Online-to-Confidence-Set Conversions and Application to Sparse Stochastic Bandits

Yasin Abbasi-Yadkori
Dept. of Computing Science
University of Alberta
abbasiya@cs.ualberta.ca

Dávid Pál
Google, Inc.
New York, NY
dpal@google.com

Csaba Szepesvári
Dept. of Computing Science
University of Alberta
szepesva@cs.ualberta.ca

Abstract

We introduce a novel technique, which we call *online-to-confidence-set conversion*. The technique allows us to construct high-probability confidence sets for linear prediction with correlated inputs given the predictions of *any* algorithm (e.g., online LASSO, exponentiated gradient algorithm, online least-squares, p-norm algorithm) targeting online learning with linear predictors and the quadratic loss. By construction, the size of the confidence set is directly governed by the regret of the online learning algorithm. Constructing tight confidence sets is interesting on its own, but the new technique is given extra weight by the fact having access tight confidence sets underlies a number of important problems. The advantage of our construction here is that progress in constructing better algorithms for online prediction problems directly translates into tighter confidence sets. In this paper, this is demonstrated in the case of linear stochastic bandits. In particular, we introduce the *sparse* variant of linear stochastic bandits and show that a recent online algorithm together with our online-to-confidence-set conversion allows one to derive algorithms that can exploit if the reward is a function of a sparse linear combination of the components of the chosen action.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

1 Introduction

A large portion of machine learning is devoted to constructing point estimates of some unknown quantity given some "noisy data". A main issue with point estimates is that they lack a description of the remaining uncertainty about the unknown quantity. Confidence sets, on the other hand, allow one to characterize the remaining uncertainty. Wasserman's maxim "never give an estimator without giving a confidence set" (Wasserman, 1998, p. vii.) clearly illustrates the importance of confidence sets. However useful confidence sets are on their own, in a number of sequential tasks they are in fact indispensable. Examples include stopping problems (Mnih et al., 2008), bandit problems (Auer et al., 2002a, Auer, 2002, Dani et al., 2008), variants of the pick-the-winner problem (Even-Dar et al., 2002, Mannor and Tsitsiklis, 2004, Mnih et al., 2008), reinforcement learning (Bartlett and Tewari, 2009, Jaksch et al., 2010), or active learning (Even-Dar et al., 2002).

In this paper we investigate the problem of constructing confidence sets for the vector of coefficients of a linear function observed at a finite number of points under martingale noise (the exact conditions will be stated in the next section). We take a very general approach to the construction of the confidence sets. In particular, we propose a new technique which we call *online-to-confidence-set conversion*. The basic idea is that the predictions of any online algorithm that predicts the responses of the chosen inputs in a sequential manner can be "converted" to a confidence set. The only assumption is that the online prediction algorithm comes with an upper bound on its regret[1] with respect to the best linear predictor using the quadratic prediction loss. The details of this conversion are explained in Section 2.

One strength of our method is that it allows one to

[1] This notion of regret, to be defined in the next section, is different from the regret of the bandit problem!

use any linear prediction algorithm as the underlying online algorithm, such as (online) least squares (regularized or constrained) (Lai et al., 1979, Auer et al., 2002b, Vovk, 2001), online LASSO, the exponentiated gradient (EG) algorithm[2] (Kivinen and Warmuth, 1997), the p-norm algorithm (Grove et al., 1997, Gentile and Littlestone, 1999), the SeqSEW algorithm (Gerchinovitz, 2011), etc. These algorithm differ in terms of what solutions they are biased to. For example, some of these algorithms are biased towards sparse solutions, some of them are biased towards sparse inputs, etc. However, *all* the algorithms just mentioned satisfy the assumptions of the conversion, i.e., they work with quadratic prediction loss and for most of these algorithms a regret bound is known. Thanks to generality of our solution we can obtain a confidence set for each of these algorithms and, in fact, for any algorithm that might be developed in the future, too. These allows one to derive confidence sets which are similarly "'biased" towards different qualities.

Conversions and reductions between machine learning tasks were studied by Langford and colleagues; see Langford (2011). Our online-to-confidence-set conversion can be compared with the online-to-batch conversions (Littlestone, 1989, Cesa-Bianchi et al., 2004, Dekel and Singer, 2006). However, there are two big differences between these two. First, online-to-batch conversions convert the predictions of a low-regret online algorithm into a single prediction with a low risk, whereas in our online-to-confidence-set conversion we combine the predictions to construct a confidence set. Second, in online-to-batch conversions one assumes that the data (i.e., covariate-response pairs) are generated in an i.i.d.[3] fashion (in fact, the risk is defined with respect to the underlying joint distribution), while in online-to-confidence-set conversion the covariates can be chosen adversarially and only responses are stochastic. In summary, we are not aware of previous results on reductions of the type we consider.

A second major contribution of this paper is the introduction and study of a variant of stochastic linear bandit problems (Dani et al., 2008), which we call *sparse stochastic linear bandits*. Sparsity, in recent years, became the line of attack for statistical problems which were previously thought unsolvable. The assumption that the underlying statistical model is sparse greatly decreases the sample size required to learn the model provided, of course, that the model is indeed sparse. Several examples of algorithms that take advantage of sparsity are the Winnow algorithm (Littlestone, 1988), the LASSO (Tibshirani, 1996) and algorithms for compressed sensing (Candès, 2006).

With sparsity in mind, we investigate the sparse variant of the linear stochastic bandit problem. The "dense" versions of this problem has been studied previously by Auer (2002), Dani et al. (2008), Abbasi-Yadkori et al. (2009), Rusmevichientong and Tsitsiklis (2010), Li et al. (2010). Recall that the linear stochastic bandit problem is a sequential decision problem, where in each round the learner chooses an action and he receives a reward, which is an unknown linear function of the action, corrupted with random zero-mean noise. The goal of the learner is to maximize his reward accumulated over the course of multiple rounds. Precise description of the model is given in Subsection 4.1. In this paper we focus on the situation when the underlying linear function is potentially sparse, i.e., many of its coefficients are zero, as can be expected to be the case in applications when the feature space is high-dimensional but only a few features are relevant (e.g., in web advertisement applications).

Sparse linear bandit problem can be viewed as sequential decision making version of the feature selection problem. Its potential applications include medical trials, web advertising, content optimization, and reinforcement learning. To the best of our knowledge this problem is new and we are not aware of any prior work.

In order to design an algorithm in this learning model we use the standard *optimism-in-the-face-uncertainty principle*. This principle has been used to design algorithms for the linear stochastic bandit problem and related problems (Auer, 2002, Dani et al., 2008, Li et al., 2010, Walsh et al., 2009). The basic idea is that the algorithm maintains a confidence set for the vector of the coefficients of the linear function, and in each round it chooses an action and a vector from the confidence set that maximize the predicted reward, i.e., the choice of the algorithm is optimistic.

2 Online-to-Confidence-Set Conversion

The confidence set construction problem is as follows: Consider a stochastic sequence $\{(X_t, Y_t)\}_{t=1}^{\infty}$, where $X_t \in \mathbb{R}^d$ are the d-dimensional inputs, $Y_t = \langle \theta_*, X_t \rangle + \eta_t$ are the real-valued responses, $\theta_* \in \mathbb{R}^d$ is an unknown parameter vector and η_t is "random noise" satisfying $\mathbf{E}[\, \eta_t \,|\, X_{1:t}, \eta_{1:t-1}] = 0$ and some tail-constraints, to be specified soon (here, $X_{1:t}$ denotes the sequence $X_1, X_2, \ldots, X_t$ and, similarly, $\eta_{1:t-1}$ denotes the sequence $\eta_1, \eta_2, \ldots, \eta_{t-1}$). The problem is to construct a confidence set $C_n \subseteq \mathbb{R}^d$ for the unknown

[2] EG is a variant of Winnow for linear prediction.

[3] The abbreviation i.i.d. stands for "independent and identically distributed".

parameter vector θ_* given data $(X_1, Y_1, \ldots, X_n, Y_n)$. In particular, given n, one is interested in constructing a (measurable) map C_n from $(\mathbb{R}^d \times \mathbb{R})^n \times (0, \delta_0)$ to the measurable subsets of $\mathbb{R}^d$ such that for any $0 < \delta < \delta_0$ and data $\{(X_t, Y_t)\}_{t=1}^n$ that satisfies the above conditions, $C_n = C_n(X_1, Y_1, \ldots, X_n, Y_n, \delta)$ satisfies $\Pr[\theta_* \in C_n] \geq 1 - \delta$. As will be explained, we shall in fact construct confidence sets that hold simultaneously: $\Pr[\theta_* \in \cap_{n \geq 1} C_n] \geq 1 - \delta$, which is much stronger and correspondingly more useful from the point of view of applications.

The tail-constraint on the noise sequence that was mentioned is as follows: We assume that $\{\eta_t\}_{t=1}^\infty$ is conditionally R-sub-Gaussian for some $R \geq 0$ in the sense that for any $t \geq 1$,

$$\forall \lambda \in \mathbb{R} \quad \mathbf{E}[e^{\lambda \eta_t} \mid X_{1:t}, \eta_{1:t-1}] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right). \quad (1)$$

The conditional sub-Gaussianity of η_t automatically implies that $\mathbf{E}[\eta_t \,|\, X_{1:t}, \eta_{1:t-1}] = 0$. Furthermore, it also implies that $\mathbf{Var}[\eta_t \,|\, X_{1:t}, \eta_{1:t-1}] \leq R^2$ and thus we can think of R^2 as (a bound on) the variance of the noise. An example of R-sub-Gaussian random variable η_t is a zero-mean Gaussian random variable with variance at most R^2, or a bounded zero-mean random variable lying in an interval of length at most $2R$.

In this section we show how to convert regret bounds of an online algorithm for online linear prediction to confidence sets for θ_*. In online linear prediction we assume that in round t an online algorithm receives $x_t \in \mathbb{R}^d$, predicts $\widehat{y}_t \in \mathbb{R}$, receives $y_t \in \mathbb{R}$ and suffers a loss $\ell_t(\widehat{y}_t)$ where $\ell_t(y) = (y - y_t)^2$ is the *quadratic prediction loss*. In online linear prediction, one makes no assumptions on the sequence $\{(x_t, y_t)\}_{t=1}^\infty$, perhaps except for bounds on the norm of x_t and magnitude of y_t. In fact, the sequence $\{(x_t, y_t)\}_{t=1}^\infty$ can be chosen in an adversarial fashion.

The task of the online algorithm is to keep its n-step cumulative loss $\sum_{t=1}^n \ell_t(\widehat{y}_t)$ as low as possible. We compare the loss of the algorithm with the loss of the strategy that uses a fixed weight vector $\theta \in \mathbb{R}^d$ and in round t predicts $\langle \theta, x_t \rangle$ – this is why the problem is called linear prediction. The difference of the losses is called the *regret with respect to* θ and formally we write it as

$$\rho_n(\theta) = \sum_{t=1}^n \ell_t(\widehat{y}_t) - \sum_{t=1}^n \ell_t(\langle \theta, x_t \rangle).$$

The construction of algorithms with "small" regret $\rho_n(\theta)$ is an important topic in the online learning literature. Examples of algorithms designed to achieve this include variants of the least squares method (projected or regularized), the exponentiated gradient algorithm, the p-norm regularized algorithm, online LASSO, SeqSEW, etc.

Suppose now that we feed an online algorithm for linear prediction with a stochastic sequence $\{(X_t, Y_t)\}_{t=1}^\infty$ generated according to the model described above. Let the sequence of predictions produced by the algorithm $\{\widehat{y}_t\}_{t=1}^\infty$. The following theorem states that from the sequence $\{\widehat{y}_t\}_{t=1}^\infty$ of predictions we can construct high-probability confidence sets C_n for θ_*. Moreover, as we will see the volume of the set C_n will be related to the regret of the algorithm; the smaller the regret of the algorithm, the smaller the volume of C_n is. The theorem below states the precise result.

Theorem 1 (Online-to-Confidence-Set Conversion)**.** *Assume that $\{F_t\}_{t=0}^\infty$ is a filtration and for any $t \geq 1$, X_t is an $\mathbb{R}^d$-valued, F_t-measurable random variable and η_t is a real-valued, F_{t+1}-measurable random variable that is conditionally R-sub-Gaussian in the sense that*

$$\forall \lambda \in \mathbb{R} \qquad \mathbf{E}[\, e^{\lambda \eta_t} \,|\, F_t] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right). \quad (2)$$

Define $Y_t = \langle \theta_*, X_t \rangle + \eta_t$, where $\theta^* \in \mathbb{R}^d$ is the true parameter. Suppose that we feed $\{(X_t, Y_t)\}_{t=1}^\infty$ into an online prediction algorithm which, for all $t \geq 0$, admits a regret bound

$$\rho_t(\theta_*) \leq B_t, \qquad \textit{(almost surely)}$$

where $\{B_t\}_{t=0}^\infty$ is some sequence of $\{F_t\}_{t=0}^\infty$-adapted non-negative random variables. Then, for any $\delta \in (0, 1/4]$, with probability at least $1 - \delta$, the true parameter θ_* lies in the intersection of the sets

$$C_n = \left\{ \theta \in \mathbb{R}^d \;:\; \sum_{t=1}^n (\widehat{y}_t - \langle \theta, X_t \rangle)^2 \leq 1 + 2B_n + 32R^2 \ln\left(\frac{R\sqrt{8} + \sqrt{1 + B_n}}{\delta}\right) \right\},$$

where $n \geq 0$.

The proof of the theorem can be found in Section 3.

Notice that, as expected, the confidence sets C_n in the theorem can be constructed from observable quantities: the data $X_1, X_2, \ldots, X_n$, $Y_1, Y_2, \ldots, Y_n$, the predictions $\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_n$ of the linear prediction algorithm, the regret bound B_n, the "variance" of the noise R^2 and the confidence parameter δ. Finally, it is not hard to see that since C_n is a sub-level set of a non-negative quadratic function in θ, it is an ellipsoid, possibly, with some of the axes infinitely long.

An important feature of the confidence sets constructed in Theorem 1 is that they are based on regret

bounds B_n which can themselves be *data-dependent bounds* on the regret. Although we will not exploit this in the later sections of the paper, in practice, the use of such data dependent bounds (which exists for a large number of the algorithms mentioned) is highly recommended.

Another important feature of the bound is that the unknown parameter vector belongs to the intersection of all the the confidence sets constructed, i.e., the confidence sets hold the true parameter vector *uniformly in time*. This property is useful both because it leads to simpler algorithm designs and also to simpler analysis. Note that usually this property is achieved by taking a union bound, where the failure probability δ at time step n would be divided by a diverging function of n in the definition of the confidence set. With our techniques, we were able to avoid this union bound, which is expected to give better results in practice. In particular, if the online algorithm is "lucky" in that its regret B_n does not grow, or grows very slowly, our confidence set shrink faster than if a union bound was used to ensure uniformity in time.

It turns out that the fact that confidence sets constructed in Theorem 1 can be unbounded, might potentially lead to trouble. To deal with this issue, we slightly modify the confidence sets: If we know a priori that $\|\theta_*\|_2 \le E$ we can add $\|\theta\|_2^2 \le E^2$ to the inequality defining C_n in the theorem. This leads to the following obvious corollary.

Corollary 2 (Regularized Confidence Sets)**.** *Assume the same as in Theorem 1 and additionally assume that $\|\theta_*\|_2 \le E$. Then, for any $\delta \in (0, 1/4]$, with probability at least $1-\delta$, the true parameter θ_* lies in the intersections of the sets*

$$C_n = \left\{ \theta \in \mathbb{R}^d \;:\; \|\theta\|_2^2 + \sum_{t=1}^{n} (\widehat{y}_t - \langle \theta, X_t \rangle)^2 \le E^2 + 1 + 2B_n + 32R^2 \ln\left(\frac{R\sqrt{8} + \sqrt{1+B_n}}{\delta} \right) \right\},$$

where $n \ge 0$.

Of course, it would be better to take intersection of the confidence sets from Theorem 1 and the set $\{\theta \;:\; \|\theta\|_2 \le E\}$ instead, since the resulting confidence set would be smaller than the confidence set constructed in the corollary. However, the resulting confidence set would no longer be an ellipsoid and this might complicate matters later. The confidence set constructed in the corollary is always a bounded non-degenerate ellipsoid and this allows a relatively simple analysis.

3 Proof of Theorem 1

To prove Theorem 1, we will need Corollary 8 from Appendix A and Propositions 9 and 10 from Appendix B.

Proof of Theorem 1. With probability one,

$$\begin{aligned} B_n &\ge \rho_n(\theta_*) \\ &= \sum_{t=1}^{n} \ell_t(\widehat{Y}_t) - \ell_t(\langle \theta_*, X_t \rangle) \\ &= \sum_{t=1}^{n} (\widehat{Y}_t - Y_t)^2 - (\langle \theta_*, X_t \rangle - Y_t)^2 \\ &= \sum_{t=1}^{n} (\widehat{Y}_t - \langle \theta_*, X_t \rangle - \eta_t)^2 - \eta_t^2 \\ &= \sum_{t=1}^{n} (\widehat{Y}_t - \langle \theta_*, X_t \rangle)^2 - 2\eta_t (\widehat{Y}_t - \langle \theta_*, X_t \rangle) \,. \end{aligned}$$

Thus, with probability one,

$$\sum_{t=1}^{n} (\widehat{Y}_t - \langle \theta_*, X_t \rangle)^2 \le B_n + 2 \sum_{t=1}^{n} \eta_t (\widehat{Y}_t - \langle \theta_*, X_t \rangle) \,. \quad (3)$$

The sequence $\{\sum_{t=1}^{n} \eta_t (\widehat{Y}_t - \langle \theta_*, X_t \rangle)\}_{n=0}^{\infty}$ is a martingale adapted to $\{F_{n+1}\}_{n=0}^{\infty}$. We upper bound its tail using Corollary 8 with $V = 1$ (in the corollary, $Z_t = \widehat{Y}_t - \langle \theta_*, X_t \rangle$). Note that instead of the self-normalized inequality we use, we could get a self-normalized form from Freedman's inequality by using a peeling/stratification argument (see, e.g., inequality (53) in the paper by Audibert et al. (2008)). The price of doing this is a $\log \log n$ factor. It does not appear to be particularly better or easier than what we do.

Corollary 8 gives that with probability at least $1-\delta$, for all $n \ge 0$

$$\left| \sum_{t=1}^{n} \eta_t (\widehat{Y}_t - \langle \theta_*, X_t \rangle) \right| \le R \sqrt{2 \left(1 + \sum_{t=1}^{n} (\widehat{Y}_t - \langle \theta_*, X_t \rangle)^2 \right) \times \ln\left(\frac{\sqrt{1 + \sum_{t=1}^{n} (\widehat{Y}_t - \langle \theta_*, X_t \rangle)^2}}{\delta} \right)} \,.$$

Combining with (3), we get

$$\begin{aligned} &\sum_{t=1}^{n} (\widehat{Y}_t - \langle \theta_*, X_t \rangle)^2 \le B_n \\ &+ 2R \sqrt{2 \left(1 + \sum_{t=1}^{n} (\widehat{Y}_t - \langle \theta_*, X_t \rangle)^2 \right)} \\ &\times \sqrt{\ln\left(\frac{\sqrt{1 + \sum_{t=1}^{n} (\widehat{Y}_t - \langle \theta_*, X_t \rangle)^2}}{\delta} \right)} \,. \end{aligned} \quad (4)$$

From this point on, we just need to "solve" this inequality. More precisely, our goal is to isolate a simple function of θ_*. We proceed as follows. We first add 1 to the both sides of the inequality and introduce the notation $z = \sqrt{1 + \sum_{t=1}^{n} (\widehat{Y}_t - \langle \theta_*, X_t \rangle)^2}$, $a = B_n + 1$ and $b = 2R\sqrt{2\ln(z/\delta)}$. With this notation, we can write the last equation equivalently in the form

$$z^2 \le a + bz \; .$$

Since $a \ge 0$ and $b \ge 0$ (since $z \ge 1$ and $\delta \in (0, 1/4]$) we can apply Proposition 9 and obtain that

$$z \le b + \sqrt{a} \; .$$

Substituting for b we have

$$z \le R\sqrt{8\ln(z/\delta)} + \sqrt{a} \; .$$

Introducing the notation $c = \sqrt{a}$ and $f = R\sqrt{8}$ we can write the last inequality equivalently as

$$z \le c + f\sqrt{\ln(z/\delta)} \; .$$

Therefore, by Proposition 10,

$$z \le c + f\sqrt{2\ln\left(\frac{f+c}{\delta}\right)} \; .$$

Substituting for c, a and f we get

$$z \le \sqrt{B_n + 1} + 4R\sqrt{\ln\left(\frac{R\sqrt{8} + \sqrt{1+B_n}}{\delta}\right)} \; .$$

Squaring both sides and using the inequality $(u+v)^2 \le 2u^2 + 2v^2$ valid for any $u, v \in \mathbb{R}$, we have

$$z^2 \le 2B_n + 2 + 32R^2 \ln\left(\frac{R\sqrt{8} + \sqrt{1+B_n}}{\delta}\right) \; .$$

Substituting for z^2 and subtracting 1 from both sides we get

$$\sum_{t=1}^{n} (\widehat{Y}_t - \langle \theta_*, X_t \rangle)^2 \le 1 + 2B_n + 32R^2 \ln\left(\frac{R\sqrt{8} + \sqrt{1+B_n}}{\delta}\right) \; .$$

This means that $\theta_* \in C_n$ and the proof is finished. □

4 Application to Sparse Stochastic Linear Bandits

In this section we first define the stochastic linear bandits and their sparse variant. Next, we review how the so-called "optimism in the face of uncertainty" principle can be applied to this problem and how the construction of the previous section gives rise to novel results.

4.1 Stochastic Linear Bandits

In each round t, the learner is given a decision set $D_t \subseteq \mathbb{R}^d$ from which he has to choose a vector X_t, which, keeping synchrony with the literature, we shall call an *action*. Subsequently, he observes the reward $Y_t = \langle X_t, \theta_* \rangle + \eta_t$, where $\theta_* \in \mathbb{R}^d$ is an unknown parameter. Here the "noise sequence", $\{\eta_t\}_{t=1}^{\infty}$, is a sequence of conditionally R-sub-Gaussian random variables, as in the previous section (cf. (1)). In particular, $\mathbf{E}[\, \eta_t \,|\, X_{1:t}, \eta_{1:t-1}] = 0$ must hold for $t = 1, 2, \ldots$.

The goal of the learner is to maximize his total expected reward $\mathbf{E}[\sum_{t=1}^{n} \langle X_t, \theta_* \rangle]$, accumulated over the course of n rounds. Clearly, with the knowledge of θ_*, the optimal strategy in round t is to choose the point $x_t^* = \operatorname{argmax}_{x \in D_t} \langle x, \theta_* \rangle$, i.e., the action that maximizes the expected reward for that round. This strategy would accumulate a total expected reward $\mathbf{E}[\sum_{t=1}^{n} \langle x_t^*, \theta_* \rangle]$. It is natural to evaluate the learner relative to this optimal strategy. The difference of the learner's total expected reward and the total expected reward of the optimal strategy is called the *expected total regret*. The expected total regret is the expected value of the algorithm's *pseudo-regret* (Audibert et al., 2009),

$$\begin{aligned} R_n &\stackrel{\text{def}}{=} \left(\sum_{t=1}^{n} \langle x_t^*, \theta_* \rangle\right) - \left(\sum_{t=1}^{n} \langle X_t, \theta_* \rangle\right) \\ &= \sum_{t=1}^{n} \langle x_t^* - X_t, \theta_* \rangle \; . \end{aligned}$$

In what follows, for simplicity, we use the word *regret* instead of the more precise pseudo-regret in connection to R_n. Note that R_m has nothing, whatsoever, to do with $\rho_n(\theta)$ of the previous section, except for sharing the same name.

The goal of the algorithm is to keep the regret R_n as low as possible. As a bare minimum, we require that the algorithm is Hannan consistent, i.e., $R_n/n \to 0$ with probability one. Our goal will be to design algorithms for which the regret is low if θ_* is sparse, that is, if most coordinates of θ_* are zero. This is what we call the *sparse* variant of the stochastic linear bandit problem.

In order to obtain meaningful upper bounds on the regret, we will place assumptions on $\{D_t\}_{t=1}^{\infty}$, θ_*. We will assume that a prior bounds are known on the norm of θ_* and the norm of actions in $\{D_t\}_{t=1}^{\infty}$.

4.2 Optimism in the Face of Uncertainty

A natural and successful way to design an algorithm is the *optimism in the face of uncertainty principle* (OFU). The basic idea is that the algo-

for $t := 1, 2, \ldots$ **do**
  $(X_t, \widetilde{\theta}_t) = \operatorname{argmax}_{(x,\theta)\in D_t \times C_{t-1}} \langle x, \theta \rangle$
  Play X_t and observe reward Y_t
  Update C_t
end for

Figure 1: OFUL ALGORITHM

rithm maintains a confidence set $C_{t-1} \subseteq \mathbb{R}^d$ for the parameter θ_*. It is required that C_{t-1} can be calculated from $(X_1, Y_1, X_2, Y_2, \ldots, X_{t-1}, Y_{t-1})$ and $(D_1, \ldots, D_{t-1})$ and "with high probability" θ_* lies in C_{t-1}. The algorithm chooses an optimistic estimate $\widetilde{\theta}_t = \operatorname{argmax}_{\theta \in C_{t-1}} (\max_{x \in D_t} \langle x, \theta \rangle)$ and then chooses action $X_t = \operatorname{argmax}_{x \in D_t} \left\langle x, \widetilde{\theta}_t \right\rangle$ which maximizes the reward according to the estimate $\widetilde{\theta}_t$. Equivalently, and more compactly, the algorithm chooses the pair

$$(X_t, \widetilde{\theta}_t) = \operatorname*{argmax}_{(x,\theta)\in D_t \times C_{t-1}} \langle x, \theta \rangle \ ,$$

which *jointly* maximizes the reward. We call the resulting algorithm the OFUL algorithm, for "optimism in the face of uncertainty linear bandit algorithm". The pseudo-code of the algorithm is given in Figure 1.

The crux of the problem is the construction of the confidence sets C_t. One method is to use our online-to-confidence-set construction. This is what we explore in the next section.

4.3 Regret Analysis of OFUL

Consider the OFUL ALGORITHM from Figure 1 that uses the confidence set C_n constructed in Corollary 2 from an online linear prediction algorithm. To keep the analysis general, we leave the underlying linear prediction algorithm unspecified and we only assume that for all $n \geq 0$ it satisfies the regret bound $\rho_n(\theta_*) \leq B_n$.

We introduce a shorthand notation for the right-hand side of the inequality in Corollary 2 specifying the confidence set C_n:

$$\beta_n(\delta) = E^2 + 1 + 2B_n + 32R^2 \ln\left(\frac{R\sqrt{8} + \sqrt{1 + B_n}}{\delta}\right) \ .$$

The next two theorems upper bound the regret R_n of the resulting OFUL ALGORITHM. The proofs, which are largely based on the work of Dani et al. (2008) and are included for completeness, can be found in Appendix C.

Theorem 3 (Regret of OFUL)**.** *Assume that $\|\theta_*\|_2 \leq E$ and assume that for all $t \geq 1$ and for all $x \in D_t$, $\|x\|_2 \leq X$ and $|\langle x, \theta_* \rangle| \leq G$. Then, for any $\delta \in (0, 1/4]$, with probability at least $1 - \delta$, for any $n \geq 0$, the regret of the* OFUL *algorithm is bounded as*

$$R_n \leq 2\max\{1, G\}\sqrt{2nd \ln\left(1 + \frac{nX^2}{d}\right) \max_{0 \leq t < n} \beta_t(\delta)} \ .$$

Similar to Dani et al. (2008), we can also have a problem dependent logarithmic regret bound when the "gap" is positive. Dani et al. (2008) defines the gap as the difference in the rewards of the best and the second best actions in the extremal points of the action set.

Theorem 4 (Problem Dependent Regret Bound of OFUL)**.** *Assume the same as in Theorem 3 and additionally assume that the gap Δ is positive, then, for any $\delta \in (0, 1/4]$, with probability at least $1 - \delta$, for any $n \geq 0$, the regret of the* OFUL *algorithm is bounded as*

$$R_n \leq \frac{8d}{\Delta} \ln\left(1 + \frac{nX^2}{d}\right) \max_{0 \leq t < n} \beta_t(\delta) \max\{1, G^2\} \ .$$

To simplify things a bit, here and in the rest of the paper we view E, X, G, R as constants. Then the problem dependent and independent regrets of OFUL are $\widetilde{O}(dB_n \ln n/\Delta)$ and $\widetilde{O}(\sqrt{ndB_n})$, respectively. [4] Consequently, *smaller regret bound for the online prediction algorithm translates (via Theorem 3) into a smaller regret bound regret of* OFUL.

As the theorems show the regret of OFUL depends on the regret of the online learning algorithm that we use as a sub-routine to construct the confidence set. In particular, in order to achieve $O(\text{polylog}(n)\sqrt{n})$ uniform regret for OFUL, one needs an online learning algorithm with $O(\text{polylog}(n))$ regret bound.

Unfortunately, for some of the popular algorithms, such as the exponentiated gradient, the p-norm algorithms, and also for online LASSO the best known regret bounds are of the order $O(\sqrt{n})$; see (Kivinen and Warmuth, 1997) and (Cesa-Bianchi and Lugosi, 2006, Chapter 11). The main reason for the mediocre $O(\sqrt{n})$ regret bounds seems to be that these algorithm use only gradient information about the quadratic prediction loss function $\ell_t(\langle \cdot, X_t \rangle)$.

Better bounds are available for, e.g., the online regularized least squares algorithm (i.e., ridge regression) that also uses Hessian information:

Theorem 5 (Regret of ridge regression (Cesa-Bianchi and Lugosi, 2006))**.** *Let $\{\theta_t\}_{t=1}^{n+1}$ be the sequence generated by the Follow the Regularized Leader algorithm*

[4] $\widetilde{O}(\cdot)$ hides polylogarithmic factors in $n, d, X, \|\theta_*\|_0$ and $\|\theta_*\|_1$.

on the quadratic loss with the quadratic regularizer $R(\theta) = \|\theta\|_2^2/2$. *The FTRL algorithm with learning rate* $\eta > 0$ *satisfies the following bound that holds for all* $n \geq 1$ *and all* $(x_1, y_1), \ldots, (x_n, y_n)$

$$\sum_{t=1}^{n} \ell_t(\widehat{y}_t) \leq \inf_{\theta \in \mathbb{R}^d} \left\{ \sum_{t=1}^{n} \ell_t(\langle \theta, x_t \rangle) + \frac{\|\theta\|_2^2}{2\eta} \right\} + \frac{L_n d}{2} \ln\left(1 + \frac{\eta X^2 n}{d}\right),$$

where $L_n = \max_{1 \leq t \leq n} \ell_t(\langle \theta_t, x_t \rangle)$.

By combining Theorems 3 and 5, we get that the regret of OFUL with ridge regression is $\widetilde{O}(d\sqrt{n})$. Note that this latter bound essentially matches the bound obtained by Dani et al. (2008).

In online linear prediction one approach to exploit sparsity (when present) is to use an online ℓ^1-regularized least-squares method. To be able to demonstrate that sparsity can indeed be exploited in stochastic linear bandits, one then needs results similar to Theorem 5 for this algorithm, under sparsity assumption. This was an open problem until recently, when Gerchinovitz (2011) proposed the SeqSEW algorithm, which is based on the sparse exponential weighting algorithm introduced by Dalalyan and Tsybakov (2007), and proved the following logarithmic regret bound for it.

Theorem 6 (Regret of SeqSEW (Gerchinovitz, 2011))**.** *The* SeqSEW *algorithm introduced by Gerchinovitz (2011) satisfies the following bound that holds for all* $n \geq 1$ *and all* $(x_1, y_1), \ldots, (x_n, y_n)$

$$\sum_{t=1}^{n} \ell_t(\widehat{y}_t) \leq \inf_{\theta \in \mathbb{R}^d} \left\{ \sum_{t=1}^{n} \ell_t(\langle \theta, x_t \rangle) + H_n(\theta) \right\} + (1 + 38 \max_{1 \leq t \leq n} y_t^2) A_n,$$

where

$$H_n(\theta) = 256 \left(\max_{1 \leq t \leq n} y_t^2 \right) \|\theta\|_0 \ln\left(e + \sqrt{\sum_{t=1}^{n} \|x_t\|^2} \right) + 64 \left(\max_{1 \leq t \leq n} y_t^2 \right) A_n \|\theta\|_0 \ln\left(1 + \frac{\|\theta\|_1}{\|\theta\|_0}\right) \quad (5)$$

and

$$A_n = 2 + \log_2 \ln\left(e + \sqrt{\sum_{t=1}^{n} \|x_t\|^2} \right).$$

Theorem 6 motivates the OFUL algorithm presented in Figure 2 that uses the SeqSEW algorithm of Gerchinovitz (2011) as an online learning sub-routine. By combining Theorems 3 and 6, we get that the regret

for $t := 1, 2, \ldots$ **do**
  Construct C_{t-1} from Corollary 2
  $(X_t, \widetilde{\theta}_t) = \operatorname{argmax}_{(x,\theta) \in D_t \times C_{t-1}} \langle x, \theta \rangle$
  Predict $\widehat{Y}_t$ from SeqSEW
  Play X_t and observe reward Y_t
  Update C_t
end for

Figure 2: OFUL with SeqSEW

of the OFUL with SeqSEW algorithm is bounded, with probability at least $1 - \delta$, as

$$R_n \leq 2 \max\{1, G\} \sqrt{2nd \log\left(1 + \frac{nX^2}{d}\right) \max_{0 \leq t < n} \beta_t(\delta)}\,, \quad (6)$$

where

$$\beta_t(\delta) = E^2 + 1 + 2B_n(\theta_*) + 32R^2 \ln\left(\frac{R\sqrt{8} + \sqrt{1 + B_n(\theta_*)}}{\delta} \right)$$

$B_n(\theta_*) = H_n(\theta_*) + (1 + 38 \max_{1 \leq t \leq n} y_t^2) A_n$ and H_n, A_n are defined as in Theorem 6.

From the sub-Gaussianity assumption (1), we have that with probability $1 - \delta$, for any time $t \leq n$,

$$|y_t| \leq G + R\sqrt{2 \log(n/\delta)}.$$

Thus the regret (6) can be compactly written as $\widetilde{O}(\sqrt{d\|\theta_*\|_0 n})$. Compared to the $\widetilde{O}(d\sqrt{n})$ bound of Dani et al. (2008), the regret bound of OFUL with SeqSEW is lower when $\|\theta_*\|_0 < d$, which is the case for sparse vectors. Similarly, by application of Theorem 6 to the problem dependent regret bound of OFUL in Theorem 4, the $\widetilde{O}(d^2 \log^3 n/\Delta)$ problem dependent bound of Dani et al. (2008) can be improved to $\widetilde{O}(d\|\theta_*\|_0 \log^2 n/\Delta)$.

Notice that the regret bound of OFUL with SeqSEW still depends on d. A slight modification of the usual lower bound for d-armed bandit (Cesa-Bianchi and Lugosi, 2006, Chapter 6) will give us that even if sparsity is $p = 1$ then the regret must be $O(\sqrt{dn})$. More specifically, we cook up d arms, so that a random arm has a small reward and all the others have reward 0. This is equivalent to having a sparse θ_* with one non-zero component. Antos and Szepesvári (2009) provide another lower bound of the same order when the action set is the unit ball. This shows that the $\sqrt{d}$ term in the regret is unavoidable, which is in contrast to sparsity regret bounds for the full information online learning problems.

5 Compressed Sensing and Bandits

In their parallel submission Carpentier and Munos (2012) employ compressed sensing techniques to estimate the support of θ_* and achieve sparsity regret bounds of order of $\widetilde{O}(p\sqrt{n})$. Their setting is different than ours in two aspects. First, they consider the case when the action set is the unit ball, which makes it possible to satisfy the isotropic conditions that are required for compressed sensing. In contrast, our results hold for any bounded action set. The second difference, that also explains why they can avoid the $\sqrt{d}$ in their upper bound, is that they assume noise "in the parameters" in the sense that their loss function takes the form of $\ell_t = \langle x_t, \theta_* \rangle + \langle x_t, \eta_t \rangle$.

6 Conclusion

The main new technical contribution of the paper is a novel regret-to-confidence-set construction, which works for linear prediction problems with martingale noise. We have also introduced stochastic sparse linear bandits, a natural framework for bandit linear optimization when the number of features is large, but many of them might be irrelevant. With our reduction, we obtained the first results for this problem: We showed that the combination of a recent algorithm for online linear prediction and our construction gives rise to an algorithm that is able to exploit the sparsity of the unknown parameter vector. In particular, for sparse parameter vectors, we can demonstrate a better bound that was possible with previous techniques. It is important to emphasize that the main advantage of our approach, being based on the idea of reductions (Langford, 2011), is that a better regret bound for online linear prediction is automatically transformed into a better regret bound for linear bandits.

Acknowledgements

This work was supported in part by NSERC, AITF and the Alberta Ingenuity Centre for Machine Learning.

References

Yasin Abbasi-Yadkori, András Antos, and Csaba Szepesvári. Forced-exploration based algorithms for playing in stochastic linear bandits. In *COLT Workshop on On-line Learning with Limited Feedback*, 2009.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2011a.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. Arxiv preprint http://arxiv.org/abs/1102.2670, 2011b.

András Antos and Csaba Szepesvári. Personal Communication, 2009.

Jean Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. In *Theoretical Computer Science-2008*, 2008.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002b.

Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence*, 2009.

Emmanuel J. Candès. Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 1433–1452, 2006.

Alexandra Carpentier and Remi Munos. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *Proceedings of fifteenth international conference on Artificial Intelligence and Statistics*, 2012.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Proceedings of the 20th Annual Conference on Learning Theory*, pages 97–111, 2007.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feed-

back. In Rocco Servedio and Tong Zhang, editors, *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, pages 355–366, 2008.

Ofer Dekel and Yoram Singer. Data-driven online to batch conversions. *NIPS 2005*, 18:267, 2006.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *Fifteenth Annual Conference on Computational Learning Theory (COLT)*, pages 255–270, 2002.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

Claudio Gentile and Nick Littlestone. The robustness of the p-norm algorithms. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 1–11, New York, NY, USA, 1999. ACM.

Sèbastien Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. In *Proceedings of the 24st Annual Conference on Learning Theory (COLT 2011)*, 2011.

Adam J. Grove, Nick Littlestone, and Dale Schuurmans. General convergence results for linear discriminant updates. In *Machine Learning*, pages 171–183. ACM Press, 1997.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563—1600, 2010.

Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, January 1997.

Tze Leung Lai, Herbert Robbins, and Ching Zong Wei. Strong consistency of least squares estimates in multiple regression. *Proceedings of the National Academy of Sciences*, 75(7):3034–3036, 1979.

John Langford. Machine learning reductions, 2011. `http://hunch.net/~jl/projects/reductions/reductions.html`.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*, pages 661–670. ACM, 2010.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988.

Nicolas Littlestone. From on-line to batch learning. In *Annual Workshop on Computational Learning Theory: Proceedings of the second annual workshop on Computational learning theory (COLT 1989)*. Association for Computing Machinery, Inc, One Astor Plaza, 1515 Broadway, New York, NY, 10036-5701, USA, 1989.

Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical bernstein stopping. In Andrew McCallum and Sam Roweis, editors, *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 672–679, 2008.

Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodolical)*, 58(1):267–288, 1996.

Vladimir Vovk. Competitive on-line statistics. *International Statistical Review*, 69:213–248, 2001.

Thomas J. Walsh, István Szita, Carlso Diuk, and Michael L. Littman. Exploring compact reinforcement-learning representations with linear regression. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pages 591–598. AUAI Press, 2009.

Larry Wasserman. *All of Nonparametric Statistics*. Springer, 1998.