Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Weighted Bandits or: How Bandits Learn Distorted Values That Are Not Expected

Aditya Gopalan
Indian Institute of Science
aditya@ece.iisc.ernet.in

Prashanth L.A., Michael Fu, Steve Marcus
University of Maryland
{prashla,mfu,marcus}@umd.edu

Abstract

Motivated by models of human decision making proposed to explain commonly observed deviations from conventional expected value preferences, we formulate two stochastic multi-armed bandit problems with distorted probabilities on the cost distributions: the classic K-armed bandit and the linearly parameterized bandit. In both settings, we propose algorithms that are inspired by Upper Confidence Bound (UCB) algorithms, incorporate cost distortions, and exhibit sublinear regret assuming Hölder continuous weight distortion functions. For the K-armed setting, we show that the algorithm, called W-UCB, achieves problem-dependent regret $O\left(L^2M^2\log n/\Delta^{\frac{2}{\alpha}-1}\right)$, where n is the number of plays, Δ is the gap in distorted expected value between the best and next best arm, L and α are the Hölder constants for the distortion function, and M is an upper bound on costs, and a problem-independent regret bound of $O((KL^2M^2)^{\alpha/2}n^{(2-\alpha)/2})$. We also present a matching lower bound on the regret, showing that the regret of W-UCB is essentially unimprovable over the class of Hölder -continuous weight distortions. For the linearly parameterized setting, we develop a new algorithm, a variant of the Optimism in the Face of Uncertainty Linear bandit (OFUL) algorithm (Abbasi-Yadkori, Pál, and Szepesvári 2011) called WOFUL (Weight-distorted OFUL), and show that it has regret $O(d\sqrt{n}\ \mathrm{polylog}(n))$ with high probability, for sub-Gaussian cost distributions. Finally, numerical examples demonstrate the advantages resulting from using distortion-aware learning algorithms.

Introduction

Consider the following two-armed bandit problem: The rewards of Arm 1 are \$1 million w.p. $1/10^6$ and 0 otherwise, while Arm 2 rewards are \$1000 w.p. $1/10^3$ and 0 otherwise. In this case, humans would usually prefer Arm 1 over Arm 2. The human preferences get flipped if we change to costs, i.e., Arm 1 loses a million with a very low probability of $1/10^6$, while Arm 2 loses \$1000 w.p. $1/10^3$. In this case, Arm 2 is preferred over Arm 1. The above example illustrates that traditional expected value falls short in explaining human preferences, and the reader is referred to the classic Allais problem (Allais 1953) that rigorously argues against expected utility theory as a model for human-based decision making systems.

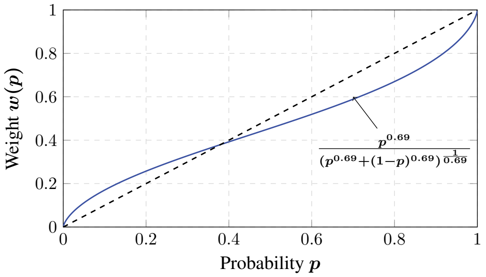


Figure 1: Graphical illustration of a typical weight function that inflates low probabilities and deflates large probabilities. The weight function used in the figure is the one recommended by Tversky and Kahneman (1992) based on empirical tests involving human subjects.

Violations of the expected value-based preferences in human-based decision making systems can be alleviated by incorporating distortions in the underlying probabilities of the system (Starmer 2000) (Quiggin 2012, Chapter 4). Probabilistic distortions have a long history in behavioral science and economics, and we bring this idea to a multi-armed bandit setup. In particular, we base our approach on rank-dependent expected utility (RDEU) (Quiggin 2012), which includes the popular *cumulative prospect theory* (CPT) of Tversky and Kahneman (1992).

The distortions happen via a weight function $w : [0, 1] \to [0, 1]$ that transforms probabilities in a nonlinear fashion. As illustrated in Figure 1, a typical weight function, say $w : [0, 1] \to [0, 1]$, has the inverted-S shape. In other words, w inflates low probabilities and deflates large probabilities and can explain human preferences well. For instance, in the example above, if we choose $w(1/10^6) > 1/10^6$ and take expectations w.r.t. the w-distorted distribution, then Arm 1 would be preferable when the problem is setup with rewards and Arm 2 for the problem with costs. The suitability of this approach, esp. with a inverted-S shaped weight function, to model human decision making (and thus preferences) has been widely documented (Prelec 1998; Wu and Gonzalez 1996; Conlisk 1989; Camerer 1989; 1992; Cherny and Madan 2009; Gonzalez and Wu 1999).

We use a traveler's route choice as a running example to illustrate the main ideas in this paper. The setting here is that a human travels from a source (e.g., home) to a destination (e.g., office), both fixed, every day. He/she has multiple routes to choose from, and each route incurs a stochastic delay with unknown distributions. The problem then is to choose a route that minimizes some function of delay. Using expected delay may not lead to a routing choice that is appealing to the human traveler. In addition to the examples mentioned earlier, intuitively humans would prefer a route with a slight excess of delay over another that has a small probability of getting into a traffic jam that takes hours to be resolved. The requirement here is for an automated routing algorithm, say one that sits as an application on the traveler's mobile device, that learns the best route for the traveler. The algorithm is online and uses the delay information for a recommended route as feedback to find the best route. We treat this problem in two regimes: first, a setting where the number of routes is small, so the traveler can afford to try each of the routes a small number of times before fixing on the "best" route; second, a big road network setting that involves a large number of routes, which prohibits an approach that requires trying the bulk of the routes before deciding which is the "best".

We formalize two probabilistically distorted bandit settings that correspond to the two routing setups mentioned above. The first is the classic K-armed setting, while the second is the linear bandit setting. In both settings, we define the weight-distorted value μ_x for any arm x in the space of arms $\mathcal{X}$ as follows:

$$\mu_x = \int_0^\infty w(\mathbb{P}\left[\mathcal{Y}_x > z\right])dz - \int_0^\infty w(\mathbb{P}\left[-\mathcal{Y}_x > z\right])dz, \quad (1)$$

where w is the weight function that satisfies $w(0) = 0$ and $w(1) = 1$ and $\mathcal{Y}_x$ is the random variable (r.v.) corresponding to the stochastic rewards from arm $x \in \mathcal{X}$. By choosing the identity weight function $w(p) = p$, we obtain $\mu_x = \mathbb{E}(\mathcal{Y}_x^+) - \mathbb{E}(\mathcal{Y}_x^-) = \mathbb{E}(\mathcal{Y}_x)$, where $y^+ = \max(y, 0)$ and $y^- = \max(-y, 0)$ denote the positive and negative parts of $y \in \mathbb{R}$, respectively. Thus, μ_x as in (1) generalizes standard expected value. As discussed earlier, w has to be chosen in a non-linear fashion, to capture human preferences, which has strong empirical support.

In our setting, the goal is find an arm x_* that maximizes (1). The problem is challenging because the current bandit solutions, for instance, the popular UCB algorithm, cannot handle distortions. This is because the environment provides samples from the distribution F_x when arm x is pulled, while the integral in (1) involves a distorted distribution. The implication is that a simple sample mean and a confidence term suggested by the Hoeffding inequality is enough to derive the UCB values for any arm in the regular setting involving expected values. On the other hand, one requires a good enough estimate of F_x to estimate μ_x. Just for the sake of example, a (α-Hölder continuous) weight function such as $w(t) := t^\alpha$, when applied to a Bernoulli(p) distribution, distorts the mean to p^α from p, and can introduce an arbitrarily large scaling for arms with real expectations close to 0. It follows that nonlinear weight distortion can, in fact, change the order of the optimal arm, resulting in a distortion-unaware algorithm like UCB converging to the *wrong* arm and incurring linear regret. The W-UCB algorithm that we propose incorporates a empirical distribution-based approach, similar to that of Prashanth et al. (2016), to estimate μ_x. However, unlike the latter, our algorithm incorporates a confidence term relying on the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Wasserman 2015) that ensures the W-UCB values are a high-probability bound on the true value μ_x. We provide upper bounds on the regret of W-UCB, assuming w is Hölder continuous and provide empirical demonstrations on a setting from Tversky and Kahneman (1992).

Next, we consider a linear bandit setting with weight distortions that can be motivated as follows: Consider a network graph $G = (V, E)$, $|E| = d$, with a source $s \in V$ and destination $t \in V$. The interaction proceeds over multiple rounds, where in each round m, the user picks a route x_m from s to t (a route is a collection of edges encoded by a vector of $0 - 1$ values in d dimensions) and experiences a stochastic delay $x_m^\intercal (\theta + N_m)$. Here $\theta \in \mathbb{R}^d$ is an underlying model parameter and $N_m \in \mathbb{R}^d$ is a random noise vector, both unknown to the learner. The physical interpretation is that the nodes represent geographical locations (say junctions) and the edges are roads that connect nodes. An edge from i to j will have an edge weight θ_{ij}, which quantifies the delay for this edge.

Notice that the observations include a noise component that scales with the route chosen – this is unlike the model followed in earlier linear bandit works (Abbasi-Yadkori, Pál, and Szepesvári 2011; Dani, Hayes, and Kakade 2008), where the noise was independent of the arm chosen. Our noise model makes practical sense because the observed delay in a road traffic network depends on the length of the route, for e.g., one would expect more noise in a detour involving ten roads than in a direct one-road route. The aim is to find a low-delay route that satisfies the user. The setting is such that the number of routes is large (so the regular K-armed bandits don't scale) and one needs to utilize the linearity in the costs (delays) to find the optimal route, where optimality is qualified in terms of a weight-distorted expectation.

For the linear bandit setting, we propose a variant of the OFUL algorithm (Abbasi-Yadkori, Pál, and Szepesvári 2011), that incorporates weight-distorted values in the arm selection step. The regret analysis of the resulting WOFUL algorithm poses novel challenges compared to that in the linear bandit problem, primarily because the instantaneous cost (and hence regret) at each round is in fact a *nonlinear* function of the features of the played arm. This occurs due to the distortion in expectation caused by the weight function, although the actual observation (e.g., network delay in the example above) is linear in expectation over the played arm's features. The weight function can not only change the optimal arm but can also potentially amplify small differences in real expected values of arms to much larger values, leading to a blowing up of overall regret. However, our analysis shows that regret in weight-distorted cost as the performance metric can be controlled at the same rate as that in standard linear bandit models, irrespective of the structure of the distortion function. More specifically, we show, using a careful analysis of the effect of weight distortion, that the regret of the WOFUL algorithm is no more than $O\left(d\sqrt{n}\,\text{polylog}(n)\right)$ in n rounds with high probability, similar to the guarantee enjoyed by the OFUL algorithm in linear bandits (note however that the identity of the optimal arm may be different due to weight distortion in costs).

Related work: The closest related previous contribution is that of Prashanth et al. (2016), where the authors bring in ideas from CPT to a reinforcement learning (RL) setting. In contrast, we formulate two multi-armed bandit models that incorporate weight-distortions. From a theoretical standpoint, we handle the exploration-exploitation tradeoff via UCB-inspired algorithms, while the focus of Prashanth et al. (2016) was to devise a policy-gradient scheme given biased estimates of a certain CPT-value defined for each policy. Moreover, we provide finite-time regret bounds for both bandit settings, while the guarantees for the policy gradient algorithm in Prashanth et al. (2016) are asymptotic in nature.

Previous works involving RDEU and CPT are huge in number; at a conceptual level, the work in this paper integrates machine learning (esp. bandit learning) with an RDEU approach that involves a probabilistic distortions via a weight function. To the best of our knowledge, RDEU/CPT papers in the literature assume model information, i.e., a setting where the distributions of the arms are known, while we have a *model-free* setting where one can only obtain sample values from the arms' distributions. Our setting makes practical sense; for instance, in the traveler's route choice problem one can only obtain sample delays for a particular route, while the distribution governing the delays for any route is not known explicitly.

K-armed bandit with weight distortion

Suppose there are K arms with unknown distributions $F_k, k = 1, \dots, K$. In each round $m = 1, \dots, n$, the algorithm pulls an arm $I_m \in \{1, \dots, K\}$ and obtains a sample cost from the distribution F_{I_m} of arm I_m.

The classic objective is to play (or pull) the arm whose expected cost is the least. In this paper, we take a different approach inspired by non-expected utility (EU) approaches and use the weight distorted-cost μ_k as the performance criterion for any arm k. The latter quantity, defined in (1), can be seen to be equivalent to the following:

$$\mu_k := \int_0^\infty w(1 - F_k(z))dz - \int_0^\infty w(F_k(-z))dz, \qquad (2)$$

where $w : [0, 1] \to [0, 1]$ is a weight function that distorts probabilities.

The optimal arm is one that minimizes the weight-distorted cost, i.e., $\mu_* = \min_k \mu_k$. The optimal arm is not necessarily unique, i.e., there may exist multiple arms with the optimal weight-distorted cost μ_*.

With the above notion of weight-distorted cost, we define the cumulative regret R_n as $R_n = \sum_{k=1}^K T_k(n)\mu_k - n\mu_*$, where $T_k(n) = \sum_{m=1}^n I(I_m = k)$ is the number of times arm k is pulled up to time n. The expected regret can be written as $\mathbb{E}R_n = \sum_{k=1}^K \mathbb{E}[T_k(n)]\Delta_k$, where $\Delta_k = \mu_k - \mu_*$ denotes the gap between the weight-distorted costs of the optimal arm and of arm k. Note that this definition of regret arises from the interpretation that each arm k is associated with a deterministic value μ_k. The least possible cumulative cost that can be suffered in n rounds is thus $n\mu_*$, while that suffered by a given strategy is $\sum_{k=1}^K T_k(n)\mu_k$. Thus, the regret as defined above is a measure of the rate at which a strategy converges to playing the optimal arm in the sense of weighted or distorted cost.

We remark that the regret performance measure, as defined above, is explicitly defined within a stochastic model for rewards. Thus, low-regret algorithms designed for the nonstochastic setting, e.g., EXP3 (Auer et al. 2003), are not inherently suitable for this problem, as they do not factor in the distortion caused in (expected) reward. A similar observation holds for conventional stochastic bandit algorithms such as UCB, and algorithms sensitive to arm reward variances such as UCB-V (Audibert, Munos, and Szepesvári 2009) – once weight distortion is incorporated, the algorithm will converge to an arm that is not weight-distorted value optimal. Thus, applying a variance-sensitive algorithm (like UCB-V) will still yield linear regret in the distorted setting.

W-UCB Algorithm

Estimating the weight-distorted cost for any arm k is challenging, and one cannot use a Monte Carlo approach with sample means because weight-distorted cost involves a distorted distribution, whereas the samples come from the undistorted distribution F_k. Thus, one needs to estimate the entire distribution, and for this purpose, we adapt the quantile-based approach of Prashanth et al. (2016).

Estimating μ_k: At time instant m, let $Y_{k,1}, \dots, Y_{k,l}$ denote the samples from the cost distribution F_k for arm k, where we have used l to denote the number of samples $T_k(m-1)$ for notational convenience. Order the samples in ascending fashion as $Y_{[k,1]} \le Y_{[k,2]} \le \cdots \le Y_{[k,l_b]} \le 0 \le Y_{[k,l_b+1]} \le \cdots \le Y_{[k,l]}$, where $l_b \in \{0, 1, 2, \dots, l\}$ denotes the index of a 'boundary' sample after which a sign change occurs. The first integral in (2) is estimated by the quantity

$$\widehat{\mu}^+_{k,l} := \sum_{i=l_b+1}^{l} Y_{[k,i]} \left(w\left(\frac{l+1-i}{l}\right) - w\left(\frac{l-i}{l}\right)\right), \qquad (3)$$

while the second integral in (2) is estimated by the quantity

$$\widehat{\mu}^-_{k,l} := \sum_{i=1}^{l_b} Y_{[k,i]} \left(w\left(\frac{i-1}{l}\right) - w\left(\frac{i}{l}\right)\right). \qquad (4)$$

We finally estimate μ_k as follows:

$$\widehat{\mu}_{k,l} = \widehat{\mu}^+_{k,l} - \widehat{\mu}^-_{k,l}. \qquad (5)$$

From (2) and (5) above, it can be seen that $\widehat{\mu}_{k,l}$ is the weight-distorted cost of the empirical distribution of samples from arm k seen thus far, i.e.,

$$\widehat{\mu}_{k,l} := \int_0^\infty w(1 - \hat{F}_{k,l}(z))dz - \int_0^\infty w(\hat{F}_{k,l}(-z))dz,$$

where $\hat{F}_{k,l}(x) := \frac{1}{l}\sum_{i=1}^{l} I_{[Y_{k,i}\leq x]}$ denotes the empirical distribution of r.v. $\mathcal{Y}_x$. In particular, the first and second integral above correspond to (3) and (4), respectively.

We next provide a sample complexity result for the accuracy of the estimator $\widehat{\mu}_{k,l}$ under the following assumptions:

(A1) The weight function w is Hölder continuous with constant L and exponent $\alpha \in (0,1]$: $\sup_{x\neq y} \frac{|w(x)-w(y)|}{|x-y|^{\alpha}} \leq L$.

(A2) The arms' costs are bounded by $M > 0$ almost surely.

(A1) is necessary to ensure that the weight-distorted value $\mu_k, k = 1, \ldots, K$ is finite. Moreover, the popular choice for the weight function, proposed by Tversky and Kahneman (1992) and illustrated in Figure 1, is Hölder continuous.

Theorem 1 (***Sample complexity of estimating distorted cost***)**.** *Assume (A1)-(A2). Then, for any $\epsilon > 0$ and any $k \in \{1, \ldots, K\}$, we have $P(|\widehat{\mu}_{k,m} - \mu_k| > \epsilon) \leq 2\exp\left(-2m(\epsilon/LM)^{2/\alpha}\right)$.*

For the special case of Lipschitz weight functions w, setting $\alpha = 1$ in the above theorem, we obtain a sample complexity of order $O\left(1/\epsilon^2\right)$ for accuracy ϵ.

B-values (weighted UCB values): At instant m, define the B-value for any arm k, as a function of the number of samples l and constants $\alpha \in [0,1], L > 0, M > 0$, as: $B_{m,l}(k) = \widehat{\mu}_{k,l} - \gamma_{m,l}$, where $\gamma_{m,l} := LM\left(\frac{3\log m}{2l}\right)^{\frac{\alpha}{2}}$. The r.v. $\widehat{\mu}_{k,l}$, defined by (5), is an estimate of μ_k that uses the $l = T_k(m-1)$ sample costs of arm k seen so far and $\gamma_{m,l}$ is the confidence width, which together with $\widehat{\mu}_{m,l}$ ensures that the true weight-distorted value μ_k lies within $[\widehat{\mu}_{k,l} - \gamma_{m,l}, \widehat{\mu}_{k,l} + \gamma_{m,l}]$ with high probability, i.e., for $k = 1, \ldots, K$, both $P\left(\widehat{\mu}_{k,l} + \gamma_{m,l} \leq \mu_k\right) \leq 2m^{-3}$ and $P\left(\widehat{\mu}_{k,l} - \gamma_{m,l} \geq \mu_k\right) \leq 2m^{-3}$.

Using the B-values defined above, the W-UCB algorithm chooses the arm I_m at instant m as follows:

If $m \leq K$, then play $I_m = m$ (initial round-robin phase),

$$\text{Else, play } I_m = \underset{k=\{1,\ldots,K\}}{\arg\min}\; B_{m,T_k(m-1)}(k). \tag{6}$$

Theorem 2 (***Regret bound***)**.** *Under (A1)-(A2), the expected cumulative regret R_n of W-UCB is bounded as follows:*

$$\mathbb{E}R_n \leq \sum_{\{k:\Delta_k>0\}} \frac{3(2LM)^{2/\alpha}\log n}{2\Delta_k^{2/\alpha-1}} + MK\left(1+\frac{2\pi^2}{3}\right).$$

The theorem above involves the gaps Δ_k. We next present a gap-independent regret bound in the following result:

Corollary 1 (***Gap-independent regret***)**.** *Under (A1)-(A2), the expected cumulative regret R_n of W-UCB satisfies the following gap-independent bound. There exists a universal constant $c > 0$ such that for all n, $\mathbb{E}R_n \leq MK^{\alpha/2}\left(\frac{3}{2}(2L)^{2/\alpha}\log n + c\right)^{\frac{\alpha}{2}}\; n^{\frac{2-\alpha}{2}}$.*

Remark 1. ***(Lipschitz weights)*** *We can recover the $O(\sqrt{n})$ regret bound (or same dependence on the gaps) as in regular UCB for the case when $\alpha = 1$, i.e., Lipschitz weights. On the other hand, when $\alpha < 1$, the regret bounds are weaker than $O(\sqrt{n})$.*

The following result shows that one cannot hope to obtain better regret than that of W-UCB (Theorem 2) over the class of Hölder-continuous weight functions, i.e., weight functions satisfying (A1)-(A2), by exhibiting a matching lower bound.

Theorem 3 (***Regret lower bound***)**.** *For any learning algorithm with sub-polynomial regret in the time horizon, there exists (1) a weight function which is monotone increasing and α-Hölder continuous with constant L, and (2) a set of cost distributions for the arms with support bounded by M, for which the algorithm's regret satisfies*

$$\mathbb{E}\left[R_n\right] = \Omega\left(\sum\nolimits_{\{k:\Delta_k>0\}} \frac{(LM)^{2/\alpha}\log n}{4\Delta_k^{2/\alpha-1}}\right).$$

A more precise statement of the above result, and the proofs of Theorems 1–3, can be found in (Gopalan et al. 2016).

Linearly parameterized bandit with weight distortion

The setting here involves arms that are given as the compact set $\mathcal{X} \subset \mathbb{R}^d$ (each element of $\mathcal{X}$ is interpreted as a vector of features associated with an arm). The learning game proceeds as follows. At each round $m = 1, 2, \ldots$, the learner

(a) plays an arm $x_m \in \mathcal{X}$, possibly depending on the history of observations thus far, and

(b) observes a stochastic, nonnegative cost given by

$$c_m := x_m^{\mathsf{T}}\left(\theta + N_m\right), \tag{7}$$

where $N_m := (N_m^1, \ldots, N_m^d)$ is a vector of i.i.d. standard Gaussian random variables, independent of the previous vectors $N_1, \ldots, N_{m-1}$, and $\theta \in \mathbb{R}^d$ is an underlying model parameter. Both θ and N_m, $m \geq 1$, are unknown to the learner.

Given a weight function $w : [0,1] \to [0,1]$, we define the weight-distorted cost $\mu(x, \theta)$ for arm $x \in \mathcal{X}$, with underlying model parameter θ, to be the quantity

$$\mu_x(\theta) := \int_0^{\infty} w(1-F_x^{\theta}(z))dz + \int_0^{\infty} w(F_x^{\theta}(-z))dz, \tag{8}$$

where $F_x^{\theta}(z) := \mathbb{P}\left[x^{\mathsf{T}}(\theta+N) \leq z\right]$, $z \in \mathbb{R}$, is the cumulative distribution function of the stochastic cost from playing arm $x \in \mathcal{X}$. An arm x is said to be optimal if its weight-distorted cost equals the least possible weight-distorted cost achieved across all arms, i.e., if $\mu_x = \mu_* := \min_{x'\in\mathcal{X}} \mu_{x'}(\theta)$. As in the K-armed setting, the performance measure is the cumulative regret R_n over n rounds, defined as $R_n = \sum_{m=1}^{n} \mu_{x_m}(\theta) - n\mu^*$, where x_m is the arm chosen by the bandit algorithm in round m.

Algorithm 1 presents the pseudocode for the proposed algorithm, which follows the general template for linear bandit algorithms (cf. ConfidenceBall in (Dani, Hayes, and Kakade 2008) or OFUL in (Abbasi-Yadkori, Pál, and Szepesvári 2011)), but deviates in the step when an arm is chosen. In particular, in any round m of the algorithm, WOFUL uses $\mu_x(\theta)$ as the decision criterion for any arm $x \in \mathcal{X}$ and $\theta \in C_m$, where $\mu_x(\theta)$ is the weight-distorted value

Algorithm 1 WOFUL

```
Input: regularization constant $\lambda \geq 0$, confidence $\delta \in (0, 1)$, norm bound $\beta$, weight function $w$.
Initialization: $A_1 = \lambda I_{d\times d}$ ($d \times d$ identity matrix), $b_1 = 0, \hat{\theta}_1 = 0$.
for $m = 1, 2, \dots$ do
    Set $C_m := \left\{ \theta \in \mathbb{R}^d : \left\| \theta - \hat{\theta}_m \right\|_{A_m} \leq D_m \right\}$ and
    $D_m := \sqrt{2 \log \left( \frac{\det(A_m)^{1/2}\ \lambda^{d/2}}{\delta} \right)} + \beta\sqrt{\lambda}$.
    Let $(x_m, \tilde{\theta}_m) := \underset{(x,\theta')\in\mathcal{X}\times C_m}{\arg\min}\ \mu_x(\theta')$.
    Choose arm $x_m$ and observe cost $c_m$.
    Update $A_{m+1} = A_m + \frac{x_m x_m^\mathsf{T}}{\|x_m\|^2}$,
    $b_{m+1} = b_m + \frac{c_m x_m}{\|x_m\|}$, and
    $\hat{\theta}_{m+1} = A_{m+1}^{-1} b_{m+1}$
end for
```

that is defined in (8) and C_m is the confidence ellipsoid that is specified in Algorithm 1. This is unlike regular linear bandit algorithms, which use $x^\mathsf{T}\theta$ as the cost for any arm $x \in \mathcal{X}$ and $\theta \in C_m$. Note that the "in-parameter" or arm-dependent noise model (7) also necessitates modifying the standard confidence ellipsoid construction of (Abbasi-Yadkori, Pál, and Szepesvári 2011) by rescaling with the arm size (the A_m and b_m variables in Algorithm 1). For a positive semidefinite matrix M and a vector x, we use the notation $\|x\|_M = \sqrt{x^\mathsf{T} M x}$ to denote the Euclidean norm of x weighted by M.

Remark 2. *(**Computation cost**) The computationally intensive step in WOFUL is the optimization of the weight-distorted value over an ellipsoid in the parameter space (the third line in the* **for** *loop). This can be explicitly solved as follows. For a fixed $x \in \mathcal{X}$, we can let $\bar{\theta}_{m,x} := \underset{\theta'\in C_m}{\arg\min}\, \mu_x(\theta') = \underset{\theta'\in C_m}{\arg\min}\, x^\mathsf{T}\theta' = \hat{\theta}_m - D_m A_m^{-1} x / \|x\|_{A^{-1}}$ This is because the weight-distorted value is monotone under translation (see Lemma 5 below). The cost-minimizing arm is thus computed as $x_m = \arg\min\{\mu_{x_1}(\bar{\theta}_{m,1}), \dots, \mu_{x_{|\mathcal{X}|}}(\bar{\theta}_{m,|\mathcal{X}|})\}$.*

Theorem 4 (***Regret bound for WOFUL***)**.** *Suppose that the weight function w satisfies $0 \leq w(p) \leq 1$, $\forall p \in (0, 1)$, $\forall x \in \mathcal{X} : x^\mathsf{T}\theta \in [-1, 1]$, and $\|\theta\|_2 \leq \beta$. Then, for any $\delta > 0$, the regret R_n of WOFUL, run with parameters $\lambda > 0$, B, δ and w, satisfies $P\left(R_n \leq \sqrt{32 d n D_n \log n}\ \ \forall n \geq 1\right) \geq 1 - \delta$.*

Remark 3. *If for all $x \in X$, $\|x\|_2 \leq \ell$, then the quantity D_n appearing in the regret bound above is $O\left(\sqrt{d \log\left(\frac{n\ell^2}{\lambda\delta}\right)}\right)$ (Abbasi-Yadkori, Pál, and Szepesvári 2011, Lemma 10); thus, the overall regret is*[1] $\tilde{O}\left(d\sqrt{n}\right)$.

Remark 4. *For the identity weight function $w(t) = t$, $0 \leq t \leq 1$ with $L = \alpha = 1$, we recover the stochastic linear bandit setting, and the associated $\tilde{O}\left(d\sqrt{n}\right)$ regret bound for linear bandit algorithms such as ConfidenceBall$_1$ and ConfidenceBall$_2$ (Dani, Hayes, and Kakade 2008), OFUL (Abbasi-Yadkori, Pál, and Szepesvári 2011). Hence, the result above is a generalization of regret bounds for standard linear bandit optimization to the case where a non-linear weight function of the cost distribution is to be optimized from linearly parameterized observations. The distortion of the cost distribution via a weight function, rather interestingly, does* not *impact the order of the bound on problem-independent regret, and we obtain $\tilde{O}\left(d\sqrt{n}\right)$ here as well.*

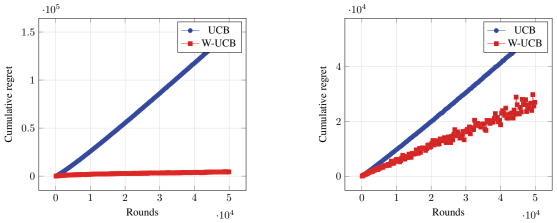


Figure 2: Cumulative regret along with standard error from 100 replications for the UCB and W-UCB algorithms for 2 stochastic K-armed bandit environments: (a) *Arm 1:* ($50 w.p. 0.1, $0 w.p. 0.9) **vs.** *Arm 2:* $7 w.p. 1. (b) *Arm 1:* ($500 w.p. 0.01, $0 w.p. 0.99) **vs.** *Arm 2:* ($250 w.p. 0.03, $0 w.p. 0.97).

Remark 5. *Note that the weight function w can be any non-linear function bounded in $[0, 1]$; unlike the K-armed setting, we do not impose a Hölder continuity assumption on w.*

Remark 6. *A lower bound of essentially the same order as Theorem 4 ($O(d\sqrt{n})$) holds for regret in (undistorted) linear bandits (Dani, Kakade, and Hayes 2007). One can show a similar lower bound argument with distortions, implying that the result of the theorem is not improvable (order-wise) across weight functions.*

Remark 7. *(**Linear bandits with arm-independent additive noise**) An alternative to modelling "in-parameter" or arm-dependent noise* (7) *is to have independent additive noise, i.e., $c_m := x_m^\mathsf{T}\theta + \eta_m$. This is a standard model of stochastic observations adopted in the linear bandit literature (Abbasi-Yadkori, Pál, and Szepesvári 2011; Dani, Hayes, and Kakade 2008). The key difference here is that, unlike the setting in* (7)*, the noise component η_m does* not *depend on the arm played x_m. In this case, Lemma 5 below shows that $\mu_{X+a} \geq \mu_X$, i.e., the distorted CPT value μ preserves order under translations of random variables. As a consequence of this fact, the WOFUL algorithm reduces to the OFUL algorithm in the standard linear bandit setting with arm-independent noise.*

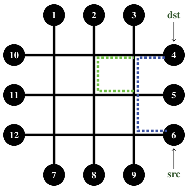
Proof sketch for Theorem 4. We upper-bound the instantaneous regret r_m as follows: Letting $\hat{x}_m = \frac{x_m}{\|x_m\|}$ and $\mathcal{N}$ to be a standard Gaussian r.v. in d dimensions, we have

$$r_m = \mu_{x_m}(\theta) - \mu_{x_*}(\theta) \leq \mu_{x_m}(\theta) - \mu_{x_m}(\tilde{\theta}_m)$$

$$= \|x_m\| \left(\mu_{W + \hat{x}_m^\mathsf{T}\theta} - \mu_{W + \hat{x}_m^\mathsf{T}\tilde{\theta}_m} \right) \qquad (9)$$

$$\leq 2\,\|x_m\| \left| \hat{x}_m^\mathsf{T}(\theta - \tilde{\theta}_m) \right|, \qquad (10)$$

[1] $\tilde{O}(\cdot)$ is a variant of the $O(\cdot)$ that ignores log-factors.



	Expected value $x_{\textbf{alg}}^{\mathsf{T}}\hat{\theta}_{\textbf{off}}$	**Distorted value** $\mu_{x_{\textbf{alg}}}(\hat{\theta}_{\textbf{off}})$
OFUL	51.71	-2.9
WOFUL	57.86	-6.75

Figure 3: Expected value $x_{\text{alg}}^{\mathsf{T}}\hat{\theta}_{\text{off}}$ and weight-distorted value $\mu_{x_{\text{alg}}}(\hat{\theta}_{\text{off}})$ for OFUL and WOFUL algorithms on a 3x3-grid network. Here $\hat{\theta}_{\text{off}}$ is a ridge regression-based estimate of the true parameter θ (see (7)) that is obtained by running an independent simulation and x_{alg} is the route that the respective algorithm converges. x_{OFUL} is the blue-dotted route in the figure, while x_{WOFUL} includes the green-dotted detour.

and the rest of the proof uses the standard confidence ellipsoid result that ensures θ resides in C_m with high probability. A crucial observation necessary to ensure (9) is that, for any r.v. X and any $a \in \mathbb{R}$, the difference in weight-distorted cost $\mu_{X+a} - \mu_X$ is a non-linear function of a (see Lemma 5 below). Thus, it is not straightforward to compute the weight-distorted cost after translation and this poses a significant challenge in the analysis of WOFUL for the arm-dependent noise model that we consider here.

Lemma 5. *Let* $\mu_X := \int_0^\infty w(\mathbb{P}[X > z])dz - \int_0^\infty w(\mathbb{P}[-X > z])dz$*. Then, for any* $a \in \mathbb{R}$*, we have* $\mu_{X+a} = \mu_X + \int_{-a}^0 [w(\mathbb{P}[X > u]) + w(\mathbb{P}[X < u])]\,du$*. Consequently, since* w *is bounded by 1, we have* $|\mu_{X+a} - \mu_X| \leq 2|a|$*, for any* $a \in \mathbb{R}$*.*

The reader is referred to (Gopalan et al. 2016) for a detailed proof. □

Numerical Experiments

We describe two sets of experiments pertaining to K-armed and linear bandit settings, respectively. For both problems, we take the S-shaped weight function $w(p) = \frac{p^\eta}{(p^\eta+(1-p)^\eta)^{1/\eta}}$, with $\eta = 0.61$, to model perceived distortions in cost. Tversky and Kahneman observe that this distortion weight function is a good fit to explain distorted value preferences among human beings.

Experiments for K-armed Bandit

We study two stylized 2-armed bandit problems which, in part, draw upon experiments carried out by (1992) on human subjects in their studies of non-EU cumulative prospect theory. Figures 2 (a) and (b) describe each problem setting in detail (following the convention of this paper of modeling costs, "\$$x$ w.p. p" is taken to mean a loss of \$$x$ suffered with a probability of p.)

For the first problem, the weight function w gives the distorted cost of arm 1 as \$$10.55$, much higher than its expected cost of \$5 and thus more expensive due to the deterministic Arm 2 with a (distorted and expected) cost of \$7. The distortion in costs thus shifts the optimal arm from Arm 1 to Arm 2, and an online learning algorithm must be aware of this effect in order to attain low regret with respect to choosing Arm 2. A similar pattern is true for the other problem involving truly stochastic arms – weight distortion favors the arm with the higher cost in true expectation.

We benchmark the cumulative regret of two algorithms – (a) the well-known UCB algorithm (Auer, Cesa-Bianchi, and Fischer 2002), and (b) W-UCB; the results are as in Figure 2. In the experiments, UCB is not aware of the distorted weighting and hence attains linear regret with respect to playing the optimal distorted arm, due to converging to essentially a 'wrong' arm. On the other hand, the W-UCB algorithm, being designed to explicitly account for distorted cost perception, estimates the distortion using sample-based quantiles and exhibits significantly lower regret.

Experiments for Linear Bandit

We study the problem of optimizing the route choice of a human traveler using Green Light District (GLD) traffic simulation software (Wiering et al. 2004). In this setup, a source-destination pair is fixed in a given road network. Learning proceeds in an online fashion, where the algorithm chooses a route in each round, and the system provides the (stochastic) delay for the chosen route. The objective is to find the "best" path that optimizes some function of delay, while not exploring too much. While traditional algorithms minimized the expected delay, in this work, we consider the distorted value (as defined in (8)) as the performance metric.

We implement both OFUL and WOFUL algorithms for this problem. Since the weight function w is non-linear, a closed form expression for $\mu_x(\hat{\theta}_m)$ is not available and we employ the empirical distribution scheme, described for the K-armed bandit setting (see (5)), for estimating the weight-distorted value. For this purpose, we simulate 25000 samples of the Gaussian distribution, as defined in (7).

Figure 3 depicts the road network considered for our experiments and presents the expected and weight-distorted values for OFUL and WOFUL. These values are calculated using a ridge regression-based estimate of the true parameter θ, which is obtained by running an independent simulation for $100,000$ steps. This is based on the assumption of a linear (additive) relationship between the delay of a route and the delays along each of its component lanes, in steady state. As expected, OFUL (resp. WOFUL) algorithm recommends a route x_{OFUL} (resp. x_{WOFUL}) with minimum mean delay (resp. weight-distorted value). As shown in Figure 3, x_{OFUL} is the shortest path, while x_{WOFUL} involves a detour. The lower value of distorted value (delay) for the longer path preferred by WOFUL, over the shorter path preferred by OFUL, is presumably due to the fact that the two routes differ in the variance of the end-to-end delay. This leads to rare events being overestimated by the weight function, ultimately making the former path more appealing to the distortion-conscious WOFUL strategy. The reader is referred to (Gopalan et al. 2016) for details of the simulation parameters and additional results on another road network.

Conclusions and Future Work

We have designed online learning algorithms to minimize *weight-distorted* cost – a generalization of expected value – in both the standard (unstructured) k-armed bandit and the linearly parameterized bandit settings. Moving forward, it is of interest to study the general online reinforcement learning problem with weight-distorted cost metrics. Existing algorithms for expected value maximization such as UCRL (Jaksch, Ortner, and Auer 2010) and PSRL (Osband, Russo, and Van Roy 2013) could be adapted for this purpose. Other interesting directions include considering contextual versions of the cost-distorted bandit problem, and perceived distortions in the observations.

Acknowledgments This work was supported by the National Science Foundation (NSF) under Grants CMMI-1434419, CNS-1446665, and CMMI-1362303, and by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-15-10050. Aditya Gopalan was partially supported by the DST INSPIRE faculty grant IFA13-ENG-69.

References

Abbasi-Yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2312–2320.

Allais, M. 1953. Le comportement de l'homme rationel devant le risque: Critique des postulats et axioms de l'ecole americaine. *Econometrica* 21:503–546.

Audibert, J.-Y.; Munos, R.; and Szepesvári, C. 2009. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.* 410(19):1876–1902.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2003. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.* 32(1):48–77.

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

Camerer, C. F. 1989. An experimental test of several generalized utility theories. *Journal of Risk and Uncertainty* 2(1):61–104.

Camerer, C. F. 1992. Recent tests of generalizations of expected utility theory. In *Utility Theories: Measurements and Applications*. Springer. 207–251.

Cherny, A., and Madan, D. 2009. New measures for performance evaluation. *Review of Financial Studies* 22(7):2571–2606.

Conlisk, J. 1989. Three variants on the Allais example. *The American Economic Review* 392–407.

Dani, V.; Hayes, T. P.; and Kakade, S. M. 2008. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, 355–366.

Dani, V.; Kakade, S. M.; and Hayes, T. P. 2007. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems 20*. MIT Press. 345–352.

Gonzalez, R., and Wu, G. 1999. On the shape of the probability weighting function. *Cognitive Psychology* 38(1):129–166.

Gopalan, A.; Prashanth, L.; Fu, M.; and Marcus, S. 2016. Weighted bandits or: How bandits learn distorted values that are not expected. *arXiv preprint arXiv:1611.10283*.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.* 11:1563–1600.

Osband, I.; Russo, D.; and Van Roy, B. 2013. (More) Efficient reinforcement learning via posterior sampling. In *Proc. Neural Information Processing Systems*. Curran Associates, Inc. 3003–3011.

Prashanth, L.; Cheng, J.; Fu, M.; Marcus, S.; and Szepesvári, C. 2016. Cumulative prospect theory meets reinforcement learning: prediction and control. In *Proceedings of the 33rd International Conference on Machine Learning*, 1406–1415.

Prelec, D. 1998. The probability weighting function. *Econometrica* 497–527.

Quiggin, J. 2012. *Generalized Expected Utility Theory: The Rank-dependent Model*. Springer Science & Business Media.

Starmer, C. 2000. Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 332–382.

Tversky, A., and Kahneman, D. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5(4):297–323.

Wasserman, L. A. 2015. *All of Nonparametric Statistics*. Springer.

Wiering, M.; Vreeken, J.; van Veenen, J.; and Koopman, A. 2004. Simulation and optimization of traffic in a city. In *IEEE Intelligent Vehicles Symposium*, 453–458.

Wu, G., and Gonzalez, R. 1996. Curvature of the probability weighting function. *Management Science* 42(12):1676–1690.