Risk-aware linear bandits with convex loss

Patrick Saux patrick.saux@inria.fr
Scool team
Inria, Univ. Lille, CNRS, Centrale Lille, UMR 9198-CRIStAL
F-59000 Lille, France

Odalric-Ambrym Maillard odalric.maillard@inria.fr
Scool team
Inria, Univ. Lille, CNRS, Centrale Lille, UMR 9198-CRIStAL
F-59000 Lille, France

Abstract

In decision-making problems such as the multi-armed bandit, an agent learns sequentially by optimizing a certain feedback. While the mean reward criterion has been extensively studied, other measures that reflect an aversion to adverse outcomes, such as mean-variance or conditional value-at-risk (CVaR), can be of interest for critical applications (healthcare, agriculture). Algorithms have been proposed for such risk-aware measures under bandit feedback without contextual information. In this work, we study contextual bandits where such risk measures can be elicited as linear functions of the contexts through the minimization of a convex loss. A typical example that fits within this framework is the expectile measure, which is obtained as the solution of an asymmetric least-square problem. Using the method of mixtures for supermartingales, we derive confidence sequences for the estimation of such risk measures. We then propose an optimistic UCB algorithm to learn optimal risk-aware actions, with regret guarantees similar to those of generalized linear bandits. This approach requires solving a convex problem at each round of the algorithm, which we can relax by allowing only approximated solution obtained by online gradient descent, at the cost of slightly higher regret. We conclude by evaluating the resulting algorithms on numerical experiments.

Keywords: Contextual bandit; UCB; risk-aware measures; elicitable risk measures; online gradient descent.

1. Intro

Contextual bandits are sequential decision-making models where at each time step an agent observes a set of possible actions, or contexts, plays one of them and receives a stochastic reward, the distribution of which is a function of the selected action. The goal of the agent is to learn an action policy in order to maximize the sum of rewards, facing the classical exploitation-exploration dilemma. A prominent example of such models is the linear bandit, which assumes a linear relationship between the action and the *mean* of the reward distribution. In this setting, a standard learning strategy consists in estimating the linear reward model by ridge regression coupled with an appropriate exploration scheme, such as the optimistic LinUCB (Abbasi-Yadkori et al., 2011), Thompson sampling (Agrawal and Goyal, 2013; Abeille and Lazaric, 2017) or information-directed sampling (Russo and Van Roy, 2014; Kirschner et al., 2021).

One limitation of this setting is that real-world agents may have a preference for other reward criterion than the mean. Indeed, while mathematically convenient, the mean is known to equally weights large positive and negative outcomes, which may lead to risky policies unsuitable to critical applications, and is also sensitive to outliers. In contrast, *risk-aware* measures emphasize different characteristics of the reward distribution, e.g by stressing out the impact of adverse outcomes (Dowd, 2007). Such measures include the mean-variance (Markowitz, 1952), conditional Value-at-Risk (Rockafellar et al., 2000), entropic risk (Maillard, 2013) and the expectiles (Newey and Powell, 1987). These risk measures, and in particular the conditional Value-at-Risk (CVaR), have been studied as alternatives to the mean criterion in classical multi-armed bandits, that is without contextual information (Galichet et al., 2013; Cassel et al., 2018; Tamkin et al., 2019; Prashanth et al., 2020; Baudry et al., 2021). In distributional reinforcement learning, quantile regression has been studied for DQN (Dabney et al., 2017). Recently, bandits with contextual mean-variance and CVaR have been applied to vehicular communication Wirth et al. (2022). Despite promising empirical results, these contributions are largely devoid of theoretical regret guarantees.

In this work, we investigate an extension of the linear bandit where a given risk measure, rather than the mean, is linearly parametrized by the chosen actions. Specifically, we consider the case of convex risk measures which can be elicited as minimizers of certain loss functions, which naturally extends the standard ridge regression. This definiton covers quantiles, expectiles and entropic risk, and can be extended to mean-variance and conditional value-at-risk using multivariate risk measures. To the best of our knowledge, this setting is new, although related to existing approaches, in particular bandits with regression oracles (Foster and Rakhlin, 2020), which contrary to the present work considers finite action sets, and generalized linear bandits (GLB) (Filippi et al., 2010; Li et al., 2017; Faury et al., 2020).

In contrast with the standard mean-linear bandit, learning the linear mapping between actions and risk measures cannot be performed sequentially and presents similar theoretical challenges to GLB. We introduce a variant of LinUCB designed to maximize the risk measure of rewards, with sublinear regret guarantees. To do so, we derive time-uniform confidence sets based on the method of mixtures (Peña et al., 2008) and introduce a condition on the curvature of the convex loss. To mitigate the numerical burden of this algorithm, which requires to solve a convex minimization problem at each time step, we also introduce an episodic version with online gradient descent (OGD) approximation, directly inspired by the recent literature on approximate Thompson sampling for GLB (Ding et al., 2021).

2. Contextual bandits with convex risk

We consider the standard contextual bandit setting where an agent sequentially observes at time $t \in \mathbb{N}$ a decision set $\mathcal{X}_t \subseteq \mathbb{R}^d$, then chooses an action $X_t \in \mathcal{X}_t$ and receives a stochastic reward Y_t, the distribution of which is dependent on X_t. More formally, let $\mathcal{X} = \cup_{t\in\mathbb{N}}\mathcal{X}_t$ and $\Phi\colon \mathcal{X} \to \mathcal{P}(\mathbb{R})$ a mapping from actions to reward distributions, so that the agent receives at time t the reward $Y_t \sim \Phi(X_t)$. We denote by $\mathcal{F}_t = \sigma\left(\mathcal{X}_1, X_1, Y_1, \ldots, \mathcal{X}_{t-1}, X_{t-1}, Y_{t-1}, \mathcal{X}_t, X_t\right)$ the σ-algebra corresponding to the information available to the agent at time t (that is after choosing the action X_t but before observing the reward Y_t). Loosely speaking, the goal of the agent is to learn a representation of the mapping Φ in order to select actions that induce high rewards. A standard inductive bias in this context is to assume a linear relation between rewards and contexts, typically of the form $Y_t = \langle \theta^*, X_t\rangle + \eta_t$, where η is a stochastic noise process. In this work, we consider a slightly more general notion of linearity by assuming instead the existence of a factorization

$$\mathcal{X} \xrightarrow{\ \ell^*\ } \mathbb{R}^p \xrightarrow{\ \varphi\ } \mathcal{P}(\mathbb{R}), \qquad \Phi = \varphi \circ \ell^* \colon \mathcal{X} \to \mathcal{P}(\mathbb{R})$$

where $\ell^*\colon \mathcal{X} \to \mathbb{R}^p$ denotes a linear map. In other words, the reward distribution (but not necessarily its mean) is linearly parametrized by the chosen action. We denote such a linear bandit by (φ, ℓ^*). When the distribution depends on a single parameter ($p = 1$), we represent the linear form by $\ell^*\colon x \in \mathcal{X} \mapsto \langle \theta^*, x\rangle$, where $\theta^* \in \mathbb{R}^d$, and we use the notation (φ, θ^*), or equivalently we say that it is represented by $Y \sim \varphi(\langle \theta^*, X\rangle)$. In the following, we also denote by $\Theta \subseteq \mathbb{R}^d$ a subset of possible values for the parameter θ^* (and assume for convenience that θ^* is in the interior of Θ).

As an example, let us consider the following Gaussian mapping $\Phi\colon x \in \mathcal{X} \mapsto \mu(x) + \sigma(x)\mathcal{N}(0, 1)$. If σ is constant and $\mu(x) = \langle \theta_\mu, x\rangle$, we recover a standard linear bandit model, in which the goal is to maximize the cumulative average rewards $\sum_{t=1}^T \mu(X_t)$. However, in many applications, the agent may exhibit a preference for actions leading to lower volatility of rewards or higher 5% quantile, in order to mitigate extreme events. This risk aversion can be encoded for instance by $\mu(X_t) - 1.64\sigma(X_t)$, where -1.64 is approximately the 5% quantile of the standard Gaussian.

2.1 Overview of risk measures

Convex loss In the bandit setting, the agent faces the classical dilemma between exploitation (playing the most promising actions) and exploration (playing other actions to gain information). For most contextual bandit algorithms, the exploitation takes the form of supervised estimation that consists in learning the mapping φ at time t from the past observations $\{(X_s, Y_s),\ 1 \leqslant s \leqslant t-1\}$. When the expected reward is parametrized by $Y_t = \langle \theta^*, X_t\rangle + \eta_t$ with $\mathbb{E}[\eta_t|\mathcal{F}_t] = 0$, a standard strategy consists in estimating θ^* by ridge regression $\min_{\theta\in\Theta} \sum_{s=1}^{t-1} \left(Y_s - \langle \theta, X_s\rangle\right)^2 + \frac{\alpha}{2}\|\theta\|_2^2$, where $\alpha > 0$ is a regularization parameter. Assuming for now the solution is in the interior of Θ, the solution

Table 1: Example of risk measures elicited by convex potentials.

Name	Potential $\psi(z)$	Risk measure $\rho_\psi(\nu)$
Mean	$z^2/2$	$\rho_\psi(\nu) = \int y\nu(dy)$
Quantile, $p \in (0,1)$	$(p - \mathbb{I}_{z<0})z$	$\int_{-\infty}^{\rho_\psi(\nu)} \nu(dy) = p$
Expectile, $p \in (0,1)$	$\lvert p - \mathbb{I}_{z<0}\rvert z^2$	$(1-p)\int_{-\infty}^{\rho_\psi(\nu)} \lvert y - \rho_\psi(\nu)\rvert \nu(dy) = p\int_{\rho_\psi(\nu)}^{\infty} \lvert y - \rho_\psi(\nu)\rvert \nu(dy)$

can be written as $\widehat{\theta}_t = (V_t^\alpha)^{-1}\sum_{s=1}^{t-1} Y_s X_s$, where $V_t^\alpha := \sum_{s=1}^{t-1} X_s X_s^\top + \alpha I_d$ is a $d \times d$ positive definite matrix. This method presents several advantages: it can be computed efficiently via sequential matrix inversion (with complexity $\mathcal{O}(d^2)$ at each step thanks to the Sherman-Morrison formula for the rank-one update $V_{t+1} = V_t + X_t X_t^\top$ starting from $V_0 = \alpha I_d$) and explicit confidence ellipsoids for θ^* can be constructed analytically around $\widehat{\theta}_t$ to tune exploration (Abbasi-Yadkori et al., 2011). The implicit limitation of this procedure is that it can only estimate the expectation $\mathbb{E}[Y] = \operatorname{argmin}_{\theta\in\Theta} \mathbb{E}[(Y - \langle\theta, X\rangle)^2]$. We call this setting the *mean-linear* bandit.

As motivated by the example above, we aim to estimate other statistics than the mean of the reward distribution. Drawing inspiration from this simple case, we consider an arbitrary convex loss function $\mathcal{L}\colon \mathbb{R}\times\mathbb{R}^p \to \mathbb{R}_+$ and define the *risk measure* associated with loss $\mathcal{L}$ for a distribution ν over $\mathbb{R}$ as $\rho_\mathcal{L}(\nu) = \arg\min_{\xi\in\mathbb{R}^p} \mathbb{E}_{Y\sim\nu}[\mathcal{L}(Y,\xi)]$. We assume here that the argmin is unique for simplicity (which is the case if $\mathcal{L}$ is strongly convex) and will sometimes use the notation $\rho_\mathcal{L}(Y)$ instead of $\rho_\mathcal{L}(\nu)$ for a random variable Y with distribution ν. Similarly, we define the conditional risk measure $\rho_\mathcal{L}(\nu|\mathcal{G}) = \arg\min_{\xi\in\mathbb{R}^p} \mathbb{E}_{Y\sim\nu}[\mathcal{L}(Y,\xi)|\mathcal{G}]$ for any event $\mathcal{G}$ with positive measure. Note that with this definition, $\rho_\mathcal{L}(\nu)$ is a vector in $\mathbb{R}^p$. When $p = 1$, we call these *scalar* risk measures. The motivation to consider vector-valued risk measures in this framework comes from the fact that not every measure of interest can be *elicited* as a scalar risk measure, which we develop in the next paragraph.

Elicitable risk measure Scalar risk measures that can be expressed as minimizers of such loss functions are known as (first-order) *elicitable* risk measures (Ziegel, 2016). Examples of such measures include the mean, the median, and more generally any quantile and expectile, which we further discuss below as special cases of risk measures associated with convex potentials. Other examples are any generalized moments $\rho(\nu) = \mathbb{E}_{Y\sim\nu}[T(Y)]$, where $T\colon \mathbb{R}\to\mathbb{R}$ is a ν-integrable mapping, and the entropic risk defined by $\rho_\mathcal{L}(\nu) = \frac{1}{\gamma}\log\mathbb{E}_{Y\sim\nu}[e^{\gamma Y}]$ (Maillard, 2013). Unfortunately, not all measures commonly encountered in the risk literature are first-order elicitable. In particular, neither the variance nor the CVaR can be expressed as scalar risk measures with respect to a convex loss (Fissler et al., 2015; Fissler and Ziegel, 2016). However, they are second-order elicitable, in the sense that the pairs (mean, variance) and (VaR, CVaR) are jointly elicitable. We refer to Appendix A for a summary and further interpretation of elicitable risk measures.

We say that the loss $\mathcal{L}$ is adapted to the linear bandit (φ, ℓ^*) if for all $x \in \mathcal{X}$, ℓ^* is a minimizer of $\mathbb{E}[\mathcal{L}(Y, \ell(x))]$ among all linear forms $\ell\colon \mathcal{X}\to\mathbb{R}^p$, where $Y \sim \varphi\circ\ell^*(x)$ denotes the reward random variable of the linear bandit when action x is played. Intuitively, this means that the risk measure $\rho_\mathcal{L}$ of the reward distribution is linearly parametrized by the actions, the same way the expected reward is a linear form of the action in the standard mean-linear bandit.

Remark 1 *The above definition is written in the general case of a vector-valued risk measure $\rho_\mathcal{L}$. In the rest of this paper, we only consider scalar risk measure and leave the extension to measures like CVaR for further work. We say that $\mathcal{L}$ is adapted to the linear bandit (φ, θ^*) if for all $x \in \mathcal{X}$, we have that $\theta^* \in \arg\min_{\theta\in\Theta} \mathbb{E}_{Y\sim\varphi(\langle\theta^*, x\rangle)}[\mathcal{L}(Y, \langle\theta, x\rangle)]$.*

Convex potential A special case of interest is when the convex loss $\mathcal{L} = \mathcal{L}_\psi$ derives from a potential ψ, that is when $\mathcal{L}_\psi(y,\xi) = \psi(y-\xi)$. This includes the ordinary least square potential associated to the mean, as well as quantiles and expectiles. We assume the reader to be familiar with the former, but perhaps less so with the latter. Following Newey and Powell (1987), we define the p-expectile of ν for $p \in (0,1)$ as $\operatorname{argmin}_{\xi\in\mathbb{R}} \mathbb{E}_{Y\sim\nu}[|p - \mathbb{I}_{Y<\xi}|(Y-\xi)^2]$. Expectiles have been studied in particular in the context of risk management (Bellini and Di Bernardino, 2017) and risk-aware Bayesian optimization (Torossian et al., 2020). Furthermore, under some symmetry conditions, quantiles and expectiles are known to coincide (Abdous and Remillard, 1995), and thus expectiles can be seen as a smooth (in particular differentiable) generalization of quantiles (see Philipps (2022) for further interpretation of the notion of expectiles). We refer the reader to Table 1 for a summary of risk measures elicited by convex potentials.

In the terminology defined above, the ordinary least square potential is adapted to the mean-linear reward model $Y_t = \langle \theta^*, X_t \rangle + \eta_t$ with $\mathbb{E}[\eta_t | \mathcal{F}_t] = 0$. More generally, such an additive decomposition exists for losses derived from potentials (see Proposition 17 in Appendix B).

Non-unicity of adapted loss In general, a risk measure can be described by multiple different adapted losses. First, the set of losses that elicit a given risk measure is a cone invariant by scalar translation, i.e. $\rho_{\alpha\mathcal{L}+\beta} = \rho_{\mathcal{L}}$ for all $\alpha > 0$ and $\beta \in \mathbb{R}$. Other less trivial examples of non-unicity arise even for the simple mean criterion. Indeed, Theorem 1 in Banerjee et al. (2005) shows that $\mathbb{E}_{Y\sim\nu}[Y] = \rho_{\mathcal{B}_\psi}(\nu)$ where ψ is any strictly convex, differentiable function and $\mathcal{B}_\psi \colon (y, \xi) \mapsto \psi(y) - \psi(\xi) - \psi'(\xi)(y - \xi)$ is the Bregman divergence associated to ψ, which generalizes the quadratic potential. In fact, every continuously differentiable loss $\mathcal{L}$ that elicits the mean has this form (Theorem 3 and 4 in Banerjee et al. (2005)). Similarly, the pairs (mean, variance) and (VaR, CVaR) can be elicited by families of differentiable, strictly convex functions (see Table 2 in Appendix A).

2.2 Contextual bandits with linear elicitable risk measures

Regret For a linear bandit (φ, θ^*), we define the pseudo-regret associated with a loss $\mathcal{L}$ and a sequence of actions $(X_t)_{1\leqslant t\leqslant T}$ as $\mathcal{R}_T = \sum_{t=1}^T \rho_{\mathcal{L}}(\varphi(\langle \theta^*, X_t^* \rangle)) - \rho_{\mathcal{L}}(\varphi(\langle \theta^*, X_t \rangle))$, where $X_t^* = \operatorname{argmax}_{x\in\mathcal{X}_t} \rho_{\mathcal{L}}(\varphi(\langle \theta^*, x \rangle))$ is the optimal action w.r.t the risk measure $\rho_{\mathcal{L}}$. By definition, if the loss $\mathcal{L}$ is adapted to the linear bandit, this notion of regret reduces to $\mathcal{R}_T = \sum_{t=1}^T \langle \theta^*, X_t^* \rangle - \langle \theta^*, X_t \rangle$, which is formally the same as the standard regret for mean-linear bandits. What differs though is the meaning of $\langle \theta^*, X_t \rangle$, which can now represent any elicitable risk measure for the reward distribution. As an example, this framework paves the way for expectile-linear bandit of the form $Y_t = \langle \theta^*, X_t \rangle + \eta_t$ where the conditional expectile of η_t is zero and expectile rewards are measured as linear forms of the actions $\langle \theta^*, X_t \rangle$.

Supervised estimation of θ^* Similarly to how ridge regression provides natural predictors of the mean, we define the estimator $\widehat{\theta}_t \in \operatorname{argmin}_{\theta\in\Theta} \sum_{s=1}^{t-1} \mathcal{L}(Y_s, \langle \theta, X_s \rangle) + \frac{\alpha}{2}\|\theta\|_2^2$, which corresponds to the empirical risk minimization associated to an adapted loss $\mathcal{L}$, with L^2 regularization parameter $\alpha > 0$. Assuming that $\mathcal{L}$ is differentiable and that $\widehat{\theta}_t$ is in the interior of Θ, this estimator is characterized by $\alpha\widehat{\theta}_t = -\sum_{s=1}^{t-1} \partial\mathcal{L}\left(Y_s, \langle \widehat{\theta}_t, X_s \rangle\right) X_s$, where $\partial\mathcal{L}(y, \xi)$ stands for the derivative of $\xi \mapsto \mathcal{L}(y, \xi)$, i.e the partial derivative of $\mathcal{L}$ with respect to the second coordinate. When $\widehat{\theta}_t$ is not in the interior of Θ, an additional projection onto Θ is necessary, which we denote by the operator Π. We also define by $H_t^\alpha(\theta) = \sum_{s=1}^{t-1} \partial^2\mathcal{L}(Y_s, \langle \theta, X_s \rangle) X_s X_s^\top + \alpha I_d$ the Hessian of the empirical loss at θ of the minimization problem, with I_d the identity matrix of $\mathbb{R}^d$ and ∂^2 the second order derivative with respect to the second coordinate, and denote by $\|x\|_P = \sqrt{\langle x, Px \rangle}$ the norm associated with a positive definite matrix P (typically $P = H_t^\alpha(\theta)$ or $P = V_t^\alpha$).

We note that when $\mathcal{L}$ derives from the quadratic potential $\psi(\xi) = \xi^2/2$, it holds that $H_t^\alpha(\theta) = V_t^\alpha$ and we thus fall back to the mean-linear case. For all other choices of the loss function $\mathcal{L}$, the Hessian H_t^α depends on θ, and in particular no closed-form expression of $\widehat{\theta}_t$ in terms of the inverse of H_t^α is available. As we detail in the next sections, this introduces technical challenges to the analysis of linear bandit algorithms and forces the use of convex programming algorithms to numerically evaluate $\widehat{\theta}_t$.

Remark 2 *Similar complications arise in the case of generalized linear bandits (GLB) $Y_t = \mu(X_t) + \eta_t$, with $\mathbb{E}[\eta_t|\mathcal{F}_t] = 0$ and μ the link function between actions and expected rewards. In fact, GLB can be seen as a special case of the risk-aware setting with $\mathcal{L}$ the negative log-likelihood loss. While formally similar, we point out that GLB is designed solely to maximize the mean reward criterion. Another difference with our setting is that the use of maximum likelihood estimation requires a parametric assumption on the dsitribution of Y_t, typically that it belongs to a one-dimensional exponential family.*

Extension of LinUCB to linear bandits with convex loss The main benefit of the formulation of risk-awareness in terms of convex losses is that it suggests a transparent generalization of the standard LinUCB algorithm (OFUL in Abbasi-Yadkori et al. (2011), chapter 19 in Lattimore and Szepesvári (2020)), essentially substituting the least-squares estimate with the empirical risk minimizer associated with $\mathcal{L}$. The general idea of such optimistic algorithms is to play at time t the action $x \in \mathcal{X}_t$ with the highest plausible reward. In the mean-linear case with ridge regression, this highest plausible reward takes the form of $\langle\widehat{\theta}_t, x\rangle + \gamma_t(x)$, where $\gamma_t(x)$ is a certain action-dependent quantity also known as the *exploration bonus*. We write the general structure of our extension of LinUCB in Algorithm 1.

Algorithm 1 LinUCB for convex risk

Input: horizon T, loss function $\mathcal{L}$, regularisation parameter α, projection Π, exploration bonus sequence $(\gamma_t)_{t\in\mathbb{N}}$.
Initialization: Observe $\mathcal{X}_1$.
for $t = 1, \ldots, T$ **do**
  $\widehat{\theta}_t \in \arg\min_{\mathbb{R}^d} \sum_{s=1}^{t-1} \mathcal{L}(Y_s, \langle\theta, X_s\rangle) + \frac{\alpha}{2}\|\theta\|_2^2$; ▷ Empirical risk minimization
  $\bar{\theta}_t = \Pi(\widehat{\theta}_t)$; ▷ Projection
  $X_t = \arg\max_{x\in\mathcal{X}_t} \langle\bar{\theta}_t, x\rangle + \gamma_t(x)$; ▷ Play arm
  Observe Y_t and $\mathcal{X}_{t+1}$.

3. Analysis of risk-aware LinUCB

The goal of this section is to derive an exploration bonus sequence $(\gamma_t)_{t\in\mathbb{N}}$ and a projection operator Π¸ that ensure sub-linear regret of the corresponding LinUCB instance.

3.1 Martingale property and concentration

A key property for the analysis of mean-linear bandits is that the sum process $\sum_{s=1}^{t-1} \eta_s X_s$ naturally defines a vector-valued martingale in $\mathbb{R}^d$ with respect to the filtration $\mathcal{F}_t$ (Abbasi-Yadkori et al., 2011). This is not the case in general for bandits associated with generic convex losses. For a given loss $\mathcal{L}$, we know that $\theta^* = \mathrm{argmin}_\theta\, \mathbb{E}[\mathcal{L}(Y_t, \langle\theta, X_t\rangle)|\mathcal{F}_t]$, which implies $\mathbb{E}[\partial\mathcal{L}(Y, \langle\theta^*, X_t\rangle)|\mathcal{F}_t] = 0$ if $\mathcal{L}$ is differentiable, since X_t is measurable with respect to $\mathcal{F}_t$. A direct consequence of this is that $S_t = \sum_{s=1}^{t-1} \partial\mathcal{L}(Y_s, \langle\theta^*, X_s\rangle)X_s$ defines a $\mathcal{F}$-martingale. This process is at the heart of the next proposition, which establishes confidence bounds using the method of mixture (see Peña et al. (2008) in general and Abbasi-Yadkori et al. (2011); Faury et al. (2020) for specific applications to contextual bandits). We detail below a helpful assumption to transform the sum process S into a nonnegative supermartingales and state the high-probability uniform deviation bound we obtain, the proof of which is deferred to Appendix C.

Assumption 3 (Supermartingale control) *There exists $\sigma > 0$ such that for any $t \in \mathbb{N}$ and $\lambda \in \mathbb{R}^d$, we have*

$$\mathbb{E}\left[\exp\left(\partial\mathcal{L}\left(Y_t, \langle\theta^*, X_t\rangle\right)\lambda^\top X_t - \frac{\sigma^2}{2}\partial^2\mathcal{L}\left(Y_t, \langle\theta^*, X_t\rangle\right)\left(\lambda^\top X_t\right)^2\right)\middle|\mathcal{F}_t\right] \leqslant 1\,.$$

Proposition 4 (Method of mixtures for convex loss regression) *Let $\beta > 0$. Under Assumption 3, it holds that*

$$\mathbb{P}\left(\exists t \in \mathbb{N},\ \|S_t\|^2_{H_t^\beta(\theta^*)^{-1}} \geqslant \sigma^2\left(2\log\frac{1}{\delta} + \log\frac{\det H_t^\beta(\theta^*)}{\det \beta I_d}\right)\right) \leqslant \delta\,.$$

Discussion on Assumption 3 This assumption may look like an ad hoc attempt to mimic the sub-Gaussian method of mixtures, however we argue that it comes rather naturally if we impose some mild restrictions to the loss function. First, if the curvature of $\mathcal{L}$ is bounded away from zero, i.e there exists $m > 0$ such that $\partial^2\mathcal{L} \geqslant m$ uniformly, then it is sufficient to ask that $\partial\mathcal{L}(Y_t, \langle\theta^*, X_t\rangle)$ is $\sqrt{m}\sigma$-sub-Gaussian. Note in the standard mean-linear setting $Y_t = \langle\theta^*, X_t\rangle + \eta_t$ with $\mathcal{L}(y, \xi) = \frac{1}{2}(y-\xi)^2$, we have that $\partial\mathcal{L}(Y_t, \langle\theta^*, X_t\rangle) = \eta_t$, which is classically assumed to be sub-Gaussian. For other losses, it may be more convenient for the modeler to make assumptions on the distribution of observable quantities such as X_t and Y_t rather than directly on $\partial\mathcal{L}(Y_t, \langle\theta^*, X_t\rangle)$. Formally, this raises the question of how the sub-Gaussian property of a random variable Z transfers to $f(Z)$ for a given mapping f. While to the best of our knowledge no complete answer is available, several partial results exist:

(i) If Z is Gaussian with variance σ^2 and f is M-Lipschitz, the Tsirelson-Ibragimov-Sudakov inequality (Theorem 5.5 in Boucheron et al. (2013)) shows that $f(Z)$ is $M\sigma$-sub-Gaussian. In particular, the Lipschitz assumption holds for $\partial\mathcal{L}$ if the loss curvature is bounded from above by M uniformly. More generally, if Z can be written as a σ-Lipschitz function of a standard Gaussian random variable, then $f(Z)$ is $M\sigma$-sub-Gaussian.

(ii) If the density of Z is strongly log-concave, then $f(Z)$ is sub-Gaussian (with parameter related to the largest eigenvalue of the Hessian of the log-density, see Theorem 5.2.15 in Vershynin (2018)).

(iii) If Z is bounded (i.e actions and rewards are bounded) and f is Lipschitz and separately convex, then $f(Z)$ is sub-Gaussian (application of the entropy method, see e.g Theorem 6.10 in Boucheron et al. (2013)). The boundedness assumption can be lifted at the cost of a slightly more stringent condition than the sub-Gaussianity of Z, see Theorem 3 in Adamczak (2005).

Confidence set for θ^* To help write the above confidence set in terms of θ^* and the empirical estimator $\widehat{\theta}_t$, we introduce the function $F_t^\alpha \colon \theta \in \Theta \mapsto \sum_{s=1}^{t-1} \partial\mathcal{L}\left(Y_s, \langle \theta, X_s\rangle\right) X_s + \alpha\theta \in \mathbb{R}^d$. As seen above, $F_t^\alpha(\widehat{\theta}_t) = 0$ and $F_t^\alpha(\theta^*) = S_t + \alpha\theta^*$. Noticing that $\|F_t^\alpha(\theta^*) - F_t^\alpha(\widehat{\theta}_t)\|^2_{H_t^\beta(\theta^*)^{-1}} = \|S_t + \alpha\theta^*\|_{H_t^\beta(\theta^*)^{-1}} \leqslant \|S_t\|_{H_t^\beta(\theta^*)^{-1}} + \alpha\|\theta^*\|_{H_t^\beta(\theta^*)^{-1}}$, we immediately derive the following result.

Corollary 5 *For $t \in \mathbb{N}$, $\delta \in (0,1)$, $\alpha, \beta > 0$, let*

$$\widehat{\Theta}_t^\delta = \left\{ \theta \in \Theta,\ \|F_t^\alpha(\theta) - F_t^\alpha(\widehat{\theta}_t)\|_{H_t^\beta(\theta)^{-1}} \leqslant \sigma\sqrt{2\log\frac{1}{\delta} + \log\frac{\det H_t^\beta(\theta)}{\det \beta I_d}} + \alpha\|\theta\|_{H_t^\beta(\theta)^{-1}} \right\}.$$

Then under Assumption 3, it holds that $\mathbb{P}\left(\forall t \in \mathbb{N},\ \theta^* \in \widehat{\Theta}_t^\delta\right) \geqslant 1 - \delta$.

We constantly use this result in the following, in particular to construct the projection operator Π. Indeed, if we define $\bar{\theta}_t \coloneqq \Pi(\widehat{\theta}_t)$ as $\mathrm{argmin}_{\theta\in\Theta}\|F_t^\alpha(\theta) - F_t^\alpha(\widehat{\theta}_t)\|_{H_t^\beta(\theta)^{-1}}$, we have the property that $\Pi(\widehat{\theta}_t) \in \widehat{\Theta}_t^\delta$ with high probability.

3.2 Optimism and local metrics

We recall here the principle of optimism in the face of uncertainty and adapt it to the framework of elicitable risk measures. We denote by $r_t = \langle \theta^*, X_t^*\rangle - \langle \theta^*, X_t\rangle$ the instantaneous regret, where $\langle \theta^*, X_t^*\rangle = \max_{x\in\mathcal{X}_t}\langle \theta^*, x\rangle$ is the optimal risk measure associated with $\mathcal{L}$ at time for the actions available at time t. Then, simple algebra shows that

$$r_t = \langle \theta^* - \bar{\theta}_t, X_t^*\rangle - \langle \theta^* - \bar{\theta}_t, X_t\rangle + \langle \bar{\theta}_t, X_t^* - X_t\rangle = \Delta(X_t^*, \bar{\theta}_t) + \Delta(X_t, \bar{\theta}_t) + \langle \bar{\theta}_t, X_t^* - X_t\rangle\,,$$

where we define for $x \in \mathcal{X}$ and $\theta \in \Theta$, $\Delta(x,\theta) = |\langle \theta^* - \theta, x\rangle|$ the absolute error made by θ with respect to the true parameter of the linear bandit θ^* in the direction of x. If we know a sequence of functions $\gamma_t \colon \mathcal{X} \to \mathbb{R}_+$ such that with high probability, for all $t \in \mathbb{N}$ and $x \in \mathcal{X}_t$, $\Delta(x, \bar{\theta}_t) \leqslant \gamma_t(x)$, then the principle of optimism recommends the action $X_t \in \mathrm{argmax}_{x\in\mathcal{X}_t}\langle \bar{\theta}_t, x\rangle + \gamma_t(x)$, i.e the one leading to the best plausible reward with respect to the confidence on the prediction error of $\bar{\theta}_t$. In this case, $r_t \leqslant \Delta(X_t^*, \bar{\theta}_t) + \Delta(X_t, \bar{\theta}_t) + \gamma_t(X_t) - \gamma(X_t^*) \leqslant 2\gamma_t(X_t)$ with high probability, and hence $\mathcal{R}_T \leqslant 2\sum_{t=1}^T \gamma_t(X_t)$. We detail below how Corollary 5 coupled with standard assumptions provides such a bound.

Bound on the prediction error We follow the standard strategy of decoupling the dependency on $\bar{\theta}_t$ and x in $\Delta(x, \bar{\theta}_t)$. Indeed, we have by Cauchy-Schwarz's inequality that, for some positive definite matrix P to be determined later,

$$\Delta(x, \bar{\theta}_t) = |\langle \theta^* - \bar{\theta}_t, x\rangle| = |\langle P^{\frac{1}{2}}(\theta^* - \bar{\theta}_t), P^{-\frac{1}{2}}x\rangle| \leqslant \|\theta^* - \bar{\theta}_t\|_P \|x\|_{P^{-1}}\,.$$

As we see below, a natural choice for P is the (average) Hessian of the empirical risk minimization problem, and therefore the term $\|x\|_{P^{-1}}$ can be handled by the elliptical potential lemma (Lemma 11 in Abbasi-Yadkori et al.

(2011)). To control the remainder term in $\theta^* - \bar{\theta}_t$, we borrow technical tools from the classical approach developed for generalized linear bandits (Filippi et al., 2010; Faury et al., 2020) and note that

$$F_t^\alpha(\theta^*) - F_t^\alpha(\bar{\theta}_t) = \bar{H}_t^\alpha(\theta^*, \bar{\theta}_t)(\theta^* - \bar{\theta}_t)\,,,$$

where $\bar{H}_t^\alpha(\theta^*, \bar{\theta}_t) = \int_0^1 H_t^\alpha(u\theta^* + (1-u)\bar{\theta}_t)du$ is the average of the Hessian matrices along the segment $[\bar{\theta}_t, \theta^*]$ (this follows from the observation that the differential of ∇F_t^α is H_t^α). Therefore the choice $P = \bar{H}_t^\alpha(\theta^*, \bar{\theta}_t)$ yields

$$\|\theta^* - \bar{\theta}_t\|_P = \|F_t^\alpha(\theta^*) - F_t^\alpha(\bar{\theta}_t)\|_{\bar{H}_t^\alpha(\theta^*,\bar{\theta}_t)^{-1}} \leqslant \|F_t^\alpha(\theta^*) - F_t^\alpha(\widehat{\theta}_t)\|_{\bar{H}_t^\alpha(\theta^*,\bar{\theta}_t)^{-1}} + \|F_t^\alpha(\bar{\theta}_t) - F_t^\alpha(\widehat{\theta})\|_{\bar{H}_t^\alpha(\theta^*,\bar{\theta}_t)^{-1}}\,.$$

To conclude, we need to find a way to relate the local metric defined by $\bar{H}_t^\alpha(\theta^*, \bar{\theta}_t)^{-1}$ to those defined by $H_t^\beta(\theta^*)^{-1}$ and $H_t^\beta(\bar{\theta}_t)^{-1}$, for which we have high confidence bounds. This motivates the following assumption.

Assumption 6 (Transportation of local metrics) *For any $\alpha > 0$, there exists $\kappa > 0$ and $\beta > 0$ such that*

$$\bar{H}_t^\alpha(\theta^*, \bar{\theta}_t) \succcurlyeq \frac{1}{\kappa} H_t^\beta(\theta^*) \quad \textit{and} \quad \bar{H}_t^\alpha(\theta^*, \bar{\theta}_t) \succcurlyeq \frac{1}{\kappa} H_t^\beta(\bar{\theta}_t)\,.$$

In a special case of generalized linear bandit, namely the logistic bandit, it is shown in (Faury et al., 2020) that a similar assumption can be satisfied thanks to self-concordance properties of the link function. As detailed below, we will show that in our case it is sufficient to bound the extreme curvature of the loss function $\mathcal{L}$.

3.3 Regret analysis

So far, we have kept the assumptions on supermartingale control (Assumption 3) and transportation of local metrics (Assumption 6) fairly general. The condition below is easy to check in practice and implies both.

Assumption 7 (Bounded loss curvature) *There exists $m, M > 0$ such that*

$$\forall y, \xi \in \mathbb{R},\ m \leqslant \partial^2 \mathcal{L}(y, \xi) \leqslant M\,.$$

We call the parameter $\kappa = \frac{M}{m}$ the conditioning of the convex loss $\mathcal{L}$.

We detail in Appendix D how this control implies Assumption 6 for $\beta = \kappa\alpha$. It also implies that Assumption 3 relaxes to a Lipschitz sub-Gaussian property as discussed in Section 3.1.

Remark 8 *This assumption is reminiscent of the standard lower bound on the derivative of the link function μ' commonly encountered in the GLB literature. Moreover, although we formulate this bounded curvature assumption globally, we note that we only require it to hold in a convex neighborhood of θ^* containing $\bar{\theta}_t$, and Corollary 5 shows that with high probability, $\|\theta^* - \bar{\theta}_t\|_2$ is bounded (going from the $H_t^\beta(\theta^*)^{-1}$ norm to the Euclidean norm can be done by simple positive definite matrix inequalities). Therefore, one could instead assume a local curvature control on $\partial\mathcal{L}(y, \langle\theta, x\rangle)$ for $x \in \mathcal{X}_t$ and θ in a ball around θ^*, in the same spirit as Assumption 1 (Li et al., 2017) for GLB.*

In addition, we make two additional standard assumptions that prior bounds are known on θ^* and on the action set $\mathcal{X}$, which is standard in the existing literature on linear bandits.

Assumption 9 (Prior bound on parameters) *All parameters are in the Euclidean ball of radius S, i.e $\Theta \subseteq \mathcal{B}_{\|\cdot\|_2}(0, S)$. In particular, this implies that $\|\theta^*\|_{H_t^\beta(\theta^*)^{-1}} \leqslant \frac{S}{\sqrt{\beta}}$ for any $\beta > 0$.*

Assumption 10 (Prior bound on actions) *All actions are in the Euclidean ball of radius L, i.e $\mathcal{X} \subseteq \mathcal{B}_{\|\cdot\|_2}(0, L)$.*

We now obtain a high probability upper bound on the regret incurred by Algorithm 1 for an explicit choice of exploration bonus sequence $(\gamma_t)_{t\in\mathbb{N}}$ and projection Π.

Theorem 11 (Regret of LinUCB for convex risk) *Let $\delta \in (0,1)$, $\alpha \geqslant \max(1, L^2)$ and define for $t \in \mathbb{N}$ the exploration bonus*

$$\gamma_t : x \in \mathcal{X}_t \mapsto c_t^\delta \|x\|_{H_t^{\kappa\alpha}(\bar{\theta}_t)^{-1}} \quad \textit{and} \quad c_t^\delta = 2\kappa \left(\sigma \sqrt{2\log\frac{1}{\delta} + d\log\frac{m}{\alpha} + \log\det V_t^{\frac{\alpha}{m}}} + \sqrt{\frac{\alpha}{\kappa}} S \right)$$

and the projection operator $\Pi : \widehat{\theta} \in \mathbb{R}^d \mapsto \operatorname{argmin}_{\theta\in\Theta} \|F_t^\alpha(\theta) - F_t^\alpha(\widehat{\theta})\|_{H_t^\beta(\theta)^{-1}}$.

Under Assumptions 3, 7, 9 and 10, with probability at least $1-\delta$, the regret of Algorithm 1 is bounded by

$$\mathcal{R}_T \leqslant 2c_T^\delta \sqrt{\frac{2Td}{m}\log\left(1 + \frac{TL^2}{d\alpha}\right)}.$$

In particular, we have $\mathcal{R}_T = \mathcal{O}\left(\frac{\kappa\sigma d}{\sqrt{m}}\sqrt{T}\log\frac{TL^2}{d}\right)$ when $T \to +\infty$.

The proof of this result follows the standard regret analysis of LinUCB, up to the modification detailled in the previous sections. We report the detailed arguments in Appendix E.

Remark 12 *We note that the regret bound scales with the conditioning κ, a phenomenon already observed for generalized linear bandits in (Filippi et al., 2010). It was also conjectured in this work that the scaling could be reduced to $\sqrt{\kappa}$. In our proof, the extra $\sqrt{\kappa}$ is due to transportation of local metrics using the curvature bound of Assumption 7. In the special case of logistic bandits, this extra factor was removed by (Faury et al., 2020) using self-concordance of the sigmoid. In our case, a sharper assumption on the transportation of local metrics may improve the scaling similarly.*

4. Approximate convex risk minimisation with online gradient descent

So far, we have shown that the standard LinUCB principle can be extended to the convex loss setting with similar regret guarantees under some curvature assumption. However, this comes at the cost of a significant computational overhead since the estimator $\widehat{\theta}_t$ needs to be calculated from scratch at each step as $\operatorname{argmin}_{\theta\in\mathbb{R}^d} \sum_{s=1}^{t-1} \mathcal{L}(Y_s, \langle\theta, X_s\rangle) + \frac{\alpha}{2}\|\theta\|_2^2$. As a reminder, in the standard mean-linear case, this estimator has an analytical expression that amounts to incrementally inverting the matrix V_t^α, which can be done efficiently from the knowledge of the inverse of V_{t-1}^α.

We propose an alternative algorithm that relies on online gradient descent (OGD) to compute a fast approximation of $\widehat{\theta}_t$. This may be of practical interest to deploy risk-aware linear bandits in time-sensitive environments, such as in real-time online recommendation systems. Moreover, it can also be used in the mean-linear setting with high dimensional action sets, where computing gradients may be more tractable than inverting a large $d \times d$ matrix.

The intuition behind Algorithm 2 is that at time $t = nh+1$, the approximation error between the OGD estimate $\bar{\theta}_n^{\mathrm{OGD}}$ and the exact minimizer of the empirical risk $\widehat{\theta}_t$ induces additional exploration, which translates into an increased regret compared to LinUCB. In other words, LinUCB-OGD trades off accuracy for computational efficiency. The episodic structure is borrowed from Ding et al. (2021) and is key to ensure sufficient convexity of the aggregate loss $\sum_{k=1}^h \partial\mathcal{L}(Y_{(n-1)h+k}, \langle\widehat{\theta}_{n-1}^{\mathrm{OGD}}, X_{(n-1)h+k}\rangle)$. This allows to leverage the strong approximation guarantees of OGD, which we extend in the following proposition by relaxing the standard boundedness requirement of the gradient (Theorem 3.3, Hazan (2019)) to a weaker sub-Gaussian control at a given parameter (proof in Appendix F).

Proposition 13 (OGD regret with sub-Gaussian gradients) *Let $\mathcal{C}$ a convex subset of $\mathbb{R}^d$ and Π the projection operator onto $\mathcal{C}$. For $j = 1, \ldots, N$, let $\ell_j : \mathcal{C} \longrightarrow \mathbb{R}_+$ a twice differentiable convex function and $a, A > 0$ such that $aI_d \preccurlyeq \nabla^2\ell_j(z) \preccurlyeq AI_d$ for all $z \in \mathcal{C}$. Define the OGD update at step j by $z_j = \Pi(z_{j-1} - \varepsilon_{j-1}\nabla\ell_j(z_{j-1}))$ and $\bar{z}_n = \arg\min_{z\in\mathcal{C}} \sum_{j=1}^n \ell_j(z)$. Assume that there exists $z^* \in \mathcal{C}$ such that $\nabla\ell_j(z^*) = g_j + \frac{\alpha}{n}z^*$ with $\alpha \geqslant 0$ and g*

Algorithm 2 LinUCB-OGD for convex risk

Input: horizon T, loss function $\mathcal{L}$, regularisation parameter α, projection Π, exploration bonus sequence $(\gamma_{t,T}^{\mathrm{OGD}})_{t\leqslant T}$, gradient descent step sequence $(\varepsilon_t)_{t\in\mathbb{N}}$, episode length $h>0$.

Initialization: Observe $\mathcal{X}_1$, set $\widehat{\theta}_0^{\mathrm{OGD}}$, $t=1$, $n=1$.

for $t=1,\dots,T$ **do**

  if $t=nh+1$ **then**

    $\widehat{\theta}_n^{\mathrm{OGD}} = \widehat{\theta}_{n-1}^{\mathrm{OGD}} - \varepsilon_{n-1}\left(\sum_{k=1}^{h}\partial\mathcal{L}(Y_{(n-1)h+k}, \langle\widehat{\theta}_{n-1}^{\mathrm{OGD}}, X_{(n-1)h+k}\rangle) + \alpha\widehat{\theta}_{n-1}^{\mathrm{OGD}}\right)$; ▷ `OGD`

    $\bar{\theta}_n^{\mathrm{OGD}} = \frac{1}{n}\sum_{j=1}^{n}\Pi(\widehat{\theta}_j^{\mathrm{OGD}})$; ▷ `Average over previous OGD steps`

    $n \leftarrow n+1$

  $X_t = \arg\max_{x\in\mathcal{X}_t}\langle\bar{\theta}_n^{\mathrm{OGD}}, x\rangle + \gamma_{t,T}^{\mathrm{OGD}}(x)$; ▷ `Play with same parameter for` h `steps`

  Observe Y_t and $\mathcal{X}_{t+1}$;

  $t \leftarrow t+1$;

a centered, $\mathbb{R}^d$-valued σ-sub-Gaussian process, and also that $\mathcal{C}$ is bounded, i.e diam$(\mathcal{C}) = \sup_{z,z'\in\mathcal{C}}\|z-z'\| < \infty$. Then with probability at least $1-\delta$, the OGD regret with step size $\varepsilon_j = \frac{3}{aj}$ is bounded for all $n \leqslant N$ by

$$\sum_{j=1}^{n}\ell_j(z_j) - \ell_j(\bar{z}_n) \leqslant \frac{9}{2a}\left(2d\sigma^2\log\frac{2dN}{\delta} + A^2 \mathit{diam}(\mathcal{C})^2 + \frac{\alpha^2}{n^2}\|z^*\|^2\right)(1+\log n)\ .$$

This can be written more concisely as $\sum_{j=1}^{N}\ell_j(z_j) - \ell_j(\bar{z}_N) = \mathcal{O}(\frac{d\sigma^2}{a}\log^2 N)$ when $N\to+\infty$. In addition, if g is uniformly bounded by a constant $G>0$, the regret with step size $\varepsilon_s = \frac{1}{aj}$ can be reduced to the almost sure bound:

$$\sum_{j=1}^{n}\ell_j(z_j) - \ell_j(\bar{z}_n) \leqslant \frac{G^2}{2a}(1+\log n)\ .$$

Our final result, which we prove in Appendix G, states that the approximation error of OGD induces at most a polylog correction in the regret of LinUCB-OGD.

Theorem 14 (Regret of LinUCB-OGD for convex risk) *In addition to the conditions of Theorem 11, assume that there exists $\varepsilon_h > 0$ such that for all $j \leqslant T/h$, $\sum_{k=1}^{h} X_{(j-1)h+k}X_{(j-1)h+k}^\top \succcurlyeq \varepsilon_h I_d$, and also that $\partial\mathcal{L}(Y_t, \langle\theta^*, X_t\rangle)$ is $\sqrt{m}\sigma$-sub-Gaussian for all $t\leqslant T$. Let $C>0$ and define the exploration bonus sequence by:*

$$\gamma_{t,T}^{OGD} : x\in\mathcal{X}_t \mapsto c_{t,T}^{OGD,\delta}\|x\|_{H_t^{\kappa\alpha}(\bar{\theta}^{OGD}_{\lfloor\frac{t-1}{h}\rfloor})^{-1}} \quad \text{and} \quad c_{t,T}^{OGD,\delta} = c_t^{\delta} + 2L\sqrt{\frac{C\kappa m d h^2\sigma^2}{\varepsilon_h t}\log\left(\frac{2dT}{h\delta}\right)\log\left(\frac{t}{h}\right)}\ .$$

Then if $C>0$ is large enough, with probability at least $1-2\delta$, the regret of Algorithm 2 with the OGD step sequence of Proposition 13 is bounded $\tilde{\mathcal{O}}\left(\sqrt{T}\right)$, where $\tilde{\mathcal{O}}$ hides polylogarithmic terms in T.

Remark 15 *The lower bound in ε_h is here to enforce strong convexity of the agregate episodic loss, together with Assumption 7. For i.i.d actions, it is known to hold with high probability provided the episode length h is large enough, (Proposition 1, Li et al. (2017)). Also, note that the union bound used in Proposition 13 imposed the knowledge of the time horizon T during the run of the algorithm (in the definition of $\gamma_{t,T}$), making Algorithm 2 not anytime.*

5. Experiments

We conducted three numerical experiments to illustrate the performance of the two risk-aware algorithms, which are reported in Figure 1. In the first two, we considered arms with respectively Gaussian and expectile-based asymmetric

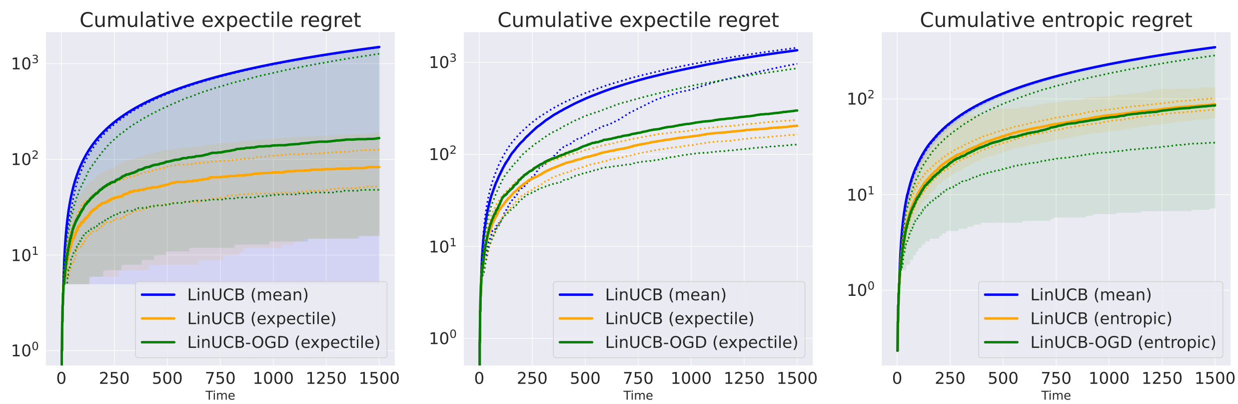

Figure 1: Left: two-armed Gaussian expectile bandit. Center: two-armed linear expectile bandit with $\mathbb{R}^3$ contexts and expectile-based asymmetric noises. Right: two-armed Bernoulli entropic risk bandit. Thick lines denote median cumulative regret over 500 independent replications. Dotted lines denote the 25 and 75 regret percentiles. Shaded areas denote the 5 and 95 percentiles.

distributions (Torossian et al., 2020) with context-dependent expectiles. In the last one, we considered Bernoulli-like arms under the entropic risk criterion. These settings were designed so that the optimal arms were different when considering the mean or the corresponding risk-aware criterion. Instances of the classical LinUCB algorithm (Abbasi-Yadkori et al., 2011) were indeed deceived and accumulated linear risk-aware regret, while Algorithms 1 and 2 exhibit milder sublinear trends. Compared to LinUCB with the exact minimization of the empirical risk, the LinUCB-OGD variant accumulated slightly more regret and showed higher variability across independent replications, at the benefit of an improved runtime. Further details of the experiments are reported in Appendix H.

6. Conclusion

We have introduced a new setting for contextual bandits, building on the recent interest for risk-awareness in multi-armed bandits. We reviewed the literature on risk measures, in particular the notion of elicitability, that allows to extend the risk minimization framework of ridge regression beyond standard mean-linear bandits. We detailed a set of generic assumptions to lift the regret analysis of optimistic algorithms in the setting of scalar risk measures $\rho_{\mathcal{L}}$ elicited by a convex loss $\mathcal{L}$, and showed that uniformly bounding the curvature of the loss is sufficient to maintain similar theoretical guarantees ($\mathcal{O}(\sqrt{T})$ worst-case regret, up to polylog terms). Going further, we believe it would be interesting to extend the linear model between actions and risk measures to generalized linear models ($\rho_{\mathcal{L}}(Y_t) = \mu(\langle \theta, X_t \rangle)$ for some link function $\mu \colon \mathbb{R} \to \mathbb{R}$), kernelized bandits ($\rho_{\mathcal{L}}(Y_t) = f(X_t)$ where f belongs to some RKHS) or neural bandits ($\rho_{\mathcal{L}}(Y_t) = f_\theta(X_t)$ where f_θ is a neural network with weights θ). Moreover, capturing well-established risk measures such as mean-variance and conditional value-at-risk would require to develop the theory of high-order elicitable measures. Finally, we believe an extended empirical study would be valuable, e.g by testing the proposed algorithms on real-world data collected by risk-averse agents.

Acknowledgments

The authors acknowledge the funding of the French National Research Agency, the French Ministry of Higher Education and Research, Inria, the MEL and the I-Site ULNE regarding project R-PILOTE-19-004-APPRENF and Bandits For Health (B4H). We thank the anonymous reviewers for their careful reading of the paper and their suggestions for improvements. Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Belkacem Abdous and Bruno Remillard. Relating quantiles and expectiles under weighted-symmetry. *Annals of the Institute of Statistical Mathematics*, 47(2):371–384, 1995.

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR, 2017.

Radoslaw Adamczak. Logarithmic sobolev inequalities and concentration of measure for convex functions and polynomial chaoses. *arXiv preprint math/0505175*, 2005.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

Julyan Arbel, Stéphane Girard, H Nguyen, and Antoine Usseglio-Carleve. Multivariate expectile-based distribution: properties, bayesian inference, and applications. 2021.

Arindam Banerjee, Xin Guo, and Hui Wang. On the optimality of conditional expectation as a bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005.

Dorian Baudry, Romain Gautron, Emilie Kaufmann, and Odalric Maillard. Optimal thompson sampling strategies for support-aware cvar bandits. In *International Conference on Machine Learning*, pages 716–726. PMLR, 2021.

Fabio Bellini and Elena Di Bernardino. Risk management with expectiles. *The European Journal of Finance*, 23(6): 487–506, 2017.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Mario Brandtner, Wolfgang Kürsten, and Robert Rischau. Entropic risk measures and their comparative statics in portfolio selection: Coherence vs. convexity. *European Journal of Operational Research*, 264(2):707–716, 2018.

Jonas Brehmer. Elicitability and its application in risk management. *arXiv preprint arXiv:1707.09604*, 2017.

Asaf Cassel, Shie Mannor, and Assaf Zeevi. A general approach to multi-armed bandits under risk criteria. In *Conference On Learning Theory*, pages 1295–1306. PMLR, 2018.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression, 2017.

Qin Ding, Cho-Jui Hsieh, and James Sharpnack. An efficient algorithm for generalized linear bandit: Online stochastic gradient descent and thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 1585–1593. PMLR, 2021.

Kevin Dowd. *Measuring market risk*. John Wiley & Sons, 2007.

Paul Embrechts, Tiantian Mao, Qiuqi Wang, and Ruodu Wang. Bayes risk, elicitability, and the expected shortfall. *Mathematical Finance*, 31(4):1190–1217, 2021.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 23, 2010.

Tobias Fissler and Johanna F Ziegel. Higher order elicitability and osband's principle. *The Annals of Statistics*, 44(4): 1680–1707, 2016.

Tobias Fissler, Johanna F Ziegel, and Tilmann Gneiting. Expected shortfall is jointly elicitable with value at risk-implications for backtesting. *arXiv preprint arXiv:1507.00244*, 2015.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

Nicolas Galichet, Michele Sebag, and Olivier Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pages 245–260. PMLR, 2013.

Elad Hazan. Introduction to online convex optimization, 2019.

Con Keating and William F Shadwick. A universal performance measure. *Journal of performance measurement*, 6 (3):59–84, 2002.

Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvári. Asymptotically optimal information-directed sampling. In *Conference on Learning Theory*, pages 2777–2821. PMLR, 2021.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017.

Odalric-Ambrym Maillard. Robust risk-averse stochastic multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 218–233. Springer, 2013.

Harry Markowitz. March 1952. portfolio selection. *Journal of finance*, 7(1):77–91, 1952.

Whitney K Newey and James L Powell. Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847, 1987.

Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.

Collin Philipps. Interpreting expectiles. *Available at SSRN 3881402*, 2022.

LA Prashanth, Krishna Jagannathan, and Ravi Kumar Kolla. Concentration bounds for cvar estimation: The cases of light-tailed and heavy-tailed distributions. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5577–5586, 2020.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Advances in Neural Information Processing Systems*, 27, 2014.

Alex Tamkin, Ramtin Keramati, Christoph Dann, and Emma Brunskill. Distributionally-aware exploration for cvar bandits. In *NeurIPS 2019 Workshop on Safety and Robustness in Decision Making; RLDM*, volume 2020, 2019.

Léonard Torossian, Victor Picheny, and Nicolas Durrande. Bayesian quantile and expectile optimisation. *arXiv preprint arXiv:2001.04833*, 2020.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

D. Williams. *Probability with Martingales*. Cambridge mathematical textbooks. Cambridge University Press, 1991. ISBN 9780521406055.

Maximilian Wirth, Anja Klein, and Andrea Ortiz. Risk-aware multi-armed bandits for vehicular communications. 2022.

Johanna F Ziegel. Coherence and elicitability. *Mathematical Finance*, 26(4):901–918, 2016.

Table 2: Example of elicitable risk measures.

Name	$\rho_{\mathcal{L}}(\nu)$	Associated loss $\mathcal{L}(y, \xi)$	Domain
Mean	$\mathbb{E}_{Y\sim\nu}[Y]$	$(y-\xi)^2$ Bregman divergence $\mathcal{B}_\psi(y,\xi)$ $\psi(y) - \psi(\xi) - \psi'(\xi)(y-\xi)$, ψ differentiable, strictly convex.	$\xi \in \mathbb{R}$
Derived from potential ψ	$\text{argmin}_{\xi\in\mathbb{R}} \mathbb{E}_{Y\sim\nu}[\psi(Y-\xi)]$	$\psi(y-\xi)$	$\xi \in dom(\psi)$
Generalized moment $T\colon \mathbb{R}\to\mathbb{R}$	$\mathbb{E}_{Y\sim\nu}[T(Y)]$	$\frac{1}{2}\xi^2 - \xi T(y)$	$\xi \in \mathbb{R}$
Entropic risk, $\gamma \neq 0$ (Example 1, Embrechts et al. (2021))	$\frac{1}{\gamma}\log \mathbb{E}_{Y\sim\nu}[e^{\gamma Y}]$	$\xi + \frac{1}{\gamma}(e^{\gamma(y-\xi)} - 1)$	$\xi \in \mathbb{R}$
(mean, variance) (Example 1.23, Brehmer (2017))	$\mu = \mathbb{E}_{Y\sim\nu}[Y]$ $\sigma^2 = \mathbb{E}_{Y\sim\nu}[Y^2] - \mu^2$	$\frac{1}{2}\xi_1^2 + \frac{1}{2}(\xi_2+\xi_1^2)^2$ $-\xi_1 y - (\xi_2+\xi_1^2)y^2$ $-\psi_1(\xi_1) - \psi_1'(\xi_1)(y-\xi_1)$ $-\psi_2(\xi_2+\xi_1^2)$ $-\psi_2'(\xi_2+\xi_1^2)(y^2-\xi_2-\xi_1^2))$, ψ_1, ψ_2 differentiable, strictly convex.	$\xi_1 \in \mathbb{R}$ $\xi_2 \geqslant 0$
$(\text{VaR}_\alpha, \text{CVaR}_\alpha)$, $\alpha \in (0,1)$ (Corollary 5.5, Fissler and Ziegel (2016))	$\text{VaR}_\alpha = \inf\{y\in\mathbb{R}, \int_{-\infty}^{y} d\nu \geqslant \alpha\}$ $\text{CVaR}_\alpha = \frac{1}{\alpha}\int_0^\alpha \text{VaR}_a da$	$(\xi_1 - y)_+ - \alpha\xi_1$ $+\xi_2(\frac{1}{\alpha}(\xi_1-y)_+ - \xi_1)$ $+\frac{1}{2}\xi_2^2$ $(\mathbb{I}_{y\leqslant\xi_1} - \alpha)\psi_1'(\xi_1)$ $-\mathbb{I}_{y\leqslant\xi_1}\psi_1'(y)$ $+\psi_2'(\xi_2)(\xi_2-\xi_1+\frac{1}{\alpha}\mathbb{I}_{y\leqslant\xi_1}(\xi_1-y))$ $-\psi_2(\xi_2) + c(y)$, ψ_1 convex, ψ_2 strictly convex and increasing, $c\colon \mathbb{R}\to\mathbb{R}$.	$\xi_1 \geqslant \xi_2$

Appendix A. Summary and interpretation of elicitable risk measures

We report in Table 2 an overview of common elicitable risk measures and their associated loss functions. We recall that for a distribution ν over $\mathbb{R}$ and a loss function $\mathcal{L}\colon \mathbb{R}\times\mathbb{R}^p \to \mathbb{R}$, we defined the risk measure elicited by $\mathcal{L}$ as $\rho_{\mathcal{L}}(\nu) = \arg\min_{\xi\in\mathbb{R}^p} \mathbb{E}_{Y\sim\nu}[\mathcal{L}(Y,\xi)]$. Note that the pairs (mean, variance) and (VaR, CVaR) are second-order elicitable but neither the variance nor the CVaR are first-order elicitable. For these pairs, we report the generic form of elicitation losses, which depend on arbitrary convex functions ψ_1 and ψ_2, as well as instances of such losses obtained for the natural choice $\psi_1(\xi) = \psi_2(\xi) = \xi^2/2$.

We provide below some intuition about these commonly used measures in risk management.

Mean-Variance Assessing the risk-reward tradeoff of an underlying distribution ν by penalizing its mean by a higher order moment (typically the variance) is perhaps the most intuitive of risk measures. Following Markowitz (1952), the mean-variance risk measure at risk aversion level $\lambda \in \mathbb{R}$ is defined by $\rho_{MV_1}(\nu) = \mu - \lambda\sigma$, where μ and σ denote the mean and standard deviation of ν. Alternatively, it can also be defined as $\rho_{MV_2}(\nu) = \mu - \frac{\lambda}{2}\sigma^2$, using the variance rather than the standard deviation in the penalization term. Both measures are especially well-suited for Gaussian distributions as μ and σ fully characterize this family.

VaR and CVaR For a distribution with continuous cdf (i.e. it has no atom), the Value-at-Risk $\text{VaR}_\alpha(\nu)$ at level $\alpha \in (0,1)$ is equivalent to the α quantile, and a simple change of variable reveals that the Conditional Value-at-Risk $\text{CVaR}_\alpha(\nu)$ is thus $\mathbb{E}\left[X \mid X \leqslant \text{VaR}_\alpha(\nu)\right]$. Intuitively, a random variable with a high CVaR_α distribution takes on average relatively high values in the "$\alpha\%$ worst-case" scenario. For $\alpha \to 1^-$, $\text{CVaR}_\alpha(\nu) \to \mathbb{E}_{Y\sim\nu}[Y]$ and thus the risk measure becomes oblivious to the tail risk; on the contrary, the case $\alpha \to 0^+$ emphasizes only the worst outcomes.

In the Gaussian case $\nu \sim \mathcal{N}(\mu, \sigma)$, using the notations ϕ and Φ respectively for the pdf and cdf of the standard normal distribution, simple calculus shows that

$$\begin{aligned}\text{VaR}_\alpha(\nu) &= \mu + \sigma\Phi^{-1}(\alpha)\,,\\ \text{CVaR}_\alpha(\nu) &= \mu - \frac{\sigma}{\alpha\sqrt{2\pi}}\phi\left(\Phi^{-1}(\alpha)\right)\,,\end{aligned}$$

i.e. $\text{CVaR}_\alpha(\nu) = \rho_{MV_1}(\nu)$ with risk aversion level $\lambda = \frac{1}{\alpha\sqrt{2\pi}}\phi\left(\Phi^{-1}(\alpha)\right)$. In particular, increasing the variance σ^2 reduces $\text{CVaR}_\alpha(\nu)$, corresponding to the intuition of higher volatility risk.

Entropic risk The non-elicitability of CVaR_α motivated the use of the entropic risk as an alternative measure. This measures rewrites as (see Brandtner et al. (2018))

$$\rho_\gamma(\nu) = \sup_{\nu' \text{ probability measure}} \left\{\mathbb{E}_{Y\sim\nu'}[Y] - \frac{1}{\gamma}\text{KL}(\nu'\|\nu)\right\}\,.$$

The intuition here is similar to the mean-variance measure, i.e penalizing the expected value by a measure of uncertainty, but differs by the use of the Kullback-Leibler divergence $\texttt{KL}(\nu'\|\nu) = \mathbb{E}_{Y\sim\nu'}[\log\frac{d\nu'}{d\nu}]$ instead of the variance. The entropic risk measure can be interpreted as the expected value that a mispecified model ν' (in place of the true underlying distribution ν) would have, where $\texttt{KL}(\nu'\|\nu)$ controls the magnitude of the mispecification.

Again, in the Gaussian case, this measure reduces to $\rho_\gamma(\nu) = \mu + \frac{\gamma}{2}\sigma^2 = \rho_{MV_2}(\nu)$ at risk aversion level $\lambda = -\gamma$.

Expectile Beyond their interpretation as generalized, smooth quantiles, expectiles can also be understood in light of the financial risk management literature. Let $e_p(\nu)$ denote the p-expectile of ν for a given probability $p \in (0,1)$. Then, simple calculus shows that

$$(1-p)\mathbb{E}_{Y\sim\nu}[(e_p(\nu) - Y)_+] = p\mathbb{E}_{Y\sim\nu}[(Y - e_p(\nu))_+]\,,$$

where $z_+ = \max(z, 0)$. If ν represents the distribution of a tradeable asset Y at time T, then the p-expectile is the strike $K = e_p(\nu)$ such that call and put on Y struck at K at maturity T are in proportion $\frac{1-p}{p}$ to each other, where we define the call and put prices (with zero time discounting) by respectively

$$\begin{aligned}C(\nu, K) &= \mathbb{E}_{Y\sim\nu}[(Y-K)_+]\,,\\ P(\nu, K) &= \mathbb{E}_{Y\sim\nu}[(K-Y)_+]\,.\end{aligned}$$

Similarly, Keating and Shadwick (2002) introduced the notion of Omega ratio as a risk-return performance measures. It is defined at level K by

$$\Omega(K) = \frac{\int_K^{+\infty}(1-F(y))\,dy}{\int_{-\infty}^{K} F(y)dy}\,,$$

where F is the cdf of ν. This ratio can also be viewed as a call-put ratio, hence another definition of the p-expectile is via the implicit equation $\Omega(K) = \frac{1-p}{p}$ for $K = e_p(\nu)$.

We now derive an elementary proof of the sensitivity of the expectile with respect to the standard deviation σ of ν, based on the call-put parity principle, which states that for any distribution ν and strike K,

$$C(\nu, K) - P(\nu, K) = \mathbb{E}_{Y\sim\nu}[Y] - K\,.$$

Assuming ν is a normal or lognormal distribution (i.e. essentially a distribution parametrized by its variance, as in the Bachelier or Black-Scholes models), we write $C(\nu, K) = C(\sigma, K)$, $P(\nu, K) = P(\sigma, K)$ and $e_p(\sigma)$. Notice that the call-parity principle implies that $\partial_\sigma C = \partial_\sigma P$. We call this quantity Vega and denote it by V. From the relation $C(\sigma, e_p(\sigma)) = (1-p)/pP(\sigma, e_p(\sigma))$, we deduce that

$$\frac{d}{d\sigma}C(\sigma, e_p(\sigma)) = \partial_\sigma C(\sigma, e_p(\sigma)) + \partial_K C(\sigma, e_p(\sigma))\frac{de_p(\sigma)}{d\sigma}\,,$$
$$\frac{d}{d\sigma}P(\sigma, e_p(\sigma)) = \partial_\sigma P(\sigma, e_p(\sigma)) + \partial_K P(\sigma, e_p(\sigma))\frac{de_p(\sigma)}{d\sigma}\,,$$

and thus

$$\frac{1-2p}{p}V + \frac{de_p(\sigma)}{d\sigma}\left(\frac{1-p}{p}\partial_K P(\sigma, e_p(\sigma)) - \partial_K C(\sigma, e_p(\sigma))\right) = 0\,.$$

Elementary option pricing principles show that $V \geqslant 0$, i.e. the call and put prices both increase with higher volatility, as well as $\partial_K C \leqslant 0$ and $\partial_K P \geqslant 0$. Therefore, we deduce that $\frac{de_p(\sigma)}{d\sigma} \leqslant 0$, i.e. the p-expectile decreases with higher volatility, thus making it suitable for risk management purposes.

In particular for $p = 1/2$, the p-expectile corresponds to the strike K at which call and put have equal prices, which by the call-put parity principle (with zero discounting) implies that $K = \mathbb{E}[Y]$, thus giving an alternative derivation of the equivalence between $1/2$-expectile and mean.

Appendix B. Properties of convex losses and potentials

Before we prove Proposition 17, we write the following technical lemma.

Lemma 16 (Risk measures ρ_ψ are additive) *Let $\psi\colon \mathbb{R} \to \mathbb{R}$ be a strongly convex, differentiable function, ν be a distribution over $\mathbb{R}$ and $c \in \mathbb{R}$. Then $\rho_\psi(\nu + c) = \rho_\psi(\nu) + c$.*

Proof For the sake of simplicity, we assume ν admits a density p (with respect to the Lebesgue measure) and that ψ and p are regular enough to allow for differentiation under the following integral. Then the risk measure associated with $\mathcal{L}_\psi$ reads $\rho_\psi(\nu) = \operatorname{argmin}_{\xi\in\mathbb{R}} \int \psi\left(y - \xi\right) p(y)dy$ and the first order condition gives $\int \psi'(y-\rho_\psi(\nu))p(y)dy = 0$. Similarly, for any $c \in \mathbb{R}$, we have $\int \psi'(y - \rho_\psi(\nu + c))p(y - c)dy = 0$ since the density of $\nu + c$ is $p(\cdot - c)$. We now deduce from a simple change of variable $z = y - c$ that $\int \psi'(z + c - \rho_\psi(\nu + c))p(z)dz = 0$, which shows that $\rho_\psi(\nu + c) - c$ is also a minimizer of $\xi \mapsto \int \psi\left(y - \xi\right) p(y)dy$. By uniqueness ($\psi$ is strongly convex), we deduce that $\rho_\psi(\nu + c) = \rho_\psi(\nu) + c$. ■

Noise additivity for losses derived from potentials

Proposition 17 *Assume $\mathcal{L}_\psi$ is adapted to the linear bandit (φ, θ^*) and ψ is strongly convex and differentiable. Then there exists a stochastic process η such that the bandit is represented at time t by $Y_t \sim \langle\theta^*, X_t\rangle + \eta_t$ and $\rho_\psi(\eta|\mathcal{F}_t) = 0$.*

Proof Define the process η at time t by $\eta_t = Y_t - \langle\theta^*, X_t\rangle$. To compute $\rho_\psi(\nu|\mathcal{F}_t)$, note that X_t is measurable with respect to $\mathcal{F}_t$, therefore by Lemma 16 and the properties of conditional expectation, we have that $\rho_\psi(\eta_t|\mathcal{F}_t) = \rho_\psi\left(Y_t|\mathcal{F}_t\right) - \langle\theta^*, X_t\rangle = \rho_\psi\left(\varphi\left(\langle\theta^*, X_t\rangle\right)|\mathcal{F}_t\right) - \langle\theta^*, X_t\rangle = 0$ by definition of $\mathcal{L}_\psi$ being adapted to the bandit (φ, θ^*).

■

Appendix C. Proof of Proposition 4

Assumption 3 (Supermartingale control) *There exists $\sigma > 0$ such that for any $t \in \mathbb{N}$ and $\lambda \in \mathbb{R}^d$, we have*

$$\mathbb{E}\left[\exp\left(\partial\mathcal{L}\left(Y_t, \langle\theta^*, X_t\rangle\right)\lambda^\top X_t - \frac{\sigma^2}{2}\partial^2\mathcal{L}\left(Y_t, \langle\theta^*, X_t\rangle\right)\left(\lambda^\top X_t\right)^2\right)\middle|\mathcal{F}_t\right] \leqslant 1\,.$$

Proposition 4 (Method of mixtures for convex loss regression) *Let $\beta > 0$. Under Assumption 3, it holds that*

$$\mathbb{P}\left(\exists t \in \mathbb{N},\ \|S_t\|^2_{H_t^\beta(\theta^*)^{-1}} \geqslant \sigma^2\left(2\log\frac{1}{\delta} + \log\frac{\det H_t^\beta(\theta^*)}{\det \beta I_d}\right)\right) \leqslant \delta\,.$$

Proof The proof follows the method of mixture techniques, popularized in bandits by Abbasi-Yadkori et al. (2011). For $\lambda \in \mathbb{R}^d$, we define the process $M_t^\lambda = \exp\left(\lambda^\top S_t - \frac{\sigma^2}{2}\|\lambda\|^2_{H_t^0(\theta^*)}\right)$. We recall the expression of the Hessian $H_t^0(\theta) = \sum_{s=1}^{t-1}\partial^2\mathcal{L}\left(Y_s, \langle\theta, X_s\rangle\right)X_sX_s^\top$ and that in particular $\|\lambda\|^2_{H_t^0(\theta^*)} = \sum_{s=1}^{t-1}\partial^2\mathcal{L}\left(Y_s, \langle\theta, X_s\rangle\right)\left(\lambda^\top X_s\right)^2$. This process is nonnegative and defines as supermartingale since

$$\begin{aligned}
\mathbb{E}\left[M_{t+1}^\lambda\middle|\mathcal{F}_t\right] &= \mathbb{E}\left[\exp\left(\lambda^\top S_{t+1} - \frac{\sigma^2}{2}\|\lambda\|^2_{H_{t+1}^0(\theta^*)}\right)\middle|\mathcal{F}_t\right] \\
&= \mathbb{E}\left[\exp\left(\lambda^\top S_t - \frac{\sigma^2}{2}\|\lambda\|^2_{H_t^0(\theta^*)} + \partial\mathcal{L}\left(Y_t, \langle\theta^*, X_t\rangle\right)\lambda^\top X_t - \frac{\sigma^2}{2}\partial^2\mathcal{L}\left(Y_t, \langle\theta^*, X_t\rangle\right)\left(\lambda^\top X_t\right)^2\right)\middle|\mathcal{F}_t\right] \\
&= \exp\left(\lambda^\top S_t - \frac{\sigma^2}{2}\|\lambda\|^2_{H_t^0(\theta^*)}\right)\mathbb{E}\left[\exp\left(\partial\mathcal{L}\left(Y_t, \langle\theta^*, X_t\rangle\right)\lambda^\top X_t - \frac{\sigma^2}{2}\partial^2\mathcal{L}\left(Y_t, \langle\theta^*, X_t\rangle\right)\left(\lambda^\top X_t\right)^2\right)\middle|\mathcal{F}_t\right] \\
&\leqslant \exp\left(\lambda^\top S_t - \frac{\sigma^2}{2}\|\lambda\|^2_{H_t^0(\theta^*)}\right) \quad \text{(Assumption 3)} \\
&= M_t^\lambda\,.
\end{aligned}$$

Now we construct a new supermartingale by mixing all the M^λ. More formally, let Λ a $\mathbb{R}^d$-valued random variable independent of the rest and define $M_t = \mathbb{E}\left[M_t^\Lambda\middle|\mathcal{F}_\infty\right]$ where $\mathcal{F}_\infty = \sigma\left(\bigcup_{t\in\mathbb{N}}\mathcal{F}_t\right)$. If Λ has density p with respect to the Lebesgue measure, this means that $M_t = \int_{\mathbb{R}^d} M_t^\lambda p(\lambda)d\lambda$. For the choice $\Lambda \sim \mathcal{N}(0, \frac{1}{\beta\sigma^2}I_d)$ with $\beta > 0$, we have, by completing the square in the exponential:

$$\begin{aligned}
M_t &= \frac{(\beta\sigma^2)^{d/2}}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\exp\left(-\lambda^\top S_t + \frac{\sigma^2}{2}\left(\lambda^\top\left(H_t^0(\theta^*) + \beta I_d\right)\lambda\right)\right)d\lambda \\
&= \frac{(\beta\sigma^2)^{d/2}}{(2\pi)^{d/2}}\exp\left(\frac{\sigma^2}{2}\bar\lambda^\top H_t^\beta(\theta^*)\bar\lambda\right)\int_{\mathbb{R}^d}\exp\left(-\frac{\sigma^2}{2}\left(\lambda - \bar\lambda\right)^\top H_t^\beta(\theta^*)\left(\lambda - \bar\lambda\right)\right)d\lambda \\
&= \left(\frac{\beta^d}{\det H_t^\beta(\theta^*)}\right)^{\frac{1}{2}}\exp\left(\frac{\sigma^2}{2}\bar\lambda^\top H_t^\beta(\theta^*)\bar\lambda\right)\,,
\end{aligned}$$

where $\bar\lambda = \frac{1}{\sigma^2}H_t^\beta(\theta^*)^{-1}S_t$ and $H_t^\beta(\theta) = H_t^0(\theta) + \beta I_d$ is the regularized Hessian, which is positive definite and hence invertible. This expression further simplifies to $M_t = \left(\frac{\det\beta I_d}{\det H_t^\beta(\theta^*)}\right)^{\frac{1}{2}}\exp\left(\frac{1}{2\sigma^2}\|S_t\|^2_{H_t^\beta(\theta^*)^{-1}}\right)$.

From there, the argument is standard: M^λ is a nonnegative supermartingale, and therefore the pointwise limit $M_\infty^\lambda = \lim_{t\to+\infty}M_t^\lambda$ exists almost surely (Doob's supermartingale convergence theorem, Chapter 11 in (Williams, 1991)). Therefore for any $\mathcal{F}$-stopping time τ, M_τ^λ is well-defined, and thus so is and M_τ. By Fatou's lemma and Doob's stopping theorem, we have that $\mathbb{E}[M_\tau] = \mathbb{E}[\liminf_{t\to+\infty} \mathbb{E}[M_{t\wedge\tau} \mid \mathcal{F}_\infty]] \leqslant \liminf_{t\to+\infty} \mathbb{E}[\mathbb{E}[M_{t\wedge\tau} \mid \mathcal{F}_\infty]] \leqslant 1$. Finally, the particular choice of $\tau = \inf\left\{t \in \mathbb{N},\ \|S_t\|^2_{H_t^\beta(\theta^*)^{-1}} \geqslant \sigma^2\left(2\log\frac{1}{\delta} + \log\frac{\det \beta I_d}{\det H_t^\beta(\theta^*)}\right)\right\}$ and a straightforward application of Markov's inequality reveals that

$$\mathbb{P}\left(\tau < \infty\right) = \mathbb{P}\left(\exists t \in \mathbb{N},\ M_\tau \geqslant \frac{1}{\delta}\right) \leqslant \mathbb{E}[M_\tau]\delta \leqslant \delta\,,$$

which is exactly the expected result. ■

Appendix D. Assumption 7 $\implies$ Assumption 6

First, we recall the two assumptions of interest.

Assumption 6 (Transportation of local metrics) *For any $\alpha > 0$, there exists $\kappa > 0$ and $\beta > 0$ such that*

$$\bar{H}_t^\alpha(\theta^*, \bar{\theta}_t) \succcurlyeq \frac{1}{\kappa} H_t^\beta(\theta^*) \quad \text{and} \quad \bar{H}_t^\alpha(\theta^*, \bar{\theta}_t) \succcurlyeq \frac{1}{\kappa} H_t^\beta(\bar{\theta}_t)\,.$$

Assumption 7 (Bounded loss curvature) *There exists $m, M > 0$ such that*

$$\forall y, \xi \in \mathbb{R},\ m \leqslant \partial^2\mathcal{L}(y, \xi) \leqslant M\,.$$

We call the parameter $\kappa = \frac{M}{m}$ the conditioning of the convex loss $\mathcal{L}$.

Now simple calculations show that:

$$\begin{aligned}
\bar{H}_t^\alpha(\theta^*, \bar{\theta}_t) &= \sum_{s=1}^{t-1} \int_0^1 \partial^2\mathcal{L}(Y_s, \langle u\theta^* + (1-u)\bar{\theta}_t, X_s\rangle) du X_s X_s^\top + \alpha I_d \\
&= \sum_{s=1}^{t-1} \int_0^1 \partial^2\mathcal{L}(Y_s, \langle \theta^*, X_s\rangle) \frac{\partial^2\mathcal{L}(Y_s, \langle u\theta^* + (1-u)\bar{\theta}_t, X_s\rangle)}{\partial^2\mathcal{L}(Y_s, \langle \theta^*, X_s\rangle)} du X_s X_s^\top + \alpha I_d \\
&\succcurlyeq \frac{m}{M} \sum_{s=1}^{t-1} \partial^2\mathcal{L}(Y_s, \langle \theta^*, X_s\rangle) X_s X_s^\top + \alpha I_d \\
&= \frac{1}{\kappa}\left(\sum_{s=1}^{t-1} \partial^2\mathcal{L}(Y_s, \langle \theta^*, X_s\rangle) X_s X_s^\top + \kappa\alpha I_d\right) \\
&= \frac{1}{\kappa} H_t^{\kappa\alpha}(\theta^*)\,,
\end{aligned}$$

which is the desired result if $\beta = \kappa\alpha$. The other inequality with $\bar{H}_t^\beta(\bar{\theta}_t)$ is derived similarly.

Appendix E. Proof of Theorem 11

In this section, we prove the main regret theorem for LinUCB with convex risk, which we restate below.

Theorem 11 (Regret of LinUCB for convex risk) *Let $\delta \in (0, 1)$, $\alpha \geqslant \max(1, L^2)$ and define for $t \in \mathbb{N}$ the exploration bonus*

$$\gamma_t \colon x \in \mathcal{X}_t \mapsto c_t^\delta \|x\|_{H_t^{\kappa\alpha}(\bar{\theta}_t)^{-1}} \quad \text{and} \quad c_t^\delta = 2\kappa\left(\sigma\sqrt{2\log\frac{1}{\delta} + d\log\frac{m}{\alpha} + \log\det V_t^{\frac{\alpha}{m}}} + \sqrt{\frac{\alpha}{\kappa}} S\right)$$

and the projection operator $\Pi \colon \widehat{\theta} \in \mathbb{R}^d \mapsto \operatorname{argmin}_{\theta \in \Theta} \|F_t^\alpha(\theta) - F_t^\alpha(\widehat{\theta})\|_{H_t^\beta(\theta)^{-1}}$.

Under Assumptions 3, 7, 9 and 10, with probability at least $1-\delta$*, the regret of Algorithm 1 is bounded by*

$$\mathcal{R}_T \leqslant 2c_T^\delta \sqrt{\frac{2Td}{m}\log\left(1+\frac{TL^2}{d\alpha}\right)}\,.$$

In particular, we have $\mathcal{R}_T = \mathcal{O}\left(\frac{\kappa\sigma d}{\sqrt{m}}\sqrt{T}\log\frac{TL^2}{d}\right)$ *when* $T \to +\infty$.

Proof We will prove the regret bound in two steps. First, we justify the choice of exploration sequence $(\gamma_t)_{t\in\mathbb{N}}$, which naturally derives from the optimistic principle and the analysis of local metrics. Then, we use a somewhat crude bound on the Hessian to simplify the analysis and reduce it to the so-called elliptic potential lemma.

Indeed, as established in Section 3.2, the cumulative regret up to time T, denoted by $\mathcal{R}_T$, is upper bounded with probability at least $1-\delta$ by $2\sum_{t=1}^T \gamma_t(X_t)$ provided that $\mathbb{P}\left(\forall t \leqslant T,\ \Delta(X_t, \bar{\theta}_t) \leqslant \gamma_t(X_t)\right) \geqslant 1-\delta$, where

$$\Delta(X_t, \theta) = |\langle \theta^* - \theta, X_t\rangle| \leqslant \|\theta^* - \bar{\theta}_t\|_{\bar{H}_t^\alpha(\theta^*,\bar{\theta}_t)} \|X_t\|_{\bar{H}_t^\alpha(\theta^*,\bar{\theta}_t)^{-1}}\,.$$

Tuning of the exploration bonus sequence The transportation of local metrics (Assumption 6, implied by the curvature bound of Assumption 7) reveals that

$$\begin{aligned}\|\theta^* - \bar{\theta}_t\|_{\bar{H}_t^\alpha(\theta^*,\bar{\theta}_t)} &\leqslant \|F_t^\alpha(\theta^*) - F_t^\alpha(\widehat{\theta}_t)\|_{\bar{H}_t^\alpha(\theta^*,\bar{\theta}_t)^{-1}} + \|F_t^\alpha(\bar{\theta}_t) - F_t^\alpha(\widehat{\theta})\|_{\bar{H}_t^\alpha(\theta^*,\bar{\theta}_t)^{-1}} \\ &\leqslant \sqrt{\kappa}\left(\|F_t^\alpha(\theta^*) - F_t^\alpha(\widehat{\theta}_t)\|_{H_t^\beta(\theta^*)^{-1}} + \|F_t^\alpha(\bar{\theta}_t) - F_t^\alpha(\widehat{\theta})\|_{\bar{H}_t^\beta(\bar{\theta}_t)^{-1}}\right)\,.\end{aligned}$$

Thanks to the supermartingale control of Assumption 3, we deduce from Corollary 5 that with probability at least $1-\delta$, the following inequalities hold for all $t \leqslant T$:

$$\begin{aligned}\|F_t^\alpha(\theta^*) - F_t^\alpha(\widehat{\theta}_t)\|_{H_t^\beta(\theta^*)^{-1}} &\leqslant \sigma\sqrt{2\log\frac{1}{\delta} + \log\frac{\det H_t^\beta(\theta^*)}{\det \beta I_d}} + \alpha\|\theta^*\|_{H_t^\beta(\theta^*)^{-1}}\,, \\ \|F_t^\alpha(\bar{\theta}_t) - F_t^\alpha(\widehat{\theta}_t)\|_{H_t^\beta(\bar{\theta}_t)^{-1}} &\leqslant \sigma\sqrt{2\log\frac{1}{\delta} + \log\frac{\det H_t^\beta(\bar{\theta}_t)}{\det \beta I_d}} + \alpha\|\bar{\theta}_t\|_{H_t^\beta(\bar{\theta}_t)^{-1}}\,.\end{aligned}$$

The prior bound on parameters (Assumption 9) yields $\|\theta\|_{H_t^\beta(\theta)^{-1}} \leqslant \frac{S}{\sqrt{\beta}}$ for $\theta \in \{\theta^*, \bar{\theta}_t\}$. Furthermore, the curvature bound (Assumption 7) implies that $H_t^\beta(\theta) \preccurlyeq MV_t^{\beta/M}$, and therefore $\det H_t^\beta(\theta) \leqslant M^d \det V_t^{\beta/M}$ for $\theta \in \{\theta^*, \bar{\theta}_t\}$. Combining this together and substituting the expression of $\beta = \kappa\alpha$, where $\kappa = \frac{M}{m}$ the conditioning of the convex loss $\mathcal{L}$, we obtain:

$$\|\theta^* - \bar{\theta}_t\|_{\bar{H}_t^\alpha(\theta^*,\bar{\theta}_t)} \leqslant 2\sqrt{\kappa}\left(\sigma\sqrt{2\log\frac{1}{\delta} + d\log\frac{m}{\alpha} + \log\det V_t^{\frac{\alpha}{m}}} + \sqrt{\frac{\alpha}{\kappa}}S\right)\,.$$

Finally, the same arguments show that $\bar{H}_t^\alpha(\theta^*, \bar{\theta}_t)^{-1} \preccurlyeq \kappa H_t^{\kappa\alpha}(\bar{\theta}_t)^{-1}$, therefore $\|X_t\|_{\bar{H}_t^\alpha(\theta^*,\bar{\theta}_t)^{-1}} \leqslant \sqrt{\kappa}\|X_t\|_{H_t^{\kappa\alpha}(\bar{\theta}_t)^{-1}}$. This shows that

$$\gamma_t(X_t) \coloneqq \underbrace{2\kappa\left(\sigma\sqrt{2\log\frac{1}{\delta} + d\log\frac{m}{\alpha} + \log\det V_t^{\frac{\alpha}{m}}} + \sqrt{\frac{\alpha}{\kappa}}S\right)}_{=:\, c_t^\delta}\|X_t\|_{H_t^{\kappa\alpha}(\bar{\theta}_t)^{-1}}$$

is a valid choice of exploration sequence.

Bounding the regret Going back to the cumulative regret $\mathcal{R}_T$, we notice that $(c_t^\delta)_{t=1,\dots,T}$ is a positive, nondecreasing sequence, therefore we have with probability at least $1-\delta$ that

$$\mathcal{R}_T \leqslant 2\sum_{t=1}^T \gamma_t(X_t) \leqslant 2c_T^\delta \sum_{t=1}^T \|X_t\|_{H_t^{\kappa\alpha}(\bar{\theta}_t)^{-1}} .$$

A priori, the direct analysis of the right-hand side is tedious due to the dependency on $\bar{\theta}_t$ in the local metric. However, we notice that the curvature bound (Assumption 7) also implies the weaker control $H_t^{\kappa\alpha}(\bar{\theta}_t)^{-1} \preccurlyeq \frac{1}{m}(V_t^{\frac{\kappa\alpha}{m}})^{-1}$, which translates to $\|X_t\|_{H_t^{\kappa\alpha}(\bar{\theta}_t)^{-1}} \leqslant \frac{1}{\sqrt{m}}\|X_t\|_{(V_t^{\frac{\kappa\alpha}{m}})^{-1}}$. This bound is less informative as it looses the local information carried by $\bar{\theta}_t$, but still sufficient to obtain sublinear regret growth. We recall the following result, which is a direct consequence of the deterministic elliptic potential lemma (Lemma 11, Abbasi-Yadkori et al. (2011)) and the Cauchy-Schwarz inequality.

Lemma 18 *Let $(x_t)_{t\in\mathbb{N}}$ denote an arbitrary sequence of vectors in Euclident ball of $\mathbb{R}^d$ with radius L, $\alpha \geqslant \max(1, L^2)$ and $v_t = \sum_{s=1}^{t-1} x_s x_s^\top + \alpha I_d \in \mathbb{R}^{d\times d}$ for $t \in \mathbb{N}$. Then*

$$\sum_{s=1}^t \|x_s\|_{v_s^{-1}} \leqslant \sqrt{2td\log\left(1+\frac{tL^2}{d\alpha}\right)} .$$

Note that this result holds in our case thanks to the prior bound on actions (Assumption 10).

Conclusion With high probability, the regret of LinUCB with convex risk is bounded by

$$\mathcal{R}_T \leqslant 2\sum_{t=1}^T \gamma_t(X_t) \leqslant 2c_T^\delta\sqrt{\frac{2Td}{m}\log\left(1+\frac{TL^2}{d\alpha}\right)} .$$

Going back to the expression of c_T^δ, it follows from simple algebra (see e.g proof of Lemma 19.4, Lattimore and Szepesvári (2020)) that $\det V_t^{\frac{\alpha}{m}} \leqslant \left(\frac{\alpha}{m} + \frac{TL^2}{d}\right)^d$, and thus $c_T^\delta = \mathcal{O}\left(\kappa\sigma\sqrt{d\log\frac{TL^2}{d}}\right)$ when $T \to +\infty$. A simpler asymptotic bound on the regret is therefore

$$\mathcal{R}_T = \mathcal{O}\left(\frac{\kappa\sigma d}{\sqrt{m}}\sqrt{T}\log\frac{TL^2}{d}\right) .$$

■

Appendix F. Proofs of OGD concentration

Proposition 13 (OGD regret with sub-Gaussian gradients) *Let $\mathcal{C}$ a convex subset of $\mathbb{R}^d$ and Π the projection operator onto $\mathcal{C}$. For $j = 1,\dots,N$, let $\ell_j\colon \mathcal{C} \longrightarrow \mathbb{R}_+$ a twice differentiable convex function and $a, A > 0$ such that $aI_d \preccurlyeq \nabla^2\ell_j(z) \preccurlyeq AI_d$ for all $z \in \mathcal{C}$. Define the OGD update at step j by $z_j = \Pi(z_{j-1} - \varepsilon_{j-1}\nabla\ell_j(z_{j-1}))$ and $\bar{z}_n = \arg\min_{z\in\mathcal{C}} \sum_{j=1}^n \ell_j(z)$. Assume that there exists $z^* \in \mathcal{C}$ such that $\nabla\ell_j(z^*) = g_j + \frac{\alpha}{n}z^*$ with $\alpha \geqslant 0$ and g a centered, $\mathbb{R}^d$-valued σ-sub-Gaussian process, and also that $\mathcal{C}$ is bounded, i.e diam$(\mathcal{C}) = \sup_{z,z'\in\mathcal{C}}\|z-z'\| < \infty$. Then with probability at least $1-\delta$, the OGD regret with step size $\varepsilon_j = \frac{3}{aj}$ is bounded for all $n \leqslant N$ by*

$$\sum_{j=1}^n \ell_j(z_j) - \ell_j(\bar{z}_n) \leqslant \frac{9}{2a}\left(2d\sigma^2\log\frac{2dN}{\delta} + A^2\mathit{diam}(\mathcal{C})^2 + \frac{\alpha^2}{n^2}\|z^*\|^2\right)(1+\log n) .$$

This can be written more concisely as $\sum_{j=1}^N \ell_j(z_j) - \ell_j(\bar{z}_N) = \mathcal{O}(\frac{d\sigma^2}{a}\log^2 N)$ when $N \to +\infty$. In addition, if g is uniformly bounded by a constant $G > 0$, the regret with step size $\varepsilon_s = \frac{1}{aj}$ can be reduced to the almost sure bound:

$$\sum_{j=1}^n \ell_j(z_j) - \ell_j(\bar{z}_n) \leqslant \frac{G^2}{2a}(1+\log n) .$$

Proof Let $j \leqslant n$. The uniform lower bound on the Hessian of ℓ_j makes it a-strongly convex, which implies

$$\ell_j(z_j) - \ell_j(\bar{z}) \leqslant \langle \nabla \ell_j(z_j), z_j - \bar{z}_n \rangle - \frac{a}{2} \|\bar{z} - z_j\|^2.$$

By definition of the OGD scheme, the following holds:

$$\begin{aligned}
\|z_{j+1} - \bar{z}_n\|^2 &= \|\Pi\left(z_j - \varepsilon_j \nabla \ell_j(z_j)\right) - \bar{z}_n\|^2 \\
&\leqslant \|z_j - \varepsilon_j \nabla \ell_j(z_j) - \bar{z}_n\|^2 \quad \text{(projection onto a convex set)} \\
&\leqslant \|z_j - \bar{z}_n\|^2 + \varepsilon_j^2 \|\nabla \ell_j(z_j)\|^2 - 2\varepsilon_j \langle \nabla \ell_j(z_j), z_j - \bar{z}_n \rangle \,,
\end{aligned}$$

from which we deduce

$$\langle \nabla \ell_j(z_j), z_j - \bar{z}_n \rangle \leqslant \frac{\|z_j - \bar{z}_n\|^2 - \|z_{j+1} - \bar{z}_n\|^2}{2\varepsilon_j} + \frac{\varepsilon_j}{2} \|\nabla \ell_j(z_j)\|^2 \,.$$

Bounded gradients

This case is covered by Theorem 3.3 in Hazan (2019). We reproduce the proof here for reference and as a first step towards the more general setting of sub-Gaussian gradients.

Let $G > 0$ be such that $\|\nabla \ell_j(z_j)\| \leqslant G$ for all $j = 1, \ldots, n$. This allows to upper bound the above equation, leading to

$$\langle \nabla \ell_j(z_j), z_j - \bar{z}_n \rangle \leqslant \frac{\|z_j - \bar{z}_n\|^2 - \|z_{j+1} - \bar{z}_n\|^2}{2\varepsilon_j} + \frac{\varepsilon_j}{2} G^2.$$

The online regret of OGD is therefore

$$\sum_{j=1}^{n} \ell_j(z_j) - \ell_j(\bar{z}_n) \leqslant \frac{1}{2} \sum_{j=1}^{n} \frac{\|z_j - \bar{z}_n\|^2 - \|z_{j+1} - \bar{z}_n\|^2}{\varepsilon_j} - a\|z_j - \bar{z}_n\|^2 + \frac{G^2}{2} \sum_{j=1}^{n} \varepsilon_j.$$

The first sum can be rewritten after a simple index shift and the convention $1/\varepsilon_0 \coloneqq 0$:

$$\begin{aligned}
\frac{1}{2} \sum_{j=1}^{n} \frac{\|z_j - \bar{z}_n\|^2 - \|z_{j+1} - \bar{z}_n\|^2}{\varepsilon_j} - a\|z_j - \bar{z}_n\|^2 &= \frac{1}{2} \sum_{j=1}^{n} \|z_j - \bar{z}_n\|^2 \left(\frac{1}{\varepsilon_j} - \frac{1}{\varepsilon_{j-1}} - a \right) - \frac{1}{\varepsilon_n} \|z_{n+1} - \bar{z}_n\|^2 \\
&\leqslant \frac{1}{2} \sum_{j=1}^{n} \|z_j - \bar{z}_n\|^2 \left(\frac{1}{\varepsilon_j} - \frac{1}{\varepsilon_{j-1}} - a \right) \\
&= 0
\end{aligned}$$

for the choice $\varepsilon_j = \frac{1}{aj}$. Consequently, the online regret can be simplified as

$$\begin{aligned}
\sum_{j=1}^{n} \ell_j(z_j) - \ell_j(\bar{z}_n) &\leqslant \frac{G^2}{2} \sum_{j=1}^{n} \varepsilon_j \\
&= \frac{G^2}{2a} \sum_{j=1}^{n} \frac{1}{j} \\
&\leqslant \frac{G^2}{2a} \left(1 + \log n\right) \,.
\end{aligned}$$

Sub-Gaussian gradients

We do not assume here that $\nabla\ell_j(z_j)$ is uniformly bounded, but instead rely on the weaker assumption that $\nabla\ell_j(z^*)$ is sub-Gaussian. The strategy is to control the variation between $\nabla\ell_j(z_j)$ and $\nabla\ell_j(z^*)$ on the one hand, and bound in high probability $\nabla\ell_j(z^*)$ on the other hand.

Notice that $\nabla\ell_j(z_j) = g_j + \frac{\alpha}{n}z^* + \nabla\ell_j(z_j) - \nabla\ell_j(z^*)$ and that there exists $\bar{z}_n \in [z_j, z^*] \subset \mathcal{C}$ such that $\nabla\ell_j(z_j) - \nabla\ell_j(z^*) = \nabla^2\ell_j(\bar{z}_n)\,(z_j - z^*)$ thanks to the mean value theorem and the convexity of $\mathcal{C}$. This yields

$$\begin{aligned}\|\nabla\ell_j(\phi_j)\|^2 &\leqslant 3\|g_j\|^2 + \frac{3\alpha^2}{n^2}\|z^*\|^2 + 3\|\nabla\ell_j(z_j) - \nabla\ell_j(z^*)\|^2 \\ &\leqslant 3\|g_j\|^2 + \frac{3\alpha^2}{n^2}\|z^*\|^2 + 3A^2\|z_j - z^*\|^2\,,\end{aligned}$$

since $\nabla\ell_j$ is A-Lipschitz. Combining this with the above yields

$$\begin{aligned}\langle\nabla\ell_j(z_j), z_j - \bar{z}_n\rangle &\leqslant \frac{3}{2}\frac{\|z_j - \bar{z}_n\|^2 - \|z_{j+1} - \bar{z}_n\|^2}{\varepsilon_j} + \frac{3}{2}\varepsilon_j\|g_j\|^2 + \frac{3}{2}\varepsilon_j\frac{\alpha^2}{n^2}\|z^*\|^2 + \frac{3}{2}\varepsilon_j A^2\|z_j - z^*\|^2 \\ &\leqslant \frac{3}{2}\frac{\|z_j - \bar{z}_n\|^2 - \|z_{j+1} - \bar{z}_n\|^2}{\varepsilon_j} + \frac{3}{2}\varepsilon_j\left(\|g_j\|^2 + \frac{\alpha^2}{n^2}\|z^*\|^2\right) + \frac{3}{2}\varepsilon_j A^2\mathrm{diam}(\mathcal{C})^2\,.\end{aligned}$$

The online regret of OGD is therefore

$$\begin{aligned}\sum_{j=1}^n \ell_j(z_j) - \ell_j(\bar{z}_n) \leqslant &\frac{3}{2}\sum_{j=1}^n \frac{\|z_j - \bar{z}_n\|^2 - \|z_{j+1} - \bar{z}_n\|^2}{\varepsilon_j} - \frac{a}{3}\|z_j - \bar{z}_n\|^2 \\ &+ \frac{3}{2}A^2\mathrm{diam}(\mathcal{C})^2\sum_{j=1}^n\varepsilon_j + \frac{3}{2}\sum_{j=1}^n\varepsilon_j\left(\|g_j\|^2 + \frac{\alpha^2}{n^2}\|z^*\|^2\right)\,.\end{aligned}$$

As in the bounded case, the choice $\varepsilon_j = \frac{3}{aj}$ makes the first sum vanish. Moreover, a simple union argument over the Chernoff bound for the σ-sub-Gaussian random variables $(g_j)_{j=1,\dots,n}$ reveals that

$$\mathcal{E}_n = \left\{\forall j = 1,\dots,n,\ \|g_j\| \leqslant \sigma\sqrt{2d\log\frac{2dn}{\delta}}\right\}$$

holds with probability at least $1-\delta$ for $\delta \in (0,1)$. Therefore, the following holds with probability at least $1-\delta$:

$$\sum_{j=1}^n \varepsilon_j\|g_j\|^2 \leqslant \sum_{j=1}^n \varepsilon_j\|g_j\|^2\mathbb{I}_{\mathcal{E}_n} \leqslant 2d\sigma^2\log\frac{2dn}{\delta}\sum_{j=1}^n\varepsilon_j\,.$$

Therefore, with probability at least $1-\delta$, we obtain the following online regret:

$$\begin{aligned}\sum_{j=1}^n \ell_j(z_s) - \ell_j(\bar{z}_n) &\leqslant \frac{3}{2}\left(2d\sigma^2\log\frac{2dn}{\delta} + A^2\mathrm{diam}(\mathcal{C})^2 + \frac{\alpha^2}{n^2}\|z^*\|^2\right)\sum_{j=1}^n\varepsilon_j \\ &= \frac{9}{2a}\left(2d\sigma^2\log\frac{2dn}{\delta} + A^2\mathrm{diam}(\mathcal{C})^2 + \frac{\alpha^2}{n^2}\|z^*\|^2\right)\sum_{j=1}^n\frac{1}{j} \\ &\leqslant \frac{9}{2a}\left(2d\sigma^2\log\frac{2dn}{\delta} + A^2\mathrm{diam}(\mathcal{C})^2 + \frac{\alpha^2}{n^2}\|z^*\|^2\right)(1 + \log n)\,.\end{aligned}$$

■

Appendix G. Proofs of Theorem 14

In this section, we adapt the regret analysis of LinUCB to the LinUCB-OGD variant that relies on online gradient approximation of the empirical risk minimizer.

Theorem 14 (Regret of LinUCB-OGD for convex risk) *In addition to the conditions of Theorem 11, assume that there exists $\varepsilon_h > 0$ such that for all $j \leqslant T/h$, $\sum_{k=1}^h X_{(j-1)h+k} X_{(j-1)h+k}^\top \succcurlyeq \varepsilon_h I_d$, and also that $\partial\mathcal{L}(Y_t, \langle\theta^*, X_t\rangle)$ is $\sqrt{m}\sigma$-sub-Gaussian for all $t \leqslant T$. Let $C > 0$ and define the exploration bonus sequence by:*

$$\gamma_{t,T}^{OGD} : x \in \mathcal{X}_t \mapsto c_{t,T}^{OGD,\delta} \|x\|_{H_t^{\kappa\alpha}(\bar{\theta}^{OGD}_{\lfloor \frac{t-1}{h} \rfloor})^{-1}} \quad \text{and} \quad c_{t,T}^{OGD,\delta} = c_t^\delta + 2L\sqrt{\frac{C\kappa m d h^2 \sigma^2}{\varepsilon_h t} \log\left(\frac{2dT}{h\delta}\right) \log\left(\frac{t}{h}\right)} .$$

Then if $C > 0$ is large enough, with probability at least $1 - 2\delta$, the regret of Algorithm 2 with the OGD step sequence of Proposition 13 is bounded $\tilde{\mathcal{O}}\left(\sqrt{T}\right)$, where $\tilde{\mathcal{O}}$ hides polylogarithmic terms in T.

Proof

Let $\ell_j(\theta) = \sum_{k=1}^h \mathcal{L}(Y_{(j-1)h+k}, \langle\theta, X_{(j-1)h+k}\rangle) + \frac{\alpha}{2n}\|\theta\|^2$. For simplicity, we assume that $\bar{\theta}_t = \widehat{\theta}_t$ for all $t \leqslant T$, i.e the empirical risk minimizer is always in the stable set of the projection operator Π. We recall that $\widehat{\theta}_t$ satisfies $\sum_{j=1}^n \nabla\ell_j(\bar{\theta}_t) = 0$ for $n = \frac{t-1}{h}$ (i.e after episode n, when $\widehat{\theta}_n^{\text{OGD}}$ is updated). In the general case, replacing $\widehat{\theta}_t$ by $\bar{\theta}_t$ induces an extra correction factor in the inequalities below which is at most polylogarithmic in T (a consequence of Corollary 5), and hence does not change the conclusion. Again, we point out that, similarly to the generalized linear bandit setting (Filippi et al., 2010; Faury et al., 2020), $\widehat{\theta}_t$ is often in the stable set of Π in practice.

We use the notations of Proposion 13 and define:

$$z_j = \widehat{\theta}_j^{\text{OGD}} ,$$
$$\bar{z}_n = \bar{\theta}_t ,$$
$$z^* = \theta^* .$$

We also denote by $\tilde{z}_n = \bar{\theta}_n^{\text{OGD}} = \frac{1}{n}\sum_{j=1}^n z_j$ the average of the past n OGD updates.

Bound on $\|\tilde{z}_n - \bar{z}_n\|$ We first note that $\nabla\ell_j(\theta^*) = g_j + \frac{\alpha}{n}\theta^*$, where $g_j = \sum_{k=1}^h \partial\mathcal{L}(Y_{(j-1)h+k}, \langle\theta^*, X_{(j-1)h+k}\rangle)$ is $\sqrt{hm}\sigma$-sub-Gaussian (sum of h terms, each of them being $\sqrt{m}\sigma$-sub-Gaussian). The assumption on ε_h makes the one-step losses ℓ_j $m\varepsilon_h$ strongly convex. Under Assumption 7, we can then apply the bound on the OGD regret of Proposion 13 with $a = m\varepsilon_h$, $A = hML^2$, namely that the *good event*

$$\forall n \leqslant N,\ \sum_{j=1}^{n-1} \ell_j(z_j) - \ell_j(\bar{z}_n) \leqslant \frac{Cdh\sigma^2}{\varepsilon_h} \log\left(\frac{2dN}{\delta}\right) \log(n) ,$$

holds with probability at least $1 - \delta$, for some constant $C > 0$ (in which we hide the dependency on h, M, L, α and S to avoid further cluttering) and $N = \lceil \frac{T-1}{h} \rceil$ the maximum number of episodes. We assume to be on this event in the rest of the proof.

The crux of the argument is similar to the proof of Lemma 2 in (Ding et al., 2021) and exploits the strong convexity of the losses ℓ_j to relate the online regret to a control on the distance $\|\tilde{z}_n - \bar{z}_n\|$. By Jensen's inequality, we have

$$\sum_{j=1}^n \ell_j(\tilde{z}_n) - \ell_j(\bar{z}_n) \leqslant \frac{Cdh\sigma^2}{\varepsilon_h} \log\left(\frac{2dN}{\delta}\right) \log(n) .$$

Strong convexity also implies the following inequality:

$$\ell_j(\tilde{z}_n) - \ell_j(\bar{z}_n) \geqslant \langle\nabla\ell_j(\bar{z}_n), \tilde{z}_n - \bar{z}_n\rangle + \frac{m\varepsilon_h}{2}\|\tilde{z}_n - \bar{z}_n\|^2 .$$

Summing over $j = 1, \ldots, n$ and exploiting the fact that the sum of gradients vanishes at $\bar{z}_n$, we obtain after some simple algebra:

$$\|\tilde{z}_n - \bar{z}_n\| \leqslant \frac{Cdh\sigma^2}{\varepsilon_h} \log\left(\frac{2dN}{\delta}\right) \log(n)\,.$$

Regret analysis of LinUCB-OGD Mirroring the regret proof of LinUCB, we see that we need

$$\forall t \leqslant T,\ \Delta(X_t, \bar{\theta}_t^{\mathrm{OGD}}) \leqslant \gamma_t(X_t)$$

holds with high probability, for a certain exploration sequence $(\gamma_t)_{t\in\mathbb{N}}$. This amount to controlling the norm

$$\begin{aligned}
\|\theta^* - \bar{\theta}_t^{\mathrm{OGD}}\|_{\bar{H}_t^\alpha(\theta^*,\bar{\theta}_t^{\mathrm{OGD}})} &\leqslant \|\theta^* - \bar{\theta}_t\|_{\bar{H}_t^\alpha(\bar{\theta}_t,\bar{\theta}_t^{\mathrm{OGD}})} + \|\bar{\theta}_t - \bar{\theta}_t^{\mathrm{OGD}}\|_{\bar{H}_t^\alpha(\bar{\theta}_t,\bar{\theta}_t^{\mathrm{OGD}})} \\
&\leqslant \|\theta^* - \bar{\theta}_t\|_{\bar{H}_t^\alpha(\theta^*,\bar{\theta}_t^{\mathrm{OGD}})} + \sqrt{M}\|\bar{\theta}_t - \bar{\theta}_t^{\mathrm{OGD}}\|_{V_t^{\frac{\alpha}{m}}} \\
&\leqslant \|\theta^* - \bar{\theta}_t\|_{\bar{H}_t^\alpha(\theta^*,\bar{\theta}_t^{\mathrm{OGD}})} + L\sqrt{M}\|\bar{\theta}_t - \bar{\theta}_t^{\mathrm{OGD}}\|\,.
\end{aligned}$$

Therefore, on the good event, the OGD approximation of $\bar{\theta}_t$ induces an additional factor of at most

$$2L\sqrt{\frac{C\kappa m d h^2\sigma^2}{\varepsilon_h t} \log\left(\frac{2dT}{h\delta}\right) \log\left(\frac{t}{h}\right)}$$

in the exploration bonus sequence compared to the factor c_t^δ used in LinUCB. Finally, we invoke a standard union bound argument to ensure both the good event and the regret bound hold with probability at least $1 - 2\delta$.

Appendix H. Experiments

We report three simple experiments, two in the expectile setting and one in the entropic risk setting.

Computing expectiles We detail two cases of distributions for which expectiles are known. For $p \in (0,1)$, we denote by $e_p(\nu)$ the p-expectile of distribution ν.

- If $\nu = \mathcal{N}(0,1)$, then, letting ϕ and Φ be the pdf and cdf of ν respectively, we obtain after simple calculus and the identity $\phi'(y) = -y\phi(y)$ the following fixed point equation:

  $$e_p(\nu) = \frac{2p\phi(e_p(\nu)) - 1}{(1-2p)\Phi(e_p(\nu)) + p}\,,$$

  from which one can estimate the value of $e_p(\nu)$ using a fast iterative scheme. The general Gaussian case $\nu = \mathcal{N}(\mu, \sigma^2)$ is then easily deduced from the relation $e_p(\nu) = \mu + \sigma e_p(\mathcal{N}(0,1))$. Expectile calculations for a few other classical distributions are covered in Philipps (2022).

- If ν is the so-called expectile based distribution (Torossian et al., 2020; Arbel et al., 2021) with asymmetric density (with respect to the Lebesgue measure) given by

  $$f_{\mu,\sigma,p}(y) = \frac{\sqrt{2p(1-p)}}{\sigma\sqrt{\pi}(\sqrt{p} + \sqrt{1-p})} \exp\left(-\frac{|p - \mathbb{I}_{y<\mu}|(y-\mu)^2}{2\sigma^2}\right)\,,$$

  then $e_p(\nu) = \mu$. In other words, these distributions offer a family parametrized directly by their expectile, generalizing the family of Gaussian distributions parametrized by their mean (for a given variance).

We recall that the p-expectile can be elicited by the convex potential $\psi(z) = |p - \mathbb{I}_{z<0}|z^2$. The second derivative of this potential is given by $\psi''(z) = (1-p)\mathbb{I}_{z<0} + p\mathbb{I}_{z>0}$, which is bounded between p and $1-p$. In particular, Assumption 7 holds with conditioning $\kappa = \frac{M}{m} = \frac{1-p}{p}$ if $p \leqslant \frac{1}{2}$ and $\kappa = \frac{M}{m} = \frac{p}{1-p}$ otherwise. Note that the two classes of distributions considered above are Gaussian or simple, log-concave transforms of Gaussian distributions, which fits the scope of the supermartingale control of Assumption 3.

Computing entropic risk For a distribution ν, the entropic risk at level $\gamma > 0$ takes the form $\rho_\gamma(\nu) = \frac{1}{\gamma}\log \mathbb{E}_{Y\sim\nu}\left[e^{\gamma Y}\right]$ and corresponds to the loss $\mathcal{L}\colon (y,\xi) \mapsto \xi + \frac{1}{\gamma}(e^{\gamma(y-\xi)} - 1)$. Derivatives of this loss satisfy the following identities, where ∂ represents the differentiation operator with respect to the second coordinate ξ:

$$\partial\mathcal{L}(y,\xi) = 1 - e^{\gamma(y-\xi)},$$
$$\partial^2\mathcal{L}(y,\xi) = \gamma e^{\gamma(y-\xi)},$$

and is thus in particular strictly convex.

For a Bernoulli-like distribution $\nu = p\delta_a + (1-p)\delta_b$, with $p \in (0,1)$, $a, b \in \mathbb{R}$, the entropic risk takes the simple form $\rho_\gamma(\nu) = \frac{1}{\gamma}\log\left(pe^{\gamma a} + (1-p)e^{\gamma b}\right)$. If ν has a bounded support with diameter $\mathcal{D}$, then it is clear that the Hessian of the loss is controlled by $m = \gamma e^{-\gamma\mathcal{D}} \leqslant \partial^2\mathcal{L} \leqslant \gamma e^{\gamma\mathcal{D}} = M$, and therefore the conditioning number of the loss κ can be bounded by $e^{2\gamma\mathcal{D}}$. Finally, ν being bounded also fits the scope of the supermartingale control of Assumption 3.

General case If a density p and a loss function $\mathcal{L}$ are known, one may resort to numerical integration to approximate the following quantity up to arbitrary precision:

$$\mathbb{E}_{Y\sim\nu}\left[\mathcal{L}(Y,\xi)\right] = \int \mathcal{L}(y,\xi)p(y)dy \approx \sum_i w_i\mathcal{L}(y_i,\xi)p(y_i),$$

where the weights (w_i) and knots (y_i) depend on the approximation routine. Then, one may simply run a minimization algorithm on the function $\xi \mapsto \sum_i w_i\mathcal{L}(y_i,\xi)p(y_i)$ to estimate $\rho_{\mathcal{L}}(\nu)$.

Experiment 1: multi-armed Gaussian bandit with expectile noise We considered $K = 2$ Gaussian arms with expectiles at level $p = 10\%$ equal to 1 and 0 respectively. This bandit can be represented by constant orthonormal actions $\mathcal{X}_t = \{[1\ 0]^\top,\ [0\ 1]^\top\}$, parameter $\theta^* = [1\ 0]^\top$ and noise distributions $\mathcal{N}\left(\mu_k, \sigma_k^2\right)$, with μ_k and σ_k chosen such that the expectile of the corresponding noise is zero for $k \in \{1,2\}$. This can be achieved with e.g. $\mu_1 \approx 0.44$, $\sigma_1 = 0.5$ and $\mu_2 \approx 2.62$, $\sigma_2 = 3$, which was the setup for this experiment. Note that for a given expectile level $p \in (0,1)$ and standard deviation σ, finding the unique mean μ such that $\mathcal{N}\left(\mu, \sigma^2\right)$ has zero p-expectile can be easily done via a numerical root search, using the formula for Gaussian expectiles described above.

The optimal arm with respect to the expectile criterion is the first one by definition. However, the expectations of these arms are in reversed order, making the second one optimal with respect to the mean criterion.

Experiment 2: linear bandit with expectile asymmetric noise We considered a second example with non-Gaussian noise and non-orthogonal features. We defined the action set at time t by $\mathcal{X}_t = \{X_t^1, X_t^2\} \subset \mathbb{R}^3$ where:

- $X_t^1 = \frac{Z_t^1}{\|Z_t^1\|}$ with $Z_t^1 \sim \mathcal{N}([1\ 0\ 0]^\top, \sigma_x I_3)$,
- $X_t^2 = \frac{Z_t^2}{\|Z_t^2\|}$ with $Z_t^1 \sim \mathcal{N}([0\ 1\ 0]^\top, \sigma_x I_3)$,
- We set the action noise to an arbitrary value $\sigma_x = 0.1$.
- $(Z_t^1, Z_t^2)_{t\in\mathbb{N}}$ are all independent random variables.

This construction results in bounded, anisotropic actions. We chose $\theta^* = [0.9\ 0\ 1]^\top$, so that $\langle\theta^*, X_t^1\rangle$ is likely higher than $\langle\theta^*, X_t^2\rangle$, thus favoring $X_t = X_t^1$ in the expectile model $Y_t = \langle\theta^*, X_t\rangle + \eta_t$. To model the zero p-expectile noise η_t with $q = 10\%$, we used the expectile based distribution presented above with $\mu_1 = \mu_2 = 0$ and $\sigma_1 = 0.5$ if action X_t^1 is played, and $\sigma_2 = 1.5$ otherwise, resulting in different mean noise $\mathbb{E}[\eta_t|\mathcal{F}_t] \approx 1.8$ and $\mathbb{E}[\eta_t|\mathcal{F}_t] \approx 3.3$ respectively. As in the previous example, this setting was designed to deceive the mean criterion by inverting the order of optimal actions.

Experiment 3: multi-armed Bernoulli bandit with entropic risk noise The last experiment consisted of $K = 2$ Bernoulli-like arms $\nu_1 = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$ and $\nu_2 = \frac{1}{4}\delta_2 + \frac{3}{4}\delta_{-2}$, which corresponds to means $\mu_1 = 0$, $\mu_2 = -1$ and entropic risk $\rho_\gamma(\nu_1) \approx 0.43$ and $\rho_\gamma(\nu_2) \approx 0.67$ at level $\gamma = 1$. Again, this setting was designed so that the best optimal arm is different under the mean and entropic risk criteria.

Results On each of the three settings, we ran an instance of Algorithm 1, i.e. LinUCB (convex risk), and Algorithm 2, i.e. LincUCB-OGD (convex risk). We also ran a standard LinUCB algorithm for the mean criterion (Abbasi-Yadkori et al., 2011). Hyperparameters m, M and κ were tuned according to the analysis above. Regularization was fixed at $\lambda = 0.1$. As is customary in bandit experiments, the parameter σ, which in the formal analysis is derived from the supermartingale control of the noise, was considered a degree of freedom to control the amount of exploration; we arbitrarily fixed it at $\sigma = 0.1$ in experiments 1 and 2 and at $\sigma = 1$ in experiment 3. For the LinUCB-OGD variant, the step size for the OGD scheme was set to $\varepsilon_n = 0.1/n$, following the linear decay suggested by Proposition 13, and the frequency of OGD update to $h = 5$. In addition, all algorithms went through an initial warmup phase where each arm was played 5 times, in order to ensure better stability of the initial estimations of θ.

In all three examples, the mean criterion algorithm was deceived and accumulated linear expectile and entropic risk regret, while both risk-aware algorithms exhibited sublinear trends. Interestingly, the LinUCB-OGD variant showed higher regrets due to the approximate minimization of the loss criterion by OGD, but remained below the mean criterion LinUCB benchmark. Figure 2 reproduces the results of each experiments across 500 independent replications. Finally, average runtimes for each algorithm are reported in Table 3. Calculations were performed on a distributed infrastructure comprised of 80 CPUs. While the values themselves are not indicative, as they would vary on a different system, their relative magnitudes illustrate the computational gain of the OGD scheme over solving the empirical risk minimization problem at each step as required in LinUCB (convex risk). Note also that the standard LinUCB with mean criterion is faster due to the sequential nature of the ridge regression estimator. Indeed, this procedure involves inverting at each step a $d \times d$ matrix subject to rank one updates, which can be calculated efficiently via the Sherman-Morrison formula. By contrast, other convex losses than the one derived from the quadratic potential loose this sequential form and require solving the corresponding regression problem from scratch at each time step.

Table 3: Runtimes for the classical LinUCB and Algorithms 1 (LinUCB for convex risk) and 2 (LinUCB-OGD for convex risk) in each experiments. Runtimes are reported in seconds as mean $\pm$ standard deviation, estimated across 500 independent replications with time horizon $T = 1500$.

Algorithm	Experiment 1	Experiment 2	Experiment 3
LinUCB (mean)	0.4 ± 0.0	37.2 ± 4.9	0.6 ± 0.0
LinUCB (convex risk)	231.0 ± 21.7	814.8 ± 88.3	519.1 ± 33.3
LinUCB-OGD (convex risk)	20.4 ± 3.9	60.2 ± 12.0	25.7 ± 4.9

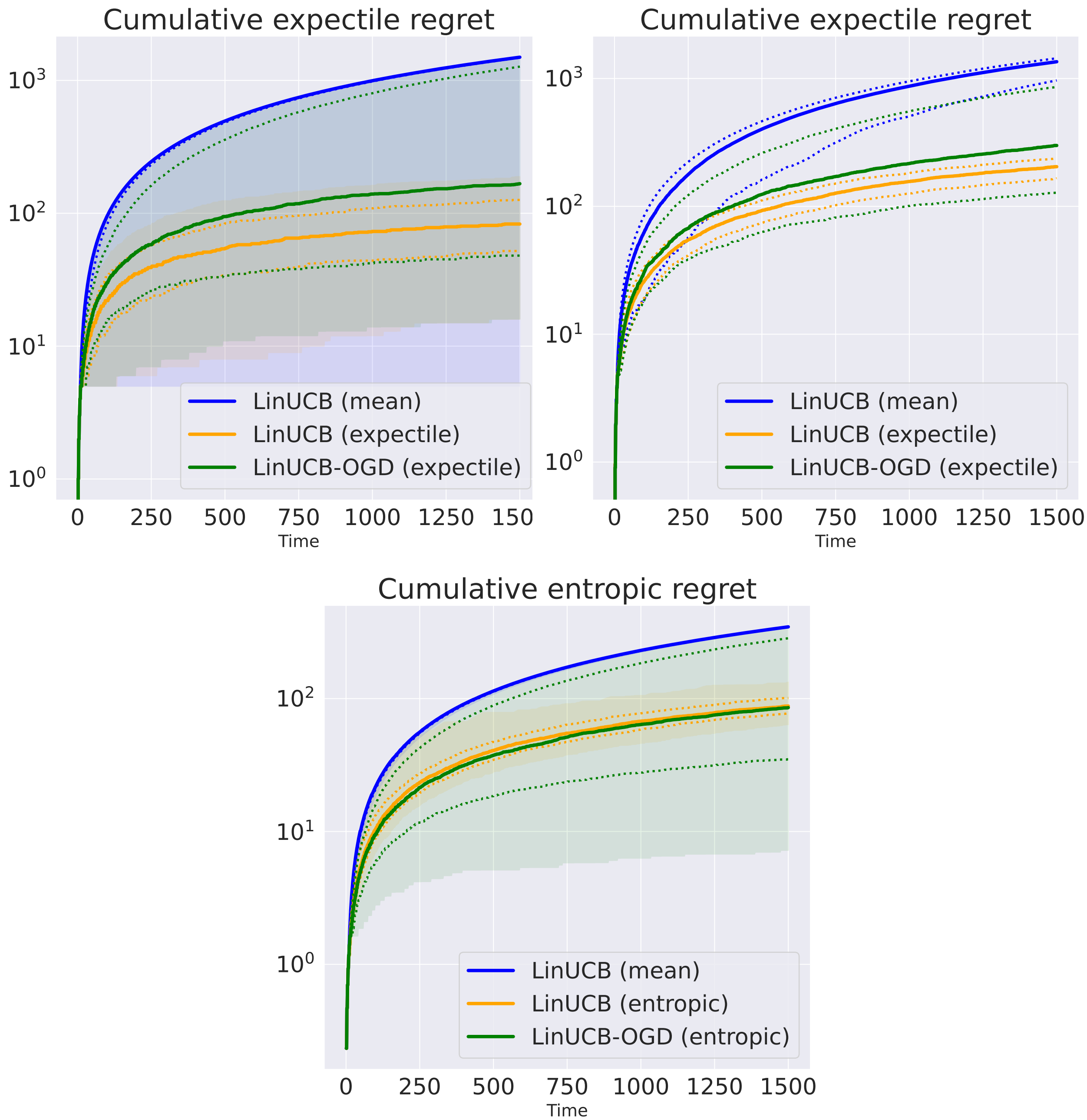

Figure 2: Left: two-armed Gaussian expectile bandit. Center: two-armed linear expectile bandit with $\mathbb{R}^3$ contexts and expectile-based asymmetric noises. Right: two-armed Bernoulli entropic risk bandit. Thick lines denote median cumulative regret over 500 independent replications. Dotted lines denote the 25 and 75 regret percentiles. Shaded areas denote the 5 and 95 percentiles.