Continuous-Armed Bandit under Mixed Smoothness: Reduction to MAB and Regret Analysis

1 Reduction to Multi-Armed Bandit via DMS Discretization

1.1 Mixed Smoothness Setup

We consider a stochastic bandit problem with a continuous d-dimensional action space (arms)

$$A \subseteq [0,1]^d.$$

For simplicity in the approximation-theoretic statements below, we present results on $[0,1]^d$; one may take $A = [0,1]^d$ (or assume f extends to $[0,1]^d$ and restrict actions to A).
We assume the unknown reward function $f : A \to \mathbb{R}$ has *dominating mixed smoothness of order 1*, i.e., all mixed weak derivatives $D^\alpha f$ with $\alpha \in \{0,1\}^d$ exist and are uniformly bounded (see Assumption 1). In particular, this implies a Lipschitz-type bound in the ℓ_1 metric. .

This structural assumption implies that f can be approximated with sparse grids or hyperbolic cross basis functions with dimension-independent accuracy. In particular, from approximation theory we have that for any budget of N interpolation nodes, one can construct an approximant $\tilde{f}_N$ such that the uniform error is bounded by[2]:

$$\sup_{x \in [0,1]^d} |f(x) - \tilde{f}_N(x)| \le C\, N^{-1} (\ln N)^\beta, \tag{1}$$

where $C > 0$ is a constant and β depends on d (often $\beta = d - 1$ for standard sparse grids)[2].

Key observation: The convergence rate N^{-1} is *independent of* d, in stark contrast to naive grids where error $O(N^{-1/d})$ deteriorates with dimension.

In other words, to achieve approximation error ε, one needs only

$$N = O\left(\varepsilon^{-1} \left(\ln \frac{1}{\varepsilon}\right)^\beta\right) \text{ points}^2,$$

whereas an isotropic Lipschitz function would require $N = O(\varepsilon^{-d})$ points (exponential in d)[4].

This mixed-smoothness property is the key to avoiding the curse of dimensionality in our bandit algorithm[2].

1.2 Discretizing the Action Space

Given this approximation result, a natural approach is to discretize the continuous action space using sparse-grid interpolation nodes.

Let

$$X_N = \{x_1, \ldots, x_N\}$$

be the set of N carefully chosen grid points (e.g., Smolyak sparse-grid nodes) such that the above approximation bound holds with $\tilde{f}_N$ constructed from samples on X_N.

Strategy: We restrict the agent to choosing actions only from this finite set X_N. Effectively, we reduce the continuous bandit to a discrete N-armed bandit problem (each grid node x_i is now treated as an arm).

Intuition: The existence of a uniform approximation bound

$$\|f - \tilde{f}_N\|_{L^\infty([0,1]^d)} \leq \varepsilon_N$$

does *not* by itself imply that the best grid point is ε_N-optimal for f. Restricting actions to X_N induces a discretization bias that must be treated explicitly.

Let

$$x^* = \arg\max_{x \in A} f(x), \qquad x^+ \in \arg\max_{x_i \in X_N} f(x_i),$$

and define the discretization bias

$$b_N := f(x^*) - \max_{x_i \in X_N} f(x_i) = f(x^*) - f(x^+).$$

The quantity $b_N \geq 0$ measures the loss incurred by restricting the action space to X_N. At this stage we do not assume any specific relation between b_N and the approximation error ε_N.

1.3 Applying UCB on the Discrete Grid

Having reduced the action space to N arms, we can now apply standard multi-armed bandit algorithms. In particular, we choose an *Upper Confidence Bound* strategy (specifically UCB1, as in Auer et al., or a variant thereof) to handle the exploration-exploitation trade-off on the discrete set X_N.

The UCB algorithm treats each grid node x_i as an independent arm with unknown mean $f(x_i)$ and draws samples (obtaining noisy rewards) to gradually concentrate estimates of each arm's value.

UCB will ensure that, with high probability, it identifies the best arm in X_N up to an uncertainty of order

$$O\left(\sqrt{\frac{\ln T}{n_i}}\right)$$

after n_i pulls of arm i.

The objective of the bandit algorithm is therefore to identify the best arm x^+ within X_N; the effect of discretization on the continuous optimum is entirely captured by the bias term b_N defined above.

1.4 Trade-off: Approximation vs. Bandit Regret

The discretization introduces a fundamental trade-off between discretization bias and bandit learning regret:

- **Finer grid (large N):** Potentially reduces the discretization bias b_N, but increases the number of arms N and thus slows down learning.
- **Coarser grid (small N):** Easier to learn over, but may induce a larger bias b_N.

Our strategy: Choose N as a function of the total time horizon T to balance these two sources of regret. We will analyze this balance formally in the next section.

Intuition: In high dimensions, DMS allows a much sparser discretization than a full grid. For example, a sparse grid may use

$$N = O(n(\ln n)^{d-1}) \text{ points}$$

to achieve the resolution of a n^d full grid[2]. This dramatic reduction in arm count will permit sublinear regret independent of d in the exponent.

In contrast, a naive uniform discretization of $[0,1]^d$ (or assumption of only Lipschitz continuity in a metric space) leads to regret bounds that deteriorate quickly with d. Indeed, the minimax regret for Lipschitz bandits is

$$R_T = \Theta\left(T^{\frac{d+1}{d+2}}\right),$$

which approaches linear regret as d grows large (a manifestation of the curse of dimensionality).

Summary: By leveraging mixed smoothness, we reduce the continuum-armed bandit to a discrete N-armed bandit problem using the interpolation nodes X_N. The regret analysis will account for a bias term b_N per round due to discretization, and a standard multi-armed bandit regret term for learning the best arm within X_N.

2 Regret Analysis for Bandit with Mixed Smoothness

2.1 Problem Setup and Assumptions

We assume the reward function f is in the DMS class W^1_{mix} (as above), and that the reward observations are noisy but sub-Gaussian with mean $f(x)$.

Reward model: When the agent plays action $x_t \in A$ at time t, it receives reward

$$r_t = f(x_t) + \eta_t,$$

where η_t is zero-mean R-sub-Gaussian noise.

Cumulative regret: We analyze the expected cumulative regret after T rounds, defined as:

$$R(T) = T \cdot f(x^*) - \mathbb{E}\left[\sum_{t=1}^{T} f(x_t)\right], \tag{2}$$

where $x^* = \arg\max_{x \in A} f(x)$ is the true optimal action in the continuum.

We will upper-bound $R(T)$ in terms of T, N, and the discretization bias b_N.

2.2 Regret Decomposition

We decompose the regret into two parts:

1. **Discretization bias** from restricting to X_N
2. **Learning regret** (not immediately knowing which arm in X_N is best)

Let x^+ be the optimal arm in the discrete set X_N. Then:

$$R(T) = T\big(f(x^*) - f(x^+)\big) + \mathbb{E}\left[\sum_{t=1}^{T} \big(f(x^+) - f(x_t)\big)\right] = Tb_N + R_{\text{MAB}}(T).$$

The first term Tb_N is the cumulative discretization bias incurred by restricting actions to X_N. The second term $R_{\text{MAB}}(T)$ is the learning regret of the bandit algorithm on the N-armed discrete problem.

2.3 Bandit Algorithm Regret (UCB1)

For the N-armed sub-problem, we use the UCB1 algorithm. Recall that UCB1 guarantees an expected regret on the order of $O(\sqrt{NT \ln T})$ in the worst case.

More precisely, a classic result is that for N arms with sub-Gaussian rewards,

$$R_{\text{MAB}}(T) \leq O\left(\sqrt{N\,T\,\ln T}\right), \tag{3}$$

up to additive constants.

For our analysis we use a bound of form

$$R_{\text{MAB}}(T) \leq c\sqrt{NT \ln T}$$

for some constant c.

2.4 Regret Bound Derivation

Combining the two components, we obtain

$$R(T) \le T b_N + c\sqrt{N\,T\,\ln T}.$$

To optimize N, we use the discretization-bias bound implied by the Smolyak construction. By Theorem 3.2,

$$b_N \le \varepsilon_N := \|f - \tilde{f}_N\|_{L^\infty([0,1]^d)}.$$

By Theorem 3.1, $\varepsilon_N \le C\,N^{-1}(\ln N)^\beta$, hence

$$R(T) \le T \cdot C\,N^{-1}(\ln N)^\beta + c\sqrt{N\,T\,\ln T}.$$

Intuition: As N increases:

- Bias term ($\propto T/N$ up to logs) *decreases*
- Learning term ($\propto \sqrt{NT}$) *increases*

Ignoring logarithmic factors for a moment, balance the two terms:

$$T/N \approx \sqrt{NT} \quad \Longrightarrow \quad N \approx T^{1/3}.$$

Precise calculation (up to log factors): Plug $N = \kappa T^{1/3}$ into the bound:

- **Bias part:**

$$T b_N \;\le\; T \cdot C(\kappa^{-1}T^{-1/3})\big(\ln(\kappa T^{1/3})\big)^\beta = O\Big(T^{2/3}(\ln T)^\beta\Big).$$

- **Bandit part:**

$$c\sqrt{NT\ln T} = c\sqrt{\kappa T^{1/3}\cdot T\ln T} = O\Big(T^{2/3}(\ln T)^{1/2}\Big).$$

Therefore,

$$R(T) = O\Big(T^{2/3}(\ln T)^{\max\{\beta,\,1/2\}}\Big) = \tilde{O}(T^{2/3}).$$

2.5 Main Theorem

Theorem 2.1 (Regret Bound for DMS Bandits)**.** *Let f satisfy Assumption 1 (dominating mixed smoothness). Under sub-Gaussian noise, construct a Smolyak sparse-grid discretization X_N with N nodes and run UCB on the resulting N-armed bandit.*

Then there exists a choice of

$$N = \Theta\left(T^{1/3}\right)$$

(up to logarithmic factors in T and d) such that the expected cumulative regret satisfies

$$R(T) = O\Big(T^{2/3}(\ln T)^{\max\{\beta,\,1/2\}}\Big) = \tilde{O}(T^{2/3}),$$

where $\beta = \beta(d)$ depends on the dimension (typically $\beta = d-1$ for standard Smolyak grids).

In particular, the regret exponent $2/3$ does not depend on d (dimension enters only through logarithmic factors).

Proof sketch. By Theorem 3.1, the uniform approximation error is $\varepsilon_N = CN^{-1}(\ln N)^\beta$. By Theorem 3.2, the discretization bias satisfies $b_N \le \varepsilon_N = O(N^{-1}(\ln N)^\beta)$. By Lemma 3.4, the MAB regret is $R_{\mathrm{MAB}}(T) = O(\sqrt{NT\ln T})$.

From the regret decomposition (Section 2.2),

$$R(T) = Tb_N + R_{\mathrm{MAB}}(T) \le C'TN^{-1}(\ln N)^\beta + c\sqrt{NT\ln T}.$$

Choosing $N = \kappa T^{1/3}$ (with κ chosen to balance the two terms) yields both terms of order $O(T^{2/3}(\ln T)^{\max\{\beta,1/2\}})$, completing the proof. □

Notation: Here $\tilde{O}(\cdot)$ hides polylogarithmic factors.

2.6 Comparison to Lipschitz Bandits

For context, in a d-dimensional Lipschitz bandit (no mixed smoothness), the optimal regret scales as

$$R(T) \sim \Theta\left(T^{\frac{d+1}{d+2}}\right).$$

For example:

- $d = 1$ yields $T^{2/3}$ (as in the classic continuum-armed bandit result by Kleinberg (2005))
- $d = 2$ gives $T^{3/4}$
- As $d \to \infty$, the exponent $(d+1)/(d+2) \to 1$, meaning regret becomes nearly linear

Our result for DMS can be viewed as achieving an *effective dimension of 1* (up to log factors) regardless of the ambient d.

Importance of sparse grid discretization: The sparse grid discretization is crucial. If we had instead discretized the space naively (e.g., picked an ε-net in Euclidean distance), we would require on the order of $(1/\varepsilon)^d$ points to cover $[0,1]^d$. Planning a priori for horizon T, one might set ε such that

$$T\varepsilon \sim \sqrt{(1/\varepsilon)^d\, T},$$

leading to regret $T^{(d+1)/(d+2)}$ – reproducing the curse of dimensionality.

The mixed smoothness assumption lets us use structured sets of size

$$N \sim (1/\varepsilon)(\ln(1/\varepsilon))^{d-1}$$

to achieve uniform approximation error ε, a drastically smaller arm set.

2.7 High-Probability Guarantees

The UCB algorithm's guarantee can also be stated in high-probability terms. With probability at least $1 - \delta$, one can show

$$R_{\mathrm{MAB}}(T) = O(\sqrt{NT \ln(T/\delta)}).$$

We omit these details, focusing on expected regret.

2.8 Adaptation to Unknown Horizon

Finally, we note that our analysis assumed knowledge of the time horizon T to choose N. In practice, one can:

- use a doubling trick over epochs and increase N over time,
- or tune N on the fly.

This adds only lower-order terms.

3 Theoretical Steps and Key Lemmas

We now outline the key steps and assumptions in a more formal manner, which also sets the stage for extending these results to reinforcement learning later.

3.1 Mixed Smoothness Function Class

Assumption 1 (Dominating mixed smoothness of order 1)**.** *There exists $L < \infty$ such that for every multi-index $\alpha \in \{0,1\}^d$, the mixed weak derivative $D^\alpha f$ exists and satisfies*

$$\|D^\alpha f\|_{L^\infty([0,1]^d)} \le L.$$

Equivalently, $f \in W^1_{\mathrm{mix}}([0,1]^d)$ (dominating mixed smoothness $r = 1$).

This implies a form of Lipschitz condition: for any x, y,

$$|f(x) - f(y)| \le L \sum_{j=1}^{d} |x_j - y_j|.$$

3.2 Smolyak Sparse Grid Approximation and Uniform Error

Theorem 3.1 (Smolyak (Sparse Grid) Approximation Rate)**.** *For f satisfying Assumption 1, for any $N \in \mathbb{N}$, there exists a Smolyak sparse-grid recovery operator (based on N nodes) producing an approximant $\tilde{f}_N$ such that*

$$\|f - \tilde{f}_N\|_{L^\infty([0,1]^d)} \le C\, N^{-1} (\ln N)^\beta,$$

for some constants $C > 0$ and $\beta = \beta(d)$[2]. In particular, an ε-uniform approximation can be achieved with

$$N = O\left(\varepsilon^{-1} \left(\ln \frac{1}{\varepsilon}\right)^\beta\right) \text{ nodes}^2.$$

Concrete construction: For concreteness, one may take $\tilde{f}_N$ to be the Smolyak sparse grid interpolant/sampling-recovery operator built from nested 1D rules (e.g., Clenshaw–Curtis) and combined by the Smolyak formula.

Remark on bias vs. approximation error: The uniform approximation bound in Theorem 3.1 controls $\|f - \tilde{f}_N\|_\infty$. For the multilinear Smolyak interpolants used here, Theorem 3.2 establishes that the discretization bias satisfies $b_N \le \varepsilon_N$, a tight bound that is crucial for the regret analysis.

3.3 Discretization Bias for Multilinear Sparse Grids

The following result establishes the link between the uniform approximation error and the discretization bias when the Smolyak recovery is piecewise multilinear on a partition whose vertices coincide with the grid nodes.

Theorem 3.2 (Sparse-Grid Discretization Bias)**.** *Let f satisfy Assumption 1. Let $X_N = \{x_1, \dots, x_N\}$ be the node set of a Smolyak sparse-grid construction based on tensor-product* piecewise-linear *1D rules. Assume the resulting recovery $\tilde{f}_N$ is continuous and* piecewise multilinear *on a partition $\mathcal{T}_N$ of $[0,1]^d$ such that every cell $Q \in \mathcal{T}_N$ has all its vertices in X_N. If*

$$\|f - \tilde{f}_N\|_{L^\infty([0,1]^d)} \le \varepsilon_N,$$

then the discretization bias satisfies

$$b_N := f(x^*) - \max_{i=1,\dots,N} f(x_i) \le \varepsilon_N.$$

Proof. Fix any cell $Q \in \mathcal{T}_N$. Since $\tilde{f}_N$ is multilinear on Q, its maximum over Q is attained at a vertex of Q. By assumption, all vertices of Q belong to X_N, hence

$$\max_{x \in Q} \tilde{f}_N(x) = \max_{v \in \mathrm{vert}(Q)} \tilde{f}_N(v) \leq \max_{i=1,\dots,N} \tilde{f}_N(x_i).$$

Taking the maximum over all cells yields

$$\max_{x \in [0,1]^d} \tilde{f}_N(x) = \max_{i=1,\dots,N} \tilde{f}_N(x_i).$$

Moreover, $\tilde{f}_N$ interpolates f on X_N, so $\tilde{f}_N(x_i) = f(x_i)$ for all i and therefore

$$\max_{x \in [0,1]^d} \tilde{f}_N(x) = \max_{i=1,\dots,N} f(x_i).$$

Finally, by the uniform approximation bound,

$$f(x^*) \leq \tilde{f}_N(x^*) + \varepsilon_N \leq \max_{x \in [0,1]^d} \tilde{f}_N(x) + \varepsilon_N = \max_{i=1,\dots,N} f(x_i) + \varepsilon_N,$$

which implies $b_N \leq \varepsilon_N$. □

Remark 3.1. *The only nontrivial structural requirement is that $\tilde{f}_N$ is piecewise multilinear on a partition whose vertices are contained in the node set X_N (so that the maximum of $\tilde{f}_N$ is attained at a node). If one uses a construction where the induced partition has additional vertices not in X_N, then the identity $\max_x \tilde{f}_N(x) = \max_i \tilde{f}_N(x_i)$ may fail and the above argument must be adapted (e.g., by enlarging X_N to include the partition vertices).*

3.4 Optional: Surrogate-based continuous recommendation

Lemma 3.3 (Surrogate maximizer is near-optimal)**.** *If $\|f - \tilde{f}_N\|_{L^\infty([0,1]^d)} \leq \varepsilon_N$ and $\hat{x} = \arg\max_{x \in [0,1]^d} \tilde{f}_N(x)$, then*

$$f(x^*) - f(\hat{x}) \leq 2\varepsilon_N.$$

Proof. Since $\tilde{f}_N(\hat{x}) \geq \tilde{f}_N(x^*)$, we have $\tilde{f}_N(x^*) - \tilde{f}_N(\hat{x}) \leq 0$. Thus

$$f(x^*) - f(\hat{x}) \leq [f(x^*) - \tilde{f}_N(x^*)] + [\tilde{f}_N(\hat{x}) - f(\hat{x})] \leq \varepsilon_N + \varepsilon_N = 2\varepsilon_N.$$

□

Remark 3.2. *Lemma 3.3 is useful if one considers a* continuous *recommendation $\hat{x}$ obtained by maximizing the surrogate $\tilde{f}_N$. In the cumulative-regret setting analyzed above, actions are restricted to X_N and the corresponding discretization loss is captured by b_N.*

3.5 Detailed Regret Decomposition

A straightforward lemma is that the discrete-bandit learning term equals the sum of suboptimality gaps times the expected number of pulls:

$$R_{\mathrm{MAB}}(T) = \sum_{i:\Delta_i > 0} \Delta_i \, \mathbb{E}[n_i(T)],$$

where $\Delta_i = f(x^+) - f(x_i)$ and $n_i(T)$ is the number of times arm i was played.

Standard UCB analysis gives

$$\mathbb{E}[n_i(T)] = O\left(\frac{\ln T}{\Delta_i^2}\right)$$

for each suboptimal arm, leading to

$$R_{\mathrm{MAB}}(T) = O\left(\sum_{i:\Delta_i > 0} \frac{\ln T}{\Delta_i}\right).$$

In the worst case (adversarially small gaps), this yields the gap-free bound $O(\sqrt{NT \ln T})$ as used above.

3.6 UCB Regret Bound (Formal Statement)

Lemma 3.4 (UCB Regret Bound)**.** *For N arms with sub-Gaussian (σ^2-sub-Gaussian) noise, UCB1 with appropriate confidence parameter $\delta_t = 1/t^2$ achieves*

$$R_{MAB}(T) \leq 8\sqrt{NT \ln T} + O(N).$$

Note: The $O(N)$ term comes from an additive constant bounded by N times the reward range; it is negligible for sublinear growth.

We now combine Lemma 3.4 with the decomposition $R(T) = Tb_N + R_{\mathrm{MAB}}(T)$. Optimizing the choice of N is carried out in Section 2 using Theorem 3.2 (and yields $N \asymp T^{1/3}$ up to logarithmic factors).

4 Relation to GLB

Generalized linear bandits have a reward functions that writes as follows

$$f(x) = \sigma\left(\sum_{i=1}^{d} \theta_i \phi(x_i)\right) \qquad \sigma(y) = \frac{\exp(y)}{1+\exp(y)},$$

(or some other $\sigma : \mathbb{R} \to \mathbb{R}$, often restricted to $0, 1$). This mimics a single-laywer neural network. Crucially, the features have often only one bounded derivative. For example for $\phi = \max\{0, 1\}$ (the RELU activations)

$$\phi'(x) = 1_{[0,+\infty]}(x), \qquad \|\phi'\|_\infty = 1, \|\phi''\|_\infty = +\infty.$$

The derivarives of the former are given as follows:

$$\begin{aligned}\frac{\partial f}{\partial x_i}(x) &= \theta_i \phi'(x_i)\sigma'(x^\top\theta) \\ &= \theta_i \phi'(x_i)(1-\sigma(x^\top\theta))\sigma(x^\top\theta).\end{aligned}$$

Thus, any mixed derivative is bounded:

$$\begin{aligned}\frac{\partial^2 f}{\partial x_i \partial x_j}(x) &= \frac{\partial}{\partial x_j}\theta_i\phi'(x_i)(1-\sigma(x^\top\theta))\sigma(x^\top\theta) \\ &= \theta_i\phi'(x_i)\frac{\partial}{\partial x_j}(1-\sigma(x^\top\theta))\sigma(x^\top\theta) \\ &= \theta_i\phi'(x_i)\frac{\partial}{\partial x_j}(\sigma(x^\top\theta)-\sigma(x^\top\theta)^2) \\ &= \theta_i\phi'(x_i)\Big(\theta_j\phi'(x_j)(1-\sigma(x^\top\theta))\sigma(x^\top\theta) - 2\theta_j\phi'(x_j)(1-\sigma(x^\top\theta))\sigma(x^\top\theta)\sigma(x^\top\theta)\Big) \\ &= \theta_i\phi'(x_i)\Big(\theta_j\phi'(x_j)(1-\sigma(x^\top\theta))\sigma(x^\top\theta)\big(1-2\sigma(x^\top\theta)\big)\Big) \\ &= \theta_i\phi'(x_i)\theta_j\phi'(x_j)(1-\sigma(x^\top\theta))\sigma(x^\top\theta)\big(1-2\sigma(x^\top\theta)\big) \\ &\leq |\theta_i||\theta_j|\|\phi'\|_\infty^2,\end{aligned}$$

(which keeps being small if we assume $\|\theta\|_\infty \leq 1$ and $\|\phi'\|_\infty \leq 1$) while taking two derivatives w.r.t. the same index leads to a $\|\phi''\|_\infty$, that is unbounded.

5 Conclusion

We have developed a continuum-armed bandit analysis leveraging mixed smoothness and Smolyak sparse grids. By discretizing actions to sparse grid nodes and employing UCB, we obtain a regret bound of the form

$$R(T) \leq Tb_N + O(\sqrt{NT \ln T}).$$

Using Theorem 3.2, which establishes $b_N \leq \varepsilon_N$ for piecewise-multilinear Smolyak interpolants, together with the approximation rate $\varepsilon_N = O(N^{-1}(\ln N)^{\beta})$ from Theorem 3.1, we obtain

$$R(T) = \tilde{O}(T^{2/3})$$

with no exponential dependence on dimension.

These steps give a foundation to connect approximation structure (via mixed smoothness and Smolyak recovery) with bandit learning on a finite action subset.

References

1. Smolyak, S. A. (1963). Quadrature and interpolation formulas for tensor products of certain classes of functions. *Soviet Mathematics, Doklady*, 4, 240–243.

2. Bungartz, H.-J., & Griebel, M. (2004). Sparse grids. *Acta Numerica*, 13, 147–269.

3. Wasilkowski, G. W., & Woźniakowski, H. (1995). Explicit cost bounds of algorithms for multivariate tensor product problems. *Journal of Complexity*, 11(1), 1–56.

4. Temlyakov, V. N. (2018). *Multivariate Approximation.* Cambridge University Press.

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

6. Kleinberg, R. D. (2005). Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems*, 17, 697–704.

7. Bubeck, S., Munos, R., Stoltz, G., & Szepesvári, C. (2011). X-armed bandits. *Journal of Machine Learning Research*, 12, 1655-1655-1695.