Risk-aware linear bandits with convex loss

Patrick Saux, Odalric-Ambrym Maillard

Risk-aware linear bandits with convex loss

Patrick Saux
Inria, Univ. Lille, CNRS, Centrale Lille,
UMR 9198-CRIStAL, F-59000 Lille, France

Odalric-Ambrym Maillard
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.

1 INTRODUCTION

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 a policy in order to maximize rewards, facing the classical exploitation-exploration dilemma. A prominent example of such models is the linear bandit, which assumes a linear relationship actions and the *mean* rewards. In this setting, a standard learning strategy consists in estimating the reward model by ridge regression coupled with an appropriate exploration scheme, e.g., optimism (Abbasi-Yadkori et al., 2011), Thompson sampling (Agrawal and Goyal, 2013; Abeille and Lazaric, 2017) or information-directed (Russo and Van Roy, 2014; Kirschner et al., 2021).

One limitation of this setting is that real-world agents may assess rewards with a different criterion than the mean. While mathematically convenient, the latter is known to equally weight large positive and negative outcomes, possibly leading 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), which is a special case of spectral risk measures (Acerbi, 2002), entropic risk (Maillard, 2013) and the expectiles (Newey and Powell, 1987). These measures, 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; Gopalan et al., 2017; Cassel et al., 2018; Tamkin et al., 2019; Prashanth et al., 2020; Pandey et al., 2021; 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 definition covers quantiles, expectiles and entropic risk, and

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

can be extended to mean-variance and conditional value-at-risk using multivariate risk measures. To our knowledge, this setting is new, although related to existing approaches, such as bandits with regression oracles (Foster and Rakhlin, 2020) and generalized linear bandits (GLB) (Filippi et al., 2010; Li et al., 2017; Faury et al., 2020), that go beyond reward linearity while still working under the mean criterion.

Contributions We introduce a generalization of LinUCB to a large class of so-called *elicitable* risk measures, which includes the expectiles and the entropic risk. In contrast with the standard mean-linear bandit, learning the linear mapping between actions and risk measures cannot be performed sequentially, which presents theoretical and numerical challenges similar to GLB. We derive time-uniform confidence sets (Proposition 1) based on the method of mixtures (Peña et al., 2008) and introduce a geometric condition (Lemma 2) to ensure sublinear regret in this new setting (Theorem 1). Using recent developments on time-uniform matrix concentration, we further strengthen the regret bound in the case of stochastic actions with a known covariance lower bound (Theorem 2). To mitigate the numerical burden, we introduce an episodic version of LinUCB with online gradient descent approximation (Theorem 3), inspired by previous works on online regression (Korda et al., 2015) and the recent literature on approximate Thompson sampling for GLB (Ding et al., 2021).

Notations We consider the Euclidean space $\left(\mathbb{R}^d, \langle\cdot,\cdot\rangle\right)$ and denote by I_d the identity matrix of $\mathbb{R}^d$. For a positive definite matrix $P \in \mathcal{S}_d^{++}(\mathbb{R})$ and a vector $x \in \mathbb{R}^d$, we define the norm $\|x\|_P = \sqrt{\langle x, Px\rangle}$. When $P = I_d$, we let $\|\cdot\|_P = \|\cdot\|_2$ be the L^2 norm. $\mathcal{B}^d_{\|\cdot\|}(x, R)$ denotes the ball in $\mathbb{R}^d$ centred on x of radius R with respect to the norm $\|\cdot\|$. $A \preccurlyeq B$ stands $B - A \in \mathcal{S}_d(\mathbb{R})$ (positive semidefinite matrix). For $K \in \mathbb{N}$, $[K]$ denotes the set $\{1, \dots, K\}$. For a set E, 2^E denotes the set of all subsets of E.

2 CONTEXTUAL BANDITS WITH 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 $\mathcal{F}_t = \sigma\left(\mathcal{X}_1, X_1, Y_1, \dots, \mathcal{X}_{t-1}, X_{t-1}, Y_{t-1}, \mathcal{X}_t, X_t\right)$ 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 of the mapping between actions and rewards:

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

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 $\Theta \subseteq \mathbb{R}^d$ parameter space.

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 be averse to high reward volatility, which can be encoded by $\mu(x) - \lambda\sigma(x)$ for some $\lambda > 0$. We detail in Appendix A how many standard risk measures (entropic, p-expectile) realize this mean-variance tradeoff.

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). In most algorithms, the exploitation takes the form of a 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, that is $\min_{\theta\in\Theta} \sum_{s=1}^{t-1} \left(Y_s - \langle\theta, X_s\rangle\right)^2 + \frac{\alpha}{2}\|\theta\|_2^2$, where $\alpha > 0$ is a regularization parameter. Assuming for now the solution is in the interior of Θ, the solution can be written as $\widehat{\theta}_t = \left(V_t^\alpha\right)^{-1} \sum_{s=1}^{t-1} Y_s X_s$, where we define the $d \times d$ positive definite matrix $V_t^\alpha \coloneqq \sum_{s=1}^{t-1} X_s X_s^\top + \alpha I_d$. 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 $V_{t+1}^\alpha = V_t^\alpha + X_t X_t^\top$ starting from $V_0^\alpha = \alpha I_d$) 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 standard 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 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 $\mathbb{E}\left[\mathcal{L}\left(Y,\ell(x)\right)\right]$ 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 $\theta^*\in\arg\min_{\theta\in\Theta}\mathbb{E}_{Y\sim\varphi(\langle\theta^*,x\rangle)}\left[\mathcal{L}\left(Y,\langle\theta,x\rangle\right)\right]$.*

Table 1: Example of Risk Measures Elicited by Convex Potentials.

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

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 $\operatorname{argmin}_{\xi\in\mathbb{R}}\mathbb{E}_{Y\sim\nu}[\lvert p-\mathbb{I}_{Y<\xi}\rvert(Y-\xi)^2]$. 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 Lemma 4 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. 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 induced by ψ, which generalizes the quadratic potential. In fact, every continuously differentiable loss 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 indexed by differentiable, strictly convex functions (Table 3, Appendix A).

2.2 Contextual Bandits with Elicitable Risk Measures

Regret For a linear bandit (φ,θ^*), we define the pseudo-regret associated to a risk mea-

sure $\rho_{\mathcal{L}}$ and a sequence of actions $(X_t)_{1\leqslant t\leqslant T}$ as $\mathcal{R}_T = \sum_{t=1}^T \rho_{\mathcal{L}}\left(\varphi(\langle\theta^*, X_t^*\rangle)\right) - \rho_{\mathcal{L}}\left(\varphi(\langle\theta^*, X_t\rangle)\right)$, where $X_t^* = \operatorname{argmax}_{x\in\mathcal{X}_t} \rho_{\mathcal{L}}\left(\varphi(\langle\theta^*, x\rangle)\right)$ 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 now represents an elicitable risk measure for the reward distribution. As an example, this paves a 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 estimators of the mean, we define $\widehat{\theta}_t \in \operatorname{argmin}_{\theta\in\Theta} \sum_{s=1}^{t-1} \mathcal{L}(Y_s, \langle\theta, X_s\rangle) + \frac{\alpha}{2}\|\theta\|_2^2$, which corresponds to the empirical risk minimization associated to 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 the equation $\alpha\widehat{\theta}_t = -\sum_{s=1}^{t-1} \partial\mathcal{L}(Y_s, \langle\widehat{\theta}_t, X_s\rangle)X_s$, where $\partial\mathcal{L}(y,\xi)$ stands for the derivative of $\xi \mapsto \mathcal{L}(y,\xi)$. 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 (such an operator is detailed in Section 3.1). We also define $H_t^\alpha(\theta) = \sum_{s=1}^{t-1} \partial^2\mathcal{L}\left(Y_s, \langle\theta, X_s\rangle\right) X_s X_s^\top + \alpha I_d$ the Hessian of the empirical loss at θ of the minimization problem.

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) $Y_t = \mu(\langle\theta^*, X_t\rangle) + \eta_t$, with $\mathbb{E}[\eta_t|\mathcal{F}_t] = 0$ and μ a nonlinear link function. Under parametric assumptions on Y_t (typically one-dimensional exponential family), GLB can be seen as a special case of the risk-aware setting with $\mathcal{L}$ the negative log-likelihood loss, with the analogy $\mu \leftrightarrow \partial\mathcal{L}$. Despite this formal similarity, GLB is designed solely to optimize the mean criterion. Another difference with our setting is that regret for GLB is commonly defined as $\sum_{t=1}^T \mu(\langle\theta^*, X_t^*\rangle) - \mu(\langle\theta^*, X_t\rangle)$, which is smaller than $\mathcal{R}_T$ when μ is contracting.*

Extension of LinUCB to Convex Losses 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), Ch.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 (CR for Convex Risk) in Algorithm 1.

Algorithm 1 LinUCB-CR

Input: regularisation parameter α, projection Π, exploration bonus sequence $(\gamma_t)_{t\in\mathbb{N}}$.
Initialization: Observe $\mathcal{X}_1$.
for $t = 1, \dots, T$ **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 LinUCB-CR

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. To this end, we introduce the following control on the curvature of the adapted loss $\mathcal{L}$.

Assumption 1 (Bounded Loss Curvature) *There exists m and M 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 $\mathcal{L}$.

Remark 3 *This assumption is reminiscent of the standard lower bound on the derivative of the link function μ' commonly encountered in the GLB literature.*

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. Instead, for a given loss $\mathcal{L}$, we know that $\theta^* = \operatorname{argmin}_\theta \mathbb{E}[\mathcal{L}(Y_t, \langle\theta, X_t\rangle)|\mathcal{F}_t]$. Assuming $\mathcal{L}$ is differentiable and using the shorthand $\partial^j\mathcal{L}_t^* = \partial^j\mathcal{L}(Y_t, \langle\theta^*, X_t\rangle)$ for $j \in \mathbb{N}$ and $t \in \mathbb{N}$, this implies $\mathbb{E}[\partial^1\mathcal{L}_t^*|\mathcal{F}_t] = 0$ since X_t is measurable with respect to $\mathcal{F}_t$. A direct consequence of this is that $S_t = \sum_{s=1}^{t-1} \partial^1\mathcal{L}_s^* X_s$ 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). To this end, we detail below a helpful transformation of the sum process S into a nonnegative supermartingales (Lemma 1) under a standard sub-Gaussian assumption (Assumption 2) and state the high-probability uniform deviation bound we obtain (Proposition 1), the proof of which is deferred to Appendix C.

Assumption 2 (Sub-Gaussian) *$\partial^1\mathcal{L}^*$ is a conditionally sub-Gaussian process, i.e., there exists $R > 0$ such that*

$$\forall t \in \mathbb{N},\ \forall \lambda \in \mathbb{R},\ \log \mathbb{E}\left[\exp\left(\lambda \partial^1 \mathcal{L}_t^*\right) | \mathcal{F}_t\right] \leqslant \frac{\lambda^2 R^2}{2}\,.$$

Lemma 1 (Supermartingale Control) *Under Assumptions 1-2, there exists $\sigma > 0$ such that for any $t \in \mathbb{N}$ and $\lambda \in \mathbb{R}^d$, the following holds:*

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

Proposition 1 (Method of Mixtures with Convex Loss) *Let $\beta > 0$. Under Assumptions 1-2, with probability at least $1-\delta$, for all $t \in \mathbb{N}$, it holds that*

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

Discussion on Lemma 1 and Assumption 2 As shown in the proof, Lemma 1 alone implies Proposition 1. While this lemma may be valid in more general settings, we show in Appendix C how it is conveniently implied by Assumption 1 and the sub-Gaussian control of Assumption 2. In the rest of the paper, in particular in the regret bounds of Theorem 1, 2 and 3, σ will refer to the parameter that appears in the supermartingale control of Lemma 1.

Regarding Assumption 2, note that for a mean-linear bandit $Y_t = \langle \theta^*, X_t\rangle + \eta_t$ with adapted loss $\mathcal{L}(y,\xi) = \frac{1}{2}(y-\xi)^2$, we have that $\partial\mathcal{L}_t^* = \eta_t$, which is classically assumed to be sub-Gaussian. For other bandits, and thus other adapted losses, it may be more convenient to make assumptions on the distribution of observable quantities such as X_t and Y_t rather than directly on $\partial\mathcal{L}_t^*$. 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 our knowledge no complete answer is available, several partial results are available in the concentration literature.

(i) If Z is Gaussian with variance σ^2 and f is M-Lipschitz, the Tsirelson-Ibragimov-Sudakov inequality (Boucheron et al., 2013, Theorem 5.5) 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. More generally, if Z can be written as a σ-Lipschitz function of a $\mathcal{N}(0,1)$, 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 Vershynin (2018, Theorem 5.2.15)).

(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., Boucheron et al. (2013, Theorem 6.10)). The boundedness assumption can be lifted at the cost of a slightly more stringent condition than the sub-Gaussianity of Z, see Adamczak (2005, Theorem 3).

In short, Assumption 2 holds in a variety of settings, under rather mild assumptions on either X_t and Y_t or the loss $\mathcal{L}$. Also of note, $\partial^1\mathcal{L}$ is M-Lipschitz under Assumption 1.

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 (note the use of a priori different regularization parameters α and β, which we exploit in the sequel).

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

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

Then under Assumptions 1-2, 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 $\operatorname{argmin}_{\theta\in\Theta}\|F_t^\alpha(\theta) - F_t^\alpha(\widehat{\theta}_t)\|_{H_t^\beta(\theta)^{-1}}$, we have the property that $\Pi(\widehat{\theta}_t) \in \widehat{\Theta}_t^\delta$ with high probability.

Remark 4 *Although we formulated the bounded curvature condition (Assumption 1) globally, we note that we only require it to hold in a convex neighborhood of θ^* containing $\bar{\theta}_t$, and Corollary 1 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 in Li et al. (2017) for GLB.*

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

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

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 $X_t \in \operatorname{argmax}_{x \in \mathcal{X}_t} \langle \bar{\theta}_t, x \rangle + \gamma_t(x)$, 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 1 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 $\Delta(x, \bar{\theta}_t)$. By Cauchy-Schwarz's inequality, we have, for some positive definite matrix P to be determined later,

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

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 $[\bar{\theta}_t, \theta^*]$[1] (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

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

[1] One could also use $\bar{H}_t^\alpha(\theta^*, \bar{\theta}_t) = \int_0^1 H_t^\alpha(\gamma_u) du$ where $\gamma \colon [0,1] \to \Theta$ is smooth, unit speed path connecting $\bar{\theta}_t$ and θ^*.

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.

Lemma 2 (Transportation of Local Metrics) *Under Assumption 1, for $\alpha > 0$, there exists $\kappa > 0, \beta > 0$ such that*

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

We detail in Appendix D that a suitable choice of parameter is $\beta = \kappa\alpha$ with $\kappa = \frac{M}{m}$ the conditioning of the loss $\mathcal{L}$, which is a direct consequence of Assumption 1. Again, we keep the formulation fairly generic as Lemma 2 may hold beyond losses with bounded curvature. For instance, in a special case of GLB, namely the logistic bandit, it is shown in Faury et al. (2020) that this lemma holds thanks to self-concordance properties of the sigmoid link function.

3.3 Examples

We conclude this section by discussing examples of standard losses and whether they satisfy the above conditions.

Example 1 (Expectile) *The expectile loss is derived from the potential $\psi_2(z) = |p - \mathbb{I}_{z<0}| z^2$, the second derivative of which is $\psi_2''(z) = 2|p - \mathbb{I}_{z<0}|$. Thus, Assumption 1 holds with $m = 2\min(p, 1-p)$ and $M = 2\max(p, 1-p)$.*

Example 2 (Quantiles) *The quantile loss is derived from the potential $\psi_1(z) = (p - \mathbb{I}_{z<0}) z$, which is piecewise linear. In particular, it is not strongly convex and thus does not satisfy Assumption 1. Bandits with quantile regression are therefore outside the scope of this work.*

3.4 Regret Analysis

We make two additional standard assumptions that prior bounds are known on θ^* and on the actions $\mathcal{X} = \bigcup_{t \in \mathbb{N}} \mathcal{X}_t$, which is standard in the existing literature on linear bandits.

Assumption 3 (Prior Bound on Parameters) *All parameters are bounded by S, i.e., $\Theta \subseteq \mathcal{B}^d_{\|\cdot\|_2}(0, S)$. In particular, this implies that $\|\theta^*\|_{H_t^\beta(\theta^*)^{-1}} \leqslant \frac{S}{\sqrt{\beta}}$ for any $\beta > 0$.*

Assumption 4 (Prior Bound on Actions) *All actions are bounded by L, i.e., $\mathcal{X} \subseteq \mathcal{B}^d_{\|\cdot\|_2}(0, L)$.*

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. As is standard for contextual bandits, this bound is *minimax* (worst-case) as it does not depend explicitly on the optimality gaps $\langle \theta^*, X_t^* \rangle - \langle \theta^*, X_t \rangle$.

Theorem 1 (Regret upper bound for LinUCB-CR - 1) *Let $\delta \in (0,1)$, $\alpha \geqslant \max(1, L^2)$ and define for $t \in \mathbb{N}$ the exploration bonus*

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

and the projection operator

$$\Pi \colon \widehat{\theta} \in \mathbb{R}^d \mapsto \operatorname*{argmin}_{\theta \in \Theta} \|F_t^\alpha(\theta) - F_t^\alpha(\widehat{\theta})\|_{H_t^{\kappa\alpha}(\theta)^{-1}}.$$

Under Assumptions 1-2-3-4, with probability at least $1-\delta$, the regret of Algorithm 1 is bounded by

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

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

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 5 *This regret bound scales with $\kappa m^{-1/2}$, where m is the minimum curvature of the loss $\mathcal{L}$ and κ the coefficient of transportation of local metrics. Under Assumption 1, this scales as $m^{-3/2}$. In the limit of flattening loss $m \to 0$, learning with this strategy becomes impossible. We show in Appendix E that a small modification of the exploration sequence reduces this dependency to $\kappa^{1/2} m^{-1/2}$, at the cost of loosing local information carried by $H_t^{\kappa\alpha}(\bar{\theta}_t)$. An analogous dependency on m^{-1} was observed for GLB in Filippi et al. (2010). In the special case of logistic bandit, Faury et al. (2020) obtained a κ independent of m using self-concordance, and even pushed the dependency on $m^{-1/2}$ to higher order terms in T using a more intricate algorithmic design. We conjecture that a similar construction could apply here but leave this open for future work.*

Remark 6 *In the mean-linear case, $m = M = \kappa = 1$ and $H_t^\alpha = V_t^\alpha$, thus this result is compatible with the minimax lower bound $\mathcal{O}(d\sqrt{T})$ for actions in $\mathcal{B}_{\|\cdot\|}^d(0,1)$ (Lattimore and Szepesvári, 2020, Theorem 24.2) and matches the standard LinUCB upper bound (Abbasi-Yadkori et al., 2011).*

Theorem 1 holds for arbitrary (potentially adversarial) sequence of action sets $(\mathcal{X}_t)_{t\in\mathbb{N}}$. If these are instead stochastically generated, the regret bound can be further tightened.

Assumption 5 (Stochastic action sets) *Let $\nu_{\mathcal{X}}$ a probability measure on $2^{\mathcal{B}_{\|\cdot\|_2}^d(0,L)}$ (i.e., samples drawn from $\nu_{\mathcal{X}}$ are sets of vectors of L^2 norm at most L).*

(i) For $t \in \mathbb{N}$, $\mathcal{X}_t \sim \nu_{\mathcal{X}}$ defines an i.i.d. sequence of random action sets.

(ii) Recall that $X_t \in \mathcal{X}_t$ denotes the action selected by the agent at time $t \in \mathbb{N}$. Then $\mathbb{E}\left[X_t X_t^\top | \mathcal{F}_{t-1}\right] \succcurlyeq \rho_{\mathcal{X}} L^2 I_d > 0$.

Theorem 2 (Regret upper bound for LinUCB-CR - 2) *Under Assumptions 1-2-3-4-5, with probability at least $1-2\delta$, the regret of Algorithm 1 is bounded by*

$$\mathcal{R}_T \leqslant 4c_T^\delta \sqrt{\frac{2T}{m\rho_{\mathcal{X}}}} \left(1 + \frac{C}{\sqrt{T}}\right),$$

where C is a constant independent of T. In particular, we have $\mathcal{R}_T = \mathcal{O}\left(\kappa\sigma\sqrt{\frac{dT}{m\rho_{\mathcal{X}}}}\log\frac{TL^2}{d}\right)$.

The lower bound on conditional covariance of actions of Assumption 5 is new, although related to more standard settings. In the case of finite action sets $\mathcal{X}_t = \{X_{k,t}, k \in [K]\}$, Li et al. (2017); Kim et al. (2022) considered a lower bound on the unconditional average covariance across arms $\mathbb{E}\left[\frac{1}{K}\sum_{k\in[K]} X_{k,t}X_{k,t}^\top\right]$. We argue that this assumption is quite mild in the sense that for non-degenerate $\nu_{\mathcal{X}}$, the conditioning is essentially irrelevant. At time t, $\mathcal{X}_t$ is drawn independently of $\mathcal{F}_{t-1}$, and $X_t \in \mathcal{X}_t$ is selected in a $\mathcal{F}_{t-1}$-measurable fashion. To violate the covariance inequality, there should exist a fixed strict subspace $\mathcal{V} \subset \mathbb{R}^d$ such that with some probability $\mathcal{X}_t \cap \mathcal{V} \neq \emptyset$ (when randomizing over the action set $\mathcal{X}_t$) and X_t should be one of the vectors in $\mathcal{V}$; however, if, e.g., $\nu_{\mathcal{X}}$ spans an open set, this almost surely cannot happen. In other words, Theorem 2 shows that if action sets $\mathcal{X}_t$ are generated with enough diversity and no adversarial bias, the regret of optimistic strategies can be slightly improved by a factor $\mathcal{O}(\log(T))$. Finally, note that Assumption 5 and $\rho_{\mathcal{X}}$ do not influence the design of Algorithm 1, only the $\mathcal{O}(\log T)$ term in its regret upper bound.

In general, $\rho_{\mathcal{X}}^{-1} \geqslant d$ and in many cases $\rho_{\mathcal{X}}^{-1} = \mathcal{O}(d)$ (see Appendix F), hence the regret upper bound scales linearly with d. Compared to Kim et al. (2022), our proof relies on line crossing arguments developed in Howard et al. (2020) rather than on a crude union bound, leading to improved constants and higher order terms (even in the mean-linear case). We refer to Appendix F for additional details.

4 APPROXIMATE STRATEGY 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 via the Sherman-Morrison formula.

We propose an alternative algorithm that exploits online gradient descent (OGD) to compute a fast approximation of the empirical risk minimizer $\widehat{\theta}_t$. 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 relevant 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. For $\widehat{\theta} \in \mathbb{R}^d$, we use the shorthand $\nabla^{\alpha}_{n,h}(\widehat{\theta}) = \sum_{k=1}^{h} \partial\mathcal{L}(Y_{(n-1)h+k}, \langle\widehat{\theta}, X_{(n-1)h+k}\rangle) + \alpha\widehat{\theta}$.

Algorithm 2 LinUCB-OGD-CR

Input: horizon T, regularisation parameter α, projection Π, exploration bonus sequence $(\gamma^{\mathrm{OGD}}_{t,T})_{t\leqslant T}$, step sequence $(\varepsilon_t)_{t\in\mathbb{N}}$, episode length $h > 0$.
Initialization: Observe $\mathcal{X}_1$, set $\widehat{\theta}^{\mathrm{OGD}}_0$, $t = 1$, $n = 1$.
for $t = 1, \dots, T$ **do**
  if $t = nh + 1$ **then**
    $\widehat{\theta}^{\mathrm{OGD}}_n = \widehat{\theta}^{\mathrm{OGD}}_{n-1} - \varepsilon_{n-1}\nabla^{\alpha}_{n,h}(\widehat{\theta}^{\mathrm{OGD}}_{n-1})$; ▷ OGD
    $\bar{\theta}^{\mathrm{OGD}}_n = \frac{1}{n}\sum_{j=1}^{n}\Pi(\widehat{\theta}^{\mathrm{OGD}}_j)$; ▷ Average
    $n \leftarrow n + 1$
  a $X_t = \arg\max_{x\in\mathcal{X}_t}\langle\bar{\theta}^{\mathrm{OGD}}_n, x\rangle + \gamma^{\mathrm{OGD}}_{t,T}(x)$; ▷ Play with same parameter for h steps
  Observe Y_t and $\mathcal{X}_{t+1}$;
  $t \leftarrow t + 1$;

The intuition behind Algorithm 2 is that at time $t = nh+1$, the approximation error between the OGD estimate $\bar{\theta}^{\mathrm{OGD}}_n$ 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 $\nabla^{\alpha}_{n,h}(\widehat{\theta})$. 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. We prove in Appendix G an extension of the following proposition, with an explicit bound on the OGD regret, which we report below in the $\mathcal{O}$ notation for the sake of concision.

Proposition 2 (OGD Regret, 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 uniformly in $N \to +\infty$ by*

$$\sum_{j=1}^{N}\ell_j(z_j) - \ell_j(\bar{z}_N) = \mathcal{O}\left(\frac{d\sigma^2}{a}\log^2 N\right).$$

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

Theorem 3 (Regret of LinUCB-OGD-CR) *Let $\varepsilon_h > 0$ and $h = \lceil\frac{2\varepsilon_h}{\rho_{\mathcal{X}}L^2} + \frac{8}{\rho^2_{\mathcal{X}}}\log\frac{2}{\delta}\rceil$. Assume that $\partial\mathcal{L}(Y_t, \langle\theta^*, X_t\rangle)$ is $\sqrt{m}\sigma$-sub-Gaussian for all $t \leqslant T$. Under Assumptions 1-2-3-4-5, there exists constants $C, C' > 0$ such that with probability at least $1 - (1 + T/h)\delta$ the regret of Algorithm 2 with exploration bonus sequence*

$$\gamma^{OGD}_{t,T} \colon x \in \mathcal{X}_t \mapsto (c^{\delta}_t + c^{OGD,\delta}_{t,T})\|x\|_{H_t^{\kappa\alpha}(\bar{\theta}^{OGD}_{\lfloor\frac{t-1}{h}\rfloor})^{-1}},$$

$$c^{OGD,\delta}_{t,T} = \sqrt{\left(L^2 + \frac{\alpha}{mMt}\right)\left(\frac{2\kappa C' dh^2\sigma^2}{\varepsilon_h^2}\log\left(\frac{2dT}{h\delta}\right)\log\left(\frac{t}{h}\right)\right)},$$

and the OGD step sequence of Proposition 2 satisfies

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

The episode length h scales as $\mathcal{O}(\rho_{\mathcal{X}}^{-2})$, which grows at least as fast as $\mathcal{O}(d^2)$ in the action dimension d. This is sufficient to bound with high probability the smallest eigenvalue of the Hessian of the aggregate losses $\nabla^{\alpha}_{n,h}$ and thus ensure their strong convexity. However, longer episodes also means less frequent updates of $\bar{\theta}^{\mathrm{OGD}}_n$, i.e., less learning, which is materialized by the additional dependency on h in the regret. In Appendix H, Lemma 7, we deduce a tighter, more intricate expression for h, although still scaling as $\mathcal{O}(\rho_{\mathcal{X}}^{-2})$. We only report the simpler expression here to avoid cluttering.

Remark 7 *The union bound used in Proposition 2 imposes the knowledge of the horizon T at runtime (in the definition of $\gamma_{t,T}$), thus making Algorithm 2 not anytime.*

5 EXPERIMENTS

We conducted three numerical experiments to illustrate the performance of both risk-aware algorithms, under expectiles and entropic risk criteria. In Figure 1, we considered an expectile-based asymmetric distribution (Torossian et al., 2020) with context-dependent p-expectiles. This distribution is log-concave, thus fitting the scope of the supermartingale control of Lemma 1. As recalled in Section 3.3, the corresponding loss satisfies Assumption 1 with $m = 2\min(p, 1-p)$ and $M = 2\max(1-p, p)$. Note that the more risk-averse ($p \to 0$), the flatter the loss ($m \to 0$)

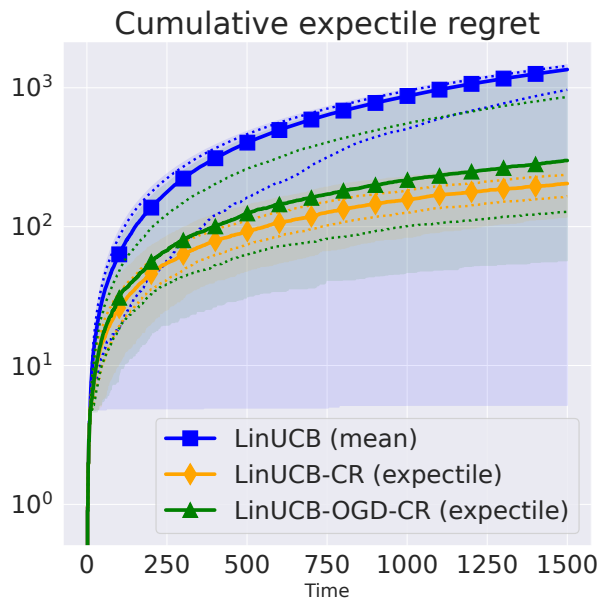

Figure 1: Two-armed linear 10%-expectile bandit with $\mathbb{R}^3$ contexts and expectile-based asymmetric noises. Thick lines denote median cumulative regret over 500 independent replications. Dotted lines denote the 25 and 75 regret percentiles. Shaded areas denote the 5 and 95 percentiles.

and thus the harder it is to learn. This matches the intuition on risk-aware measures: by focusing on the more extreme events, they require more samples to reach the same statistical accuracy. More details, including the other two experiments, are postponed to Appendix I. As far as we are aware, no algorithm exists for the expectile criterion; for entropic risk, Maillard (2013) analyzes a variant of KL-UCB but only for the non-contextual multi-armed bandit problem (and without numerical evidences).

The settings of these numerical experiments were designed so that the optimal arms were different depending on the criterion of interest (mean versus risk-aware). 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 exhibited milder sublinear trends. As expected, the LinUCB-OGD variant accumulated slightly more regret and showed higher variability across independent replications compared to LinUCB with the exact minimization of the empirical risk, at the benefit of improved runtimes (Table 2).

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 be-

Table 2: Runtimes for the Classical LinUCB and Algorithms 1 (LinUCB for Convex Risk) and 2 (LinUCB-OGD for Convex Risk), Reported in Seconds as Mean $\pm$ Standard Deviation, Estimated Across 500 Independent Replications with Time Horizon $T = 1500$.

Algorithm	Runtime
LinUCB (mean)	37.2 ± 4.9
LinUCB-CR (expectile)	814.8 ± 88.3
LinUCB-OGD-CR (expectile)	60.2 ± 12.0

yond standard mean-linear bandits. To lift the regret analysis of optimistic algorithms to the setting of scalar risk measures $\rho_{\mathcal{L}}$ elicited by a convex loss $\mathcal{L}$, we showed that uniformly bounding the curvature of the loss (Assumption 1) is sufficient to maintain satisfying theoretical guarantees ($\mathcal{O}(\sqrt{T})$ worst-case regret, up to polylog terms (Theorem 1 and 2). More precisely, we identified two key conditions, namely a supermartingale control (Lemma 1) and a transportation inequality (Lemma 2), that guarantee sublinear regret; while these are direct consequences of the bounded curvature assumption, they may hold in different settings, as was recently discovered in GLB.

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, conditional value-at-risk or quantiles would require to adapt the theory to high-order elicitable measures and to non-smooth losses. Finally, we believe the technical results developed here, in particular the supermartingale control and the transportation inequality, can pave a way for the design and analysis of Thompson sampling strategies in the contextual risk-aware setting.

Acknowledgements

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). We also thank the organizers and attendees of the European Workshop on Reinforcement Learning (EWRL) 2022, where this work was initially presented.

References

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

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

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

C. Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.

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

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

J. Arbel, S. Girard, H. Nguyen, and A. Usseglio-Carleve. Multivariate expectile-based distribution: properties, bayesian inference, and applications. 2021.

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

H. Bastani, M. Bayati, and K. Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

D. Baudry, R. Gautron, E. Kaufmann, and O. Maillard. Optimal thompson sampling strategies for support-aware cvar bandits. In *International Conference on Machine Learning*, pages 716–726. PMLR, 2021.

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

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

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

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

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

W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos. Distributional reinforcement learning with quantile regression, 2017.

Q. Ding, C.-J. Hsieh, and J. 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.

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

P. Embrechts, T. Mao, Q. Wang, and R. Wang. Bayes risk, elicitability, and the expected shortfall. *Mathematical Finance*, 31(4):1190–1217, 2021.

L. Faury, M. Abeille, C. Calauzènes, and O. Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 23, 2010.

T. Fissler and J. F. Ziegel. Higher order elicitability and osband's principle. *The Annals of Statistics*, 44(4):1680–1707, 2016.

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

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

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

A. Gopalan, L. Prashanth, M. Fu, and S. Marcus. Weighted bandits or: How bandits learn distorted values that are not expected. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

E. Hazan. Introduction to online convex optimization, 2019.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

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

W. Kim, G.-s. Kim, and M. C. Paik. Doubly robust thompson sampling with linear payoffs. *Advances in Neural Information Processing Systems*, 34:15830–15840, 2021.

W. Kim, K. Lee, and M. C. Paik. Double doubly robust thompson sampling for generalized linear contextual bandits. *arXiv preprint arXiv:2209.06983*, 2022.

J. Kirschner, T. Lattimore, C. Vernade, and C. Szepesvári. Asymptotically optimal information-directed sampling. In *Conference on Learning Theory*, pages 2777–2821. PMLR, 2021.

N. Korda, L. Prashanth, and R. Munos. Fast gradient descent for drifting least squares regression, with application to bandits. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

T. L. Lai and C. Z. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

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

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

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

R. C. Merton. Theory of rational option pricing. *The Bell Journal of economics and management science*, pages 141–183, 1973.

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

A. K. Pandey, L. Prashanth, and S. P. Bhat. Estimation of spectral risk measures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12166–12173, 2021.

V. H. Peña, T. L. Lai, and Q.-M. Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.

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

I. Pinelis. An approach to inequalities for the distributions of infinite-dimensional martingales. In *Probability in Banach Spaces, 8: Proceedings of the Eighth International Conference*, pages 128–134. Springer, 1992.

I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

L. Prashanth, K. Jagannathan, and R. K. 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. T. Rockafellar, S. Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

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

A. Tamkin, R. Keramati, C. Dann, and E. Brunskill. Distributionally-aware exploration for cvar bandits. In *NeurIPS 2019 Workshop on Safety and Robustness in Decision Making; RLDM*, volume 2020, 2019.

L. Torossian, V. Picheny, and N. Durrande. Bayesian quantile and expectile optimisation. *arXiv preprint arXiv:2001.04833*, 2020.

R. 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.

M. Wirth, A. Klein, and A. Ortiz. Risk-aware multi-armed bandits for vehicular communications. 2022.

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

A SUMMARY AND INTERPRETATION OF ELICITABLE RISK MEASURES

We report in Table 3 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 $\rho_{\mathcal{L}}(\nu) = \arg\min_{\xi \in \mathbb{R}^p} \mathbb{E}_{Y\sim\nu}\left[\mathcal{L}(Y,\xi)\right]$. 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 $\mathrm{VaR}_\alpha(\nu)$ 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 $\mathrm{CVaR}_\alpha(\nu)$ is thus $\mathbb{E}\left[X \mid X \leqslant \mathrm{VaR}_\alpha(\nu)\right]$. Intuitively, a random variable with a high CVaR_α distribution takes on average relatively high values in the "$\alpha\%$ worst-case" scenario. For $\alpha \to 1^-$, $\mathrm{CVaR}_\alpha(\nu) \to \mathbb{E}_{Y\sim\nu}[Y]$ 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

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

i.e., $\mathrm{CVaR}_\alpha(\nu) = \rho_{MV_1}(\nu)$ 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 $\mathrm{CVaR}_\alpha(\nu)$, 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))

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

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 $\mathtt{KL}(\nu'\|\nu) = \mathbb{E}_{Y\sim\nu'}[\log\frac{d\nu'}{d\nu}]$ instead of the variance. The entropic risk measure can be interpreted as the largest expected value that a mispecified model ν' (in place of the true underlying distribution ν) may have, where $\mathtt{KL}(\nu'\|\nu)$ 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

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

Table 3: 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	$\operatorname{argmin}_{\xi\in\mathbb{R}}\mathbb{E}_{Y\sim\nu}[\psi(Y-\xi)]$	$\psi(y-\xi)$	$\xi\in dom(\psi)$
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, $c\colon\mathbb{R}\to\mathbb{R}$.	$\xi_1\geqslant\xi_2$

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 $K = e_p(\nu)$.

Contrary to the previous risk measures, it may not be clear from this definition alone that expectiles do encode a notion of aversion to risk. The next proposition shows that p-expectiles of many distributions, including normal and adjusted lognormal, are decreasing functions of their variances when $p < \frac{1}{2}$, thus penalizing more volatile distributions, making them suitable for risk management. We provide an elementary proof using the tools of the financial mathematics literature, where such risk measures were extensively studied.

Proposition 3 *Let $\mathcal{I} \subseteq \mathbb{R}_+^*$ an open set and $\{\nu_\sigma, \sigma \in \mathcal{I}\}$ a family of probability distributions such that*

(i) the expectation mapping $\sigma \in \mathcal{I} \mapsto \mathbb{E}_{Y\sim\nu_\sigma}[Y]$ is constant,

(ii) both the call and put mappings $C(\cdot, K)\colon \sigma \in \mathcal{I} \mapsto \mathbb{E}_{Y\sim\nu_\sigma}\left[(Y-K)_+\right]$ and $P(\cdot, K)\colon \sigma \in \mathcal{I} \mapsto \mathbb{E}_{Y\sim\nu_\sigma}\left[(K-Y)_+\right]$ are differentiable and nondecreasing, for any $K \in \mathbb{R}$.

For $p \in (0,1)$, we let $e_p(\sigma) = e_p(\nu_\sigma)$. Then $\operatorname{sign} \frac{d}{d\sigma} e_p(\sigma) = \operatorname{sign}(p - \frac{1}{2})$.

Before we proceed to the proof, let us note that two classical families of distributions satisfy these assumptions (see (Merton, 1973, Theorem 8) for a a general result).

- **Normal**: for $\mu_0 \in \mathbb{R}$, $\{\nu_\sigma = \mathcal{N}(\mu_0, \sigma^2),\ \sigma \in \mathbb{R}_+^*\}$, for which $\mathbb{E}_{Y\sim\nu_\sigma}[Y] = \mu_0$.
- **Adjusted lognormal**: for $\mu_0 \in \mathbb{R}$, $\{\nu_\sigma = \exp\left(\mathcal{N}(\mu_0, \sigma^2) - \frac{\sigma^2}{2}\right),\ \sigma \in \mathbb{R}_+^*\}$, for which $\mathbb{E}_{Y\sim\nu_\sigma}[Y] = e^{\mu_0}$.

In particular in the normal case, it follows from Lemma 3 that $e_p(\nu) = \mu_0 + \sigma e_p(\mathcal{N}(0,1)) = \rho_{MV_1}(\nu_\sigma)$ at risk aversion level $\lambda = -e_p(\mathcal{N}(0,1))$ (positive if $p < \frac{1}{2}$).

Proof We first recall the call-put parity principle, which states that for any distribution ν_σ and strike $K \in \mathbb{R}$, the following equality holds:

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

where we write $C(\nu_\sigma, K) = C(\sigma, K)$ and $P(\nu_\sigma, K) = P(\sigma, K)$.

Notice that the call-parity principle and the assumption that $\frac{d}{d\sigma}\mathbb{E}_{Y\sim\nu_\sigma}[Y] = 0$ implies that $\partial_\sigma C = \partial_\sigma P$. We denote this quantity by V. First, note that $\sigma \in \mathcal{I} \mapsto e_p(\sigma)$ is differentiable (implicit function theorem). From the equation $C(\sigma, e_p(\sigma)) = (1-p)/pP(\sigma, e_p(\sigma))$, we deduce that

$$\begin{aligned}
\frac{d}{d\sigma} C(\sigma, e_p(\sigma)) &= \partial_\sigma C(\sigma, e_p(\sigma)) + \partial_K C(\sigma, e_p(\sigma)) \frac{d}{d\sigma} e_p(\sigma)\,,\\
\frac{d}{d\sigma} P(\sigma, e_p(\sigma)) &= \partial_\sigma P(\sigma, e_p(\sigma)) + \partial_K P(\sigma, e_p(\sigma)) \frac{d}{d\sigma} e_p(\sigma)\,,
\end{aligned}$$

and thus

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

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 $\frac{d}{d\sigma} e_p(\sigma) \leqslant 0$. ■

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.

B PROPERTIES OF CONVEX LOSSES AND POTENTIALS

Before we prove Lemma 4, we write the following technical lemma.

Lemma 3 (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 $\rho_\psi(\nu) = \operatorname{argmin}_{\xi\in\mathbb{R}} \int \psi(y-\xi)\, p(y)dy$ 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 $\xi \mapsto \int \psi(y-\xi)\, p(y)dy$. By uniqueness ($\psi$ is strongly convex), we deduce that $\rho_\psi(\nu + c) = \rho_\psi(\nu) + c$. ■

Noise Additivity for Losses Derived From Potentials

Lemma 4 *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 3 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^*).

■

C PROOF OF LEMMA 1 AND PROPOSITION 1

Lemma 1 (Supermartingale Control) *Under Assumptions 1-2, there exists $\sigma > 0$ such that for any $t \in \mathbb{N}$ and $\lambda \in \mathbb{R}^d$, the following holds:*

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

Proof Assumption 1 implies that $\partial^2\mathcal{L}_t^* \geqslant m$, therefore it is sufficient to show that there exists $\sigma > 0$ such that

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

Since X_t is $\mathcal{F}_t$-measurable, this is equivalent to

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

which follows from the sub-Gaussian property of the process $\partial^1\mathcal{L}^*$ (Assumption 2).

Proposition 1 (Method of Mixtures with Convex Loss) *Let $\beta > 0$. Under Assumptions 1-2, with probability at least $1 - \delta$, for all $t \in \mathbb{N}$, it holds that*

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

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

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

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, Ch. 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. ■

D ASSUMPTION 1 $\implies$ LEMMA 2

First, we recall the two assumptions of interest.

Assumption 1 (Bounded Loss Curvature) *There exists m and M 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 $\mathcal{L}$.

Lemma 2 (Transportation of Local Metrics) *Under Assumption 1, for $\alpha > 0$, there exists $\kappa > 0, \beta > 0$ such that*

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

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.

E PROOF OF THEOREM 1

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

Theorem 1 (Regret upper bound for LinUCB-CR - 1) *Let $\delta \in (0,1)$, $\alpha \geqslant \max(1, L^2)$ and define for $t \in \mathbb{N}$ the exploration bonus*

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

and the projection operator

$$\Pi \colon \widehat{\theta} \in \mathbb{R}^d \mapsto \operatorname*{argmin}_{\theta\in\Theta} \|F_t^\alpha(\theta) - F_t^\alpha(\widehat{\theta})\|_{H_t^{\kappa\alpha}(\theta)^{-1}} \,.$$

Under Assumptions 1-2-3-4, with probability at least $1-\delta$, the regret of Algorithm 1 is bounded by

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

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

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 (Lemma 2, implied by the curvature bound of Assumption 1) reveals that

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

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

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

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

The prior bound on parameters (Assumption 3) yields $\|\theta\|_{H_t^\beta(\theta)^{-1}} \leqslant \frac{S}{\sqrt{\beta}}$ for $\theta \in \{\theta^*, \bar{\theta}_t\}$. Furthermore, the curvature bound (Assumption 1) 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 $\kappa = \frac{M}{m}$ is the conditioning of the convex loss $\mathcal{L}$, we obtain:

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

By the same arguments, it holds that $\bar{H}_t^\alpha(\theta^*,\bar{\theta}_t)^{-1} \preccurlyeq \kappa H_t^{\kappa\alpha}(\bar{\theta}_t)^{-1}$ and 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

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

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 1) 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 8 (Deterministic elliptic potential) *Let $(x_t)_{t\in\mathbb{N}}$ denote an arbitrary sequence of vectors in $\mathcal{B}_{\|\cdot\|}^d(0,L)$, $\varepsilon > 0$ and $v_t = \sum_{s=1}^{t-1} x_s x_s^\top + \varepsilon I_d \in \mathcal{S}_d(\mathbb{R})$ for $t \in \mathbb{N}$. Then*

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

Note that this result holds in our case (with $\varepsilon = \frac{\kappa\alpha}{m}$) thanks to the prior bound on actions (Assumption 4).

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

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

Going back to the expression of c_T^δ, it follows from simple algebra (see e.g., Lattimore and Szepesvári (2020, proof of Lemma 19.4)) 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

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

■

Impact of using the local Hessian metric H_t versus the global metric V_t To highlight the benefit of using local metrics, we detail here the regret bound obtained using the above proof with the natural global metric induced by $V^{\alpha/m}$ (independent of the local point θ). Instantiating the positive definite matrix P to $V_t^{\alpha/m}$ instead of $\bar{H}^\alpha(\theta^*, \bar{\theta}_t)$ in the bound on the prediction error of Section 3.2 yields

$$\begin{aligned}
\Delta(x, \bar{\theta}_t) &\leqslant \|\theta^* - \bar{\theta}_t\|_{V_t^{\alpha/m}} \|x\|_{(V_t^{\alpha/m})^{-1}} \\
&= \|F_t^\alpha(\theta^*) - F_t^\alpha(\bar{\theta}_t)\|_{\bar{H}_t^\alpha(\theta^*,\bar{\theta}_t)^{-1} V_t^{\alpha/m} \bar{H}_t^\alpha(\theta^*,\bar{\theta}_t)^{-1}} \|x\|_{(V_t^{\alpha/m})^{-1}} \\
&\leqslant \frac{1}{\sqrt{m}} \|F_t^\alpha(\theta^*) - F_t^\alpha(\bar{\theta}_t)\|_{\bar{H}_t^\alpha(\theta^*,\bar{\theta}_t)^{-1}} \|x\|_{(V_t^{\alpha/m})^{-1}} \\
&\leqslant 2\sqrt{\frac{\kappa}{m}} \left(\sigma\sqrt{2\log\frac{1}{\delta} + d\log\frac{m}{\alpha} + \log\det V_t^{\frac{\alpha}{m}}} + \sqrt{\frac{\alpha}{\kappa}} S\right) \|x\|_{(V_t^{\alpha/m})^{-1}}.
\end{aligned}$$

Similarly to the above proof, this shows that

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

is also a valid choice of exploration sequence. Finally, a straightforward application of Lemma 8 shows the regret of the corresponding LinUCB-CR strategy is upper bounded with probability at least $1-\delta$ by

$$4\sqrt{\frac{\kappa}{m}} \left(\sigma\sqrt{2\log\frac{1}{\delta} + d\log\frac{m}{\alpha} + \log\det V_t^{\frac{\alpha}{m}}} + \sqrt{\frac{\alpha}{\kappa}} S\right) \sqrt{2Td\log\left(1 + \frac{mTL^2}{d\alpha}\right)}.$$

Compared to the local analysis, this improves the scaling of the regret in κ by a factor $\sqrt{\kappa}$. However, it forces the use of the global metric $\|\cdot\|_{(V_t^{\alpha/m})^{-1}}$ instead of the local one $\|\cdot\|_{H_t^{\kappa\alpha}(\bar{\theta}_t)^{-1}}$, thus ignoring the precise shape of the loss function $\mathcal{L}$.

Looking at our proof, we see that κ and m fulfil two different roles. $\sqrt{\kappa}$ is the price to pay in order to transport local metrics $H_t^\alpha(\theta)$ between $\theta = \theta^*$ (true parameter) and $\theta = \bar{\theta}_t$ (estimate); it is paid once to bound the prediction error $\Delta(X_t, \theta)$ using the concentration bound of Proposition 1, and it is also paid a second time if local metrics are used in the exploration bonus when moving from $\bar{H}_t(\theta^*, \bar{\theta}_t)^{-1}$ to $\bar{H}_t(\bar{\theta}_t)^{-1}$ (the former cannot be used directly in the algorithm as it depends on the a priori unknown paramter θ^*). On the other hand, $m^{-1/2}$ is the price paid in both the local and global analyses to move from $H_t(\bar{\theta}_t)^{-1}$ to V_t^{-1} in order to apply the elliptic potential lemma, which is in general incompatible with local metrics. A similar phenomenon is observed in the analysis of Faury et al. (2020): the regret of their algorithm Logistic-UCB-1 (global) scales as $\sqrt{\kappa}$ while that of Logistic-UCB-2 (local) scales as κ. In addition to local metrics, Logistic-UCB-2 also makes use of an intricate projection step that allows for a new elliptic potential lemma compatible with local metrics, thus removing the factor m^{-1} (at least from the first order contribution to the regret in T). We conjecture that a similar analysis could be unlocked in the present risk-aware setting and leave it open for future investigation.

We reiterate that in the logistic setting, κ is derived from self-concordance properties of the link function and is in particular independent of the curvature lower bound represented by m (it is in fact equal to $1 + 2S$ where S is an upper bound on the parameter space Θ, as in Assumption 3). By analogy with logistic bandits, we argue that the exact scaling in κ is likely not too harmful for the practical performances of Algorithm 1 and we therefore recommend the use of local metrics instead.

F PROOF OF THEOREM 2

In this section, we prove the regret bound of Theorem 2 in the stochastic i.i.d. actions setting of Assumption 5. We first state the full regret bound with an explicit higher order term.

Theorem 4 (Regret of LinUCB-CR with stochastic actions) *Let $\delta \in (0,1)$ and $t_0 = \lceil \frac{8}{\rho_{\mathcal{X}}^2} \log \frac{2}{\delta} - \frac{2\beta}{m\rho_{\mathcal{X}} L^2} \rceil$. Under Assumptions 1-2-3-4-5, for $T \geqslant t_0$, with probability at least $1-2\delta$, the regret of Algorithm 1 is bounded by*

$$\mathcal{R}_T \leqslant 4c_T^{\delta} \sqrt{\frac{2T}{m\rho_{\mathcal{X}}}} \left(1 + \frac{C}{\sqrt{T}}\right),$$

where

$$C = \frac{1}{L^2}\sqrt{\frac{\kappa\alpha - 4\frac{mL^2}{\rho_{\mathcal{X}}}\log\frac{2}{\delta}}{2m\rho_{\mathcal{X}}}} - \frac{1}{2L}\sqrt{t_0 - 1 + \frac{2(\kappa\alpha - \frac{4mL^2}{\rho_{\mathcal{X}}}\log\frac{2}{\delta})}{m\rho_{\mathcal{X}}L^2}} + \frac{1}{2}\max\left(1, L\sqrt{\frac{m}{\kappa\alpha}}\right)\sqrt{\rho_{\mathcal{X}} d t_0 \log\left(1 + \frac{mL^2 t_0}{d\kappa\alpha}\right)}.$$

In particular, we have $\mathcal{R}_T = \mathcal{O}\left(\kappa\sigma\sqrt{\frac{dT}{m\rho_{\mathcal{X}}}\log\frac{TL^2}{d}}\right)$.

The main difference with the proof of Theorem 1 is the use of an alternative stochastic elliptic potential lemma, mirroring the classical result of Lemma 8, that exploits the lower bound on the covariance of actions (Assumption 5). This proof technique is adapted from Kim et al. (2022), although we use a different, sharper concentration result (Proposition 4 below).

Lemma 5 (Stochastic elliptic potential lemma) *Let $\beta > 0$, $(\theta_t)_{t\in\mathbb{N}}$ a sequence of vectors in $\Theta \subseteq \mathbb{R}^d$ and $(X_t)_{t\in\mathbb{N}}$ a sequence of random variables in $\mathbb{R}^d$. Recall that $H_t^{\beta}(\theta_t) = \sum_{s=1}^{t-1} \partial^2\mathcal{L}(Y_s, \langle X_s, \theta_t\rangle) X_s X_s^\top + \beta I_d$ and $V_t^{\beta} = \sum_{s=1}^{t-1} X_s X_s^\top + \beta I_d$ for $t \in \mathbb{N}$. Under Assumptions 4 and 5, let $\delta \in (0,1)$ and $t_0 = \lceil \frac{8}{\rho_{\mathcal{X}}^2}\log\frac{2}{\delta} - \frac{2\beta}{m\rho_{\mathcal{X}}L^2}\rceil$. For $T \geqslant t_0$, with probability at least $1-\delta$, it holds that*

$$\sum_{t=1}^{T} \|X_t\|_{H_t^{\beta}(\theta_t)^{-1}} \leqslant 2\sqrt{\frac{2T}{m\rho_{\mathcal{X}}}}\left(1 + \frac{C}{\sqrt{T}}\right),$$

where

$$C = \frac{1}{L^2}\sqrt{\frac{\beta - 4\frac{mL^2}{\rho_{\mathcal{X}}}\log\frac{2}{\delta}}{2m\rho_{\mathcal{X}}}} - \frac{1}{2L}\sqrt{t_0 - 1 + \frac{2(\beta - \frac{4mL^2}{\rho_{\mathcal{X}}}\log\frac{2}{\delta})}{m\rho_{\mathcal{X}}L^2}} + \frac{\sqrt{\rho_{\mathcal{X}}}}{2}\max\left(1, L\sqrt{\frac{m}{\beta}}\right)\sqrt{t_0 d \log\left(1 + \frac{mL^2 t_0}{d\beta}\right)}.$$

The intuition about this result is the following: if the matrix norms induced by $H_t^{\beta}(\theta_t)$ grow at least linearly in t, then the left-hand side should scale like $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} = \mathcal{O}(\sqrt{T})$, without the extra $\mathcal{O}(\sqrt{\log T})$ factor present in Lemma 8. The lower curvature bound of Assumption 1 shows that it is enough to look at the norms induced by $V_t^{\beta/m}$, at the cost of an extra $m^{-1/2}$ factor (in particular Lemma 5 holds for *any* sequence $(\theta_t)_{t\in\mathbb{N}}$, not just the sequence of estimators used in the bandit algorithms). Because of the stochastic sampling of actions (Assumption 5), it is likely that the sequence $(X_t)_{t\in\mathbb{N}}$ spans all directions of $\mathbb{R}^d$ quite fast; in other words, each new $X_t X_t^\top$ will contribute at least a fixed amount to the sum that defines $V_t^{\beta/m}$, leading to the linear growth of the induced norms.

We formalize this intuition in Lemma 6 below, which relies on the following concentration bound in Hilbert spaces.

Proposition 4 (Time-uniform line crossing inequality for martingales with bounded increments in a Hilbert space) *Let $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ a Hilbert space and $(M_t)_{\in\mathbb{N}}$ a $\mathcal{H}$-valued martingale (with respect to a filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$) such that $M_0 = 0$. Assume that there exists a sequence of positive scalars $(c_t)_{t\in\mathbb{N}}$ such that $\|M_{t+1} - M_t\| \leqslant c_t$ for all $t \in \mathbb{N}$, where $\|\cdot\|$ denotes the norm induced by the scalar product. Then for any $\eta > 0$ and $\delta \in (0,1)$, it holds that*

$$\mathbb{P}\left(\exists t \in \mathbb{N},\ \|M_t\| \geqslant \frac{1}{2\eta}\log\frac{2}{\delta} + \eta\sum_{s=1}^{t} c_s^2\right) \leqslant \delta\,.$$

Interestingly, this concentration bound does not depend on the dimension of $\mathcal{H}$, and in particular remains valid even if the ambient space is infinite-dimensional. Moreover, this bound controls the probability that *any* deviation occurs in the sequence $(\|M_t\|)_{t\in\mathbb{N}}$, which is much stronger than controlling the deviation probability individually at each time $t \in \mathbb{N}$. The proof relies on martingale arguments rather than a crude union bound over a finite set of individual deviation probabilities, which yields anytime ($t \in \mathbb{N}$ rather than $t \leqslant T$ for some known horizon T) and typically tighter bounds.

Proof This result is directly taken from Howard et al. (2020). More precisely, Howard et al. (2020, Theorem 1) shows a variety of equivalent time-uniform line crossing inequalities for martingales, and Howard et al. (2020, Corollary 10) applies this generic result to concentration of norm-like operators in Banach spaces. In order to get the most convenient form for our problem, we derive Proposition 4 from the generic theorem rather than the specific corollary.

The proofs of Pinelis (1992, Theorem 3) and Pinelis (1994, Theorem 3) reveals that for any $\lambda \in \mathbb{R}$, the exponential process $L_t = \cosh(\lambda \|M_t\|) \exp(-\frac{\lambda^2}{2} \sum_{s=1}^{t} c_s)$ is a nonnegative $\mathcal{F}$-supermartingale. Therefore, Howard et al. (2020, Theorem 1, (a)) shows that for any $a, b > 0$,

$$\mathbb{P}\left(\exists t \in \mathbb{N},\ S_t \geqslant a + bV_t\right) \leqslant 2e^{-aD(b)}\,,$$

where $S_t = \|M_t\|$, $V_t = \sum_{s=1}^{t} c_s^2$ and $D(b) = 2b$. Equating the right-hand side to δ and letting $b = \eta$ concludes the proof, with $a = \frac{1}{2\eta} \log \frac{2}{\delta}$. ■

In the next lemma, we show that the smallest eigenvalue of $H_t^\beta(\theta_t)$, which provides a lower bound to the corresponding induced norm, does indeed grow linearly with t on an event of high probability. For a given symmetric matrix $A \in \mathcal{S}_d(R)$, we denote by $\lambda_{\min}(A)$ its smallest eigenvalue.

Lemma 6 (Smallest eigenvalue of $H_t^\beta(\theta_t)$ grows linearly with t with high probability) *Under Assumptions 1-4-5, it holds that*

$$\mathbb{P}\left(\exists t \in \mathbb{N},\ \lambda_{\min}\left(H_{t+1}^\beta(\theta_{t+1})\right) \leqslant \beta - \frac{4mL^2}{\rho_{\mathcal{X}}} \log \frac{2}{\delta} + \frac{m\rho_{\mathcal{X}} L^2}{2} t\right) \leqslant \delta\,.$$

Proof First notice that Assumption 1 implies

$$\lambda_{\min}\left(H_{t+1}^\beta(\theta_{t+1})\right) \geqslant m\lambda_{\min}\left(V_{t+1}^0\right) + \beta\,.$$

The idea is to relate $\lambda_{\min}\left(V_{t+1}^0\right)$ to the norm of some martingale in order to apply Proposition 4. In the stochastic actions setting (Assumption 5), a natural martingale is defined the following sum of random matrices:

$$M_t = \sum_{s=1}^{t} X_s X_s^\top - \mathbb{E}\left[X_s X_s^\top | \mathcal{F}_{s-1}\right] = V_{t+1}^0 - \bar{V}_{t+1}^0\,,$$

where we defined $\bar{V}_{t+1}^0 = \sum_{s=1}^{t} \mathbb{E}\left[X_s X_s^\top | \mathcal{F}_{s-1}\right]$. We recall that a consequence of Weyl's inequality on eigenvalues is that for any $A, B \in \mathcal{S}_d(\mathbb{R})$, the following inequality holds:

$$\lambda_{\min}(A) + \lambda_{\min}(B) \leqslant \lambda_{\min}(A + B)\,.$$

Applying this to $A = M_t$ and $B = \bar{V}_{t+1}^0$ yields $\lambda_{\min}(M_t) + \lambda_{\min}\left(\bar{V}_{t+1}^0\right) \leqslant \lambda_{\min}\left(V_{t+1}^0\right)$. Now, notice that $\lambda_{\min}(A) = -\lambda_{\max}(A) \geqslant -\|A\|$ where $\|\cdot\|$ is the matrix norm induced by the scalar product $\langle A, B\rangle = \mathrm{Tr}\left(A^\top B\right)$ (also known as the Frobenius norm). Moreover, the conditional covariance lower bound of Assumption 5 and another application of Weyl's inequality imply that

$$\lambda_{\min}\left(\bar{V}_{t+1}^0\right) \geqslant \sum_{s=1}^{t} \lambda_{\min}\left(\mathbb{E}\left[X_s X_s^\top | \mathcal{F}_{s-1}\right]\right) \geqslant \rho_{\mathcal{X}} L^2 t\,.$$

Combining these together, we obtain that for arbitrary $a \in \mathbb{R}$ and $b > 0$, the following inequality holds:

$$\mathbb{P}\left(\exists t \in \mathbb{N},\ \lambda_{\min}\left(H_{t+1}^\beta(\theta_{t+1})\right) \leqslant a + bt\right) \leqslant \mathbb{P}\left(\exists t \in \mathbb{N},\ \|M_t\| \geqslant \frac{\beta - a}{m} + \left(\rho_{\mathcal{X}} L^2 - \frac{b}{m}\right) t\right)\,. \tag{1}$$

Notice that $\|M_{t+1} - M_t\| = \|X_t X_t^\top - \mathbb{E}\left[X_t X_t^\top | \mathcal{F}_{t-1}\right]\| \leqslant 2L^2$ (Assumption 4). Now if we choose $b = \frac{1}{2} m\rho_{\mathcal{X}} L^2$ and $a = \beta - \frac{4mL^2}{\rho_{\mathcal{X}}} \log \frac{2}{\delta}$, the bound from Proposition 4 holds with $\eta = \frac{\rho_{\mathcal{X}}}{8L^2}$ and $\delta \in (0, 1)$ and $c_t = 2L^2$, thus proving the result.

■

We are now ready to prove the stochastic elliptic potential lemma.

Proof of Lemma 5 We fix $t_0 \in \mathbb{N}$ arbitrarily for now and let $T \geqslant t_0$. We start by splitting the sum in two and by applying the deterministic elliptic potential lemma (Lemma 8) up to time t_0:

$$\begin{aligned}
\sum_{t=1}^{T} \|X_t\|_{H_t^\beta(\theta_t)^{-1}} &= \sum_{t=1}^{t_0} \|X_t\|_{H_t^\beta(\theta_t)^{-1}} + \sum_{t=t_0+1}^{T} \|X_t\|_{H_t^\beta(\theta_t)^{-1}} \\
&\leqslant \frac{1}{\sqrt{m}} \sum_{t=1}^{t_0} \|X_t\|_{\left(V_t^{\beta/m}\right)^{-1}} + \sum_{t=t_0+1}^{T} \|X_t\|_{H_t^\beta(\theta_t)^{-1}} \quad \text{(Assumption 1)} \\
&\leqslant \left(\frac{1}{\sqrt{m}}, \frac{L}{\sqrt{\beta}}\right) \sqrt{2t_0 d \log\left(1 + \frac{mL^2 t_0}{d\beta}\right)} + \sum_{t=t_0}^{T-1} \|X_{t+1}\|_{H_{t+1}^\beta(\theta_{t+1})^{-1}} \quad \text{(Assumption 4)}\,.
\end{aligned}$$

Now let $\mathcal{E}_t^\delta = \left\{\forall t' \geqslant t,\ \lambda_{\min}\left(H_{t+1}^\beta(\theta_{t+1})\right) > \beta - \frac{4mL^2}{\rho_{\mathcal{X}}} \log\frac{2}{\delta} + \frac{m\rho_{\mathcal{X}}L^2}{2} t\right\}$. It is clear that $\mathcal{E}_{t_0}^\delta \subseteq \mathcal{E}_0^\delta$, and thus by Lemma 6, $\mathbb{P}\left(\mathcal{E}_{t_0}^\delta\right) \geqslant 1 - \delta$. The choice $t_0 = \left\lceil \frac{8}{\rho_{\mathcal{X}}^2} \log\frac{2}{\delta} - \frac{2\beta}{m\rho_{\mathcal{X}}L^2} \right\rceil$ implies that the right-hand side in the definition of $\mathcal{E}_{t_0}$ is positive. On this event, we bound the second sum as follows:

$$\begin{aligned}
\sum_{t=t_0}^{T-1} \|X_{t+1}\|_{H_{t+1}^\beta(\theta_{t+1})^{-1}} &\leqslant L \sum_{t=t_0}^{T-1} \frac{1}{\sqrt{a+bt}} \\
&\leqslant \frac{2L}{b}\left(\sqrt{a+b(T-1)} - \sqrt{a+b(t_0-1)}\right) \\
&\leqslant \frac{2L}{b}\left(\sqrt{bT} + \sqrt{a} - \sqrt{a+b(t_0-1)}\right)\,,
\end{aligned}$$

where we use the shorthand $a = \beta - \frac{4mL^2}{\rho_{\mathcal{X}}} \log\frac{2}{\delta}$ and $b = \frac{1}{2} m\rho_{\mathcal{X}} L^2$ (the penultimate line comes from sum-integral comparison while the last one follows from the inequality $\sqrt{x+y} \leqslant \sqrt{x} + \sqrt{y}$). After collecting the dominating term in $\sqrt{T}$, the two sums give the following upper bound:

$$\sum_{t=1}^{T} \|X_t\|_{H_t^\beta(\theta_t)^{-1}} \leqslant 2L\sqrt{\frac{T}{b}}\left(1 + \frac{C}{\sqrt{T}}\right)$$

with

$$C = \sqrt{\frac{a}{b}} - \sqrt{\frac{a+b(t_0-1)}{b}} + \frac{\sqrt{b}}{2L}\left(\frac{1}{\sqrt{m}}, \frac{L}{\sqrt{\beta}}\right)\sqrt{2t_0 d \log\left(1 + \frac{mL^2 t_0}{d\beta}\right)}\,.$$

Substituting a and b with their expressions yields the result.

■

We finally prove the regret bound in the stochastic i.i.d. actions setting.

Proof of Theorem 4 We follow the exact same steps as with Theorem 1 in order to bound the regret by

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

with probability at least $1 - \delta$. The sum on the left-hand side is controlled by Lemma 5 also with probability at least $1 - \delta$. A simple union argument over both events concludes the proof, resulting in a regret upper bound with probability at least $1 - 2\delta$.

■

We conclude this section with two remarks.

Remark 9 (Dependency of $\rho_{\mathcal{X}}$ on d) *We recall that in the stochastic actions setting (Assumption 5), $\rho_{\mathcal{X}}$ is a lower bound on the conditional covariance of actions, which can be equivalently formulated as $\rho_{\mathcal{X}} L^2 \leqslant \lambda_{\min}\left(\mathbb{E}\left[X_t X_t^\top | \mathcal{F}_{t-1}\right]\right)$ for*

all $t \in \mathbb{N}$. Following the argument of Kim et al. (2021) in the case of unconditional covariance control, we obtain the following bound on $\rho_{\mathcal{X}}$:

$$d\rho_{\mathcal{X}} L^2 \leqslant d\lambda_{\min}\left(\mathbb{E}\left[X_t X_t^\top | \mathcal{F}_{t-1}\right]\right) \leqslant \sum_{\lambda \in \mathcal{S}_t} \lambda = \mathrm{Tr}\left(\mathbb{E}\left[X_t X_t^\top | \mathcal{F}_{t-1}\right]\right) = \mathbb{E}\left[\mathrm{Tr}\left(X_t X_t^\top\right) | \mathcal{F}_{t-1}\right] \leqslant L^2\,,$$

where $\mathcal{S}_t$ denotes the spectrum of the symmetric matrix $\mathbb{E}\left[X_t X_t^\top | \mathcal{F}_{t-1}\right]$ and Tr *the trace operator. Therefore $\rho_{\mathcal{X}} \leqslant d^{-1}$. Moreover, Bastani et al. (2021); Kim et al. (2021, 2022) identified two families of examples where $\rho_{\mathcal{X}} = \mathcal{O}(d^{-1})$ in the unconditional case:*

- *If the distribution of $X \in \mathcal{X}_t$ (marginal distribution of a each action) admits a density p with respect to the Lebesgue measure supported in $\mathcal{B}^d_{\|\cdot\|_2}(0, L)$ and such that $p(x) \geqslant p_{\min} > 0$ for all $x \in \mathcal{B}^d_{\|\cdot\|_2}(0, L)$, then $\rho_{\mathcal{X}} = \frac{p_{\min}}{(d+2)} \mathrm{vol}\left(\mathcal{B}^d_{\|\cdot\|_2}(0,1)\right)$ is a suitable choice (Kim et al., 2022, Lemma C.1). In general, the volume of the Euclidean unit ball in $\mathbb{R}^d$ is $\mathrm{vol}\left(\mathcal{B}^d_{\|\cdot\|_2}(0,1)\right) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} \sim \frac{1}{\sqrt{d\pi}}\left(\frac{2\pi e}{d}\right)^{\frac{d}{2}}$, which goes to 0 when $d \to +\infty$. In certain cases though, such as the uniform and truncated Gaussian distributions, $p_{\min}$ is proportional to $\mathrm{vol}\left(\mathcal{B}^d_{\|\cdot\|_2}(0,1)\right)$, thus leading to $\rho_{\mathcal{X}} = \mathcal{O}(d^{-1})$.*
- *If the covariance matrix $\mathbb{E}\left[XX^\top\right]$ exhibits a certain structure, for instance AR(1), tridiagonal or block diagonal, then $\rho_{\mathcal{X}} = \mathcal{O}(d^{-1})$, regardless of the marginal distributions.*

Remark 10 (Previous results about the growth of the smallest eigenvalue) *Kim et al. (2021, 2022) prove similar results on the linear growth of the smallest eigenvalues of a different sequence of Hessian matrices. More precisely, they consider a fixed number K of arms, i.e., action sets of the form $\mathcal{X}_t = \{X_{k,t}, k \in [K]\}$ and Hessian matrices constructed from all actions $V_{t+1}^{[K]} = \sum_{k\in[K]} \sum_{s=1}^t X_{k,s} X_{k,s}^\top$, instead of using only the actions played at previous time steps. This is made possible in their analyses by resorting to a doubly robust imputation of unobserved rewards associated to unplayed actions, which is significantly different from our approach. One theoretical benefit of their method is that the sequence $(V_{t+1}^{[K]})_{t\in\mathbb{N}}$ can be more easily transformed into a $\mathcal{F}$-martingale using only the unconditional lower bound on the covariance, as opposed to the conditional one of Assumption 5.*

Of note, Li et al. (2017) also questions the feasibility of a linear lower bound on the smallest eigenvalue of V_{t+1}^0 but concludes that it requires more stringent assumption as Lai and Wei (1982, Example 1) seemingly provides a counterexample of sublinear growth in the context of a regression problem. However, this counterexample studies autoregressive actions instead of i.i.d. action sets, which leads to $\mathbb{E}\left[X_t X_t^\top | \mathcal{F}_{t-1}\right] \to 0$ when $t \to +\infty$. Therefore, this is different from what we consider in Assumption 5 and does not invalidate our analysis.

G PROOFS OF OGD CONCENTRATION

We reformulate below Proposition 2 in full details and provide a proof of the OGD regret.

Proposition 5 (OGD Regret, 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 toward the more general setting of sub-Gaussian gradients.

Let $G > 0$ be such that $\|\nabla \ell_j(z_j)\| \leqslant G$ for all $j = 1, \dots, n$. 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:

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

■

H PROOFS OF THEOREM 3

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 3 (Regret of LinUCB-OGD-CR) *Let $\varepsilon_h > 0$ and $h = \lceil \frac{2\varepsilon_h}{\rho_{\mathcal{X}} L^2} + \frac{8}{\rho_{\mathcal{X}}^2} \log \frac{2}{\delta} \rceil$. Assume that $\partial \mathcal{L}(Y_t, \langle \theta^*, X_t \rangle)$ is $\sqrt{m}\sigma$-sub-Gaussian for all $t \leqslant T$. Under Assumptions 1-2-3-4-5, there exists constants $C, C' > 0$ such that with probability at least $1 - (1 + T/h)\delta$ the regret of Algorithm 2 with exploration bonus sequence*

$$\gamma_{t,T}^{OGD} \colon x \in \mathcal{X}_t \mapsto (c_t^\delta + c_{t,T}^{OGD,\delta}) \|x\|_{H_t^{\kappa\alpha}(\bar{\theta}^{OGD}_{\lfloor \frac{t-1}{h} \rfloor})^{-1}} ,$$

$$c_{t,T}^{OGD,\delta} = \sqrt{\left(L^2 + \frac{\alpha}{mMt}\right)\left(\frac{2\kappa C' d h^2 \sigma^2}{\varepsilon_h^2} \log\left(\frac{2dT}{h\delta}\right) \log\left(\frac{t}{h}\right)\right)},$$

and the OGD step sequence of Proposition 2 satisfies

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

Proof Let $\ell_j(\theta) = \sum_{k=1}^{h} \mathcal{L}(Y_{(j-1)h+k}, \langle \theta, X_{(j-1)h+k} \rangle) + \frac{\alpha}{2N} \|\theta\|_2^2$, where $N = \lceil \frac{T-1}{h} \rceil$ denotes the total number of episodes of length h. 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 $\sum_{j=1}^{n} \nabla \ell_j(\bar{\theta}_t) = 0$ for $n = \frac{t-1}{h}$ (i.e., after episode n, when $\widehat{\theta}_n^{\mathrm{OGD}}$ 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 1), 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 2 and define:

$$\begin{aligned} z_j &= \widehat{\theta}_j^{\mathrm{OGD}} , \\ \bar{z}_n &= \bar{\theta}_t , \\ z^* &= \theta^* . \end{aligned}$$

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

Bound on $\|\tilde{z}_n - \bar{z}_n\|_2$ Without loss of generality, we assume here that $\partial \mathcal{L}^*$ is a $\sqrt{m}\sigma$ sub-Gaussian process (this follows in variety of settings from the discussion of Assumption 2 and Lemma 1; the conclusions are essentially unchanged when assuming only Assumption 2, at the cost of slightly heavier notations).

We first note that $\nabla \ell_j(\theta^*) = g_j(\theta^*) + \frac{\alpha}{N}\theta^*$, where $g_j(\theta^*) = \sum_{k=1}^{h} \partial \mathcal{L}^*_{(j-1)h+k}$ and $j \in [N]$, is $\sqrt{hm}\sigma$-sub-Gaussian (sum of h random variables, each of them being drawn from a $\sqrt{m}\sigma$-sub-Gaussian distribution). Setting the episode length to $h = \lceil \frac{2\varepsilon_h}{\rho_{\mathcal{X}} L^2} + \frac{8}{\rho_{\mathcal{X}}^2} \log \frac{2}{\delta} \rceil$ makes the one-step losses ℓ_j $m\varepsilon_h$-strongly convex with high probability. Indeed, let us define for $h' \in \mathbb{N}$ the function $f(h') = \frac{\rho_{\mathcal{X}} L^2}{2} h' - \frac{4L^2}{\rho_{\mathcal{X}}} \log \frac{2}{\delta}$ and the event $\mathcal{E}_{j,h'} = \left\{ \lambda_{\min}\left(\sum_{k=1}^{h} X_{(j-1)h'+k} X_{(j-1)h'+k}^\top \right) > f(h') \right\}$.

First, notice that $\mathcal{E}_{j,h} \supseteq \bigcap_{h' \in \mathbb{N}} \mathcal{E}_{j,h'}$ and that $\sum_{k=1}^{h} X_{(j-1)h'+k} X_{(j-1)h'+k}^\top$ has the same distribution as $V_{h'+1}^0$ by Assumption 5. We deduce from Lemma 6 applied to V_{h+1}^0 (that is with $\beta = 0$ and $m = 1$), that

$$\mathbb{P}(\mathcal{E}_{j,h}) \geqslant \mathbb{P}\Big(\bigcap_{h' \in \mathbb{N}} \mathcal{E}_{j,h'}\Big) = \mathbb{P}\left(\forall h' \in \mathbb{N},\ \lambda_{\min}\left(V_{h'+1}^0\right) > -\frac{4L^2}{\rho_{\mathcal{X}}} \log \frac{2}{\delta} + \frac{\rho_{\mathcal{X}} L^2}{2} h' \right) \geqslant 1 - \delta .$$

In particular for the value of h defined above, we have $\mathcal{E}_{j,h} \subseteq \left\{\lambda_{\min}\left(\sum_{k=1}^{h} X_{(j-1)h'+k}X_{(j-1)h'+k}^\top\right) \geqslant \varepsilon_h\right\}$, which gives the $m\varepsilon_h$-strong convexity of ℓ_j by the usual minoration $\partial^2\mathcal{L} \geqslant m$ (Assumption 1). In the rest of this proof, we assume to be on the event $\bigcap_{j\in[N]} \mathcal{E}_{j,h}$, the probability of which is at least $1 - N\delta$ by a simple union argument.

Now, we apply the bound on the OGD regret of Proposition 5 with $a = m\varepsilon_h$, $A = hML^2$, namely that the *good event*

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

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). We assume to be on this event in the rest of the proof, which we combine to the previous events with a union argument, leading to a probability ot at least $1 - (N+1)\delta$.

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:

$$\ell_j(\tilde{z}_n) - \ell_j(\bar{z}_n) \geqslant \langle \nabla\ell_j(\bar{z}_n), \tilde{z}_n - \bar{z}_n\rangle + \frac{m\varepsilon_h}{2}\|\tilde{z}_n - \bar{z}_n\|_2^2\,.$$

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

$$\|\tilde{z}_n - \bar{z}_n\|_2^2 \leqslant \frac{2Cdh\sigma^2}{m\varepsilon_h^2 n} \log\left(\frac{2dN}{\delta}\right)\log(n)\,.$$

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

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

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

$$\begin{aligned}
\|\theta^* - \bar{\theta}_n^{\mathrm{OGD}}\|_{\bar{H}_t^\alpha(\theta^*, \bar{\theta}_n^{\mathrm{OGD}})} &\leqslant \|\theta^* - \bar{\theta}_t\|_{\bar{H}_t^\alpha(\bar{\theta}_t, \bar{\theta}_n^{\mathrm{OGD}})} + \|\bar{\theta}_n^{\mathrm{OGD}} - \bar{\theta}_t\|_{\bar{H}_t^\alpha(\bar{\theta}_t, \bar{\theta}_n^{\mathrm{OGD}})} \\
&\leqslant \|\theta^* - \bar{\theta}_t\|_{\bar{H}_t^\alpha(\theta^*, \bar{\theta}_n^{\mathrm{OGD}})} + \sqrt{M}\|\bar{\theta}_n^{\mathrm{OGD}} - \bar{\theta}_t\|_{V_t^{\frac{\alpha}{m}}} \\
&\leqslant \|\theta^* - \bar{\theta}_t\|_{\bar{H}_t^\alpha(\theta^*, \bar{\theta}_n^{\mathrm{OGD}})} + \sqrt{M\left(L^2 t + \frac{\alpha}{m}\right)}\|\bar{\theta}_n^{\mathrm{OGD}} - \bar{\theta}_t\|_2\,.
\end{aligned}$$

The first term can be controlled by transportation of local metrics in the same way as in the proof of Theorem 1, i.e.,

$$\|\theta^* - \bar{\theta}_t\|_{\bar{H}_t^\alpha(\theta^*, \bar{\theta}_n^{\mathrm{OGD}})} \leqslant \sqrt{\kappa}\left(\|F_t^\alpha(\theta^*) - F_t^\alpha(\widehat{\theta}_t)\|_{H_t^\beta(\theta^*)^{-1}} + \|F_t^\alpha(\bar{\theta}_t) - F_t^\alpha(\widehat{\theta})\|_{H_t^\beta(\bar{\theta}_t)^{-1}}\right)\,,$$

and thus this term adds the same contribution to the design of the exploration bonus sequence and to the regret bound. For the second term, we apply the previous bound on $\|\tilde{z}_n - \bar{z}_n\|_2 = \|\bar{\theta}_n^{\mathrm{OGD}} - \bar{\theta}_t\|_2$. Combining these two inequalities results in the following control:

$$\|\theta^* - \bar{\theta}_n^{\mathrm{OGD}}\|_{\bar{H}_t^\alpha(\theta^*, \bar{\theta}_n^{\mathrm{OGD}})} \leqslant \underbrace{c_t^\delta}_{\substack{\text{same as in} \\ \text{the proof of} \\ \text{Theorem 1}}} + \underbrace{\sqrt{\left(L^2 + \frac{\alpha}{mMt}\right)\left(\frac{2\kappa Cdh^2\sigma^2}{\varepsilon_h^2}\log\left(\frac{2dT}{h\delta}\right)\log\left(\frac{t}{h}\right)\right)}}_{=:\, c_{t,T}^{\mathrm{OGD},\delta} = \mathcal{O}\left(\sigma L\sqrt{\kappa d}\log T\right)}\,.$$

The rest of the proof is now identical to that of Theorem 4, i.e., we use the exploration bonus sequence

$$\gamma_{t,T}^{\mathrm{OGD}} \colon x \in \mathcal{X}_t \mapsto (c_t^\delta + c_{t,T}^{\mathrm{OGD},\delta}) \|x\|_{H_t^{\kappa\alpha}(\bar{\theta}^{\mathrm{OGD}}_{\lfloor \frac{t-1}{h} \rfloor})^{-1}} \,,$$

and control $\sum_{t=1}^{T} \|X_t\|_{H_t^{\kappa\alpha}(\bar{\theta}^{\mathrm{OGD}}_{\lfloor \frac{t-1}{h} \rfloor})^{-1}}$ using an elliptical potential lemma (since we operate under Assumption 5 for the strong convexity of the episodic losses ℓ_j, we use the strong regret guarantee provided by Lemma 5; note that the high probability event on which this lemma applies is already included in the above events, for a total probability of at least $1-(N+1)\delta$).

■

Remark 11 (The importance of strong convexity) *The key argument behind the proof of Theorem 3 is that the aggregated loss over an episode $\ell_n \colon \theta \in \Theta \mapsto \sum_{k=1}^{h} \mathcal{L}(Y_{(n-1)h+k}, \langle \theta, X_{(n-1)h+k} \rangle) + \frac{\alpha}{2N}\|\theta\|_2^2$ is $m\varepsilon_h$ strongly convex. With simple convexity only, the online regret guarantee of the OGD approximation scales like $\mathcal{O}(\sqrt{T})$ instead of logarithmically. This would only guarantee $\|\bar{\theta}_n^{OGD} - \bar{\theta}_t\|_2 = \mathcal{O}(\sqrt{t})$, resulting in linear $\mathcal{O}(T)$ bandit regret after multiplying this term with the contribution of the elliptic potential lemma. Moreover, although ℓ_n is always trivially at least $\frac{\alpha}{N}$-strongly convex, it is necessary to ensure non-vanishing strong convexity when $T \to +\infty$ (we recall that $N = \lceil \frac{T-1}{h} \rceil$). Indeed, substituting ε_h with $\frac{\alpha}{N}$ in the regret bound above gives $\mathcal{R}_T \leqslant \mathcal{O}(\varepsilon_h^{-1}) = \mathcal{O}(T)$.*

Scaling of episode length h As shown in the proof, non-vanishing strong convexity of ℓ_n can be deduced from a fixed lower bound on the smallest eigenvalue of the Hessian of ℓ_n. By Lemma 6, this holds with high probability provided the episode length h is high enough, which translates to $h = \lceil \frac{2\varepsilon_h}{\rho_{\mathcal{X}} L^2} + \frac{8}{\rho_{\mathcal{X}}^2} \log \frac{2}{\delta} \rceil$. Using the typical bound $\rho_{\mathcal{X}} = \mathcal{O}(d^{-1})$, we see that h scales like $\mathcal{O}(d^2)$ in the action dimension. By comparison, the only similar OGD scheme for generalized linear bandit scales like $\mathcal{O}(d^3)$ (Ding et al., 2021, Lemma 2 and Remark 2), thus suffering to a greater extent from the curse of dimensionality. Note the practical tradeoff on h faced by the agent running Algorithm 2: the higher h, the more likely it is that the OGD estimator $\bar{\theta}^{\mathrm{OGD}}$ well approximates the true empirical risk minimizer $\bar{\theta}$ (because of stronger convexity of the episodic losses); however, it also means longer episodes and thus less frequent updates of $\bar{\theta}^{\mathrm{OGD}}$, i.e., less learning.

Note that the value of h is derived from a concentration bound that is *uniform* in h (Lemma 6). However, since h is kept constant throughout the run of the algorithm, a similar, non-uniform result would actually be sufficient (we chose to use Lemma 6 mainly for the sake of convenience, since we already assumed to be on the corresponding good event in order to mirror the regret analysis of Theorem 4). It is actually possible to tighten the lower bound on h using a finer, non-uniform concentration result, which we state below.

Lemma 7 (Tighter bound on the episode length h) *Under Assumptions 4 and 5, let $\varepsilon_h > 0$, $\delta \in (0,1)$ and define*

$$h = \left\lceil \frac{1}{4\rho_{\mathcal{X}}^2} \left(\sqrt{2(1+\gamma_\delta) \log\left(\frac{2}{\delta}\sqrt{1+\frac{1}{\gamma_\delta}} \right)} + \sqrt{\sqrt{2(1+\gamma_\delta) \log\left(\frac{2}{\delta}\sqrt{1+\frac{1}{\gamma_\delta}} \right)} + \frac{\rho_{\mathcal{X}}\varepsilon_h}{L^2}} \right)^2 \right\rceil \,,$$

where $\gamma_\delta = \frac{-1}{1+W_{-1}(-\frac{\delta^2}{4e})}$ and W_{-1} is the first lower branch of the Lambert W function, i.e., the smallest real solution for $z = [-\frac{1}{e}, 0)$ of the equation $W_{-1}(z)e^{W_{-1}(z)} = z$. It holds that

$$\mathbb{P}\left(\lambda_{\min}\left(V_{h+1}^0 \right) \leqslant \varepsilon_h \right) \leqslant \delta \,.$$

Proof The idea is similar to that of the proof of Lemma 6, i.e., relate the deviation of the smallest eigenvalue to that of a matrix martingale. The difference lies in the choice of the concentration bound for this martingale. Fix $\eta > 0$, Howard et al. (2021, Corollary S1(a)) with a normal mixture bound shows that a martingale $(M_t)_{t\in\mathbb{N}}$ taking values in a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ with uniformly bounded increments $\|M_{t+1} - M_t\| \leqslant c$ for some $c > 0$ and all $t \in \mathbb{N}$ satisfies

$$\mathbb{P}\left(\exists t \in \mathbb{N},\ \|M_t\| \geqslant c\sqrt{2(t+\eta)\log\left(\frac{2}{\delta}\sqrt{1+\frac{t}{\eta}} \right)} \right) \leqslant \delta \,.$$

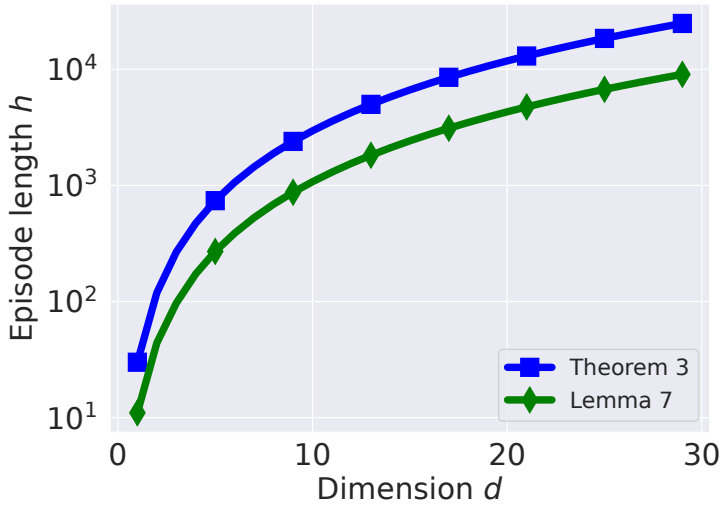

Figure 2: Recommended episode length $h = \lceil \frac{2\varepsilon_h}{\rho_{\mathcal{X}} L^2} + \frac{8}{\rho_{\mathcal{X}}^2} \log \frac{2}{\delta} \rceil$ for LinUCB-OGD-CR (Algorithm 2) according to Theorem 3 and Lemma 7 as a function of $d = \rho_{\mathcal{X}}^{-1}$, for typical values $L = 1$, $\varepsilon_h = 0.1$ and $\delta = 5\%$.

Now, we fix $h \in \mathbb{N}$. Although this bound is uniform in $t \in \mathbb{N}$, we can make use of the free parameter η to optimize it for $t = h$. This procedure is standard (see e.g., Howard et al. (2021, Proposition 3)) and yields $\eta = \gamma_\delta h$ with $\gamma_\delta = \frac{-1}{1+W_{-1}(-\frac{\delta^2}{4e})}$. In particular, we have

$$\mathbb{P}\left(\|M_h\| \geqslant c\sqrt{2h(1+\gamma_\delta)\log\left(\frac{2}{\delta}\sqrt{1+\frac{1}{\gamma_\delta}}\right)} \right) \leqslant \delta\,.$$

Following the steps of Lemma 6, we have that

$$\mathbb{P}\left(\lambda_{\min}\left(V_{h+1}^0\right) \leqslant \varepsilon_h\right) \leqslant \mathbb{P}\left(\|M_h\| \geqslant \rho_{\mathcal{X}} L^2 h - \varepsilon_h\right) \leqslant \delta\,,$$

with $M_h = V_{h+1}^0 - \mathbb{E}\left[V_{h+1}^0\right]$, which is a martingale with increments bounded by $c = 2L^2$. Equating both bounds on $\|M_h\|$ yields

$$\rho_{\mathcal{X}} L^2 h - 2L^2\sqrt{2h(1+\gamma_\delta)\log\left(\frac{2}{\delta}\sqrt{1+\frac{1}{\gamma_\delta}}\right)} - \varepsilon_h = 0\,.$$

This expression is a quadratic equation in $H = \sqrt{h}$. Let $a = \rho_{\mathcal{X}} L^2$ and $b = L^2\sqrt{2h(1+\gamma_\delta)\log\left(\frac{2}{\delta}\sqrt{1+\frac{1}{\gamma_\delta}}\right)}$. Notice that both a and b are positive and that the discriminant of $aH^2 + 2bH - \varepsilon_h = 0$ is $4b^2 + 4a\varepsilon_h$, which is also positive. The only positive solution is thus given by $\sqrt{h} = H = \frac{b+2\sqrt{b^2+a\varepsilon_h}}{2a}$, which concludes the proof.

■

The value of h recommended by Lemma 7 scales with $\rho_{\mathcal{X}}^{-2}$ at the first order, and thus $h = \mathcal{O}(d^2)$, just like in Theorem 4. However, it is typically smaller, and thus more practical (allows for more frequent OGD updates with the same theoretical guarantees). We report in Figure 2 the numerical values for both expressions of h for typical choices of the parameters L, ε_h and δ as a function the dimension d. In practice, even smaller values of h may ensure enough convexity of the episodic losses to observe sublinear regret, which we empirically witnessed in the experiments (see Appendix I). For the practitioner, h may be viewed as a hyperparameter to be tuned manually, potentially on an instance-dependent basis, with Theorem 4 and Lemma 7 giving only worst-case guarantees.

I 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 $\nu = \mathcal{N}\left(\mu, \sigma^2\right)$ 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 $\psi''(z) = 2(1-p)\mathbb{I}_{z<0} + 2p\mathbb{I}_{z>0}$, which is bounded between $2p$ and $2(1-p)$. In particular, Assumption 1 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 log-concave, which fits the scope of the supermartingale control of Lemma 1.

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 Lemma 1.

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 $p = 10\%$, 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 2, 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 3 reproduces the results of each experiments across 500 independent replications. Finally, average runtimes for each algorithm are reported in Table 4. 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 4: 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-CR (convex risk)	231.0 ± 21.7	814.8 ± 88.3	519.1 ± 33.3
LinUCB-OGD-CR (convex risk)	20.4 ± 3.9	60.2 ± 12.0	25.7 ± 4.9

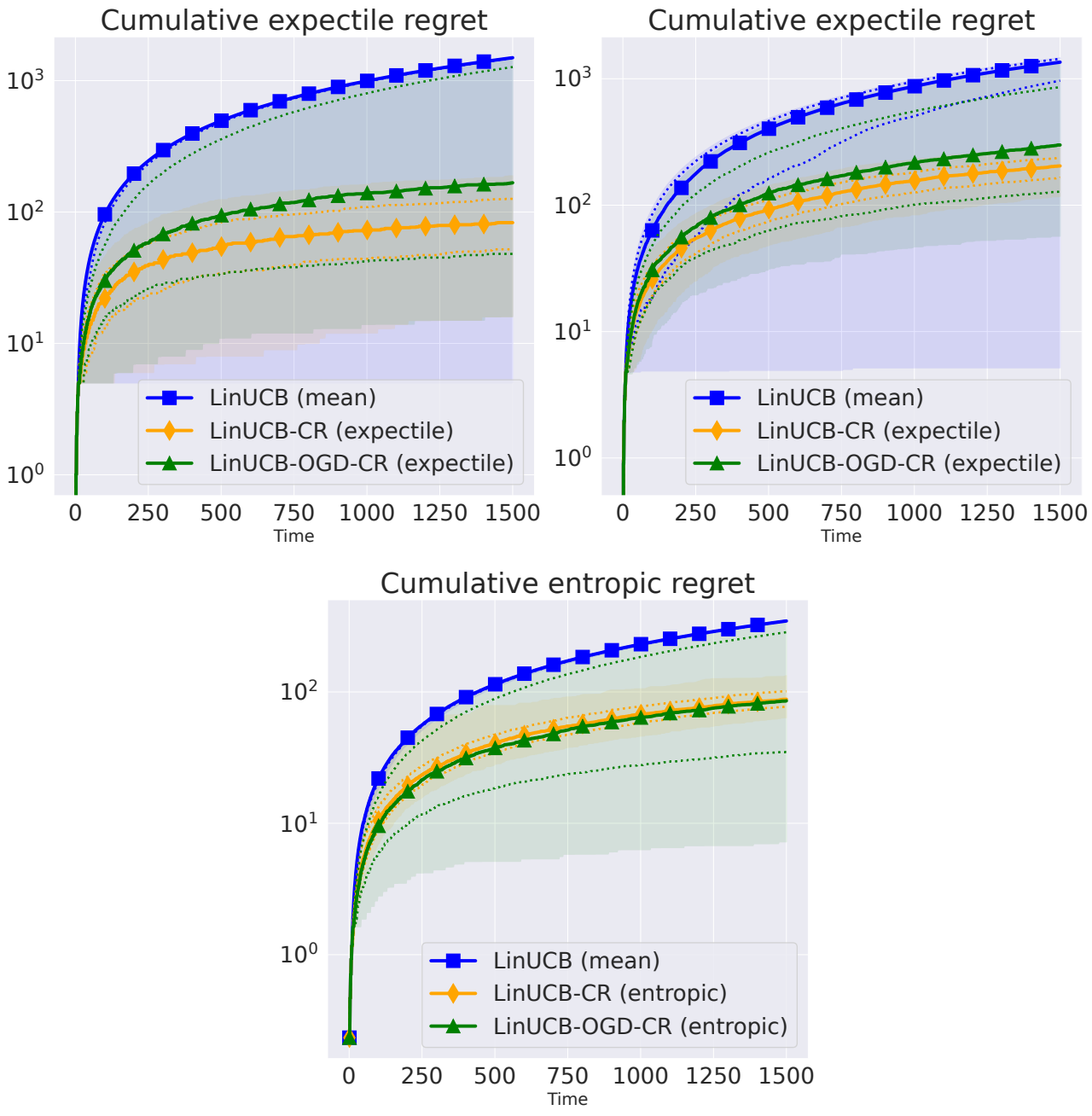

Figure 3: Left: two-armed Gaussian expectile bandit. Center: two-armed linear expectile bandit with $\mathbb{R}^3$ contexts and expectile-based asymmetric noises. Right: two-armed Bernoulli entropic risk bandit. Thick lines denote median cumulative regret over 500 independent replications. Dotted lines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