsymbol{e}_j : j \in [d]\} \cup \{\mathbf{0}\}$, the set of scaled standard basis vectors, together with the origin. Observe that $\tilde{c}_t \in \mathrm{conv}(\mathcal{D})$ by our definition of unbiased estimates (in fact, it is only a scaling of one coordinate).

We now reason conditionally on x_t. As before, let $\boldsymbol{\epsilon}_s \in \{\pm 1\}^d$ denote a vector of independent Rademacher random variables. Let us abbreviate by $\boldsymbol{\rho} = (\boldsymbol{\epsilon}_{t+1:n}, x_{t+1:n})$, a draw of independent Rademacher variables and covariates from P_x for the "future rounds", as part of the random playout procedure. Together with the estimates $\tilde{c}_s$ for $s < t$, we may now construct $\tilde{Y}^{(t)}$ and M matrices and define the randomized prediction algorithm as

$$q_t^*(\boldsymbol{\rho}) = \operatorname*{argmin}_{q \in \Delta_d} \sup_{\tilde{c} \in \mathcal{D}} \left\{ q^\intercal \tilde{c} + \sup_{M \in \mathcal{M}[x_{1:n}]} - \sum_{s \ne t} M_s^\intercal \tilde{Y}_s^{(t)} - M_t^\intercal \tilde{c} \right\} \tag{19}$$

$$= \operatorname*{argmin}_{q \in \Delta_d} \sup_{\widehat{y}_t, q_t'} \max_{c_t} \left\{ q^\intercal \tilde{c}_t(c_t, q_t', \widehat{y}_t) + \sup_{M \in \mathcal{M}[x_{1:n}]} - \sum_{s \ne t} M_s^\intercal \tilde{Y}_s^{(t)} - M_t^\intercal \tilde{c}_t(c_t, q_t', \widehat{y}_t) \right\}. \tag{20}$$

We remark that x_t enters the above definition of $q_t^*(\boldsymbol{\rho})$, but we leave this dependence implicit until the end of the proof. For the purposes of the proof also define

$$q_t(\boldsymbol{\rho}) = (1-d\gamma)\cdot q_t^*(\boldsymbol{\rho}) + \gamma\mathbf{1}, \tag{21}$$

a version of $q_t^*(\boldsymbol{\rho})$ that is shifted away from the boundary of the simplex (a step that allows for estimation of c_t). Also define $q_t = \mathbb{E}_{\boldsymbol{\rho}}[q_t(\boldsymbol{\rho})]$ and $q^* = \mathbb{E}_{\boldsymbol{\rho}}[q_t^*(\boldsymbol{\rho})]$. Observe that

$$\mathbb{E}_{\widehat{y}_t\sim q_t}[c_t(\widehat{y}_t)] = q_t^\intercal c_t \le (q_t^*)^\intercal c_t + \gamma\mathbf{1}^\intercal c_t \le \mathbb{E}_{\widehat{y}_t\sim q_t}[(q_t^*)^\intercal \tilde{c}_t(c_t,q_t,\widehat{y}_t)] + d\gamma. \tag{22}$$

Hence,

$$\begin{aligned}
&\max_{c_t\in[0,1]^d} \mathbb{E}_{\widehat{y}_t\sim q_t}\left\{c_t(\widehat{y}_t) + \mathbf{Rel}\left(I_1,\ldots,I_t\right)\right\}\\
&\le \max_{c_t\in[0,1]^d} \mathbb{E}_{\widehat{y}_t\sim q_t}\left\{(q_t^*)^\intercal\tilde{c}_t(c_t,q_t,\widehat{y}_t) + \mathbf{Rel}\left(I_{1:t-1}, I_t(x_t,q_t,\widehat{y}_t,c_t)\right)\right\} + d\gamma\\
&\le \sup_{\widehat{y}_t\in[d],q_t'} \max_{c_t\in[0,1]^d}\left\{(q_t^*)^\intercal\tilde{c}_t(c_t,q_t',\widehat{y}_t) + \mathbf{Rel}\left(I_{1:t-1}, I_t(x_t,q_t',\widehat{y}_t,c_t)\right)\right\} + d\gamma.
\end{aligned} \tag{23}$$

In the last expression, the supremum is over q_t' of the form $(1-d\gamma)\cdot q + \gamma\mathbf{1}$, $q\in\Delta_d$. This last upper bound holds because q_t is one of such distributions. The importance of this upper bound is that it decouples the q_t^* from q_t' in the first term, a step that yields a simple optimization problem that defines $q_t^*(\boldsymbol{\rho})$. Writing out the form of the relaxation, the last expression is equal to

$$\begin{aligned}
&\sup_{\widehat{y}_t,q_t'}\max_{c_t}\left\{(q_t^*)^\intercal\tilde{c}_t(c_t,q_t',\widehat{y}_t) + \mathbb{E}_{\boldsymbol{\rho}}\sup_{M\in\mathcal{M}} -\sum_{s\ne t} M_s^\intercal \tilde{Y}_s^{(t)} - M_t^\intercal\tilde{c}_t(c_t,q_t',\widehat{y}_t)\right\} + (n-t+1)d\gamma\\
&\qquad\le \sup_{\tilde{c}_t\in\mathrm{conv}(\mathcal{D})}\left\{(q_t^*)^\intercal\tilde{c}_t + \mathbb{E}_{\boldsymbol{\rho}}\sup_{M\in\mathcal{M}} -\sum_{s\ne t} M_s^\intercal \tilde{Y}_s^{(t)} - M_t^\intercal\tilde{c}_t\right\} + (n-t+1)d\gamma
\end{aligned}$$

since $\tilde{c}_t(c_t,q_t',\widehat{y}_t)\in\mathrm{conv}(\mathcal{D})$. The expression inside the supremum is a convex function of $\tilde{c}_t$, and thus the supremum is achieved at a vertex, an element of $\mathcal{D}$. Since $q_t^* = \mathbb{E}_{\boldsymbol{\rho}}[q_t^*(\boldsymbol{\rho})]$, we upper bound the last expression via Jensen's inequality (omitting $(n-t+1)d\gamma$ to simplify the exposition) by

$$\mathbb{E}_{\boldsymbol{\rho}}\sup_{\tilde{c}_t\in\mathcal{D}}\left\{q_t^*(\boldsymbol{\rho})^\intercal\tilde{c}_t + \sup_{M\in\mathcal{M}} -\sum_{s\ne t} M_s^\intercal \tilde{Y}_s^{(t)} - M_t^\intercal\tilde{c}_t\right\}. \tag{24}$$

Since $q_t^*(\boldsymbol{\rho})$ is precisely defined to be the minimizer (given $\boldsymbol{\rho}$) of the supremum in (24), the preceding expression is equal to

$$\mathbb{E}_{\boldsymbol{\rho}}\inf_{q\in\Delta_d}\sup_{\tilde{c}_t\in\mathcal{D}}\left\{q^\intercal\tilde{c}_t + \sup_{M\in\mathcal{M}} -\sum_{s\ne t} M_s^\intercal \tilde{Y}_s^{(t)} - M_t^\intercal\tilde{c}_t\right\}$$

The rest of the upper bounds will be derived conditionally on $\boldsymbol{\rho}$. Observe that

$$\inf_{q\in\Delta_d}\sup_{\tilde{c}_t\in\mathcal{D}}\left\{q^\intercal\tilde{c}_t + \sup_{M\in\mathcal{M}} -\sum_{s\ne t} M_s^\intercal \tilde{Y}_s^{(t)} - M_t^\intercal\tilde{c}_t\right\} = \sup_{p_t}\inf_{q}\mathbb{E}_{\tilde{c}_t\sim p_t}\left\{q^\intercal\tilde{c}_t + \sup_{M\in\mathcal{M}} -\sum_{s\ne t} M_s^\intercal \tilde{Y}_s^{(t)} - M_t^\intercal\tilde{c}_t\right\}$$

by the minimax theorem, where p_t ranges over the set of distributions on $\mathcal{D}$. By linearity of expectation, the preceding expression is equal to

$$\begin{aligned}
&\sup_{p_t}\inf_{q}\left\{q^\intercal\mathbb{E}_{\tilde{c}_t\sim p_t}[\tilde{c}_t] + \mathbb{E}_{\tilde{c}_t\sim p_t}\sup_{M\in\mathcal{M}}\left[-\sum_{s\ne t} M_s^\intercal \tilde{Y}_s^{(t)} - M_t^\intercal\tilde{c}_t\right]\right\}\\
&= \sup_{p_t}\left\{\min_{j\in[d]}\boldsymbol{e}_j^\intercal\mathbb{E}_{\tilde{c}_t\sim p_t}[\tilde{c}_t] + \mathbb{E}_{\tilde{c}_t\sim p_t}\sup_{M\in\mathcal{M}}\left[-\sum_{s\ne t} M_s^\intercal \tilde{Y}_s^{(t)} - M_t^\intercal\tilde{c}_t\right]\right\}.
\end{aligned} \tag{25}$$

Observe that for any $M\in\mathcal{M}$, $\sum_{j=1}^d M_{j,t} = 1$ and the elements of M_t are nonnegative. Thus

$$\min_j \boldsymbol{e}_j^\intercal\mathbb{E}_{\tilde{c}_t\sim p_t}[\tilde{c}_t] \le M_t^\intercal\mathbb{E}_{\tilde{c}_t\sim p_t}[\tilde{c}_t].$$

Therefore, (25) is equal to

$$\begin{aligned}
&\sup_{p_t} \mathbb{E}_{\tilde{c}_t \sim p_t} \sup_{M \in \mathcal{M}} \left\{ -\sum_{s \neq t} M_s^\intercal \tilde{Y}_s^{(t)} + \min_{j \in [d]} \boldsymbol{e}_j^\intercal \mathbb{E}_{\tilde{c}_t \sim p_t}[\tilde{c}_t] - M_t^\intercal \tilde{c}_t \right\} \\
&\leq \sup_{p_t} \mathbb{E}_{\tilde{c}_t \sim p_t} \sup_{M \in \mathcal{M}} \left\{ -\sum_{s \neq t} M_s^\intercal \tilde{Y}_s^{(t)} + M_t^\intercal \mathbb{E}_{\tilde{c}_t \sim p_t}[\tilde{c}_t] - M_t^\intercal \tilde{c}_t \right\} \\
&= \sup_{p_t} \mathbb{E}_{\tilde{c}_t, \tilde{c}_t' \sim p_t} \sup_{M \in \mathcal{M}} \left\{ -\sum_{s \neq t} M_s^\intercal \tilde{Y}_s^{(t)} + M_t^\intercal (\tilde{c}_t' - \tilde{c}_t) \right\}.
\end{aligned}$$

Since exchanging $\tilde{c}_t$ and $\tilde{c}_t'$ switches the sign in the last term, we may introduce an independent Rademacher random variable δ_t via the standard technique of symmetrization. The last expression is then equal to

$$\begin{aligned}
&\sup_{p_t} \mathbb{E}_{\tilde{c}_t, \tilde{c}_t' \sim p_t} \mathbb{E}_{\delta_t} \sup_{M \in \mathcal{M}} \left\{ -\sum_{s \neq t} M_s^\intercal \tilde{Y}_s^{(t)} + \delta_t M_t^\intercal (\tilde{c}_t' - \tilde{c}_t) \right\} \\
&\leq \sup_{p_t} \mathbb{E}_{\tilde{c}_t \sim p_t} \mathbb{E}_{\delta_t} \sup_{M \in \mathcal{M}} \left\{ -\sum_{s \neq t} M_s^\intercal \tilde{Y}_s^{(t)} + 2\delta_t M_t^\intercal \tilde{c}_t \right\}.
\end{aligned}$$

The above inequality follows by splitting the supremum into two parts equal parts. Let us now reason conditionally on $\tilde{c}_t$. There are two cases: either $\tilde{c}_t = \mathbf{0}$ or $\tilde{c}_t = \gamma^{-1} \boldsymbol{e}_j$ for some coordinate $j \in [d]$. Let us consider the second case, and the first follows from the same reasoning. Take Z to be a random vector with independent coordinates and values in $\{-\gamma^{-1}, \gamma^{-1}\}^d$. For the jth coordinate, Z_j is identically γ^{-1}, while for all other coordinates $i \neq j$ the distribution Z_i is symmetric. Clearly, $\mathbb{E}Z = \tilde{c}_t$. By Jensen's inequality,

$$\mathbb{E}_{\delta_t} \sup_{M \in \mathcal{M}} \left\{ -\sum_{s \neq t} M_s^\intercal \tilde{Y}_s^{(t)} + 2\delta_t M_t^\intercal \tilde{c}_t \right\} \leq \mathbb{E}_{\delta_t} \mathbb{E}_Z \sup_{M \in \mathcal{M}} \left\{ -\sum_{s \neq t} M_s^\intercal \tilde{Y}_s^{(t)} + 2\delta_t M_t^\intercal Z \right\}.$$

It is not hard to see that the distribution of $\delta_t Z$ is uniform on $\{-\gamma^{-1}, \gamma^{-1}\}^d$, and we can write it as $\gamma^{-1} \boldsymbol{\epsilon}_t$, a scaled vector of independent Rademacher random variables. The overall bound (together with the omitted term $(n - t + 1)d\gamma$) is then

$$\begin{aligned}
\max_{c_t \in [0,1]^d} \mathbb{E}_{\widehat{y}_t \sim q_t} \left\{ c_t(\widehat{y}_t) + \mathbf{Rel}\left(I_1, \ldots, I_t\right) \right\} &\leq \mathbb{E}_{\boldsymbol{\rho}} \sup_{p_t} \left\{ \mathbb{E}_{\tilde{c}_t \sim p_t} \mathbb{E}_{\boldsymbol{\epsilon}_t} \sup_{M \in \mathcal{M}} -\sum_{s \neq t} M_s^\intercal \tilde{Y}_s^{(t)} + 2\gamma^{-1} M_t^\intercal \boldsymbol{\epsilon}_t \right\} + (n - t + 1)d\gamma \\
&= \mathbb{E}_{\boldsymbol{\rho}} \mathbb{E}_{\boldsymbol{\epsilon}_t} \sup_{M \in \mathcal{M}} \left\{ -\sum_{s \neq t} M_s^\intercal \tilde{Y}_s^{(t)} + 2\gamma^{-1} M_t^\intercal \boldsymbol{\epsilon}_t \right\} + (n - t + 1)d\gamma
\end{aligned}$$

since the expression no longer depends on p_t and $\tilde{c}_t$. The above inequality holds for any x_t. Hence, we may take expectation on both sides, yielding

$$\begin{aligned}
\mathbb{E}_{x_t} \max_{c_t \in [0,1]^d} \mathbb{E}_{\widehat{y}_t \sim q_t} \left\{ c_t(\widehat{y}_t) + \mathbf{Rel}\left(I_1, \ldots, I_t\right) \right\} &\leq \mathbb{E}_{\boldsymbol{\epsilon}_{t:n}, x_{t:n}} \sup_{M \in \mathcal{M}[x_{1:n}]} \left\{ -\sum_{s \neq t} M_s^\intercal \tilde{Y}_s^{(t)} + 2\gamma^{-1} M_t^\intercal \boldsymbol{\epsilon}_t \right\} + (n - t + 1)d\gamma \\
&= \mathbf{Rel}\left(I_{1:t-1}\right)
\end{aligned}$$

because $\boldsymbol{\rho} = (\boldsymbol{\epsilon}_{t+1:n}, x_{t+1:n})$. This proves admissibility.

Omitting 0 from objective Examining the algorithm in (19), we note that the optimization problem may be taken over $\tilde{c} \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$; that is, the argmin over q does not change upon the removal of $\mathbf{0}$. To see this, suppose that $q_t^*(\boldsymbol{\rho})$ is the optimal response when $\tilde{c} \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$. Then it is also an optimal response to $\tilde{c} \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\} \cup \{\mathbf{0}\}$ since for $\tilde{c} = \mathbf{0}$ the value of q does not make any difference in terms of the value. This proves our claim, and is reflected in the definition of Algorithm 1.

Regret bound The final bound is given by

$$\mathbf{Rel}\left(\varnothing\right) = \mathbb{E}_x \mathbb{E}_\epsilon \sup_{M \in \mathcal{M}[x_{1:n}]} -\sum_{t=1}^n M_t^\intercal \tilde{Y}_t^{(0)} + nd\gamma = \frac{2}{\gamma} \mathbb{E}\mathfrak{R}(\mathcal{F}; x_{1:n}) + nd\gamma = 2\sqrt{2dn\mathbb{E}\mathfrak{R}(\mathcal{F}; x_{1:n})}.$$